REVIEW

A Review of Machine Learning for Convective Weather

AMY MCGOVERN,[a,b,c] RANDY J. CHASE,[a,b,c] MONTGOMERY FLORA,[d] DAVID J. GAGNE II,[e,c] RYAN LAGERQUIST,[f] COREY K. POTVIN,[g,b,c] NATHAN SNOOK,[b,c] AND ERIC LOKEN[g,d]

[a] *School of Computer Science, University of Oklahoma, Norman, Oklahoma*
[b] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*
[c] *NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography, Norman, Oklahoma*
[d] *Cooperative Institute for Severe and High-Impact Weather Research and Operations, University of Oklahoma, Norman, Oklahoma*
[e] *National Center for Atmospheric Research, Boulder, Colorado*
[f] *Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado*
[g] *National Severe Storms Laboratory, Norman, Oklahoma*

(Manuscript received 22 October 2022, in final form 28 March 2023, accepted 9 May 2023)

ABSTRACT: We present an overview of recent work on using artificial intelligence (AI)/machine learning (ML) techniques for forecasting convective weather and its associated hazards, including tornadoes, hail, wind, and lightning. These high-impact phenomena globally cause both massive property damage and loss of life, yet they are very challenging to forecast. Given the recent explosion in developing ML techniques across the weather spectrum and the fact that the skillful prediction of convective weather has immediate societal benefits, we present a thorough review of the current state of the art in AI and ML techniques for convective hazards. Our review includes both traditional approaches, including support vector machines and decision trees, as well as deep learning approaches. We highlight the challenges in developing ML approaches to forecast these phenomena across a variety of spatial and temporal scales. We end with a discussion of promising areas of future work for ML for convective weather, including a discussion of the need to create trustworthy AI forecasts that can be used for forecasters in real time and the need for active cross-sector collaboration on testbeds to validate ML methods in operational situations.

SIGNIFICANCE STATEMENT: We provide an overview of recent machine learning research in predicting hazards from thunderstorms, specifically looking at lightning, wind, hail, and tornadoes. These hazards kill people worldwide and also destroy property and livestock. Improving the prediction of these events in both the local space as well as globally can save lives and property. By providing this review, we aim to spur additional research into developing machine learning approaches for convective hazard prediction.

KEYWORDS: Convective storms; Artificial intelligence; Deep learning; Machine learning; Neural networks

1. Introduction

High-impact convective weather events cause significant loss of life and property annually. With population growth and climate changes, the number of billion dollar events has been growing (NOAA National Centers for Environmental Information 2022). While we cannot control such events, we can improve our resiliency to such high-impact phenomena by improving both prediction and understanding of these events. This paper provides an overview of the current research in developing artificial intelligence (AI) and machine learning (ML), referred to as AI/ML throughout the paper, techniques to improve prediction and understanding of convective weather and its associated hazards including tornadoes, wind, hail, and lightning.

The prediction of convective hazards cannot be optimized solely by improving existing numerical weather prediction (NWP) models. Thunderstorms and their hazards are not fully resolved at the resolutions achievable by real-time models in the foreseeable future (e.g., Bryan et al. 2003; Potvin and Flora 2015; Bauer et al. 2021). Ensuring that the subgrid physics parameterizations remain realistic as model resolution increases is challenging (e.g., Morrison et al. 2020; Angevine et al. 2020). New sensing capabilities, such as the shift in the national radar network to dual polarization (Kennedy et al. 2014; Loeffler and Kumjian 2018; Kuster et al. 2019), can improve assimilation and prediction of thunderstorms (e.g., Putnam et al. 2021), but these improvements will be constrained by model limitations. Optimizing the prediction of thunderstorms and

Corresponding author: Amy McGovern, amcgovern@ou.edu

DOI: 10.1175/AIES-D-22-0077.1 e220077

their hazards requires postprocessing techniques that leverage previously learned statistical relationships between observations and NWP output and convective events.

AI, including ML and deep learning (LeCun et al. 1998, 2015), is poised to make significant breakthroughs across many fields in science and society. Deep learning (DL) has recently demonstrated this in fields including computer gaming, physics, and medicine (e.g., Silver et al. 2017; Rakhlin et al. 2018; Krizhevsky et al. 2012; Dieleman et al. 2015; Karras et al. 2018; Suwajanakorn et al. 2017; Zhao et al. 2018). ML and deep learning have recently been recognized as critical to geosciences in general (Reichstein et al. 2019; Karpatne et al. 2019; Gil et al. 2019), including atmospheric sciences (Boukabara et al. 2019), ocean science (Malde et al. 2020), and computational sustainability (Gomes et al. 2019). While discussing all of the reasons why AI truly is poised to make significant breakthroughs across many fields of science and society is really a paper by itself, we list a few reasons here to highlight the need to further research into AI methods for all types of high-impact phenomena, including our focus here on convective weather.

With the introduction of DL, AI/ML methods are able to learn with complex spatial and temporal patterns without the need for significant feature engineering. This can facilitate the discovery of new features or interactions that the developers did not originally envision. With more traditional ML methods, such as decision trees or regression, the AI method was learning only on the features that the developers identified, which limits the hypothesis space of the model. With the ability to identify its own features, a DL model may be able to find new features that were not previously considered by the domain scientists. AI methods provide a way to examine the data with fewer preconceived ideas about what the answer should be. We wish to note that this is both a potential blessing and a curse, as AI methods lack a fundamental understanding of the laws of physics and may identify nonphysical features. Several of the authors discuss this in more length in our recent article on the need for ethical and responsible development of AI for the environmental sciences (McGovern et al. 2022).

AI is also able to handle larger datasets than a human can feasibly analyze. The amount of data being generated by high-resolution NWP models and new sensing systems is overwhelming, and much larger than a human could examine in a reasonable amount of time. AI can help to sift through these data and identify the most relevant parts for a human to study. To give two concrete examples in the domain of convective weather, one could imagine generating hundreds of high-resolution NWP storms of a specific phenomenon, such as hail. This would likely generate many terabytes of data, far too much for a human to analyze. However, an AI system could then sift through the storms, looking for features that predict the phenomena being analyzed and potentially identifying new features. As a second example, imagine being a forecaster looking at the high-resolution satellite data, the high-resolution data from mobile radar trucks, and a field campaign flying above a storm, all at once. While all of these data are extremely valuable, they are overwhelming for a single person to analyze, and AI can be used to highlight the critical areas for a human forecaster to study in more depth.

Another reason AI is poised to create breakthroughs across fields is the explosion is the combination of easily available toolkits for AI and the increasing availability of resources to run AI methods. In the past few years, the number of AI and DL toolkits have grown exponentially. As more toolkits become available, their ability to be used by nonexperts has increased, especially as new books and tutorials become available. Likewise, cloud-based computing resources are making the ability to train larger AI and DL models more available (there is still plenty of work to be done in making it truly inclusive), which facilitates new people getting involved in developing and applying AI models to a wide variety of phenomena and will also stimulate new science.

AI has a rich history within the atmospheric sciences (e.g., Haupt et al. 2008; McGovern et al. 2017a; Monteleoni et al. 2013) including regression (Malone 1955; Kitzmiller et al. 1995; Billet et al. 1997), tree-based methods (Mecikalski et al. 2015; Williams et al. 2008a,b; Gagne et al. 2009; McGovern et al. 2014; Williams 2014; McGovern et al. 2015; Clark et al. 2015; Elmore and Grams 2016), Bayesian nets (Cintineo et al. 2014b), genetic algorithms (Allen et al. 2007), and support vector machines (Trafalis et al. 2003; Adrianto et al. 2005, 2009). DL has recently been used in meteorology to estimate sea ice concentration (Wang et al. 2016) and tropical-cyclone intensity (Wimmers et al. 2019), detect extreme-weather patterns in model output (Racah et al. 2017; Kurth et al. 2018; Lagerquist et al. 2019) and replace subgrid-scale parameterizations in numerical models (Rasp and Lerch 2018; Brenowitz and Bretherton 2018, 2019; Kochkov et al. 2021).

The authors of this paper have extensive experience in developing ML models for convective weather prediction (McGovern et al. 2017b; Lagerquist et al. 2017, 2020; McGovern et al. 2017a, 2019; Gagne et al. 2017, 2019; Burke et al. 2020; Loken et al. 2020, 2022a; Flora et al. 2021). Our models can be used in conjunction with human forecasters to produce more skillful forecasts and to reduce the cognitive overload of preparing forecasts with a deluge of data. The contribution of this paper is to summarize and review the state of the art in developing ML and DL approaches for multiple convective hazards and the challenges in the development of these models. We also discuss the challenges in creating models that are trusted by human forecasters and call for future research in creating and validating such models. This paper differs from a review paper on extreme weather that recently appeared (Salcedo-Sanz et al. 2022). Their focus is on extreme events in general, while our paper specifically covers ML for convective weather with a higher level of specificity and depth. Note that this paper makes the assumption that the reader is already familiar with AI/ML methods. If that is not the case, we highly recommend the ML tutorials recently developed by several of the authors (Chase et al. 2022, 2023).

2. AI/ML for convective hazards

Convective storms generate a variety of high-impact phenomena including wind, hail, tornadoes, and lightning. We break the review into categories, focusing first on the initiation

of the convective storms themselves, then on each hazard in turn and finally on approaches that examine the prediction of multiple hazards.

a. *Forecasting the timing and location of convection*

Convection, even of the nonsevere variety, has a major impact on activities such as large outdoor gatherings (e.g., concerts and sporting events), aviation, and wind energy. For such activities, predicting the timing and location of convection, even at short lead times, allows for preventive actions that reduce both human and economic losses (Wilson and Mueller 1993; Mueller et al. 1993; Ahijevych et al. 2016). In the ML literature, some work has focused solely on convective *initiation* (CI; the development of new storms), while some work has focused more generally on convective *occurrence* (CO; which includes modeling convective initiation/decay and the advection of existing storms). Also, most ML work has focused on the nowcasting horizon (lead times of 0–3 h), as forecasting CI/CO at longer lead times without NWP is very difficult. Early work in forecasting CI/CO included mostly expert systems applied to radar data, while later work has applied more sophisticated ML algorithms to satellite data. Satellites can help detect thunderstorms earlier in their life cycle (i.e., before they develop enough precipitation to produce a radar echo) and cover a much larger percentage of the globe than do radars, especially over the oceans. Some of these ML algorithms incorporate information from more traditional object-tracking techniques, such as optical flow (Sun et al. 2019; Su et al. 2020) or atmospheric motion vectors (Mecikalski and Bedka 2006). Also, for CO nowcasting, some authors measure the performance of their ML against a persistence model ("existing thunderstorms will stay at their current location forever"), which is a common baseline model for nowcasting problems (e.g., Lagerquist et al. 2021a).

An early success in CI nowcasting is an expert system called Satellite Convection Analysis and Tracking (SATCAST; Mecikalski and Bedka 2006). SATCAST used eight predictors [called "interest fields" in Mecikalski and Bedka (2006) and much of the other literature], all based on infrared Geostationary Operational Environmental Satellite (GOES) data, to forecast CI at 0–1-h lead times. SATCAST improved upon earlier efforts to forecast CI/CO—which used only radar data (Mueller and Wilson 1989; Mueller et al. 1993; Wilson and Mueller 1993) or simpler satellite data (Roberts and Rutledge 2003)—by using multispectral satellite data, including band differences (e.g., 13.3- minus 10.7-μm brightness temperature) and temporal changes thereof. Mecikalski and Bedka (2006) remains an influential study, as the predictors they developed have since been widely used for CI and CO forecasting. SATCAST was followed by two more influential expert systems for CI forecasting: the University of Wisconsin Convective Initiation (UWCI; Sieglaff et al. 2011) algorithm and SATCASTv2 (Walker et al. 2012). These newer systems differed from SATCAST in their object-tracking approaches, used to compute temporal changes in brightness temperature. Veillette et al. (2013) builds on the SATCAST system by integrating both satellite and NWP data to predict CI using random forests. Williams et al. (2008c) also used random forests to predict CI, based on both satellite fields and NWP model data.

In the GOES-R era, the first CI algorithm that we are aware of is Mecikalski et al. (2015). Their use of ML—specifically logistic regression and random forests—allowed for more skillful forecasts than earlier approaches and for probabilistic, rather than binary, forecasts. The GOES-R CI algorithm had another crucial advantage: it used predictors from NWP, allowing for a dramatic decrease in false alarms, which have traditionally been a problem in CI forecasting. Mecikalski et al. (2015) found that the four most important predictors were all from NWP data, rather than satellite data: surface-based convective inhibition (CIN), most unstable CIN, surface-based convective available potential energy (CAPE), and most unstable CAPE. However, one caveat to this finding is that most NWP models assimilate satellite data, so the CIN and CAPE variables might have been influenced by satellite data. Continuing ML efforts, Lee et al. (2017) used random forests to develop a CO-forecasting algorithm for eastern Asia and the western Pacific, based on 12 infrared predictors from the *Himawari-8* satellite. Han et al. (2017) added a nowcasting method based on real-time radar data and predicting radar echoes greater than 30 dBZ in the next 30 min. Han et al. (2019) also expanded upon Mecikalski et al. (2015) with a procedure to iteratively expand the training set, adding cases similar to those that yielded the worst random-forest forecasts. Although most of their performance metrics were worse than in Lee et al. (2017), the model of Han et al. (2019) detected convective storms earlier in their life cycle, resulting in a longer lead time for CI.

To our knowledge, the first deep learning application in this domain was Lee et al. (2021), who developed a convolutional neural network (CNN; LeCun et al. 1998, 2015) to detect ongoing convection. A major advantage of Lee et al. (2021) is that they used a sophisticated echo-classification algorithm to create their labels, which uses richer radar information than a simple reflectivity threshold, the labeling method used by most studies. The labels (or "ground truth") consist of a 0 or 1 at each pixel, indicating the absence of presence of convection, respectively.

Lagerquist et al. (2021a) also used deep learning, specifically U-nets (Ronneberger et al. 2015), to forecast CO at 0–2-h lead times. Like Lee et al. (2017) and Han et al. (2019), they used infrared predictors from the *Himawari-8* satellite but focused on a smaller region near Taiwan. They, like Lee et al. (2021), used a sophisticated echo-classification algorithm to create labels, called Storm Labeling in Three Dimensions (SL3D; Starzec et al. 2017). Lagerquist et al. (2021a) also experimented with newer and more complex deep learning architectures, specifically the U-net++ (Z. Zhou et al. 2020) and a U-net for time series (Chiu et al. 2020), but found that a simple U-net performed better. Furthermore, Lagerquist et al. (2021a) used a spatial verification (Gilleland et al. 2009) method, called the fractions skill score (FSS; Roberts and Lean 2008), as the loss function for training models. As discussed in Gilleland et al. (2009), the FSS avoids common problems with pixelwise verification, such as the double penalty (where a small displacement between forecast and observed convection is counted as both a false positive and false negative). Lagerquist et al. (2021a) is one of just a few

studies to use spatial verification in the loss function for an ML model, and in follow-on work they use CO forecasting as a testbed for more spatially enhanced loss functions (Lagerquist and Ebert-Uphoff 2022). They have also used CO forecasting as a testbed for uncertainty quantification (UQ) with ML (Haynes et al. 2023), which has been identified as a key research direction in ML applied to the geosciences (Reichstein et al. 2019; Gil et al. 2019). Specifically, they trained U-nets with either Monte Carlo dropout or quantile regression, methods that allow the U-net to estimate its own certainty; they found that quantile regression leads to better-calibrated uncertainty estimates.

b. *Hail*

Hail is the costliest severe weather hazard as measured by property damage (NOAA National Centers for Environmental Information 2022). In the United States, hail annually causes billions of dollars in insured losses to buildings, vehicles, and agriculture, and occasionally injures and kills people. Outside the United States, hail more frequently injures and kills people (Púčik et al. 2019). With accurate predictions of hail with lead times of hours to days, hail losses can be significantly mitigated by relocating vulnerable people, vehicles, and animals into protective shelters ahead of storms. For large businesses, such as car factories and dealerships, farms, airports, or zoos, relocating vulnerable assets can take hours to complete and requires employees to be reassigned from their normal work to intervene. Therefore, accurate forecasts of hail location, timing, and intensity are vital to help mitigate these losses while minimizing false alarms.

Hail formation is not easily modeled by NWP models as formation is dependent on multiple causal factors that must combine to produce large hail that reaches the ground. First, there must be a thunderstorm containing a strong updraft, hail embryos, and sufficient amounts of supercooled liquid water in the vicinity of the updraft that can freeze onto hail embryos lofted by the updraft (Gagne et al. 2019; Dennis and Kumjian 2017). The residence time of hailstones in the supercooled liquid water region is one of the most vital factors in generating large hailstones. This residence time is dependent on both the strength and shape of the storm's updraft, which is influenced by the large-scale environment and small-scale interactions with other storms (Dennis and Kumjian 2017). Even if a storm produces large hail aloft, the hail can melt before it reaches the surface. Given the complexity of large-hail growth, it is not surprising that NWP models struggle with predicting severe hail occurrence, with forecast errors largely attributed to poor representation of hail processes by microphysics schemes (e.g., Snook et al. 2016; Labriola et al. 2017) One approach to address difficulty of explicit prediction of hail within NWP models has been to couple a one-dimensional hail growth model with three-dimensional NWP output (HAILCAST; Adams-Selin and Ziegler 2016).

AI/ML models have had success in forecasting hail at multiple spatial and temporal scales. Because of the linkage of severe hail occurrence to the large-scale environment, AI/ML postprocessing of NWP model output as well as satellite and radar data can provide added skill by conditioning its predictions on variations in mesoscale and storm-scale features. Initial hail machine learning models focused on nowcasting with data from radars (Marzban and Witt 2001) and soundings (López et al. 2007; Manzato 2013), which trained models to link a large set of convective parameters to hail occurrence based on observed hail reports and hail pads. Gagne et al. (2017) extended machine learning hail prediction to day-ahead lead times at a coarser spatial scale by utilizing convection-allowing model output and radar-derived hail size estimates to link predicted storms in the model to their potential to produce severe and significant severe hail. Burke et al. (2020) extended this work with a focus on forecaster trust and ensuring that the AI/ML model developed met the needs of the targeted end users. Yao et al. (2020) also developed a random forest nowcasting approach for hail in China and Czernecki et al. (2019) extended Gagne et al.'s (2017) work to predicting hail in Poland.

All the previously mentioned work has mainly leveraged traditional machine learning methods (e.g., random forest) with some use of neural networks. Since then, there have been several investigations of how CNNs could assist in hail detection and prediction. Namely in 2018, Wang et al. (2018) used CNNs to label radar images that had hail, demonstrating that CNNs could outperform the probability of severe hail method (i.e., Witt et al. 1998). In 2019, Gagne et al. (2019) found that CNNs could statistically outperform all other machine learning methods in predicting future simulated hail (>25 mm) within numerical model output and that the CNN identified linkages between the probability of severe hail and storm mode. The AI/ML algorithms in (Gagne et al. 2019) linked to a microphysics-derived hail size estimate rather than observed hail reports, so the results may not generalize to mapping between simulated storms and observed hail. Gagne et al. hypothesizes that the ability for DL to outperform the traditional ML methods stems from the ability of DL methods, even shallow CNNs, to identify complex spatial patterns in the input data. Pullman et al. (2019) also explored whether CNNs could detect hail using geostationary satellite and reanalysis data, finding good performance in comparison with storm reports, as measured using the critical success index (CSI > 0.4) using their limited dataset. Pulukool et al. (2020) used the Tropical Rainfall Measuring Mission (TRMM) Ku-band radar to label hail locations globally; these data could then be used to train various machine learning methods to detect hail locations from only reanalysis data. Pulukool et al. (2020) found that their random forest outperformed a CNN, but their paper lacks detail on how the random forests were trained and compared (whereas they explained the CNN architecture in more detail). Because their evaluation is done over a large spatial scale, their choice of data also differs significantly from many other hail AI/ML methods discussed in this paper. Given the paucity of hail data available outside the United States, many methods train on NWP output from models over the United States while they used environmental variables derived from the TRMM data.

Hail forecasting with deep learning is a relatively new approach that has seen some development in recent years. For example, during the 2021 NOAA Hazardous Weather Testbed Spring Forecasting Experiment, the developers of the

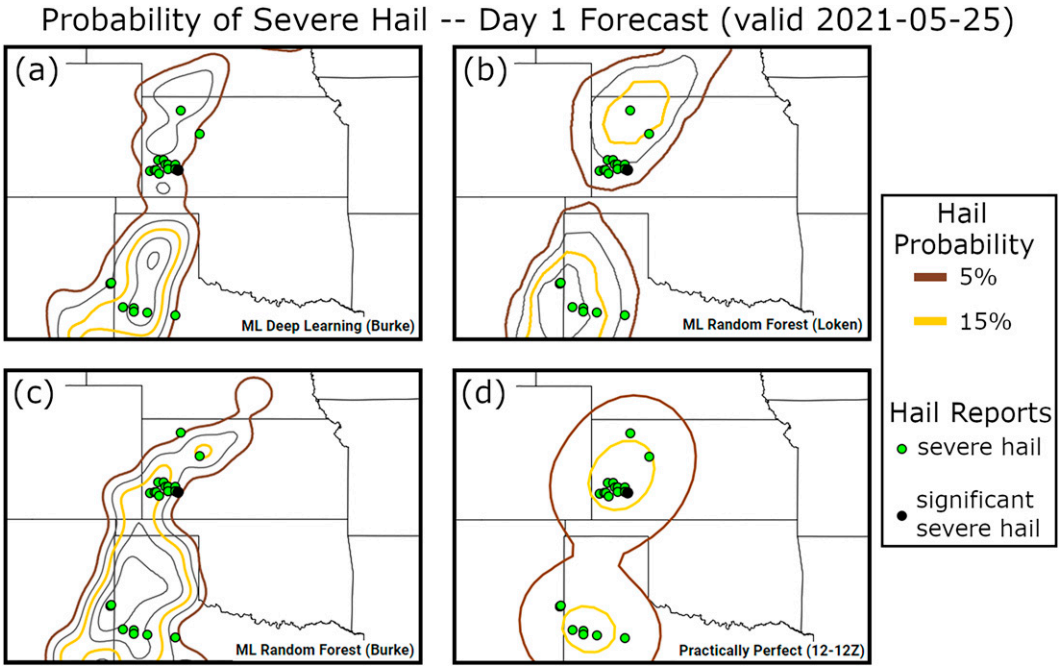

FIG. 1. Day-ahead ensemble probabilistic forecasts of hail produced by (a) a U-net model based upon Burke et al. (2020), (b) the random forest model of L20, and (c) the random forest model of Burke et al. (2020). For comparison, also shown is (d) a practically perfect forecast (an "optimal" forecast calculated from the actual locations of observed hail reports). Contours indicate predicted probability of hail at 5% intervals (5% and 15% contours are highlighted). Observed hail reports are overlaid on each panel, with green circles indicating reports of severe hail [diameter 1–2 in. (25.4–50.8 mm)], and black circles indicating reports of significant severe hail (diameter 2 in. or greater).

Burke et al. (2020) random forest hail prediction model produced real-time hail forecasts via a U-net model using the same set of input variables, obtaining results comparable to that of the random forest model with limited tuning. An example of the day-ahead forecasts produced by these systems using data from the High Resolution Ensemble Forecast (HREF) is shown in Fig. 1; the U-net model (Fig. 1a) highlights similar regions of hail threat as both an independent random forest model (i.e., from Loken et al. 2020; Fig. 1b) and the Burke et al. (2020) random forest model, which uses the same inputs (Fig. 1c). All of the ML models capture the observed hail reports within regions of predicted hail threat and produce forecast probabilities similar in magnitude and extent to those of the practically perfect forecast (an algorithmically generated field intended to represent the optimal forecast that could be drawn if a forecaster knew ahead of time where hail reports would occur; Fig. 1d). More recently, there have been efforts to do near-term forecasts using U-nets where numerical weather prediction is used as an input and radar is used to label hail (Spychalla et al. 2022; Schmidt et al. 2023).

c. *Tornadoes*

Tornadoes are another very costly weather phenomenon, in terms of both life and property (Insurance Information Institute 2019). Tornadoes remain challenging to predict and, while tornado warning time increased throughout the 1990s and into the early 2000s, there has been a recent stagnation in skill due a variety of changes in how warnings are issued and measured (i.e., CSI; Brooks and Correia 2018). Some of the uncertainty in tornado prediction exists because tornadoes continue to be unresolved features within operational NWP models. Thus, the primary method of diagnosing tornadic potential has been the use of environmental parameters (e.g., storm relative helicity) from atmospheric measurements and NWP forecasts.

The use of machine learning for tornadoes is not new, with the first applications in the 1990s (e.g., Marzban and Stumpf 1996). In general, there are two main tasks in ML associated with tornadoes: tornado detection and tornado prediction. Both tasks are on a small spatial scale (tornadoes can be as small as tens of meters wide) and a short temporal scale. Improving tornado detection is useful not only for improving tornado warnings and tornado forecast verification but also to improve detection of tornadoes in low-population areas where they might otherwise go undetected Potvin et al. (2019) suggests that more than half of tornadoes go unreported.

Radar imagery of ongoing convective storms is the principal tool used by NWS forecasters for detecting tornadoes and issuing warnings. Current operational techniques are largely based on expert systems that detect areas of strong rotation (Mitchell et al. 1998). Early researchers explored using machine learning for detecting tornadic rotation signatures (e.g., Marzban 2000; Trafalis et al. 2003; Wang and Yu 2015), but they were restricted by small datasets. A priori, given enough

data, ML should be well suited for tornado prediction on three counts. First, ML can seamlessly leverage data from many sources—unlike the data-assimilation methods used in NWP, which notoriously struggle to leverage satellite data (https://www.ecmwf.int/en/about/media-centre/news/2020/experts-chart-way-ahead-satellite-cloud-and-precipitation) and surface wind observations (e.g., Bédard et al. 2015), for example. Second, because tornadoes are rapidly evolving phenomena—and because just a few minutes of lead time can be enough to allow people to escape harm from these highly localized phenomena—tornado prediction is often pursued as a nowcasting problem. NWP models struggle with nowcasting, due to the well-known spinup problem (e.g., Heng et al. 2020), which does not exist for ML. Third, the genesis, evolution, and decay of tornadoes (and their parent thunderstorms) is highly nonlinear, which makes ML better-suited than traditional statistical methods such as linear regression.

Recently, using a random forest, Sandmæl et al. (2023) trained a tornado detection method on over 10 000 tornadic and nontornadic storms as part of the Probabilistic Hazards Information (PHI) Hazard Services project. The model, called the New Tornado Detection Algorithm (NTDA), starts from automatically detected regions of high rotation and uses multiple nearby radar variables to predict the probability of the rotation being tornadic. The product has been tested in the Hazardous Weather Testbed Spring Forecasting Experiment and has shown promising results.

Like many of the other hazards discussed in this paper, there are two main time scales of tornado prediction: short-term/nowcasting (0–3 h) and next day (or longer). The initial use of machine learning for solely tornado prediction was on the nowcasting time scale, diagnosing whether a storm would produce a tornado in the next 20 min (Marzban and Stumpf 1996); the results indicated that the neural network could outperform the other methods tested (e.g., logistic regression and discriminate analysis), achieving a CSI of 0.3 on Doppler-radar-identified circulations. From there Lakshmanan et al. (2005); Lakshmanan et al. (2007) and Adrianto et al. (2009) used fuzzy logic and support vector machines and neural networks to deviate from the storm-centric approach and produce maps of 0–30-min tornado potential. Adrianto et al. (2009) found improved CSI scores ranging from 0.33 to 0.57 on their test dataset. While these early results looked promising, they were produced using somewhat small datasets, and their test datasets were generally balanced to be nearly 50/50 tornadoes and nontornadoes. This is a problem because, in operations, the actual ratio of tornadoes to nontornadoes is likely less than 1 to 100. Thus, the aforementioned metrics are likely overestimates of true performance. To address this class balance issue, Trafalis et al. (2014) found that adjusting the probability in which events are labeled as tornadic or nontornadic allowed for improved accuracy in predicting tornadoes with support vector machines. Since Trafalis et al. (2014) the latest efforts in nowcasting tornadoes comes from Lagerquist et al. (2020) where CNNs were used to identify tornadic storms from storm-centered radar measurements. Lagerquist et al. (2020) showed good skill (CSI $>$ 0.3) on a dataset of thousands of storms but initial explainable AI (XAI; e.g., Rudin 2018; McGovern et al. 2019; Molnar et al. 2020) showed that the CNN might not perform well for quasi-linear convective system (QLCS) tornadoes.

Beyond nowcasting, other studies began considering the potential of ML to assist in forecasting next-day tornado outbreaks. Mercer et al. (2009) explored whether logistic regression and support vector machines could be leveraged to discriminate between tornado outbreak days and nonoutbreak days. Their results were skillful and indicated that 0–1-km energy helicity index (EHI) was the most important predictor of next-day tornado outbreaks. One caveat Mercer et al. (2009) is that they did not consider the seasonality of tornado outbreaks, where most were in spring and fall, while nontornadic outbreaks were more common during the summer months. As a follow on, Shafer et al. (2010) looked to reevaluate ML methods in light of the seasonality issue and found that, when these seasonal considerations were included, ML performed worse than reported in Mercer et al. (2009). From there, Shafer et al. (2012) extended Shafer et al. (2010) by using more environmental predictors and considering the areal extent of the environmental reports that intersected severe weather reports. The results were largely the same as those of Shafer et al. (2010), but Shafer et al. (2012) showed the significant tornado parameter and EHI were better predictors than all the others they explored. After Shafer et al. (2012), the next exploration of supervised machine learning approaches for prediction of tornado outbreaks came from McGuire and Moore (2022), which used CNNs to diagnose tornadic synoptic regimes within the North American Regional Reanalysis (Mesinger et al. 2006). McGuire and Moore (2022) found that CNNs could accurately (~94% accuracy) diagnose tornado outbreak days (>20 tornadoes) from maps of CAPE, helicity, CIN, 850-hPa heights, and 500-hPa heights.

While we noted earlier that operational models do not have sufficiently fine grid spacing to resolve tornadoes, there are some NWP results capable of resolving tornado-like vortices in a research setting, including Orf et al. (2017) using 30-m horizontal grid spacing and Snook et al. (2019) using 50-m horizontal grid spacing. One study by Steinkruger et al. (2020) leveraged high-resolution models (200 m horizontal grid spacing) and machine learning (logistic regression, random forest, and gradient boosted trees) to automate the forecast of tornado-like vortices in order to simulate tornado warning within the numerical weather prediction model (Fig. 2). Steinkruger et al. (2020) found this to be a skillful and flexible system, but its transition to operation is unclear given that it is trained on high-resolution model data only and current operational NWP models are unable to provide forecasts at such a high resolution.

The discussion thus far has been centered on supervised methods that could be used quasi-autonomously. There have been a few studies that have considered unsupervised methods in order to identify patterns that humans could leverage in an operational setting. Specifically, Nowotarski and Jensen (2013) used self-organizing maps to cluster soundings from near-storm environments and found that self-organizing maps did cluster soundings that lead to tornadogenesis (greater than 70% of the tornadic examples). Another unsupervised machine learning method looked at *k*-means clustering of 500-hPa heights to identify weather regimes that lead to

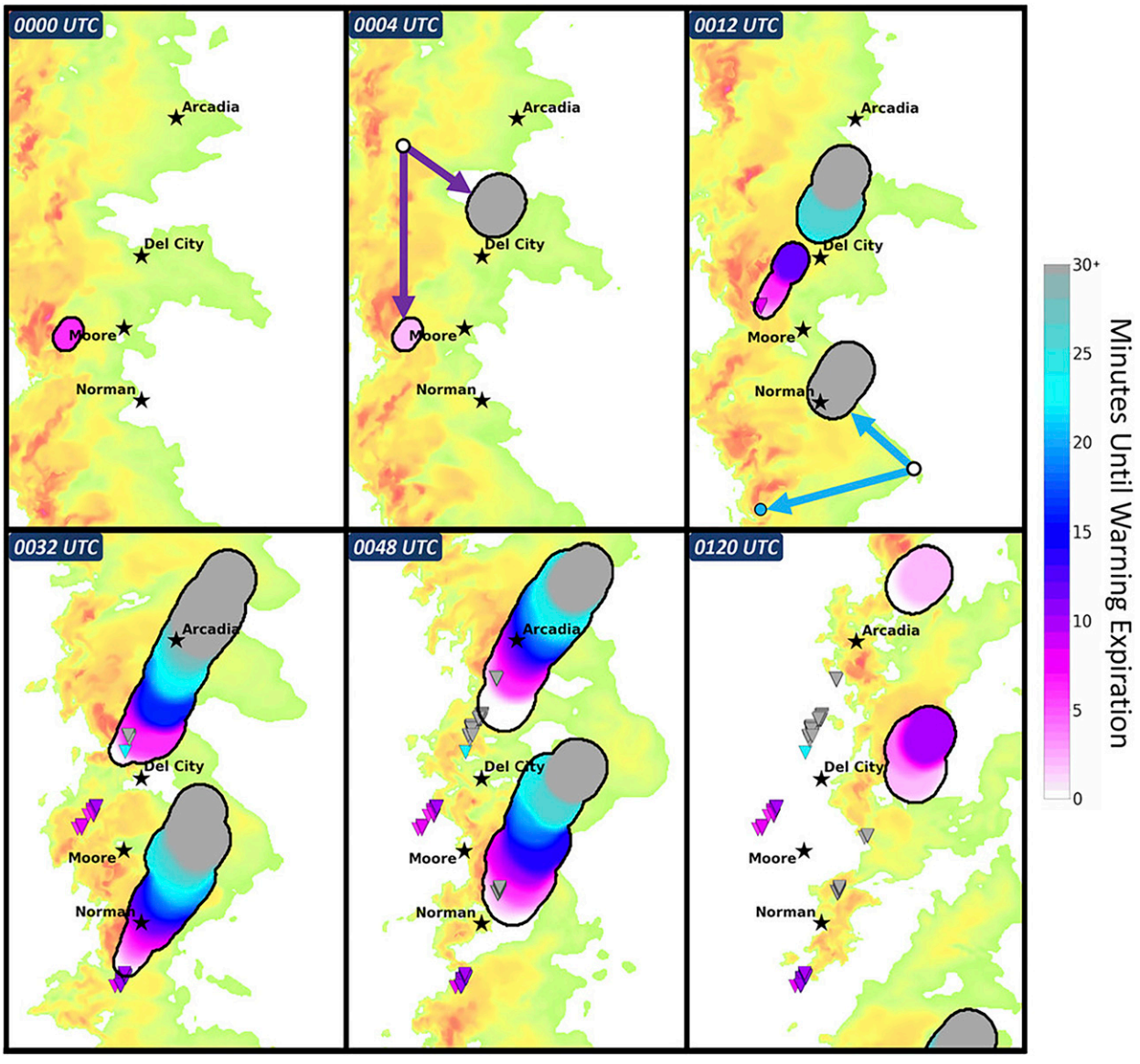

FIG. 2. Figure 14 from Steinkruger et al. (2020), showing an example of localized tornado warning output from an AI system.

tornadoes and tornado outbreaks. Miller et al. (2020) found that for tornado outbreaks, two weather patterns were persistent: a prolonged trough across the Rockies and northern United States with a ridge centered over the east coast (North Carolina and Virginia), and a strong dipole over the United States, with a western trough and an eastern ridge.

d. Straight-line wind

This section focuses on nontornadic severe convective wind, often called "straight-line wind." Although straight-line wind receives less attention than tornadic wind, it is a major hazard, causing multiple deaths and billions of dollars in damage per year in the United States alone. The most costly straight-line wind events are derechos; a derecho is a long-lived bow echo or series of bow echoes, typically associated with a mesoscale convective system (MCS; Coniglio et al. 2004; Corfidi et al. 2016). Some notable examples are the derechos of August 2020 (https://www.weather.gov/dvn/summary_081020), December 2021 (https://www.weather.gov/dmx/StormyandWindyWednesdayDecember152021), and June 2022 (https://www.weather.gov/cle/event_20220613_severe_weather_derecho), all of which occurred over the central United States. When using ML to predict straight-line wind, many studies simply predict "severe weather," which includes straight-line wind in addition to other hazard types, without separating hazard types. There are two reasons for this tradition. First, severe weather is already a rare event (most thunderstorms do not produce severe weather), and each constituent hazard type is an even rarer event than severe weather in aggregate (any severe).

Because rare events are difficult for ML to predict, separating by hazard type makes the prediction task even more difficult. Second, human reports of straight-line wind, often used as labels for ML, are notoriously unreliable, as discussed later in this section.

The earliest work using ML to predict straight-line wind was the Severe Weather Potential (SWP; Kitzmiller et al. 1995) algorithm, which predicted any severe weather (straight-line or tornado or hail) for a given storm cell. The SWP algorithm used linear regression with five predictors, all based on radar-estimated vertically integrated liquid (VIL). The output variable was an index typically ranging from [0, 40] and was positively correlated with severe-weather probability. Marzban and Stumpf (1998) used a neural network to predict any severe wind (both straight-line and tornadic) for a given storm cell with a mesocyclone. Their neural network used 23 radar-derived predictors. Unlike the SWP algorithm, this neural network output a confidence score ranging from [0, 1], which could be more directly interpreted as a probability. Alexiuk et al. (1999) was the first ML study to separate straight-line wind from other hazard types. They experimented with four ML algorithms—neural networks, decision trees, *k* nearest neighbors, and fuzzy clustering—ultimately choosing fuzzy clustering with 22 radar-derived predictors.

More recent work in this domain has used predictors from multiple sources, resulting in better performance than radar predictors only. For example, Cintineo et al. (2014a) used naïve Bayes to create an algorithm called ProbSevere, which forecasts the probability of any severe weather (straight-line wind or tornado or hail) for a given storm cell. The five predictors include maximum estimated size of hail (MESH) from radar data, CAPE and wind shear from NWP data, and cloud glaciation rate and cloud emissivity from satellite data. However, their dataset omits storm cells that produce only straight-line wind and no other hazard types, thus emphasizing tornadoes and hail. Cintineo et al. (2018) updated the ProbSevere algorithm by including lightning as a predictor. Although performance improved overall, they found that on days with primarily straight-line wind throughout the United States (as opposed to the other hazard types), ProbSevere failed to outperform official NWS warnings. Cintineo et al. (2020) developed ProbSevere 2.0, a major update to the ProbSevere algorithm that includes dozens more predictors and separates by hazard type: ProbTor, ProbWind, and ProbHail.

Other recent work includes Lagerquist et al. (2017), who compared several ML algorithms—logistic regression, neural networks, decision trees, random forests, and gradient-boosted forests—for predicting straight-line wind only. Overall, the tree-based methods (random forests and gradient boosted forests) performed the best. They used 431 predictor variables, based on radar images, proximity soundings from NWP, and storm shape/motion derived from a storm-cell-tracking algorithm. The main advantage of Lagerquist et al. (2017) is that they used observations at surface weather stations (in addition to human reports) to create labels, resulting in many more severe-wind observations for both training and evaluating ML models. The main disadvantage is that their sampling technique, which discarded storm cells with too few associated wind observations and no severe-wind observations, probably led to an overestimate of the event frequency (percentage of storms that produce severe wind) and underestimate of the false-alarm ratio (FAR). The forecasting system of Lagerquist et al. (2017) was evaluated during the 2017 NOAA Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE). Last, Pal et al. (2022) developed an ML algorithm to quality-control human reports of severe wind in the NWS local storm reports (LSR) database, removing false reports, which do not exceed the official severe criterion of 25.7 m s^{-1} [50 kt (1 kt $\approx$ 0.51 m s^{-1})]. False reports are a major problem with the LSR database (Doswell et al. 2005). This algorithm was tested during the 2020 (Clark et al. 2021) and 2021 (Clark et al. 2022d) HWT SFE, receiving positive reviews. Iowa State University's ML severe wind probabilities were also tested in the 2022 HWT SFE (Clark et al. 2022a,c). According to the 2022 SFE operations plan(Clark et al. 2022a), these probabilities are based on textual damage reports, 31 parameters from the Storm Prediction Center's (SPC) mesoanalysis page, population density, elevation, and land-use data.

Other recent related work includes Mounier et al. (2022), who investigated the detection of bow echoes via U-nets, with an aim to improve prediction of severe wind. Although they did not focus on convective gusts, Coburn and Pryor (2022) developed a neural network approach to predicting wind gusts in general, with a focus on airports. Also looking at wind gusts in general and not only from convection, Schulz and Lerch (2022) provide an overview of ML methods for ensemble postprocessing geared toward wind prediction.

e. Lightning

While it is sometimes overlooked as a serious thunderstorm hazard, cloud-to-ground (CG) lightning kills more people than tornadoes in the United States and makes every thunderstorm potentially deadly. Emergency managers and event planners rely upon CG lightning forecasts to safeguard sporting events, fairs, and other large outdoor gatherings.

Lightning ML model developers have leveraged both traditional and DL approaches, with much heavier emphasis on the latter in recent years. The traditional learning algorithms of choice have been the random forest (e.g., Blouin et al. 2016; Meyer et al. 2017; Ringhausen et al. 2021; Chmieleweski et al. 2021; Fata et al. 2022) and gradient boosting (e.g., Mostajabi et al. 2019; Leinonen et al. 2022b). Deep learning models for lightning prediction have used CNN-based architectures, many of them augmented to better incorporate temporal correlations in NWP forecast output (Geng et al. 2019, 2021; Lin et al. 2019; K. Zhou et al. 2020; Shrestha et al. 2021; Cintineo et al. 2022; Leinonen et al. 2022a; Zhou et al. 2022). Most lightning ML models nowcast (0–60-min lead times) the occurrence of either a CG stroke or any lightning flash [CG or in cloud (IC)], with forecast intervals ranging from 60 min (e.g., Cintineo et al. 2022) to 5 min (e.g., Chmieleweski et al. 2021). Other models produce longer-range predictions of occurrence, with lead times out to 6 h (Geng et al. 2021), 12 h (Zhou et al. 2022), or even 24 h (Blouin et al. 2016). At least two models have been developed to predict not just lightning occurrence, but flash rate (Chmieleweski et al. 2021;

Shrestha et al. 2021). Another model has been developed to classify flash type into CG versus IC (Ringhausen et al. 2021). The feature sets for lightning ML models range in complexity from a handful of surface observations (Mostajabi et al. 2019) to dozens of predictors from radar, satellite, lightning, and NWP output (e.g., Leinonen et al. 2022b). Predictions may be valid for particular surface stations (Mostajabi et al. 2019), on a grid (e.g., Cintineo et al. 2022), or for storm-based objects (Meyer et al. 2017; Chmieleweski et al. 2021; Leinonen et al. 2022b). Several of the aforementioned lightning models are described in more detail below.

Meyer et al. (2017) trained a random forest (RF) model to predict CG occurrence at lead times of 0–30 min. The predictions are valid over storm-based objects produced by applying the enhanced *k*-means watershed technique to Multi-Radar Multi-Sensor (MRMS) system reflectivity at −10°C. Input features include RUC-based variables characterizing the near-storm environment, MRMS MESH (Witt et al. 1998), reflectivity, azimuthal shear, and vertically integrated liquid variables, and recent lightning observations. The model was positively reviewed by forecasters and emergency managers participating in the NOAA Hazardous Weather Testbed Probabilistic Hazard Information Experiment and subsequently incorporated into the operational MRMS product suite.

Inspired by the success of the Meyer et al. (2017) model, Chmieleweski et al. (2021) developed a set of RF models for application to the NSSL Warn-on-Forecast (Stensrud et al. 2009, 2013) System (WoFS; Wheatley et al. 2015; Jones et al. 2016), a rapidly updating Convective Allowing Model (CAM) ensemble designed to provide high-resolution thunderstorm guidance at 0–6-h lead times, currently planned for operationalization after 2025. To our knowledge, this is the only lightning ML model to leverage CAM ensemble output. Using the approach of Skinner et al. (2018), storm objects are generated from the WoFS and MRMS composite reflectivity fields, then MRMS objects are matched to WoFS objects. Intrastorm and near-storm environment predictors are generated from WoFS output within/near each WoFS storm object that is matched to an MRMS storm object. The RF models predict CG density (rate normalized by object area) within each storm object over the next 90 or 180 min. The RF models are found to be more accurate than even the most sophisticated lightning parameterizations for CG densities of less than 2 strokes per 5 minutes per 25 kilometers squared.

Cintineo et al. (2022) trained a CNN U-net to predict lightning occurrence (CG or IC) over the next 60 min on a 2-km grid (Fig. 3). This model, known as ProbSevere LightningCast, is unique in that it uses predictors solely from the GOES-R Advanced Baseline Imager (ABI). The model was found to generalize well outside of the regions in which it was trained and to provide nearly 20 min of lead time to lightning onset in about half of true-positive cases. The model has been evaluated in the HWT Satellite Proving Ground, a testbed at the NWS Ocean Prediction Center, and at NWS Southern Region Weather Forecast Offices (WFO). While not yet fully operational, experimental LightningCast output is available in near–real time at the University of Wisconsin–Madison/Cooperative Institute for Meteorological Satellite Studies and is already used by many WFOs.

The Geng et al. (2021) LightNet+ model uses a combined CNN and long-short-term memory (LSTM) (ConvLSTM) architecture (Shi et al. 2015) to predict hourly lightning occurrences at 0–6-h lead times on a 4-km grid. The ConvLSTM combines the CNN and LSTM (Goodfellow et al. 2016) methods, facilitating identification of temporal patterns in spatially correlated data. Model features are generated from Weather Research and Forecasting (WRF) Model storm-based variables and surface and lightning observations. LightNet+ is shown to substantially improve upon popular lightning forecasting schemes as well as a model using the StepDeep (Shen et al. 2018) architecture, which takes a different approach (3D convolution) to dealing with spatiotemporal data.

f. Multiple hazard prediction

Some ML methods predict severe hail, wind, and tornadoes based on a common set of predictors. The general approach has been to use many correlated NWP and/or observation-based variables as predictors and observed storm reports as targets. Using the same predictors for each hazard is convenient and allows an ML algorithm to "learn" on its own which variables are most relevant for each hazard, but it sacrifices ML model lightness and possibly skill. Recent studies following this framework use different ML algorithms, preprocessing techniques, underlying dynamical models, predictors, and lead times.

One of the best-known multihazard severe weather prediction algorithms is ProbSevere (PS; Cintineo et al. 2014a, 2018, 2020). The original version of PS (Cintineo et al. 2014a) predicts the probability of an observed storm cell producing a severe weather report (of any hazard type) within 120 min. Predictors include the MESH from radar data, CAPE and wind shear from NWP data, and cloud emissivity and glaciation rate from satellite data. Since its inception, PS has undergone several updates, with Cintineo et al. (2018) adding lightning as a predictor, and Cintineo et al. (2020) establishing ProbSevere v2.0 (PS v2.0), which considers dozens more predictors and makes probabilistic forecasts for each hazard individually. PS v2.0 has been evaluated in the 2017–19 HWT Experimental Warning Program (EWP) Experiments, where most forecasters have indicated the product helps increase warning confidence and lead time (Cintineo et al. 2020).

Other multihazard methods, which predict for longer lead times, rely exclusively on NWP-based predictors. These methods are useful because they provide a way to convert NWP data into explicit severe weather hazard probabilities (e.g., Fig. 4). These methods have frequently been designed to predict for time and spatial scales used by the SPC (i.e., lead times of 1–8 days and within 40 km of a point). Hill et al. (2020, hereinafter H20) and Loken et al. (2020, hereinafter L20) are two of the earliest methods that use RFs to create SPC-style severe weather hazard probabilities based on NWP ensemble data. Both methods predict the probability of severe weather occurring within 40 km of a point, but H20 take predictors from the Second Generation Global Ensemble

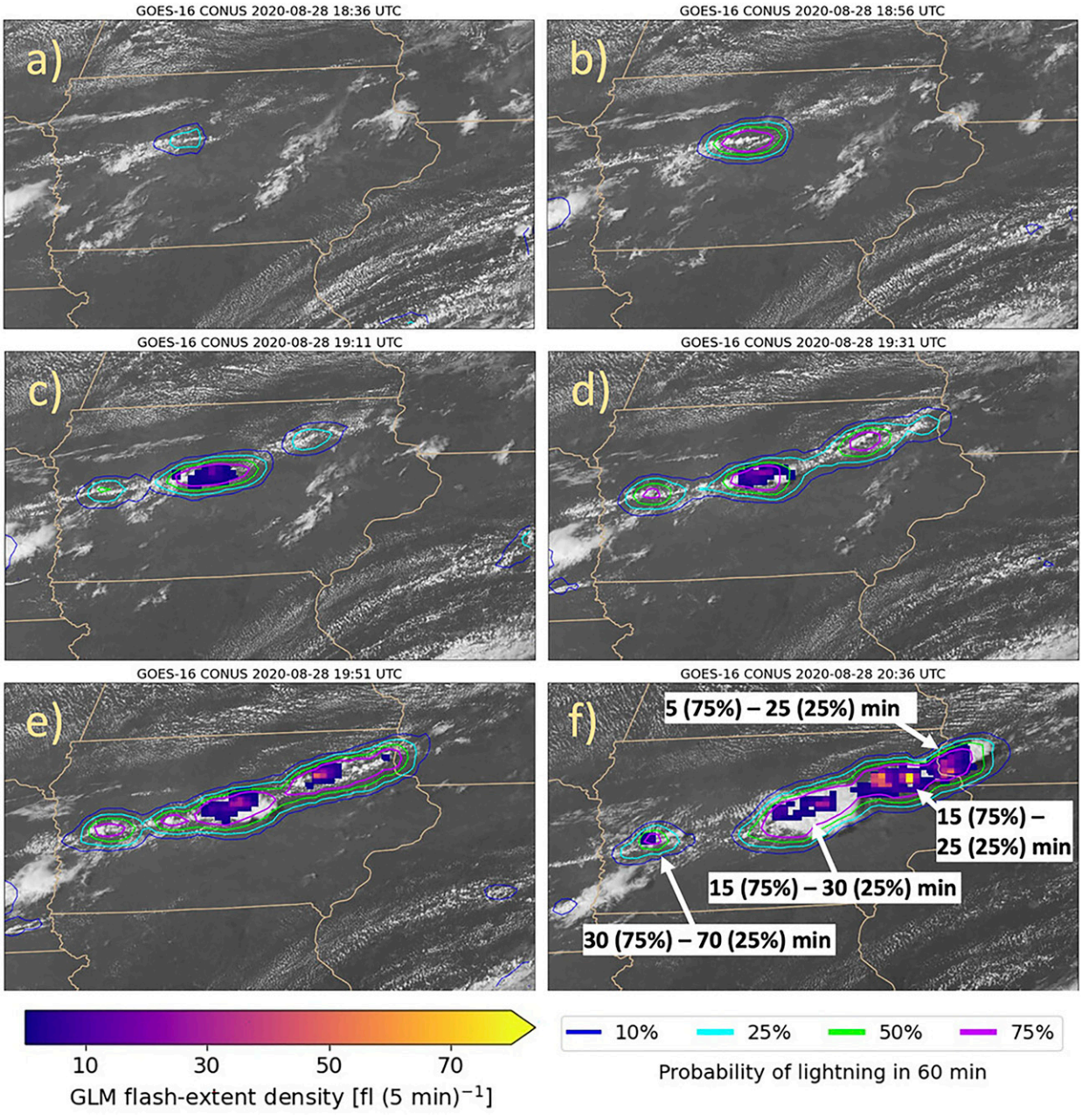

FIG. 3. (a)–(f) A select sequence of images depicting the evolution of LightningCast probabilities along a cold front in Iowa, superimposed on 0.64-μm reflectance (grayscale) and GLM flash-extent density from *GOES-16*. Lead times to the initial GLM flashes for several areas of interest are annotated in (f), showing lead times in minutes from both the 75% and 25% probability thresholds [reproduced, with permission, from Fig. 14 in Cintineo et al. (2022)].

Forecast System Reforecast (GEFS/R), while L20 use predictors from the Storm Scale Ensemble of Opportunity (SSEO) and, subsequently, the HREF (e.g., Loken et al. 2022a, hereinafter L22). Specifically, predictors in H20 are GEFS/R median forecast variables extracted every 3 h (for day-1 and day-2 forecasts) or 6 h (for day-3 forecasts) at the point of prediction and up to 3 horizontal points away. In L20, predictors are derived by first aggregating SSEO forecast variables in time (e.g., by taking a period maximum, minimum, or mean), then upscaling to an 80 km grid, and finally extracting the ensemble mean at the point of prediction and the eight closest 80 km points. Figure 4 shows a real-time example from 11 May 2022 of how preprocessed HREF data are converted to multiple hazard probability forecasts using the method described in L20 and L22. H20 finds that their RF-based method outperforms SPC forecasts for days 2 and 3 but not day 1, while L20

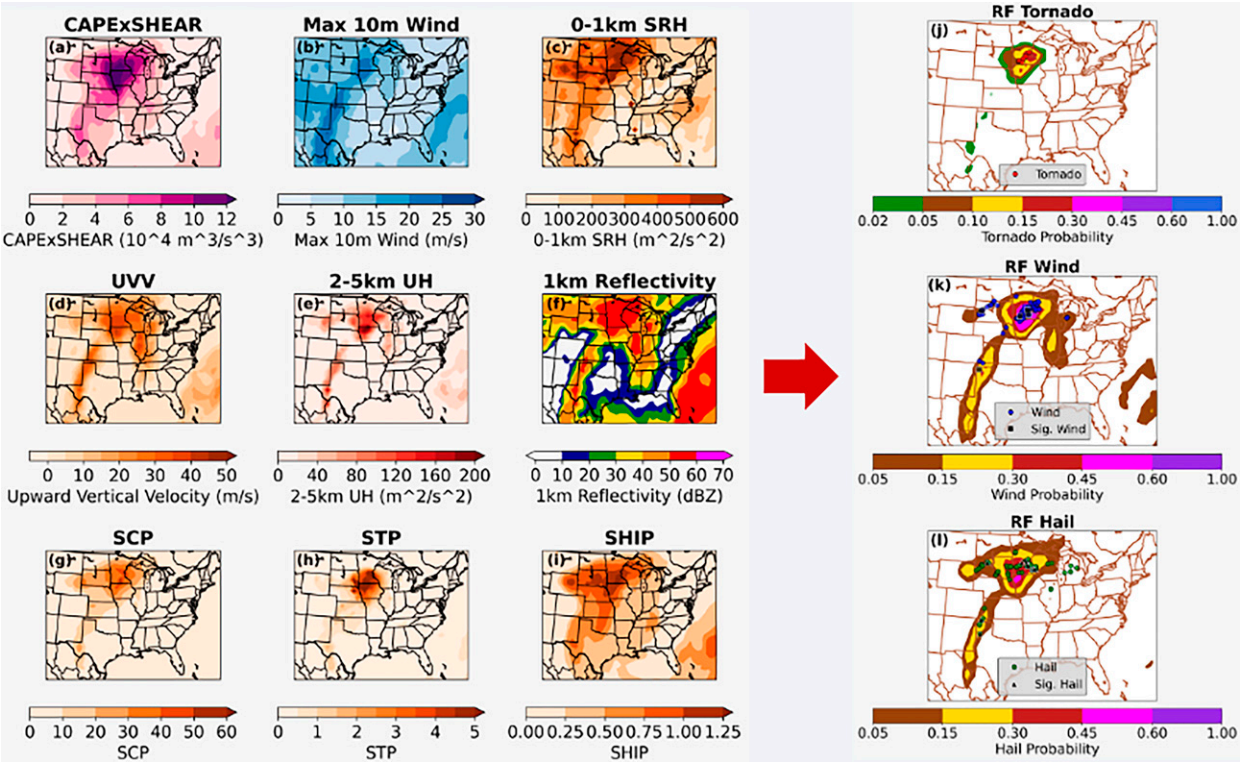

FIG. 4. The 24-h maximum HREF ensemble mean (a) most-unstable CAPE × 10 m–500-hPa wind shear, (b) 10-m wind speed, (c) 0–1-km storm relative helicity (SRH), (d) upward vertical velocity, (e) 2–5-km updraft helicity, (f) 1-km reflectivity, (g) supercell composite parameter (SCP), (h) significant tornado parameter (STP), and (i) significant hail parameter from the 0000 UTC 11 May 2022 initialization. These are used (along with other variables, not shown) to produce RF-based (j) tornado, (k) wind, and (l) hail probabilities, valid from 1200 UTC 11 to 1200 UTC 12 May 2022 using the procedure outlined in L20, L22. In (j)–(l), observed SPC tornado, wind, significant wind, hail, and significant hail reports are indicated by red dots, blue dots, black squares, green dots, and black triangles, respectively.

only analyzes day-1 lead times (i.e., forecast hours 12–36) but finds that their RF forecasts tend to outperform day-1 SPC forecasts for most hazards in most regions and times of the year, presumably due to the inclusion of explicit storm attribute predictors (L22).

Nadocast (Hempel 2022, hereinafter H22), so named because it was originally designed to predict tornadoes, is a more recently developed ML method that now predicts severe hail, wind, and tornadoes at next-day and 4-h lead times. Nadocast uses gradient boosted decision trees (GBDT) and derives most of its predictors from the HREF and Short-Range Ensemble Forecast System (SREF). It uses more than 17 000 and 18 000 predictors from HREF and SREF, respectively. These are broken down into ensemble mean fields, spatial means [within 25-, 50-, and 100-mi. (40, 80, and 161 km, respectively) radii], spatial gradients in multiple directions, and forecast variables at different times. Nadocast originally predicts the likelihood of an observed storm report occurring within 1-h time windows and trains three separate GBDTs per hazard, with each corresponding to a different set of lead times. The hourly predictions are ultimately combined and calibrated using logistic regression to create 24- and 4-h hazard forecasts. While the hail and wind forecasts have not been formally evaluated, the 24-h tornado-predicting version of Nadocast received the greatest mean participant rating of all tornado methods in the 2022 HWT SFE (Clark et al. 2022b).

While many multihazard prediction methods involve tree-based algorithms, Sobash et al. (2020, hereinafter S20) used feed-forward neural networks (NNs) to generate 4-h multiple-hazard severe weather probabilities. They obtained predictors from the deterministic 3-km WRF model [and, subsequently, the High-Resolution Rapid Refresh (HRRR) model; Clark et al. 2022a]. As in L20, NWP forecast variables were first upscaled to 80 km. Predictors were ensemble mean variables at the point of prediction as well as over multiple time and space windows, similar to H22. Unlike Nadocast, however, predictands were observed storm reports falling within 4-h time windows. S20 found their NN hazard forecasts to be more skillful than spatiotemporally calibrated 2–5-km updraft helicity forecasts for all hazards.

Other methods have been applied to predict for multiple severe weather hazards at intraday lead times using WoFS (Wheatley et al. 2015; Jones et al. 2016). Flora et al. (2021, hereinafter F21) develop an object-based method for predicting severe hail, wind, and tornadoes within sliding 30-min windows, up to 150 min of lead time based on WoFS forecast data. Of the three hazards, guidance for severe wind has performed best, especially when compared

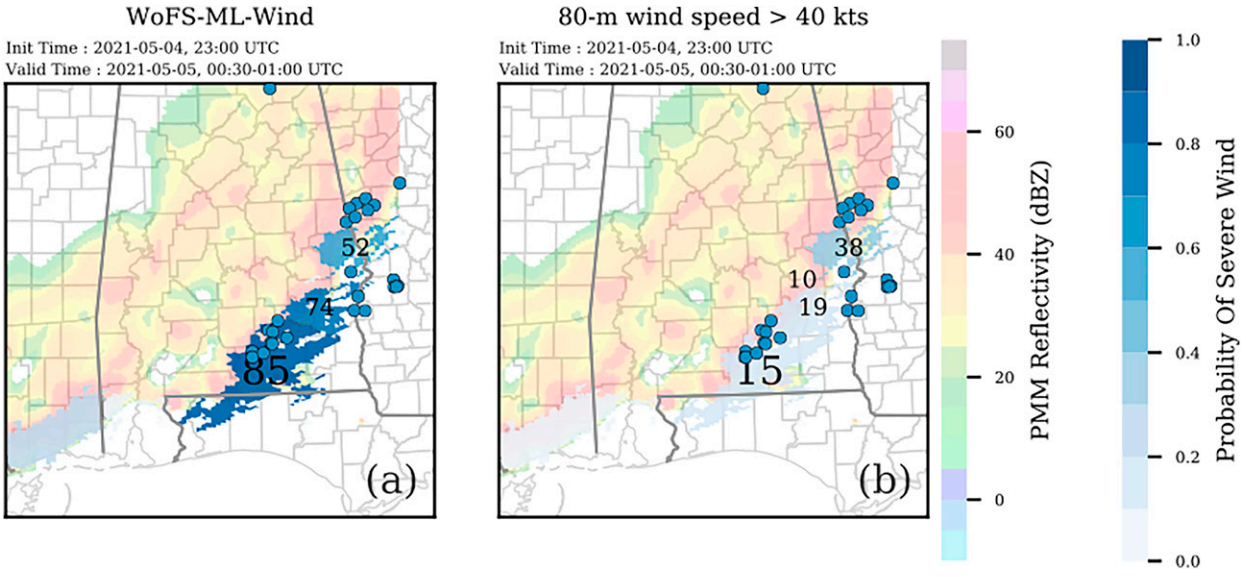

FIG. 5. Representative example comparing the WoFS–ML–wind product vs a carefully calibrated probability of exceedance (80-m wind speed > 40 kt) during 2021 HWT SFE. This forecast is valid for 0030–0100 UTC 5 May 2021. Objects shaded in blue are ensemble storm tracks (Flora et al. 2019, F21) where the probability value is the likelihood of a severe wind report occurring within the object. For reference, the WoFS probability-matched mean of composite reflectivity is shown as a shaded underlay and observed severe wind reports (blue dots) are overlaid.

with a calibrated baseline system based on WoFS's 80-m wind speed prediction (Fig. 5). The object-based framework allows for the creation of sharper probabilities but requires that observed storms are represented in WoFS. Instead of predicting the probability of an observed report at a given point in space (as in H20; L20; S20; H22), F21 predicts the probability that an observed report will fall inside of an ensemble storm track (i.e., the region depicting the ensemble's uncertainty of a storm's location) within the 30-min forecast period plus or minus a 15-min buffer. F21 preprocess WoFS data by taking a 30-min maximum or minimum for storm-related variables and using environmental fields from the beginning of the forecast period. Predictors include spatial means of ensemble means and standard deviations of WoFS variables at each point in the storm track (spatial predictors), ensemble means and standard deviations of each member's 90th-percentile (or 10th percentile) storm-related variables from the points in the storm track (amplitude predictors), and ensemble storm track morphology characteristics (e.g., eccentricity, orientation, minor and major axis length). F21 train RFs, gradient-boosted trees, and logistic regression models to obtain final predictions. They find that the predictions from all algorithms substantially outperform hazard-specific baseline forecasts (i.e., ensemble probabilities of threshold exceedance for 2–5-km updraft helicity, HAILCAST maximum hail diameter, and 80-m wind speed). This improved performance stems from the fact that ML utilizes multiple predictors, while the baseline relies on a single predictor. In terms of specific hazards, they found that WoFS forecasts of 80-m AGL wind speed can underpredict the strength of near-surface winds, which could possibly explain the poor severe wind baseline. Ultimately, they acknowledged that future research is required.

Clark and Loken (2022, hereinafter CL22) also use WoFS to construct short-term ML-based severe weather probabilities. However, unlike F21, CL22 operates in a point-based framework, which allows for nonzero ML probabilities in locations without WoFS storms and includes misses (i.e., null forecasts of observed events) during verification. One cost of operating in a point-based framework is lower precision to achieve the same forecast probability sharpness. Thus, the method of CL22 predicts the probability of any severe weather report occurring within a 39-km radius of a point over the next 3 h. Like many other methods, CL22 use RFs for these predictions. Preprocessing in CL22 is similar to L20 and involves taking the 3-h maximum of storm-related variables and the mean over forecast hours 1, 2, and 3 for environment-related variables at each point. CL22 examine multiple predictor configurations and find the most skillful RFs use predictors from: ensemble mean environment fields; individual-member 2–5 km updraft helicity forecasts; and ensemble maximum, 90th-percentile, and spatially smoothed ensemble mean values of the other storm-related fields. CL22 find this top-performing configuration has significantly better Brier skill score when compared with thresholded and smoothed 2–5-km updraft-helicity-based WoFS probabilities.

While each of the above methods is skillful in isolation, ongoing research is exploring the use of machine learning to combine existing prediction systems for multiple hazard prediction. For example, current work at Cooperative Institute for Severe and High-Impact Weather Research and Operations (CIWRO) and the NSSL is using RFs with predictors

from both PS v2.0 and WoFS to predict the probability of severe weather hazards at lead times between 30 min and 3 h (Loken et al. 2022b).

3. Discussion and future directions

We have provided a thorough review of AI/ML techniques as developed for and applied to convective weather. These include traditional machine learning as well as deep learning for individual convective hazards including wind, lightning, hail, tornadoes, and convective initiation, as well as for multihazard prediction. We discussed methods that span multiple spatial and temporal scales. These methods range from broadscale prediction across wide areas (primarily across CONUS) that are also typically at least one to multiple days in advance. More targeted approaches tend to focus on smaller spatial and temporal scales, focusing on specific regions or on short-range time scales.

Our goal with this review was to first provide an overview of the state of the art and then to motivate readers about where to dive into developing new AI methods for convective and severe weather. As such, the remainder of the paper focuses on specific challenges and motivation for future work.

a. Creating trustworthy AI/ML for convective weather

The end users for weather forecasting products range from other researchers to forecasters, emergency managers, and the general public. While most academic research papers tend to focus on the foundational science that is being addressed (e.g., improving understanding or prediction of a phenomena), there is a critical need to develop trustworthy AI for end users beyond just other researchers. Forecasters and emergency managers are making life-or-death decisions in short time periods and, if they are to use an AI/ML method to help streamline the process of data overload, they must fully trust the product or it will simply be ignored (Hoffman 2017; Hoffman et al. 2017; Karstens et al. 2018; Demuth et al. 2020).

Many of the authors of this paper are part of a large research institute focused on creating trustworthy AI for weather and climate, the NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES; https://www.ai2es.org/). We take a convergent approach to creating trustworthy AI, working across computer science, atmospheric, and ocean sciences, and risk communication/social science. All of these disciplines work together to ensure that the end users we study actually trust the AI methods we develop and to ensure we learn why a method may not be trusted and what needs to change.

One of the key approaches that must be followed in creating trustworthy AI that will actually be used by end users is that of coproduction or codevelopment (Hoffman et al. 2010; Harrison 2022). If ML researchers are creating models that are intended for use by actual end users, for example, they are developing with the end goal of creating an operational model to assist at the forecasting task, they need to include the targeted end users from the beginning of the project. The needs of the end users vary widely, and they drive the choices of the model. This provides a critical challenge for many AI/ML researchers as they must create a working relationship with the end users of the product they are working to create. Often the academic or operational funding calls do not line up to create such relationships, and thus it is a challenge that should be overcome for all types of AI/ML for weather, not just for convective weather.

Involving the end users from the beginning will improve trust in the final product and direct the development of the model itself. The end users know what the specifics of the prediction problems are that must be solved. They know what is hard (and what is not), and where ML assistance would be most beneficial. And, as the experts on the domain, the end users can guide the ML developers on what are the best input variables as well as the best sources of labels (which is not always straightforward for severe weather).

As an example of ensuring that AI/ML researchers are meeting the needs of their targeted end users, in the work of the authors of this paper, we focus on creating trustworthy ML methods that will eventually be used for operational forecasting. This focus is different than many ML researchers who are only focused on creating the most highly skilled model. In some cases, the most highly skilled model is not the model that is most trusted or preferred by forecasters. An interesting example of this arose on 18 May 2022 during the HWT SFE. When evaluating two versions of the NCAR neural network probabilistic tornado guidance (Fig. 6; S20), many participants who were forecasters stated they preferred version 2 (Fig. 6b), which had higher tornado probabilities, even though no tornadoes—only tornado warnings—were observed in the SFE domain. Although version 1 (Fig. 6a) had the greater objective performance, version 2 offered guidance that was more consistent with the forecasters' desired messaging (i.e., that this was a day with a nonzero tornado threat). By working with the forecasters, AI/ML researchers can ensure that they are developing and providing products that will be most useful to them as opposed to products that purely maximize skill scores.

A related challenge is we have not definitively established what ML does really well versus what humans do really well, and how their strengths might complement each other. McCloskey et al. (2022) takes the first step with this, but we need more studies that can tackle this question to obtain the most useful, trustworthy AI systems. Human–AI teams are being studied across a wide range of applications and have immense import in the weather world (Stuart et al. 2022).

Many of the remaining challenges discussed below also relate to creating trustworthy AI, but we felt they were also challenges that need to be addressed on their own as well and thus we put them in separate sections.

b. Ensuring AI/ML methods follow the laws of physics

Ensuring that AI/ML models respect the laws of physics is a key unsolved challenge for AI/ML model development for all aspects of weather and climate forecasting. We highlight it here especially for the convective scale, where there are many small-scale processes that are critical to forecasting convection and must be simulated in a physics-based manner in order to ensure that the output is realistic. While many ML models can be trained based entirely on observational data, in

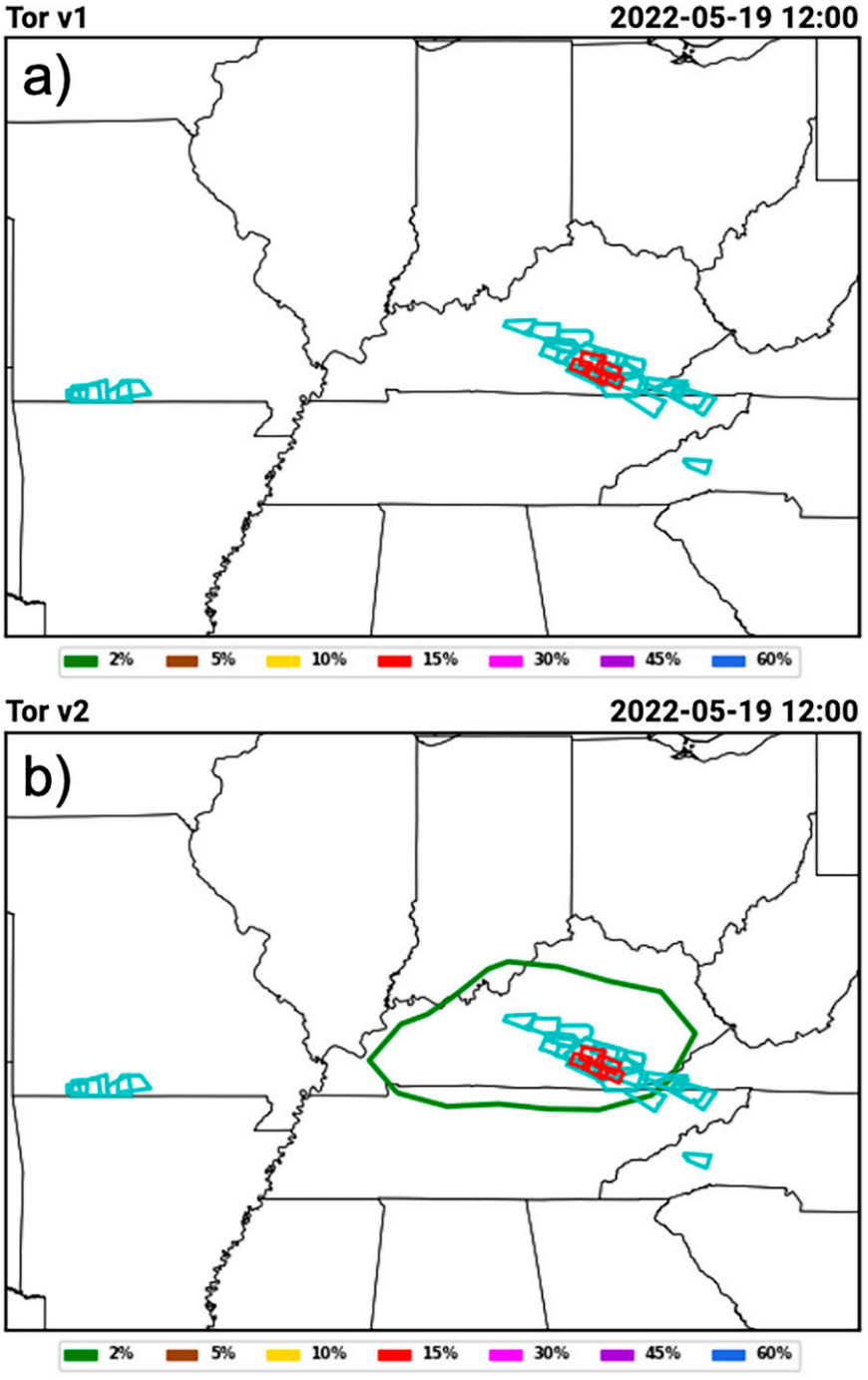

FIG. 6. Contoured (a) version 1 and (b) version 2 NCAR NN-based tornado probabilities (S20) tested during the 2022 HWT SFE, valid from 1200 UTC 18 to 1200 UTC 19 May 2022. Severe thunderstorm and tornado warning polygons are overlaid in blue and red, respectively. The figure is adapted from the SFE Viewer website (SPC 2022).

our experience, the models must also be constrained by the laws of physics, or they can learn patterns that cannot occur in the real world. With the goal of creating trustworthy ML, it is critical that the models respect the laws of physics both in their successful predictions and in their unsuccessful ones.

Some methods already exist for incorporating physical laws into ML. These include custom loss functions (Ebert-Uphoff et al. 2021), which punish the ML model for violating physical laws, and custom architectures, which use equations to enforce physical laws inside the model (e.g., Lagerquist et al.

2021b; Beucler et al. 2019a,b). See Beucler et al. (2019b) for a comparison of the two approaches.

c. Explainable AI

Another open challenge in AI/ML for convective weather that is related to both creating trustworthy models and to creating models that follow the laws of physics is that of XAI (Ras et al. 2018; Molnar 2018; Mueller et al. 2019; Miller 2019; Samek et al. 2019). XAI provides end users a way to look inside the traditionally "black box" AI/ML model. This is critical for end users in weather and climate (McGovern et al. 2019) as it increases trust in the model and allows end users to verify that the model is performing as expected in a variety of situations.

As an example, the reason many of the authors initially chose to use random forests and gradient boosted trees for hail prediction instead of deep learning was because there are well-known XAI methods for these algorithms. Using these methods, in conjunction with testing our methods in NOAA's HWT enabled us to create a method that was more transparent to the end users and was validated to be more trustworthy in the HWT (Clark et al. 2022d).

d. Uncertainty quantification

Forecasters and emergency managers already understand that no model is perfect (Cains et al. 2022), especially when forecasting convection and related hazards. However, forecasters do want to know when they should conditionally trust the predictions of a model and when they should not. Prediction uncertainty estimates provide a relatively concise way to synthesize the uncertainty propagated from variations in possible model realizations (epistemic uncertainty) and variation in the training data (aleatoric uncertainty). For severe weather forecasting, the relationship between the larger-scale atmospheric environment and the microscale hazards of tornadoes and hail is inherently noisy (aleatoric uncertainty), and some combinations of large-scale environmental parameters and severe hazard occurrence are sparsely sampled in the historical record, so a wide range of models could fit the data equally well (epistemic uncertainty). There is additional uncertainty introduced by biases in how LSR are collected by the National Weather Service (Doswell et al. 1999), increasing the aleatoric uncertainty.

To incorporate sources of uncertainty into the ML training process, many groups (Gagne et al. 2014, 2017, 2019; L20; F21) have trained ML postprocessing models from NWP ensemble output, either from individual members or ensemble averages. The NWP ensemble accounts for uncertainties in the initial conditions and sometimes in the physical assumptions, if a multiphysics or multimodel ensemble is used. Most applications of ML to NWP ensembles postprocess the ensemble information into a single probability of hazard occurrence, which aligns with the NWS Storm Prediction Center Convective Outlook. If the probability is calibrated, then it can serve as an estimate of the uncertainty for binary classification problems (Burke et al. 2020). However, that probability estimate has its own associated uncertainties that could be estimated by or by utilizing ML models that produce parametric uncertainty estimates (e.g., Rasp and Lerch 2018; Scheuerer et al. 2020; Grönquist et al. 2021).

Other groups have combined ML with physical ensembles in a way that preserves or even improves uncertainty estimates, but to our knowledge there are no such applications to severe weather. This could prove to be fruitful research given the inherent uncertainty to severe weather. For example, uncertainty quantification techniques for severe weather hazards have been more extensively applied to the problem of estimating societal risk to different severe weather hazards. More research is needed on the best uncertainty quantification methods for severe weather (and geoscience applications in general; Gil et al. 2019; Reichstein et al. 2019; Haynes et al. 2023), as well as the best ways to communicate uncertainty with end users. This latter is especially crucial for convective weather, where end users are making time-critical decisions.

e. Research to operations: Cross sector collaborations

Another challenge to developing AI/ML for convective weather is to actually validate the models with the targeted end users, many of whom are making critical decisions in very short time periods. Since many of the operational end users have access to real-time data feeds that are different than the archived data available to train ML models, this can be a significant challenge. As with challenges discussed above, funding can be a hindrance in that there is often little to no funding for testing operational models nor rewards in the academic system for producing such a system. Instead of suggesting we reframe the academic reward system, we propose that creating cross-sector testbeds such as the HWT will jumpstart research all the way from foundations to operations. AI/ML researchers can observe how products are actually used, facilitating coproduction as discussed above.

The experience of the authors and others in NOAA testbeds has shown that forecasters often find uses for experimental guidance that were not envisioned by the developers. Forecasters are particularly adept at finding value in imperfect model predictions, and identifying scenarios where the guidance has especially high or low value. This feedback enables the developers to improve the guidance, and in turn, the developers guide the forecasters in using the guidance. The developers also play the important role of demystifying the novel products, thereby enhancing forecaster trust.

f. AI for science discovery

Humans are overloaded by data and new sensors, and this is where AI/ML methods can really shine. For example, AI/ML can be used to process hundreds of simulations of an event to identify common patterns, something that would be very challenging for a human to do. However, the promise of using AI/ML methods to identify *fundamental* new science discoveries in convective weather, such as identifying why one storm is tornadic and another similar storm never generates a tornado (e.g., McGovern et al. 2014, 2017b), remains elusive. Sophisticated ML methods that are strongly tailored to the phenomena of interest (e.g., McGovern et al. 2014, 2017b)

may be required to advance our understanding of severe weather processes that have been studied for decades. This area is prime for additional research.

For example, ML tornado applications have primarily focused upon detection and prediction, not characterization. Many relationships have been identified between tornado properties and storm mode, environment, and radar- and satellite-measured attributes (e.g., Thompson et al. 2003; Grams et al. 2012; Bodine et al. 2013; Van Den Broeke and Jauernic 2014; Kingfield and LaDue 2015; Smith et al. 2015; Gibbs 2016; Thompson et al. 2017; Marion et al. 2019; Sessa and Trapp 2020; French and Kingfield 2021). ML could exploit these and perhaps yet unknown relationships between tornado attributes and regularly measured atmospheric variables to produce real-time estimates and short-term predictions of tornado intensity, damage, width, and duration.

g. Other challenges and areas for future work

Another challenge in training ML for severe weather forecasting is the rarity/sparsity of the data as well as the imperfections/biases in the data themselves. This comes from both the rarity of events such as tornadoes (over a 3D volume of the United States, at any given time, there are almost never tornadoes occurring) and from the rarity of a well-curated database of labels. For example, hail reports are known to occur more frequently in larger cities and along highways and are underreported if larger-scale hazards such as tornadoes occur (Allen and Tippett 2015). Tornado reports are similarly biased (Potvin et al. 2019, 2022). Wind reports have also been shown to contain nonmeteorological artifacts, including clustering of reports at wind speeds ending in 0 or 5 mi h^{-1} or kt and spatial discontinuities in the distribution of reports (Edwards et al. 2018) If a developer does not handle the data bias issues, the ML model may not be able to be trained to be a reliable predictor of the phenomena being studied (McGovern et al. 2022). Evaluation of the ML model will similarly be inhibited by data bias. This motivates development of methods to improve the severe weather database, perhaps using historical NEXRAD data (e.g., the MRMS azimuthal shear and MESH products), or to account for the potential for missing reports, for example, by weighting the training loss function or verification metrics proportionally to population density.

In a related example, multiple ongoing challenges will continue to affect machine learning hail forecasting efforts. Chief among them is the quality and trade-offs of different hail report datasets. The current standard dataset in the United States is the National Weather Service Storm Data hail reports, which offers the longest continuous record of hail reports and the broadest coverage but is subject to numerous biases, including population biases near cities and roads, reports assigned to city centers, underestimation of maximum hailstone size, and limited reporting of subsevere hail (Allen and Tippett 2015; Blair et al. 2017). Crowdsourced reporting datasets, such as that of the mobile Precipitation Identification Near the Ground project (mPING; Elmore et al. 2014), provide better coverage of subsevere hail reports but also have significant population biases. Hail pads and other fixed location networks (Manzato 2013) can capture diameter more accurately but will likely miss the largest hailstones in an area. Radar-derived hail estimates provide better coverage but have been found to not always discriminate hail size well when compared with high-resolution reports (Cintineo et al. 2012). Newer radar calibration algorithms that use a larger and more quality-controlled report dataset (Murillo and Homeyer 2019) and that use dual polarimetric radar variables to account for the thermodynamic properties of hail (Brook et al. 2021) should further refine radar-based hail location and size estimates to enable finer-scale hail forecasting at all lead times.

Another challenge to training ML for severe weather is the cost (both economic and societal) for missing a forecast. While this cost exists to some degree in many traditional ML problems, it is literally often life-or-death in severe weather forecasting. For example, missing a tornado has a much higher cost than misclassifying a digit in a handwriting recognition system. Even if the systems are designed to work together with human forecasters, they should not miss frequently or the humans will lose trust. This ties back into the first statement, that the systems should be developed alongside the targeted end users and always ensuring the needs of those end users are met.

The cost for missing an event even differs from tornado to tornado, with long-lived significant tornadoes passing through populated areas having much more of a penalty than brief squall line spinups in open fields. This poses a challenge because most current ML algorithms treat all reports as identical, not only in quality, but also potential impact during training and real-time prediction. Yet, it is unclear how or even if ML should account for differences between reports of the same class.

Studies of potential changes in severe weather frequency under climate change have generally leveraged convection-allowing simulations downscaled from climate model projections (e.g., Gensini and Mote 2015; Hoogewind et al. 2017; Trapp et al. 2019; Molina et al. 2021; Ashley et al. 2022) or empirical relationships between severe weather occurrence and large-scale environment (e.g., Trapp et al. 2009; Marsh et al. 2009; Diffenbaugh et al. 2013; Gensini et al. 2014). ML can facilitate the latter approach by generating complex statistical models of severe weather potential from numerous environmental variables (e.g., Gensini et al. 2021), and may therefore enable better estimates of severe weather activity trends over the next century.

Studies that have included prior lightning data as input to lightning ML classifiers and analyzed feature importance have found most of the prediction skill is attributable to prior lightning data (Meyer et al. 2017; Geng et al. 2021; Leinonen et al. 2022b). This is expected, since a persistence forecast will generally verify well for lightning; for example, once lightning begins, it is likely to persist for tens of minutes. However, it is particularly valuable (albeit challenging) to predict when lightning will begin and end. We therefore recommend that lightning model verification place more focus on lightning onset and cessation to provide a truer assessment of operational value. The importance of focusing model evaluation on hazard

transitions has recently been discussed in the context of fog nowcasting by Vorndran et al. (2022).

The authors are continuing research on the topic of creating trustworthy ML for severe weather prediction, taking into account the challenges presented above. We challenge all readers of the review paper to join us on the challenging open tasks of developing trustworthy AI/ML methods for all varieties of weather hazards, to save lives and property in high-impact weather.

Acknowledgments. Funding for authors Flora and Potvin was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce. For authors McGovern, Chase, Gagne, and Snook, this material is based upon work supported by the National Science Foundation under Grant ICER-2019758. This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement 1852977.

Data availability statement. As a review paper, this paper has no data produced.

REFERENCES

Adams-Selin, R. D., and C. L. Ziegler, 2016: Forecasting hail using a one-dimensional hail growth model within WRF. *Mon. Wea. Rev.*, **144**, 4919–4939, https://doi.org/10.1175/MWR-D-16-0027.1.

Adrianto, I., T. M. Smith, K. A. Scharfenberg, and T. B. Trafalis, 2005: Evaluation of various algorithms and display concepts for weather forecasting. *21st Int. Conf. on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, San Diego, CA, Amer. Meteor. Soc., 5.7, https://ams.confex.com/ams/pdfpapers/83928.pdf.

——, T. B. Trafalis, and V. Lakshmanan, 2009: Support vector machines for spatiotemporal tornado prediction. *Int. J. Gen. Syst.*, **38**, 759–776, https://doi.org/10.1080/03081070601068629.

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

Alexiuk, M., N. Pizzi, and W. Pedrycz, 1999: Classification of volumetric storm cell patterns. *1999 IEEE Canadian Conf. on Electrical and Computer Engineering*, Edmonton, AB, Canada, IEEE, 1081–1085, https://doi.org/10.1109/CCECE.1999.808201.

Allen, C. T., S. E. Haupt, and G. S. Young, 2007: Source characterization with a genetic algorithm–coupled dispersion–backward model incorporating SCIPUFF. *J. Appl. Meteor. Climatol.*, **46**, 273–287, https://doi.org/10.1175/JAM2459.1.

Allen, J. T., and M. K. Tippett, 2015: The characteristics of United States hail reports: 1955–2014. *Electron. J. Severe Storms Meteor.*, **10** (3), https://ejssm.com/ojs/index.php/site/article/view/60.

Angevine, W. M., J. Olson, J. J. Gristey, I. Glenn, G. Feingold, and D. D. Turner, 2020: Scale awareness, resolved circulations, and practical limits in the MYNN-EDMF boundary layer and shallow cumulus scheme. *Mon. Wea. Rev.*, **148**, 4629–4639, https://doi.org/10.1175/MWR-D-20-0066.1.

Ashley, W. S., A. M. Haberlie, and V. A. Gensini, 2022: The future of supercells in the United States. *Bull. Amer. Meteor. Soc.*, **104**, E1–E21, https://doi.org/10.1175/BAMS-D-22-0027.1.

Bauer, P., P. D. Dueben, T. Hoefler, T. Quintino, T. C. Schulthess, and N. P. Wedi, 2021: The digital revolution of Earth-system science. *Nat. Comput. Sci.*, **1**, 104–113, https://doi.org/10.1038/s43588-021-00023-0.

Bédard, J., S. Laroche, and P. Gauthier, 2015: A geo-statistical observation operator for the assimilation of near-surface wind data. *Quart. J. Roy. Meteor. Soc.*, **141**, 2857–2868, https://doi.org/10.1002/qj.2569.

Beucler, T., M. Pritchard, S. Rasp, J. Ott, P. Baldi, and P. Gentine, 2019a: Enforcing analytic constraints in neural-networks emulating physical systems. arXiv, 1909.00912v5, https://doi.org/10.48550/arXiv.1909.00912.

——, S. Rasp, M. Pritchard, and P. Gentine, 2019b: Achieving conservation of energy in neural network emulators for climate modeling. arXiv, 1906.06622v1, https://doi.org/10.48550/arXiv.1906.06622.

Billet, J., M. DeLisi, B. G. Smith, and C. Gates, 1997: Use of regression techniques to predict hail size and the probability of large hail. *Wea. Forecasting*, **12**, 154–164, https://doi.org/10.1175/1520-0434(1997)012<0154:UORTTP>2.0.CO;2.

Blair, S. F., and Coauthors, 2017: High-resolution hail observations: Implications for NWS warning operations. *Wea. Forecasting*, **32**, 1101–1119, https://doi.org/10.1175/WAF-D-16-0203.1.

Blouin, K. D., M. D. Flannigan, X. Wang, and B. Kochtubajda, 2016: Ensemble lightning prediction models for the province of Alberta, Canada. *Int. J. Wildland Fire*, **25**, 421–432, https://doi.org/10.1071/WF15111.

Bodine, D. J., M. R. Kumjian, R. D. Palmer, P. L. Heinselman, and A. V. Ryzhkov, 2013: Tornado damage estimation using polarimetric radar. *Wea. Forecasting*, **28**, 139–158, https://doi.org/10.1175/WAF-D-11-00158.1.

Boukabara, S.-A., V. Krasnopolsky, J. Q. Stewart, S. G. Penny, R. N. Hoffman, and E. Maddy, 2019: Artificial intelligence may be key to better weather forecasts. *Eos*, **100**, https://doi.org/10.1029/2019EO129967.

Brenowitz, N. D., and C. S. Bretherton, 2018: Prognostic validation of a neural network unified physics parameterization. *Geophys. Res. Lett.*, **45**, 6289–6298, https://doi.org/10.1029/2018GL078510.

——, and ——, 2019: Spatially extended tests of a neural network parametrization trained by coarse-graining. *J. Adv. Model. Earth Syst.*, **11**, 2728–2744, https://doi.org/10.1029/2019MS001711.

Brook, J. P., A. Protat, J. Soderholm, J. T. Carlin, H. McGowan, and R. A. Warren, 2021: HailTrack—Improving radar-based hailfall estimates by modeling hail trajectories. *J. Appl. Meteor. Climatol.*, **60**, 237–254, https://doi.org/10.1175/JAMC-D-20-0087.1.

Brooks, H. E., and J. Correia Jr., 2018: Long-term performance metrics for national weather service tornado warnings. *Wea Forecasting*, **33**, 1501–1511, https://doi.org/10.1175/WAF-D-18-0120.1.

Bryan, G. H., J. C. Wyngaard, and J. M. Fritsch, 2003: Resolution requirements for the simulation of deep moist convection. *Mon. Wea. Rev.*, **131**, 2394–2416, https://doi.org/10.1175/1520-0493(2003)131<2394:RRFTSO>2.0.CO;2.

Burke, A., N. Snook, D. J. Gagne II, S. McCorkle, and A. McGovern, 2020: Calibration of machine learning-based probabilistic hail predictions for operational forecasting. *Wea. Forecasting*, **35**, 149–168, https://doi.org/10.1175/WAF-D-19-0105.1.

Cains, M. G., and Coauthors, 2022: NWS forecasters' perceptions and potential uses of trustworthy AI/ML for hazardous

weather risks. *21st Conf. on Artificial Intelligence for Environmental Science*, Houston, TX, Amer. Meteor. Soc., 1.3, https://ams.confex.com/ams/102ANNUAL/meetingapp.cgi/Paper/393121.

Chase, R. J., D. R. Harrison, A. Burke, G. M. Lackmann, and A. McGovern, 2022: A machine learning tutorial for operational meteorology. Part I: Traditional machine learning. *Wea. Forecasting*, **37**, 1509–1529, https://doi.org/10.1175/WAF-D-22-0070.1.

——, ——, G. Lackmann, and A. McGovern, 2023: A machine learning tutorial for operational meteorology. Part II: Neural networks and deep learning. *Wea. Forecasting*, https://doi.org/10.1175/WAF-D-22-0187.1, in press.

Chiu, H., E. Adeli, and J. C. Niebles, 2020: Segmenting the future. *IEEE Robot. Autom. Lett.*, **5**, 4202–4209, https://doi.org/10.1109/LRA.2020.2992184.

Chmieleweski, V. C., C. Potvin, P. S. Skinner, A. E. Reinhart, E. R. Mansell, and K. M. Calhoun, 2021: How well can we forecast cloud-to-ground lightning rates within the NSSL Experimental Warn-on-Forecast system using machine learning? *10th Conf. on the Meteorological Application of Lightning Data*, Online, Amer. Meteor. Soc., 5.4, https://ams.confex.com/ams/101ANNUAL/meetingapp.cgi/Paper/380582.

Cintineo, J. L., T. M. Smith, V. Lakshmanan, H. E. Brooks, and K. L. Ortega, 2012: An objective high-resolution hail climatology of the contiguous United States. *Wea. Forecasting*, **27**, 1235–1248, https://doi.org/10.1175/WAF-D-11-00151.1.

——, M. J. Pavolonis, J. M. Sieglaff, and D. T. Lindsey, 2014a: 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 ——, 2014b: 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. J. Pavolonis, J. M. Sieglaff, L. Cronce, and J. 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.

——, ——, and ——, 2022: ProbSevere LightningCast: A deep-learning model for satellite-based lightning nowcasting. *Wea. Forecasting*, **37**, 1239–1257, https://doi.org/10.1175/WAF-D-22-0019.1.

Clark, A. J., and E. D. Loken, 2022: Machine learning-derived severe weather probabilities from a Warn-on-Forecast system. *Wea. Forecasting*, **37**, 1721–1740, https://doi.org/10.1175/WAF-D-22-0056.1

——, A. MacKenzie, A. McGovern, V. Lakshmanan, and R. Brown, 2015: An automated, multi-parameter dryline identification algorithm. *Wea. Forecasting*, **30**, 1781–1794, https://doi.org/10.1175/WAF-D-15-0070.1

——, and Coauthors, 2021: A real-time, virtual spring forecasting experiment to advance severe weather prediction. *Bull. Amer. Meteor. Soc.*, **102**, E814–E816, https://doi.org/10.1175/BAMS-D-20-0268.1

——, and Coauthors, 2022a: Spring Forecasting Experiment 2022 Conducted by the Experimental Forecast Program of the Hazardous Weather Testbed: Program overview and operations plan. NOAA, 43 pp., https://hwt.nssl.noaa.gov/sfe/2022/docs/HWT_SFE2022_operations_plan.pdf.

——, and Coauthors, 2022b: Spring forecasting experiment 2022 conducted by the experimental forecast program of the NOAA hazardous weather testbed: Preliminary findings and results. NOAA, 95 pp., https://hwt.nssl.noaa.gov/sfe/2022/docs/HWT_SFE_2022_Prelim_Findings_FINAL.pdf.

——, and Coauthors, 2022c: The third real-time, virtual spring forecasting experiment to advance severe weather prediction capabilities. *Bull. Amer. Meteor. Soc.*, **104**, E456–E458, https://doi.org/10.1175/BAMS-D-22-0213.1.

——, and Coauthors, 2022d: The second real-time, virtual spring forecasting experiment to advance severe weather prediction. *Bull. Amer. Meteor. Soc.*, **103**, E1114–E1116, https://doi.org/10.1175/BAMS-D-21-0239.1.

Coburn, J., and S. C. Pryor, 2022: Do machine learning approaches offer skill improvement for short-term forecasting of wind gust occurrence and magnitude? *Wea. Forecasting*, **37**, 525–543, https://doi.org/10.1175/WAF-D-21-0118.1.

Coniglio, M. C., D. J. Stensrud, and M. B. Richman, 2004: An observational study of derecho-producing convective systems. *Wea. Forecasting*, **19**, 320–337, https://doi.org/10.1175/1520-0434(2004)019<0320:AOSODC>2.0.CO;2.

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

Czernecki, B., M. Taszarek, M. Marosz, M. Półrolniczak, L. Kolendowicz, A. Wyszogrodzki, and J. Szturc, 2019: Application of machine learning to large hail prediction—The importance of radar reflectivity, lightning occurrence and convective parameters derived from ERA5. *Atmos. Res.*, **227**, 249–262, https://doi.org/10.1016/j.atmosres.2019.05.010.

Demuth, J. L., and Coauthors, 2020: Recommendations for developing useful and usable convection-allowing model ensemble information for NWS forecasters. *Wea. Forecasting*, **35**, 1381–1406, https://doi.org/10.1175/WAF-D-19-0108.1.

Dennis, E. J., and M. R. Kumjian, 2017: The impact of vertical wind shear on hail growth in simulated supercells. *J. Atmos. Sci.*, **74**, 641–663, https://doi.org/10.1175/JAS-D-16-0066.1.

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

Diffenbaugh, N. S., M. Scherer, and R. J. Trapp, 2013: Robust increases in severe thunderstorm environments in response to greenhouse forcing. *Proc. Natl. Acad. Sci. USA*, **110**, 16 361–16 366, https://doi.org/10.1073/pnas.1307758110.

Doswell, C. A., III, A. R. Moller, and H. E. Brooks, 1999: Storm spotting and public awareness since the first tornado forecasts of 1948. *Wea. Forecasting*, **14**, 544–557, https://doi.org/10.1175/1520-0434(1999)014<0544:SSAPAS>2.0.CO;2.

——, H. E. Brooks, and M. P. Kay, 2005: Climatological estimates of daily local nontornadic severe thunderstorm probability for the United States. *Wea. Forecasting*, **20**, 577–595, https://doi.org/10.1175/WAF866.1.

Ebert-Uphoff, I., R. Lagerquist, K. Hilburn, Y. Lee, K. Haynes, J. Stock, C. Kumler, and J. Q. Stewart, 2021: CIRA guide to custom loss functions for neural networks in environmental sciences—Version 1. arXiv, 2106.09757v1, https://doi.org/10.48550/arXiv.2106.09757.

Edwards, R., J. T. Allen, and G. W. Carbin, 2018: Reliability and climatological impacts of convective wind estimations. *J. Appl. Meteor. Climatol.*, **57**, 1825–1845, https://doi.org/10.1175/JAMC-D-17-0306.1.

Elmore, K. L., and H. Grams, 2016: Using mPING data to generate random forests for precipitation type forecasts. *14th Conf. on Artificial and Computational Intelligence and its Applications to the Environmental Sciences*, New Orleans, LA, Amer. Meteor. Soc., 4.2, https://ams.confex.com/ams/96Annual/webprogram/Paper289684.html.

——, Z. L. Flamig, V. Lakshmanan, B. T. Kaney, V. Farmer, H. D. Reeves, and L. P. Rothfusz, 2014: MPING: Crowd-sourcing weather reports for research. *Bull. Amer. Meteor. Soc.*, **95**, 1335–1342, https://doi.org/10.1175/BAMS-D-13-00014.1.

Fata, A. L., F. Amato, M. Bernardi, M. D'Andrea, R. Procopio, and E. Fiori, 2022: Horizontal grid spacing comparison among random forest algorithms to nowcast cloud-to-ground lightning occurrence. *Stochastic Environ. Res. Risk Assess.*, **36**, 2195–2206, https://doi.org/10.1007/s00477-022-02222-1.

Flora, M. L., P. S. Skinner, C. K. Potvin, A. E. Reinhart, T. A. Jones, N. Yussouf, and K. H. Knopfmeier, 2019: Object-based verification of short-term, storm-scale probabilistic mesocyclone guidance from an experimental Warn-on-Forecast system. *Wea. Forecasting*, **34**, 1721–1739, https://doi.org/10.1175/WAF-D-19-0094.1.

——, C. K. Potvin, P. S. Skinner, S. Handler, and A. McGovern, 2021: Using machine learning to generate storm-scale probabilistic guidance of severe weather hazards in the Warn-on-Forecast system. *Mon. Wea. Rev.*, **149**, 1535–1557, https://doi.org/10.1175/MWR-D-20-0194.1.

French, M. M., and D. M. Kingfield, 2021: Tornado formation and intensity prediction using polarimetric radar estimates of updraft area. *Wea. Forecasting*, **36**, 2211–2231, https://doi.org/10.1175/WAF-D-21-0087.1.

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

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

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

——, 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.

Geng, Y.-a., and Coauthors, 2019: LightNet: A dual spatiotemporal encoder network model for lightning prediction. *KDD'19: Proc. 25th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Anchorage, AK, Association for Computing Machinery, 2439–2447, https://doi.org/10.1145/3292500.3330717.

——, and Coauthors, 2021: A deep learning framework for lightning forecasting with multi-source spatiotemporal data. *Quart. J. Roy. Meteor. Soc.*, **147**, 4048–4062, https://doi.org/10.1002/qj.4167.

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

——, C. Ramseyer, and T. L. Mote, 2014: Future convective environments using NARCCAP. *Int. J. Climatol.*, **34**, 1699–1705, https://doi.org/10.1002/joc.3769.

——, C. Converse, W. S. Ashley, and M. Taszarek, 2021: Machine learning classification of significant tornadoes and hail in the United States using ERA5 proximity soundings. *Wea. Forecasting*, **36**, 2143–2160, https://doi.org/10.1175/WAF-D-21-0056.1.

Gibbs, J. G., 2016: A skill assessment of techniques for real-time diagnosis and short-term prediction of tornado intensity using the WSR-88D. *J. Oper. Meteor.*, **4**, 170–181, https://doi.org/10.15191/nwajom.2016.0413.

Gil, Y., and Coauthors, 2019: Intelligent systems for geosciences: An essential research agenda. *Commun. ACM*, **62**, 76–84, https://doi.org/10.1145/3192335.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, https://doi.org/10.1175/2009WAF2222269.1.

Gomes, C., and Coauthors, 2019: Computational sustainability: Computing for a better world and a sustainable future. *Commun. ACM*, **62**, 55–65, https://doi.org/10.1145/3339399.

Goodfellow, I., Y. Bengio, and A. Courville, 2016: *Deep Learning*. MIT Press, 800 pp.

Grams, J. S., R. L. Thompson, D. V. Snively, J. A. Prentice, G. M. Hodges, and L. J. Reames, 2012: A climatology and comparison of parameters for significant tornado events in the United States. *Wea. Forecasting*, **27**, 106–123, https://doi.org/10.1175/WAF-D-11-00008.1.

Grönquist, P., C. Yao, T. Ben-Nun, N. Dryden, P. Dueben, S. Li, and T. Hoefler, 2021: Deep learning for post-processing ensemble weather forecasts. *Philos. Trans. Roy. Soc.*, **A379**, 20200092, https://doi.org/10.1098/rsta.2020.0092.

Han, D., J. Lee, J. Im, S. Sim, S. Lee, and H. Han, 2019: A novel framework of detecting convective initiation combining automated sampling, machine learning, and repeated model tuning from geostationary satellite data. *Remote Sens.*, **11**, 1454, https://doi.org/10.3390/rs11121454.

Han, L., J. Sun, W. Zhang, Y. Xiu, H. Feng, and Y. Lin, 2017: A machine learning nowcasting method based on real-time reanalysis data. *J. Geophys. Res. Atmos.*, **122**, 4038–4051, https://doi.org/10.1002/2016JD025783.

Harrison, D., 2022: Machine learning co-production in operational meteorology. Ph.D. dissertation, School of Meteorology, University of Oklahoma, 196 pp., https://shareok.org/handle/11244/335971.

Haupt, S. E., A. Pasini, and C. Marzban, 2008: *Artificial Intelligence Methods in the Environmental Sciences*. 1st ed. Springer, 424 pp.

Haynes, K., R. Lagerquist, M. McGraw, K. Musgrave, and I. Ebert-Uphoff, 2023: Creating and evaluating uncertainty estimates with neural networks for environmental-science applications. *Artif. Intell. Earth Syst.*, **2**, e220061, https://doi.org/10.1175/AIES-D-22-0061.1.

Hempel, B., 2022: Nadocast—CONUS severe weather probabilities via feature engineering and gradient boosted decision trees. GitHub, https://github.com/brianhempel/nadocast.

Heng, B. C. P., and Coauthors, 2020: SINGV-DA: A data assimilation system for convective-scale numerical weather prediction over Singapore. *Quart. J. Roy. Meteor. Soc.*, **146**, 1923–1938, https://doi.org/10.1002/qj.3774.

Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting severe weather with random forests. *Mon. Wea. Rev.*, **148**, 2135–2161, https://doi.org/10.1175/MWR-D-19-0344.1.

Hoffman, R. R., 2017: A taxonomy of emergent trusting in the human-machine relationship. *Cognitive Systems Engineering: The Future for a Changing World*, P. J. Smith and R. R. Hoffman, Eds., CRC Press, 137–164.

——, S. V. Deal, S. Potter, and E. M. Roth, 2010: The practitioner's cycles, Part 2: Solving envisioned world problems. *IEEE Intell. Syst.*, **25**, 6–11, https://doi.org/10.1109/MIS.2010.89.

——, D. S. LaDue, H. M. Mogil, P. J. Roebber, and J. G. Trafton, 2017: *Minding the Weather: How Expert Forecasters Think*. MIT Press, 488 pp.

Hoogewind, K. A., M. E. Baldwin, and R. J. Trapp, 2017: The impact of climate change on hazardous convective weather in the United States: Insight from high-resolution dynamical downscaling. *J. Climate*, **30**, 10 081–10 100, https://doi.org/10.1175/JCLI-D-16-0885.1.

Insurance Information Institute, 2019: Facts + statistics: Tornadoes and thunderstorms. Insurance Information Institute, https://www.iii.org/fact-statistic/facts-statistics-tornadoes-and-thunderstorms.

Jones, T. A., K. Knopfmeier, D. Wheatley, G. Creager, P. Minnis, and R. Palikonda, 2016: Storm-scale data assimilation and ensemble forecasting with the NSSL Experimental Warn-on-Forecast system. Part II: Combined radar and satellite data experiments. *Wea. Forecasting*, **31**, 297–327, https://doi.org/10.1175/WAF-D-15-0107.1.

Karpatne, A., I. Ebert-Uphoff, S. Ravela, H. A. Babaie, and V. Kumar, 2019: Machine learning for the geosciences: Challenges and opportunities. *IEEE Trans. Knowl. Data Eng.*, **31**, 1544–1554, https://doi.org/10.1109/TKDE.2018.2861006.

Karras, T., T. Aila, S. Laine, and J. Lehtinen, 2018: Progressive growing of GANs for improved quality, stability, and variation. arXiv, 1710.10196v3, https://doi.org/10.48550/arXiv.1710.10196.

Karstens, C. D., and Coauthors, 2018: Development of a human-machine mix for forecasting severe convective events. *Wea. Forecasting*, **33**, 715–737, https://doi.org/10.1175/WAF-D-17-0188.1.

Kennedy, P. C., S. A. Rutledge, B. Dolan, and E. Thaler, 2014: Observations of the 14 July 2011 Fort Collins hailstorm: Implications for WSR-88D-based hail detection and warnings. *Wea. Forecasting*, **29**, 623–638, https://doi.org/10.1175/WAF-D-13-00075.1.

Kingfield, D. M., and J. G. LaDue, 2015: The relationship between automated low-level velocity calculations from the WSR-88D and maximum tornado intensity determined from damage surveys. *Wea. Forecasting*, **30**, 1125–1139, https://doi.org/10.1175/WAF-D-14-00096.1.

Kitzmiller, D. H., W. E. McGovern, and R. F. Saffle, 1995: The WSR-88D severe weather potential algorithm. *Wea. Forecasting*, **10**, 141–159, https://doi.org/10.1175/1520-0434(1995)010<0141:TWSWPA>2.0.CO;2.

Kochkov, D., J. A. Smith, A. Alieva, Q. Wang, M. P. Brenner, and S. Hoyer, 2021: Machine learning accelerated computational fluid dynamics. *Proc. Natl. Acad. Sci.*, **118**, e2101784118, https://doi.org/10.1073/pnas.2101784118.

Krizhevsky, A., I. Sutskever, and G. E. Hinton, 2012: Imagenet classification with deep convolutional neural networks. *Commun. ACM*, **62**, 84–90, https://doi.org/10.1145/3065386.

Kurth, T., and Coauthors, 2018: Exascale deep learning for climate analytics. *Proc. Int. Conf. for High Performance Computing, Networking, Storage, and Analysis*, Dallas, TX, IEEE, 1–12, https://dl.acm.org/doi/10.5555/3291656.3291724.

Kuster, C. M., J. C. Snyder, T. J. Schuur, T. T. Lindley, P. L. Heinselman, J. C. Furtado, J. W. Brogden, and R. Toomey, 2019: Rapid-update radar observations of Z_{DR} column depth and its use in the warning decision process. *Wea. Forecasting*, **34**, 1173–1188, https://doi.org/10.1175/WAF-D-19-0024.1.

Labriola, J., N. Snook, Y. Jung, B. Putnam, and M. Xue, 2017: Ensemble hail prediction for the storms of 10 May 2010 in south-central Oklahoma using single- and double-moment microphysical schemes. *Mon. Wea. Rev.*, **145**, 4911–4936, https://doi.org/10.1175/MWR-D-17-0039.1.

Lagerquist, R., and I. Ebert-Uphoff, 2022: Can we integrate spatial verification methods into neural-network loss functions for atmospheric science? *Artif. Intell. Earth Syst.*, **1**, e220021, https://doi.org/10.1175/AIES-D-22-0021.1.

——, A. McGovern, and T. Smith, 2017: Machine learning for real-time prediction of damaging straight-line convective wind. *Wea. Forecasting*, **32**, 2175–2193, https://doi.org/10.1175/WAF-D-17-0038.1.

——, ——, and D. J. Gagne II, 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.

——, J. Q. Stewart, I. Ebert-Uphoff, and C. Kumler, 2021a: Using deep learning to nowcast the spatial coverage of convection from *Himawari-8* satellite data. *Mon. Wea. Rev.*, **149**, 3897–3921, https://doi.org/10.1175/MWR-D-21-0096.1.

——, D. Turner, I. Ebert-Uphoff, J. Stewart, and V. Hagerty, 2021b: Using deep learning to emulate and accelerate a radiative transfer model. *J. Atmos. Oceanic Technol.*, **38**, 1673–1696, https://doi.org/10.1175/JTECH-D-21-0007.1.

Lakshmanan, V., I. Adrianto, T. Smith, and G. Stumpf, 2005: A spatiotemporal approach to tornado prediction. *Proc. 2005 IEEE Int. Joint Conf. on Neural Networks*, Montreal, QC, Canada, IEEE, 1642–1647, https://doi.org/10.1109/IJCNN.2005.1556125.

——, K. L. Ortega, and T. M. Smith, 2007: Creating spatio-temporal tornado probability forecasts using fuzzy logic and motion variability. *Fifth Conf. on Artificial Intelligence Applications to Environmental Science*, San Antonio, TX, Amer. Meteor. Soc., 2.3, https://ams.confex.com/ams/87ANNUAL/techprogram/paper_119456.htm.

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner, 1998: Gradient-based learning applied to document recognition. *Proc. IEEE*, **86**, 2278–2324, https://doi.org/10.1109/5.726791.

——, Y. Bengio, and G. Hinton, 2015: Deep learning. *Nature*, **521**, 436–444, https://doi.org/10.1038/nature14539.

Lee, S., H. Han, J. Im, E. Jang, and M.-I. Lee, 2017: Detection of deterministic and probabilistic convection initiation using Himawari-8 advanced Himawari Imager data. *Atmos. Meas. Tech.*, **10**, 1859–1874, https://doi.org/10.5194/amt-10-1859-2017.

Lee, Y., C. D. Kummerow, and I. Ebert-Uphoff, 2021: Applying machine learning methods to detect convection using *Geostationary Operational Environmental Satellite-16* (*GOES-16*) advanced baseline imager (ABI) data. *Atmos. Meas. Tech.*, **14**, 2699–2716, https://doi.org/10.5194/amt-14-2699-2021.

Leinonen, J., U. Hamann, and U. Germann, 2022a: Seamless lightning nowcasting with recurrent-convolutional deep learning. arXiv, 2203.10114v3, https://doi.org/10.48550/arXiv.2203.10114.

——, ——, ——, and J. R. Mecikalski, 2022b: Nowcasting thunderstorm hazards using machine learning: The impact of data sources on performance. *Nat. Hazards Earth Syst. Sci.*, **22**, 577–597, https://doi.org/10.5194/nhess-22-577-2022.

Lin, T., and Coauthors, 2019: Attention-based dual-source spatiotemporal neural network for lightning forecast. *IEEE Access*, **7**, 158 296–158 307, https://doi.org/10.1109/ACCESS.2019.2950328.

Loeffler, S. D., and M. R. Kumjian, 2018: Quantifying the separation of enhanced Z_{dr} and K_{dp} regions in nonsupercell tornadic storms. *Wea. Forecasting*, **33**, 1143–1157, https://doi.org/10.1175/WAF-D-18-0011.1.

Loken, E. D., A. J. Clark, and C. D. Karstens, 2020: Generating probabilistic next-day severe weather forecasts from convection-allowing ensembles using random forests. *Wea. Forecasting*, **35**, 1605–1631, https://doi.org/10.1175/WAF-D-19-0258.1.

——, ——, and A. McGovern, 2022a: Comparing and interpreting differently designed random forests for next-day severe weather hazard prediction. *Wea. Forecasting*, **37**, 871–899, https://doi.org/10.1175/WAF-D-21-0138.1.

——, K. A. Wilson, T. Sandmael, A. J. Clark, K. M. Calhoun, A. E. Reinhart, P. Skinner, and P. C. Burke, 2022b: Improving probabilistic watch-to-warning severe hazard guidance by merging the Warn-on-Forecast system with observations-based products using machine learning. *30th Conf. on Severe Local Storms*, Santa Fe, NM, Amer. Meteor. Soc., 42, https://ams.confex.com/ams/30SLS/meetingapp.cgi/Paper/407634.

López, L., E. García-Ortega, and J. L. Sánchez, 2007: A short-term forecast model for hail. *Atmos. Res.*, **83**, 176–184, https://doi.org/10.1016/j.atmosres.2005.10.014.

Malde, K., N. O. Handegard, L. Eikvil, and A.-B. Salberg, 2020: Machine intelligence and the data-driven future of marine science. *ICES J. Mar. Sci.*, **77**, 1274–1285, https://doi.org/10.1093/icesjms/fsz057.

Malone, T. F., 1955: Application of statistical methods in weather prediction. *Proc. Natl. Acad. Sci. USA*, **41**, 806–815, https://doi.org/10.1073/pnas.41.11.806.

Manzato, A., 2013: Hail in northeast Italy: A neural network ensemble forecast using sounding-derived indices. *Wea. Forecasting*, **28**, 3–28, https://doi.org/10.1175/WAF-D-12-00034.1.

Marion, G. R., R. J. Trapp, and S. W. Nesbitt, 2019: Using overshooting top area to discriminate potential for large, intense tornadoes. *Geophys. Res. Lett.*, **46**, 12 520–12 526, https://doi.org/10.1029/2019GL084099.

Marsh, P. T., H. E. Brooks, and D. J. Karoly, 2009: Preliminary investigation into the severe thunderstorm environment of Europe simulated by the Community Climate System Model 3. *Atmos. Res.*, **93**, 607–618, https://doi.org/10.1016/j.atmosres.2008.09.014.

Marzban, C., 2000: A neural network for tornado diagnosis: Managing local minima. *Neural Comput. Appl.*, **9**, 133–141, https://doi.org/10.1007/s005210070024.

——, and G. J. Stumpf, 1996: A neural network for tornado prediction based on Doppler radar-derived attributes. *J. Appl. Meteor.*, **35**, 617–626, https://www.jstor.org/stable/26187858.

——, and ——, 1998: A neural network for damaging wind prediction. *Wea. Forecasting*, **13**, 151–163, https://doi.org/10.1175/1520-0434(1998)013%3C0151:ANNFDW%3E2.0.CO;2.

——, and A. Witt, 2001: A Bayesian neural network for severe-hail size prediction. *Wea. Forecasting*, **16**, 600–610, https://doi.org/10.1175/1520-0434(2001)016<0600:ABNNFS>2.0.CO;2.

McCloskey, S., E. D. Loken, C. Karstens, B. T. Smith, and D. E. Jahn, 2022: Examining when and how random forests add value to next-day Storm Prediction Center hail forecasts. *31st Conf. on Weather Analysis and Forecasting/27th Conf. on Numerical Weather Prediction*, Houston, TX, Amer. Meteor. Soc., J8.3, https://ams.confex.com/ams/102ANNUAL/meetingapp.cgi/Paper/397442.

McGovern, A., D. J. Gagne II, J. K. Williams, R. A. Brown, and J. B. Basara, 2014: Enhancing understanding and improving prediction of severe weather through spatiotemporal relational learning. *Mach. Learn.*, **95**, 27–50, https://doi.org/10.1007/s10994-013-5343-x.

——, ——, J. Basara, T. M. Hamill, and D. Margolin, 2015: Solar energy prediction: An international contest to initiate interdisciplinary research on compelling meteorological problems. *Bull. Amer. Meteor. Soc.*, **96**, 1388–1395, https://doi.org/10.1175/BAMS-D-14-00006.1.

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

——, C. Potvin, and R. A. Brown, 2017b: Using large-scale machine learning to improve our understanding of the formation of tornadoes. *Large-Scale Machine Learning in the Earth Sciences*, A. Srivastava et al., Eds., Chapman and Hall, 95–112.

——, R. Lagerquist, D. J. Gagne II, 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.

——, I. Ebert-Uphoff, D. J. Gagne II, and A. Bostrom, 2022: Why we need to focus on developing ethical, responsible, and trustworthy artificial intelligence approaches for environmental science. *Environ. Data Sci.*, **1**, e6, https://doi.org/10.1017/eds.2022.5.

McGuire, M. P., and T. W. Moore, 2022: Prediction of tornado days in the United States with deep convolutional neural networks. *Comput. Geosci.*, **159**, 104990, https://doi.org/10.1016/j.cageo.2021.104990.

Mecikalski, J. R., and K. M. Bedka, 2006: Forecasting convective initiation by monitoring the evolution of moving cumulus in daytime GOES imagery. *Mon. Wea. Rev.*, **134**, 49–78, https://doi.org/10.1175/MWR3062.1.

——, J. K. Williams, C. P. Jewett, D. Ahijevych, A. LeRoy, and J. R. Walker, 2015: Probabilistic 0–1-h convective initiation nowcasts that combine geostationary satellite observations and numerical weather prediction model data. *J. Appl. Meteor. Climatol.*, **54**, 1039–1059, https://doi.org/10.1175/JAMC-D-14-0129.1.

Mercer, A. E., C. M. Shafer, C. A. Doswell III, L. M. Leslie, and M. B. Richman, 2009: Objective classification of tornadic and nontornadic severe weather outbreaks. *Mon. Wea. Rev.*, **137**, 4355–4368, https://doi.org/10.1175/2009MWR2897.1.

Mesinger, F., and Coauthors, 2006: North American Regional Reanalysis. *Bull. Amer. Meteor. Soc.*, **87**, 343–360, https://doi.org/10.1175/BAMS-87-3-343.

Meyer, T. C., K. M. Calhoun, D. M. Kingfield, and C. Karstens, 2017: Using random forest to generate cloud-to-ground lightning probabilities. *38th Conf. on Radar Meteorology*, Chicago, IL, Amer. Meteor. Soc., 5B.6, https://ams.confex.com/ams/38RADAR/meetingapp.cgi/Paper/320597.

Miller, D. E., Z. Wang, R. J. Trapp, and D. S. Harnos, 2020: Hybrid prediction of weekly tornado activity out to week 3: Utilizing weather regimes. *Geophys. Res. Lett.*, **47**, e2020GL087253, https://doi.org/10.1029/2020GL087253.

Miller, T., 2019: Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, **267**, 1–38, https://doi.org/10.1016/j.artint.2018.07.007.

Mitchell, E. D. W., S. V. Vasiloff, G. J. Stumpf, A. Witt, M. D. Eilts, J. T. Johnson, and K. W. Thomas, 1998: The national severe storms laboratory tornado detection algorithm. *Wea. Forecasting*,

13, 352–366, https://doi.org/10.1175/1520-0434(1998)013<0352:TNSSLT>2.0.CO;2.

Molina, M. J., D. J. Gagne, and A. F. Prein, 2021: A benchmark to test generalization capabilities of deep learning methods to classify severe convective storms in a changing climate. *Earth Space Sci.*, **8**, e2020EA001490, https://doi.org/10.1029/2020EA001490.

Molnar, C., 2018: Interpretable machine learning: A guide for making black box models explainable. Leanpub, accessed, https://christophm.github.io/interpretable-ml-book/.

——, G. Casalicchio, and B. Bischl, 2020: Interpretable machine learning—A brief history, state-of-the-art and challenges. arXiv, 2010.09337v1, https://doi.org/10.48550/arXiv.2010.09337.

Monteleoni, C., G. A. Schmidt, and S. McQuade, 2013: Climate informatics: Accelerating discovering in climate science with machine learning. *Comput. Sci. Eng.*, **15**, 32–40, https://doi.org/10.1109/MCSE.2013.50.

Morrison, H., and Coauthors, 2020: Confronting the challenge of modeling cloud and precipitation microphysics. *J. Adv. Model. Earth Syst.*, **12**, e2019MS001689, https://doi.org/10.1029/2019MS001689.

Mostajabi, A., D. L. Finney, M. Rubinstein, and F. Rachidi, 2019: Nowcasting lightning occurrence from commonly available meteorological parameters using machine learning techniques. *npj Climate Atmos. Sci.*, **2**, 41, https://doi.org/10.1038/s41612-019-0098-0.

Mounier, A., L. Raynaud, L. Rottner, M. Plu, P. Arbogast, M. Kreitz, L. Mignan, and B. Touzé, 2022: Detection of bow echoes in kilometer-scale forecasts using a convolutional neural network. *Artif. Intell. Earth Syst.*, **1**, e210010, https://doi.org/10.1175/AIES-D-21-0010.1.

Mueller, C., and J. Wilson, 1989: Evaluation of the TDWR aviation nowcasting experiment. *Conf. on Radar Meteorology*, Tallahassee, FL, Amer. Meteor. Soc, 224–227.

——, ——, and N. A. Crook, 1993: The utility of sounding and mesonet data to nowcast thunderstorm initiation. *Wea. Forecasting*, **8**, 132–146, https://doi.org/10.1175/1520-0434(1993)008<0132:TUOSAM>2.0.CO;2.

Mueller, S. T., R. R. Hoffman, W. Clancey, A. Emrey, and G. Klein, 2019: Explanation in human-AI systems: A literature meta-review synopsis of key ideas and publication and bibliography for explainable AI. arXiv, 1902.01876v1, https://doi.org/10.48550/arXiv.1902.01876.

Murillo, E. M., and C. R. Homeyer, 2019: Severe hail fall and hailstorm detection using remote sensing observations. *J. Appl. Meteor. Climatol.*, **58**, 947–970, https://doi.org/10.1175/JAMC-D-18-0247.1.

NOAA National Centers for Environmental Information, 2022: Billion-dollar weather and climate disasters: Table of events. NOAA NCEI, accessed 22 October 2022, https://www.ncdc.noaa.gov/billions/.

Nowotarski, C. J., and A. A. Jensen, 2013: Classifying proximity soundings with self-organizing maps toward improving supercell and tornado forecasting. *Wea. Forecasting*, **28**, 783–801, https://doi.org/10.1175/WAF-D-12-00125.1.

NSSL/NWS SPC, 2022: Experimental outlook verification. NOAA, accessed 22 October 2022, https://hwt.nssl.noaa.gov/sfe_viewer/2022/outlook_verification/.

Orf, L., R. Wilhelmson, B. Lee, C. Finley, and A. Houston, 2017: Evolution of a long-track violent tornado within a simulated supercell. *Bull. Amer. Meteor. Soc.*, **98**, 45–68, https://doi.org/10.1175/BAMS-D-15-00073.1.

Pal, S., E. Tirone, S. Dutta, W. A. Gallus, R. Maitra, J. Newman, and E. Weber, 2022: "Blowin' in the wind"—Diagnosing the probability that a severe thunderstorm wind report is truly due to severe intensity wind event. *21st Conf. on Artificial Intelligence for Environmental Science*, Houston, TX, Amer. Meteor. Soc., 5A.1, https://ams.confex.com/ams/102ANNUAL/meetingapp.cgi/Paper/395627.

Potvin, C. K., and M. L. Flora, 2015: Sensitivity of idealized supercell simulations to horizontal grid spacing: Implications for Warn-on-Forecast. *Mon. Wea. Rev.*, **143**, 2998–3024, https://doi.org/10.1175/MWR-D-14-00416.1.

——, C. Broyles, P. S. Skinner, H. E. Brooks, and E. Rasmussen, 2019: A Bayesian hierarchical modeling framework for correcting reporting bias in the U.S. tornado database. *Wea. Forecasting*, **34**, 15–30, https://doi.org/10.1175/WAF-D-18-0137.1.

——, ——, ——, and ——, 2022: Improving estimates of U.S. tornado frequency by accounting for unreported and underrated tornadoes. *J. Appl. Meteor. Climatol.*, **61**, 909–930, https://doi.org/10.1175/JAMC-D-21-0225.1.

Púčik, T., C. Castellano, P. Groenemeijer, T. Kühne, A. T. Rädler, B. Antonescu, and E. Faust, 2019: Large hail incidence and its economic and societal impacts across Europe. *Mon. Wea. Rev.*, **147**, 3901–3916, https://doi.org/10.1175/MWR-D-19-0204.1.

Pullman, M., I. Gurung, M. Maskey, R. Ramachandran, and S. A. Christopher, 2019: Applying deep learning to hail detection: A case study. *IEEE Trans. Geosci. Remote Sens.*, **57**, 10 218–10 225, https://doi.org/10.1109/TGRS.2019.2931944.

Pulukool, F., L. Li, and C. Liu, 2020: Using deep learning and machine learning methods to diagnose hailstorms in large-scale thermodynamic environments. *Sustainability*, **12**, 10499, https://doi.org/10.3390/su122410499.

Putnam, B. J., Y. Jung, N. Yussouf, D. Stratman, T. A. Supinie, M. Xue, C. Kuster, and J. Labriola, 2021: The impact of assimilating Z_{DR} observations on storm-scale ensemble forecasts of the 31 May 2013 Oklahoma storm event. *Mon. Wea. Rev.*, **149**, 1919–1942, https://doi.org/10.1175/MWR-D-20-0261.1.

Racah, E., C. Beckham, T. Maharaj, S. E. Kahou, Prabhat, and C. Pal, 2017: Extreme weather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. *Proc. 31st Int. Conf. on Neural Information Processing Systems*, Long Beach, CA, Association for Computing Machinery, 3405–3416, https://dl.acm.org/doi/10.5555/3294996.3295099.

Rakhlin, A., A. Shvets, V. Iglovikov, and A. Kalinin, 2018: Deep convolutional neural networks for breast cancer histology image analysis. arXiv, 1802.00752v2, https://doi.org/10.48550/arXiv.1802.00752.

Ras, G., M. van Gerven, and P. Haselager, 2018: Explanation methods in deep learning: Users, values, concerns and challenges. *Explainable and Interpretable Models in Computer Vision and Machine Learning*, Springer, 19–36.

Rasp, S., and S. Lerch, 2018: Neural networks for postprocessing ensemble weather forecasts. *Mon. Wea. Rev.*, **146**, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1.

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, 2019: Deep learning and process understanding for data-driven Earth system science. *Nature*, **566**, 195–204, https://doi.org/10.1038/s41586-019-0912-1.

Ringhausen, J., P. Bitzer, W. Koshak, and J. Mecikalski, 2021: Classification of GLM flashes using random forests. *Earth Space Sci.*, **8**, e2021EA001861, https://doi.org/10.1029/2021EA001861.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

Roberts, R. D., and S. Rutledge, 2003: Nowcasting storm initiation and growth using GOES-8 and WSR-88D data. *Wea. Forecasting*, **18**, 562–584, https://doi.org/10.1175/1520-0434(2003)018<0562:NSIAGU>2.0.CO;2.

Ronneberger, O., P. Fischer, and T. Brox, 2015: U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, N. Navab et al., Eds., Lecture Notes in Computer Science, Vol. 9351, Springer, 234–241.

Rudin, C., 2018: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. arXiv, 1811.10154v3, https://doi.org/10.48550/arXiv.1811.10154.

Salcedo-Sanz, S., and Coauthors, 2022: Analysis, characterization, prediction and attribution of extreme atmospheric events with machine learning: A review. arXiv, 2207.07580v1, https://doi.org/10.48550/arXiv.2207.07580.

Samek, W., G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Muller, Eds., 2019: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Lecture Notes in Computer Science, Vol. 11700, Springer, 439 pp.

Sandmæl, T. N., and Coauthors, 2023: The tornado probability algorithm: A probabilistic machine learning tornadic circulation detection algorithm. *Wea. Forecasting*, **38**, 445–466, https://doi.org/10.1175/WAF-D-22-0123.1.

Scheuerer, M., M. B. Switanek, R. P. Worsnop, and T. M. Hamill, 2020: Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. *Mon. Wea. Rev.*, **148**, 3489–3506, https://doi.org/10.1175/MWR-D-20-0096.1.

Schmidt, T., and Coauthors, 2023: 1-2 hour hail nowcasting using time-resolving 3-dimensional UNets. *22nd Conf. on Artificial Intelligence for Environmental Science*, Denver, CO, Amer. Meteor. Soc., 5A.2, https://ams.confex.com/ams/103ANNUAL/meetingapp.cgi/Paper/418273.

Schulz, B., and S. Lerch, 2022: Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Mon. Wea. Rev.*, **150**, 235–257, https://doi.org/10.1175/MWR-D-21-0150.1.

Sessa, M. F., and R. J. Trapp, 2020: Observed relationship between tornado intensity and pretornadic mesocyclone characteristics. *Wea. Forecasting*, **35**, 1243–1261, https://doi.org/10.1175/WAF-D-19-0099.1.

Shafer, C. M., A. E. Mercer, L. M. Leslie, M. B. Richman, and C. A. Doswell, 2010: Evaluation of WRF model simulations of tornadic and nontornadic outbreaks occurring in the spring and fall. *Mon. Wea. Rev.*, **138**, 4098–4119, https://doi.org/10.1175/2010MWR3269.1.

——, ——, M. B. Richman, L. M. Leslie, and C. A. Doswell III, 2012: An assessment of areal coverage of severe weather parameters for severe weather outbreak diagnosis. *Wea. Forecasting*, **27**, 809–831, https://doi.org/10.1175/WAF-D-11-00142.1.

Shen, B., X. Liang, Y. Ouyang, M. Liu, W. Zheng, and K. Carley, 2018: Stepdeep: A novel spatial-temporal mobility event prediction framework based on deep neural network. *Proc. 24th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, London, United Kingdom, Association for Computing Machinery, 724–733, https://dl.acm.org/doi/10.1145/3219819.3219931.

Shi, X., Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, 2015: Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Proc. 28th Int. Conf. on Neural Information Processing Systems*, Montreal, QC, Canada, Association for Computing Machinery, 802–810, https://dl.acm.org/doi/10.5555/2969239.2969329.

Shrestha, Y., Y. Zhang, R. Doviak, and P. W. Chan, 2021: Lightning flash rate nowcasting based on polarimetric radar data and machine learning. *Int. J. Remote Sens.*, **42**, 6762–6780, https://doi.org/10.1080/01431161.2021.1933243.

Sieglaff, J. M., L. M. Cronce, W. F. Feltz, K. M. Bedka, M. J. Pavolonis, and A. K. Heidinger, 2011: Nowcasting convective storm initiation using satellite-based box-averaged cloud-top cooling and cloud-type trends. *J. Appl. Meteor. Climatol.*, **50**, 110–126, https://doi.org/10.1175/2010JAMC2496.1.

Silver, D., and Coauthors, 2017: Mastering the game of Go without human knowledge. *Nature*, **550**, 354–359, https://doi.org/10.1038/nature24270.

Skinner, P. S., and Coauthors, 2018: Object-based verification of a prototype Warn-on-Forecast system. *Wea. Forecasting*, **33**, 1225–1250, https://doi.org/10.1175/WAF-D-18-0020.1.

Smith, B. T., R. L. Thompson, A. R. Dean, and P. T. Marsh, 2015: Diagnosing the conditional probability of tornado damage rating using environmental and radar attributes. *Wea. Forecasting*, **30**, 914–932, https://doi.org/10.1175/WAF-D-14-00122.1.

Snook, N., Y. Jung, J. Brotzge, B. Putnam, and M. Xue, 2016: Prediction and ensemble forecast verification of hail in the supercell storms of 20 May 2013. *Wea. Forecasting*, **31**, 811–825, https://doi.org/10.1175/WAF-D-15-0152.1.

——, M. Xue, and Y. Jung, 2019: Tornado-resolving ensemble and probabilistic predictions of the 20 May 2013 Newcastle–Moore EF5 tornado. *Mon. Wea. Rev.*, **147**, 1215–1235, https://doi.org/10.1175/MWR-D-18-0236.1.

Sobash, R. A., G. S. Romine, and C. S. Schwartz, 2020: A comparison of neural-network and surrogate-severe probabilistic convective hazard guidance derived from a convection-allowing model. *Wea. Forecasting*, **35**, 1981–2000, https://doi.org/10.1175/WAF-D-20-0036.1.

Spychalla, L. K., J. K. Robinson, R. Chase, A. McGovern, J. T. Allen, J. K. Williams, and N. Snook, 2022: Next-hour hail prediction from numerical weather prediction models using U-nets. *21st Conf. on Artificial Intelligence for Environmental Science*, Houston, TX, Amer. Meteor. Soc., 15.1, https://ams.confex.com/ams/102ANNUAL/meetingapp.cgi/Paper/393566.

Starzec, M., C. R. Homeyer, and G. L. Mullendore, 2017: Storm labeling in three dimensions (SL3D): A volumetric radar echo and dual-polarization updraft classification algorithm. *Mon. Wea. Rev.*, **145**, 1127–1145, https://doi.org/10.1175/MWR-D-16-0089.1.

Steinkruger, D., P. Markowski, and G. Young, 2020: An artificially intelligent system for the automated issuance of tornado warnings in simulated convective storms. *Wea. Forecasting*, **35**, 1939–1965, https://doi.org/10.1175/WAF-D-19-0249.1.

Stensrud, D. J., and Coauthors, 2009: Convective-scale Warn-on-Forecast system. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1500, https://doi.org/10.1175/2009BAMS2795.1.

——, and Coauthors, 2013: Progress and challenges with Warn-on-Forecast. *Atmos. Res.*, **123**, 2–16, https://doi.org/10.1016/j.atmosres.2012.04.004.

Stuart, N. A., and Coauthors, 2022: The evolving role of humans in weather prediction and communication. *Bull. Amer. Meteor. Soc.*, **103**, E1720–E1746, https://doi.org/10.1175/BAMS-D-20-0326.1.

Su, A., H. Li, L. Cui, and Y. Chen, 2020: A convection nowcasting method based on machine learning. *Adv. Meteor.*, **2020**, 5124274, https://doi.org/10.1155/2020/5124274.

Sun, F., D. Qin, M. Min, B. Li, and F. Wang, 2019: Convective initiation nowcasting over China from Fengyun-4A measurements based on TV-L_1 optical flow and BP_Adaboost neural network algorithms. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **12**, 4284–4296, https://doi.org/10.1109/JSTARS.2019.2952976.

Suwajanakorn, S., S. Seitz, and I. Kemelmacher-Shlizerman, 2017: Synthesizing Obama: Learning lip sync from audio. *ACM Trans. Graph.*, **36** (4), 1–13, https://doi.org/10.1145/3072959.3073640.

Thompson, R. L., R. Edwards, J. A. Hart, K. L. Elmore, and P. Markowski, 2003: Close proximity soundings within supercell environments obtained from the Rapid Update Cycle. *Wea. Forecasting*, **18**, 1243–1261, https://doi.org/10.1175/1520-0434(2003)018<1243:CPSWSE>2.0.CO;2.

——, and Coauthors, 2017: Tornado damage rating probabilities derived from WSR-88D data. *Wea. Forecasting*, **32**, 1509–1528, https://doi.org/10.1175/WAF-D-17-0004.1.

Trafalis, T. B., H. Ince, and M. B. Richman, 2003: Tornado detection with support vector machines. *Computational Science—ICCS 2003*, P. M. A. Sloot et al., Eds., Lecture Notes in Computer Science, Vol. 2660, Springer, 289–298.

——, I. Adrianto, M. B. Richman, and S. Lakshmivarahan, 2014: Machine-learning classifiers for imbalanced tornado data. *Comput. Manage. Sci.*, **11**, 403–418, https://doi.org/10.1007/s10287-013-0174-6.

Trapp, R. J., N. S. Diffenbaugh, and A. Gluhovsky, 2009: Transient response of severe thunderstorm forcing to elevated greenhouse gas concentrations. *Geophys. Res. Lett.*, **36**, L01703, https://doi.org/10.1029/2008GL036203.

——, K. A. Hoogewind, and S. Lasher-Trapp, 2019: Future changes in hail occurrence in the United States determined through convection-permitting dynamical downscaling. *J. Climate*, **32**, 5493–5509, https://doi.org/10.1175/JCLI-D-18-0740.1.

Van Den Broeke, M. S., and S. T. Jauernic, 2014: Spatial and temporal characteristics of polarimetric tornadic debris signatures. *J. Appl. Meteor. Climatol.*, **53**, 2217–2231, https://doi.org/10.1175/JAMC-D-14-0094.1.

Veillette, M. S., H. Iskenderian, P. M. Lamey, and L. J. Bickmeier, 2013: Convective initiation forecasts through the use of machine learning methods. *11th Conf. on Artificial and Computational Intelligence and its Applications to the Environmental Sciences*, Austin, TX, Amer. Meteor. Soc., TJ37.4, https://ams.confex.com/ams/93Annual/webprogram/Paper221560.html.

Vorndran, M., A. Schütz, J. Bendix, and B. Thies, 2022: Current training and validation weaknesses in classification-based radiation fog nowcast using machine learning algorithms. *Artif. Intell. Earth Syst.*, **1**, e210006, https://doi.org/10.1175/AIES-D-21-0006.1.

Walker, J. R., W. M. MacKenzie Jr., J. R. Mecikalski, and C. P. Jewett, 2012: An enhanced geostationary satellite–based convective initiation algorithm for 0–2-h nowcasting with object tracking. *J. Appl. Meteor. Climatol.*, **51**, 1931–1949, https://doi.org/10.1175/JAMC-D-11-0246.1.

Wang, L., K. A. Scott, L. Xu, and D. Clausi, 2016: Sea ice concentration estimation during melt from dual-pol SAR scenes using deep convolutional neural networks: A case study. *IEEE Trans. Geosci. Remote Sens.*, **54**, 4524–4533, https://doi.org/10.1109/TGRS.2016.2543660.

Wang, P., W. Lv, C. Wang, and J. Hou, 2018: Hail storms recognition based on convolutional neural network. *2018 13th World Congress on Intelligent Control and Automation (WCICA)*, Changsha, China, IEEE, 1703–1708, https://doi.org/10.1109/WCICA.2018.8630701.

Wang, Y., and T.-Y. Yu, 2015: Novel tornado detection using an adaptive neuro-fuzzy system with S-band polarimetric weather radar. *J. Atmos. Oceanic Technol.*, **32**, 195–208, https://doi.org/10.1175/JTECH-D-14-00096.1.

Wheatley, D. M., K. H. Knopfmeier, T. A. Jones, and G. J. Creager, 2015: Storm-scale data assimilation and ensemble forecasting with the NSSL Experimental Warn-on-Forecast system. Part I: Radar data experiments. *Wea. Forecasting*, **30**, 1795–1817, https://doi.org/10.1175/WAF-D-15-0043.1.

Williams, J. K., 2014: Using random forests to diagnose aviation turbulence. *Mach. Learn.*, **95**, 51–70, https://doi.org/10.1007/s10994-013-5346-7.

——, D. Ahijevych, S. Dettling, and M. Steiner, 2008a: Combining observations and model data for short-term storm forecasting. *Proc. SPIE*, **7088**, 708805, https://doi.org/10.1117/12.795737.

——, R. Sharman, J. Craig, and G. Blackburn, 2008b: Remote detection and diagnosis of thunderstorm turbulence. *Proc. SPIE*, **7088**, 708804, https://doi.org/10.1117/12.795570.

——, D. A. Ahijevych, C. J. Kessinger, T. R. Saxen, M. Steiner, and S. Dettling, 2008c: A machine-learning approach to finding weather regimes and skillful predictor combinations for short-term storm forecasting. *13th Conf. on Aviation, Range and Aerospace Meteorology*, New Orleans, LA, Amer. Meteor. Soc., J1. 4, https://ams.confex.com/ams/88Annual/techprogram/session_20816.htm.

Wilson, J. W., and C. K. Mueller, 1993: Nowcasts of thunderstorm initiation and evolution. *Wea. Forecasting*, **8**, 113–131, https://doi.org/10.1175/1520-0434(1993)008<0113:NOTIAE>2.0.CO;2.

Wimmers, A., C. Velden, and J. H. Cossuth, 2019: Using deep learning to estimate tropical cyclone intensity from satellite passive microwave imagery. *Mon. Wea. Rev.*, **147**, 2261–2282, https://doi.org/10.1175/MWR-D-18-0391.1.

Witt, A., M. D. Eilts, G. J. Stumpf, J. Johnson, E. D. W. Mitchell, and K. W. Thomas, 1998: An enhanced hail detection algorithm for the WSR-88D. *Wea. Forecasting*, **13**, 286–303, https://doi.org/10.1175/1520-0434(1998)013<0286:AEHDAF>2.0.CO;2.

Yao, H., X. Li, H. Pang, L. Sheng, and W. Wang, 2020: Application of random forest algorithm in hail forecasting over Shandong Peninsula. *Atmos. Res.*, **244**, 105093, https://doi.org/10.1016/j.atmosres.2020.105093.

Zhao, M., T. Li, M. A. Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, 2018: Through-wall human pose estimation using radio signals. *2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, IEEE, 7356–7365, https://doi.org/10.1109/CVPR.2018.00768.

Zhou, K., Y. Zheng, W. Dong, and T. Wang, 2020: A deep learning network for cloud-to-ground lightning nowcasting with multi-source data. *J. Atmos. Oceanic Technol.*, **37**, 927–942, https://doi.org/10.1175/JTECH-D-19-0146.1.

Zhou, X., Y.-A. Geng, H. Yu, Q. Li, L. Xu, W. Yao, D. Zheng, and Y. Zhang, 2022: LightNet+: A dual-source lightning forecasting network with bi-direction spatiotemporal transformation. *Appl. Intell.*, **52**, 11 147–11 159, https://doi.org/10.1007/s10489-021-03089-5.

Zhou, Z., M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, 2020: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging*, **39**, 1856–1867, https://doi.org/10.1109/TMI.2019.2959609.