Physically Explainable Deep Learning for Convective Initiation Nowcasting Using GOES-16 Satellite Observations

Da Fan[a], Steven J. Greybush[a,b], Eugene E. Clothiaux[a], and David John Gagne II[c]

[a] *Department of Meteorology and Atmospheric Science, The Pennsylvania State University, University Park, Pennsylvania*

[b] *Institute for Computational and Data Sciences, The Pennsylvania State University, University Park, Pennsylvania*

[c] *National Center for Atmospheric Research, Boulder, Colorado*

Corresponding author: Da Fan, dxf424@psu.edu

Early Online Release: This preliminary version has been accepted for publication in *Artificial Intelligence for the Earth Systems*, may be fully cited, and has been assigned DOI 10.1175/AIES-D-23-0098.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT: Convective initiation (CI) nowcasting remains a challenging problem for both numerical weather prediction models and existing nowcasting algorithms. In this study, an object-based probabilistic deep learning model is developed to predict CI based on multichannel infrared GOES-16 satellite observations. The data come from patches surrounding potential CI events identified in Multi-Radar Multi-Sensor Doppler weather radar products over the Great Plains region from June and July 2020 and June 2021. An objective radar-based approach is used to identify these events. The deep learning model significantly outperforms the classical logistic model at lead times up to 1 hour, especially on the false alarm ratio. Through case studies, the deep learning model exhibits dependence on the characteristics of clouds and moisture at multiple altitudes. Model explanation further reveals that the contribution of features to model predictions is significantly dependent on the baseline, a reference point against which the prediction is compared. Under a moist baseline, moisture gradients in the lower and middle troposphere contribute most to correct CI forecasts. In contrast, under clear-sky baselines, correct CI forecasts are dominated by cloud-top features, including cloud-top glaciation, height, and cloud coverage. Our study demonstrates the advantage of using different baselines in further understanding model behavior and gaining scientific insights.

1. Introduction

Convective initiation (CI) remains a significant and challenging forecasting problem within the meteorological community. Accurately predicting the location and onset times of convection remains difficult for both empirical and numerical weather prediction (NWP) models (e.g., Mecikalski et al. 2015; Lawson et al. 2018; Cintineo et al. 2020a). The failure to forecast CI causes delayed warnings of convective hazards like heavy rainfall, hail, and tornadoes, and disruptions to outdoor activities and travel (Brooks et al. 2003; Brooks and Dotzek 2008; Dixon et al. 2011). Given its socioeconomic impacts, more accurate and timely CI nowcasts and a more thorough understanding of physical processes underlying CI are needed.

Multiple satellite-based algorithms have been developed to make use of cloud characteristics to enhance the forecast skill of CI (Sieglaff et al. 2011; Walker et al. 2012; Nisi et al. 2014; Lee et al. 2017; Zhuge and Zou 2018; Han et al. 2019). The University of Wisconsin Convective Initiation (UWCI) nowcasting algorithm (Sieglaff et al. 2011) was developed to nowcast CI based on box-average cloud-top characteristics evident within Geostationary Operational Environmental Satellite (GOES) observations. The Satellite Convection Analysis and Tracking, version 2, (SATCASTv2) algorithm was developed by Walker et al. (2012) to track cumulus clouds and nowcast the probability of CI in the cloud objects. Cloud-top features, like cloud-top cooling rate and phase change, were employed to predict CI (Sieglaff et al. 2011; Walker et al. 2012). Using similar features from the Himawari-8 satellite, Lee et al. (2017) developed a random forest, a machine learning (ML) model, to predict CI for tracked cloud objects. Han et al. (2019) extended the framework in Lee et al. (2017) by integrating a procedure to iteratively expand the training dataset, resulting in a decline in the overall CI forecast skill but slightly better skill at longer lead times. However, the high false alarm ratio remains a significant issue in these algorithms (Sieglaff et al. 2011; Walker et al. 2012; Lee et al. 2017; Han et al. 2019). Another issue is that a substantial number of CI events blocked by thick cirrus clouds were ignored. In these studies, spatial variations of cloud features surrounding potential CI events were not used to improve the forecast skill.

Fine-resolution infrared observations are now available every few minutes from the current generation of geostationary satellites. These high spatiotemporal resolution satellite observations have increased our ability to better represent cloud-top characteristics associated with convection (e.g., Senf and Deneke 2017; Apke et al. 2018; Fan et al. 2022). Enhanced spatial patterns of environmental information have been found critical for predicting severe hailstorms (Gagne et al. 2019) and tornadoes (Lagerquist et al. 2019) using deep learning methods. However, spatial features in these new satellite observations have not yet received much attention for increasing CI forecast skill.

ML and deep learning methods have recently gained popularity as a powerful tool in convective weather detection and forecasting (Cintineo et al. 2014, 2018; McGovern et al. 2019; Gagne et al. 2019; Lee et al. 2020; Cintineo et al. 2020a,b; Mecikalski et al. 2021; Leinonen et al. 2022). Mecikalski et al. (2021) showed that a random forest model improved the forecast skill of severe storms with satellite, radar, and lightning data as predictors. Leinonen et al. (2022), using a gradient-boosted tree algorithm, demonstrated that satellite data are beneficial for ML-based severe weather nowcasting, while radar data are the most important predictor overall. Lee et al. (2020) developed a convolutional neural network (CNN; Lecun et al. 2015), a popular deep learning method that encodes spatial features within data to enhance forecast skill, for detecting convective regions from satellite observations with improved accuracy. Lagerquist et al. (2021) further demonstrated that deep learning provides skillful forecasts of the spatial coverage of convection at lead times up to 120 minutes using infrared satellite data. Sun et al. (2023) developed a convolutional recurrent neural network that leverages spatiotemporal features from satellite and radar data to predict convective weather, and their model showed good forecast skill in several CI cases at lead times up to 30 min. However, Sun et al. (2023) lacks an in-depth statistical evaluation of CI forecast skill. The successful applications of deep learning to detect and forecast convection with satellite observations hold promise for our study on developing deep learning models for CI nowcasting using infrared satellite data.

Despite the increase of ML's successful applications in meteorology, it is often criticized by forecasters and domain scientists as a "black box" technique because of our inability to readily interpret its decision-making process in physical terms. Thus, explainable artificial intelligence (XAI) has received a lot of attention in both the meteorology and ML communities (Olah et al. 2017; Lipton 2018; McGovern et al. 2019; Toms et al. 2020; Molnar 2020). XAI encapsulates and approximates intricate relations between inputs and model predictions inherent in the decision-making process, enabling domain scientists to gain trust in the model, as well as understand its limitations. This facilitates application of the model to ideal scenarios (McGovern et al. 2019) and withholding it from inappropriate ones. XAI is becoming increasingly important as ML methods outperform current NWP models in some applications.

Successful applications of XAI include Toms et al. (2020), who identified the spatial patterns for two dominant modes of El Nino variability using layer-wise relevance propagation (LRP) and backwards optimization, two XAI methods. Mayer and Barnes (2021) demonstrated that neural networks are able to identify tropical hot spots that are important for subseasonal predictions in the North Atlantic through model explanations using LRP. Mamalakis et al. (2022) objectively assessed the performance of different explanation methods on a large benchmark dataset and discussed their reliability and limitations compared to the ground truth.

The purpose of this study is to characterize nowcasting skill of CI obtained through two ML models trained on GOES-16 satellite infrared observations and to explore the radiative features that lead to skill in forecasting CI through model explanation and visualization. A CNN model is optimized and evaluated against logistic regression, a classical statistical method. False positive prediction is a crucial issue in previous CI forecasting algorithms (Mecikalski et al. 2015; Apke et al. 2015), so particular attention is paid to this challenge. The rest of this paper is organized as follows. Section 2 describes CI identification and data preprocessing. Section 3 describes model architectures, optimization, evaluation, and XAI method. Section 4 evaluates the CNN and logistic regression models through performance statistics and case examples. Section 5 explains radiative features in the decision-making process of the CNN with different choices of the baseline. Section 6 presents the main findings and limitations of the study, and includes concluding remarks.

2. Data

We use GOES-16 Advanced Baseline Imager (ABI) data to generate predictors for CI events obtained from the Multi-Radar Multi-Sensor (MRMS; Lakshmanan et al. 2006, 2007) dataset. This study focuses on the Great Plains in the United States (Fig. 1), following Apke et al. (2015) and Walker et al. (2012), because of the importance of CI to this region and the availability of dense radar observations within it.

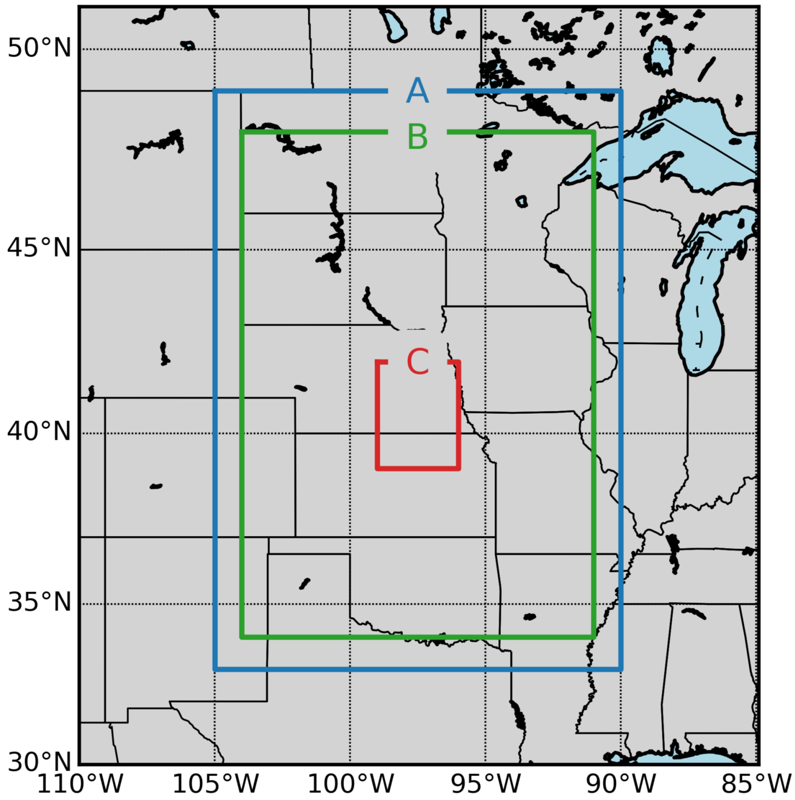

Fig. 1. Total area of study (blue box labeled with A), track-corrected area of study (green box labeled with B, and ranging from 91°W to 104°W and from 34°N to 48°N), and validation domain for manually identified CI true clusters (red box labeled with C).

a. CI definition and identification

Radars used to produce the MRMS dataset make azimuth scans at a number of elevation angles (Lakshmanan et al. 2006). These radar data are then linearly interpolated onto a uniform grid for postprocessing. Composite reflectivity, defined as the maximum reflectivity in a column on a 0.01° spatial grid and 2-min temporal grid in the MRMS dataset, is used for storm tracking and CI identification. Use of a column-max value mitigates against the impacts of terrain blocking radar beams and increases the probability of detecting convective cells in their early stages as hydrometeors form at higher levels (e.g., Matthee et al. 2014; Apke et al. 2015; Senf and Deneke 2017; Henderson et al. 2021). A radar reflectivity threshold of 35 dBZ is employed to distinguish between convective and nonconvective regions (Mecikalski and Bedka 2006; Kain et al. 2013). Pixels are defined as being part of a CI cluster when the following conditions (adapted from Colbert et al. 2019) are met:

(1) composite reflectivity ≥ 35 dBZ;

(2) in the preceding 11 min, no points within 15 km exhibit composite reflectivity ≥ 35 dBZ; and

(3) in the preceding 30 min, no points within 5 km exhibit composite reflectivity ≥ 35 dBZ.

The first condition identifies pixels associated with convection either just initiated or advected to the pixel from the surroundings. The second and third conditions eliminate pixels related to preexisting convection. Storm tracking and CI cluster identification are done in three steps. First, storm cells are identified and clustered into tracks iteratively in time using the w2segmotionll algorithm (Lakshmanan and Smith 2010), which is part of the Warning Decision Support System–Integrated Information (WDSS-II) suite of algorithms. Combined K-means and enhanced watershed methods (Lakshmanan and Smith 2009) are used for this purpose. Second, the tracks are corrected by a modified best-track algorithm (Lakshmanan et al. 2015) using post-event tracking, which fits the storm cells to a best-fit Theil-Sen trajectory and removes falsely truncated tracks. Third, the first storm cells within the final tracks are identified as CI clusters. CI clusters are validated through two examples selected from the area C in Fig. 1. (See the video in the online supplemental material.)

b. Feature engineering

An object-based forecasting method is designed to identify localized environments within which to predict CI, thereby largely reducing the data volume. Both CI and non-CI events must be identified for the dataset. CI events are 48-km by 48-km square patches centered on at least one CI cluster, whereas most (~91%) non-CI events, called Near-Miss (NM) events, are 48-km by 48-km square patches that are nearest neighbors to CI events and contain no CI cluster of their own. The rest (~9%) of the non-CI events, called RandoM (RM) events, are 48-km by 48-km square patches randomly extracted across the Great Plains area and neither contain a CI cluster nor are a nearest neighbor to a patch that does. To avoid the impacts of class imbalance on model performance (Ukkonen and Mäkelä 2019), non-CI events, the majority class, are undersampled to produce a balanced dataset that consists of 58% CI, 38% NM, and 4% RM events. We use binary labels to classify these events, so that 1 indicates a CI event and 0 a non-CI event. Our entire dataset consists of 94,618 samples. 45,077, and 19,320 samples from June and July 2020 are for training

TABLE 1. Central wavelength, channel number, and description of BTs from seven GOES-16 infrared channels used as predictors.

Central wavelength (μm)	Channel number	description
6.2	8	Upper-level tropospheric water vapor
6.95	9	Mid-level tropospheric water vapor
7.3	10	Lower-level tropospheric water vapor
8.4	11	Cloud-top phase
9.6	12	Ozone
10.35	13	Cloud-top/Surface temperature
11.2	14	Cloud-top/Surface temperature

and validation, respectively, whereas 30,221 samples collected in June 2021 are for testing. Predictors for CI and non-CI events are brightness temperatures (BTs; Table 1) from seven GOES-16 ABI infrared channels with ~2-km native horizontal resolution available every 5 minutes over the continental United States (CONUS). Predictors are extracted from 48-km by 48-km square patches (i.e., around 24 columns by 24 rows in a GOES-16 ABI image) at lead times from 60 minutes down to 10 minutes before the occurrence time of an event. As GOES-16 ABI views clouds across CONUS slantwise, their surface referenced latitudes and longitudes in the database differ from their vertically projected latitudes and longitudes, with the difference largest for the highest altitude clouds. This displacement, called parallax error, is comparable to the scale of the clouds during thunderstorm initiation (Zhang et al. 2019) and thus not negligible. Following Zhang et al. (2019), parallax errors are corrected using the cloud-top height (ACHA) product of GOES-16 to improve the quality of cloud locations.

The depth and performance of CNNs are largely limited by the size of the input (Thambawita et al. 2021; Sabottke and Spieler 2020), with small-size inputs usually leading to shallow CNNs with limited ability to encode complex spatial features. Thus, GOES-16 ABI BTs are remapped to a 1.5-km mesh through linear interpolation, so that the 48-km by 48-km square patches contain 32×32 input BT values. Each predictor set of channel BTs is standardized using its mean and standard deviation (Table 1) to a set of values with zero mean and standard deviation of one prior to being fit by the two ML models.

3. Methods

This section introduces the architecture of the logistic regression and CNN models, a hyperparameter optimization method, and model explanation approaches.

a. Logistic regression

Logistic regression is a nonlinear transformation with a sigmoid function applied to the weighted sum of the predictors (x_i):

$$p = \frac{1}{1+e^{-z}}, \text{ where } z = \beta_0 + \sum_{i=1}^{N} \beta_i x_i, \tag{1}$$

p is the prediction in the range between zero and one, N is the total number of predictors, β_i is the i^{th} learned weight, and β_0 is the bias term. Predictions from logistic regressions are often used to estimate probabilities for classification problems with monotonic relationships between predictors and predictands. The model weights are iteratively adjusted by minimizing the binary cross-entropy

$$C\sum_{j=1}^{M}[y_j \log_2(p_j) + (1-y_j)\log_2(1-p_j)] + \lambda \sum_{j=1}^{M} |\beta_j| + \frac{1-\lambda}{2}\sum_{j=1}^{M} |\beta_j|^2 \tag{2}$$

between the true labels (y_j) and the predictions (p_j), where the two additional terms, known as elastic-net penalties, are for regularization, M is the number of samples, C is the inverse of the regularization strength, and λ is the mixing parameter that controls the strengths of the two regularization terms. The second term in Eq.(2) is known as the lasso penalty, or L_1 regularization, and rewards small weights by penalizing the sum of absolute values of the weights (Tibshirani 1996). The third term is known as the ridge penalty, or L_2 regularization, and reduces the impacts of multicollinearity, i.e., correlations between predictors, by adding additional penalties to large weights (Hoerl and Kennard 1988).

Despite being a simple ML model, logistic regression performs well on some problems in weather forecasting, like distinguishing between lightning and non-lightning days (Bates et al. 2018) and predicting CI using satellite observations (Mecikalski et al. 2015). In our study, the baseline logistic model feeds flattened GOES-16 predictors into a logistic regression to predict the probability of

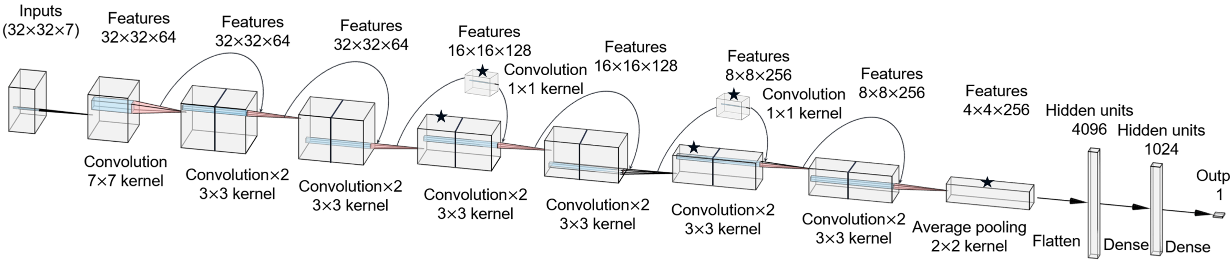

Fig. 2. Schematic of the ResNet architecture. The curves connecting the input and output of residual blocks indicate residual layers. The blocks marked by a star involve downsampling of features from the preceding layer. The architecture parameters and components shown produced the highest validation score during the hyperparameter optimization.

CI. The logistic model is implemented in version 0.20 of the scikit-learn (Pedregosa et al. 2011) library.

b. Residual neural network

We also trained a deep CNN architecture to predict the probability of CI. Deep neural networks often perform better than shallow networks by encoding spatial features across multiple scales through higher-order abstraction. A drawback of deep learning networks is potential difficulty with convergence resulting from vanishing gradients during the optimization process (Glorot and Bengio 2010). Gradients decrease exponentially as they are propagated back to the early layers so that optimization of the weights and biases in the early layers becomes problematic during training. To avoid this issue, we used the residual neural network (ResNet) architecture (He et al. 2016) with its residual connections. In this architecture, the output of earlier layers is added to the output of later layers to preserve high-resolution information and thus preserve the gradients. The ResNet method performed well in medium-range weather forecasting in the WeatherBench challenge, and its skill is comparable to the baseline physical model at a similar resolution (Rasp and Thuerey 2021).

A schematic of the ResNet architecture used in this study is shown in Fig. 2. For the first layer, a single two-dimensional (2D) convolutional block extracts features from the inputs using a kernel size of (7, 7) to broaden the view of the field. The convolutional block is defined as a sequence of 2D convolution layer → Leaky rectified linear (Leaky ReLU; max(0.138x,x)) activation → Batch normalization → Dropout. The 2D convolution layer extracts spatial features from the inputs by transforming them through a number of filters, which are small patches of weights and biases. Then, the Leaky ReLU activation function is applied to feature maps to enable a small, non-zero gradient for negative signals, mitigating the vanishing gradient problem by preventing inactive neurons (Maas et al. 2013). Batch normalization (Ioffe and Szegedy 2015) is subsequently used to rescale the values of the features in order to maintain a more stable structure during training and to enable faster convergence of model errors with higher learning rates. Dropout regularization then randomly sets the values of certain features to zero with a fixed probability to prevent overfitting (Srivastava et al. 2014).

After the first convolutional block, there are six residual blocks with each block consisting of two 2D convolutional blocks with a 3×3 kernel and a residual layer. The residual layer adds the features of the preceding layer to the features of the residual block to increase the magnitude of the gradients. In the third and fifth residual blocks, the features from the preceding layer are downsampled via a residual layer with a 1×1 kernel and a stride of 2 and a convolutional block with a 3×3 kernel and a stride of 2. The number of convolutional filters increases by a factor of two from the second to third and fourth to fifth residual blocks. This is done to offset the loss of information from a decrease in spatial resolution at these steps. The spatial resolutions of the feature maps decrease with increasing depth, and the feature maps evolve to contain different levels of abstraction. The average pooling layer reduces the dimensions of the feature maps by a factor of 2 via a convolution with a 2×2 kernel and a stride of 2, thereby refining the features used for prediction. The resulting features are then flattened into a one-dimensional feature vector. The vector is then condensed through two fully connected dense layers. Each feature of a dense layer is a weighted sum of features from the previous layer. The outputs of the final dense layer are transformed through a sigmoid activation function into the probability of CI.

We trained the ResNet model using binary cross-entropy as the loss function, and we used area-under-curve (AUC) scoring as the metric to track model performance on the validation data during training. An Adam optimizer was used with an initial learning rate. The learning rate was decreased by a factor of two after validation losses did not decrease across three training epochs. We terminated training when the validation losses did not decrease across ten training epochs. We built the ResNet model using the Keras library (Chollet et al. 2015) with a Tensorflow low-level backend (Abadi et al. 2016).

c. Model optimization and evaluation

To find the optimal configuration for both the logistic regression model and the ResNet, we performed a guided search over a range of hyperparameters using Earth Computing Hyperparameter Optimization (ECHO: `https://doi.org/10.5281/zenodo.7787022`). The search is based on the Tree-structured Parzen Estimator (TPE) sampler, which samples the next hyperparameters based on the ranking information of previous experiments. We performed 200 hyperparameter searches and selected the model with the highest AUC score on the validation data. The AUC score is the area under the receiver operating curve (ROC; Mason 1982) and assesses a model's ability to discriminate between classes. An AUC score of 0.5 indicates a no-skill forecast model, while a score of 1.0 is an indication of a perfect discriminator. The selected hyperparameters, search space, and optimal values are shown in Table A2.

We then evaluated the optimized logistic regression and ResNet models on the testing dataset. Most performance metrics were derived from the relationship between CI/non-CI observations and binary deterministic predictions ("yes"/"no") converted from probabilistic forecasts using a probability threshold. The four possible outcomes are: 1) hits: correctly forecast CI occurrences, 2) false alarms: CI forecast where no CI occurred, 3) correct negatives: correctly forecast non-CI occurrences, 4) misses: non-CI forecast where CI occurred. Commonly used metrics for deterministic forecasts include probability of detection [POD; h/(h+m)], probability of false detection [POFD; f/(f+c)], false alarm ratio [FAR; f/(f+h)], success ratio [SR; h/(h+f)], frequency bias [(h+f)/(h+m)] and critical success index [CSI; h/(h+m+f)], where h, f, c, and m are the frequency of hits, false alarms, correct negatives, and misses, respectively. The probability thresholds were selected to maximize the CSI for both the logistic regression and ResNet models. The models were individually trained using the same architecture at lead times from 60 minutes to 10 minutes prior to the event, in steps of 10 minutes. Then, the models were evaluated and compared to each other.

The final skill score in this study is the Brier skill score (BSS; Wilks 2019). The BSS is a measure of the improvement of the forecast skill relative to climatology based on how well the probabilistic forecast agrees with the observed event frequency. The BSS is decomposed into two terms with a scaling factor:

$$BSS = \frac{Resolution - Reliability}{Uncertainty} = \frac{\frac{1}{N}\sum_{k=1}^{K} n_k (y_k - \overline{y})^2 - \frac{1}{N}\sum_{k=1}^{K} n_k (p_k - y_k)^2}{\overline{y}(1-\overline{y})} \tag{3}$$

where N is the total number of samples, K is the number of bins, n_k is the number of samples in the k^{th} probability bin, y_k is the conditional event frequency given the probability in the k^{th} bin, p_k is the mean probability in the k^{th} bin, and $\overline{y}$ is the climatological event frequency. The two terms in the numerator are known as the resolution and the reliability, while the scaling factor in the denominator is known as the uncertainty. The resolution measures the difference between conditional event frequencies and the observed climatological frequency, whereas the reliability measures how close the forecast probabilities are to the observed frequency at the corresponding probabilities. The uncertainty term rescales the BSS score based on the class proportion. A BSS score of 1 represents a perfect model, whereas a score less than 0 indicates the model is worse than climatology.

d. Explainable AI

To explain the encoding underlying the complex structures of the ResNet model, we employed a model-agnostic method called SHapley Additive exPlanations (SHAP; Lundberg and Lee 2017). SHAP explains individual predictions as a game played by features and fairly distributes the payout among the features (Molnar 2020). A player of the game is an individual feature value or a group of feature values. To explain an image input, pixels are grouped into superpixels and contributions to the prediction are distributed among them. The SHAP method estimates the Shapley value of each feature as its contribution to the prediction. The Shapley value is the only explanation method with a solid theory (Young 1985) that satisfies symmetry, local accuracy (also known as additivity), and consistency, properties not pertinent to other XAI methods.

4. Performance evaluation

We break down model performance into two parts. First, we present summary statistics of model performance to provide an overview of how each model performed. Then, we present examples of what the models got right and wrong.

Table 2. Performance scores of the baseline logistic regression and ResNet models on the testing dataset at lead times of 10-, 30-, and 60-min. The probability threshold was selected to generate the optimal CSI for the validation dataset. Bold numbers indicate the best value for each metric.

Model	Lead time (min)	AUC	CSI	FAR	BSS
Logistic Regression	10	0.758	0.666	0.302	0.238
ResNet	10	**0.855**	**0.723**	**0.236**	**0.382**
Logistic Regression	30	0.675	0.612	0.343	0.107
ResNet	30	0.744	0.647	0.328	0.182
Logistic Regression	60	0.629	0.587	0.385	0.051
ResNet	60	0.677	0.607	0.375	0.058

a. General performance

The primary performance statistics at lead times of 10-, 30-, and 60-min are presented in Table 2. The ResNet substantially outperforms the baseline logistic regression model in terms of the AUC score (Table 2). While the AUC score is considered overly optimistic and less informative for rare event predictions because it weighs positive and negative events equally (Flora et al. 2021; Leinonen et al. 2022), it's still a useful metric for our study because our testing dataset has been downsampled to be balanced. ROCs of POD versus POFD, which are a measure of forecast skill at different probability thresholds at lead times of 10, 30, and 60 minutes, are illustrated in Fig. 3a. Across all probability thresholds and lead times, the Resnet consistently performs better than the logistic regression model, with the magnitude of their differences diminishing as lead time increases. At the 10-min lead time, the maximum value of the Pierce skill score (PSS; defined as POD-POFD) of 0.536 for ResNet is higher than 0.381 for the logistic regression.

Forecast skill is further evaluated with a performance diagram (Fig. 3b) at lead times of 10, 30, and 60 minutes. The performance diagram contains the POD versus the SR, thereby emphasizing a model's ability to predict positive events while ignoring correct negative events (Roebber 2009). Frequency bias (black dashed lines) and CSI (filled contours) are also displayed in the performance diagram. Frequency bias, the ratio of total positive forecasts to total positive events, is a measure of bias resulting from class imbalance; a balanced dataset has a value of 1. The CSI is a significant metric for severe weather prediction because events like CI and tornado occurrences hold greater importance than non-events. For the 10-min lead time, the optimal threshold that maximizes CSI (Fig. 3b, stars) on the validation dataset almost maximizes CSI on the testing dataset for both

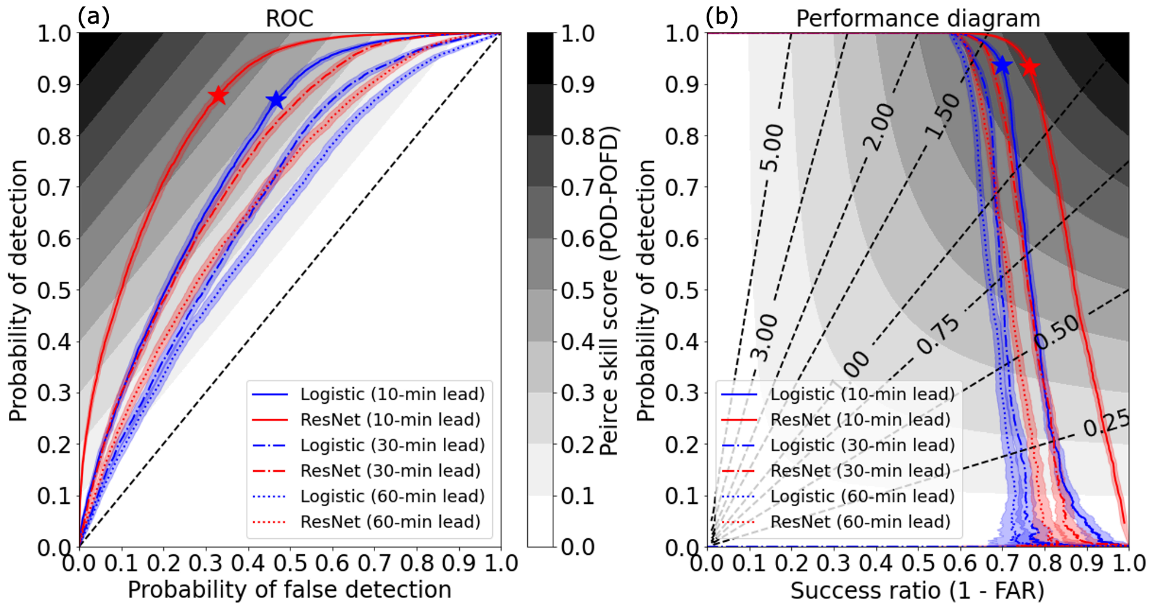

Fig. 3. Performance of the logistic regression (blue) and ResNet (red) models on the testing dataset at lead times of 10 (solid), 30 (dash-dotted), and 60 (dotted) minutes. Each curve shows the means for thresholds ranging from 0 to 1. Light shading around each line shows the 95% confidence intervals determined by bootstrapping the testing samples 1000 times. (a) ROCs with the diagonal black dashed line indicating a no-skill random classifier. Filled contours are the Pierce skill scores (PSS; defined as POD - POFD). For the 10-min lead time, the threshold that maximizes the PSS (0.536 for ResNet and 0.381 for logistic regression) on the validation data is marked by a star on each curve. (b) Performance diagrams with the black dashed lines representing the frequency bias. Filled contours are the critical success indices (CSIs). For the 10-min lead time, the threshold that maximizes the CSI on the validation data is marked by a star on each curve.

models, suggesting that general characteristics, like class proportion and input feature distribution, of the validation and testing datasets are highly consistent. At the optimal threshold, the ResNet demonstrates a POD above 0.90, which is significantly higher than the SR. This indicates that the ResNet model's performance is primarily influenced by false alarms rather than misses. Brooks and Correia (2018) have found that for rare event forecasts, achieving a certain amount of decrease in false alarms requires a much larger increase in misses, consistent with our results. Given that both the ROC and performance diagrams show separation between the logistic regression and ResNet models, we conclude that the ResNet is utilizing additional information in the data that the logistic regression does not. The ResNet produces a higher CSI with a lower FAR than the baseline logistic

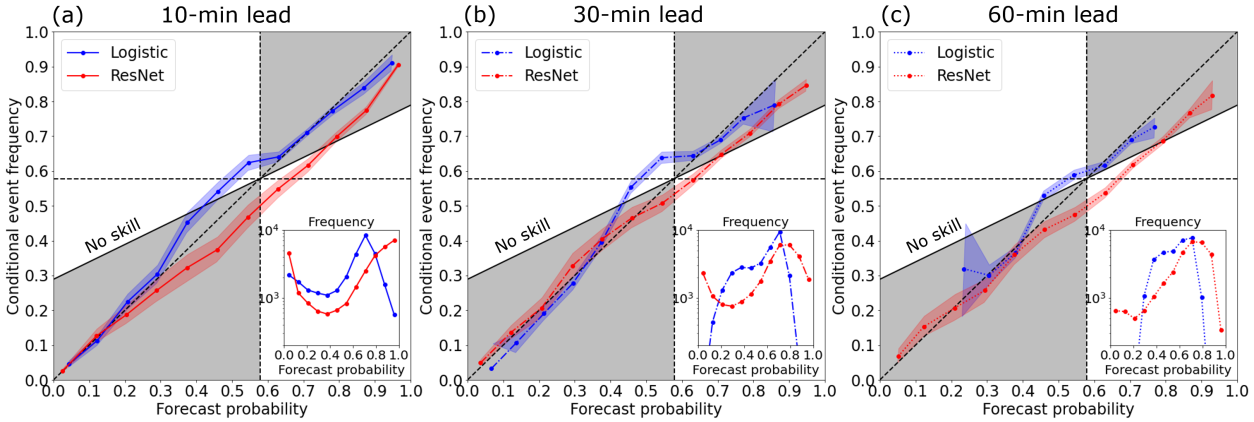

Fig. 4. Attribute diagrams for the logistic regression (blue) and ResNet (red) models on the testing dataset at lead times of (a) 10, (b) 30, and (c) 60 minutes. Each curve shows the means achieved over all thresholds from 0 to 1. Light shading around each curve shows the 95% confidence intervals determined by bootstrapping the testing samples 1000 times. The diagonal black dashed line indicates perfect reliability, and the horizontal black dashed line represents the climatological event frequency. The gray shaded areas indicate regions where points on the curves produce positive BSSs, whereas the white areas indicate regions where points on the curve generate negative BSSs. The inset panel shows the binned frequencies of the forecast probabilities for each model.

regression model (Table 2) at the 10-min lead time. As the lead time increases, the advantage of ResNet over logistic regression gradually decreases. High FAR is a noticeable issue in previous satellite-based CI nowcasting algorithms (Mecikalski et al. 2015). These results indicate that the additional complexity of the ResNet model encodes localized spatial features that help to reduce FAR.

The BSS of ResNet is better than for the logistic regression model (Table 2) at the 10-min lead time. Elements of the reliability and resolution terms in the BSS at lead times of 10, 30, and 60 minutes are illustrated in the attribute diagrams of Fig. 4, which show the conditional event frequency against the forecast probability. Reliability measures how close the forecast probabilities are to the observed frequency at the corresponding probabilities, while resolution measures how the conditional event frequencies differ from the climatological event frequency. The two gray areas represent the regions of positive BSSs, where the resolution term is greater than the reliability term. At all lead times, the reliability curves of both models are close to the perfect reliability curve (Fig. 4, diagonal black dashed line), and thus their differences in the reliability term are relatively small. Consistently across all lead times, the ResNet has an over-forecasting bias at probabilities over 0.4,

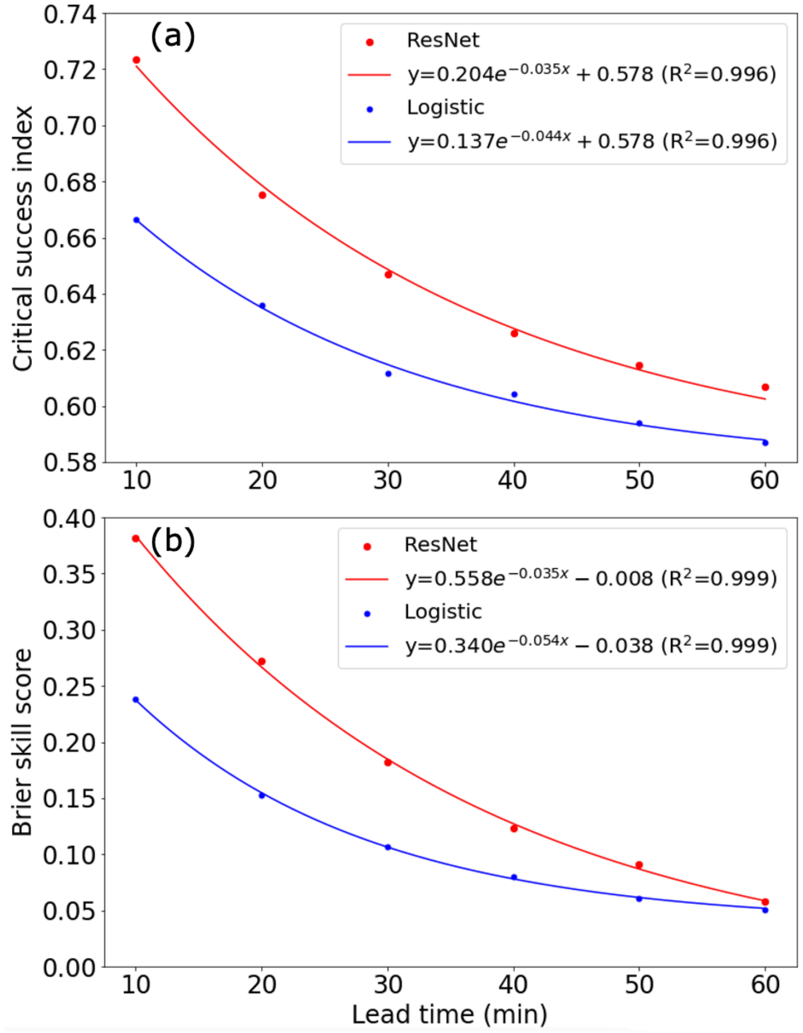

FIG. 5. (a) CSIs and (b) BSSs for the ResNet (red) and logistic regression (blue) models at lead times from 60 min to 10 min in 10-min steps. Each solid line represents an exponential fit using $y = ae^{-\lambda x} + b$ (b is preset to 0.578 for fitting CSI). R-squared (R^2) values are also provided to assess the goodness of fit.

whereas the logistic regression model consistently displays an under-forecasting bias starting at probabilities around 0.4. At the 10-min lead (Fig. 4a), the ResNet has more forecasts at probabilities less than 0.2 and greater than 0.8 compared to the logistic regression model, leading to a higher resolution term compared to the logistic regression model. Thus, the higher BSS of ResNet at the 10-min lead time is mainly attributable to the sharper forecast probability distribution than for the logistic regression model. As the lead time increases (Fig. 4b-c), the number of forecasts from both models at probabilities less than 0.2 and greater than 0.8 decreases, signifying a reduction in the resolution term at longer lead times.

The skills of both models are further evaluated at additional lead times. Figure 5 shows CSIs and BSSs at lead times from 60 min to 10 min in 10-min steps along with exponential fits $y = ae^{-\lambda x} + b$ to them. The performances of both models degrade with increasing lead time as expected. For both models, the high correlation coefficients between both CSIs and BSSs and their exponential fits indicate that both scores decrease exponentially with lead time. With increasing lead time, CSIs for both models trend asymptotically towards a lower limit close to the CSI (0.578) of the climatological forecast using the event frequency of 0.578. BSSs for both models approach 0 asymptotically with increasing lead time, suggesting that both models degrade towards climatology. ResNet performance decreases much faster than logistic regression for both scores, and their differences largely disappear at longer lead times. Thus, local spatial features, as the major advantage of the ResNet, likely become less important with increasing lead time.

b. Example cases

We chose three characteristic examples of good and poor forecasts from the ResNet at lead times from 60 min to 10 min. The ResNet model, as the examples will show, is sensitive to a variety of spatial features, including water vapor amounts and the location of clouds. The probabilistic predictions of the logistic regression model are also included to explain the skill differences between the ResNet and the logistic regression models at different lead times.

The first example in Fig. 6 shows a hit case, a growing cloud object matched to a CI event with high probabilities (close to 1) for both models at the lead time t = −10 min. The CI probability for the logistic regression is low at t = −60 min likely because of the relatively homogeneous water vapor amounts in the scene. In contrast, the CI probability for the ResNet is high, probably due to the weak pre-cloud signal observed in the bottom of the scene in the 10.35-μm BT. The CI probability for the ResNet remains high when the cumulus cloud seed first appeared in the 10.35-μm BT at t = −50 min. From t = −50 min to −20 min, the probability for the ResNet increases gradually to 0.993, while moisture convergence in the lower and middle troposphere likely intensified and the cumulus cloud grew transporting hydrometeors upward to colder temperatures. At t = −20 min, the much cooler cloud-top BTs of the cloud object indicate the occurrence of hydrometeors here,

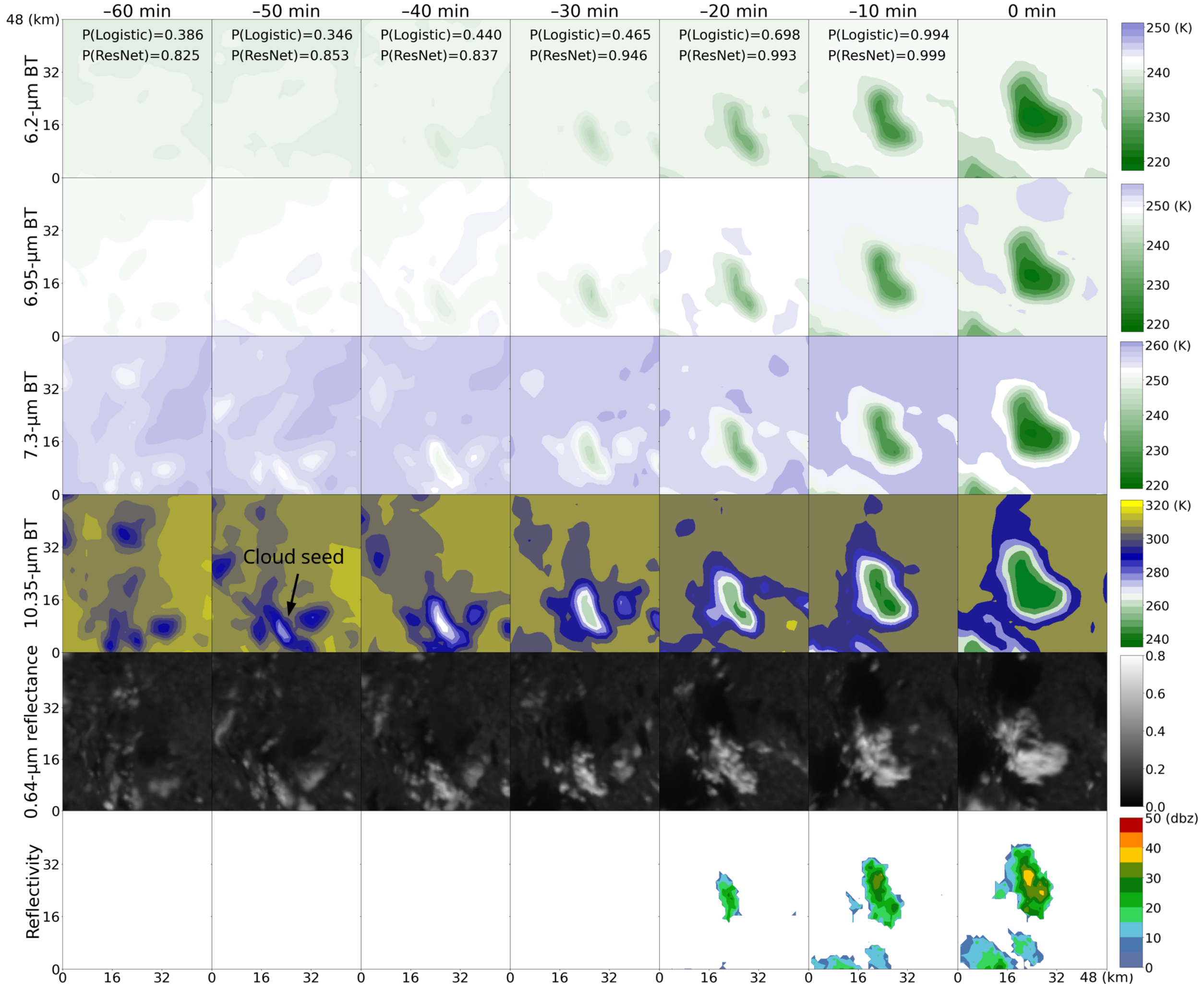

Fig. 6. A hit example of a growing cloud object matched to a CI event at 2259 UTC 7 June 2021 with high probabilities for the ResNet and the logistic regression models at a 10-min lead time. The top four rows of images show the observed GOES-16 6.2-, 6.95-, 7.3-, and 10.35-μm BT, respectively, at lead times indicated on the top of each column. The fifth row of images shows the observed GOES-16 0.64-μm reflectance for visual interpretation. The bottom row of images shows the observed MRMS composite reflectivity. The probabilistic predictions of the ResNet and the logistic regression models at different lead times are indicated just underneath the lead times at the top. The annotation highlights the cloud object seed associated with CI.

consistent with the observed 20-dBZ reflectivity. From t = −20 min to t = −10 min, the cumulus cloud expanded, moved towards the center, and grew deeper with more moisture transported from the lower troposphere to the upper troposphere. The location and coverage of the clouds shown in the infrared observations match well with the radar reflectivity and visible reflectance observations.

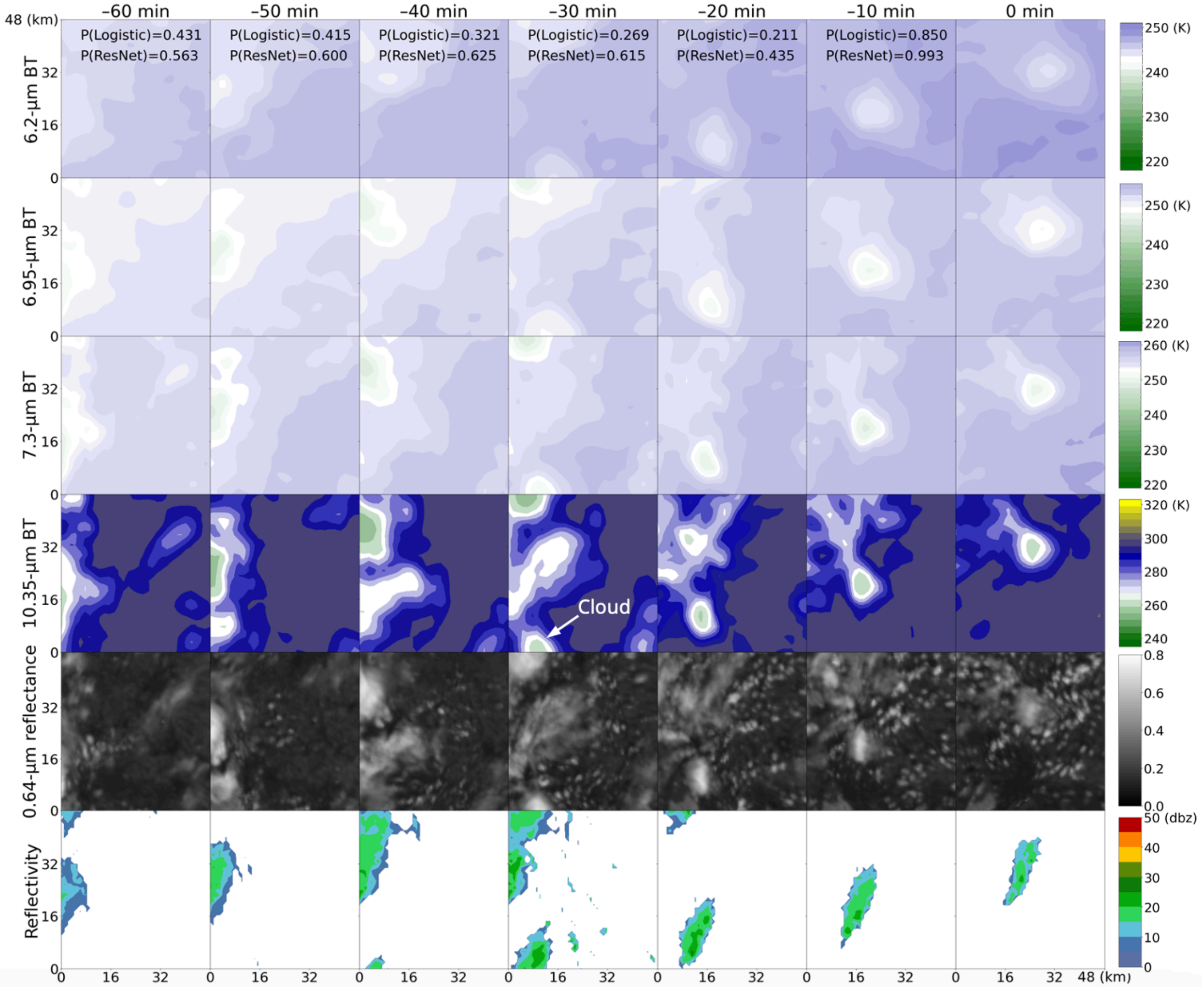

FIG. 7. As in Fig. 6, but for a false alarm example of a moving cloud object not matched to any CI event at 1942 UTC 25 June 2021 with high probabilities for both models at a 10-min lead time. The annotation highlights the preexisting cloud object associated with high probabilities for CI.

This example demonstrates that the ResNet is sensitive to the location, height, and coverage of the clouds, and water vapor amounts at different heights. Before condensed water forms, the ResNet is perhaps using water vapor features indicating what will become the cloud object to predict CI. In contrast, CI probability for the logistic regression model is less than 0.5 for lead times from 60 min to 30 min, quickly increasing from t = −30 min to t = −10 min. Thus, the logistic regression model is likely sensitive to the lowest BTs in all channels and the number of cold BTs within the cloud object. Note that these speculations on features leading to the evolution of CI probability for

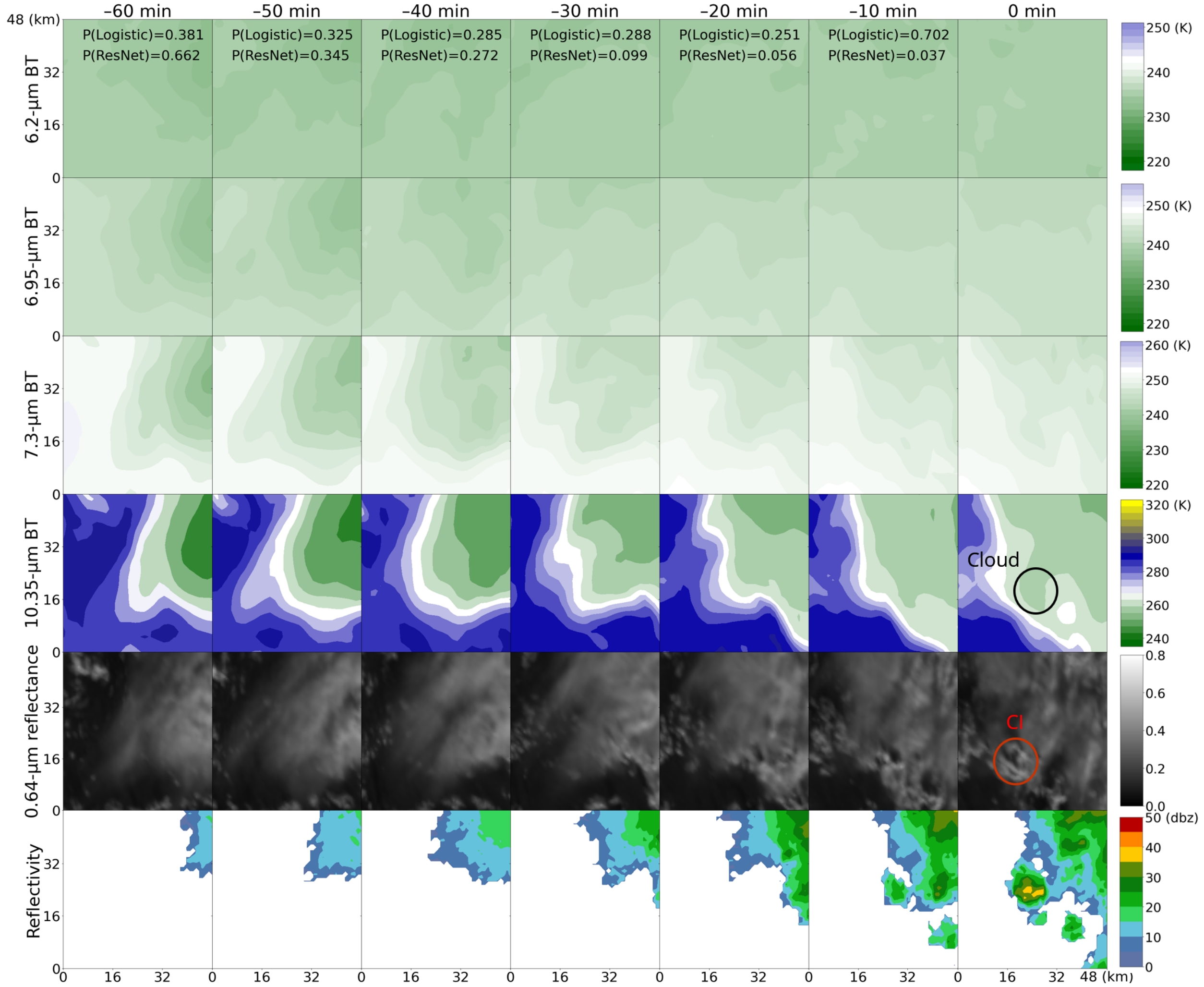

FIG. 8. As in Fig. 6, but for a miss example of a CI event at 0113 UTC 28 June 2021 obscured by cirrus clouds with a low probability for the ResNet and a high probability for the logistic regression model at a 10-min lead time. The annotation highlights the weak signature associated with the upcoming CI beneath the cirrus cloud. The black circle in the 10.35-μm BT indicates the cloud object associated with CI, while the red circle in the 0.64-μm reflectance indicates the visual CI signatures.

ResNet is generally consistent with the model explanation results of the hit case shown in Fig. A1. Details of model explanation methods are introduced in section 5.

In contrast to the first example, Fig. 7 shows a false alarm example of a moving cloud object not matched to any CI event with high probabilities for both models at a 10-min lead time. Between t = −60 min and t = −30 min, a preexisting cloud object covered the upper left quadrant of the scene, corresponding to moderate CI probabilities around 0.6 for the ResNet. From t = −30 min to t = −10 min, a second cloud object moved from the bottom left of the scene to its center. The high probability for the ResNet at t = −10 min indicates the ResNet is sensitive to the location and the low 10.35-μm BT values of cloud objects. The low probability for the ResNet at t = −20 min might be associated with the limited cloud coverage and the location of the cloud away from the center. The speculations are consistent with the model explanation results of this false alarm case in Fig. A2. The increase of the probability for the logistic regression model from t = −20 min to t = −10 min indicates that it might be sensitive to the lowest BTs in the lower and middle troposphere. The slight increase in the cloud-top temperature of the second cloud object from t = −20 min to t = −10 min in the 10.35-μm BT implies that temporal variations, not encoded in this study, might be useful for reducing false alarms for this and similar cases.

Unlike the previous two examples, Fig. 8 shows a miss example of a CI event obscured by cirrus clouds. The cirrus anvil from a preexisting storm obscured a large region of the scene starting from the top right at t = −60 min. Signatures associated with CI are obscured by the thick cirrus clouds until t = 0 min. The CI probability for the ResNet is below 0.3 from t = −30 min to −10 min, probably due to the almost homogeneous water vapor amount in the lower troposphere (see 7.3-μm BTs). These speculations are consistent with the model explanation results of this miss case in Fig. A3. Note that the model explanation of these three cases (Fig. A1, A2, and A3) suggests that water vapor in the upper troposphere (see 6.2-μm BTs) has a minor contribution to CI probabilities compared to the features in the other channels. Interestingly, the logistic regression model produced a high CI probability over 0.6 at t = −10 min. The logistic regression model is likely sensitive to the number of cold BTs. In pre-CI environments, it's common for growing cumulus clouds in the lower troposphere to be obscured by cirrus anvils from pre-existing convection. However, previous CI nowcasting algorithms (Mecikalski et al. 2015; Apke et al. 2015) focused on predicting CI for cumulus cloud objects identified from satellite observations, ignoring CI events whose radiative signatures were partially or completely obscured by cirrus clouds. While the ResNet is less skillful for the cases obscured by cirrus clouds, this example indicates that temporal variations of cloud-top BT might be essential for enhancing the predictive skill of similar cases, as done in prior studies (e.g., Nisi et al. 2014; Mecikalski et al. 2015; Cintineo et al. 2020b).

These manually selected examples do not cover the wide variety of cases in the testing dataset and the real world. However, they demonstrate how the models respond to different environments

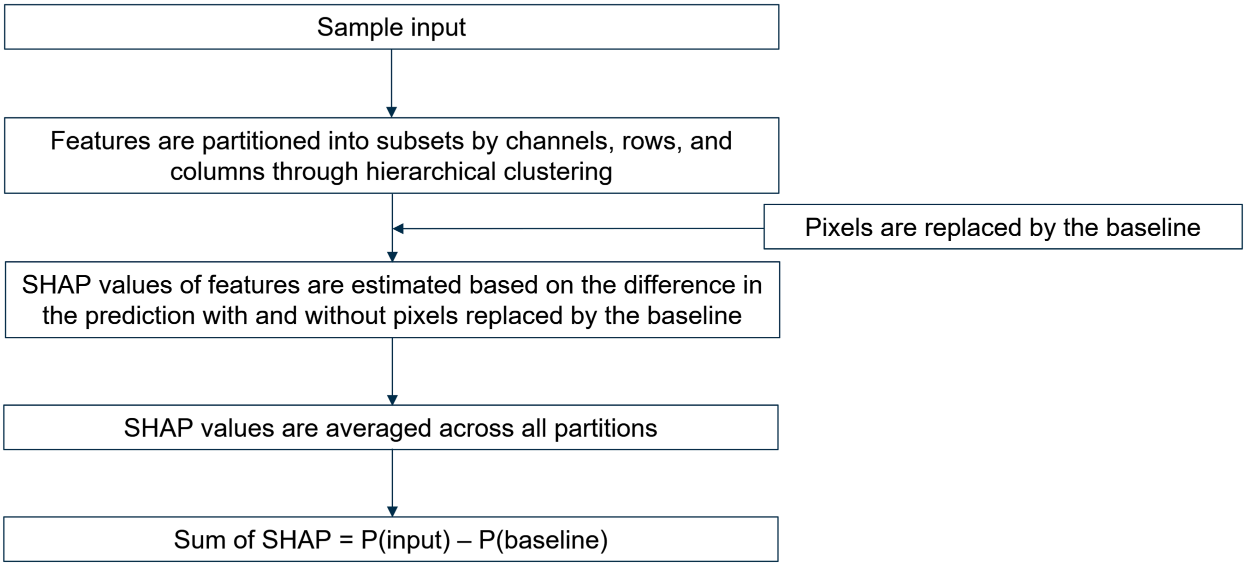
Sample input
Features are partitioned into subsets by channels, rows, and columns through hierarchical clustering
Pixels are replaced by the baseline
SHAP values of features are estimated based on the difference in the prediction with and without pixels replaced by the baseline
SHAP values are averaged across all partitions
Sum of SHAP = P(input) – P(baseline)

FIG. 9. Flowchart for the PartitionSHAP method. Each box represents an individual action.

at different lead times and how to overcome clear limitations in the current methodology, thereby motivating specific improvements in future work.

5. Model explanation

In order to explain features encoded within the ResNet model, SHAP values were estimated using PartitionSHAP by Krell (2021). The procedure of PartitionSHAP is shown in Figure 9. PartitionSHAP applies feature partitioning to explain the contribution of features to the prediction. Each sample input is recursively divided by channels, rows, and columns to generate a partition tree. Correlated features are grouped into superpixels by the partition. For each partition, the SHAP values, a measure of the contribution to the prediction, are estimated for superpixels while considering interactions within the superpixels via the difference between the model's predictions for a specific sample with and without pixels replaced by the baseline. Then, the SHAP values are averaged across partitions to obtain a single set of SHAP values as the measure of the feature attribution. Given the additivity property of SHAP values, the sum of SHAP values for all features approximates the prediction difference between the sample input and the baseline. Features with higher positive SHAP values have a larger positive contribution to the prediction difference, while features with higher negative SHAP values have a larger negative contribution to the prediction difference.

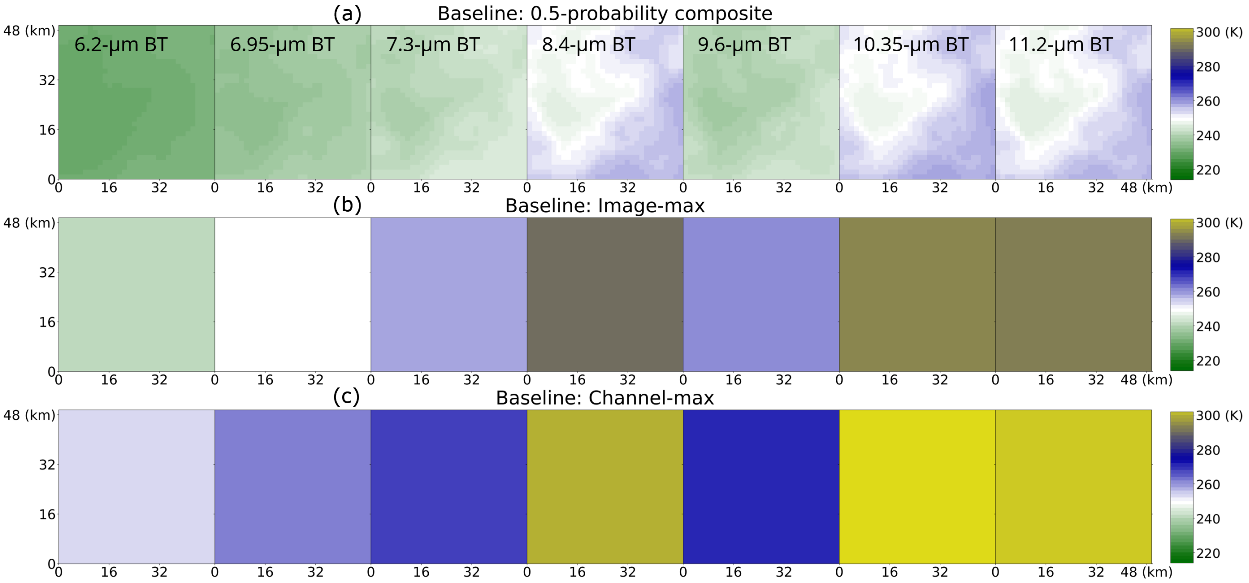

FIG. 10. Baselines used in SHAP calculations. (a): 0.5-probability composite baseline is the PMM composite of 100 0.5-probability samples with a specific percentile range from 36.6% to 63.4% to uphold the composite's 0.5 probability. (b): The ensemble mean of the 100 image-max baselines for the 100 cases. For each case, the image-max baseline comprises seven 32×32 channel-specific uniform arrays representing the maximum BT found in the image of each channel. (c): The channel-max baseline comprises seven 32×32 channel-specific uniform arrays representing the maximum BT found across all 100 images for each channel.

SHAP output is a heat map overlaid on the deviation of the input from the baseline to reveal the additive contribution of each pixel to the prediction difference between the input and the baseline. The baseline is a reference against which changes in predictions using sample inputs are interpreted by comparison. Mamalakis et al. (2022) demonstrated that model explanation is highly dependent on the baseline and different baselines can be used to answer different science questions. In our study, we computed the SHAP values for 100 individual cases and showed both the composite SHAP and deviation to demonstrate their correlation. We first explored the important radiative features behind the 100 best hit cases, CI cases with probabilities close to 1, using three different baselines (Fig. 10). The first baseline is a composite of 0.5-probability samples (Fig. 10a, hereafter 0.5-probability composite), which represents a relatively moist environment. Here, we used a probability-matched mean (PMM; Ebert 2001) composite of 100 0.5-probability samples with a specific percentile range of BTs from 36.6% to 63.4% to uphold the composite's 0.5 probability. PMMs preserve spatial structures better than simply taking the mean of inputs. The large towering

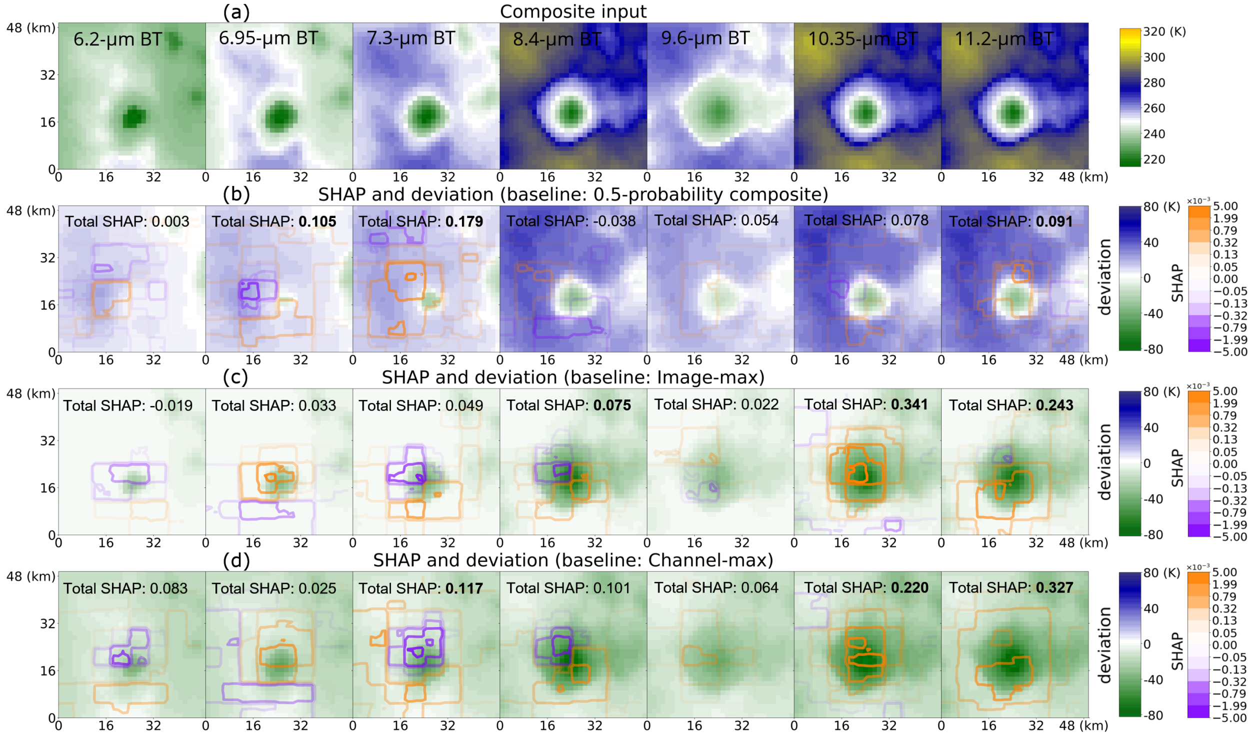

FIG. 11. (a): PMM BT composite inputs for seven infrared channels generated from the 100 best hit cases at a 10-min lead time. (b-d): Composite SHAP heat maps (contours) and composite deviations (shading) of the inputs from the (b) 0.5-probability composite, (c) image-max, and (d) channel-max baselines. All of the composites are created by applying the PMM method to the images from the 100 best hit cases. Positive SHAP values indicate positive contributions to the prediction difference between the input and the baseline, whereas negative SHAP values indicate negative contributions. The total SHAP of each channel is indicated at the top of each image. The top three maximum total SHAPs for each baseline across channels are highlighted in bold.

cumulus cloud near the center of Fig. 10a is consistent with a moderate CI probability generated by the ResNet. We call the second baseline the image-max baseline (Fig. 10b), a typical clear-sky baseline. For each case, the image-max baseline comprises seven 32×32 channel-specific uniform arrays representing the warmest BT found in the image of each channel. Figure 10b shows the ensemble mean of the image-max baselines over the 100 cases. Finally, the channel-max baseline (Fig. 10c), a dry clear-sky baseline, comprises seven 32×32 channel-specific uniform arrays representing the warmest BT found across all 100 images for each channel.

Figure 11 shows composite inputs of the 100 best hit cases (Fig. 11a) and the SHAP and deviation values for the three baselines (Fig. 11b-d). For each baseline, the top three maximum total SHAPs across the seven infrared channels are highlighted in bold, indicating the channels that receive the highest attention from the ResNet model. The best hits composite (Fig. 11a) is characterized by a cloud object near the center with moisture accumulated in the lower (7.3-μm BTs) and middle (6.95-μm BTs) troposphere, localized moisture accumulation in the upper troposphere (6.2-μm BTs), and cloud coverage observed in the window channels (10.35-/11.2-μm BTs). With the 0.5-probability composite baseline, the question to be addressed is as follows: "Which features made the model predict CI compared to a relatively moist environment with a cloud object?" Based on SHAP results (Fig. 11b), CI probability mainly comes from moisture gradients between the cloud object and the environment in the lower troposphere (7.3-μm BTs), moisture gradients surrounding the cloud object in the middle troposphere (6.95-μm BTs), and the BT gradient at cloud boundaries observed in 11.2-μm BTs. Negative contributions are mainly from the drier areas in the middle troposphere near the center. Thus, the results indicate that the model has learned that relative to the 0.5-probability composite baseline, CI is mostly determined by moisture gradients near the cloud object, possibly associated with moisture convergence, in the lower and middle troposphere.

We then use the image-max baseline to answer the following question: "Which features made the model predict CI compared to a typical clear-sky environment?" The SHAP results (Fig. 11c) highlight positive contributions mainly from the cloud-top height observed in 10.35-μm BTs, gradients at cloud boundaries in 11.2-μm BTs, and cloud-top glaciation near the cloud object in 8.4-μm BTs. While weak positive contributions arise from moisture gradients in the vicinity of the cloud object in the lower troposphere (7.3-μm BTs), they are largely counterbalanced by negative contributions from the central regions. These negative SHAPs might arise from the model's knowledge gained from other CI cases, especially those obscured by cirrus clouds. With the image-max baseline, the ResNet is more focused on cloud-top height and glaciation as well as cloud coverage. These critical features behind CI forecasts are consistent with previous studies (Mecikalski et al. 2011; Han et al. 2019).

Model explanation is further explored with the channel-max baseline to gain insights on the following question: "Which features made the model predict CI as opposed to a dry clear-sky environment?" The SHAP results (Fig. 11d) highlight the positive contributions from cloud-top height (10.35-/11.2-μm BTs), cloud coverage (10.35-/11.2-μm BTs), and the moisture environment in the lower troposphere (7.3-μm BTs). More positive contributions stem from the moisture within

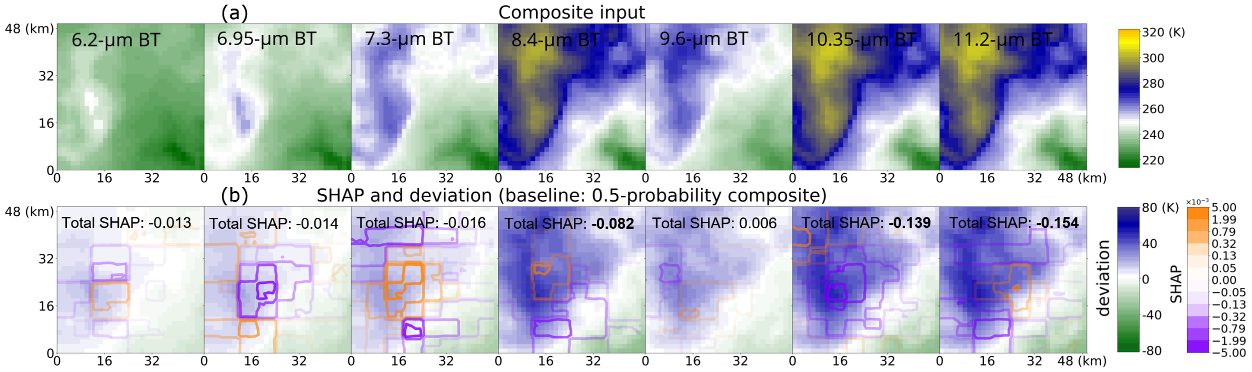

FIG. 12. (a) PMM composite of inputs over 100 worst miss cases and (b) the SHAP and deviation values relative to the 0.5-probability composite baseline.

the lower troposphere compared to the results with the image-max baseline, possibly attributed to a generally higher contrast of moisture content in the environment. The comparison demonstrates that the model prediction is not only dependent on the characteristics of cloud objects, but also takes environmental moisture into account.

We further explored the model explanation using the 0.5-probability composite baseline on the 100 worst miss cases. We aim to understand why the model missed some CI events through the SHAP analysis. Figure 12 displays the PMM composite of inputs over the 100 worst miss cases and the SHAP and deviation values from the 0.5-probability composite baseline. According to the composite input (Fig. 12a), areas in the upper-left are in clear-sky conditions whereas in the lower-right area it is cloudy. These conditions might obscure signals like moisture gradients in the lower and middle troposphere. According to the SHAP and deviation values (Fig. 12b), the model generated positive contributions from moisture gradients in the lower troposphere (7.3-μm BTs) in the central region. Negative contributions stem from the wide clear-sky areas in the window channels (10.35-/11.2-μm BTs) and the cloud-top temperature gradient near the cloud boundaries observed in 8.4-μm BTs. Combined with the third case example (Fig. 8), we hypothesize that the ResNet fails to generate correct forecasts for these misses likely because the signatures in the middle and lower troposphere were blocked by the anvil clouds from preexisting convection. Surface observations and satellite derived cloud property data would help infer these missing signatures from convergence and differential heating near the surface (Weckwerth et al. 2011; Mecikalski et al. 2013).

In the Appendix, we further explored the model explanation on worst false alarm and best correct null cases. The worst false alarms results (Fig. A4) are similar to the best hits results (Fig. 11), suggesting that the model might be unable to distinguish between growing convective clouds and mature or nonconvective clouds. The best correct null results (Fig. A5) are similar to the worst misses results (Fig. 12), highlighting the negative contributions from the clear-sky regions.

6. Discussion and Conclusions

Convective initiation nowcasting from satellite observations has proved challenging for existing algorithms, yet this work demonstrates improvements in forecast skill and explainability. We presented a data-driven method for CI nowcasting at lead times up to 1 hour using a ResNet architecture for encoding spatial features of GOES-16 satellite observations. The ResNet model was compared against the classical logistic regression model to evaluate improvements to skill added by spatial encodings. The ResNet model significantly outperforms the logistic regression model in multiple evaluation metrics at lead times up to 1 hour, especially for the false alarm ratio. However, improvements in prediction skill via encoding of spatial features quickly decreases with increasing lead time, indicating that spatial features associated with CI might be statistically weaker or omitted altogether for fast-moving cloud objects at longer lead times. Interestingly, the performance of both models decreases exponentially towards climatology with increasing lead time. Through case studies, we found that the logistic regression model is sensitive to the lowest BTs and the number of cold BTs, whereas the ResNet model is sensitive to the location, height, and coverage of clouds, and moisture amounts at different altitudes. We also found that the ResNet model is unable to correctly forecast CI events whose signatures are obscured by overlying cirrus anvil clouds and non-CI events associated with mature or nonconvective clouds.

We suggest that model explanation answers different science questions based on the choice of baseline. We employed the PartitionSHAP method to better estimate contributions from feature interactions. With the 0.5-probability composite baseline, a moist baseline with a moderate CI probability, CI is mostly determined by moisture gradients near the cloud object, possibly associated with moisture convergence in the lower and middle troposphere. With the image-max baseline, a typical clear-sky baseline, the model focused attention on cloud-top height and glaciation as well as cloud coverage. With the channel-max baseline, a dry clear-sky baseline, contributions from cloud-top height, cloud coverage, and the moisture environment in the lower troposphere were emphasized. Our study demonstrates the advantage of using different baselines in further understanding the ResNet model's decision-making processes and gaining potential scientific insights. The explanation results on worst miss cases indicate that the failure of the model in these instances is likely caused by an inability to detect signatures in the lower and middle troposphere due to obscuration by preexisting upper level clouds.

Though this work is on only a single component of an envisioned operational CI forecasting system based on ML methods, it is an important one in demonstrating extraction of the physical processes encoded in the model that impact CI forecasting. Subsequent work will focus on incorporating predictors across multiple timesteps and additional meteorological information into CI nowcasting with an emphasis on forecast skill, understanding the encoded physical processes, and operational resilience.

While these results are promising, there are some limitations that must be considered. First, our dataset might omit cloud signals associated with CI for fast-moving cloud objects at longer lead times. Based on our estimates, the patch size (48-km by 48-km) is able to capture CI-related cloud features of moving cloud objects in most conditions, but it might not sufficiently capture fast-moving cloud objects, as exemplified in Fig. 7. Second, we didn't track cumulus cloud objects or use a cloud-following patch. Given the high spatiotemporal resolution of GOES-16 observations, the tracking of cumulus cloud objects using satellite images is feasible and has been done in previous studies (Mecikalski et al. 2015; Han et al. 2019), even for those cumulus cloud objects obscured by cirrus clouds (Mecikalski et al. 2013). Third, our study is focused on CI in the U.S. Great Plains region and the findings, both evaluation and model explanation results, might be biased by CI processes of the region. Fourth, because non-CI events have been downsampled to be comparable to the number of CI events to make a balanced dataset, the class proportions of our dataset are different from realistic class proportions in the real atmosphere. Thus, performance evaluation against climatology, like the BSS score, might not be reliable. Fifth, the model explanation is still affected by interactions between correlated features. Although PartitionSHAP was initially designed to better estimate contributions from the interactions between features, the results, especially the negative SHAP values, are still affected by interactions of localized features. Feature correlation might have been encoded into the model during training. The model might have learned how to utilize correlated features to maximize its skill. For example, the difference between 8.4- and 10.35-μm BTs is usually used to provide information about cloud-top glaciation. ResNet might have encoded this signature in inferring the timing of CI.

Acknowledgments. The authors thank David J. Stensrud and Yunji Zhang for discussions about CI definition and identification. This research benefited from the National Center for Atmospheric Research's (NCAR's) Computational and Information Systems Laboratory visitor program. The machine learning training, evaluation, and explanation were performed on an Institute for Computational and Data Sciences supercomputer provided by Penn State and the Computational and Information Systems Laboratory's Cheyenne supercomputer provided by NCAR. This research was partially funded by the Penn State College of Earth and Mineral Sciences, a Penn State Institutes of Energy and the Environment seed grant, and the National Science Foundation (NSF) under Grant No. ICER-2019758. This material is based upon ongoing work at NCAR, which is a major facility sponsored by the NSF under Cooperative Agreement No. 1852977. Author contributions are as follows: conceptualization: D.F., D.J.G., and S.J.G.; methodology: D.F., D.J.G., and S.J.G.; investigation: all authors; supervision: S.J.G. and D.J.G.; writing—original draft: D.F.; and writing—review and editing: all authors. The authors thank three anonymous reviewers for their thoughtful reviews and comments on this manuscript.

Data availability statement. The machine learning and analysis software used in this paper can be accessed from the CIML library available at `https://github.com/dxf424/CIML`. Processed training and testing data are available online at (`https://doi.org/10.26208/6Y59-0R80`).

APPENDIX

a. Storm-Tracking configuration

The w2segmotionll algorithm, a WDSS-II executable, and the modified best-track algorithm are used for storm identification and tracking from radar MRMS radar dataset. Table A1 shows the configuration options for these algorithms.

TABLE A1. Configuration options used for storm tracking and identification for w2segmotionll and post-event track correction for best-track.

Parameter	Flag	Option
	Storm tracking and identification for w2segmotionll	
trackedProductName	-T	MergedReflectivity QCComposite
"min max incr maxdepth"	-d	"35 57 5 -1"
prunerSizeParameters	-p	40, 200, 300, 0:0, 0, 0
smoothing filters	-k	percent:50:1:0:1, percent:75:1:0:1
clusterIDMatchingMethod	-m	MULTISTAGE: 2:10:0
	Post-event track correction for best-track	
Buffer distance	-bd	16 (km)
Buffer time	-BT	11 (min)

b. Hyperparameter tuning

Table A2 shows the selected hyperparameters, their search space, and optimal values for the baseline logistic regression model and ResNet.

TABLE A2. Selected hyperparameters, their search space, and optimal values for the baseline logistic regression model and ResNet.

Hyperparameter	Search space	Optimal value
	Logistic regression	
C	0.0001-1.0 (log-uniform)	0.212
λ (mixing parameter)	0.0001-1.0 (log-uniform)	0.104
	ResNet	
Leaky alpha	0.0-0.4	0.138
Learning rate	0.0000001-0.001 (log-uniform)	0.0009
Initial number of filters	10-100	64
Batch size	256-2048	256
Dropout alpha	0.0-0.4	0.074

c. SHAP results for case examples

Figure A1, A2, and A3 show the SHAP and deviation of 6.2-, 6.95-, 7.3- and 10.35-μm BT relative to the image-max baseline at lead times from 10 min to 60 min for the case shown in Fig. 6, 7, and 8 respectively. Note that for the false alarm case (Fig. A2), the negative SHAP related to the lower-tropospheric water vapor (7.3-μm) correctly lowered the CI probability, especially at lead times longer than 20 min.

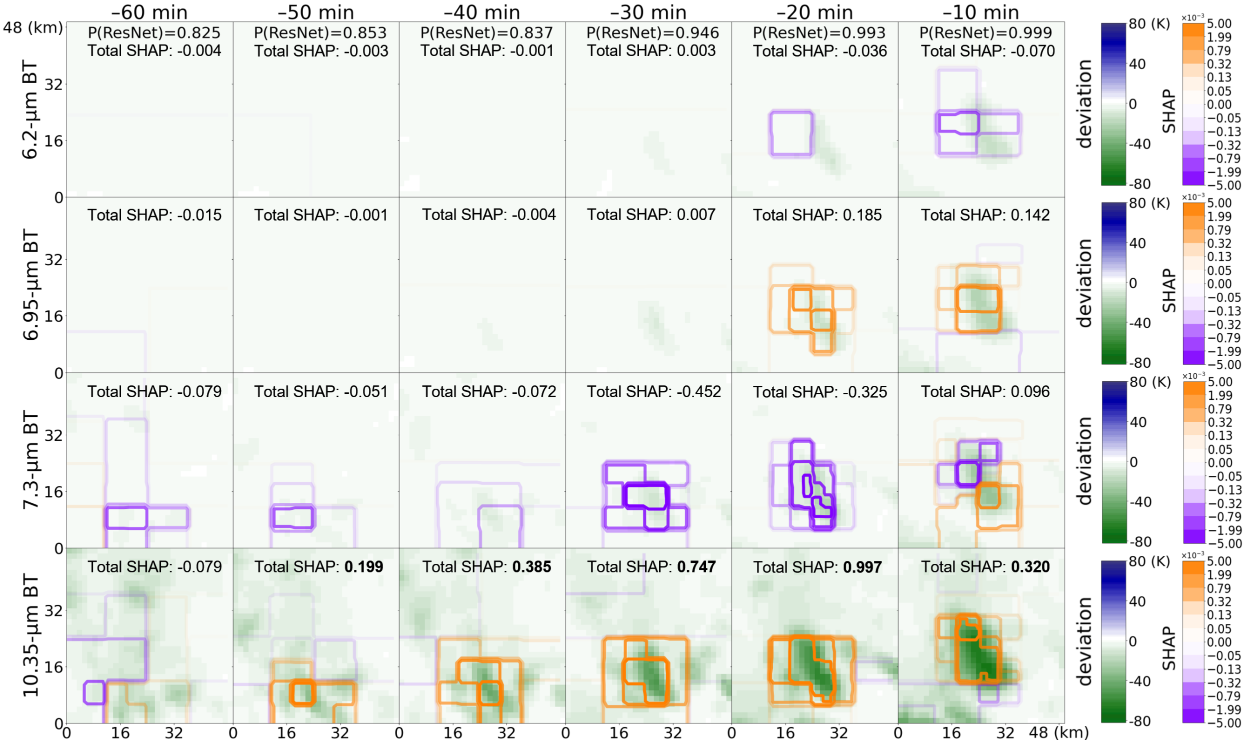

FIG. A1. SHAP and deviation of 6.2-, 6.95-, 7.3- and 10.35-μm BT relative to the image-max baseline at lead times from 10 min to 60 min for the hit case shown in Fig. 6.

d. SHAP results for false alarms and correct nulls

Figure A4 illustrates the composite inputs over 100 worst false alarm cases and SHAP results and deviation values relative to the 0.5-probability composite baseline. Similar to the best hits results (Fig. 11), the composite inputs of worst false alarms are characterized by moisture accumulated near the center in the troposphere as well as cloud coverage observed in the window channels (10.35- and 11.2-μm). The SHAP results still highlights the positive contribution from the moisture

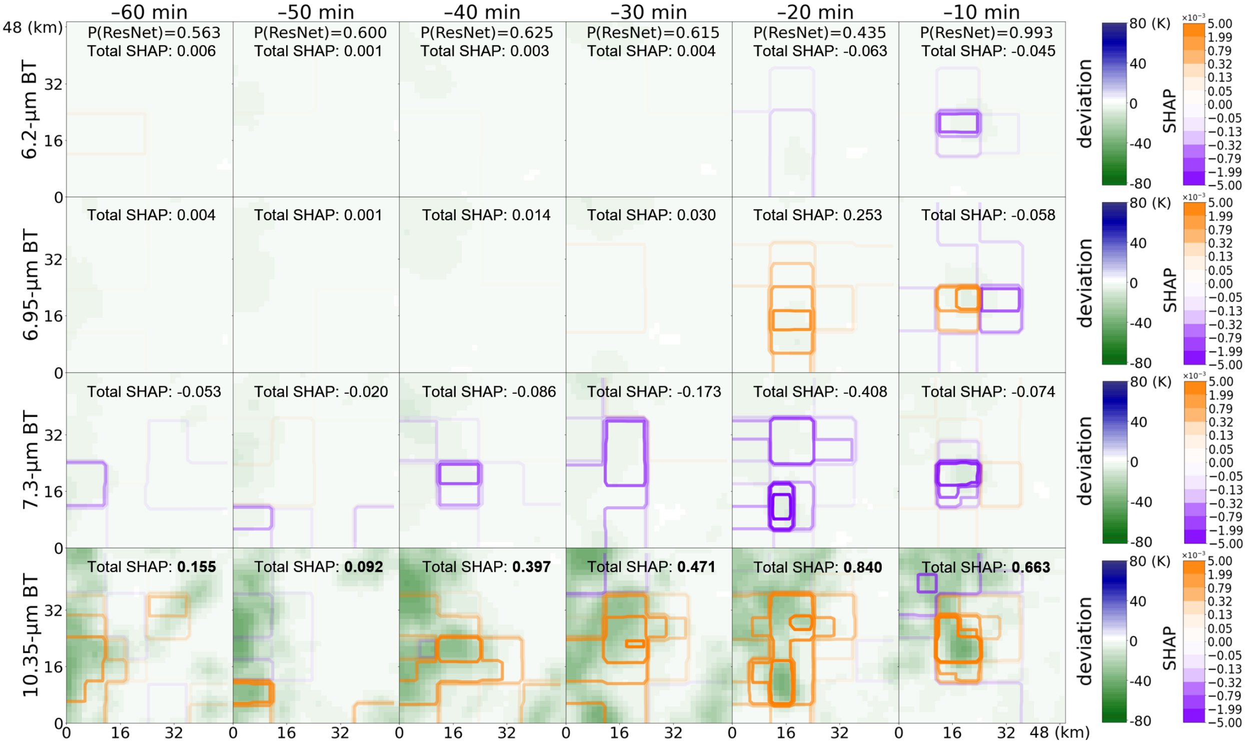

FIG. A2. SHAP and deviation of 6.2-, 6.95-, 7.3- and 10.35-μm BT relative to the image-max baseline at lead times from 10 min to 60 min for the false alarm case shown in Fig. 7.

gradients between the cloud object and the surrounding area in the lower troposphere (7.3-μm) and middle troposphere (6.95-μm) as well as the cloud-top height (10.35- and 11.2-μm). The much wider cloud coverage than the cloud of best hit cases (Fig. 11a) indicates the cloud might be a decaying mature cloud or a nonconvective cloud advected from the surroundings, consistent with the second case example (Fig. 7). Thus, the results suggest that the model might not be able to distinguish between growing convective clouds and mature or nonconvective clouds.

Figure A5 shows the composite inputs over 100 best correct null cases and SHAP results and deviation values relative to the 0.5-probability composite baseline. Similar to worst miss cases (Fig. 12), negative contributions are mostly from the wide clear-sky regions in 8.4-, 10.35-, and 11.2-μm.

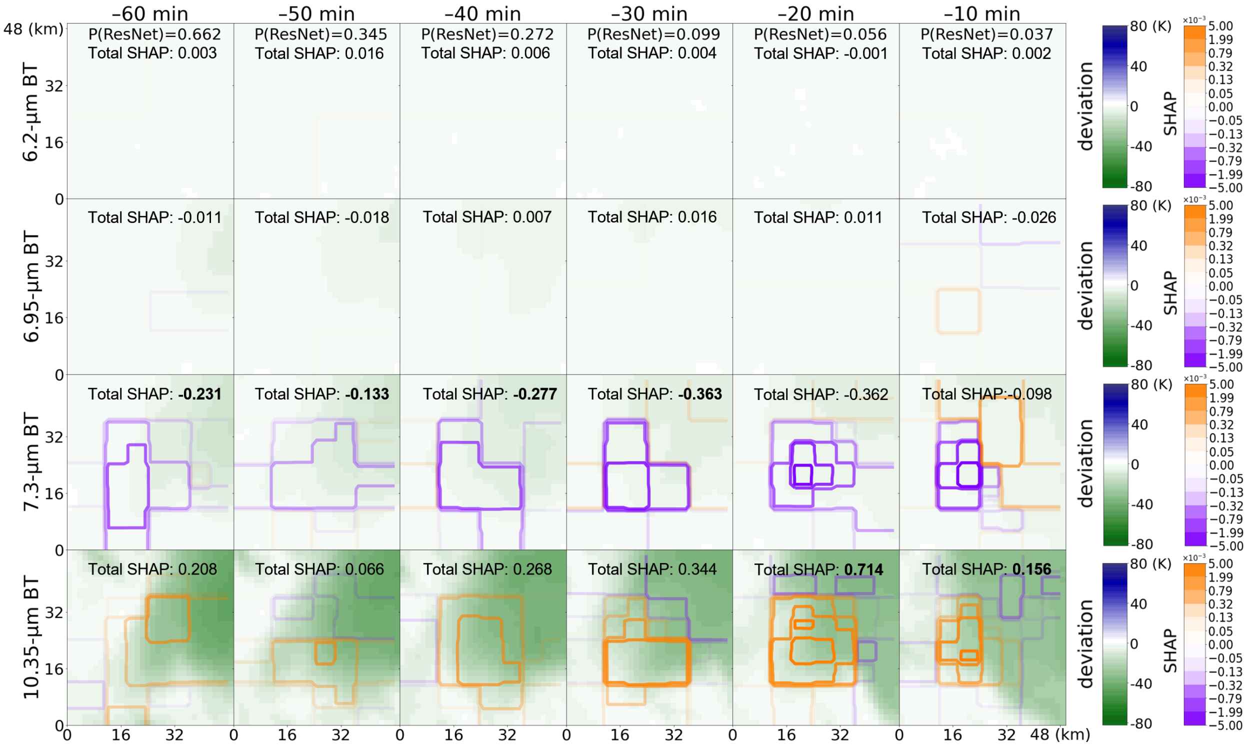

Fig. A3. SHAP and deviation of 6.2-, 6.95, 7.3- and 10.35-μm BT relative to the image-max baseline at lead times from 10 min to 60 min for the miss case shown in Fig. 8.

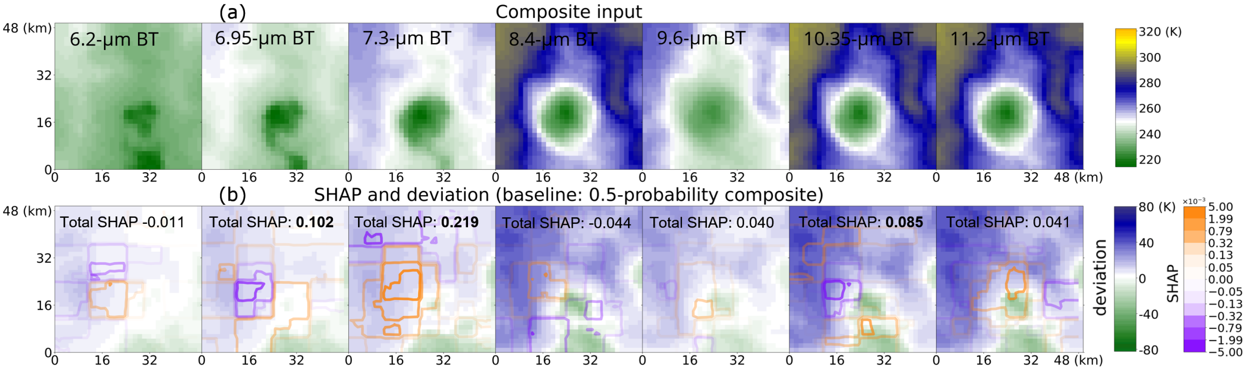

Fig. A4. (a) PMM composite of inputs over 100 worst false alarm cases and (b) the SHAP and deviation values relative to the 0.5-probability composite baseline.

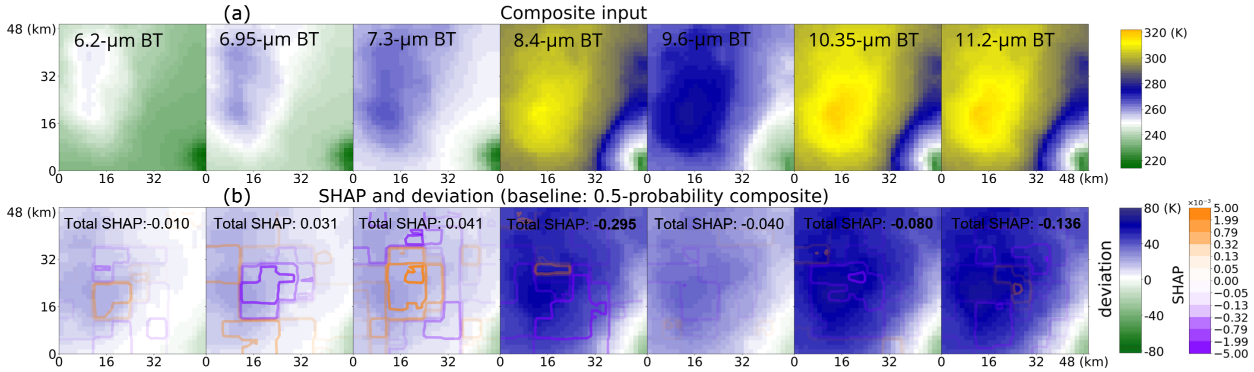

Fig. A5. (a) PMM composite of inputs over 100 best correct null cases and (b) the SHAP and deviation values relative to the 0.5-probability composite baseline.

References

Abadi, M., and Coauthors, 2016: TensorFlow: A system for large-scale machine learning. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016*, 265–283, 1605.08695.

Apke, J. M., J. R. Mecikalski, K. Bedka, E. W. Mccaul, C. R. Homeyer, and C. P. Jewett, 2018: Relationships between deep convection updraft characteristics and satellite-based super rapid scan mesoscale atmospheric motion vector-derived flow. *Monthly Weather Review*, **146 (10)**, 3461–3480, https://doi.org/10.1175/MWR-D-18-0119.1.

Apke, J. M., D. Nietfeld, and M. R. Anderson, 2015: Environmental analysis of GOES-R proving ground convection-initiation forecasting algorithms. *Journal of Applied Meteorology and Climatology*, **54 (7)**, 1637–1662, https://doi.org/10.1175/JAMC-D-14-0190.1.

Bates, B. C., A. J. Dowdy, and R. E. Chandler, 2018: Lightning prediction for Australia using multivariate analyses of large-scale atmospheric variables. *Journal of Applied Meteorology and Climatology*, **57 (3)**, 525–534, https://doi.org/10.1175/JAMC-D-17-0214.1.

Brooks, H. E., and J. Correia, 2018: Long-term performance metrics for National Weather Service Tornado warnings. *Wea. Forecast.*, **33 (6)**, 1501–1511, https://doi.org/10.1175/WAF-D-18-0120.1.

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

Brooks, H. E., and N. Dotzek, 2008: The spatial distribution of severe convective storms and an analysis of their secular changes. *Climate Extremes and Society*, **9780521870**, 35–53, https://doi.org/10.1017/CBO9780511535840.005.

Chollet, F., and Coauthors, 2015: Keras. available at: https://github.com/fchollet/keras.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, L. Cronce, and J. Brunner, 2020a: Noaa probsevere v2.0—probhail, probwind, and probtor. *Weather and Forecasting*, **35 (4)**, 1523–1543, https://doi.org/10.1175/WAF-D-19-0242.1.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, and D. T. Lindsey, 2014: An empirical model for assessing the severe weather potential of developing convection. *Wea. Forecast.*, **29 (3)**, 639–653, https://doi.org/10.1175/WAF-D-13-00113.1.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, A. Wimmers, J. Brunner, and W. Bellon, 2020b: A Deep-Learning Model for Automated Detection of Intense Midlatitude Convection Using Geostationary Satellite Images. *Weather and Forecasting*, **35 (6)**, 2567–2588, https://doi.org/10.1175/WAF-D-20-0028.1.

Cintineo, J. L., and Coauthors, 2018: The NOAA/CIMSS ProbSevere model: Incorporation of total lightning and validation. *Wea. Forecast.*, **33 (1)**, 331–345, https://doi.org/10.1175/WAF-D-17-0099.1.

Colbert, M., D. J. Stensrud, P. M. Markowski, and Y. P. Richardson, 2019: Processes associated with convection initiation in the North American Mesoscale Forecast System, version 3 (NAMv3). *Weather and Forecasting*, **34 (3)**, 683–700, https://doi.org/10.1175/WAF-D-18-0175.1.

Dixon, P. G., A. E. Mercer, J. Choi, and J. S. Allen, 2011: Tornado risk analysis: Is dixie alley an extension of Tornado alley. *Bulletin of the American Meteorological Society*, **92 (4)**, 433–441, https://doi.org/10.1175/2010BAMS3102.1.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Weather Rev.*, **129 (10)**, 2461–2480, https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Fan, D., S. J. Greybush, X. Chen, Y. Lu, F. Zhang, and G. S. Young, 2022: Exploring the Role of Deep Moist Convection in the Wavenumber Spectra of Atmospheric Kinetic Energy and Brightness Temperature. *Journal of the Atmospheric Sciences*, **79 (10)**, 2721–2737, https://doi.org/10.1175/JAS-D-21-0285.1.

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. *Monthly Weather Review*, **149 (5)**, 1535–1557, https://doi.org/10.1175/MWR-D-20-0194.1, 2012.00679.

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

Glorot, X., and Y. Bengio, 2010: Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research*, **9**, 249–256.

Han, D., J. Lee, J. Im, S. Sim, S. Lee, and H. Han, 2019: A novel framework of detecting convective initiation combining automated sampling, machine learning, and repeated model tuning from geostationary satellite data. *Remote Sensing*, **11 (12)**, 1454, https://doi.org/10.3390/rs11121454.

He, K., X. Zhang, S. Ren, and J. Sun, 2016: Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 770–778, https://doi.org/10.1109/CVPR.2016.90, 1512.03385.

Henderson, D. S., J. A. Otkin, and J. R. Mecikalski, 2021: Evaluating convective initiation in high-resolution numerical weather prediction models using GOES-16 infrared brightness temperatures. *Monthly Weather Review*, **149 (4)**, 1153–1172, https://doi.org/10.1175/MWR-D-20-0272.1.

Hoerl, A. E., and R. W. Kennard, 1988: Ridge Regression. *Encyclopedia of Statistical Sciences*, S. Kotz, Ed., Vol. 8, John Wiley and Sons, 129–136.

Ioffe, S., and C. Szegedy, 2015: Batch normalization: Accelerating deep network training by reducing internal covariate shift. *32nd International Conference on Machine Learning, ICML 2015*, **1**, 448–456, 1502.03167.

Kain, J. S., and Coauthors, 2013: A feasibility study for probabilistic convection initiation forecasts based on explicit numerical guidance. *Bulletin of the American Meteorological Society*, **94 (8)**, 1213–1225, https://doi.org/10.1175/BAMS-D-11-00264.1.

Krell, E., 2021: Partitionshap Multiband Demo. https://github.com/conrad-blucher-institute/partitionshap-multiband-demo.

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

Lagerquist, R., D. Turner, I. Ebert-Uphoff, J. Stewart, and V. Hagerty, 2021: Using deep learning to emulate and accelerate a radiative transfer model. *Journal of Atmospheric and Oceanic Technology*, **38 (10)**, 1673–1696, https://doi.org/10.1175/JTECH-D-21-0007.1.

Lakshmanan, V., B. Herzog, and D. Kingfield, 2015: A method for extracting postevent storm tracks. *Journal of Applied Meteorology and Climatology*, **54 (2)**, 451–462, https://doi.org/10.1175/JAMC-D-14-0132.1.

Lakshmanan, V., and T. Smith, 2009: Data mining storm attributes from spatial grids. *Journal of Atmospheric and Oceanic Technology*, **26 (11)**, 2353–2365, https://doi.org/10.1175/2009JTECHA1257.1.

Lakshmanan, V., and T. Smith, 2010: An objective method of evaluating and devising storm-tracking algorithms. *Weather and Forecasting*, **25 (2)**, 701–709, https://doi.org/10.1175/2009WAF2222330.1.

Lakshmanan, V., T. Smith, K. Hondl, G. J. Stumpf, and A. Witt, 2006: A real-time, three-dimensional, rapidly updating, heterogeneous radar merger technique for reflectivity, velocity, and derived products. *Weather and Forecasting*, **21 (5)**, 802–823, https://doi.org/10.1175/WAF942.1.

Lakshmanan, V., T. Smith, G. Stumpf, and K. Hondl, 2007: The warning decision support system-integrated information. *Weather and Forecasting*, **22 (3)**, 596–612, https://doi.org/10.1175/WAF1009.1.

Lawson, J. R., J. S. Kain, N. Yussouf, D. C. Dowell, D. M. Wheatley, K. H. Knopfmeier, and T. A. Jones, 2018: Advancing from convection-allowing nwp to warn-on-forecast: Evidence of progress. *Weather and Forecasting*, **33 (2)**, 599–607, https://doi.org/10.1175/waf-d-17-0145.1.

Lecun, Y., Y. Bengio, and G. Hinton, 2015: Deep learning. Nature Publishing Group, 436–444 pp., https://doi.org/10.1038/nature14539.

Lee, S., H. Han, J. Im, E. Jang, and M. I. Lee, 2017: Detection of deterministic and probabilistic convection initiation using Himawari-8 Advanced Himawari Imager data. *Atmospheric Measurement Techniques*, **10 (5)**, 1859–1864, https://doi.org/10.5194/amt-10-1859-2017.

Lee, Y., C. Kummerow, and I. Ebert-Uphoff, 2020: Applying machine learning methods to detect convection using GOES-16 ABI data. *Atmospheric Measurement Techniques Discussions*, **(November)**, 1–28, https://doi.org/10.5194/amt-2020-420.

Leinonen, J., U. Hamann, and U. Germann, 2022: Seamless Lightning Nowcasting with Recurrent-Convolutional Deep Learning. *Artificial Intelligence for the Earth Systems*, **1 (4)**, https://doi.org/10.1175/AIES-D-22-0043.1, 2203.10114.

Lipton, Z. C., 2018: The mythos of model interpretability. *Communications of the ACM*, **61 (10)**, 35–43, https://doi.org/10.1145/3233231, 1606.03490.

Lundberg, S. M., and S. I. Lee, 2017: A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA, NIPS'17, Vol. 2017-Decem, 4766–4775, 1705.07874.

Maas, A. L., A. Y. Hannun, and A. Y. Ng, 2013: Rectifier nonlinearities improve neural network acoustic models. *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, **28**.

Mamalakis, A., E. A. Barnes, and I. Ebert-Uphoff, 2022: Investigating the Fidelity of Explainable Artificial Intelligence Methods for Applications of Convolutional Neural Networks in Geoscience. *Artificial Intelligence for the Earth Systems*, **1 (4)**, https://doi.org/10.1175/aies-d-22-0012.1, 2202.03407.

Mason, I., 1982: A model for assessment of weather forecasts. 291–303 pp.

Matthee, R., J. R. Mecikalski, L. D. Carey, and P. M. Bitzer, 2014: Quantitative differences between lightning and nonlightning convective rainfall events as observed with polarimetric radar and MSG satellite data. *Monthly Weather Review*, **142 (10)**, 3651–3665, https://doi.org/10.1175/MWR-D-14-00047.1.

Mayer, K. J., and E. A. Barnes, 2021: Subseasonal Forecasts of Opportunity Identified by an Explainable Neural Network. *Geophysical Research Letters*, **48 (10)**, e2020GL092 092, https://doi.org/10.1029/2020GL092092.

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

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

Mecikalski, J. R., P. Minnis, and R. Palikonda, 2013: Use of satellite derived cloud properties to quantify growing cumulus beneath cirrus clouds. *Atmos. Res.*, **120-121**, 192–201, https://doi.org/10.1016/j.atmosres.2012.08.017.

Mecikalski, J. R., T. N. Sandmal, E. M. Murillo, C. R. Homeyer, K. M. Bedka, J. M. Apke, and C. P. Jewett, 2021: A random-forest model to assess predictor importance and nowcast severe storms using high-resolution radar goes satellite lightning observations. *Mon. Weather Rev.*, **149 (6)**, 1725–1746, https://doi.org/10.1175/MWR-D-19-0274.1.

Mecikalski, J. R., P. D. Watts, and M. Koenig, 2011: Use of Meteosat Second Generation optimal cloud analysis fields for understanding physical attributes of growing cumulus clouds. *Atmos. Res.*, **102 (1-2)**, 175–190, https://doi.org/10.1016/j.atmosres.2011.06.023.

Mecikalski, J. R., J. K. Williams, C. P. Jewett, D. Ahijevych, A. LeRoy, and J. R. Walker, 2015: Probabilistic 0-1-h convective initiation nowcasts that combine geostationary satellite observations and numerical weather prediction model data. *Journal of Applied Meteorology and Climatology*, **54 (5)**, 1039–1059, https://doi.org/10.1175/JAMC-D-14-0129.1.

Molnar, C., 2020: Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. *Book*, 247, https://christophm.github.io/interpretable–ml–book.

Nisi, L., P. Ambrosetti, and L. Clementi, 2014: Nowcasting severe convection in the Alpine region: The COALITION approach. *Quarterly Journal of the Royal Meteorological Society*, **140 (682)**, 1684–1699, https://doi.org/10.1002/qj.2249.

Olah, C., A. Mordvintsev, and L. Schubert, 2017: Feature Visualization. *Distill*, **2 (11)**, e7, https://doi.org/10.23915/distill.00007.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830, https://doi.org/10.5555/1953048.2078195, 1201.0490.

Rasp, S., and N. Thuerey, 2021: Data-Driven Medium-Range Weather Prediction With a Resnet Pretrained on Climate Simulations: A New Model for WeatherBench. *Journal of Advances in Modeling Earth Systems*, **13 (2)**, https://doi.org/10.1029/2020MS002405, 2008.08626.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Weather and Forecasting*, **24 (2)**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

Sabottke, C. F., and B. M. Spieler, 2020: The effect of image resolution on deep learning in radiography. *Radiology: Artificial Intelligence*, **2 (1)**, https://doi.org/10.1148/ryai.2019190015.

Senf, F., and H. Deneke, 2017: Satellite-based characterization of convective growth and glaciation and its relationship to precipitation formation over central Europe. *Journal of Applied Meteorology and Climatology*, **56 (7)**, 1827–1845, https://doi.org/10.1175/JAMC-D-16-0293.1.

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

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, 2014: Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, **15 (56)**, 1929–1958.

Sun, F., B. Li, M. Min, and D. Qin, 2023: Toward a Deep-Learning-Network-Based Convective Weather Initiation Algorithm from the Joint Observations of Fengyun-4A Geostationary Satellite and Radar for 0-1h Nowcasting. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **16 (Ci)**, 3455–3468, https://doi.org/10.1109/JSTARS.2023.3262557.

Thambawita, V., I. Strümke, S. A. Hicks, P. Halvorsen, S. Parasa, and M. A. Riegler, 2021: Impact of image resolution on deep learning performance in endoscopy image classification: An experimental study using a large dataset of endoscopic images. *Diagnostics*, **11 (12)**, https://doi.org/10.3390/diagnostics11122183.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0098.1.

Tibshirani, R., 1996: Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society.*, **58 (1)**, 267–288.

Toms, B. A., E. A. Barnes, and I. Ebert-Uphoff, 2020: Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability. *Journal of Advances in Modeling Earth Systems*, **12 (9)**, e2019MS002 002, https://doi.org/10.1029/2019MS002002, 1912.01752.

Ukkonen, P., and A. Mäkelä, 2019: Evaluation of Machine Learning Classifiers for Predicting Deep Convection. *Journal of Advances in Modeling Earth Systems*, **11 (6)**, 1784–1802, https://doi.org/10.1029/2018MS001561.

Walker, J. R., W. M. Mackenzie, J. R. Mecikalski, and C. P. Jewett, 2012: An enhanced geostationary satellite-based convective initiation algorithm for 0-2-h nowcasting with object tracking. *Journal of Applied Meteorology and Climatology*, **51 (11)**, 1931–1949, https://doi.org/10.1175/JAMC-D-11-0246.1.

Weckwerth, T. M., J. W. Wilson, M. Hagen, T. J. Emerson, J. O. Pinto, D. L. Rife, and L. Grebe, 2011: Radar climatology of the COPS region. *Q. J. R. Meteorol. Soc.*, **137 (SUPPL. 1)**, 31–41, https://doi.org/10.1002/qj.747.

Wilks, D. S., 2019: Statistical Methods in the Atmospheric Sciences, Fourth Edition. *Stat. Methods Atmos. Sci. Fourth Ed.*, 1–818, https://doi.org/10.1016/C2017-0-03921-6.

Young, H. P., 1985: Monotonic solutions of cooperative games. *International Journal of Game Theory*, **14 (2)**, 65–72, https://doi.org/10.1007/BF01769885.

Zhang, Y., 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. *Monthly Weather Review*, **147 (12)**, 4389–4409, https://doi.org/10.1175/MWR-D-19-0163.1.

Zhuge, X., and X. Zou, 2018: Summertime convective initiation nowcasting over southeastern China based on advanced himawari imager observations. *Journal of the Meteorological Society of Japan*, **96 (4)**, 337–353, https://doi.org/10.2151/jmsj.2018-041.