A Benchmark to Test Generalization Capabilities of Deep Learning Methods to Classify Severe Convective Storms in a Changing Climate

Maria J. Molina¹, David John Gagne¹, and Andreas F. Prein¹

¹National Center for Atmospheric Research, Boulder, CO, USA

Abstract This is a test case study assessing the ability of deep learning methods to generalize to a future climate (end of 21st century) when trained to classify thunderstorms in model output representative of the present-day climate. A convolutional neural network (CNN) was trained to classify strongly rotating thunderstorms from a current climate created using the Weather Research and Forecasting model at high-resolution, then evaluated against thunderstorms from a future climate and found to perform with skill and comparatively in both climates. Despite training with labels derived from a threshold value of a severe thunderstorm diagnostic (updraft helicity), which was not used as an input attribute, the CNN learned physical characteristics of organized convection and environments that are not captured by the diagnostic heuristic. Physical features were not prescribed but rather learned from the data, such as the importance of dry air at mid-levels for intense thunderstorm development when low-level moisture is present (i.e., convective available potential energy). Explanation techniques also revealed that thunderstorms classified as strongly rotating are associated with learned rotation signatures. Results show that the creation of synthetic data with ground truth is a viable alternative to human-labeled data and that a CNN is able to generalize a target using learned features that would be difficult to encode due to spatial complexity. Most importantly, results from this study show that deep learning is capable of generalizing to future climate extremes and can exhibit out-of-sample robustness with hyperparameter tuning in certain applications.

Plain Language Summary As temperatures and water vapor continue increasing due to climate change, models that were trained using past data may no longer perform with skill. Here, we explored whether the performance of a machine learning model was sensitive to a changing climate. The purpose of the machine learning model was to classify thunderstorms that were created using a high-resolution numerical model into two groups: potentially severe thunderstorms and potentially non-severe thunderstorms. Potentially severe thunderstorms were of interest because they have a greater likelihood of producing tornadoes and large hail, which cause billions of losses and dozens of fatalities every year. Results show that the machine learning model was able to classify thunderstorms with skill in both the present day and future climates partly due to the architecture of the machine learning model. We also explored the reasons behind the machine learning model’s skill and found that it was able to learn thunderstorm characteristics and weather information from data. These results provide us with added confidence that machine learning models can learn physical relationships from weather and climate data and perform with skill in a changing climate in certain applications.

1. Introduction

The recent success of convolutional neural networks (CNNs; Fukushima & Miyake, 1982) in Earth science applications is largely due to their ability to capture nonlinear and translation invariant details among input variables. This class of deep learning models (LeCun et al., 2015) has proven skillful in various atmospheric science tasks, including the detection of weather and climate features (Biard & Kunkel, 2015; Lagerquist et al., 2019; Y. Liu et al., 2016; Toms et al., 2019), emulation of complex model processes (Rasp et al., 2018), and prediction of extreme weather and climate phenomena (Gagne II et al., 2019; Ham et al., 2019; Jorgensen et al., 2020; Lagerquist et al., 2020; Sobash et al., 2020; Zhou et al., 2019). This study focuses on convection over the central and eastern contiguous United States (CONUS), which at extremes can produce severe hazards (e.g., hail and tornadoes) that pose societal danger. CNNs have already shown skill in

© 2021, The Authors. Earth and Space Science published by Wiley Periodicals LLC on behalf of American Geophysical Union. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
classification and prediction of convective storms in the present climate (Gagne II et al., 2019), modeled using the Weather Research and Forecasting model (WRF; Skamarock & Klemp, 2008) at high-resolution (4 km). However, as the climate continues to warm, some future thunderstorms may be outliers in the baseline climate used for training (Trapp & Hoogewind, 2016), and these extreme events may be more difficult for CNNs to identify. This article explores the ability of CNNs to classify convection of a future climate modeled with WRF, along with the physical reasons for the resultant performance.

Climate change is altering the large-scale atmospheric landscape over North America, resulting in changes to the frequency and intensity of organized convection (Prein et al., 2017; K. L. Rasmussen et al., 2017). Future changes to thermodynamic and kinematic fields can impact climatological distributions of convection morphology and associated severe hazards (e.g., tornadoes and large hail; Diffenbaugh et al., 2013; Trapp et al., 2007, 2009). Studies have shown a climate change imprint on various aspects of severe thunderstorms and associated environments (Allen, 2018), including increases in thermodynamic buoyancy and thunderstorm frequency (Brooks, 2013; Hoogewind et al., 2017), increases in convective inhibition (Taszarek et al., 2020), more societal exposure (Ashley & Strader, 2016), and an eastward geographic shift of environments over the United States favorable for severe hazards (Gensini & Brooks, 2018). However, discerning the interplay between thermodynamic and kinematic components on future convection has been more challenging (Brooks, 2013), given that subtle changes to either field can alter the potential of a thunderstorm to produce severe hazards (Doswell et al., 1996). This complex interplay, and varying seasonal and geographical trends, limit the broader conclusions that can be derived from climate studies of severe convective storms.

In current forecasting applications, advancements in delineating thunderstorms capable of producing specific hazards have included the development of environmental proxies and composite indices that take kinematic and thermodynamic factors into account (Gropp & Davenport, 2018; E. N. Rasmussen, 2003; R. L. Thompson et al., 2003, 2007, 2012). Updraft helicity (UH) is an example of a diagnostic parameter, which estimates the magnitude of rotation within a thunderstorm's updraft using vertical wind speeds and vorticity (Kain et al., 2008). Strongly rotating thunderstorms with high magnitudes of UH (e.g., ≥75 m s⁻²) have a greater likelihood to be of supercell morphology (Clark et al., 2013; Sobash et al., 2016), a type of thunderstorm that observations have shown to be more likely to produce severe hazards (Bunkers et al., 2006; Duda & Gallus Jr, 2010). Scalar thresholds for UH have been used to classify model-simulated convection, with thunderstorms that exceed the predetermined threshold classified as severe (Molina, Allen, & Prein, 2020; Sohah et al., 2011). These dichotomous assignments derived from UH have been used in kilometer-scale climate simulations to estimate changes to severe hazards in a future climate (Gensini & Mote, 2015; Trapp et al., 2011). However, the use of a heuristic to delineate non-severe and severe convection can result in incorrect categorizations of thunderstorms that fall near the predetermined threshold. UH values representative of severe convection also vary seasonally and regionally, based on the climatological environments that drive severe convection activity (Molina, Allen, & Prein, 2020; Sobash & Kain, 2017). Recently, Sobash et al. (2020) trained a CNN to forecast severe hazard potential using severe thunderstorm parameters derived from WRF, showing that a CNN can learn from diagnostics. The focus herein lies in evaluating a CNN's ability to classify convection and its out-of-sample robustness to a future climate.

CNNs are a class of deep learning models canonically used for computer vision tasks because of the capability of processing multiple layers of information to detect nonlinearities and translation invariant details of features (Krizhevsky et al., 2012; LeCun et al., 1998). Various techniques have been developed to prevent deep learning models from overfitting and to improve training stability, such as dropout and batch normalization (Ioffe & Szegedy, 2015; Srivastava et al., 2014), which help CNNs generalize relationships among input features and increase prediction accuracy. However, explaining the reasons for model skill has been challenging, due to the complex architecture of CNNs that include many trained weights and biases within hidden layers and feature maps. Various CNN techniques have been recently developed to create and improve explanations of machine learning predictions and classifications (Barnes et al., 2019; McGovern et al., 2019). These explanation techniques include saliency maps (Simonyan et al., 2013) and permutation feature importance (Breiman, 2001; Lakshmanan et al., 2015), which have been shown to help explain skillful CNN predictions of convective hazards (Gagne II et al., 2019). Identifying reliable reasons for model
Earth and Space Science

1. Are future strongly rotating thunderstorms classified skillfully by a CNN that was trained under current climate conditions?
2. Which input features and spatial patterns are identified to be most important by the deep CNN for classification?
3. What are the reasons (explanations) for incorrect classifications?

2. Data and Methods

2.1. Thunderstorm Identification in Climate Simulations

A set of two convection-permitting model simulations created by the Water System Program of the National Center for Atmospheric Research were used to extract thunderstorm objects for this study (C. Liu et al., 2017). The two simulations were created using WRF at 4 km grid spacing over the CONUS. The WRF simulations cover 13 years each and represent a retrospective climate period (October 2000–September 2013) and a future climate period (end of the 21st century). Initial and boundary conditions for both simulations were driven by the 6-hourly and 0.7° ERA-Interim (Dee et al., 2011), which is a global climate reanalysis data set produced by the European Centre for Medium-Range Weather Forecasts. A pseudo-global warming (PGW) perturbation signal (Schär et al., 1996), representative of an end of the 21st century business as usual climate scenario, was added to state variables of the future climate simulation. The PGW signal was derived from a set of 19 Coupled Model Intercomparison Project Phase 5 (CMIP5) models (Taylor et al., 2012) generated with a Representative Concentration Pathway of 8.5 W m⁻² (RCP8.5) radiative forcing, which is a very high greenhouse gas concentration pathway (Moss et al., 2010). To prevent drifting of the 4 km regional simulation from the reanalysis boundary conditions, large-scale spectral nudging of moderate strength was applied above the planetary boundary layer (von Storch et al., 2000), which provided synoptic-scale fidelity to past weather events yet allowed the mesoscale to evolve with some freedom. These model simulations allow us to isolate thermodynamic signals from kinematic influences on the future climate. Simulation details are available in Table 1 and additional specifications can be found in C. Liu et al. (2017).

The watershed transform (Lakshmanan et al., 2009) was used to identify high-intensity updrafts that constitute thunderstorms from the convection-permitting climate simulations. The watershed transform, as employed herein, identified thunderstorms using a simulated radar reflectivity minimum threshold of 40 dBZ, which is a quantity proportional to the number of drops per unit volume and provides an estimate of convective precipitation (Trapp et al., 2011). Grid cells adjacent to the detected local maxima that also exceeded a minimum threshold of 20 dBZ were then treated as a part of the thunderstorm object. This process was repeated iteratively and surrounding grid cells were continually associated with a thunderstorm until performance can increase the trust of atmospheric scientists in machine learning and foster further discovery of the physical processes driving societally impactful weather and climate extremes.

Using deep learning and explanation techniques, the following questions will be analyzed in this study:

1. Are future strongly rotating thunderstorms classified skillfully by a CNN that was trained under current climate conditions?
2. Which input features and spatial patterns are identified to be most important by the deep CNN for classification?
3. What are the reasons (explanations) for incorrect classifications?
values were either below a minimum threshold of 20 dBZ or exceeded a predetermined thunderstorm object spatial extent of 128 km (32 grid cells × 4 km grid spacing). Each thunderstorm was saved as an object spanning 128 × 128 km containing the thunderstorm and the adjacent environment, which influences thunderstorm characteristics (R. L. Thompson et al., 2012). Thunderstorms were extracted over land and east of the Rocky Mountains (Figure 1), where severe thunderstorms have a greater climatological likelihood of occurrence (Brooks et al., 2003). The temporal focus of this study was limited to winter (December, January, and February; DJF) and spring months (March, April, and May; MAM). Other seasons were omitted due to a simulated dry bias during summer months across the central CONUS, which was partly associated with land-surface feedbacks (Barlage et al., 2018).

Similar to Gagne II et al. (2019), meteorological state variables were extracted from the WRF simulations to train a CNN after creating the thunderstorm objects. Five variables were extracted and interpolated onto four different vertical levels, resulting in a total of 20 input attributes used for training the CNN (Figure 2). The five variables are pressure (P; hPa), temperature (T; K), water vapor mixing ratio (q_w; g kg⁻¹), and zonal (u) and meridional (v) winds (m s⁻¹). Variables were then interpolated onto the following heights above ground level (AGL): 1, 3, 5, and 7 km. AGL heights were preferred over constant pressure surfaces because pressure surfaces might be below ground across portions of the High Plains and AGL heights are more likely to sample similar parts of a thunderstorm updraft. UH (m² s⁻²) was also extracted and is quantified as

\[ \text{UH} = \int_{z=2 \text{ km}}^{z=5 \text{ km}} w \zeta \, dz, \]

where the integral of the product of vertical velocity (w) and vertical vorticity (\( \zeta \)) is computed from 2 to 5 km AGL (Kain et al., 2008). UH was used as ground truth to create training labels, but was not used as an input attribute (i.e., variable) for the CNN. A high-magnitude UH threshold (e.g., 75 m² s⁻²) was used to delineate convection more likely to be of supercell morphology (Sobash et al., 2011). The 1D vector containing labels for CNN training and testing was created using binary assignment (i.e., integer encoding) derived from UH and encoded as 0 or 1. Values exceeding the UH threshold (the potentially severe category) were assigned a label of 1, whereas values below the threshold were assigned a label of 0. Thunderstorm objects were then split into two subsets prior to CNN training: 60% for training and 40% for testing. Stratified sampling of
thunderstorm objects that exceeded or did not exceed the delineated UH threshold was conducted to ensure that training and testing data contained the same percentage of majority and minority classes. Since the meteorological variables contain different dynamic ranges, the training data were standardized by subtracting the training set variable's mean and then dividing by its standard deviation. The testing data were also standardized for model evaluation using the mean and standard deviation values extracted from the current climate training data.

### 2.2. CNN Architecture and Explanations

The deep learning model used in this study was a CNN (LeCun et al., 1990) that consisted of three convolutional layers (similar to Gagne II et al., 2019). The 20 input attributes described in the previous sub-section were fed into the first convolutional layer (Figure 2). A stride length of 1 was used for the filter windows, which consisted of $5 \times 5$ grid cells, with zero padding also applied to the edges of each feature map. The rectified linear unit (ReLU; max(0, $x$)) activation function was used for each feature map (except for the last dense layer), which preserved the magnitude of positive signals and negated negative signals when propagated forward through the network (LeCun et al., 2015). Max pooling was performed after each convolutional layer by extracting maximum values of the feature maps within a sliding $2 \times 2$ filter window with stride length of 1. Max pooling added translation invariance and allowed the model to learn higher-level features (i.e., increase in kilometers) in deeper layers. After the three convolution and pooling operations, the resultant data were flattened into a 1D vector, passed through a dense layer (Figure 2), and a ReLU activation function was applied to its output. The 1D vector was then passed through a final dense layer, with a sigmoid activation function applied to produce the model's output as a value between 0 and 1, which was interpreted as the probability that the input attributes contained a strongly rotating thunderstorm.

The weights of the CNN were trained to minimize mean squared error (MSE) using the Adam optimization algorithm (Kingma & Ba, 2014) via backpropagation with a learning rate of 0.0001. Glorot uniform was used as the layer weight initializer (Glorot & Bengio, 2010). Sensitivity to random weight initialization was assessed by training several models using different random initializations and skill was comparable across models, potentially due to the large training sample size or the large number of input attributes used during training. During training, a batch size of 128 was used, randomly pulled from the training data population and passed forward through the CNN. To prevent overfitting of weights during training, Ridge (L2 norm; 0.001) regularization was added as a penalty term to reduce the magnitude of the weights at each convolutional layer. Batch normalization was also applied after each convolutional layer and the first dense layer (i.e., before each pooling layer), which involved standardizing layer outputs by subtracting the batch mean and dividing by the batch standard deviation, in effect reducing covariance shift (Ioffe & Szegedy, 2015). 2D spatial dropout (30% in this study; Srivastava et al., 2014) was also employed after batch normalization, which increased the robustness of learned features. We note that numerous training iterations were run.
with the order of batch normalization and spatial dropout reversed, but results were more skillful with batch normalization preceding spatial dropout in this application. A validation data set was used during training, consisting of 10% of the available training data, which provided insight into the skill of the model during training. Cross-validation on fivefolds was also performed to assess result sensitivity to the underlying test data distribution and differences were minimal. The final model settings were selected based on the lowest resultant test data MSE from a hyperparameter grid search that resulted in over 128 independently trained CNNs trained using 20 epochs. The classification output of the lowest MSE was evaluated using probabilistic and nonprobabilistic skill metrics that will be further detailed within the results. For more details about CNNs, see Goodfellow et al. (2016).

To explore the relative importance of specific meteorological variables on the CNN classification performance, we used the permutation feature importance (PFI; Breiman, 2001) analysis, specifically the single-pass forward test variant. PFI ranks variables based on how much randomizing them impacts error during testing, with larger magnitude decreases in skill associated with greater importance. Higher relative importance suggests that the respective variables have greater relevance to the classification due to the larger magnitude weights associated with them within the CNN architecture. Five hundred permutations were completed for each of the 20 variables to capture uncertainty associated with shuffling order in PFI. Permutated fields for a set of examples were also visualized to further explain variable importance results. The chosen examples consist of cases that were originally classified as one class by the CNN but switched to another class due to PFI. Certain classified thunderstorms that were switched to incorrect classifications (according to the ground truth label) also consistently appeared in larger skill reductions, and these were used to narrow down the subset of thunderstorms for visualization. To explain model reasoning within a spatial context, image-specific saliency maps (Simonyan et al., 2013) and input*gradient (Shrikumar et al., 2016) were used. Input*gradient is computed as the product of the local gradient and input (Mamalakis et al., 2021), which shows the variable contribution to the prediction.

3. Results

3.1. Thunderstorm Classification in a Future Climate

Thunderstorms identified within the future climate model simulation contain warmer temperatures and higher moisture content than thunderstorms identified within the current climate model simulation at all vertical levels (1, 3, 5 and 7 km; Table 2), which is consistent with the applied PGW signal (C. Liu et al., 2017). Table 2 shows that future thunderstorms contain about 1.3 g kg\(^{-1}\) more low-level (at 1 km) water vapor mixing ratio and are about 2.4 K warmer at low levels (at 1 km) than thunderstorms of the current climate (both statistically significant at the 95th percentile confidence level). These results are also consistent with the Clausius-Clapeyron equation that estimates a 7% increase in saturation vapor pressure per +1°C. Using the Clausius-Clapeyron equation, a water vapor mixing ratio of future thunderstorms should be about 8.76 g kg\(^{-1}\) at 1 km, which is comparable to the 8.9 g kg\(^{-1}\) contained in thunderstorms extracted from the future climate model simulation (Table 2). Extremes within the future climate were also of interest. These extremes were classified as “outlier cases” and were selected as thunderstorms containing 1 km water vapor mixing ratio exceeding the 99th-percentile of thunderstorms from the future climate. The focus of outlier cases lies on the 1-km water vapor mixing ratio because increased low-level moisture and thermodynamic buoyancy can result in more intense vertical winds related to stronger thunderstorm updrafts. Added low-level moisture and warmth provide additional thermodynamic buoyancy and vertical instability that could lead to more intense convection in the future (Prein et al., 2017; K. L. Rasmussen et al., 2017). The increased moisture and warmth could also pose the CNN with added difficulty in performing the thunderstorm classification task. Table 2 shows little change in zonal (u) and meridional (v) thunderstorm winds between the current and future climate model simulations. Since the classification task being performed by the CNN is related to winds, the relative consistency in wind magnitude may result in little change in classification skill between the current and future climate model simulations.

Here, we evaluate probabilistic forecasts generated by the CNN, which are probabilities that the thunderstorm objects contain a strongly rotating or non-strongly rotating thunderstorm. Strongly rotating thunder-
storms are associated with a higher probability magnitude and non-strongly rotating thunderstorms are associated with a lower probability magnitude. The large imbalance between the majority and minority classes was important to consider during the evaluation of the CNN classification skill (Table 3). Therefore, the performance diagram and metrics that are more useful for evaluating forecasts of rare events were used (Roebber, 2009). The minority class in this case consists of strongly rotating thunderstorms, which are rare events that comprise approximately 3% of all thunderstorms in the convection-permitting model simulations. Performance diagrams summarize the probability of detection (POD; ratio of hits to the total of hits and false alarms), critical success index (CSI; ratio of hits to the total of hits, false alarms, and misses), and bias (ratio of false alarms to misses). Success ratio (SR) is also summarized, which is 1–false alarm ratio (FAR; ratio of false alarms to the total of hits and false alarms). The curves shown on the performance diagrams were created by varying the probability threshold between 0 and 1 to convert probabilistic forecasts into binary forecasts and show how skill changes based on the probability threshold used (Figures 3a, 3c and 3e).

The performance diagrams (Figures 3a, 3c and 3e) show that despite being trained with thunderstorm objects extracted from the current climate model simulation, CNN skill remains consistent and high (0.69 max CSI) when classifying thunderstorms of the future climate model simulation (Figure 3c). These results suggest that a CNN is capable of learning spatial representations and variable relationships that are transferable to a warmer and moister climate. Figure 4a shows that correctly classified strongly rotating thunderstorms (according to the ground truth label) of the future climate contained approximately 4 g kg\(^{-1}\) more low-level moisture (at 1 km) than correctly classified strongly rotating thunderstorms of the cur-

| Table 2 | Median of Thunderstorm Variables Extracted From the Current and Future Climate Simulations |
|---------|------------------------------------------------------------------------------------------------------------------|
|         | Current climate                                                                                                     | Future climate                                                                 | Future outliers                                                                 |
|         | 1-km  | 3-km | 5-km | 7-km | 1-km  | 3-km | 5-km | 7-km | 1-km  | 3-km | 5-km | 7-km |
| Temperature (K) | 283.7 | 272.7 | 261.0 | 247.4 | 286.2* | 275.6* | 264.6* | 251.9* | 295.7* | 284.7* | 272.5* | 260.5* |
| v-winds (m \(s^{-1}\)) | 5.3  | 7.8  | 9.7  | 11.2  | 5.1  | 7.4* | 9.5* | 11.3  | 4.8* | 4.0* | 3.8* | 4.2*  |
| u-winds (m \(s^{-1}\)) | 3.1  | 9.9  | 14.1 | 17.2  | 2.8* | 10.0 | 14.4* | 17.8* | 2.6* | 7.0* | 9.8* | 12.0* |
| Water vapor mixing ratio (g kg\(^{-1}\)) | 7.6  | 4.8  | 2.1  | 0.7   | 8.9* | 5.7* | 2.8* | 1.0*  | 17.2* | 8.0* | 3.5* | 1.4*  |
| Pressure (hPa) | 868.7 | 679.9 | 526.4 | 402.3 | 869.0 | 681.7* | 529.5* | 406.3* | 889.2* | 704.1* | 551.5* | 427.0* |

*Note. Environments surrounding the thunderstorms were omitted for these statistics. Future thunderstorms with higher low-level moisture content than most cases in the future climate (i.e., outlier cases with \(\geq 99^{th}\) percentile of 1-km water vapor mixing ratio in the future climate), are also shown. Statistically significant values of the future climate and future outliers are indicated in boldface with asterisks (*) and computed using confidence intervals of 2.5th and 97.5th percentile of a 1,000-member bootstrap from a total sample of 454,242 thunderstorm objects extracted from the current climate simulation.
Table 3
Table Contains Various Skill Metrics Used for Evaluation of Convolutional Neural Network (CNN) Performance During the Current and Future Climate

| Skill Metric | Current Climate | Future Climate | Future Outliers |
|--------------|-----------------|----------------|-----------------|
| True positives | 9,089 | 10,984 | 601 |
| False positives | 1,633 | 3,420 | 280 |
| False negatives | 2,250 | 1,954 | 34 |
| True negatives | 441,270 | 440,109 | 3,652 |
| AUC | 0.90 | 0.92 | 0.94 |
| CSI | 0.70 | 0.67 | 0.66 |
| Hit Rate | 0.80 | 0.85 | 0.95 |
| Bias | 0.95 | 1.11 | 1.39 |
| BSS | 0.74 | 0.70 | 0.59 |
| Resolution | 0.02 | 0.02 | 0.08 |
| Uncertainty | 0.02 | 0.03 | 0.11 |

Note. Also shown are the total number of true positive (i.e., hits), false positive (i.e., false alarms), false negative (i.e., misses), and true negative predictions made by the CNN. Future thunderstorms that have higher low-level moisture content than most cases in the future climate (i.e., outlier cases with ≥ 99th percentile of 1 km water vapor mixing ratio in the future climate), are also shown. Metrics were computed using a 0.5 forecast probability threshold for the current, future, and outlier thunderstorms.

rent climate. The consistency in CNN skill could be partly related to bulk wind shear (1–5 km) distributions that remained relatively stationary between both climate model simulations (Figure 4a). Despite the imbalance between the majority and minority classes, the CNN was able to perform the classification task skillfully, suggesting that techniques to augment minority classes may not always be necessary (e.g., Chawla et al., 2002). However, model bias exhibits some sensitivity to the forecast threshold used. We note that the same forecast threshold values were evaluated for current, future, and outlier thunderstorms (10,000 values evenly spaced between 0 and 1) in Figure 3 and only a threshold of 0.5 was used for all thunderstorm classes in Table 3. Max CSI and lower bias were achieved when evaluating model skill using a probability threshold of approximately 0.6 in the future climate and 0.5 in the current climate (Figure 3c). A probability threshold of 0.5 results in a small over forecasting bias (>1) of strongly rotating thunderstorms of the future (Table 3), which shows that the CNN generally has lower skill in classifying strongly rotating thunderstorms of the warmer and more moist climate.

Performance metrics were also computed for the outlier future thunderstorms, which were characterized by a higher low-level moisture content than the 99th percentile of the future climate, in order to further quantify the out-of-sample robustness of the CNN (Table 2). Results show that CNN classification skill with outlier thunderstorms of the future climate remains high, with a max CSI of 0.73 (Figure 3e), which is comparable to the current and future climate subsets (Figures 3a and 3c). These results further substantiate that a CNN can exhibit out-of-sample robustness in climate applications. Results also suggest that deep learning can sufficiently generalize relationships among input variables and remain skillful with extreme events. However, over forecasting of strongly rotating outlier thunderstorms was identified (bias >1) with a probability threshold of 0.5 (Figure 3e; Table 3), which implies overconfidence in classifying thunderstorms with extreme low-level moisture. Like thunderstorms of the future climate, the consistency in CNN skill could be partly related to bulk wind shear (1–5 km) distributions that remained relatively stationary between current climate and future outlier thunderstorms (Figure 5a). However, in addition to high-end low-level moisture content, future outlier thunderstorms also contained substantially higher moisture content in mid-to-upper levels, as shown in Figure 5 for 5 km maximum water vapor mixing ratio. This result suggests that the spatial arrangement of meteorological fields likely also plays an important role in CNN prediction skill in addition to variable relative magnitudes. We note that a model consisting of the same architecture was trained with class imbalance addressed (i.e., equal sized potentially severe and non-severe classes) and resulted in more substantial bias (Table 4) than the model trained with large class imbalance (Table 3).

The Brier skill score (BSS) was used as an additional evaluation metric and can be visualized with the attributes diagram (Figures 3b, 3d and 3f), which shows forecast probabilities against observed relative frequencies (Hsu & Murphy, 1986; Wilks, 2011). An attributes diagram provides a measure of forecast reliability, where the dashed 45° line represents perfect reliability. Attributes diagrams show forecast probabilities (between 0 and 1), which are plotted against the observed relative frequency for that forecast probability (Wandishin et al., 2005). An example of “perfect reliability” for a 0.6 forecast probability (e.g., x-axis in Figures 3b, 3d and 3f) is when that forecast probability corresponds to a similar observed relative frequency (e.g., y-axis in Figures 3b, 3d, and 3f). The solid horizontal line in Figures 3b, 3d, and 3f shows the climatological probability of strongly rotating thunderstorms occurring within the respective climate sample, which is higher in outlier cases than in the current and future climates. Since attributes diagrams consider climatological and forecast probability frequency, they also show how different forecasts are from climatology (i.e., resolution). The gray shading in Figures 3b, 3d and 3f show areas contributing to positive BSS, which are areas where BSS resolution exceeds reliability (Gagne II et al., 2019). Inset plots (Figures 3b, 3d and 3f)
show the frequency of forecast probabilities for each climate subset, which in this case features a bi-modal distribution, with peaks at low (<0.05) and high (>0.95) forecast probabilities. This bimodal distribution is most pronounced for outlier cases (Figure 3f). The attributes diagram curve closely parallels the dashed 45-degree diagonal line across all forecast probabilities for the current climate (Figure 3b), which conveys high forecast reliability. However, future and outlier cases have lower reliability between the 0.2–0.7 forecast probabilities and higher reliability at low (<0.2) and high (>0.7) forecast probabilities (Figures 3d and 3f).

Figure 3. Performance diagrams (a, c, and e) show curves that represent convolutional neural network skill as a function of the probability of detection (POD) and success ratio (1-FAR [false alarm ratio]) across various probability thresholds. The grayscale-filled contours show the critical success index (CSI), the dashed lines display the bias, and circles along the curves display probability thresholds (a, c, and e). Attributes diagrams are also displayed, which show forecast probabilities against observed relative frequency, using a forecast probability bin size of 0.05 (b, d) and 0.1 (f). Inset panels in the top left show the frequency of forecast probabilities and the gray shading shows regions where resolution exceeds reliability (b, d, f). Ninety-fifth percentile confidence intervals (two-tailed) computed from a 1,000-member bootstrap shown with shading (a–f).
These results corroborate the performance diagram results, which show that the CNN has an over-forecasting bias for future and outlier thunderstorms.

3.2. CNN Interpretation

Permutation feature importance (PFI) was conducted to determine the relative importance of input variables on CNN prediction skill. The area under the receiver operating characteristic curve (AUC; Mason, 1982) was used, which is a scalar that represents model performance encompassing the probability of detection and false detection. Using AUC, PFI reveals that zonal ($u$) and meridional ($v$) winds at 3 km have the highest relative importance for CNN prediction (Figures 6a and 6d). PFI is consistent for predictions generated using the current climate and future climate thunderstorms (Figures 6a and 6d), which shows that mid-level
kinematic fields play an important role in the proper classification of rotating convective storms. This result is physically reasonable given that UH (computed from 2 to 5 km AGL) was used to create the thunderstorm labels that were subsequently used to train the CNN. The climatological homogeneity between current and future climate mid-level winds (Table 2) also likely contributed to the consistency in variable importance across climate subsets. Zonal and meridional winds at 1- and 5 km were also identified as relatively important (Figures 6a and 6d). Several thermodynamic variables also ranked in the top 50th percentile in importance, suggesting that the CNN also relies on characteristics of physical variables that were not included in the UH computation. These relatively higher-ranking thermodynamic variables include, temperature at 5 km and water vapor mixing ratio at 1- and 7 km (Figures 6a and 6d).

Additional skill metrics were used for PFI in order to explore the sensitivity of the analysis to the respective evaluation method. PFI using CSI, which is a skill evaluation metric that neglects true negative events (as described earlier), further emphasizes the relative importance of mid-level kinematic fields (Figures 6b).
BSS was also used for PFI (Figures 6c and 6f) and results generally align with AUC and CSI results in regard to the relative importance of mid-level kinematic fields. Interestingly, however, moisture at 5 km was ranked most important when evaluating the CNN classification skill for current and future climate (Figure 6c) thunderstorms. This result suggests that mid-level moisture is an important variable for classification of strongly rotating thunderstorms, given the lower ranking found using AUC, which also takes into account correct classification of non-strongly rotating thunderstorms (according to the ground truth label).

PFI offers insight into the relative importance of variables based on modulations to the CNN prediction skill, but the method does not provide reasons for the rankings. For instance, it is not immediately clear why water vapor mixing ratio at 5 km has greater relative importance than at 1 km. To explore the reasons for PFI rankings, visualizations were created of thunderstorms that were initially classified as strongly rotating but switched to a non-strongly rotating classification as a result of the permuted variable (Figure 7). Figure 7c shows an example of a strongly rotating thunderstorm. Its associated water vapor mixing ratio at 5 km (Figure 7a) was permuted to a field that had a greater overall magnitude of moisture (Figure 7b) and peak moisture values that were offset from the thunderstorm locations (Figure 7c), which resulted in the non-strongly rotating classification. Various other thunderstorms also had a similar pattern; higher overall moisture content and shifted peak value locations in the permuted field resulted in non-strongly rotating classifications (not shown). Supercells generally form in environments characterized by moist low-levels and drier mid-to-upper levels, while stratiform precipitation or less organized convection could be characterized by higher and more homogeneous moisture profiles (Bunkers et al., 2006; R. L. Thompson et al., 2012). These examples show that moisture characteristics of vertical atmospheric profiles are likely a learned feature by the CNN. In regard to the high importance of zonal and meridional winds, thunderstorms that were classified as non-strongly rotating during PFI were generally due to the uniformity of zonal or meridional winds in the permuted fields (Figures 7e and 7h), as opposed to the overall magnitude of the horizontal winds. These results show that the CNN learned that wind directional shifts over a small region located near the thunderstorm core were indicative of strong rotation.

Visualizations were also created for thunderstorms that were initially classified as non-strongly rotating, but switched to a strongly rotating classification during PFI (Figure 8). The permuted moisture fields for these examples (e.g., Figure 8b) were generally drier and contained large magnitude gradients in space that represented isolated and intense convection. Regarding kinematic fields, the original zonal (Figure 8d) and meridional (Figure 8g) winds lacked rotational characteristics for the respective thunderstorms (Figures 8f and 8i). However, the permuted fields contained strong rotational features (Figures 8e and 8h), which likely resulted in the changed classification.

Individual thunderstorms from the future climate model simulation were chosen to visualize areas of saliency for predictions made by the CNN. Simulated radar reflectivity of the respective examples are shown in Figure 9, which contains a true positive, false positive, false negative, and true negative case. High values of simulated radar reflectivity (>65) are evident near the thunderstorm core of the true positive case (Figure 9a), which represents a region of high precipitation intensity. The false positive and false negative examples also contain thunderstorms with high reflectivity (>65; Figures 9b and 9c), but the most intense region for the false negative case is located near the southern edge of the image. The true negative case (Figure 9d) contains lower maximum reflectivity magnitudes than the other examples (<65), and convection that is smaller in size and less organized, which possibly contributed to the true negative classification by the CNN.

Saliency maps highlight the thunderstorm object areas of input features that are of relevance to the CNN prediction. For water vapor mixing ratio (right two columns in Figure 10), positive gradients demarcate
the respective pixels that are relevant to the model prediction. Moisture at low- and mid-level heights for
the true positive case located near the thunderstorm core are relevant to the prediction of strongly rotating
thunderstorms (Figures 10c and 10d). While high moisture content may not be related to thunderstorm
rotation and horizontal kinematics, it does show that the CNN identified the thunderstorm core (region
of high precipitation intensity, and thus moisture content) as relevant for the strongly rotating prediction.
Non-salient regions of respective variables are zero gradients and therefore, correspond to pixels that are
not relevant to the model prediction. In the case of zonal and meridional winds at 3 km (left two columns
in Figure 10), winds of higher absolute magnitude with opposing signs in close proximity to each other
represent regions of rotational winds within the region of the thunderstorm's updraft. These features are
evident at the thunderstorm core location, suggesting that the rotation signature is relevant to the strongly
rotating prediction. Similar gradient patterns are present in the maps of the false-positive and false-neg-
ative examples at thunderstorm core locations (Figures 10e–10l). Saliency maps for true negative cases
are substantially different (Figures 10m–10p)—gradients are no longer present across a small and focused
region near the thunderstorm core, but rather across broad areas of the thunderstorm object. Additionally,
gradients from zonal and meridional winds generally no longer align to form an organized circulation (Fig-
ures 10m and 10n).

Another local and post hoc ML explanation method is the input*gradient method, which highlights con-
tributions to the prediction, computed as the product of the local gradient with the input itself (Shriku-
mar et al., 2016). Mamalakis et al. (2021) conducted an attribution benchmark using attribution known

Figure 6. Permutation feature importance (PFI) analysis for the current climate (a–c) and future climate (d–f) thunderstorms shown using box and whisker
plots. The median of 500 permutations is represented by the vertical line within the box and the whiskers represent all 500 measured changes in skill. PFI was
conducted using various skill metrics, including area under the receiver operating characteristic curve (AUC; a, d), critical success index (CSI; b, e), and Brier
skill score (BSS; c, f). Changes in skill were normalized by the maximum change in the respective climate subset and skill metric.
a priori (i.e., ground truth) for regression problems, and found the input*gradient method to produce more skillful explanations than saliency maps. While the benchmark used global sea surface temperatures and consisted of a very different spatial scale as compared to our application, we decided to also use the input*gradient method for further exploration of our model’s explanations. Explanations were generally consistent between the saliency map and input*gradient methods, with a few minor exceptions (Figure 11). For instance, the wind rotational signature was mostly of one sign (positive) near the thunderstorm core using the input*gradient method for hits, false alarms, and misses (Figures 11a, 11b).

Figure 7. Three example cases (c, f, i) that were classified as non-strongly rotating thunderstorms during the permutation feature importance (PFI) analysis. The convolutional neural network classified these future climate thunderstorms as strongly rotating prior to PFI. The top row shows the original water vapor mixing ratio ($q_w$) field (a) for a pair of strongly rotating thunderstorms (c), and the $q_w$ field that replaced the original during PFI (b). The center and bottom rows show fields for other example thunderstorms, but for perturbed meridional ($v$) and zonal ($u$) winds, respectively. Updraft helicity exceeding 75 m$^{-2}$s$^{-2}$ is indicated with black contours (c, f, i). Vectors show winds at 3 km corresponding to the respective thunderstorm image (c, f, i). The $q_w$ fields were normalized by the maximum value of both plots (a, b).
The spatial extent of the contour explanations were reduced for true negative cases and primarily of negative magnitude for all variables shown (Figures 11m–11p). The moisture fields at 1- and 5-km for the miss (false negative) example contain mixed signals (positive and negative) that were displaced away from the thunderstorm core, potentially explaining why the CNN failed to classify the supercell as strongly rotating (Figures 11k and 11l). A challenge with comparing ML explanation methods without attribution ground truth is that we cannot quantify explanation skill and rely heavily on our domain knowledge to assess skill subjectively. It is also possible that certain explanation methods are more useful for certain spatial and temporal scales yet not others. These questions are beyond the scope of this study, but should be explored in future work.

Figure 8. Same as Figure 7, but for a subset of thunderstorms that were classified as strongly rotating. The convolutional neural network classified these thunderstorms as non-strongly rotating prior to permutation feature importance. No black contours are included in the thunderstorm plots (c, f, i) because updraft helicity did not exceed $75 \text{ m}^{-2} \text{s}^{-2}$ for the shown examples.
Results generated using saliency maps and input*gradients can be corroborated by visualizing the frequency of maximum UH for all thunderstorms. Figure 12a shows that the CNN is able to identify strongly rotating thunderstorms across a broad range of UH values that exceed 75 \( \text{m}^2\text{s}^{-2} \) and is therefore able to capture a variety of thunderstorm rotation intensities. Thunderstorms that were classified as strongly rotating, but evaluated as false alarms because the corresponding UH did not exceed 75 \( \text{m}^2\text{s}^{-2} \), are heavily skewed toward high UH values (mostly contained UH values that exceeded 40 \( \text{m}^2\text{s}^{-2} \); Figure 12b), which past studies have found to also be representative of supercellular convection (Trapp et al., 2011). Missed classifications of strongly rotating thunderstorms generally do not consist of large UH magnitudes (<100 \( \text{m}^2\text{s}^{-2} \)), with most thunderstorms characterized by UH values close to 75 \( \text{m}^2\text{s}^{-2} \) (Figure 12c). Most true negative cases consist of UH values below 40 \( \text{m}^2\text{s}^{-2} \), which is characteristic of less organized convective storms (Figure 12d). These results further demonstrate that a CNN is able to generalize a target derived from a heuristic using learned features in the data that would be difficult to encode due to spatial complexity. There is sensitivity, however, to the thunderstorm location within the thunderstorm object, which can be visualized with 2D histograms that contain the frequency of UH exceeding 75 \( \text{m}^2\text{s}^{-2} \), typically located near the thunderstorm core (Figure 13). Correctly classified strongly rotating thunderstorms (according to the ground truth label) contained regions of high rotation (UH>75 \( \text{m}^2\text{s}^{-2} \)) near the center of the thunderstorm object (Figures 13a and 13c), while missed classifications are located near the edges of the thunderstorm object (Figures 13b and 13d) for thunderstorms during both the current and future climate. This comparison shows that CNNs can struggle with classifications of features located near the edges of a spatial region of interest, resulting in missed events.

Figure 9. Simulated radar reflectivity for representative examples of thunderstorms extracted from the future climate simulation, evaluated as a hit (a), false alarm (b), miss (c), and correct negative (d).
4. Conclusions

A CNN was trained to learn relationships and identify features among meteorological state variables in order to classify convection types, with a focus on rotation within the updraft core of a thunderstorm. Strong rotation and associated storm morphology could result in a higher likelihood of convection producing severe hazards, such as tornadoes and large hail, which are a dangerous threat to the public. We hypothesized that due to climate change, a trained CNN may fail to classify and identify convection that lies outside of the climatological distribution of data used for training. We conducted a test-case study to address our hypothesis. Using a thermodynamically driven future climate model simulation, our results show that a CNN can remain skillful in classifying convective storms using learned representations of physical variables.

Figure 10. Saliency maps for representative examples of future climate thunderstorms, including true-positive (a–d), false-positive (e–h), false-negative (i–l), and true negative (m–p) cases. Variables shown include several denoted as important by the PFI analysis, such as 3 km zonal (a, e, i, m) and meridional (b, f, j, and n) winds, and water vapor mixing ratio ($q_w$) at 5-km (d, h, l, and p). Mixing ratio ($q_w$) at 1 km is also shown (c, g, k, and o).
The key results that provide answers to the questions posed in the introduction follow:

1. A CNN trained using a current climate model simulation can skillfully classify out-of-sample (with regard to moisture content) thunderstorms in a thermodynamically driven future climate. This is possibly partly due to the use of batch normalization and spatial dropout; an equivalent model trained without batch normalization and spatial dropout results in an under-forecasting bias (about 0.84) in the current and future climate.

2. Kinematic fields and mid-level moisture were identified as important variables for skillful classification by the CNN. Spatially, wind rotation signatures with concurrently overlaid sharp mid-level moisture gradients were also important.

3. Thunderstorm classifications that were incorrect according to the associated ground truth label included cases that were near the thunderstorm object edge or had a UH value that was near but on the opposite side of the predefined threshold.

**Figure 11.** The same as Figure 10, but input*gradient values are displayed instead of saliency maps.
Key result 1 shows that a CNN is robust to out-of-sample cases during convection classification, which is a promising result given the changes already occurring to large-scale environments and moisture advection patterns associated with severe thunderstorms (Gensini & Brooks, 2018; Molina & Allen, 2020). Key results 2 and 3 also show that a CNN can learn complex relationships among input features using labels derived from heuristics. Physical features were not prescribed but rather learned from the data, such as the importance of dry air at mid-levels for intense thunderstorm development when low-level moisture is present (i.e., convective available potential energy). We emphasize that UH was only used to create the labels and was not used as an input attribute into the CNN during training. Unlike computer vision classification tasks (Russakovsky et al., 2015), humans can bypass generating a large number of hand-labeled data for training models to perform atmospheric feature classifications, which would also pose challenges given conflicting definitions of atmospheric phenomena in the scientific literature. Additionally, results show that large imbalances in labeled data may be overcome with sufficient hyperparameter tuning. Overall, results show that the CNN can classify thunderstorms as strongly rotating that were near the UH threshold and appeared supercellular, learning to generalize prescribed UH labels.

There are several limitations that are important to acknowledge. The focus in this study lies on a future climate that was thermodynamically driven in order to isolate competing thermodynamic and kinematic signals, but it is possible that a CNN may not generalize well with a future climate that accounts for both changes in the thermodynamic and large-scale dynamics. We do note that there is a 14% increase in future strongly rotating thunderstorms as compared to the current climate, which is a substantial increase and an

Figure 12. Histograms show the frequency of maximum updraft helicity (UH) for future climate thunderstorms separated into four subsets: hits (a), false alarms (b), misses (c), and true negatives (d). Frequency of true negatives (d) are $10^3$ magnitude. For comparison, the frequencies of the maximum UH for current climate thunderstorms are shown with blue lines and for future outlier thunderstorms in the inset plots.
indication that changes in large-scale dynamics may not pose a significant issue to the CNN. An additional limitation is that physical interpretation methods require substantial human interpretation, making it possible to miss important features or fail to discover new physical relationships. However, this is a broader issue within machine learning explanations (i.e., interpretability), as it introduces the potential for confirmation bias from human scientists attempting to explain results. Future work should explore incorporating feature uncertainty or physics within the CNN model architecture to explore the differences to results contained herein. Additionally, methods to ameliorate missed classifications near the edges of study domains should be explored. As societal exposure to severe hazards continues to increase (e.g., Ashley & Strader, 2016), it is important to continue better identifying and understanding severe hazards within climate model simulations. The use of deep learning methods that do not impose rigid thresholds or expert systems decisions should continue to be explored, since meteorological phenomena generally do not neatly fit into predefined classes. Deep learning offers a viable avenue to continue to better understand weather and climate extremes.

**Data Availability Statement**

The climate data set (R. Rasmussen & Liu, 2017) can be freely obtained from the NCAR Research Data Archive hosted by CISL at [https://rda.ucar.edu/datasets/ds612.0/](https://rda.ucar.edu/datasets/ds612.0/). The software developed for this research is available on Github [https://github.com/mariajmolina/deep-conus](https://github.com/mariajmolina/deep-conus). Processed data shown in the figures are

**Figure 13.** Spatial histograms show the frequency of updraft helicity (UH) exceeding 75 m$^3$s$^{-2}$ normalized by maximum frequency for thunderstorm objects classified as hits (a, c) and misses (b, d) during the current (a, b) and future climates (c, d).
Acknowledgments

Portions of this study were supported by the Regional and Global Model Analysis (RGMA) component of Earth and Environmental Systems Modeling in the Earth and Environmental Systems Sciences Division of the U.S. Department of Energy’s Office of Biological and Environmental Research (BER) through the National Science Foundation IA 1947282. This work also supported by the National Center for Atmospheric Research (NCAR), which is a major facility supported by the National Science Foundation (NSF) under Cooperative Agreement No. 1852977. This material is also based upon work supported by the NSF under Grant No. 1ICER-2019758 and part of the NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES). M. J. Molina conducted this research as part of the NCAR Advanced Study Program. High-performance computing support on Cheyenne and Casper (https://doi.org/10.5065/D6RX99IX) was provided by the NCAR Computational and Information Systems Laboratory (CISL), some of which was from the CISL Analytics and Integrative Machine Learning (CISL) group allocation.

We are grateful to Imme Ebert-Uphoff (Colorado State University) and an anonymous reviewer for their insightful and valuable feedback that improved the quality of the manuscript.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). Tensorflow: A system for large-scale machine learning. In 12th USENIX symposium on operating systems design and implementation (OSDI) 16 (pp. 265–283).

Allen, J. T. (2018). Climate change and severe thunderstorms. Oxford Research Encyclopedia of Climate Science. https://doi.org/10.1093/acrefore/9781002028620.013.62

Ashley, W. S., & Strader, S. M. (2016). Recipe for disaster: How the dynamic ingredients of risk and exposure are changing the tornado disaster landscape. Bulletin of the American Meteorological Society, 97(5), 767–786. https://doi.org/10.1175/BAMS-D-15-00150.1

Barlage, M., Chen, F., Miguez-Macho, G., Liu, C., Liu, X., & Niyogi, D. (2018). Enhancing hydrologic processes in the Noah-MP land surface model to improve seasonal forecast skill. In 90th American Meteorological Society Annual Meeting.

Barnes, E. A., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2019). Viewing forced climate patterns through an AI lens. Geophysical Research Letters, 46(22), 13389–13398. https://doi.org/10.1029/2019GL084944

Binting, C., & Kunkel, K. E. (2019). Automated detection of weather fronts using a deep learning neural network. Advances in Statistical Climatology, Meteorology and Oceanography, 5(2), 147–160. https://doi.org/10.5194/ascmo-5-147-2019

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/a:1010934043424

Brooks, H. E. (2013). Severe thunderstorms and climate change. Atmospheric Research, 123, 129–138. https://doi.org/10.1016/j.atmosres.2012.04.002

Brooks, H. E., Doswell, C. A., III, & Kay, M. P. (2003). Climatological estimates of local daily tornado probability for the United States. Weather and Forecasting, 18(4), 626–640. https://doi.org/10.1175/1520-0434(2003)018<0626:CEOLDT>2.0.CO;2

Bunkers, M. J., Hjelmfelt, M. R., & Smith, P. L. (2006). An observational examination of long-lived supercells. Part I: Characteristics, evolution, and demise. Weather and Forecasting, 21(5), 673–688. https://doi.org/10.1175/WAF949.1

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Chollet, F. (2015). Keras documentation. Retrieved from keras.io

Clark, A. J., Gao, J., Marsh, P. T., Smith, T., Kain, J. S., Correia, J., Jr, et al. (2013). Tornado pathlength forecasts from 2010 to 2011 using ensemble updraft helicity. Weather and Forecasting, 28(2), 387–407. https://doi.org/10.1175/WAF-D-12-00038.1

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. Quarterly Journal of the Royal Meteorological Society, 137(653), 559–597. https://doi.org/10.1002/qj.828

Diffenbaugh, N. S., Scherer, M., & Trapp, R. J. (2013). Robust increases in severe thunderstorm environments in response to greenhouse forcing. Proceedings of the National Academy of Sciences, 110(41), 16361–16366. https://doi.org/10.1073/pnas.1307758110

Doswell, C. A., III, Brooks, H. E., & Maddox, R. A. (1996). Flash flood forecasting: An ingredients-based methodology. Weather and Forecasting, 11(4), 560–581. https://doi.org/10.1175/1520-0434(1996)011<0560:FFFAOB>2.0.CO;2

Duda, J. D., & Gallus, W. A., Jr (2010). Spring and summer midwestern severe weather reports in supercells compared to other morphologies. Weather and Forecasting, 25(1), 190–206. https://doi.org/10.1175/2009WAF2223381.1

Fukushima, K., & Miyake, S. (1982). Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. Pattern Recognition, 15(6), 455–469. https://doi.org/10.1016/0031-3203(82)90038-2

Gage, D. J., H. Haupt, S. E., Nychka, D. W., & Thompson, G. (2019). Interpretable deep learning for spatial analysis of severe hailstorms. Monthly Weather Review, 147(8), 2827–2845. https://doi.org/10.1175/MWR-D-18-0316.1

Gensini, V. A., & Brooks, H. E. (2018). Spatial trends in United States tornado frequency. npj Climate and Atmospheric Science, 1, 1. https://doi.org/10.1038/s41612-018-0048-2

Gensini, V. A., & Mote, T. L. (2015). Downscaled estimates of late 21st century severe weather from CCSM3. Climate Change, 129(1–2), 307–321. https://doi.org/10.1007/s10584-014-1320-z

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics (pp. 249–256).

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1). MIT Press.

Gropp, M. E., & Davenport, C. E. (2018). The impact of the nocturnal transition on the lifetime and evolution of supercell thunderstorms in the Great Plains. Weather and Forecasting, 33(4), 1045–1061. https://doi.org/10.1175/WAF-D-17-0190.1

Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019). Deep learning for multi-year ENSO forecasts. Nature, 573(7775), 568–572. https://doi.org/10.1038/s41586-019-1559-7

Hong, S.-Y., Noh, Y., & Duddia, J. (2006). A new vertical diffusion package with an explicit treatment of entrainment processes. Monthly Weather Review, 134(9), 2318–2341. https://doi.org/10.1175/MWR3199.1

Hoogwind, K. A., Baldwin, M. E., & Trapp, R. J. (2017). The impact of climate change on hazardous convective weather in the United States: Insight from high-resolution dynamical downscaling. Journal of Climate, 30(24), 10081–10100. https://doi.org/10.1175/JCLI-D-16-0885.1

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

Jørgensen, G. E., McGovern, A., Lagerquist, R., & Smith, T. (2020). Classifying convective storms using machine learning. Weather and Forecasting, 35(2), 537–559. https://doi.org/10.1175/WAF-D-19-0170.1
Kain, J. S., Weiss, S. J., Bright, D. R., Baldwin, M. E., Levit, J. J., Carbin, G. W., et al. (2008). Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Weather and Forecasting*, 23(5), 931–952. https://doi.org/10.1175/WAF207106.1

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105). Retrieved from http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

Lagerquist, R., McGovern, A., & Gagne, D. J., II (2019). Deep learning for spatially explicit prediction of synoptic-scale fronts. *Weather and Forecasting*, 34(4), 1137–1160. https://doi.org/10.1175/WAF-D-18-0183.1

Lagerquist, R., McGovern, A., Homeyer, C. R., Gagne, D. J., & Smith, T. (2020). Deep learning on three-dimensional multiscale data for next-hour tornado prediction. *Monthly Weather Review*, 148(7), 2837–2861. https://doi.org/10.1175/MWR-D-19-0372.1

Lakshmanan, V., Hondl, K., & Rabin, R. (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., Karstens, C., Krause, J., Elmore, K., Ryzhkov, A., & Berkseth, S. (2015). Which polarimetric variables are important for weather/no-weather discrimination? *Journal of Atmospheric and Oceanic Technology*, 32(6), 1209–1223. https://doi.org/10.1175/JTECH-D-13-00205.1

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. https://doi.org/10.1038/nature14539

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., & Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems* (pp. 396–404).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. https://doi.org/10.1109/5.726791

Liu, C., Ikeda, K., Rasmussen, R., Barlage, M., Newman, A. J., Prein, A. F., et al. (2017). Continental-scale convection-permitting modeling of the current and future climate of North America. *Climate Dynamics*, 49(1–2), 71–95. https://doi.org/10.1007/s00382-016-3327-9

Liu, Y., Racah, E., Correa, J., Khosrowshahi, A., Lavers, D., Kunkel, K., et al. (2016). Application of deep convolutional neural networks for detecting extreme weather in climate datasets. arXiv preprint arXiv:1605.01156

Mamalakis, A., Ebert-Uphoff, I., & Barnes, E. A. (2021). Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. arXiv preprint arXiv:2103.10005

Mason, I. (1982). A model for assessment of weather forecasts. *Australian Meteorological Magazine*, 30(4), 291–303.

McGovern, A., Lagerquist, R., John Gagne, D., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, 100(11), 2175–2199. https://doi.org/10.1175/BAMS-D-18-0195.1

Molina, M. J., & Allen, J. T. (2020). Regionally-stratified tornadoes: Moisture source physical reasoning and climate trends. *Weather and Climate Extremes*, 28, 100244. https://doi.org/10.1016/j.wace.2020.100244

Molina, M. J., Allen, J. T., & Prein, A. F. (2020). Moisture attribution and sensitivity analysis of a winter tornado outbreak. *Weather and Forecasting*, 35(4), 1263–1288. https://doi.org/10.1175/WAF-D-19-0240.1

Molina, M. J., Gagne, D. J., & Prein, A. F. (2020). DATA: Deep learning classification of potentially severe convective storms in a changing climate [Data set]. Zenodo. https://doi.org/10.5281/zenodo.405285

Moss, R. H., Edmonds, J. A., Hibbard, K. A., Manning, M. R., Rose, S. K., Van Vuuren, D. P., et al. (2010). The next generation of scenarios for global change research and assessment. *Nature*, 468(7328), 747–756. https://doi.org/10.1038/nature09823

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., et al. (2011). The community Noah land surface model with multiple-parameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *Journal of Geophysical Research*, 116(D12). https://doi.org/10.1029/2010JD015139

Prein, A. F., Liu, C., Ikeda, K., Trier, S. B., Rasmussen, R. M., Holland, G. J., & Clark, M. P. (2017). Increased rainfall volume from future convective storms in the US. *Nature Climate Change*, 7(12), 880–884. https://doi.org/10.1038/s41558-017-0067-7

Rasmussen, E. N. (2003). Refined supercell and tornado forecast parameters. *Weather and Forecasting*, 18(3), 530–535. https://doi.org/10.1175/1520-0434(2003)18<530:Rsastp>2.0.co;2

Rasmussen, K. L., Prein, A. F., Rasmussen, R. M., Ikeda, K., & Liu, C. (2017). Changes in the convective population and thermodynamic environments in convection-permitting regional climate simulations over the United States. *Climate Dynamics*, 55, 1–26. https://doi.org/10.1007/s00382-017-4009-7

Rasmussen, R., & Liu, C. (2017). High resolution WRF simulations of the current and future climate of North America. *Research Data Archive* at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. https://doi.org/10.5065/D4V405XP

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, 115(39), 9684–9689. https://doi.org/10.1073/pnas.1802861115

Roeber, P. J. (2009). Visualizing multiple measures of forecast quality. *Weather and Forecasting*, 24(2), 601–608. https://doi.org/10.1175/2008WAF2221159.1

Rusakovskiy, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Schär, C., Frei, C., Lüthi, D., & Davies, H. C. (1996). Surrogate climate-change scenarios for regional climate models. *Geophysical Research Letters*, 23(6), 669–672. https://doi.org/10.1029/96GL0265

Shrikumar, A., Greenside, P., Shcherbina, A., & Kundaje, A. (2016). Not just a black box: Learning important features through propagating activation differences. arXiv preprint arXiv:1605.01713

Shrinivasan, K., Vedalai, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualizing image classification models and saliency maps. arXiv preprint arXiv:1312.6034

Skamarock, W. C., & Klemp, J. B. (2008). A time-split nonhydrostatic atmospheric model for weather research and forecasting applications. *Journal of Computational Physics*, 227(7), 3465–3485. https://doi.org/10.1016/j.jcp.2007.01.037

Sobash, R. A., & Kain, J. S. (2017). Seasonal variations in severe weather forecast skill in an experimental convection-allowing model. *Weather and Forecasting*, 32(5), 1885–1902. https://doi.org/10.1175/WAF-D-17-0043.1

Sobash, R. A., Kain, J. S., Bright, D. R., Dean, A. R., Coniglio, M. C., & Weiss, S. J. (2011). Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Weather and Forecasting*, 26(5), 714–728. https://doi.org/10.1175/WAF-D-10-05046.1
Sobash, R. A., Romine, G. S., & Schwartz, C. S. (2020). A comparison of neural-network and surrogate-severe probabilistic convective hazard guidance derived from a convection-allowing model. *Weather and Forecasting, 35*(5), 1981–2000. https://doi.org/10.1175/WAF-D-20-0036.1

Sobash, R. A., Schwartz, C. S., Romine, G. S., Fossell, K. R., & Weisman, M. L. (2016). Severe weather prediction using storm surrogates from an ensemble forecasting system. *Weather and Forecasting, 31*(1), 255–271. https://doi.org/10.1175/WAF-D-15-0138.1

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*(1), 1929–1958. https://doi.org/10.5555/2627435.2670313

Taszarek, M., Allen, J. T., Brooks, H. E., Pilguj, N., & Czernecki, B. (2020). Differing trends in United States and European severe thunderstorm environments in a warming climate. *Bulletin of the American Meteorological Society, 102*, 1–51. https://doi.org/10.1175/BAMS-D-20-0004.1

Taylor, K. E., Stoulifer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society, 93*(4), 485–498. https://doi.org/10.1175/BAMS-D-11-00094.1

Thompson, R. L., Mead, C. M., & Edwards, R. (2007). Effective storm-relative helicity and bulk shear in supercell thunderstorm environments. *Weather and Forecasting, 22*(1), 102–115. https://doi.org/10.1175/WAF969.1

Thompson, R. L., Smith, B. T., Grams, J. S., Dean, A. R., & Broyles, C. (2012). Convective modes for significant severe thunderstorms in the contiguous United States. Part II: Supercell and QLCS tornado environments. *Weather and Forecasting, 27*(5), 1136–1154. https://doi.org/10.1175/WAF-D-11-00116.1

Toms, B. A., Kashinath, K., Prabhat, & Yang, D. (2019). Deep learning for scientific inference from geophysical data: The Madden-Julian Oscillation as a test case. arXiv preprint arXiv:1902.04621

Trapp, R. J., Diffenbaugh, N. S., Brooks, H. E., Baldwin, M. E., Robinson, E. D., & Pal, J. S. (2007). Changes in severe thunderstorm environment frequency during the 21st century caused by anthropogenically enhanced global radiative forcing. *Proceedings of the National Academy of Sciences, 104*(50), 19719–19723. https://doi.org/10.1073/pnas.0705494104

Trapp, R. J., Diffenbaugh, N. S., & Gluhovsky, A. (2009). Transient response of severe thunderstorm forcing to elevated greenhouse gas concentrations. *Geophysical Research Letters, 36*(1). https://doi.org/10.1029/2008GL036203

Trapp, R. J., & Hoogewind, K. A. (2016). The realization of extreme tornadic storm events under future anthropogenic climate change. *Journal of Climate, 29*(14), 5251–5265. https://doi.org/10.1175/JCLI-D-15-0623.1

Wandishin, M. S., Baldwin, M. E., Mullen, S. L., & Cortinas, J. V., Jr (2005). Short-range ensemble forecasts of precipitation type. *Weather and Forecasting, 20*(4), 609–626. https://doi.org/10.1175/WAF871.1

Wilks, D. S. (2011). Statistical methods in the atmospheric sciences (Vol. 100). Academic Press.

Zhou, K., Zheng, Y., Li, B., Dong, W., & Zhang, X. (2019). Forecasting different types of convective weather: A deep learning approach. *Journal of Meteorological Research, 33*(5), 797–809. https://doi.org/10.1007/s13351-019-8162-6