Gridded Severe Hail Nowcasting Using 3D U-Nets, Lightning Observations, and the Warn-on-Forecast System

TOBIAS G. SCHMIDT,[a,b,c,d] AMY MCGOVERN,[a,c,e] JOHN T. ALLEN,[c,g] COREY K. POTVIN,[b,a,c] RANDY J. CHASE,[f,c] CHAD M. WILEY,[a,b,c,d] WILLIAM R. MCGOVERN-FAGG,[c] MONTGOMERY L. FLORA,[d,b,c] CAMERON R. HOMEYER,[a,c] AND JOHN K. WILLIAMS[c,h]

[a] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*
[b] *NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma*
[c] *NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography, University of Oklahoma, Norman, Oklahoma*
[d] *Cooperative Institute for Severe and High-Impact Weather Research and Operations, University of Oklahoma, Norman, Oklahoma*
[e] *School of Computer Science, University of Oklahoma, Norman, Oklahoma*
[f] *Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado*
[g] *Department of Earth Sciences, Central Michigan University, Mount Pleasant, Michigan*
[h] *The Weather Company, Brookhaven, Georgia*

(Manuscript received 28 March 2024, in final form 3 October 2024, accepted 22 October 2024)

ABSTRACT: Hailstorms cause billions of dollars in damage across the United States each year. Part of this cost could be reduced by increasing warning lead times. To contribute to this effort, we developed a nowcasting machine learning model that uses a 3D U-Net to produce gridded severe hail nowcasts for up to 40 min in advance. The three U-Net dimensions uniquely incorporate one temporal and two spatial dimensions. Our predictors consist of a combination of output from the National Severe Storms Laboratory Warn-on-Forecast System (WoFS) numerical weather prediction ensemble and remote sensing observations from Vaisala's National Lightning Detection Network (NLDN). Ground truth for prediction was derived from the maximum expected size of hail calculated from the gridded NEXRAD WSR-88D radar (GridRad) dataset. Our U-Net was evaluated by comparing its test set performance against rigorous hail nowcasting baselines. These baselines included WoFS ensemble Hail and Cloud Growth Model (HAILCAST) and a logistic regression model trained on WoFS 2–5-km updraft helicity. The 3D U-Net outperformed both these baselines for all forecast period time steps. Its predictions yielded a neighborhood maximum critical success index (max CSI) of ~0.48 and ~0.30 at forecast minutes 20 and 40, respectively. These max CSIs exceeded the ensemble HAILCAST max CSIs by as much as ~0.35. The NLDN observations were found to increase the U-Net performance by more than a factor of 4 at some time steps. This system has shown success when nowcasting hail during complex severe weather events, and if used in an operational environment, may prove valuable.

KEYWORDS: Hail; Lightning; Nowcasting; Machine learning; Artificial intelligence; Neural networks

1. Introduction

Hailstorms cause billions of dollars in damage each year in the United States (Gunturi and Tippett 2017). For example, on 28 April 2021, a single hailstorm passing over Norman, Oklahoma, caused one billion U.S. dollars in damage (NSSL 2021). The accurate forecasting of hail is important for increasing warning lead time, which can in turn give the populace time to anticipate the hazard, shelter, and protect vulnerable property. In particular, advancements to the near-term forecasting of hail may play a significant role in decreasing hail damage. This near-term (0–3 h) forecasting is referred to as nowcasting (American Meteorological Society 2022). Specifically, the present study aims to develop a model for the probabilistic nowcasting of severe hail (hail with diameter ≥ 1 in.) as hail larger than this threshold is the most hazardous and destructive.

Nowcasting hail is a challenging topic (e.g., Foster and Bates 1956; Brimelow et al. 2002; Adams-Selin and Ziegler 2016; Gagne et al. 2017; Adams-Selin et al. 2019). One difficulty is that the definition of a successful hail model itself is inconclusive in the hail community (Adams-Selin et al. 2023). Comparing models must be done with extra care and consideration due to this complication. Another difficulty is that the exact physical processes behind hail formation and growth continue to be an area of exploration for leading hail experts (Allen et al. 2020). What is generally accepted is that hail formation begins with the development of a hail embryo that subsequently moves through a storm updraft. Eventual hail diameter is a function of the time this embryo spends within the growth region on the periphery of the storm updraft and the availability of supercooled water (Nelson 1983; Dennis and Kumjian 2017; Kumjian and Lombardo 2020). These concepts are well summarized in Allen et al. (2020).

Hail processes are typically encapsulated within a set of complex interactions referred to as microphysics (Labriola et al. 2019; Allen et al. 2020; Morrison et al. 2020). These physical processes are in constant competition with one another, and their relative contributions to hail growth are incompletely known given the lack of direct observations (Morrison et al. 2020). Microphysical processes occur at scales

Corresponding author: Tobias G. Schmidt, tgschmidt@shaw.ca

DOI: 10.1175/AIES-D-24-0026.1 e240026

far less than 1 m. However, computational limits keep operational numerical weather prediction (NWP) model grid cells at the scale of a few kilometers at their smallest (Hong and Dudhia 2012; Yano et al. 2018). This greatly exceeds the scale necessary to resolve these microphysical processes. As such, most current methods for the forecasting/nowcasting of hail must use a statistics-based or physical surrogate of these processes at the scale of a single NWP model grid cell (Milbrandt and Yau 2005; Stensrud et al. 2009, 2013; Labriola et al. 2019; Heinselman et al. 2024).

Many different statistical and physics-based hail models have found limited success in forecasting/nowcasting hail for various lead times (Adams-Selin et al. 2023). Some methods are exclusively based on environmental variables (e.g., Gensini et al. 2021), while other methods use storm variables within ongoing convection in model simulations (Adams-Selin and Ziegler 2016; Gagne et al. 2019). Producing methods based exclusively on environmental variables is limiting since our understanding of how these variables impact hail microphysics is incomplete (Allen et al. 2020). This has resulted in a wide spread of different physics-based approaches to hail forecasting/nowcasting using environmental variables (e.g., Thompson et al. 2003; Allen et al. 2011; Mohr and Kunz 2013; Johnson and Sugden 2014; Tuovinen et al. 2015; Taszarek et al. 2020).

One such physics-based approach that has shown more success is the Hail and Cloud Growth Model (HAILCAST) (Brimelow et al. 2002; Jewell and Brimelow 2009; Adams-Selin and Ziegler 2016). In particular, the version that often runs in convection-allowing models is known as Weather Research and Forecasting (WRF)-HAILCAST (Adams-Selin and Ziegler 2016). This version works by first examining the directly simulated updrafts produced within the WRF Model simulation (Skamarock et al. 2008). Provided these updrafts are persistent (>15 min) and sufficiently strong, the WRF column properties are one-way coupled to the time-dependent WRF-HAILCAST. It then creates a one-dimensional simulation where five embryos are injected and allowed to rise, fall, and grow before being returned in terms of a maximum hail diameter and standard deviation. Versions of HAILCAST have seen extensive use, including in real-time operations (e.g., Jewell and Brimelow 2009; Adams-Selin and Ziegler 2016; Dyson et al. 2021; Malečić et al. 2022; Adams-Selin et al. 2023). Despite being one of the more popular hail prediction models, it is still constrained by the aforementioned grid size limitations and incomplete physical understanding of hail growth.

Another approach to hail nowcasting is to use machine learning. One advantage of machine learning is that it is excellent at finding complex relationships necessary for solving large-data problems without the need to pre-engineer any physical process into the model. Not requiring this engineering implies that these methods can potentially help circumvent some of the limitations posed by the various physics-based approaches to hail growth. Indeed, in recent years, machine learning has become a popular method for solving severe weather problems including hail forecasting (Gagne et al. 2017; McGovern et al. 2017; Czernecki et al. 2019; Gagne et al. 2019; Flora et al. 2021; McGovern et al. 2023). It has likewise seen a growth in use throughout the broader meteorological community (Chase et al. 2022). Some examples of successful machine learning uses in hail forecasting/nowcasting include using random forests, logistic regression, and *k*-means clustering for all-hazard severe weather prediction (McGovern et al. 2017), next-day hail forecasting using random forests trained on NWP models (Gagne et al. 2017), and using convolutional neural networks to predict hail from convection-allowing NWP models for an hour-long period (Gagne et al. 2019).

The generation of machine learning forecasts for hail requires a high-quality dataset to provide predictive information. A source for such information is the rapidly available short-term forecast ensemble known as the Warn-on-Forecast System (WoFS) (Stensrud et al. 2009, 2013; Gallo 2017; Gallo et al. 2022, 2024; Heinselman et al. 2024). This system has shown skill in providing probabilistic severe weather guidance to the National Weather Service and has been leveraged in prior studies by the application of random forests (Flora et al. 2021). WoFS includes a rapidly refreshed data assimilation system which increases model performance, especially at the shortest lead times (Hu and Xue 2007; Stensrud et al. 2013). By rapidly assimilating radar and satellite data, the WoFS can generate accurate predictions of individual thunderstorms, especially storms that are preexisting and persistent (Stensrud et al. 2009; Guerra et al. 2022).

WoFS is designed for high-resolution severe weather nowcasting purposes and thus provides a useful source of model input. To best leverage these inputs, we elected to use U-Nets, which are a type of neural network (Ronneberger et al. 2015; Çiçek et al. 2016; Huang et al. 2020). These models have also been shown to perform well in many recent thunderstorm tasks (Lagerquist et al. 2020; McGovern et al. 2023) as well as other atmospheric applications (Justin et al. 2023). They have an advantage over more traditional machine learning methods such as random forests as they do not treat each point of data (or grid point) as independent from one another. Rather, they use small filters known as kernels to translate adjacent spatial features (e.g., a spatial wind field gradient) into numbers interpretable by machine learning. Additionally, U-Nets use gridded data for both input and output, implying additional postprocessing is not required to convert the model output to the gridded format necessary for visualization in most operational meteorology applications.

One further advantage of machine learning over physics-based methods is that machine learning is capable of combining inputs from traditionally disparate sources together to find useful relationships. We looked to exploit this advantage with a combination of real-time observations and the WoFS predictors. Some severe weather nowcasting/forecasting studies have found success in exclusively using real-time observations as the input predictors for a machine learning model (e.g., Billet et al. 1997; Huang et al. 2019). Others have exclusively used NWP model output to create their input datasets (e.g., Gagne et al. 2017; Flora et al. 2021; Gensini et al. 2021; McGovern et al. 2023), and some researchers combine both these methods together, yielding stronger results than what is obtainable solely using real-time observations or numerical weather prediction data (e.g., Czernecki et al. 2019; Lagerquist et al. 2020; Scarino et al. 2023). This hybrid approach is attractive

TABLE 1. Machine learning data sources summarized. W-up refers to updraft speed in the upward direction. CAPE is convective available potential energy. CIN is convective inhibition. MU stands for most unstable. SFC means surface based. SCP is the supercell composite parameter. SRH is storm relative helicity. The U direction is longitudinal, and the V direction is latitudinal. The term T_d is the dewpoint temperature. LFC refers to the level of free convection. LCL is the lifting condensation level.

NWP predictors (Warn-on-Forecast System)				
Updraft and thermodynamics	Severe weather composites	Kinematics	Humidity	Storm alt
W-up	Hail (WRF graupel)	UH 2–5 km	T_d (2 m)	Freezing level
MU CAPE	HAILCAST	SRH 0–1 km	—	MU LFC
MU CIN	SCP	SRH 0–3 km	—	MU LCL
SFC CAPE	—	V shear 0–6 km	—	—
SFC CIN	—	U shear 0–6 km	—	—
Observation predictors (Vaisala lightning network)				
Lightning event count per 3 × 3 km bin				
GridRad truth labels				
MESH				

as it theoretically exploits the accuracy of real-time observations, while using variables from NWP models that do not necessarily require surrogates for microphysics (e.g., nonconvection variables). An additional advantage of a hybrid approach is that it can compensate for the delays in forecast availability that occur due to data assimilation, model integration, and forecast output postprocessing. As such, we implemented a hybrid approach using, as input, a combination of high-resolution U.S. lightning observations from Vaisala's National Lightning Detection Network (NLDN) dataset (Murphy et al. 2021) and WoFS output. We theorized that the use of a U-Net along with hybrid NWP/lightning observation predictors could provide an effective and novel hail nowcasting model.

2. Methods

a. Datasets

1) WARN-ON-FORECAST SYSTEM (NWP)

Our machine learning–based nowcasting solution adopts the hybrid NWP/observation approach to selecting predictors. Our chosen NWP source was the WoFS ensemble. Operationally, these forecasts are initialized every 30 min, with output every 5 min. Each 5-min interval is included as input for our U-Net. Radar and satellite data are assimilated into WoFS every 15 min, and conventional observations are assimilated every hour. This high rate of data assimilation allows for detailed convective information to enter WoFS in a timely manner, which is essential in convection time scales. The WoFS ensemble is made up of 18 members, each with a horizontal grid resolution of 3 km. All of our data are sampled from the storm events in 2017–21 that were covered simultaneously by WoFS, the observations, and our labels. The WoFS domain had 300 × 300 grid points in 2019–21 and 250 × 250 grid points in 2017–18.

For all machine learning models examined in this study, WoFS predictors were produced using a member-agnostic approach to the 18-member ensemble, where all ensemble members are treated the same and converted to bulk samples without an additional data channel. For example, a single WoFS run produces 18 separate results from all 18 ensemble members, and when using a member-agnostic approach, 18 samples are produced for this single run as opposed to 1 sample with 18 sets of features. This approach was done so the AI could better generalize to WoFS output. Both the models with lightning observations included and those without are trained with this approach. For comparison purposes, several of the NWP-only machine learning models are used to independently make test set predictions on each WoFS ensemble member before taking an ensemble average of their outputs. This was used to examine the advantages of exploiting an ensemble-based NWP product such as WoFS for machine learning predictors.

Our WoFS input is comprised of 17 fields (Table 1). The fields we chose to use from WoFS can be broken into five categories: updraft and thermodynamic variables, WoFS hail/severe weather composites, kinematic variables, a humidity variable, and storm altitude information. Derived fields were chosen over raw fields because these variables reduce dimensionality for the U-Net, which in turn may lessen the amount of physical understanding the U-Net must learn. We theorized that this would make it easier for the U-Net to learn the relationships needed for hail nowcasting. Using these fields also helped control the collinearity that may have occurred across the numerous raw fields that would be required in place of a smaller number of derived variables.

One WoFS field of particular importance is the HAILCAST field (Adams-Selin and Ziegler 2016). HAILCAST is used both as a predictor and as a baseline for evaluating our model. We make the WoFS HAILCAST baseline (not the predictor) probabilistic by taking the fraction of all 18 ensemble members with forecast hail greater than 1 in. in diameter. This was done so a more direct comparison could be made to our U-Net's probabilistic output. As discussed, our U-Net was trained using a member-agnostic approach to its WoFS predictors and therefore uses deterministic input (without accompanying ensemble statistics) to produce its own probabilistic output. Due to this, when averaged across the test set, the HAILCAST baseline gains a performance advantage over the member-agnostic (deterministic) U-Net. The HAILCAST baseline needs to use the complete ensemble distribution when we convert it to a probabilistic product and thus does have access to these ensemble statistics. This is an important consideration when evaluating the test set results shown later.

2) Vaisala lightning (observations)

WoFS forecasts do not become available until ~20 min after WoFS initialization due to the time required for data assimilation, model integration, and forecast output postprocessing. To address this, additional lightning observations were used to assist in bridging this gap in the early stages of the forecast. Twenty minutes of observations were available from this wait time and were fed into the U-Net. As this WoFS forecast is initialized 20 min before it becomes available, 20 min of WoFS hindcasting is also available to be set alongside the lightning data. Our machine learning forecast can begin after these 20 min have passed, so our machine learning forecast begins 20 min after WoFS initialization time (called machine learning forecast minute 0). In summary, before machine learning forecast minute 0, the input data are made up of 20 min of lightning and WoFS hindcasting data, while after machine learning minute 0, it is only made up of future WoFS data.

Although NWP can offer some amount of forecast skill to the U-Net on its own, the inclusion of observations should allow for a further boost to the U-Net's performance. In particular, radar, satellite, and other observations have been shown to add considerable skill to machine learning models used for nowcasting (Czernecki et al. 2019; Scarino et al. 2023). Vaisala's NLDN dataset was chosen to be the observations used in this study (Murphy et al. 2021). The primary reason for this selection was because Vaisala's NLDN dataset has a global counterpart of similar quality, ideally allowing for future versions of our model to be scaled to the global domain. Including other observations such as radar or satellite data would limit future versions to the CONUS domain and as such were withheld for this particular study. Vaisala's dataset also has high resolution and scalability, making it ideal for hail applications. In general, this dataset appeared optimal because of the strong relationship between lightning activity and hail formation (Changnon 1992; Feng et al. 2007). A binning algorithm was used to group lightning counts together into a gridded product. Specifically, the number of lightning events that occur over a 5-min period in each 3 × 3 km grid point was used as a predictor. Twenty minutes of these data were used for every forecast run to align with the estimated WoFS latency time.

3) GridRad maximum expected size of hail (truth labels)

Our truth labels were extracted from the gridded NEXRAD WSR-88D radar (GridRad) dataset (Murillo and Homeyer 2019; Murillo et al. 2021; School of Meteorology/University of Oklahoma 2021). Specifically, we used the maximum expected size of hail (MESH) (Witt et al. 1998) calculated from the GridRad-severe distribution of this product, which used version 4.2 of the GridRad algorithm (Murphy et al. 2023). MESH is a hail size estimate that is drawn directly from radar observations. It has data for every 5 min across CONUS in 1-km grid cells. MESH was originally created by fitting hail size from hail reports to the severe hail index (SHI) (Witt et al. 1998) via a power law. MESH is known to struggle with hail size characterization at the smallest and largest diameters (Cintineo et al. 2012; Ortega 2018). However, recent updates to MESH using GridRad with larger training samples have shown improvements in performance (Murillo and Homeyer 2019).

GridRad includes two separate hail-to-SHI power-law fits: one to the 75th percentile of the hail distribution (MESH_{75}) and the other to the 95th percentile (MESH_{95}) (Murillo and Homeyer 2019). Several studies have tested the relationship between observed MESH and hail sizes at the ground and found that thresholds around 30 mm for severe hail and 51 mm for significant severe hail were ideal (Murillo and Homeyer 2019; Wendt and Jirak 2021; Murillo et al. 2021). For the applications here, MESH_{95} was chosen, as it is focused on the characterization of larger hail. This power-law fit was used as the basis for our severe hail truth labels, and subsequently, all grid points with $\mathrm{MESH}_{95} > 25.4$ mm (>1 in. for severe hail) were labeled as positive, while all other grid points were labeled as negative.

One motivation for choosing GridRad MESH is that a radar-based product does not have human bias, as opposed to using a confirmation-based source such as storm reports (Murillo et al. 2021). Another is that its gridded format means no major interpolation or preprocessing is required to ensure it can be used in pixelwise evaluations (such as what would be required when using storm reports). Resampling is done to convert its 1-km grid cells to the 3-km grid used in WoFS so that direct evaluations are possible.

b. Machine learning architecture (U-Net)

The particular variant of U-Net we selected for this study was a three-dimensional U-Net 3+ (Ronneberger et al. 2015; Çiçek et al. 2016; Huang et al. 2020). Three-dimensional U-Nets were originally created for the detection of unhealthy tissue within medical imaging. Most commonly, they were used to label unhealthy pixels in spatial three-dimensional images of human brains. Three-dimensional U-Nets were optimal for this task as their gradient-detecting kernels would ensure that complete biological structures (such as a tumor) were resolved by the model rather than treating each pixel as independent data points. In meteorology, it is possible they could also be used successfully when labeling pixels of a spatial dataset, for example, when labeling different structures of a three-dimensional snapshot of a supercell.

For a nowcasting task, it was decided that the time dimension was of greater importance than a third spatial dimension. This is because, in nowcasting, the state of future meteorological variables is heavily dependent on their values in the proceeding minutes. As such, we elected to populate our three-dimensional U-Net with meteorological data made up of two horizontal spatial dimensions and the temporal dimension. With this system, the gradient-detecting kernels can find adjacency relationships across both space and time rather than

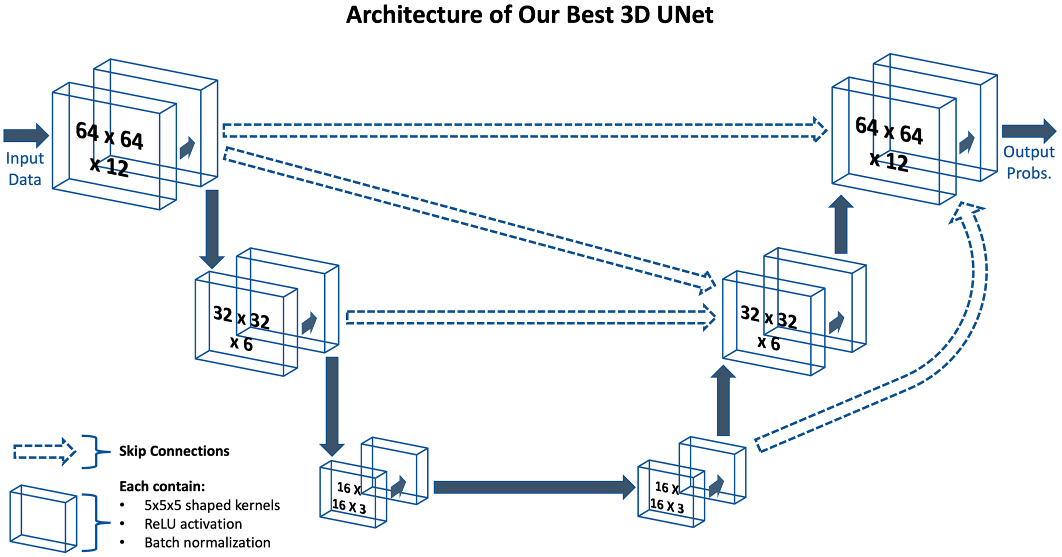

FIG. 1. Schematic diagram of the architecture for our best model. Note that all convolutional layers are 3D. There are two convolutional layers per network level and skip connections at each level (the latter is a defining characteristic of the U-Net++/U-Net 3+ variant). Each convolutional layer has batch normalization and a ReLU activation function. From the top, each network level has 8, 16, and 24 5 × 5 × 5 kernels, respectively. All hyperparameters for this model are available in Table A1 in the appendix.

being constrained to only space. More traditionally, a nowcasting task would use a two-dimensional U-Net with the time dimension relegated to additional feature channels; however, this does not offer the same built-in kernel advantages as what would be present in our system. It has been shown that this method of using the third U-Net dimension for time can provide increased model skill over the standard two-dimensional U-Net method (e.g., Bansal et al. 2022). In particular, it was hypothesized that a 3D U-Net with a time dimension would produce better hail predictions than a standard 2D U-Net since hail growth is very temporally dependent.

This temporal dimension also allowed the simultaneous output of multiple forecasted time steps without the need for multiple machine learning models or a more complex architecture. Each data sample comprised a sequence of twelve 64 × 64 patches valid at 5-min increments. Thus, each sample spanned 60 min, the latter 40 min of which was forecast time. The U-Net was trained with 3368 training samples and 846 validation samples (which with all 12 time steps comprised 40 416 and 10 152 patches, respectively). Finally, the 3+ variant of U-Net was selected because the characteristic skip connections present in this variant assist in reducing model overfitting issues and increase the quality of feature processing (Huang et al. 2020). The detailed structure of our U-Net is given in Fig. 1. The set of hyperparameters used in our final model is displayed in Table A1. These hyperparameters were found using a grid search.

For operational environment considerations, training took ~3 days (with a complete hyperparameter search) on four NVIDIA A100 graphics processing units (GPUs) and the prediction step took ~74 s for one complete WoFS domain on an Intel Xeon E5-2670 V3 2.6-GHz CPU. Note that this number is only the prediction time itself, and it does not include the time required for preprocessing tasks such as minimum–maximum normalization, data slicing, data filtering, and patch creation.

c. *Dataset preprocessing and evaluation metrics*

To prepare our data for partitioning into training, validation, and test sets, we clustered samples from the 2017–21 period together by storm event using the density-based clustering algorithm known as density-based spatial clustering of applications with noise (DBSCAN) (Birant and Kut 2007). A storm event is defined as a set of samples with WoFS initialization times that are less than 6 h apart. In total, 68 storm events were produced, with each containing hundreds to thousands of patches. 20% of these events were then randomly selected to produce the test set, which was set aside for the final model evaluation. The training and validation sets were then produced using stratified grouped fivefold cross validation (Stone 1974) from the remaining storm events. To enforce partition independence, grouping was used to ensure samples from the same storm event could not be present across the partitions. The stratification ensured similar base rates of severe hail across the training and validation sets so that the training set performance would be more representative of the validation set performance.

Each sample was then created with all 18 predictors across 12 time steps for different WoFS initialization times. In real-world applications, the lightning data would only be available up to the time when the U-Net is run, which occurs 20 min after WoFS initialization due to the discussed WoFS latency. Therefore, we placed these observations exclusively in the

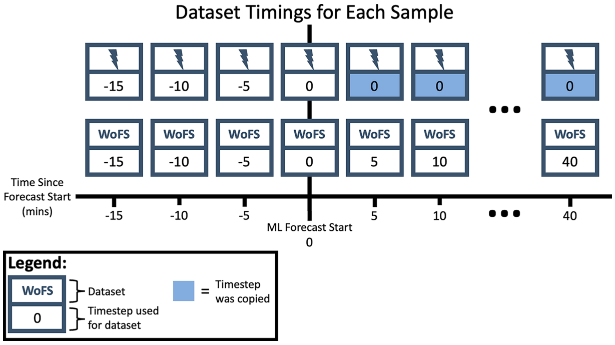

FIG. 2. Diagram outlining the timing of the two datasets used in each sample. Note that the lightning data are copied forward to fill out the forecast period.

first few time steps of each sample (prior to the defined forecast period start of "minute 0"). In total, for each sample, there were 17 WoFS features with 12 time steps (60 min) and 1 lightning observation feature with 4 time steps (20 min). Since U-Nets require that all input predictors have the same dimensions, the lightning observations must have the same number of time steps as the WoFS predictors. This issue was resolved by copying forward the final time step of the lightning data throughout the remaining 40 min. This method was chosen to encourage the U-Net to use the lightning observations in all time steps of the forecast period, not just in the first few steps resolved by the size of the U-Net's kernel. This was deemed particularly important as observation predictors could provide predictive value in all time steps of the short time period covered in a nowcasting problem. The layout of each sample is visualized in Fig. 2.

After our dataset was partitioned using cross validation and the samples were generated, normalization was applied using the minimum–maximum scaler:

$$x_{\text{scaled}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \tag{1}$$

Finally, a Gaussian expansion (Earnest et al. 2023) was applied exclusively to the training set. This expansion worked by converting isolated hail labels (or 1 values) to Gaussian distributions of 1 values surrounded by rings of 0.66 values which were in turn surrounded by rings of 0.33 values (Fig. A1). This expansion was applied to both the space and time dimensions. The objective was to train the model to accommodate modest phase errors in the hail predictions. This phase error tolerance was expected to be particularly useful for this study given the highly localized nature of severe hail when compared to the gridcell size and due to the displacement of U-Net predicted cells relative to GridRad cells, owing to model predictions not necessarily matching observation locations with increasing lead time.

After this preprocessing was concluded, training was performed with a hyperparameter grid search. Once a trained U-Net was selected from the search, predictions were generated for all patches in the test set. For the case studies, a full set of patches was necessary to fill in each complete WoFS domain. However, if no action is taken, there would be erroneous predictions present along the boundaries of the stitched-together patches that make up each domain. This is corrected by first producing three additional domains for each case study. These domains are made up of patches shifted by 32 grid points to the east, south, and southeast, respectively. All four domains are then averaged together to remove the noise that would normally exist along the internal patch boundaries of the original domain. This process would also most likely be required when running this U-Net operationally.

The primary metric used for forecast evaluation in this study was the critical success index (CSI). The equation for CSI is given in Eq. (2), where TP is the true positives, FN is the false negatives, and FP is the false positives. CSI ranges from 0 to 1. Zero CSI indicates no events are correctly predicted, while 1.0 CSI indicates that all events are correctly predicted. The CSI is calculated as follows:

$$\text{CSI} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}. \tag{2}$$

Both pixelwise and neighborhood CSIs are used in this study. Neighborhood CSI includes a 6-km radius tolerance for forecast hits. It is produced by expanding the truth labels by 6 km for exclusively the true-positive and false-positive calculations. The true and false negatives were not expanded to avoid double counting. To calculate the contingency table statistics for CSI, the probabilistic output from the U-Net must be binarized. The maximum CSI (max CSI) is calculated by taking the maximum CSI across all possible probability thresholds for this binarization. Max CSI is often displayed in performance diagrams (as seen in later figures).

3. Results

a. *Performance evaluation*

To evaluate the U-Net, we compared its test set performance to multiple baselines throughout the 40-min forecast period (Fig. 3). To consider multiple tolerances, we examined both pixelwise and neighborhood maximum CSI derived for a variety of model configurations and optimizations. Across all predictive time steps and in both metrics, we found that the U-Net with lightning observations included outperformed the baselines by a sizable margin. For the higher tolerance 6-km radius neighborhood max CSI, this margin was as high as ~0.50, ~0.25, and ~0.15 at 0, 20, and 40 min, respectively. For both metrics, the U-Net with lightning outperformed all baselines by at least a factor of 2 and by more than a factor of 4 prior to minute 15. In the earlier forecast time steps, the larger performance margin over baselines observed with the lightning-containing U-Net was likely because this U-Net had access to real-time observations at the start, while the WoFS-based methods did not.

As there is considerable overlap between the 95% confidence intervals of the no-lightning U-Nets and the best logistic regression baseline, it is clear that the architecture is not the primary driver of the model's increased performance. The lightning observations are responsible for the majority of this

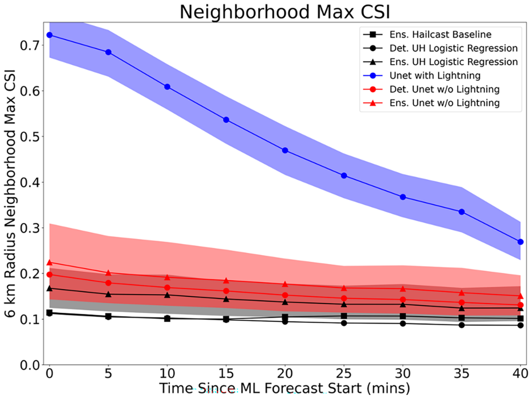

FIG. 3. Max CSI of U-Nets and several baselines for each time step of the test set's forecast period. Black indicates the various baselines, red indicates the U-Nets without lightning observations, and the blue line indicates the U-Net with lightning observations. Triangle markers indicate models that use the complete ensemble distribution of WoFS, circle markers indicate models that only use members of the WoFS ensemble deterministically, and square markers indicate the probabilistic WoFS HAILCAST that uses the complete ensemble. The shaded regions indicate 95% confidence intervals for the top model in each category (baselines, no lightning U-Nets, and lightning U-Net). (left) Max CSI time series with 6-km radius neighborhooding to allow for some tolerance. (right) Exact pixelwise max CSI with only the best models shown to reduce clutter.

performance increase. The improvement due to the lightning observations drops off rapidly as the forecast proceeds. This rapid performance drop-off may be the result of regular NWP forecast quality decay present in the WoFS predictors (Guerra et al. 2022), or it may indicate that the model overweights the lightning observations in later forecast time steps. It may also indicate that the forward copying of the last lightning observations is an inadequate system. Some combination of these three issues is also a possibility. This concept is discussed further in the conclusions.

Evaluation of the models with performance diagrams further reinforces the lightning-containing U-Net's skill (Fig. 4). This U-Net shows a performance lead over the NWP-only U-Nets and all baselines for both forecast time steps. These results also highlight that although the NWP-only models include ensemble-derived output, the purely deterministic

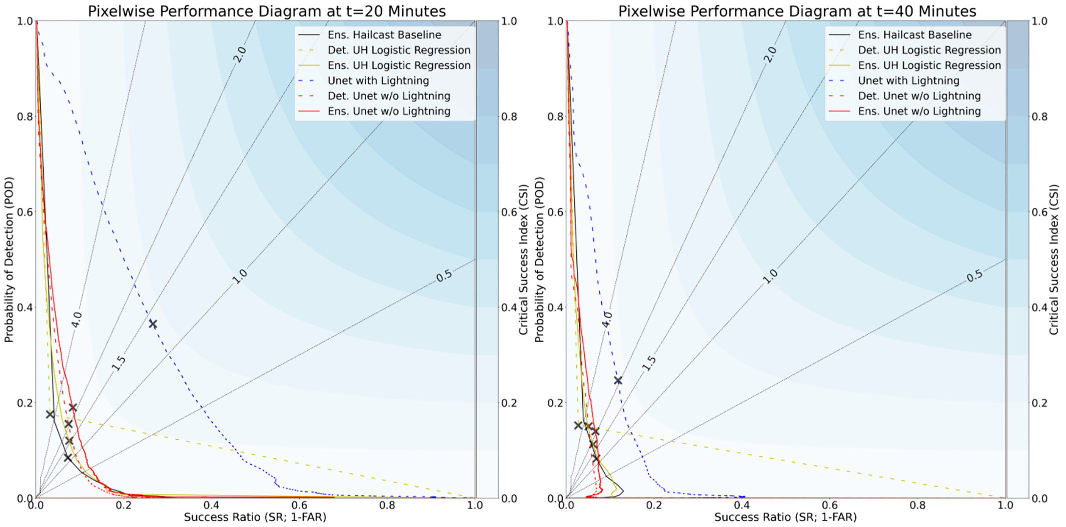

FIG. 4. Pixelwise performance diagrams at minutes 20 and 40 of the test set forecast period. The term POD = $t_p/(t_p + f_n)$, where t_p is the number of TPs and f_n is the number of FNs. The term SR = $t_p/(t_p + f_p)$, where f_p is the number of FPs. (left) The performance at minute 20. All data sources that used the entire WoFS ensemble are indicated with solid lines, while purely deterministic sources are indicated with dashed lines. The logistic regression sources are shown in yellow, HAILCAST is shown in black, the U-Nets without lightning observations are shown in red, and the U-Net with the observations is shown in blue. The max CSI values are marked with x values for each curve. (right) As in (left), but at minute 40.

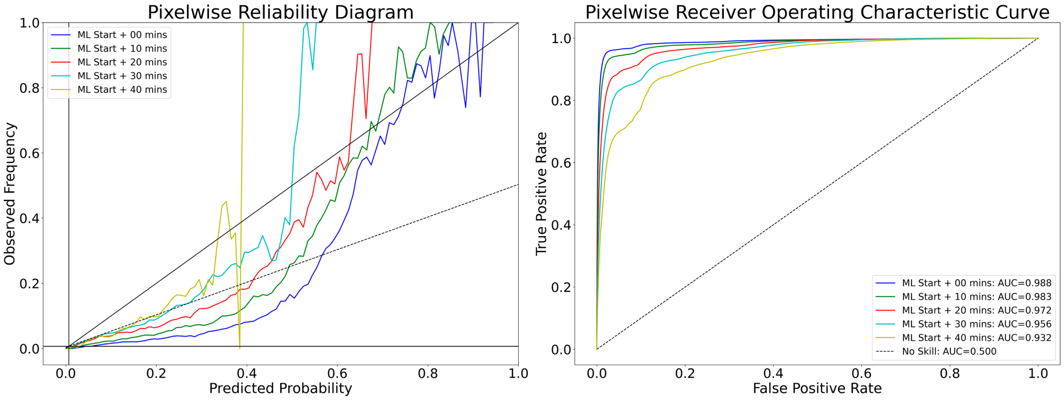

FIG. 5. (left) Test set pixelwise reliability color coded by forecast time step. The solid black vertical and horizontal lines represent the lines of climatology and no resolution, respectively. The dashed black line indicates the Briar skill score "no skill" line. True-positive rate = number of TPs/total number of positives. False-positive rate = number of FPs/total number of negatives. Reliability curves closer to the solid black diagonal line indicate greater reliability. (right) Test set pixelwise ROC curves color coded by forecast time step. Curves closer to the top left of the graph are more desirable. Areas under the curve (AUCs) are displayed in the legend.

lightning-including U-Net still outperforms them at each time step. Furthermore, these results indicate that this U-Net outperforms the baselines across nearly the entire success ratio (SR) × POD space and not just for the maximum CSIs. Therefore, regardless of the binarization probability thresholds chosen for the CSI, the lightning-including U-Net will outperform the baselines. However, despite the indicated model skill, these results also reveal a consistent overforecasting bias. An evaluation of model reliability across additional time steps is required to examine this further.

The reliability diagram in Fig. 5 reveals a consistent trend across time steps. Again, earlier time steps tend to overforecast hail, while later time steps tend closer to greater reliability. However, later time steps lose forecast confidence relative to starting magnitudes. Some of this confidence loss can be explained by the natural NWP quality decay of the WoFS predictors expected in later time steps. However, this can also be at least partially explained as an artifact of the U-Net structure. At the edge of each dimension resolved by a U-Net, the kernel incorporates fewer data since it is unable to view nonexistent adjacent grid points outside the patch domain. U-Nets therefore tend to produce poorer results near patch edges. In our U-Net, which includes the time dimension, the first and last time steps have cutoff issues where less desirable output is produced. The first time steps in the forecast period do not correspond with the first steps of the U-Net's time dimension (see Fig. 2), so this artifact is only observed in the last time steps. This problem is believed to also partially explain the weaker probabilities seen in the last time steps of the case studies shown in the following section. The receiver operating characteristic (ROC) curves (Fig. 5) show the expected decay of forecast skill with time. It should be noted that both the reliability diagram and ROC curves use pixelwise comparisons, and therefore, they show an evaluation with no tolerance.

b. Case studies

Bulk performance metrics can be misleading, so a view into model behavior through an interface resembling what might be used in an operational environment may be beneficial. Several case studies were selected from the test set for further evaluation. To highlight the full capacity of the model, each case study was chosen such that both success and failure cases could be considered. A success case is defined as when the U-Net predicts severe hail within close proximity to grid points of GridRad MESH > 1 in. A failure case is either when there are grid points of MESH > 1 in. without any nearby predicted severe hail contours from the U-Net or if there are contours of predicted severe hail without any grid points of MESH > 1 in. nearby. These cases can also be interpreted as nonbinary with ranging degrees of success or failure. This range is defined by how much spatial overlap occurs between the U-Net prediction contours and the GridRad MESH > 1 grid points. It is also defined by the magnitude of the probabilities of severe hail from the U-Net at each forecast time step.

1) 18 MAY 2017 1915 UTC

The first case study was selected to showcase a scenario filled mostly with success cases, but which also contained a few failure cases (Fig. 6). This was to highlight that the U-Net could perform well, but still have some limitations during a single severe weather event. During the forecast period starting at 1915 UTC, three isolated storms near the southwestern corner of Oklahoma produced many severe and significant severe hail reports. These storms are labeled B, C, and D. The U-Net generally shows good skill in predicting higher severe hail probabilities in the vicinity of these hail-producing storms. However, these results also show that the evolution of severe hail for storms B and C was not predicted correctly by the U-Net and that these predictions display considerable spatial displacement error by the end-of-forecast period. This arises from the differences between storm motion in reality (which produces the MESH labeling) versus the U-Net predictions driven by WoFS storm motion and an extrapolation of the lightning observations. These two storms are examples

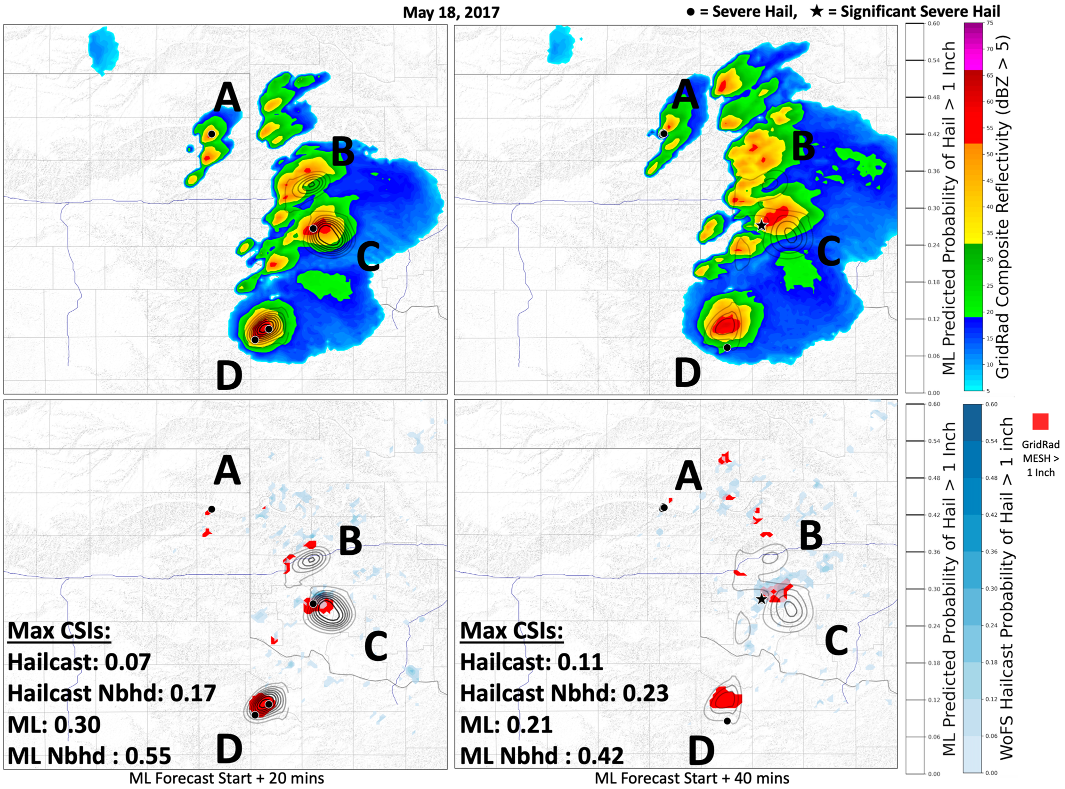
May 18, 2017
● = Severe Hail, ★ = Significant Severe Hail
A
B
C
D
Max CSIs:
Hailcast: 0.07
Hailcast Nbhd: 0.17
ML: 0.30
ML Nbhd : 0.55
Max CSIs:
Hailcast: 0.11
Hailcast Nbhd: 0.23
ML: 0.21
ML Nbhd : 0.42
ML Forecast Start + 20 mins
ML Forecast Start + 40 mins
ML Predicted Probability of Hail > 1 Inch
GridRad Composite Reflectivity (dBZ > 5)
WoFS Hailcast Probability of Hail > 1 inch
🟥 GridRad MESH > 1 Inch

FIG. 6. Case study of 18 May 2017 with U-Net forecast start time at 1915 UTC. On all plots, the black contour lines indicate the U-Net output. Greater probabilities of hail are represented with increasing opacity. (top) U-Net forecast contours overlaid onto GridRad composite reflectivity observations for reference. (bottom) U-Net forecast contours overlaid on blue shaded contours that indicate the WoFS ensemble HAILCAST forecast and red grid points where GridRad MESH > 1 in. (left) 20 min into the forecast period. (right) 40 min into the forecast period. Severe and significant severe hail reports are labeled with black dots and stars, respectively. Each storm cell cluster of note is labeled with a letter for identification needed in discussion.

of partial success cases, while the other isolated storm was a strong success case.

These results additionally suggest that the U-Net mostly avoids overpredicting hail in areas of reflectivity not associated with hail reports or GridRad MESH. In general, the U-Net's severe hail predictions align more accurately with MESH severe hail occurrences when compared to the ensemble WoFS HAILCAST product. In the north, storm A, which produced a severe hail report, does not have any hail forecast by the U-Net. This seems to indicate a possible failure case; however, only a small amount of severe hail was estimated by the GridRad MESH at this location. Alternatively, it is possible that MESH spuriously produces a weak signal for this storm due to its distance from the nearest radar site (i.e., is still a false negative).

The storms that develop during the course of the forecast period to the north of storm B are not well forecasted by the U-Net. At minute 20, these cells do not have much MESH associated with them; however, at minute 40, they do have a considerable MESH signal. It should be noted that these storms do not include storm reports so it is possible MESH may be overestimating severe hail in this area. Assuming MESH is accurate, this highlights how the U-Net can underpredict storms that develop late into the forecast period because they do not have strong lightning activity during the earlier 20-min observation time. Ideally, the U-Net would gain more information from the WoFS predictors at this late stage so it could predict more severe hail with these late developing cells.

2) 28 MAY 2019 2245 UTC (SOUTHERN DOMAIN)

The second case study was selected to highlight a scenario where all storms were obvious severe hail producers (Fig. 7). The two cells of significance in this case study both presented success cases for the U-Net by minute 20; however, the U-Net's predicted location of the northern storm's hail lagged to the southwest by minute 40. This northern storm can still be considered a weak success case in minute 40 as the U-Net predictions partially cover the grid points of GridRad MESH > 1 in. at this time step. This lagging behavior appears similar to the behavior observed in the first case study where the predictions of more northern hail have considerable spatial error by the end of the forecast. This is likely due to the NWP forecast decay in the

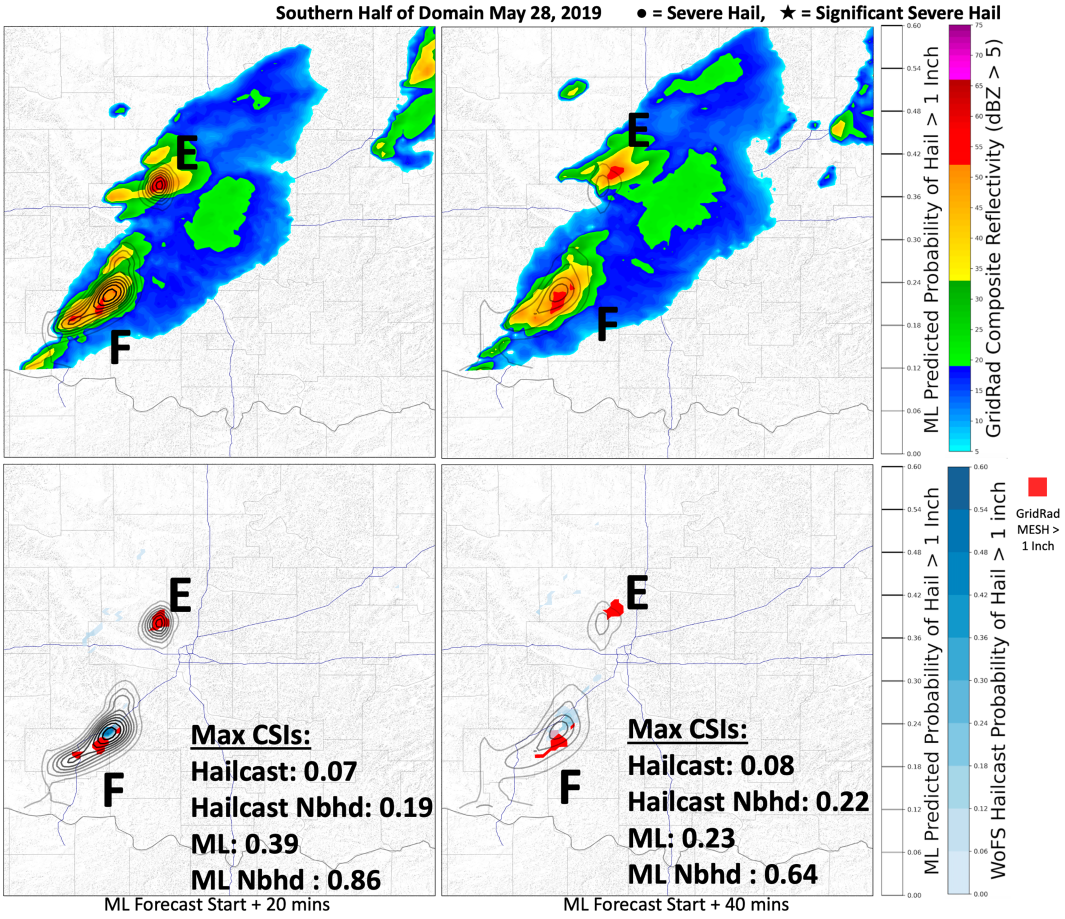
Southern Half of Domain May 28, 2019
● = Severe Hail, ★ = Significant Severe Hail
E
F
E
F
ML Predicted Probability of Hail > 1 Inch
GridRad Composite Reflectivity (dBZ > 5)
E
F
Max CSIs:
Hailcast: 0.07
Hailcast Nbhd: 0.19
ML: 0.39
ML Nbhd : 0.86
E
F
Max CSIs:
Hailcast: 0.08
Hailcast Nbhd: 0.22
ML: 0.23
ML Nbhd : 0.64
ML Predicted Probability of Hail > 1 Inch
WoFS Hailcast Probability of Hail > 1 inch
🟥 GridRad MESH > 1 Inch
ML Forecast Start + 20 mins
ML Forecast Start + 40 mins

FIG. 7. As in Fig. 6, but for the southern half of the WoFS domain on 28 May 2019. The U-Net forecast start time was at 2245 UTC.

WoFS predictors, the end-of-forecast cutoff issues experienced by our U-Nets, or it may be associated with the increased difficulty of forecasting hail in storms experiencing interactions with the outflow of the neighboring storms to the south.

Despite the issues with the northern storm's U-Net forecast, this area of severe hail was not predicted whatsoever by WoFS HAILCAST. Additionally, the U-Net accurately predicted the severe hail associated with the lagging behind, emerging cell to the southwest of the southern storm. This smaller area of severe hail was missed by the WoFS HAILCAST product, again highlighting the improvement offered using the U-Net prediction model.

3) 28 MAY 2019 2245 UTC (NORTHERN DOMAIN)

The final case study was chosen to analyze a noisy and widespread severe weather event (Fig. 8). Overall, the U-Net performs well for this case but with more mixed success. All storms except those labeled "H" and "J" are mostly well predicted throughout the forecast period. The eastern of these two storms has limited coverage of GridRad MESH > 1 in. with some MESH grid points present at minute 20 of the forecast; however, these dissipate by minute 40. This MESH trend aligns with the steady decay in reflectivity observed in this area during the same period. Despite this limited coverage, the U-Net produced confident severe hail predictions for this location, which could be considered a possible false positive. The western storm H is a clearer failure case. GridRad MESH > 1 in. persists for this small cell throughout the forecast period despite very limited forecasting from the U-Net. The storm was given a ~0.06 probability of severe hail by the U-Net exclusively at minute 20. It is possible the U-Net failed to deliver a more confident forecast for this storm because such a small cell may have produced only small quantities of lightning in the early portions of the forecast period.

In general, relative to their starting values, the probabilities reveal a trend to smaller magnitudes in the final time steps (especially minute 40) of all three case studies. This appears to be further evidence of the U-Net cutoff issue discussed previously; however, this again must be balanced with consideration for the effect of natural NWP quality decay present in the WoFS predictors. Finally, the U-Net produced a false positive in the northeastern portion of the domain where there was no GridRad MESH > 1 in., limited reflectivity, and no hail reports. One possible explanation for this false positive is

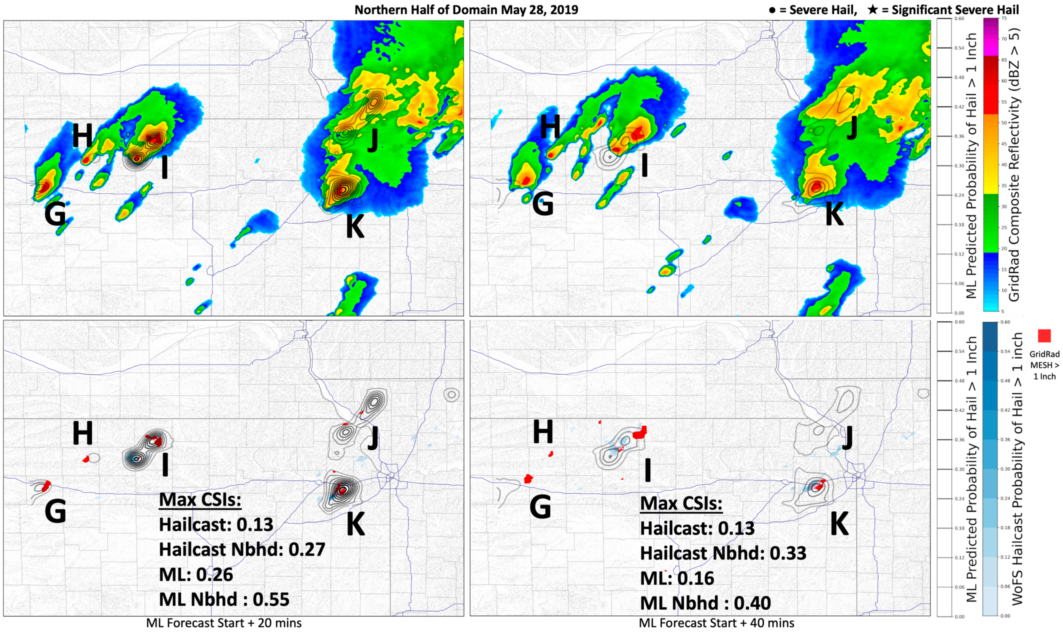
Northern Half of Domain May 28, 2019
● = Severe Hail, ★ = Significant Severe Hail
H
I
J
G
K
Max CSIs:
Hailcast: 0.13
Hailcast Nbhd: 0.27
ML: 0.26
ML Nbhd : 0.55
ML Forecast Start + 20 mins
Max CSIs:
Hailcast: 0.13
Hailcast Nbhd: 0.33
ML: 0.16
ML Nbhd : 0.40
ML Forecast Start + 40 mins
ML Predicted Probability of Hail > 1 Inch
GridRad Composite Reflectivity (dBZ > 5)
WoFS Hailcast Probability of Hail > 1 Inch
GridRad MESH > 1 Inch

Fig. 8. As in Fig. 6, but for the northern half of the WoFS domain on 28 May 2019. The U-Net forecast start time was at 2245 UTC.

a brief occurrence of lightning activity in its vicinity at the start of the forecast period without subsequent storm intensification.

4. Conclusions and future work

In this study, we developed a severe hail nowcasting machine learning model for real-time applications, aimed at the 0–1-h time frame. Through the development of this model, we have reiterated the importance of incorporating real-time observations for nowcasting problems and augmenting NWP data. Our benchmarks for evaluation were the Warn-on-Forecast System (WoFS) HAILCAST ensemble-derived probabilistic severe hail and the probability of severe hail sourced from updraft helicity (UH)-based logistic regression. These were used for a comparison of physics-based and machine learning methods, respectively. Our model outperformed both benchmarks for all time steps in the forecast period (Fig. 3).

Many studies have found varying degrees of success when using traditional physics-based approaches to hail modeling (e.g., Thompson et al. 2003; Jewell and Brimelow 2009; Allen et al. 2011; Mohr and Kunz 2013; Johnson and Sugden 2014; Tuovinen et al. 2015; Adams-Selin and Ziegler 2016; Taszarek et al. 2020). HAILCAST in particular has been popular for some time as a primary means of hail nowcasting/forecasting (e.g., Jewell and Brimelow 2009; Adams-Selin and Ziegler 2016; Trapp et al. 2019; Dyson et al. 2021; Malečić et al. 2022). The U-Net outperformed the WoFS HAILCAST both with and without the lightning observations. This highlights the value of a machine learning solution to this problem, even without the additional observation-based predictor considerations.

Other studies have explored some of these machine learning solutions to the hail forecasting/nowcasting problem (e.g., Gagne et al. 2017; McGovern et al. 2017; Gagne et al. 2019; Flora et al. 2021). Flora et al. (2021) in particular used WoFS predictors and machine learning methods that differed from U-Nets in a nongridded solution. Our U-Net outperformed the logistic regression benchmark representing less complex machine learning models using WoFS predictors and while keeping all data gridded. This allowed for useful nowcasting data that are immediately verifiable in the gridded format often used in an operational environment without needing additional postprocessing.

Another approach that may offer a more skillful forecast is to use an advanced machine learning model such as a vision transformer. These results highlighting a positive trend in performance across progressively more modern models support the validity of this approach. It is possible transformers would offer a skill increase without needing to change any of the predictors. Additionally, transformers offer an enticing alternative to U-Nets as they have recurrent logic built into their architecture, rendering additional modifications so the model can resolve temporal trends in weather (such as what was done with our third U-Net dimension) unnecessary.

Last, the vast increase in performance observed when adding the Vaisala lightning observations to our predictors provides evidence supporting the claim that observations are critical to nowcasting problems. This lightning predictor multiplied the model's max CSI by as much as ~3 times early in the nowcasting period (compared to the nonobservation using U-Net). The benefits of incorporating observational predictors have been demonstrated in several other studies (Czernecki et al. 2019; Leinonen et al. 2022). One such example is the model produced for Czernecki et al. (2019), where CSI performance values nearly doubled when radar observations and ERA5 reanalysis were combined for use in a large hail machine learning model.

In general, our model performance is similar to what has been observed in other machine learning studies, especially studies which use similar forms of neural networks (Gagne et al. 2019; Flora et al. 2021).

Despite the successes of our model, we must note several limitations that may be addressable in future studies. Most issues relate to the decay in model performance late in the forecast period observed in the results. These issues must all be balanced with a consideration for the natural decay in forecast quality we would expect from numerical model forecast data, a property seen in WoFS forecasts (Guerra et al. 2022). However, since several limitations have been identified that are related to the U-Net end-of-forecast quality, it is plausible at least some of the U-Net quality decay can be explained by them. One explanation for this decay is what was referred to as the cutoff issue. This was caused by the U-Net's kernel being unable to resolve data past the edges of each patch, thus causing quality drop-offs along the boundaries of all three dimensions, including time. A possible solution to this is extending the forecast period by increasing the size of the U-Net's time dimension. This might not require much additional effort as U-Nets are highly scalable and would have the added benefit of a deeper U-Net. A deeper U-Net may learn more complex relationships between the predictors and the potential for severe hail. The last few time steps of the expanded U-Net could be simply discarded to avoid their reduced quality without sacrificing earlier desired forecast time steps.

Another explanation for the forecast decay may be found in the U-Net's relationship with its NWP (WoFS) predictors. It has been observed that the U-Net overrelies on the lightning observations predictor in later forecast time steps (Schmidt 2023). A more optimized U-Net would value the WoFS forecasted predictors in later time steps over the lightning observations. One way to address this limitation would be to train the U-Net on WoFS predictors made up of the ensemble mean and standard deviation as opposed to the member-agnostic approach. These predictors generally represent higher-quality forecasts (over what is produced by the individual members). Therefore, it is probable the U-Net would value them to a greater degree and use them more in later time steps. Introducing a more refined loss function to the U-Net may be a further solution. This loss function could weigh predictors differently depending on the time step and therefore may encourage greater NWP predictor use for later time steps.

The explanation for why the WoFS predictors are less relied on may also be simpler. The forward copying of the lightning observations to fill out the U-Net's time dimension may encourage the U-Net to overrely on them in the later time steps. The dimension had to be entirely filled by all predictors, even the predictors which were only available in the first few time steps. The motivation for copying was to ensure the U-Net did not ignore the value of the observations completely in the last few time steps, but this compensation may have been overly strong. Instead, a solution may be to fill the remaining lightning time steps with zeros or decreasing weighted probabilities to discourage overuse in later forecast time steps.

Another future task may be to increase the use of lightning observations in later WoFS forecast time steps. The most up-to-date lightning observations could be included in later time steps to simulate the updating of an older WoFS run with newer observations. This may be prudent to maximize the advantages of the lightning observations for every possible forecast use scenario. Alternatively, consideration could be given to other observation predictors. This study used the Vaisala NLDN dataset exclusively (without secondary observations) to evaluate the NLDN data's contribution to the forecast skill in isolation. This was to ensure our model could be scaled to a global domain using Vaisala's comparable global lightning dataset, without being limited by the domains of other observational products. Using GridRad MESH as a predictor is an obvious step to immediately improve the forecast quality; however, it would change the scope of the problem we set out to solve to one focused on CONUS. In a separate study, adding GridRad MESH itself as a predictor may be prudent insofar as extra care is taken to ensure time steps evaluated against the MESH labels do not overlap with the observation time steps. Additionally, adding satellite or radar products such as cloud cover or reflectivity may offer an additional boost to model performance; however, as these data sources are assimilated as part of WoFS, this may cause a double counting issue depending on how they are used. With the success of the NLDN dataset's use in our framework, we envision that our severe hail nowcasting U-Net could be applied outside the CONUS once paired with other regional (or even global) convection-allowing models or ensembles. It is possible that using a U-Net/lightning predictor framework similar to ours could produce an operational worldwide severe hail nowcasting model with similar performance to what has been illustrated herein.

Acknowledgments. We thank Lydia Spychalla for inspiring this project with her related undergraduate work. This material is based on work supported by the National Science Foundation under Grant ICER-2019758. Funding was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement NA21OAR4320204, U.S. Department of Commerce. The computing for this project was performed at the OU Supercomputing Center for Education and Research (OSCER) at the University of Oklahoma (OU). The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of NOAA or the U.S. Department of Commerce.

Data availability statement. The experimental WoFS ensemble forecast data used in this study are not currently available in a publicly accessible repository. However, the data used to generate the results herein are available from the authors upon request. The NLDN data used in this paper belong to Vaisala Inc. These data are available for purchase or license from Vaisala for research use at https://www.vaisala.com/en/lp/request-vaisala-lightning-data-research-use. GridRad-Severe data are publicly available in the NCAR Research Data Archive at https://doi.org/10.5065/2B46-1A97. GitHub for all the code used in this project is found at https://github.com/ai2es/hail and https://github.com/ai2es/auto-patcher.

APPENDIX

Label Expansions and Hyperparameters

Figure A1 and Table A1 show the Gaussian label expansion method and a summary of all hyperparameters used, respectively. The Gaussian expansion system described in Fig. A1 is used to address the sparseness of hail labels. Table A1 lists all optimal, found hyperparameters for the best performing *U*-Net.

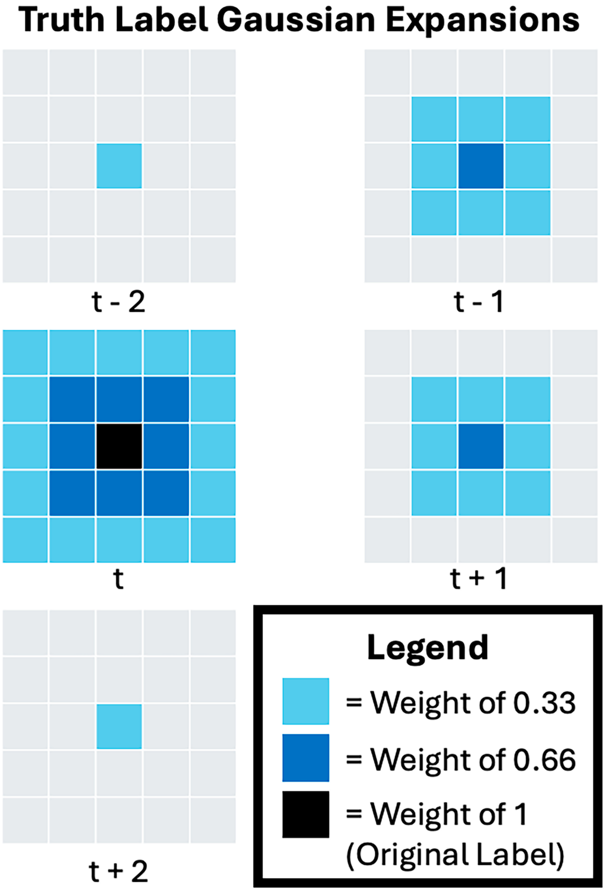

FIG. A1. Overview of how a Gaussian expansion of our labels is performed in both space and time across adjacent pixels. Time "*t*" indicates the time step at which the original label occurs.

TABLE A1. Optimal found, hyperparameters of best-performing 3D U-Net, and their search spaces. The symbol * means it had to be 3 because of 3D U-Net 3+ and sample dimension constraints.

	U-Net hyperparameters	
Name	Search space	Chosen
Convolutional layers	1, 2, 3	2
Kernel size	3, 5, 7	5 (5 × 5 × 5)
Activation	ELU, ReLU	ReLU
Number of kernels	4, 8, 16, 32	8
Depth	3*	3
Optimizer	Adam, Adagrad, SGD, RMSprop	SGD
Batch norm	Yes, no	Yes
3 plus	Yes, no	Yes
Batch size	32, 64, 128, 256	32
Learning rate	0.01, 0.001	0.001
L2 regularization	0.1, 0.05, 0.01, 0.005, 0.001, 0.0001, 0.000 01	0.01
L1 regularization	0.1, 0.05, 0.01, 0.005, 0.001, 0.0001, 0.000 01	0.001
Loss	Binary cross entropy	Binary cross entropy

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.

——, A. J. Clark, C. J. Melick, S. R. Dembek, I. L. Jirak, and C. L. Ziegler, 2019: Evolution of WRF-HAILCAST during the 2014–16 NOAA/Hazardous Weather Testbed Spring Forecasting Experiments. *Wea. Forecasting*, **34**, 61–79, https://doi.org/10.1175/WAF-D-18-0024.1.

——, and Coauthors, 2023: Just what is "good"? musings on hail forecast verification through evaluation of FV3-HAILCAST hail forecasts. *Wea. Forecasting*, **38**, 371–387, https://doi.org/10.1175/WAF-D-22-0087.1.

Allen, J. T., D. J. Karoly, and G. A. Mills, 2011: A severe thunderstorm climatology for Australia and associated thunderstorm environments. *Aust. Meteor. Oceanogr. J.*, **61**, 143–158, https://doi.org/10.22499/2.6103.001.

——, I. M. Giammanco, M. R. Kumjian, H. Jurgen Punge, Q. Zhang, P. Groenemeijer, M. Kunz, and K. Ortega, 2020: Understanding hail in the Earth system. *Rev. Geophys.*, **58**, e2019RG000665, https://doi.org/10.1029/2019RG000665.

American Meteorological Society, 2022: Nowcast definition. Glossary of Meteorology, https://glossary.ametsoc.org/wiki/Nowcast.

Bansal, A. S., Y. Lee, K. Hilburn, and I. Ebert-Uphoff, 2022: Tools for extracting spatio-temporal patterns in meteorological image sequences: From feature engineering to attention-based neural networks. arXiv, 2210.12310v2, https://doi.org/10.48550/arXiv.2210.12310.

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.

Birant, D., and A. Kut, 2007: ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data Knowl. Eng.*, **60**, 208–221, https://doi.org/10.1016/j.datak.2006.01.013.

Brimelow, J. C., G. W. Reuter, and E. R. Poolman, 2002: Modeling maximum hail size in Alberta thunderstorms. *Wea. Forecasting*, **17**, 1048–1062, https://doi.org/10.1175/1520-0434(2002)017<1048:MMHSIA>2.0.CO;2.

Changnon, S. A., 1992: Temporal and spatial relations between hail and lightning. *J. Appl. Meteor.*, **31**, 587–604, https://doi.org/10.1175/1520-0450(1992)031<0587:TASRBH>2.0.CO;2.

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.

Çiçek, Ö., A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, 2016: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016*, S. Ourselin et al., Eds., Lecture Notes in Computer Science, Vol. 9901, Springer, 424–432, https://doi.org/10.1007/978-3-319-46723-8_49.

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.

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.

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.

Dyson, L. L., N. Pienaar, A. Smit, and A. Kijko, 2021: An ERA-Interim HAILCAST hail climatology for southern Africa. *Int. J. Climatol.*, **41**, 262–277, https://doi.org/10.1002/joc.6619.

Earnest, B., A. McGovern, I. L. Jirak, and C. Karstens, 2023: Examining the role of the wildfire triangle in predicting wildfire occurrence for CONUS with the Unet3+ model. *Machine Learning for Wildfire Prediction, Modeling, and Processes I*, Denver, CO, Amer. Meteor. Soc., 10B.5, https://ams.confex.com/ams/103ANNUAL/meetingapp.cgi/Paper/419373.

Feng, G., X. Qie, T. Yuan, and S. Niu, 2007: Lightning activity and precipitation structure of hailstorms. *Sci. China*, **50D**, 629–639, https://doi.org/10.1007/s11430-007-2063-8.

Flora, M. L., 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.

Foster, D. S., and F. C. Bates, 1956: A hail size forecasting technique. *Bull. Amer. Meteor. Soc.*, **37**, 135–141, https://doi.org/10.1175/1520-0477-37.4.135.

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

Gagne, D. J., 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.

Gallo, B. T., 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Wea. Forecasting*, **32**, 1541–1568, https://doi.org/10.1175/WAF-D-16-0178.1.

——, and Coauthors, 2022: Exploring the watch-to-warning space: Experimental outlook performance during the 2019 Spring Forecasting Experiment in NOAA's Hazardous Weather Testbed. *Wea. Forecasting*, **37**, 617–637, https://doi.org/10.1175/WAF-D-21-0171.1.

——, A. J. Clark, I. Jirak, D. Imy, B. Roberts, J. Vancil, K. Knopfmeier, and P. Burke, 2024: WoFS and the wisdom of the crowd: The impact of the Warn-on-Forecast System on hourly forecasts during the 2021 NOAA Hazardous Weather Testbed Spring Forecasting Experiment. *Wea. Forecasting*, **39**, 485–500, https://doi.org/10.1175/WAF-D-23-0033.1.

Gensini, V. A., 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.

Guerra, J. E., P. S. Skinner, A. Clark, M. Flora, B. Matilla, K. Knopfmeier, and A. E. Reinhart, 2022: Quantification of NSSL Warn-on-Forecast System accuracy by storm age using object-based verification. *Wea. Forecasting*, **37**, 1973–1983, https://doi.org/10.1175/WAF-D-22-0043.1.

Gunturi, P., and M. Tippett, 2017: Managing severe thunderstorm risk: Impact of ENSO on U.S. tornado and hail frequencies. Willis Re Inc. Tech. Rep., 5 pp.

Heinselman, P. L., and Coauthors, 2024: Warn-on-Forecast System: From vision to reality. *Wea. Forecasting*, **39**, 75–95, https://doi.org/10.1175/WAF-D-23-0147.1.

Hong, S.-Y., and J. Dudhia, 2012: Next-generation numerical weather prediction: Bridging parameterization, explicit clouds, and large eddies. *Bull. Amer. Meteor. Soc.*, **93**, ES6–ES9, https://doi.org/10.1175/2011BAMS3224.1.

Hu, M., and M. Xue, 2007: Impact of configurations of rapid intermittent assimilation of WSR-88D radar data for the 8 May 2003 Oklahoma City tornadic thunderstorm case. *Mon. Wea. Rev.*, **135**, 507–525, https://doi.org/10.1175/MWR3313.1.

Huang, H., and Coauthors, 2020: UNet 3+: A full-scale connected UNet for medical image segmentation. *ICASSP 2020—2020 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, Institute of Electrical and Electronics Engineers, 1055–1059, https://doi.org/10.1109/ICASSP40776.2020.9053405.

Huang, W., Y. Jiang, X. Liu, Y. Pan, X. Li, R. Guo, Y. Huang, and B. Duan, 2019: Classified early-warning and nowcasting of hail weather based on radar products and random forest algorithm. *2019 Int. Conf. on Meteorology Observations (ICMO)*, Chengdu, China, Institute of Electrical and Electronics Engineers, 1–3, https://doi.org/10.1109/ICMO49322.2019.9026039.

Jewell, R., and J. Brimelow, 2009: Evaluation of Alberta hail growth model using severe hail proximity soundings from the United States. *Wea. Forecasting*, **24**, 1592–1609, https://doi.org/10.1175/2009WAF2222230.1.

Johnson, A. W., and K. E. Sugden, 2014: Evaluation of sounding-derived thermodynamic and wind-related parameters associated with large hail events. *Electron. J. Severe Storms Meteor.*, **9** (5), https://doi.org/10.55599/ejssm.v9i5.57.

Justin, A. D., C. Willingham, A. McGovern, and J. T. Allen, 2023: Toward operational real-time identification of frontal boundaries using machine learning. *Artif. Intell. Earth Syst.*, **2**, e220052, https://doi.org/10.1175/AIES-D-22-0052.1.

Kumjian, M. R., and K. Lombardo, 2020: A hail growth trajectory model for exploring the environmental controls on hail size: Model physics and idealized tests. *J. Atmos. Sci.*, **77**, 2765–2791, https://doi.org/10.1175/JAS-D-20-0016.1.

Labriola, J., N. Snook, Y. Jung, and M. Xue, 2019: Explicit ensemble prediction of hail in 19 May 2013 Oklahoma City thunderstorms and analysis of hail growth processes with several multimoment microphysics schemes. *Mon. Wea. Rev.*, **147**, 1193–1213, https://doi.org/10.1175/MWR-D-18-0266.1.

Lagerquist, R., A. McGovern, 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.

Leinonen, J., U. Hamann, U. Germann, and J. R. Mecikalski, 2022: 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.

Malečić, B., M. Telišman Prtenjak, K. Horvath, D. Jelić, P. Mikuš Jurković, K. Ćorko, and N. S. Mahović, 2022: Performance of HAILCAST and the Lightning Potential Index in simulating hailstorms in Croatia in a mesoscale model—Sensitivity to the PBL and microphysics parameterization schemes. *Atmos. Res.*, **272**, 106143, https://doi.org/10.1016/j.atmosres.2022.106143.

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

——, R. J. Chase, M. Flora, D. J. Gagne, II, R. Lagerquist, C. K. Potvin, N. Snook, and E. Loken, 2023: A review of machine learning for convective weather. *Artif. Intell. Earth Syst.*, e220077, https://doi.org/10.1175/AIES-D-22-0077.1.

Milbrandt, J. A., and M. K. Yau, 2005: A multimoment bulk microphysics parameterization. Part I: Analysis of the role of the spectral shape parameter. *J. Atmos. Sci.*, **62**, 3051–3064, https://doi.org/10.1175/JAS3534.1.

Mohr, S., and M. Kunz, 2013: Recent trends and variabilities of convective parameters relevant for hail events in Germany and Europe. *Atmos. Res.*, **123**, 211–228, https://doi.org/10.1016/j.atmosres.2012.05.016.

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.

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.

——, ——, and J. T. Allen, 2021: A 23-year severe hail climatology using GridRad MESH observations. *Mon. Wea. Rev.*, **149**, 945–958, https://doi.org/10.1175/MWR-D-20-0178.1.

Murphy, A. M., C. R. Homeyer, and K. Q. Allen, 2023: Development and investigation of GridRad-severe, a multiyear severe event radar dataset. *Mon. Wea. Rev.*, **151**, 2257–2277, https://doi.org/10.1175/MWR-D-23-0017.1.

Murphy, M. J., J. A. Cramer, and R. K. Said, 2021: Recent history of upgrades to the U.S. National Lightning Detection Network. *J. Atmos. Oceanic Technol.*, **38**, 573–585, https://doi.org/10.1175/JTECH-D-19-0215.1.

Nelson, S. P., 1983: The influence of storm flow structure on hail growth. *J. Atmos. Sci.*, **40**, 1965–1983, https://doi.org/10.1175/1520-0469(1983)040<1965:TIOSFS>2.0.CO;2.

NSSL, 2021: Warn-on-Forecast case studies. NOAA National Severe Storms Laboratory, https://www.nssl.noaa.gov/projects/wof/casestudies/hail-oktx-apr2021/.

Ortega, K. L., 2018: Evaluating multi-radar, multi-sensor products for surface hailfall diagnosis. *Electron. J. Severe Storms Meteor.*, **13** (1), https://doi.org/10.55599/ejssm.v13i1.69.

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, https://doi.org/10.1007/978-3-319-24574-4_28.

Scarino, B., K. Itterly, K. Bedka, C. R. Homeyer, J. Allen, S. Bang, and D. Cecil, 2023: Deriving severe hail likelihood from satellite observations and model reanalysis parameters using a deep neural network. *Artif. Intell. Earth Syst.*, **2**, e220042, https://doi.org/10.1175/AIES-D-22-0042.1.

Schmidt, T., 2023: Gridded hail nowcasting using UNets, lightning observations, and the Warn-on-Forecast System. M.S. thesis, Dept. of School of Meteorology, University of Oklahoma, 119 pp., https://shareok.org/handle/11244/338834.

School of Meteorology/University of Oklahoma, 2021: GridRad-severe—Three-dimensional gridded NEXRAD WSR-88D radar data for severe events. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, accessed 5 May 2023, https://rda.ucar.edu/datasets/d841006/.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., https://doi.org/10.5065/D68S4MVH.

Stensrud, D. J., and Coauthors, 2009: Convective-scale warn-on-forecast system: A vision for 2020. *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.

Stone, M., 1974: Cross-validatory choice and assessment of statistical predictions. *J. Roy. Stat. Soc.*, **36B**, 111–133, https://doi.org/10.1111/j.2517-6161.1974.tb00994.x.

Taszarek, M., J. T. Allen, T. Púčik, K. A. Hoogewind, and H. E. Brooks, 2020: Severe convective storms across Europe and the United States. Part II: ERA5 environments associated with lightning, large hail, severe wind, and tornadoes. *J. Climate*, **33**, 10 263–10 286, https://doi.org/10.1175/JCLI-D-20-0346.1.

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

Trapp, R. J., 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.

Tuovinen, J.-P., J. Rauhala, and D. M. Schultz, 2015: Significant-hail-producing storms in Finland: Convective-storm environment and mode. *Wea. Forecasting*, **30**, 1064–1076, https://doi.org/10.1175/WAF-D-14-00159.1.

Wendt, N. A., and I. L. Jirak, 2021: An hourly climatology of operational MRMS MESH-diagnosed severe and significant hail with comparisons to *Storm Data* hail reports. *Wea. Forecasting*, **36**, 645–659, https://doi.org/10.1175/WAF-D-20-0158.1.

Witt, A., M. D. Eilts, G. J. Stumpf, J. T. 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.

Yano, J.-I., and Coauthors, 2018: Scientific challenges of convective-scale numerical weather prediction. *Bull. Amer. Meteor. Soc.*, **99**, 699–710, https://doi.org/10.1175/BAMS-D-17-0125.1.