Development and Interpretation of a Neural-Network-Based Synthetic Radar Reflectivity Estimator Using GOES-R Satellite Observations

KYLE A. HILBURN,[a] IMME EBERT-UPHOFF,[b] AND STEVEN D. MILLER[a]

[a] *Cooperative Institute for Research in the Atmosphere, Fort Collins, Colorado*
[b] *Colorado State University, Fort Collins, Colorado*

(Manuscript received 31 March 2020, in final form 7 October 2020)

ABSTRACT: The objective of this research is to develop techniques for assimilating GOES-R series observations in precipitating scenes for the purpose of improving short-term convective-scale forecasts of high-impact weather hazards. Whereas one approach is radiance assimilation, the information content of GOES-R radiances from its Advanced Baseline Imager saturates in precipitating scenes, and radiance assimilation does not make use of lightning observations from the GOES Lightning Mapper. Here, a convolutional neural network (CNN) is developed to transform GOES-R radiances and lightning into synthetic radar reflectivity fields to make use of existing radar assimilation techniques. We find that the ability of CNNs to utilize spatial context is essential for this application and offers breakthrough improvement in skill compared to traditional pixel-by-pixel based approaches. To understand the improved performance, we use a novel analysis method that combines several techniques, each providing different insights into the network's reasoning. Channel-withholding experiments and spatial information–withholding experiments are used to show that the CNN achieves skill at high reflectivity values from the information content in radiance gradients and the presence of lightning. The attribution method, layerwise relevance propagation, demonstrates that the CNN uses radiance and lightning information synergistically, where lightning helps the CNN focus on which neighboring locations are most important. Synthetic inputs are used to quantify the sensitivity to radiance gradients, showing that sharper gradients produce a stronger response in predicted reflectivity. Lightning observations are found to be uniquely valuable for their ability to pinpoint locations of strong radar echoes.

KEYWORDS: Radars/radar observations; Satellite observations; Data assimilation; Deep learning; Machine learning; Neural networks

1. Introduction

Geostationary Operational Environmental Satellite (GOES) imagery is a key element of U.S. operational weather forecasting, supporting the need for high-resolution, rapidly refreshing imagery for situational awareness (Line et al. 2016). While used extensively by human forecasters, its usage in data assimilation (DA) for numerical weather prediction (NWP) models is limited. Instead DA makes greater usage of microwave and infrared sounder data on low-Earth-orbiting satellites (Lin et al. 2017). Sounders provide more vertically resolved information than imagers, which is advantageous for characterizing the three-dimensional model state, but are carried almost exclusively on low-Earth-orbiting satellites—providing global coverage but at the expense of coarse temporal resolution and latency from sensor to NWP center that can reach 1.5 h or more. Geostationary imagers provide much faster temporal refresh [now 10 min for full disk and 5 min over the contiguous United States (CONUS)] and very low latency over a limited field of regard. Thus, there is an opportunity for operational DA to benefit from the high volume of low-latency, complementary data coming from the global constellation of geostationary imagers.

Operational DA for convective-scale NWP has made steady scientific advances (Gustafsson et al. 2018), but all-sky assimilation of infrared radiances has yet to be operationally demonstrated (Geer et al. 2018). This means that the most dynamic areas from the standpoint of precipitation, having significant impacts on human activities, are also the areas that have the least amount of data to constrain estimates of the current atmospheric state. One approach is radiance assimilation (RA), which has the advantage of being physically based, making it simpler to interpret. Okamoto et al. (2019), Honda et al. (2018a,b), and Sawada et al. (2019) tested assimilation of *Himawari-8* water vapor absorption bands, finding improvements for heavy rain cases. Otkin and Potthast (2019) assimilate a water vapor band on SEVIRI, finding that the all-sky radiance bias correction is critical to making a positive impact on analyses. Demonstration of *GOES-16* Advanced Baseline Imager (ABI) RA was provided by Zhang et al. (2018, 2019), and Jones et al. (2020). These studies make different assumptions about how to inflate observation and background errors and how to weight information in the vertical. Errors in model microphysics and radiative transfer will be inherited by RA, and the land surface will come into play for window channels. Jones et al. (2020) find improved convective initiation forecasts with all-sky RA, but their best results come from using clear-sky radiances and cloud property retrievals. RA cannot be used for assimilating lightning observations, so an observation operator is required to convert observables into control variable increments. Kong et al. (2020) demonstrate improvements from assimilating GOES Lightning Mapper

Denotes content that is immediately available upon publication as open access.

Corresponding author: Kyle Hilburn, kyle.hilburn@colostate.edu

DOI: 10.1175/JAMC-D-20-0084.1

(GLM), using an observation operator that takes advantage of the strong physical relationship between lightning and graupel mass and volume.

A limitation of infrared RA in cloudy and precipitating pixels is saturation of information content. For GOES ABI, the information content of individual pixels saturates around optical depths of 160 and 8 during day and night, respectively, which are the maximum values reported by the retrieval algorithm (Walther et al. 2013). For warm-season convection over CONUS, we find that these values roughly correspond to composite reflectivity (REFC; the vertical maximum radar reflectivity in the column) of 20–25 dBZ for day and 0–5 dBZ for night (Rutledge et al. 2020). This truncated sensitivity means, in turn, that infrared RA holds only limited information about precipitating scenes. This limitation is also present with physically based cloud property retrievals (Jones et al. 2015). The machine learning (ML) technique of convolutional neural networks (CNNs) has the advantage of using the information content present in image gradients, which we will show provides reliable information content up to REFC of about 50 dBZ. Moreover, ML provides an effective framework for using lightning information together with radiance information. So, in this work ML serves as an observation operator for DA, but there are many potential applications of ML to DA, for example, quality control, bias correction, observation thinning, and postprocessing to name a few. This unique ability of CNNs to capture spatial information—together with the large quantity, high quality, high resolution, and low latency of GOES-R data—is justification for exploring the capabilities of ML to enhance DA. Moreover, human forecasters are bombarded with an increasing quantity of information and have limited bandwidth, and exploration of ML methods can help meteorologists to extract maximum value from the firehose of GOES-R observations.

The objective of this research is to ingest GOES-R series observations from the ABI (Schmit et al. 2017) and GLM (Goodman et al. 2013) in precipitating scenes for the purpose of improving short-term convective-scale forecasts of high-impact weather hazards. The Rapid Refresh Forecast System (RRFS) has long used radar reflectivity to estimate latent heating in order to spin up convection in the models. (Benjamin et al. 2016). Using this pathway for GOES information would require producing 3D fields of radar reflectivity. We will treat this problem as vertically separable, first estimating the spatial distribution of REFC, and then estimating the vertical profile in a second step. This paper will tackle the REFC part of the problem, focus on convective-scale applications, and consider warm-season convection over eastern CONUS where radar coverage is best. We describe the development of a CNN for that purpose, including architecture selection and a novel approach to design a loss function to deal with class imbalances of REFC values. Performance is evaluated using metrics including the mean-square error (MSE), coefficient of determination R^2, categorical metrics (probability of detection, false-alarm rate, critical success index, and categorical bias) at various output threshold levels, and evaluation of the root-mean-square difference (RMSD) binned over the range of true output values.

A potential disadvantage of ML is that it is statistically based, making it harder to interpret. So, besides producing a trained and evaluated model, part of the focus of this paper is on developing tools for the interpretation and explanation of the strategies for how CNNs make predictions. This paper is concerned specifically with tools for the GOES-radar translation problem, but a more general review for image-to-image translation problems is provided by Ebert-Uphoff and Hilburn (2020). Using ML to transform satellite data inputs has the potential to introduce errors arising from uncertainties related to the connection between observed cloud-top features and lightning with estimates of latent heating vertical profiles, so it is very important that we understand how the ML makes its predictions and to characterize the errors. McGovern et al. (2019) provides a thorough review of many approaches for understanding ML predictions. However, the focus of that paper is on methods for analyzing networks for image classification tasks, that is, networks that take images as inputs and produce one scalar value as the output. In this study, the network is performing image-to-image translation, taking images as inputs and producing images as outputs. This some techniques in McGovern et al. (2019) are not directly applicable to image-to-image translation problems, and we explored several other methods. For interpreting image-to-image translation CNNs, layerwise relevance propagation (LRP; Montavon et al. 2018; Lapuschkin et al. 2019) was found to provide very useful information (section 3d). This paper uses a novel analysis method combining LRP (section 3d) together with target architecture experiments (section 3b) and synthetic inputs (section 3e) to gain insights on strategies learned by the ML model that produce good skill.

The ML model developed in this paper is envisioned for DA applications, but there are other related research efforts with aviation and nowcasting applications. Veillette et al. (2018) derived a CNN to predict radar vertically integrated liquid (VIL) from satellite data for aviation applications. This is a similar problem to the one tackled by this paper; however, they use a somewhat unconventional architecture where features are extracted from each input variable separately and then combined in fusion layers. An interesting question about that architecture is whether it allows the network to learn to use lightning data to focus its attention on specific IR features as in this work (section 3d). Another major difference with Veillette et al. (2018) is in how they handle the class imbalance issue. Herein, we use a weighted loss function approach, while Veillette et al. (2018) deliberately sample data to create a balanced training dataset with roughly equal portions of zero, low, and high-intensity VIL. Ayzel et al. (2020), Agrawal et al. (2019), and Samsi et al. (2019) trained CNNs with a similar U-Net architecture as in this paper for the problem of nowcasting using radar data. Su et al. (2020) approached the nowcasting problem using a recurrent architecture, which should better capture temporally evolving features than a standard feedforward architecture. There are a number of commercial entities seeking to provide proxy global radar datasets. The Earth Networks company has developed PulseRad, which uses their ground-based lightning detection network to create global proxy radar maps. The ClimaCell company has merged data from several low-Earth-orbiting and geostationary satellites to create a global precipitation layer product. The interpretation methods developed in this paper could be applied to other CNN models for global radar or nowcasting to potentially improve upon the models and make them more explainable.

We will begin with short descriptions of the ''source'' observations from the GOES-R ABI (section 2a) and GLM (section 2b), followed by our ''target'' observations from the Multi-Radar Multi-Sensor (MRMS; section 2c). The approach for constructing the ML training and validation datasets is described in section 2d. The CNN architecture is described in section 2e, and the approach for constructing a weighted loss function is given in section 2f. The resulting CNN prototype has been dubbed GOES Radar Estimation via Machine Learning to Inform NWP (GREMLIN). In section 3a, we begin with an overall characterization of the performance of GREMLIN, finding remarkably good performance, even at higher REFC values. To explain how GREMLIN makes such predictions, in section 3b we selectively disable specific abilities of this model, resulting in a progression of simpler models, and analyze their results. By examining the predictions from various models (withholding certain channels and/or withholding spatial information), many insights can be gleaned. To examine the use of spatial information, we discuss and visualize the effective receptive field of GREMLIN (section 3c). To understand how the network is making its predictions, and in particular how it uses radiance information and lightning together, we apply the LRP attribution method (section 3d). Then, we construct synthetic inputs representing different meteorological scenarios to probe the network's response and gain further insights into the use of spatial information by the network and to characterize its sensitivity (section 3e). Section 4 presents a summary and future work.

2. Data and method

a. ABI

This study is making use of level-L1b radiances from the GOES-R ABI (Schmit et al. 2017) on *GOES-16*. We are taking advantage of the higher spatial resolution (2 km) and faster temporal refresh (5 min over CONUS) relative to the previous generation of GOES imagers. To produce a unified day–night algorithm, we are focusing on just infrared channels, and, for maximum portability and compatibility to legacy observing systems, we are using the ''heritage'' channels:

- channel 7 (3.9-μm, shortwave infrared window),
- channel 9 [6.9-μm, midlevel water vapor (~442 hPa)], and
- channel 13 (10.3-μm, clean longwave infrared window).

The conversion and calibration of observed radiances Rad to brightness temperatures T_B for GOES ABI follows Schmit et al. (2010):

$$T_B = \frac{c_2}{\ln\left(\frac{c_1}{\mathrm{Rad}} + 1\right)} \quad \text{and} \tag{1a}$$

$$T_{B,C} = \frac{T_B - b_1}{b_2}, \tag{1b}$$

where c_1 and c_2 are the wavenumber-dependent coefficients used to compute the monochromatic T_B, and b_1 and b_2 are spectral bandpass correction offset and scale for calculating the calibrated brightness temperature $T_{B,C}$. These coefficients are provided in the GOES L1b NetCDF data files. We note that, during the daytime, use of the optical depth information from the red band (ABI band 2; 0.64 μm) reflectance and the cloud particle size and phase information from ABI band 6 (near infrared; 2.2 μm) reflectance provide additional skill; however, use of these bands is beyond the scope of this paper.

Two angular quantities are especially relevant to the interpretation of ABI imagery: satellite viewing zenith angle and solar zenith angle. This study makes use of *GOES-16* data from 2019, positioned in its operational east position (75.2°W). In this slot, the satellite viewing zenith angle increases from 35° in northern Florida to 60° in North Dakota. Since we are focusing over just CONUS, we can ignore viewing zenith angle dependence because the limb cooling effect (Elmer et al. 2016) is small in the atmospheric window bands we are considering. We will also consider an example of storms over Colorado in 2017 when *GOES-16* was in its initial check-out position (89.5°W), which had satellite viewing zenith angles around 45°, as compared with 50° in the operational East position. ABI band 7 (3.9 μm) has a daytime solar reflective component and we tested adding solar zenith angle as an input, but found it only made small improvements, so we left it out of the version 1 GREMLIN model. In section 3a we consider the skill of the model as a function of the solar zenith angle, which was calculated following NOAA NESDIS (1998).

In a traditional pixel-based retrieval, correcting the effect of parallax [Vicente et al. (2002) and appendix A in Miller et al. (2018)] is essential for matching up satellite data with radar data on these scales. The main uncertainty with parallax correction is estimating the height of the cloud. One can assume a fixed height, such as 10 km, to substantially reduce the error, at least for the deep clouds that are most relevant; or one can use a cloud-top height product, but this can introduce blank spots in the parallax corrected imagery when low and high clouds are next to each other. To remove parallax offsets to first order, we assumed a height of 10 km. Besides residual parallax errors, there are other reasons for spatial displacements, namely, vertical wind shear, and the CNN seems to learn to apply additional spatial displacements on its own based on what it sees in the training data.

b. GLM

The other major advancement provided by the GOES-R series is real-time lightning observation from the GLM (Goodman et al. 2010, 2013). Lightning is incredibly useful in constructing synthetic radar fields because of its association with the locations of strong updrafts within an embedded convective complex. The physical basis for this association is the strong spatial relationship between lightning flash rates, updraft vertical velocity W, and latent heat release. If the terminal velocity of a raindrop goes as the square root of the

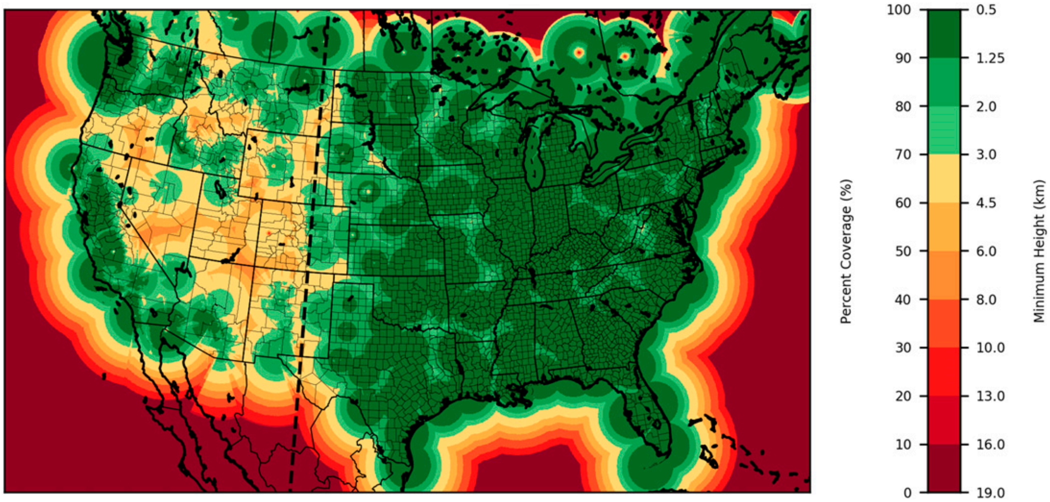

FIG. 1. MRMS radar coverage in terms of the percent of MRMS levels available at each location and the minimum height at that location; 105°W is indicated by the dashed black line.

diameter, then it can be shown that mass (and latent heating) goes as W^6 and (linear) radar reflectivity factor goes as W^{12}. Meanwhile, using simple electrostatic arguments (Price and Rind 1992, Boccippio 2002), one can derive that lightning flash rate goes as W^5 for continental thunderstorms.

Much of the research on using GLM for severe weather has focused on the temporal variability, in particular lightning jumps (Schultz et al. 2009, 2015). However, temporal variability of optically sensed lightning can provide misleading signals. This seems to be due to time-varying detection efficiency effects related to the production of cloud ice (Rutledge et al. 2020), and also possibly to the unsteady nature of updrafts. Instead spatial variability contains more reliable information content, supplementing missing information at very high optical depths, and is especially useful at night. While there is spatial variability in GLM detection efficiency (Marchand et al. 2019), our CNN is more sensitive to the presence of lightning rather than the magnitude of lightning activity, which makes it less sensitive to GLM detection efficiency issues.

GLM maps total lightning with a spatial resolution of 8 km at nadir to 14 km at the limb. The basic unit of data, called an ''event,'' is a gridded quantity, integrating all lightning pulses within the grid box over a 2-ms time window. The Lightning Cluster and Filter Algorithm (LCFA) combines adjacent lightning pixels into ''groups,'' which are then clustered into ''flashes'' using a 330-ms temporal window and a 16.5-km spatial window. Thus, groups and flashes are represented as point observations consisting of a latitude, longitude, time, and area. The LCFA also performs filtering to reduce false alarms. Examination of a few sample storms found the best results (in terms of correlation with REFC) occur when using GLM groups, because they provide more ''filled in'' maps than using flashes. For this work we create group-extent density maps using the group area, assuming it is circular, and accumulating data over 15-min intervals. We tested 5-min accumulation periods but found that this finer temporal granularity produced stratiform areas that flicker on and off from frame to frame. The lighting data units are given as groups per 5 min per kilometer squared.

c. MRMS dataset

The target dataset to which we are training is the quality-controlled composite reflectivity from the MRMS product (Smith et al. 2016). The vertical coverage of MRMS as a function of location is given in Fig. 1, which was created using the 3D reflectivity MRMS fields. Our region of interest for this study is CONUS, east of the Rocky Mountains, over which radar beam blockage issues are minimal. As the radar beam propagates away from the transmitter it is progressively higher above Earth's surface due to both the curvature of Earth and the nonzero elevation angle of the beam itself [minimum of 0.5° for the operational Next-Generation Radar (NEXRAD)]. A comparison of REFC for Hurricane Dorian off the Florida coast with GOES observations indicated that when the vertical coverage falls below 70%, implying that only echoes above 3 km can be measured, the estimate of REFC becomes questionable. When only 50% of the vertical levels are present, this implies that only echoes above 6 km can be measured, and it appears that REFC provides very little reliable information. Over the Great Plains, where dewpoint depressions are large and cloud bases are higher than in the tropical environments of hurricanes, the reliability of REFC might fall off with distance more slowly. To use the best-quality radar data, we are restricting our domain of interest to east of 105°W, for which nearly all locations have 70% coverage and most areas (by virtue of their population) have 90% coverage, or a minimum height of 1.25 km.

d. Dataset construction

The first step in constructing a dataset for training ML is to resample all the inputs and outputs to a common grid. Since the goal of this work is to use the results for data assimilation, we have chosen the 3-km HRRR mass grid as the target grid. The

TABLE 1. Projection and grid parameters for each dataset.

GOES		MRMS		HRRR	
Parameter	Value	Parameter	Value	Parameter	Value
Projection	Geostationary	Projection	Cylindrical	Projection	Lambert conformal conic
Alt	35 786 023.0 m	Lower-left lon	−130°E	Reference lon	262.5°E
Equatorial radius	6 378 137.0 m	Lower-left lat	20°N	Reference lat	38.5°N
Polar radius	6 356 752.314 14 m	Lon scale	0.01	Std parallel	38.5°N
Center lon	−75.0°E	Lon dimension	7000	X scale	3.0 km
X scale	5.6×10^{-5}	Lat scale	0.01	X dimension	1799
X offset	−0.101 332	Lat dimension	3500	Y scale	3.0 km
X dimension	2500			Y dimension	1059
Y scale	-5.6×10^{-5}			Earth radius	6370 km
Y offset	0.128 212				
Y dimension	1500				

projection and grid parameters are provided in Table 1, the formulas used for constructing the Lambert conformal conic and cylindrical grids are given by Snyder (1987), and the formulas for the geostationary projection are provided by Harris Corporation (2016). The MRMS grid is nominally 0.01° or roughly 1 km, and the GOES grid for the infrared bands used in this study is 2 km, so resampling to 3 km has minimal distortion. GOES and MRMS pixels were averaged into their corresponding HRRR grid cell. We note that, because of averaging, after resampling MRMS to 3-km REFC the occurrence of values above 60 dBZ is very rare. The second step in preparing the data for training a CNN is to scale the inputs and outputs to the range 0–1. The scaling parameters for each variable are given in Table 2 and were based on histograms of the variables. We found that the training results were not very sensitive to the exact values of the scaling parameters; however, the channel importance coming from LRP (section 3d) was sensitive.

To reduce data volume and have the CNN focus on scenes of interest, Storm Prediction Center (SPC) filtered storm reports are used to automatically define regions and times of interest in order to maximize the number of storm reports (tornado, hail, and wind). We selected samples from the 92-day period from 17 April 2019 to 17 July 2019 during which there was abundant severe weather. The samples consisted of 256 × 256-pixel images on 3-km HRRR grid (768 × 768 km) and 6-h periods with 15-min refresh. A histogram of the number of storm reports per day has a mode between 20 and 50 reports per case. Each case represents a 6-h period on each day, which may span 0000 UTC. Figure 2a shows that this construction approach results in a geographic preference for the upland South and southern Great Plains. Figure 2b shows a temporal preference for mid- to late-afternoon into the early evening. We split the data using a chronological 80%–20% split for training–validation. Based on this split, the July cases were used for validation and April–June was used for training. We have a total of 1798 samples for training and 448 samples for validation. In this paper we are restricting the focus to warm-season convection to benchmark ML performance for this particular phenomenon and identify the strategies learned by ML. Testing on wintertime precipitation, it is clear that extending the model to synoptic-scale systems will require a deeper model with a larger receptive field (section 3c). Future work training on a larger dataset will use the results of this paper to ensure that the model can be extended without losing performance for warm-season convection.

e. *Selection of CNN architecture*

This particular ML problem takes images as inputs and returns images as outputs, making this an image-to-image translation problem. The U-Net architecture is ideally suited (Ronneberger et al. 2015) to this problem type, and Fig. 3 shows the model architecture we used. The model is drawn with optional skip connections, which concatenate information from the encoder side to reduce the loss of high-resolution spatial information. However, for the results we will present we turned those connections off because they only provided small improvements and complicate the visualization (section 3d). For this particular application, the GOES data provide mostly cloud-top information, while the radar provides information from deeper inside the cloud, thus the high-resolution spatial information that skip connections provide is not necessarily helpful.

The CNN depicted in Fig. 3 has three encoding and three decoding blocks. The encoding portion maps the inputs

TABLE 2. Scaling parameters for each variable. Each scaling is linear, and inverted scaling maps maximum values to 0 and minimum values to 1.

Channel	Min	Max	Inverted
C07	200 K	300 K	True
C09	200 K	250 K	True
C13	200 K	300 K	True
GLM	0.1 groups $(5\,\mathrm{min})^{-1}\,\mathrm{km}^{-2}$	50 groups $(5\,\mathrm{min})^{-1}\,\mathrm{km}^{-2}$	False
MRMS	0 dBZ	60 dBZ	False

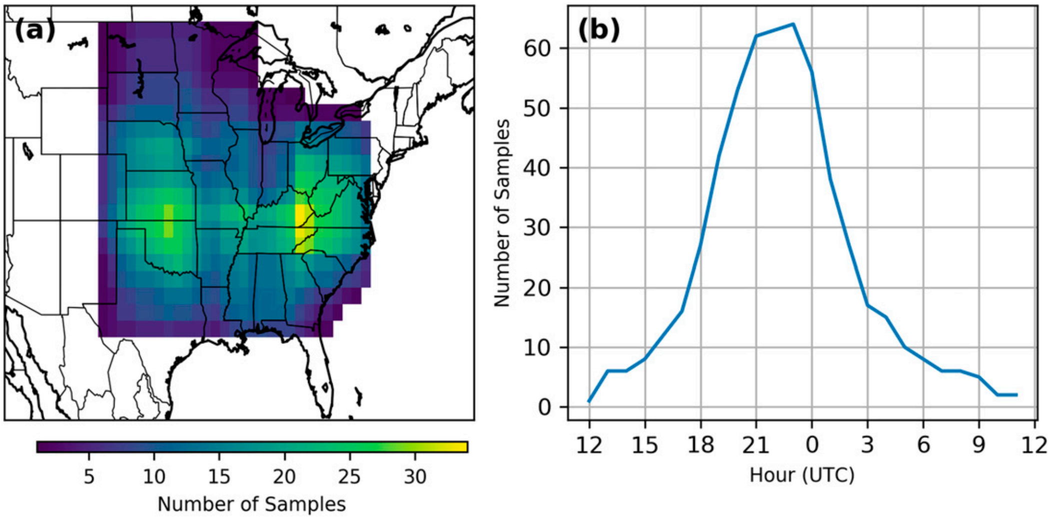

FIG. 2. (a) Spatial and (b) temporal distribution of samples.

(images) to a feature space, and the decoder maps the representation in feature space back into images. Each of the three encoding blocks consists of a convolution layer followed by a pooling layer. A pooling layer reduces resolution and allows the subsequent layers to detect patterns of larger spatial extent. Each decoder block consists of a convolution layer followed by an upsampling layer. Upsampling layers can be thought of as the (imperfect) inverse of a pooling layer, namely, increasing resolution and using interpolation to generate an approximation. The convolutional filters are 3 × 3 kernels that the network learns during training. While U-Nets often double the number of filters per convolution layer going down the encoding branch, and likewise halve the filters going up the decoding branch, we found this produced very small improvements. Instead, we used a constant number of filters, namely, 32 filters per convolution layer. Using more than 32

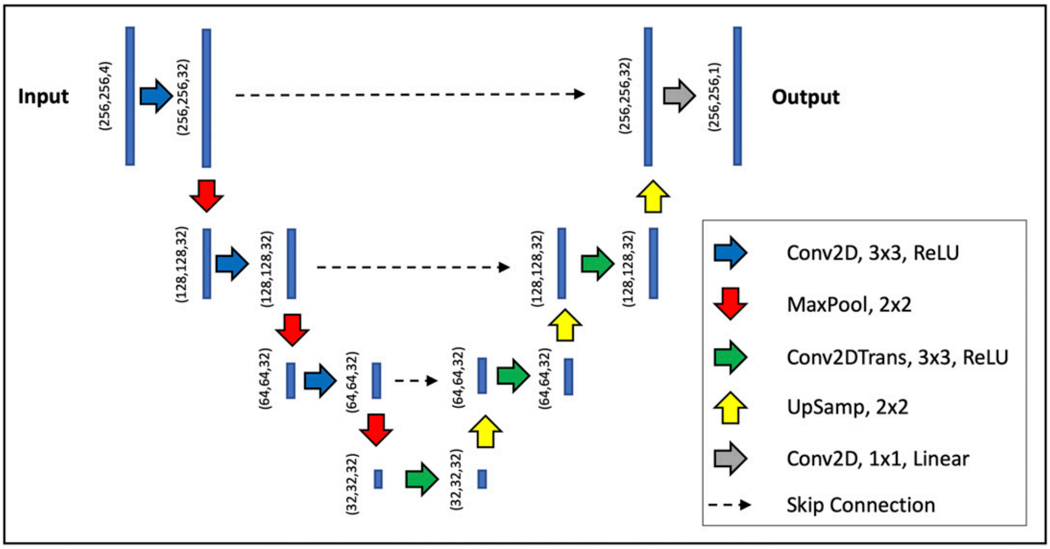

FIG. 3. U-Net architecture for a model with 47 457 trainable parameters. The images are 256 × 256 pixels with four input channels (ABI C07, C09, C13, and GLM group extent density) and one output channel (MRMS composite reflectivity). The convolutional layers (blue and green arrows) each have 32 filters of size 3 × 3 and use a rectified linear unit activation function. The final convolutional layer (gray arrow) combines results from all filters into one output channel using one 1 × 1 filter and linear activation. The encoding branch (left side) uses 2 × 2 maximum pooling to reduce the image dimensions, and the decoding branch (right side) uses 2 × 2 upsampling to increase the image dimensions. The skip connections (dashed black arrows), which concatenate channels across the network, are turned off.

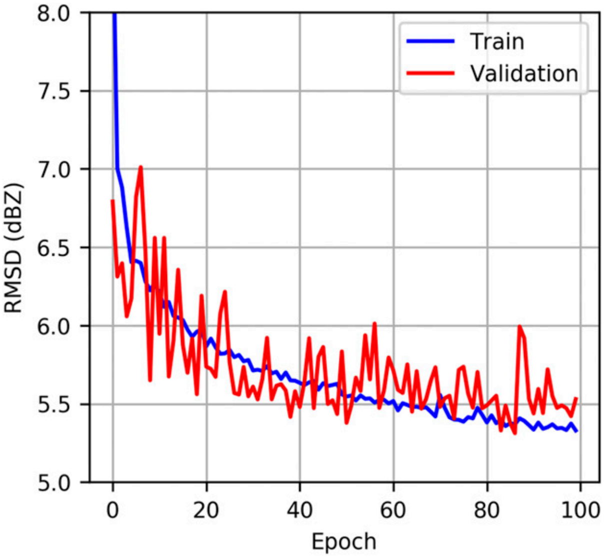

FIG. 4. Training history for GREMLIN in terms of the RMSD with MRMS.

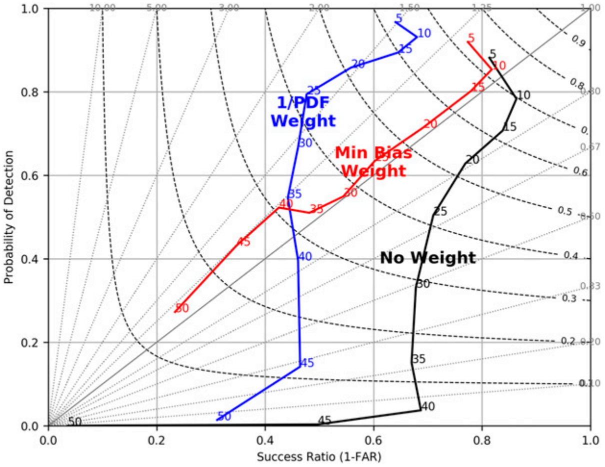

FIG. 5. Performance diagram for REFC categories 5, 10, . . . , 50 dBZ. Dashed black contours are critical success index, and gray dotted lines are categorical bias. The solid black line is performance using an unweighted MSE loss function, the solid blue line uses a 1/PDF weighted MSE loss function, and the solid red line uses weights that produce the minimum categorical bias (GREMLIN).

filters per layer was unnecessary and leaves many filters inactivated. Using fewer filters per layer, such as 16, gave similar overall statistics as 32, but the outputs were noticeably blurrier. The final layer of the network is a convolutional layer that does a pixelwise (1 × 1 filter) linear combination of the 32 filters into one output field. We note that the combination we use of an upsampling layer with nearest-neighbor interpolation followed by a 1 × 1 convolution produces identical pixel values in 2 × 2 blocks in the output field. As a small future improvement, we will include additional 3 × 3 convolutional layers to obtain an interpolated result within the 2 × 2 blocks.

As noted above, there are three encoder and three corresponding decoder layers. Based on an analysis of training and validation losses, we found that going deeper resulted in overfitting. Note, also, our choice of using only one convolution layer per encoder/decoder block, while U-Nets often use two convolution layers per block. Using two convolution layers per block doubles the number of trainable parameters, also making the chance of overfitting more likely. We are concerned with warm-season convection, a phenomenon that is inherently small scale (e.g., mesoγ to the smaller end of mesoα), and a network of this depth and architecture performs well. However, for larger spatial phenomena, such as hurricanes and synoptic-scale frontal precipitation, a deeper network will be required. In such cases, more samples would be needed for training. When additional real samples are unavailable, data augmentation is the next best approach. As a side note, we found that we could obtain results that were similar to those shown in this paper with a training dataset of 1/10 original sample by doing 10× augmentation, done by adding random noise to the real samples. However, the results shown herein used no data augmentation.

The model was trained on a single NVIDIA Tesla P100 GPU for 100 epochs, which took 15 min of wall-clock time. Using a batch size of 18, the model had a memory footprint of 0.5 gigabytes, and the data required 8 gigabytes of memory. The final model stored in HDF5 is only 625 kilobytes. The training history is shown in Fig. 4. Training beyond 100 epochs we observed the validation loss flattened while the training loss continued decreasing, indicating that further training would produce overfitting. The loss function is described in section 2f. The final version-1 GREMLIN model has validation statistics against MRMS observations of RMSD = 5.53 dBZ and $R^2 = 0.740$.

f. *Design of loss function to address REFC class imbalance*

In ML, the loss (or cost) function quantifies the difference between the model predicted values and the actual true value. The process of training a model involves changing the neural network's (NN's) weights to minimize the loss function. An important consideration in training the NN is the choice of loss function since radar reflectivity fields suffer from a class imbalance issue with an exponentially decreasing distribution for high values. In this section we discuss a new way to design a loss function to balance good performance for the rare (but important) high values with good performance for small values.

Training the NN using the standard unweighted pixelwise MSE loss function results in suboptimal performance at high REFC (Fig. 5). High radar reflectivity values are relatively less common: if y represents the scaled radar reflectivity (scaling 0–60 dBZ linearly into the range 0–1), then the probability density function is closely approximated by $P(y) \propto e^{-5y}$ with an $R^2 = 0.80$. We use a performance diagram (Fig. 5) to select loss function weights that produce the minimum categorical bias. Categorical statistics and contingency tables are discussed in Wilks (2006) and performance diagrams are discussed in Roebber (2009). The binary categories are created by evaluating whether the true and predicted REFC are greater than a threshold. While minimizing the categorical bias does not guarantee that the results will also have maximal critical success index, we found that in practice this was the case.

Our approach is related to using an area under the receiver operating characteristic curve as a loss function but avoids the problem of derivatives not existing for a discontinuous function. The approach also acts as a global constraint on the realism of the resulting fields by balancing overprediction and underprediction of reflectivity at all values. We define weights W for the MSE loss function L according to a generalized exponential:

$$L(y_{\text{true}}, y_{\text{pred}}) = \frac{1}{N}\sum_{j=1}^{N} W(y_{\text{true}})(y_{\text{pred}} - y_{\text{true}})^2 \quad \text{and} \tag{2a}$$

$$W(y_{\text{true}}) = e^{b y_{\text{true}}^{c}}\,, \tag{2b}$$

where y_{true} and y_{pred} are the true and predicted values of y and N is the number of training samples. We then vary b and c in a grid search, training a NN model for each combination, to find the optimal model producing the minimum categorical bias. Values of the categorical bias are calculated at each REFC threshold i from 5 to 50 dBZ in steps of 5 dBZ, and the best-matching model is found by taking the parameter combination k with

$$\min_{k}\left[\operatorname*{mean}_{i}(|1 - \text{bias}_{i,k}|)\right]. \tag{3}$$

To get reliable results, we also train several versions of the model (20 versions) that differ only in their random seeds and then select the model minimizing Eq. (3). During training we observed that errors for low REFC values settled down first, as evidenced by a categorical bias near 1, and that errors for high REFC values settled down last. We used only the training samples to perform model selection, to keep the validation samples independent. While the intuitive 1/PDF weights would give $b = 5$ and $c = 1$, we found that the minimum categorical bias weights were $b = 5$ and $c = 4$ for the MSE loss and $b = 5$ and $c = 3$ for the mean-absolute-error (MAE) loss. The disparity suggests there might be a way to choose coefficients from first principles based on the PDF, but we note that the best results require a much heavier weighting of the high values than would be implied by direct usage of the inverse of the PDF.

3. Results and discussion

a. Baseline network performance

The overall performance of our final neural network, GREMLIN, is shown as the red line in Fig. 5. To understand the abilities of GREMLIN to produce synthetic radar reflectivity, it is helpful to consider a specific example. Figure 6 compares MRMS REFC (Figs. 6a,c,e) with GREMLIN REFC (Figs. 6b,d,f) at three times during the event (2100, 2300, 0100 LT), noting that the first large hail reports were at 2050 LT and lasted until 2130 LT. This case is notable because of its severe impact on the Denver, Colorado, metropolitan area; the storms produced up to baseball-sized hail [2.75 in. (~7 cm)] and was the costliest weather catastrophe in Colorado—producing $1.4 billion in insured losses (Svaldi 2017). In addition to its human impact, this case poses challenges for both infrared imagers and optically sensed lightning. It is an example of Great Plains thunderstorms with abundant cloud water concentrations (e.g., Williams et al. 2005) that produce large anvils that obscure the convective cores in infrared imagery. While these conditions also lead to very high lightning rates, Rutledge et al. (2020) show these conditions also produce storms for which the lighting flash height is relatively low, making for large optical paths between the lightning source and the upper cloud boundary along the GLM sensor line of sight (both in general and for this particular case). This regionally common ''inverted'' charge structure causes a relative minimum in lightning detection efficiency over the Great Plains (Marchand et al. 2019; Fuchs et al. 2018).

Despite the challenges, Fig. 6 shows that GREMLIN performs well for this case. In the early stages (Figs. 6a,b) GREMLIN captures the three distinct convective cores near Denver, Greeley, and Fort Morgan, Colorado. It correctly represented the location of the strongest echoes, although it also tended to overestimate them, and the finescale structure of the cores is not captured. Two hours later (Figs. 6c,d) as the storms began to transition to a convective line morphology, the GREMLIN estimates captured that transition well. GREMLIN properly located the strong echoes, although small areas that were distinct in MRMS tended to get merged in GREMLIN. After dark (Figs. 6e,f) and as the convection transitioned from distinct cells to lines, GREMLIN captured the basic shape and curvature of the lines but tended to merge lines that were separate in MRMS.

Characterizing the spatiotemporal performance of the technique is complicated by the natural variation of convective morphology. In our training dataset, convection tends to be more widespread in the eastern United States, while isolated convective cells are more common in the west. Since RMSD statistics are sensitive to the echo coverage fraction F, care must be taken to separate true regional biases from artificial biases that arise from natural regional variations in these properties. Figure 7a shows the RMSD versus the echo coverage fraction F, defined here by the 20-dBZ radar reflectivity contour. It can be seen that more scattered precipitation with smaller F can be more accurately estimated with smaller RMSD. It also shows that eastern U.S. regions tend to have both larger F and larger RMSD. However, the easternmost locations do have errors greater than the average (black line given by RMSD = $2.2F^{0.36}$). Given that our training samples have a fairly uniform distribution from east to west (Fig. 2a), the fact that the predictions exhibit an ''Oklahoma centric'' bias is notable and may be a consequence of using a loss function that is heavily weighted toward higher REFC values.

The typical life cycle is for convection to initiate with the heating of the daytime, and then grow upscale overnight. One might expect the large echo structures at night to validate better since the GREMLIN estimates tend to be more smoothed out than MRMS REFC. To look for biases in time, Fig. 7b gives the RMSD versus F as a function of the solar zenith angle, where sunset is 90°. It does show a population of samples that have both large F and small RMSD; however,

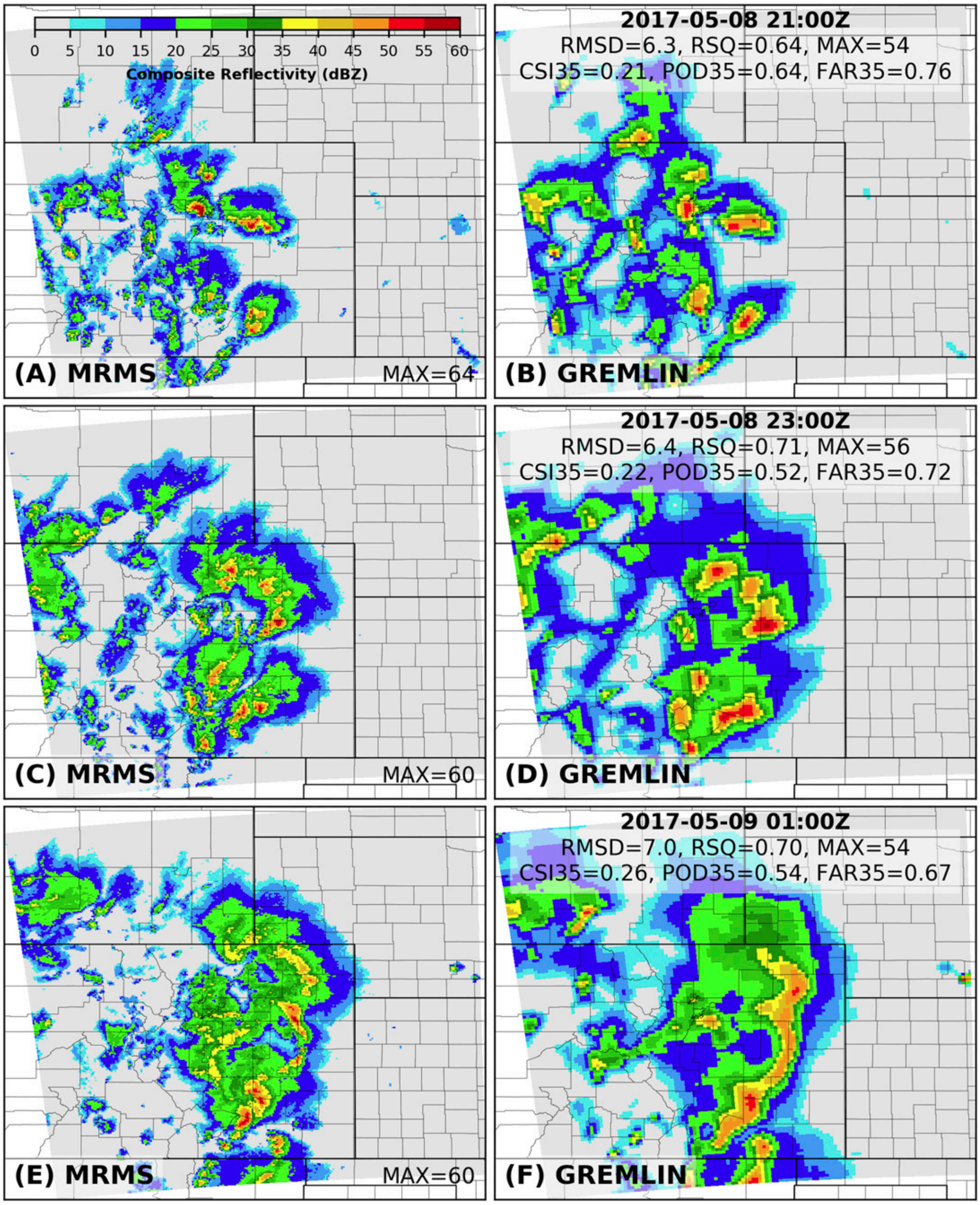
0 5 10 15 20 25 30 35 40 45 50 55 60
Composite Reflectivity (dBZ)
(A) MRMS
MAX=64
2017-05-08 21:00Z
RMSD=6.3, RSQ=0.64, MAX=54
CSI35=0.21, POD35=0.64, FAR35=0.76
(B) GREMLIN
(C) MRMS
MAX=60
2017-05-08 23:00Z
RMSD=6.4, RSQ=0.71, MAX=56
CSI35=0.22, POD35=0.52, FAR35=0.72
(D) GREMLIN
(E) MRMS
MAX=60
2017-05-09 01:00Z
RMSD=7.0, RSQ=0.70, MAX=54
CSI35=0.26, POD35=0.54, FAR35=0.67
(F) GREMLIN

FIG. 6. Colorado 8 May 2017 case for (a),(c),(e) MRMS and (b),(d),(f) GREMLIN prediction. Statistics are provided for RMSD, coefficient of determination (RSQ), maximum REFC value (MAX), critical success index at 35 dBZ (CSI35), probability of detection at 35 dBZ (POD35), and false-alarm rate at 35 dBZ (FAR35).

most nighttime samples are below the average line, even at smaller F. This good performance at night is notable given that our training samples emphasize late afternoon and early evening (Fig. 2b). It is possible this is a result of GLM having a 20% higher detection efficiency at night than during the day (Marchand et al. 2019). Not all daytime retrievals have lower skill, and the day/night distinction in skill is less clean than the east/west distinction. However, since daytime retrievals do have room for improvement, this argues that the solar reflective bands, visible and cloud particle phase/size bands in particular, should be used. Overall, GREMLIN performs well. In particular, GREMLIN is able to accurately locate areas of strong echoes, which have been difficult to capture with heritage methods (e.g., Arkin and Meisner 1987).

b. Targeted architecture experiments

A key question raised by the results shown in section 3a is, “What is the network learning to produce such good skill?” We

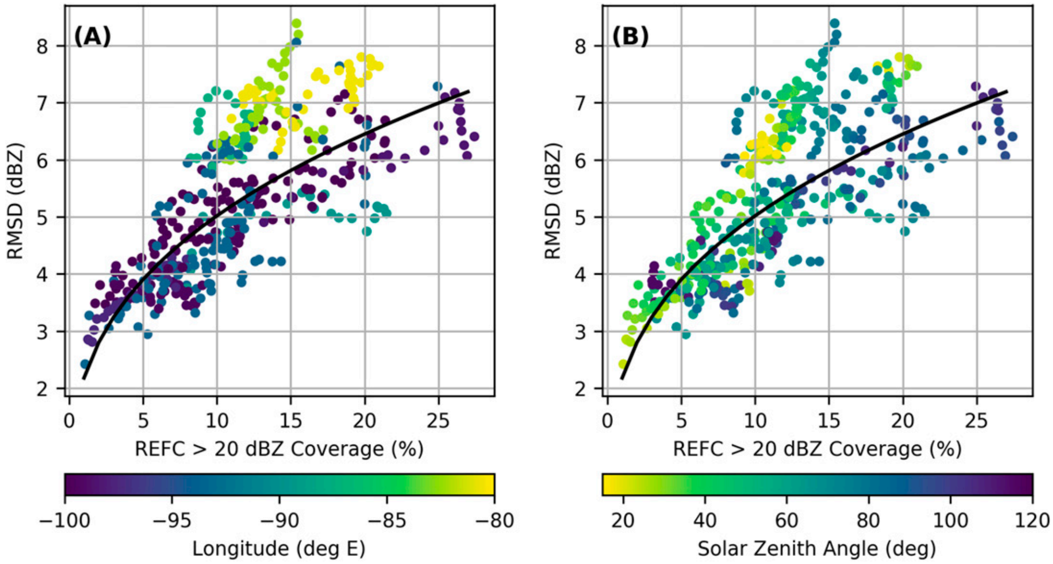

FIG. 7. RMSD vs the percentage coverage of radar echoes > 20 dBZ where the color indicates the mean (a) longitude and (b) solar zenith angle for each sample.

use several different methods to answer this question, starting with targeted architecture experiments. Namely, we modify the GREMLIN architecture by removing specific capabilities. Analyzing the performance of the resulting restricted NNs tells us which capabilities of GREMLIN are most essential for its success and sheds light on how they are used.

We begin by removing the capability of GREMLIN to utilize information from radiance gradients and spatial context used by the network—done by replacing all 3 × 3 filters by 1 × 1 filters. Second, we trained models while withholding sets of channels. Figure 8 provides results for a representative validation sample. For simplicity, we focus on the impact of gradients in channel 13, which is the most important channel (section 3d), and of lightning information. The C13 T_Bs (Fig. 8a) exhibit very sharp spatial gradients from clear areas with $T_B > 275$ K to areas with radar echo with $T_B \sim 220$ K. In comparing with the spatial pattern of REFC (Fig. 8c) it can be seen that cold T_Bs are generally a good predictor that a particular pixel has REFC > 15 dBZ, but there is a low spatial correlation between the coldest $T_B < 215$ K and the higher REFC values > 35 dBZ. These areas of strong echoes correlate well with lightning (Fig. 8b), although the lightning is a bit smoother than REFC and there are spatial displacements. The latter may be due to a combination of residual parallax displacement errors and the effects of vertical wind shear.

Figures 8d–i show the progression of results for six NN models with increasing capabilities, from the most restricted model (Fig. 8d) to the full model, GREMLIN (Fig. 8i). The 1 × 1 filter experiments are shown in the middle row (Figs. 8d–f), which represents the performance that could be expected from a traditional pixel-based retrieval. With C13 alone (Fig. 8d), the areas of REFC > 15 dBZ are reasonably well delineated, but it completely lacks any echoes > 35 dBZ. Combining GLM with C13 (Fig. 8e) shows huge improvements in the representation of echoes > 35 dBZ, although the spatial extent is a bit too large. Bringing in the other two channels (C07 and C09) in Fig. 8f does help reduce the errors a bit. So, without the use of spatial gradient information, lightning information is critical to obtaining any skill for higher REFC values.

Figures 8g–i show the results using 3 × 3 filters. Even with C13 alone, the use of gradient information and spatial context (Fig. 8g), produces marked improvements in skill, especially at the high REFC end. When compared with the 1 × 1 experiment (Fig. 8d) the probability of detection (POD) of 35-dBZ reflectivity jumps from 0 to 0.24, and the false-alarm rate (FAR) of 0.56 is slightly better than using all channels with no spatial information (Fig. 8f). Adding lightning information (Fig. 8h) more than doubles the POD and also reduces the FAR. Adding the other channels (Fig. 8i) helps as well, producing significant improvements in RMSD and R^2, also resulting in higher POD and lower FAR. We hypothesize that results of this quality (Fig. 8i) are sufficiently good to produce a positive impact on data assimilation.

The results for this example are consistent with those across all validation samples (Fig. 9 and Table 3). Without the benefit of spatial information and lightning (black and green lines in Fig. 9a), the RMSD at high REFC is as large as 25 dBZ. Note that removing spatial context but adding lightning (blue line Fig. 9a) makes the RMSD slightly worse for REFC in the range 20–35 dBZ but produces large improvements above 35 dBZ, bringing the RMSD down to 15 dBZ. Adding spatial context yields additional large improvements (Fig. 9d). Combining spatial information and lightning produces the best results, with RMSD of 12 dBZ at the highest REFC. Without spatial information, lightning shows obvious value in increasing the POD (Fig. 9b) and reducing the FAR (Fig. 9c). In the absence of lightning information, adding the water vapor channel (green line in Figs. 9b,c) does provide some improvements in POD and FAR, but not as much as lightning. Based on

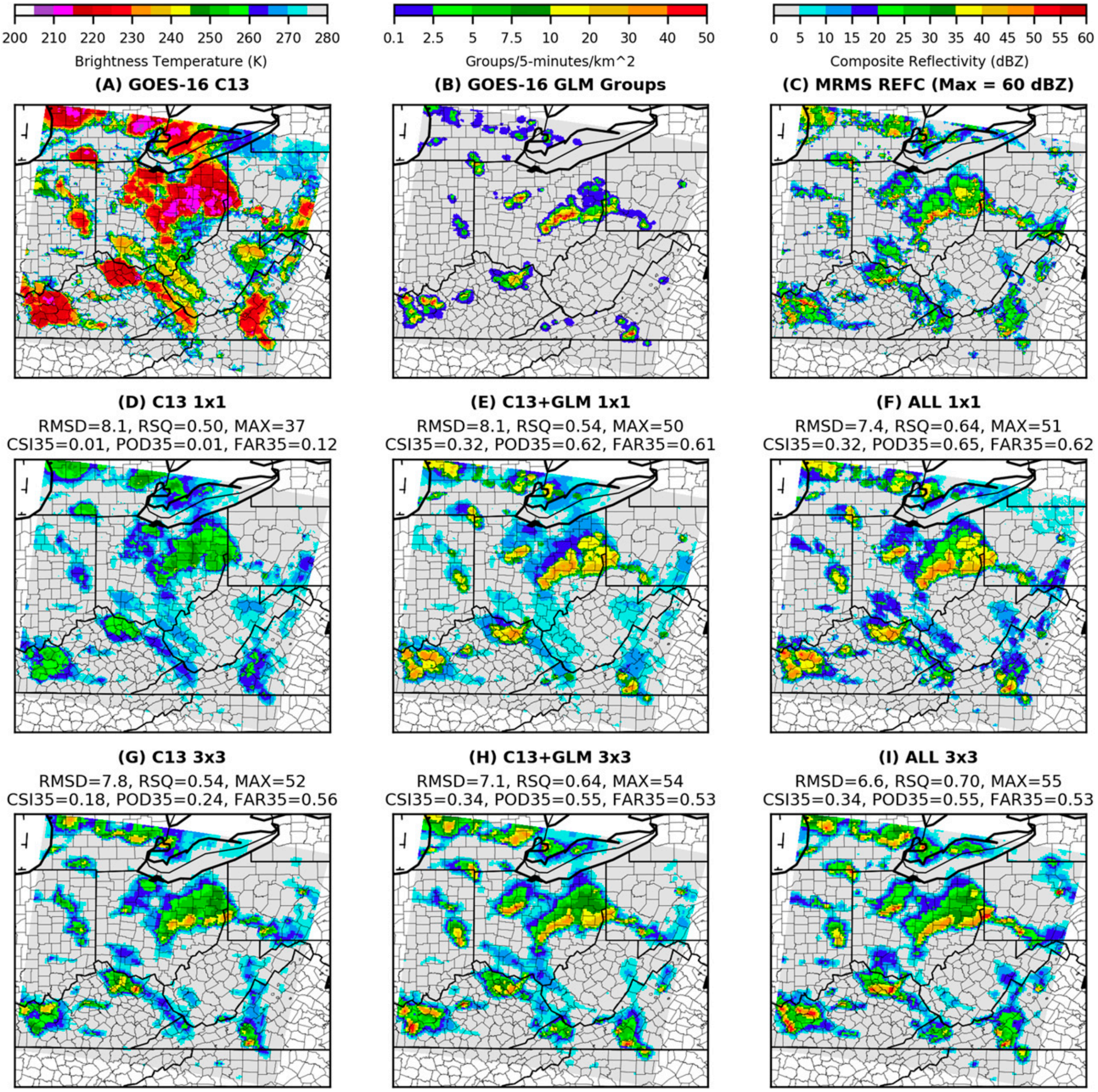
200 210 220 230 240 250 260 270 280
Brightness Temperature (K)
(A) GOES-16 C13
0.1 2.5 5 7.5 10 20 30 40 50
Groups/5-minutes/km^2
(B) GOES-16 GLM Groups
0 5 10 15 20 25 30 35 40 45 50 55 60
Composite Reflectivity (dBZ)
(C) MRMS REFC (Max = 60 dBZ)
(D) C13 1x1
RMSD=8.1, RSQ=0.50, MAX=37
CSI35=0.01, POD35=0.01, FAR35=0.12
(E) C13+GLM 1x1
RMSD=8.1, RSQ=0.54, MAX=50
CSI35=0.32, POD35=0.62, FAR35=0.61
(F) ALL 1x1
RMSD=7.4, RSQ=0.64, MAX=51
CSI35=0.32, POD35=0.65, FAR35=0.62
(G) C13 3x3
RMSD=7.8, RSQ=0.54, MAX=52
CSI35=0.18, POD35=0.24, FAR35=0.56
(H) C13+GLM 3x3
RMSD=7.1, RSQ=0.64, MAX=54
CSI35=0.34, POD35=0.55, FAR35=0.53
(I) ALL 3x3
RMSD=6.6, RSQ=0.70, MAX=55
CSI35=0.34, POD35=0.55, FAR35=0.53

FIG. 8. Validation sample inputs for 2330 UTC 2 Jul 2019: (a) GOES C13 and (b) GOES GLM; truth: (c) MRMS; and prediction for progression of six models with increasing capabilities: (d) 1 × 1 filters for C13 only, (e) 1 × 1 filters for C13 + GLM, (f) 1 × 1 filters for all channels, (g) 3 × 3 filters for C13 only, (h) 3 × 3 filters for C13 + GLM, and (i) 3 × 3 filters for all channels (GREMLIN). Panels (d)–(i) provide the following statistics: RMSD (dBZ), RSQ, MAX (dBZ), CSI35, POD35, and FAR35.

examining predictions, it appears the network correlates smaller differences between C13 and C09 with higher REFC. However, those areas of small C13 − C09 difference tend to be more spatially extensive than REFC, with the result being that POD is improved, but FAR is slightly worse. This finding demonstrates the unique benefits of lightning information to pinpoint the areas of strong updrafts and high REFC. When spatial information is used, the value of lightning is relatively less, but it still makes significant improvements in POD (Fig. 9e) and FAR (Fig. 9f). Further insights into how the network is using lightning and spatial information together is provided by use of attribution methods (section 3d).

GREMLIN predictions can be seen to have overly broad convective cores (e.g., Fig. 8), or, when using a continuous color scale, predictions look blurrier than real radar images. This is an intrinsic aspect of CNNs related to the perception–distortion trade-off (Blau and Michaeli 2018) for image-generating methods, which is a trade-off between producing images that look sharp but are less accurate (better perception) versus images that look blurry but are more accurate (less

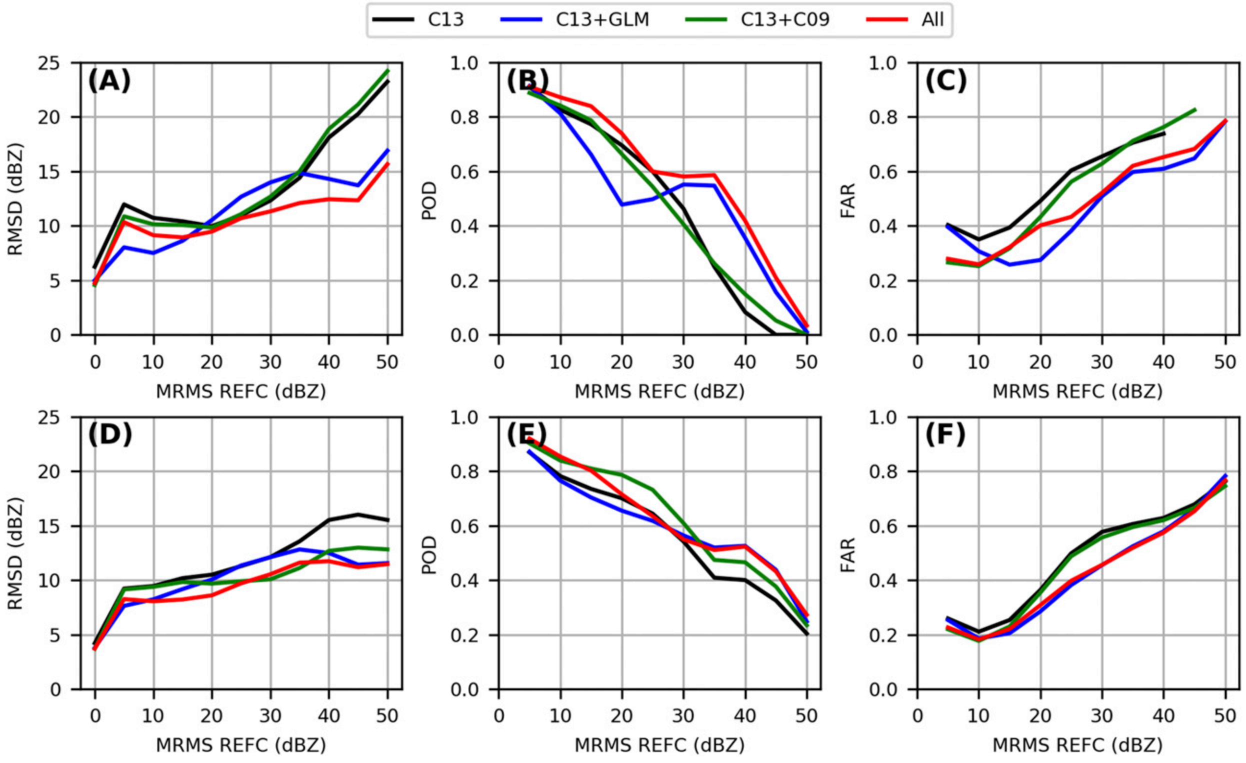

FIG. 9. Statistics for (top) 1 × 1 filters and (bottom) 3 × 3 filters for (a),(d) RMSD; (b),(e) POD; and (c),(f) FAR vs REFC for various experiments (line colors).

distortion). CNNs specialize in maximizing accuracy, for example, minimizing the mean-squared error, but results are blurry. CNN outputs are somewhat analogous to an ensemble mean field—it might be the best answer in a statistical sense, but it may not look physically realistic. In contrast, a different type of neural networks, generative adversarial networks (GANs), can produce results that are less statistically accurate, but could more closely resemble actual radar fields. GAN outputs are somewhat analogous to producing a *single* ensemble member. Stengel et al. (2020) apply GANs applied to a wind and solar data superresolution application and discuss the trade-off in detail. In summary, increased uncertainty results in increased blurriness in CNN-generated images, while resulting in a larger potential spread between different GAN-generated images. Our interpretation of the broad convective cores generated by our CNN are thus that they are a result of uncertainty yielding blurry outputs, as outlined above. Specifically, our hypothesis is that the overly broad cores provide an indication of positional uncertainty translating cloud-top features into features deep inside the cloud. In our future work we plan to try out GANs and compare results with CNNs in terms of accuracy versus blurriness.

c. *Examining the effective receptive field*

GREMLIN is a *purely* convolutional neural network, that is, it does not have any fully connected (aka dense) layers. This means that any individual output neuron, that is, any pixel of the estimated MRMS image, is connected to only a small group of input neurons corresponding to a small spatial neighborhood of the output pixel in the input channels. This small area is known as a CNN's *receptive field* (Luo et al. 2016). For our application, the receptive field tells us the maximal spatial context size and thus the maximal size of a meteorological feature that can be recognized and utilized by GREMLIN to determine the value of a single pixel of the estimated MRMS image.

One can calculate the maximal extent of the receptive field, aka the theoretical receptive field (TRF), from the CNN architecture using a closed-form expression that depends on the filter sizes and strides for each layer (Araujo et al. 2019). Results for GREMLIN's TRF are provided in Ebert-Uphoff

TABLE 3. Categorical performance statistics for GREMLIN: POD, FAR, CSI, and categorical bias for various REFC thresholds.

REFC (dBZ)	POD	FAR	CSI	Bias
5	0.92	0.23	0.72	1.19
10	0.85	0.18	0.72	1.04
15	0.80	0.22	0.65	1.03
20	0.71	0.31	0.54	1.03
25	0.63	0.40	0.45	1.05
30	0.55	0.46	0.38	1.01
35	0.51	0.57	0.33	1.06
40	0.52	0.57	0.31	1.23
45	0.43	0.65	0.24	1.24
50	0.37	0.77	0.14	1.17

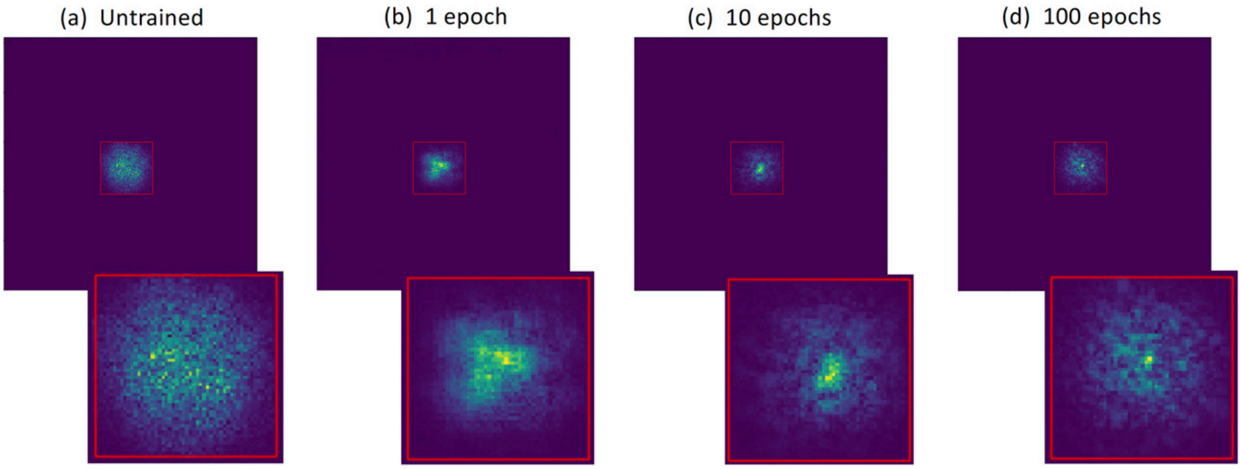

FIG. 10. ERF approximation for four different models with identical architecture (architecture of GREMLIN), but different lengths of training, ranging from (a) no training to (d) fully trained model, GREMLIN. For each image we show the ERF in the original 256 × 256 pixel (768 × 768 km) space of the input channels and a zoom-in of a 53 × 53 pixel (159 km × 159 km) region (red box). Results are for sample 68 and output pixel (125, 125). (Note that the four models did not start out with the same random seed and thus cannot strictly be seen as a progression of training toward the final model, but rather should be seen as independently trained models with different training lengths.)

and Hilburn (2020). However, pixels at the center of the receptive field have the largest impact, with impact decreasing rapidly for pixels further away in a roughly Gaussian distribution (Luo et al. 2016). Here we take the approach to sample the actual distribution of the receptive field, the *effective receptive field* (ERF; Luo et al. 2016), to understand which size neighborhood truly has a significant impact. The ERF, which depends on the network's weights, changes during training. Thus, it cannot be calculated from architecture alone. Here, we develop an ERF approximation based on the SmoothGrad algorithm (Smilkov et al. 2017). The approximation is described in detail in appendix A. The similarities between the receptive field and the radius of influence in DA applications suggests that the receptive field size—either the TRF or ERF—could potentially be used as an indication for choosing the radius of influence.

Figure 10 shows our approximation of ERF for GREMLIN for different lengths of training, ranging from an untrained model with random weights (Fig. 10a) to the final model trained for 100 epochs (Fig. 10d). Each ERF image in Fig. 10 shows the cumulative results across all four channels. Note that the ERF consistently occupies a region of less than 53 × 53 pixels or 159 km × 159 km (red squares in Fig. 10) with the region of highest impact actually much smaller than that, especially in the trained models. The ERF of the untrained model is the most spread out (Fig. 10a). Early training (Figs. 10b,c) seems to make the model put more emphasis toward the center, potentially as a sort of first-order approximation. The final model retains some focus in the center, but also spreads out more—potentially moving beyond the first-order approximation and taking additional detail into account. While the results in Fig. 10 are only ERF approximations (details in appendix A), and vary across considered samples, output pixels, and random seeds used to train the CNN, we conducted many more experiments and found the trends in Fig. 10 to be representative of the overall behavior of the ERF distributions. Please see the detailed comments in appendix A on the interpretation of such ERF approximations.

d. Applying attribution methods to identify NN strategies

To learn more about the underlying logic GREMLIN uses to derive its estimates, we use the method of LRP. Given an input sample and an output pixel, LRP reveals where the neural network was primarily looking when deriving the output pixel's estimate. We find that LRP is better suited for this purpose than standard gradient-based methods because LRP takes a global view of this decision-making process, rather than just taking a local derivative as gradient-based methods do. Details of LRP are provided in appendix B.

Figure 11 shows LRP results for GREMLIN for the same sample as in Fig. 10, but in this case focusing on a different output pixel, chosen for its close proximity to strong lightning activity. All panels in Fig. 11 are zoomed into a neighborhood of the chosen output pixel. The first row shows the input channels and corresponding desired output (i.e., the MRMS observations). Because we suspected that the neural network was heavily reliant upon the gradient of the input channels, we show an approximation of the input channel gradient magnitudes in the second row. These gradient magnitudes were calculated by applying a Sobel operator (Gonzalez and Woods 2002) to the input channels. The gradient estimates are not fed into the neural network; they are provided here simply to highlight the locations of the strongest gradients. The third row of Fig. 11 shows the first set of results, namely, the LRP maps of where in the input channels the neural network pays attention in order to estimate the value of the chosen output pixel for this sample, along with the estimated MRMS results.

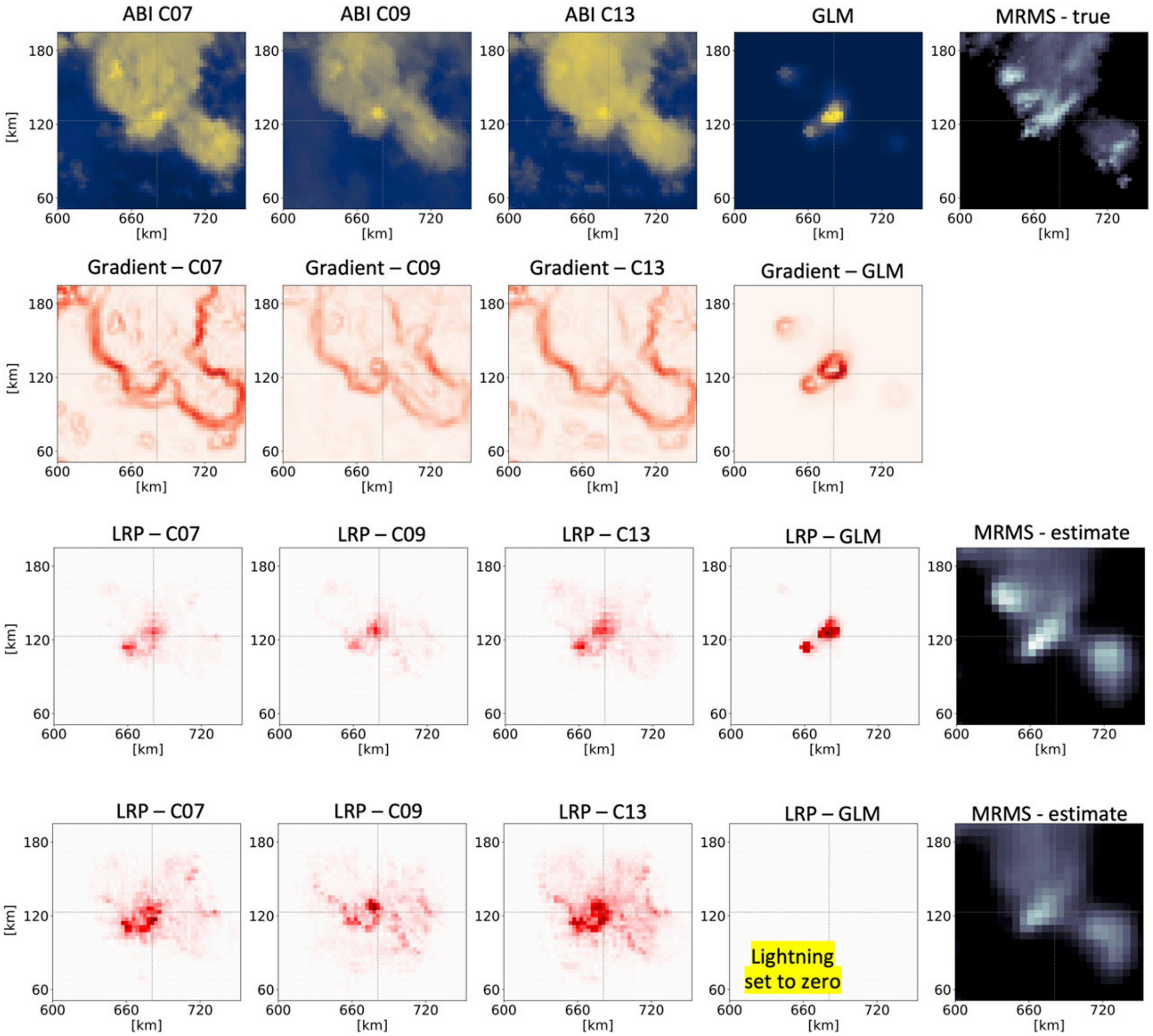

FIG. 11. LRP results for GREMLIN for sample 68 and output pixel (227, 41). (top) The four input channels (left to right: ABI C07, ABI C09, ABI C13, GLM groups) and the corresponding MRMS image (true values). (top middle) The gradient of the input channels calculated by applying a Sobel operator. (bottom middle) LRP results for the original four input channels and the chosen output pixel, and the MRMS estimate. (bottom) The equivalent of the bottom-middle row, but after all values of the GLM channel were set to zero. Note that all images are zoomed in to a region centered at the pixel of interest.

The LRP result for the GLM channel shows that the NN focused only on regions where lightning was present in that channel. The LRP results for the other channels show that even in those channels the NN's attention was drawn to focus on regions where lightning was present. We then performed a new experiment by modifying the input sample to have all lightning removed, that is, the GLM channel was set to all zero values. For this case LRP showed us that the network's focus shifted entirely to the first three input channels, that is, the ABI channels, as expected. More importantly, the focus shifted to two types of locations, namely, areas where the ABI input channels either have (i) a *large gradient* or (ii) *high brightness* (cold temperatures), as can be seen by comparing the three leftmost panels of the first, second, and fourth row. In fact, near the center of the fourth-row panels, it can be seen that the LRP patterns of the three ABI channels represent the union of the strongest gradient lines in the second row and the locations of strongest brightness in the first row. LRP vanishes further away from the center location, as expected given the nature of the ERF properties.

These results indicate the following strategy used by GREMLIN: whenever lightning is present near the output pixel, the NN primarily focuses on the values of input pixels where lightning is present, not only in the GLM channel, but in all four input channels. It seems that the network has learned that locations containing lightning are good indicators of

MRMS behavior, even in the other input channels. In the absence of any lightning, the NN focuses on (i) locations where the gradient is strong (primarily cloud boundaries) or (ii) locations of very cold cloud tops. It seems to have learned that those locations have the highest predictive power for estimating the output. Additional experiments confirmed these three strategies (lightning, cloud boundaries, cold cloud tops) of the final neural network for a wide selection of samples and output pixels.

e. Synthetic inputs to quantify sensitivity to radiance gradients

The use of architecture experiments (section 3b) and attribution methods (section 3d) have demonstrated the importance of radiance gradients for retrieving high REFC values. In this section, we construct synthetic inputs and probe the network's response to quantify that sensitivity. For this purpose, we enlist a sum of generalized elliptical Gaussians (GEG) model. This model assumes an outer Gaussian G_o that represents the thunderstorm anvil, and an inner Gaussian G_i that represents the overshooting top. The synthetic brightness temperature T is a function of (x, y) with the following parameters: location x_0 and y_0, amplitude A, size S, aspect α, orientation θ, and sharpness (exponent) p for the outer and inner Gaussians, denoted with subscripts o and i:

$$\hat{x}_{o,i} = (x - x_{0,o,i})\cos\theta_{o,i} - (y - y_{0,o,i})\sin\theta_{o,i}, \tag{4a}$$

$$\hat{y}_{o,i} = (x - x_{0,o,i})\sin\theta_{o,i} + (y - y_{0,o,i})\cos\theta_{o,i}, \tag{4b}$$

$$T_{o,i} = \exp\left\{-1\left[\frac{\hat{x}^2_{o,i}}{2S^2_{o,i}} + \frac{\hat{y}^2_{o,i}}{2(S_{o,i}\alpha_{o,i})^2}\right]^{p_{o,i}}\right\}, \quad \text{and} \tag{4c}$$

$$T = A_oT_o + A_iT_i. \tag{4d}$$

Evaluating thousands of different parameter settings, the spatial patterns that most strongly activates the network, based on the maximum REFC, all resemble Fig. 12a. What the strongly activating patterns have in common, and what is different from the weakly activating patterns, are very large p_o and large p_i, meaning that the anvil and overshooting top have very sharp T_B gradients. We evaluated p_o and p_i ranging from 0.1 to 10. The other traits that the strongly activating patterns have in common are that G_i is located near the edge of G_o and that $S_i \ll S_o$. We evaluated S_i ranging from 0 to S_o. The patterns producing a weak response tend to look unphysical from a meteorological perspective, indicating that the network has learned about realistic-looking overshooting-top signatures. This is a desirable property: rather than responding strongly to unphysical outlier inputs, it only responds strongly to patterns that look meteorological, although that does not rule out the possibility that the network could be fooled by a cleverly constructed counterexample. We explored outer sizes from 1 to 128 pixels, outer and inner aspects from 0.1 to 10, and outer and inner orientations 0°–360°. Of all the parameters of the GEG model, the ones that are most influential in producing high REFC values are p_o and p_i, and Fig. 12b characterizes the maximum REFC as a function of those parameters. The emergence of 35-dBZ echoes requires p_o to be 1 or greater or p_i to be 3 or greater. Thus, the CNN does not just respond to gradients, but calibrates its response based on the sharpness of the brightness temperature gradient. Related to these idealized synthetic input experiments, future work will consider using observation system simulation experiments to quantify errors associated with transferring from satellite observations to latent heat profiles.

4. Summary and future work

In this paper, we report on the training and evaluation of a CNN that uses ABI infrared channels and GLM lightning data to estimate MRMS REFC over eastern CONUS during the warm season. Since REFC follows an exponentially decreasing distribution, to get good performance at high values, we used a weighted loss function. This paper demonstrated that the network is learning physically meaningful strategies to predict radar reflectivity from satellite radiances and lightning. A variety of approaches were examined to investigate what the network learned and how it makes its predictions. Channel-withholding experiments showed that geostationary lightning observations are uniquely valuable for their ability to pinpoint locations of strong updrafts. Experiments that withhold spatial information demonstrated that radiance gradients carry more information about high REFC values than the radiance values themselves. Layerwise relevance propagation established that the CNN uses the information from ABI and GLM in a synergistic manner, where it interprets ABI radiance gradients in the context of whether GLM indicates the presence of lightning. Synthetic input experiments confirmed that the sharper the gradient, the stronger the CNN response, at least for patterns that have an appearance reminiscent of meteorological convection.

Having established that the horizontal spatial patterns of radar reflectivity can be accurately estimated using GOES data, the next step in this research is to produce full 3D profiles of radar reflectivity for use as an input to data assimilation systems. Here, we may leverage ongoing research to estimate cloud geometric thickness (Noh et al. 2017) and vertical structure (Miller et al. 2014) via empirically based methods. The current nonvariational technique for initializing RAP/HRRR with radar reflectivity does not require characterization of uncertainty; however, uncertainty information is required for variational approaches. Future work includes training and validation with a much larger dataset that includes samples from all times of year and using a three-way training–validation–testing split. We will also seek to provide a measure of confidence or uncertainty for use in data assimilation procedures. We also plan to try out GANs and compare results with CNNs in terms of accuracy versus blurriness. We emphasize this paper is exploratory research and the current GREMLIN, version 1, model is not suitable for estimating radar reflectivity for conditions outside of warm-season convection over CONUS.

Over CONUS the results are easy to validate using retrospective simulation experiments in which the actual radar data are withheld and replaced by the GREMLIN estimates.

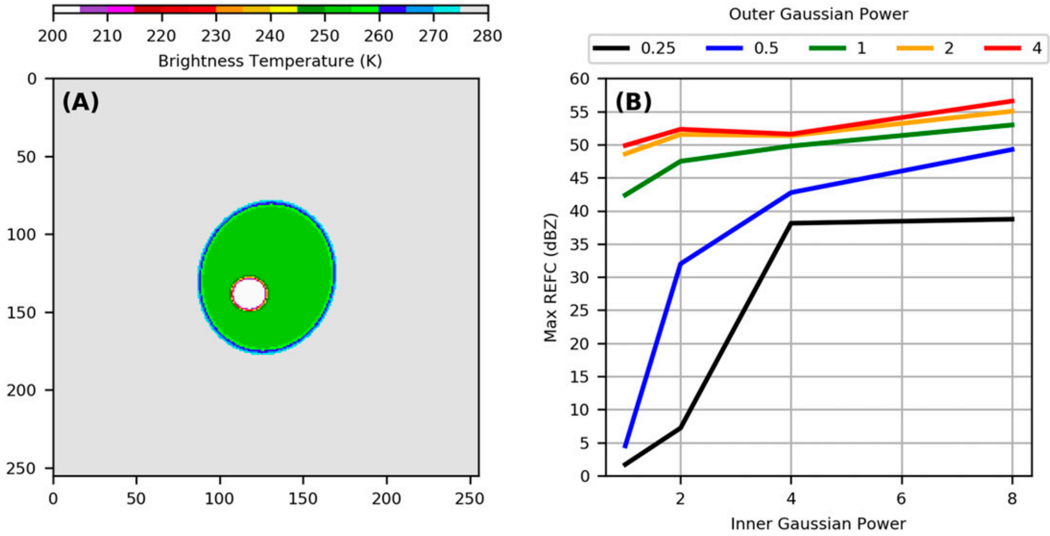

FIG. 12. (a) Synthetic C13 T_B that produces the maximum REFC response for GREMLIN. This corresponds to parameters $x_o = 128$, $y_o = 128$, $A_o = 0.5$, $S_o = 30$, $\alpha_o = 1.2$, $\theta_o = 170°$, $p_o = 10$, $x_i = x_o + dx_i \cos(\phi_i)$, $y_i = y_o + dy_i \sin(\phi_i)$, $A_i = 0.5$, $S_i = \rho_i S_o$, $\alpha_i = 1$, $\theta_i = 0°$, $p_i = 10$, $dx_i = 15$, $dy_i = 15$, $\phi_i = 135°$, and $\rho_i = 0.25$. (b) Maximum REFC as a function of inner Gaussian power (x axis) and outer Gaussian power (line color).

However, the real value of the technique will come from its ability to fill in locations that lack radar coverage because of terrain blockage, which are mostly over the western United States and coastal/oceanic locations. Evaluating results in these locations is much more difficult because of a lack of observations. However, MRMS sectors over the Caribbean Sea (*GOES-16*), Hawaii (*GOES-17*), and Guam (*Himawari-8*) do provide observations, as do spaceborne radar reflectivity observations from the Global Precipitation Measurement (GPM) Dual-Frequency Precipitation Radar (DPR). How well the model derived in this paper will generalize to meteorological regimes outside of the training set is an open question. However, it is known that both lightning and storm characteristics are different over land versus ocean (Nag and Cummins 2017; Bang and Zipser 2015). Thus, additional contextual information that is geographic or meteorological in nature may be needed, along with a deeper network to accurately depict features at the upper end of mesoalpha to synoptic scales.

Acknowledgments. We thank the GOES-R Program for supporting this research with award NA19OAR4320073 and the National Science Foundation for supporting this research through grant OAC-1934668 (to author Ebert-Uphoff). We thank NOAA RDHPCS for access to the Fine Grain Architecture System on Hera, without which this research would not have been possible.

Data availability statement. This study uses publicly available datasets. The L1b ABI data files used in this study are available from NOAA CLASS (https://www.bou.class.noaa.gov/saa/products/search?datatype_family=GRABIPRD). The L2 GLM data files used in this study are also available from NOAA CLASS (https://www.avl.class.noaa.gov/saa/products/search?datatype_family=GRGLMPROD). The MRMS composite reflectivity data files are available from NCEP (https://mrms.ncep.noaa.gov/data/).

APPENDIX A

Method for Approximating the ERF

To get an estimate of the ERF, we want to calculate and visualize how much each location in the input channels affects a specific output pixel in a considered neural network. A simple way to do so for a given input sample and chosen output pixel is to calculate the gradient of the output neuron with respect to the neurons in the input channels. Calculating this gradient is a common task in neural networks and built-in routines are readily available in neural network computing environments. However, the results tend to be noisy, and we thus use a modification of this approach, namely, the SmoothGrad algorithm by Smilkov et al. (2017). SmoothGrad calculates the gradient with respect to the input neurons several times, each time adding Gaussian noise to each pixel of each input channel before calculating the gradient, and then returns the average result. This approach, as the title of Smilkov et al. (2017) aptly states, removes noise (in the results) by adding noise (in the input channels).

We use the SmoothGrad implementation of the ''tf-explain'' package (see https://tf-explain.readthedocs.io/en/latest/) with 100 samples and a noise level of 1.0. Note that this noise level is chosen to be extremely large on purpose (keep in mind that our inputs are scaled to values between just 0 and 1), because that makes the results less dependent on the specific sample that was chosen for the estimation. When interpreting the resulting ERF estimates for a neural network model one should keep in mind that the results vary on the basis of i) chosen input sample, ii) chosen output pixel, and iii) random noise generated by SmoothGrad. Thus, it is important to generate estimates for variations of all these parameters and ensure that results are representative of the general trends. A property we noticed varying across those parameters is the presence of a few high-intensity pixels in the resulting maps. Their number and location can vary and thus should not be assigned special meaning. Aside from such details the overall distribution is fairly consistent, namely, how diffuse the ERF is and how far it stretches out from the center. More generally, *results from this ERF approximation method should be seen as a random sample drawn from a given distribution, rather than each pixel value being given specific meaning.*

APPENDIX B

LRP

A key idea of layerwise relevance propagation is that it seeks to track *relevance* backward from an output neuron to the input image, by tracking backward which neurons in the prior layer were most responsible for the values of a neuron in the later layer. To do so LRP does not use any of the built-in backpropagation rules of neural networks and develops instead its own set of customized rules. By applying those rules iteratively, an overall estimate of relevance in the input space is obtained. LRP is a fairly complex topic and the details are beyond the scope of this paper. For a detailed introduction see Bach et al. (2015), Montavon et al. (2018), or Toms et al. (2019).

We are using the implementation of LRP in the "investigate" package for Tensorflow (see https://innvestigate.readthedocs.io/en/latest/). We are using the alpha–beta rule [Eq. (60) in Bach et al. (2015)] with alpha = 1 and beta = 0, to only approximate positive attribution, that is, to identify locations for which higher activation values tend to make high values at the output *more* likely. We had to use a few tricks to make this implementation work for our purpose. First, we flattened the output layer of the NN into a vector to be able to prescribe at which output pixel we want to look. Second, we did not use the standard heat-map visualization provided by the package but instead split the heatmap result for LRP into its separate channels and plotted them separately. *For the interpretation of LRP results one needs to keep in mind that LRP uses approximation rules and that it was specifically designed for classification tasks, and not regression tasks, and therefore results should always be interpreted as showing overall trends but should not be interpreted on a pixel-by-pixel level.*

REFERENCES

Agrawal, S., L. Barrington, C. Bromberg, J. Burge, C. Gazen, and J. Hickey, 2019: Machine learning for precipitation nowcasting from radar images. arXiv 1912.12132, 6 pp., https://arxiv.org/pdf/1912.12132.pdf.

Araujo, A., W. Norris, and J. Sim, 2019: Computing receptive fields of convolutional neural networks. *Distill*, **4**, e21, https://doi.org/10.23915/distill.00021.

Arkin, P. A., and B. N. Meisner, 1987: The relationship between large-scale convective rainfall and cold cloud over the Western Hemisphere during 1982–84. *Mon. Wea. Rev.*, **115**, 51–74, https://doi.org/10.1175/1520-0493(1987)115<0051:TRBLSC>2.0.CO;2.

Ayzel, G., T. Scheffer, and M. Heistermann, 2020: RainNet v1.0: A convolutional neural network for radar-based precipitation nowcasting. *Geosci. Model Dev.*, **13**, 2631–2644, https://doi.org/10.5194/gmd-13-2631-2020.

Bach, S., A. Binder, G. Montavon, F. Klauschen, K. R. Müller, and W. Samek, 2015: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, **10**, e0130140, https://doi.org/10.1371/journal.pone.0130140.

Bang, S. D., and E. J. Zipser, 2015: Differences in size spectra of electrified storms over land and ocean. *Geophys. Res. Lett.*, **42**, 6844–6851, https://doi.org/10.1002/2015GL065264.

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

Blau, Y., and T. Michaeli, 2018: The perception-distortion tradeoff. *Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, CVPR, 6228–6237, https://openaccess.thecvf.com/content_cvpr_2018/html/Blau_The_Perception-Distortion_Tradeoff_CVPR_2018_paper.html.

Boccippio, D. J., 2002: Lightning scaling relations revisited. *J. Atmos. Sci.*, **59**, 1086–1104, https://doi.org/10.1175/1520-0469(2002)059<1086:LSRR>2.0.CO;2.

Ebert-Uphoff, I., and K. Hilburn, 2020: Evaluation, tuning, and interpretation of neural networks for working with images in meteorological applications. *Bull. Amer. Meteor. Soc.*, https://doi.org/10.1175/BAMS-D-20-0097.1, in press.

Elmer, N. J., E. Berndt, and G. J. Jedlovec, 2016: Limb correction of MODIS and VIIRS infrared channels for the improved interpretation of RGB composites. *J. Atmos. Oceanic Technol.*, **33**, 1073–1087, https://doi.org/10.1175/JTECH-D-15-0245.1.

Fuchs, B. R., S. A. Rutledge, B. Dolan, L. D. Carey, and C. Schultz, 2018: Microphysical and kinematic processes associated with anomalous charge structures in isolated convection. *J. Geophys. Res. Atmos.*, **123**, 6505–6528, https://doi.org/10.1029/2017JD027540.

Geer, A., and Coauthors, 2018: All-sky satellite data assimilation at operational weather forecasting centres. *Quart. J. Roy. Meteor. Soc.*, **144**, 1191–1217, https://doi.org/10.1002/qj.3202.

Gonzalez, R. C., and R. E. Woods, 2002: *Digital Image Processing*. 2nd ed. Prentice-Hall, 793 pp.

Goodman, S., D. Mach, W. Koshak, and R. Blakeslee, 2010: GLM lightning cluster-filter algorithm, version 2.0. NOAA NESDIS STAR Doc., 70 pp.

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

Gustafsson, N., and Coauthors, 2018: Survey of data assimilation methods for convective-scale numerical weather prediction at operational centres. *Quart. J. Roy. Meteor. Soc.*, **144**, 1218–1256, https://doi.org/10.1002/qj.3179.

Harris Corporation, 2016: Product definition and user's guide (PUG). Harris Corp. Vol. 5, Level 2+ Products, DCN-7035538, Revision-E, 699 pp.

Honda, T., and Coauthors, 2018a: Assimilating all-sky *Himawari-8* infrared radiances: A case of Typhoon Soudelor (2015). *Mon. Wea. Rev.*, **146**, 213–229, https://doi.org/10.1175/MWR-D-16-0357.1.

——, S. Kotsuki, G.-Y. Lien, Y. Maejima, K. Okamoto, and T. Miyoshi, 2018b: Assimilation of *Himawari-8* all-sky radiances every 10 minutes: Impact on precipitation and flood risk prediction. *J. Geophys. Res. Atmos.*, **123**, 965–976, https://doi.org/10.1002/2017JD027096.

Jones, T. A., D. Stensrud, L. Wicker, P. Minnis, and R. Palikonda, 2015: Simultaneous radar and satellite data storm-scale assimilation using an ensemble Kalman filter approach for 24 May 2011. *Mon. Wea. Rev.*, **143**, 165–194, https://doi.org/10.1175/MWR-D-14-00180.1.

——, and Coauthors, 2020: Assimilation of *GOES-16* radiances and retrievals into the Warn-on-Forecast system. *Mon. Wea. Rev.*, **148**, 1829–1859, https://doi.org/10.1175/MWR-D-19-0379.1.

Kong, R., M. Xue, A. O. Fierro, Y. Jung, C. Liu, E. R. Mansell, and D. R. MacGorman, 2020: Assimilation of GOES-R Geostationary Lightning Mapper flash extent density data in GSI EnKF for the analysis and short-term forecast of a mesoscale convective system. *Mon. Wea. Rev.*, **148**, 2111–2133, https://doi.org/10.1175/MWR-D-19-0192.1.

Lapuschkin, S., S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, 2019: Unmasking Clever Hans predictors and assessing what machines really learn. *Nat. Commun.*, **10**, 1096, https://doi.org/10.1038/s41467-019-08987-4.

Lin, J., S. S. Weygandt, S. G. Benjamin, and M. Hu, 2017: Satellite radiance data assimilation within the hourly updated Rapid Refresh. *Wea. Forecasting*, **32**, 1273–1287, https://doi.org/10.1175/WAF-D-16-0215.1.

Line, W. E., T. J. Schmit, D. T. Lindsey, and S. J. Goodman, 2016: Use of geostationary super rapid scan satellite imagery by the Storm Prediction Center. *Wea. Forecasting*, **31**, 483–494, https://doi.org/10.1175/WAF-D-15-0135.1.

Luo, W., Y. Li, R. Urtasun, and R. Zemel, 2016: Understanding the effective receptive field in deep convolutional neural networks. *Advances in Neural Information Processing Systems*, The MIT Press, 4898–4906.

Marchand, M., K. Hilburn, and S. D. Miller, 2019: Geostationary Lightning Mapper and Earth Networks lightning detection over the contiguous United States and dependence on flash characteristics. *J. Geophys. Res. Atmos.*, **124**, 11 552–11 567, https://doi.org/10.1029/2019JD031039.

McGovern, A., 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.

Miller, S. D., and Coauthors, 2014: Estimating three-dimensional cloud structure via statistically blended active and passive sensor observations. *J. Appl. Meteor. Climatol.*, **53**, 437–455, https://doi.org/10.1175/JAMC-D-13-070.1.

——, M. A. Rogers, J. M. Haynes, M. Sengupta, and A. K. Heidinger, 2018: Short-term solar irradiance forecasting via satellite/model coupling. *Sol. Energy*, **168**, 102–117, https://doi.org/10.1016/j.solener.2017.11.049.

Montavon, G., W. Samek, and K. R. Müller, 2018: Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.*, **73**, 1–15, https://doi.org/10.1016/j.dsp.2017.10.011.

Nag, A., and K. L. Cummins, 2017: Negative first stoke leader characteristics in cloud-to-ground lightning over land and ocean. *Geophys. Res. Lett.*, **44**, 1973–1980, https://doi.org/10.1002/2016GL072270.

NOAA NESDIS, 1998: Earth Location User's Guide (ELUG), Revision 1. NOAA NESDIS Doc. DRL 504-11, NOAA/OSD3-1998-015R1UD0, 99 pp.

Noh, Y.-J., and Coauthors, 2017: Cloud base height estimation from VIIRS, Part II: A statistical algorithm based on A-train satellite data. *J. Atmos. Oceanic Technol.*, **34**, 585–598, https://doi.org/10.1175/JTECH-D-16-0110.1.

Okamoto, K., Y. Sawada, and M. Kunii, 2019: Comparison of assimilating all-sky and clear-sky infrared radiances from *Himawari-8* in a mesoscale system. *Quart. J. Roy. Meteor. Soc.*, **145**, 745–766, https://doi.org/10.1002/qj.3463.

Otkin, J. A., and R. Potthast, 2019: Assimilation of all-sky SEVIRI infrared brightness temperatures in a regional-scale ensemble data assimilation system. *Mon. Wea. Rev.*, **147**, 4481–4509, https://doi.org/10.1175/MWR-D-19-0133.1.

Price, C., and D. Rind, 1992: A simple lightning parameterization for calculating global lightning distributions. *J. Geophys. Res.*, **97**, 9919–9933, https://doi.org/10.1029/92JD00719.

Roebber, P., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

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

Rutledge, S. A., K. Hilburn, A. Clayton, B. Fuchs, and S. D. Miller, 2020: Evaluating Geostationary Lightning Mapper flash rates within intense convective storms. *J. Geophys. Res. Atmos.*, **125**, e2020JD032827, https://doi.org/10.1029/2020JD032827.

Samsi, S., C. J. Mattioli, and M. S. Veillette, 2019: Distributed deep learning for precipitation nowcasting. *IEEE High Performance Extreme Computing Conf.*, Waltham, MA, IEEE, https://doi.org/10.1109/HPEC.2019.8916416.

Sawada, Y., K. Okamoto, M. Kunii, and T. Miyoshi, 2019: Assimilating every-10-minute *Himawari-8* infrared radiance to improve convective predictability. *J. Geophys. Res. Atmos.*, **124**, 2546–2561, https://doi.org/10.1029/2018JD029643.

Schmit, T., M. Gunshor, G. Fu, T. Rink, K. Bah, and W. Wolf, 2010: Cloud and Moisture Imagery Product (CMIP), version 2.3. NOAA NESDIS STAR GOES-R Advanced Baseline Imager (ABI) Algorithm Theoretical Basis Doc., 62 pp.

——, P. Griffith, M. M. Gunshor, J. M. Daniels, S. J. Goodman, and W. J. Lebair, 2017: A closer look at the ABI on the GOES-R series. *Bull. Amer. Meteor. Soc.*, **98**, 681–698, https://doi.org/10.1175/BAMS-D-15-00230.1.

Schultz, C. J., W. A. Petersen, and L. D. Carey, 2009: Preliminary development and evaluation of lightning jump algorithms for the real-time detection of severe weather. *J. Appl. Meteor. Climatol.*, **48**, 2543–2563, https://doi.org/10.1175/2009JAMC2237.1.

——, L. D. Carey, E. V. Schultz, and R. J. Blakeslee, 2015: Insight into the kinematic and microphysical processes that control lightning jumps. *Wea. Forecasting*, **30**, 1591–1621, https://doi.org/10.1175/WAF-D-14-00147.1.

Smilkov, D., N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, 2017: SmoothGrad: Removing noise by adding noise. arXiv 1706.03825, 10 pp., https://arxiv.org/pdf/1706.03825.pdf.

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

Snyder, J. P., 1987: Map -projections—A working manual. U.S. Geological Survey Profession Paper 1395, 397 pp.

Stengel, K., A. Glaws, D. Hettinger, and R. N. King, 2020: Adversarial super-resolution of climatological wind and solar data. *Proc. Nat. Acad. Sci. USA*, 16 805–16–815, https://doi.org/10.1073/pnas.1918964117.

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.

Svaldi, A., 2017: Hailstorm that hammered west metro Denver May 8 is costliest ever for Colorado. *Denver Post*, https://www.denverpost.com/2017/05/23/hailstorm-costliest-ever-metro-denver/.

Toms, B. A., E. A. Barnes, and I. Ebert-Uphoff, 2019: Physically interpretable neural networks for the geosciences: Applications to earth system variability. arXiv 1912.01752, 28 pp., https://arxiv.org/pdf/1912.01752.pdf.

Veillette, M. S., E. P. Hassey, C. J. Mattioli, H. Iskenderian, and P. M. Lamey, 2018: Creating synthetic radar imagery using convolutional neural networks. *J. Atmos. Oceanic Technol.*, **35**, 2323–2338, https://doi.org/10.1175/JTECH-D-18-0010.1.

Vicente, G. A., J. C. Davenport, and R. A. Scofield, 2002: The role of orographic and parallax corrections on real time high resolution satellite rainfall rate distribution. *Int. J. Remote Sens.*, **23**, 221–230, https://doi.org/10.1080/01431160010006935.

Walther, A., W. Straka, and A. K. Heidinger, 2013: ABI algorithm theoretical basis Document for Daytime Cloud Optical and Microphysical Properties (DCOMP). NOAA/NESDIS/STAR Algorithm Theoretical Basis Doc., 61 pp., https://www.goes-r.gov/products/ATBDs/baseline/Cloud_DCOMP_v2.0_no_color.pdf.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 627 pp.

Williams, E., V. Mushtak, D. Rosenfeld, S. Goodman, and D. Boccippio, 2005: Thermodynamic conditions favorable to superlative thunderstorm updraft, mixed phase microphysics and lightning flash rate. *Atmos. Res.*, **76**, 288–306, https://doi.org/10.1016/j.atmosres.2004.11.009.

Zhang, Y., F. Zhang, and D. J. Stensrud, 2018: Assimilating all-sky infrared radiances from *GOES-16* ABI using an ensemble Kalman filter for convection-allowing severe thunderstorms prediction. *Mon. Wea. Rev.*, **146**, 3363–3381, https://doi.org/10.1175/MWR-D-18-0062.1.

——, D. J. Stensrud, and F. Zhang, 2019: Simultaneous assimilation of radar and all-sky satellite infrared radiance observations for convection-allowing ensemble analysis and prediction of severe thunderstorms. *Mon. Wea. Rev.*, **147**, 4389–4409, https://doi.org/10.1175/MWR-D-19-0163.1.