Precipitation Vertical Structure Characterization: A Feature-Based Approach🖉

MALARVIZHI ARULRAJ,[a] VELJKO PETKOVIC,[a] RALPH R. FERRARO,[a] AND HUAN MENG[b]

[a] *Cooperative Institute for Satellite Earth System Studies, Earth System Science Interdisciplinary Center, University of Maryland, College Park, College Park, Maryland*
[b] *Center for Satellite Applications and Research, NOAA/NESDIS, College Park, Maryland*

ABSTRACT: The three-dimensional (3D) structure of precipitation systems is highly dependent on hydrometeor formation processes and microphysics. This study aims to characterize distinct vertical profiles of precipitation regimes by relying on the availability of a high-quality, spatially dense radar network and its capability to observe the 3D structure of the storms. A deep-learning-based framework, coupled with unsupervised clustering methods, is developed to identify types of precipitation structures irrespective of their physical properties. A 6-month period of 3D reflectivity profiles from the Multi-Radar Multi-Sensor (MRMS) network is used to identify different regimes and investigate their properties with respect to the underlying environmental conditions. Dominant features retrieved from radar reflectivity profiles using convolutional neural-network-based autoencoders are employed to identify similar-looking vertical structures using coupled k-means and agglomerative clustering algorithms. The k-means method identifies distinct groups, while the agglomerative clustering visualizes intercluster relationships. The framework identifies 18 clusters that can be broadly combined into five groups of varied echo-top heights. The 18 clusters demonstrate variability with respect to structural features and precipitation rate/type, implying that profiles in each group belong to a physically different precipitation regime. An independent analysis of the regime properties is conducted by matching the MRMS reflectivity profiles with environmental parameters derived from the High-Resolution Rapid Refresh model forecasts. The distribution of the environmental variables confirms cluster-specific feature properties, confirming the physics-based regime separation across the clusters and their dependence on the vertical structure. The identified precipitation regimes can assist in developing physics-guided retrievals and studying precipitation regimes.

SIGNIFICANCE STATEMENT: This study proposes a systematic model to identify precipitation profiles of distinct vertical structures and evaluate their dependence on environmental conditions. The model was developed using ground-based radar observations; however, there is potential to extend this model to reflectivity profiles from both ground- and satellite-based sensors. In addition, the identified precipitation regime clusters could be a proxy for the vertical structure of precipitation systems and assist in determining the structural variability within traditional precipitation type classification (e.g., convective versus stratiform). Moreover, identifying the precipitation regimes could also be used to improve satellite-based precipitation retrievals. Finally, a better understanding of precipitation structure would also help improve the initialization of climate models.

---

1. Introduction

Quantitative estimation and understanding of precipitation processes has increased vastly in the past three decades due to advanced remote sensing techniques, improved ground-based and spaceborne sensors, and retrieval techniques (Hong et al. 2019; Levizzani and Cattani 2019). Active and passive microwave sensors, such as precipitation radars and radiometers, measure the upwelling of radiation from Earth's surface to the top of the atmosphere. At the same time, quantitative precipitation estimates (QPEs) from these sensors are derived indirectly from the measurements. Active sensors can observe the three-dimensional (3D) distribution of hydrometeors explicitly (Mace et al. 1998; Shupe et al. 2001). At the same time, the retrievals from passive microwave sensors also depend on the total column-integrated distribution of hydrometeors and atmospheric composition, thereby implicitly relying on the vertical structure (Shi et al. 2015). Thus, the observations from satellite-based active and passive microwave sensors depend on the 3D structure of precipitation systems impacting the retrievals. For example, instantaneous precipitation QPEs from Global Precipitation Measurement (GPM; Hou et al. 2014) mission's Dual-Frequency Precipitation Radar (DPR; Iguchi et al. 2010), constraints the precipitation rate estimates with precipitation type and retrieved microphysics (Liao and Meneghini 2022). Passive microwave retrievals such as the Goddard profiling algorithm (GPROF; Kummerow et al. 2015) also use hydrometeor profiles from GPM DPR as an a priori knowledge when estimating the near-surface precipitation rate.

Besides, the 3D structure of precipitation also plays a critical role in other applications, including understanding precipitation microphysics and dynamics (Williams et al. 2007; Prat and Barros 2010), formation of convective processes and tropical

---

Corresponding author: Malarvizhi Arulraj, marulraj@umd.edu

cyclones (Hence and Houze 2011; Hu and Ryzhkov 2022), and lightning flash rates (Liu et al. 2012). Specifically, Dolan et al. (2018) highlighted that distinct clusters of near-surface drop size distributions (DSDs) could be related to specific reflectivity profiles and precipitation regimes. Radar reflectivity factor can also be used to initialize numerical weather prediction models to improve short-term forecasting of convective-scale precipitation events (Gustafsson et al. 2018). In particular, Schumacher et al. (2004) initialized a climate model to study the formation of large-scale tropical circulations using climatology of latent heating profiles derived from the precipitation radar on board Tropical Rainfall Measurement Mission (TRMM).

Precipitation-type classification is commonly used as a general representation of the vertical structure of the precipitation systems. For instance, precipitation systems are classified as deep and shallow or stratiform and convective systems based on the formation, thermodynamics, and vertical distribution of hydrometeors (Houze 2014). However, it should be noted that the 3D structure of precipitating events shows much higher variability within the traditional precipitation classes. For instance, midlatitude precipitation systems and extratropical cyclones can be represented as a combination of stratiform and convective events (Houze 1981). Additionally, observations and models revealed the presence of multilayered and collision–coalescence processes dominated systems are a combination of deep and shallow regimes (Wilson and Barros 2014; Verlinde et al. 2013; Porcacchia et al. 2019). Such systems exhibit higher variability in the vertical structure and are challenging to be classified, e.g., as either convective or stratiform systems. At the same time, these system-specific properties influence the near-surface precipitation rate retrievals. Therefore, grouping precipitation systems based on the profile structure should help identify different precipitation regimes and assist physics-guided precipitation retrievals (You et al. 2015) as well as initializing numerical weather prediction models (Smedsmo et al. 2005).

The advances in observational techniques measuring 3D profiles from satellite- and ground-based sensors have contributed to studies attempting to identify distinct precipitation regimes. However, this is an arduous task given the number of unique precipitation regimes is undetermined. Therefore, most studies leverage unsupervised clustering techniques to identify precipitation regimes. For example, Zhang et al. (2007) cluster histograms of reflectivity profiles from the *CloudSat* Cloud Profiling Radar (CPR) to derive distinct cloud regimes over tropics using *k*-means clustering techniques (Anderberg 1973) identifying five distinct groups related to atmospheric characteristics. Similarly, Luo et al. (2017) developed a methodology to identify cloud and precipitation regimes from coincident multifrequency observations from the *CloudSat* CPR and TRMM PR reflectivity profiles. The authors identified three distinct clusters further subdivided into three cloud regimes. Elsaesser et al. (2010) also aimed to identify different precipitation regimes from TRMM PR measurements. However, the authors extracted five features from reflectivity profiles, such as the number of cloud pixels with certain precipitation top height, average, and the ratio of convective precipitation rate, to identify three distinct regimes, namely, shallow, midlevel, and deep precipitation. Similar feature

extraction coupled with a *k*-means clustering approach has been followed by Arulraj and Barros (2021) where specific features from reflectivity profiles are used to distinguish low-level enhanced precipitation systems.

Current techniques for identifying precipitation regimes utilize reflectivity profiles, histograms of profiles within a region, or regime-specific features and parameters. However, using such a high-dimensional input to cluster results in sparse distribution of data points in the high-dimensional space, with traditional distance metrics, such as Euclidean distance, resulting in inefficient clustering (Johnstone and Titterington 2009; Singh et al. 2014). A potential solution for this problem can be found in reducing the number of dimensions through a feature selection or some of the extraction methods. The feature selection is based on selecting specific features of interest, while feature extraction corresponds to transforming the high-dimensional data into a lower-dimensional latent space. Being straightforward, these techniques are not optimal. Choosing specific features highly depends on domain knowledge and may result in selecting highly correlated features. Moreover, there is a possibility of missing features that provide nuanced information to identify distinct clusters. Principal component analysis (PCA) is a common statistical method to reduce dimension; however, PCA ignores nonlinear features. With the advent of high-performance computing and the availability of big data, the limitations of linear dimension reduction methods can be addressed using deep-learning-based methods such as autoencoders.

The objective of this work is to leverage deep-learning-based methods and unsupervised learning techniques to identify distinct precipitation regimes using ground-based reflectivity observations. Specifically, the study utilizes the 3D observations from ground-radar network to perform feature extraction, using a deep-learning-based autoencoder, and cluster specific features to identify profiles of similar vertical structures. The reflectivity profiles for liquid phase events are obtained from the Multi-Radar Multi-Sensor (MRMS) network over the eastern continental United States (CONUS). Finally, independent estimations of environmental variables from High-Resolution Rapid Refresh (HRRR) model forecasts are used to understand the physical meaning of the identified precipitation structure clusters.

The manuscript is organized as follows: section 2 describes the datasets and the study region. Section 3 introduces the machine learning framework used to extract features and the clustering techniques followed to identify distinct precipitation regimes. Section 4 assesses the performance of the feature extraction and clustering methods and links the precipitation structure clusters to model-derived environmental parameters. Section 5 investigates the relationship between machine learning–extracted features and physical variables. The conclusions and summary are provided in section 6.

2. Data description and study region

Observations from ground-based radars and independent model forecasts are used in this study. Detailed descriptions of the datasets used and the region of interest are provided below:
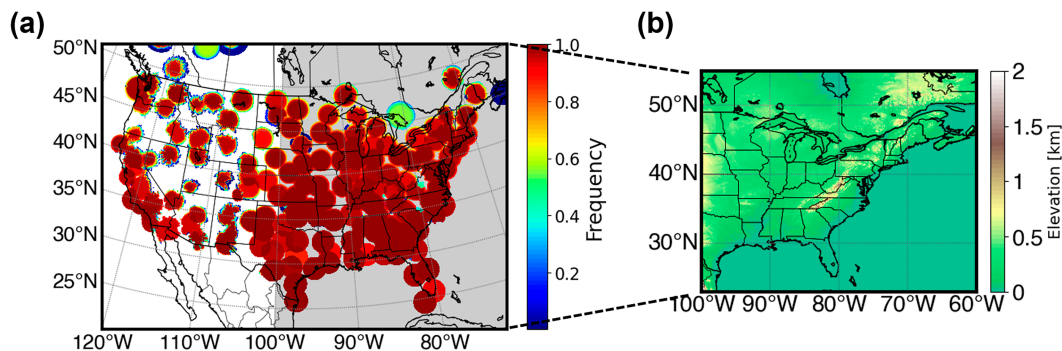
FIG. 1. (a) Frequency of available MRMS data with RQI greater than 0.8. The gray shaded region denotes the study region boundary. (b) Digital elevation model of the study region.

a. MRMS

The MRMS network ingests observations from 180 operational radars, including the S-band (3 GHz) dual-polarization Weather Surveillance Radar-1988 Doppler (WSR-88D) and Canadian radar network across the CONUS and southern Canada (Zhang et al. 2016). A seamless 3D mosaics are derived from the radar observations at a spatial resolution of $0.01° \times 0.01°$ (approximately 1 km × 1 km) and temporal resolution of 2 min (Qi and Zhang 2017). The data are available operationally at the National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Prediction (NCEP) platform. In this study, we utilize MRMS Version-12 data. We use the following parameters: MergedReflectivityQC, PrecipRate, PrecipFlag, and RadarQualityIndex. PrecipRate provides the QPEs from radar observations, and PrecipFlag inputs the precipitation type information. Seven precipitation types are identified based on the reflectivity profiles from radars and operational NWP model temperature profiles (Rapid Refresh; RAP). The precipitation types include warm stratiform rain, snow, convective, hail, tropical/stratiform mix, tropical/convective mix, and cool stratiform (Hong and Gourley 2018). Reflectivity values at 33 vertical levels, with height ranging from 0.5 to 19 km above mean sea level, and a variable vertical resolution (between 250 m and 1 km) are available under the MergedReflectivityQC variable. The quality of the radar observations is determined by RadarQualityIndex [radar quality index (RQI)]. The values of RQI range from 0 to 1, with 1 denoting the best-quality data. The quality control process on MRMS product ensures improved detection of melting layer and removal of missing, corrupt, and contaminated data, as well as mitigation of ground/sea clutter (Tang et al. 2020). Two-minute temporal resolution MRMS data between May 2021 and October 2021 are considered for analysis.

b. HRRR

The HRRR model (Dowell et al. 2022) is a cloud-resolving and convection-allowing weather research forecast model running near real time at the NOAA/NCEP. The model has 3-km horizontal grid spacing and initialization that includes assimilation of radar reflectivity observations. The model forecasts of 0–18-h lead time are archived for the period from

2014 to present and are publicly accessible from various cloud storage platforms including Google and Amazon Web Services cloud storage. The objective of the model is to improve short-scale forecasts of convective systems with a focus on the storm life cycle. This study utilizes surface variables from model forecasts with a lead time of 0 h. In particular, the following environmental variables are considered for analysis: surface pressure, air temperature at 2 m above the surface, relative humidity, the 0° isotherm level, precipitable water, total column integrated cloud water, convective available potential energy (CAPE), and convective inhibition (CIN). The data are extracted for the same period as the MRMS data and will be used in section 4d to investigate connections between rainfall physical properties and the vertical structure.

c. Study region

The MRMS network predominantly covers the CONUS and southern Canada. Figure 1a shows the percentage of MRMS data with RQI ≥ 0.8 over the MRMS coverage area between May and October 2021. Given the sparse coverage over the western CONUS, the study is focused on the region east of 100°W (gray shaded region in Fig. 1a; 60°–100°W, 20°–50°N). The digital elevation model (DEM) for the study region is depicted in Fig. 1b. The eastern CONUS has heterogeneous DEM and includes the southern Appalachian Mountains. However, most of the study region lies below 2 km above sea level, and thus the MRMS observations do not suffer from significant beam blockage.

3. Methods and model description

Precipitation regime identification is set as an unsupervised machine learning problem since the distinct precipitation vertical structure is unknown. Therefore, a data-driven framework is defined to include a feature extraction and clustering component to systematically identify and cluster the reflectivity profiles based on their shape. The framework consists of three major steps: data preprocessing, feature extraction, and clustering.

a. Preprocessing

Preprocessing includes data filtering and normalization. The study uses reflectivity profiles with an RQI of at least 0.8,

a minimum MRMS precipitation rate of 0.1 mm h^{-1}, and precipitation types that are not flagged as snow. To eliminate arbitrary values and to be consistent with the varying terrain height, the reflectivity profile height is carried in above ground level (AGL) notation across the entire region of interest. The reflectivity values are converted to an equal vertical resolution of 250 m using linear interpolation for levels from 1 to 15 km AGL. This interpolation step converts the 33-level profiles from MRMS data into 57-level profiles with heights ranging between 1 and 15 km AGL. As an additional step for quality control purposes, the profiles are checked for discontinuity and attenuation. Reflectivity profiles with values lower than 10 dBZ within 2 km AGL are not considered for analysis, as these profiles are considered to be of low quality or do not correspond to liquid precipitation type (Kirstetter et al. 2015; Wen et al. 2017). The proposed quality control eliminates approximately 3% of the total MRMS reflectivity profiles obtained within the study region for the period May–October 2021, resulting in around 500 million profiles for analysis.

The extracted reflectivity profiles are split into three independent subsets to enable the training, validation, and testing of machine learning algorithms. The training dataset consists of a total of 5.5 million profiles randomly chosen from the days 26–31 for every month (26–30 for June and September). Similarly, the validation dataset is constructed from randomly chosen profiles from days 21–25 over the study time period. The validation data used in this study has approximately 3.5 million profiles. All the data that are not considered for training and validation are independent (i.e., test dataset) and used to demonstrate the performance of the machine learning methods. A large enough training and validation datasets were chosen to train the autoencoder while a much larger independent test dataset was chosen to demonstrate the application.

b. Feature extraction

The proposed model targets extracting physically meaningful features from the input data. Each pixel of the MRMS gridded dataset is considered as an individual data point. Yet, considering a reflectivity profile with 57 vertical levels leads to an overall high number of dimensions with interdependent features. This could lead to inefficient clustering (the next step described in the following subsection) suggesting the need for feature extraction (Johnstone and Titterington 2009; Singh et al. 2014). Thus, this work leverages dimension-reduction techniques for feature extraction, such as hand-crafting physical variables, linear algorithms (e.g., PCA), and nonlinear methods (e.g., autoencoders). Due to the complexity of the profile shapes and the need to capture intrinsic variability within the profile, we rely on deep-learning-based methods to extract features from the input data.

Specifically, we use the convolutional neural network (CNN) based autoencoders for dimension reduction and feature extraction. CNN-based autoencoders belong to lossy dimension reduction methods that are primarily data driven and consist of two modules: encoder and decoder. The output of the encoder is the derived features, while the output of the decoder is the reconstructed input data. The encoder allows for learning nonlinear features and properties from the input data and transforming the data to a lower-dimension space. The extracted features from the encoder module are passed as input to the decoder module. The decoder is trained to reconstruct the input data from the reduced-dimension features. The output of the decoder module is the reconstructed reflectivity profiles. In a well-constructed model, the features extracted from the encoder provide an optimal feature-to-information link, representing the input data with minimal reconstruction losses. Even though the deep-learning-based methods are computationally intensive, they could efficiently capture microdetails within the input data. The recent advances in high-performance computing facilities [e.g., graphics processing units (GPUs)] enable us to overcome the shortcomings of computational complexity.

The encoder module consists of convolutional layers followed by fully connected (FC) layers. Each convolutional layer consists of convolutional filters followed by batch normalization, activation function, and pooling layer. The input data are one dimensional; thus, 1D convolutional filters are used in the model. The output of the convolutional filters is transformed using an activation function, after which the pooling layer performs downsampling by either averaging (low-pass filter) or extracting the maximum value (high-pass filter). On the other side, the decoder module consists of fully connected layers followed by deconvolutional layers. A deconvolutional layer has transposed convolutional filters followed by batch normalization and activation layers. Here, the encoder and decoder models are trained in tandem to minimize the reconstruction loss of input data.

The schematic of the autoencoder model used in this study is shown in Fig. 2a. The model is implemented using the PyTorch platform (an open-source machine learning framework) and trained on a workstation with 2x NVIDIA RTX 2080 Ti GPUs. Upon extensive hyperparameter tuning, the architecture shown in Table 1 and Fig. 2a were identified as optimal. The encoder module has two layers, each with 1D convolutional filters, batch normalization, activation functions, and a pooling layer. The batch normalization step is applied between the layers of the neural network architecture. The exponential linear unit (ELU) activation function is considered in the architecture to translate the output from convolutional layers. Then, the maximum pooling operation is performed to retain the element with the highest value within the region of the convolutional filter. Five FC layers follow the convolutional layers of the encoder module, with the final FC layer of the encoder module providing the extracted lower-dimensional features. The ELU is used as the activation function for the FC layers. The number of nodes in the final FC layer of the encoder module is modified to fit the number of features.

Similarly, the decoder module includes a set of FC layers followed by the deconvolutional layers with an ELU activation function. A standard gradient descent optimizer is used to train the model with mean squared error (MSE) as the loss function. The loss function is calculated between the input
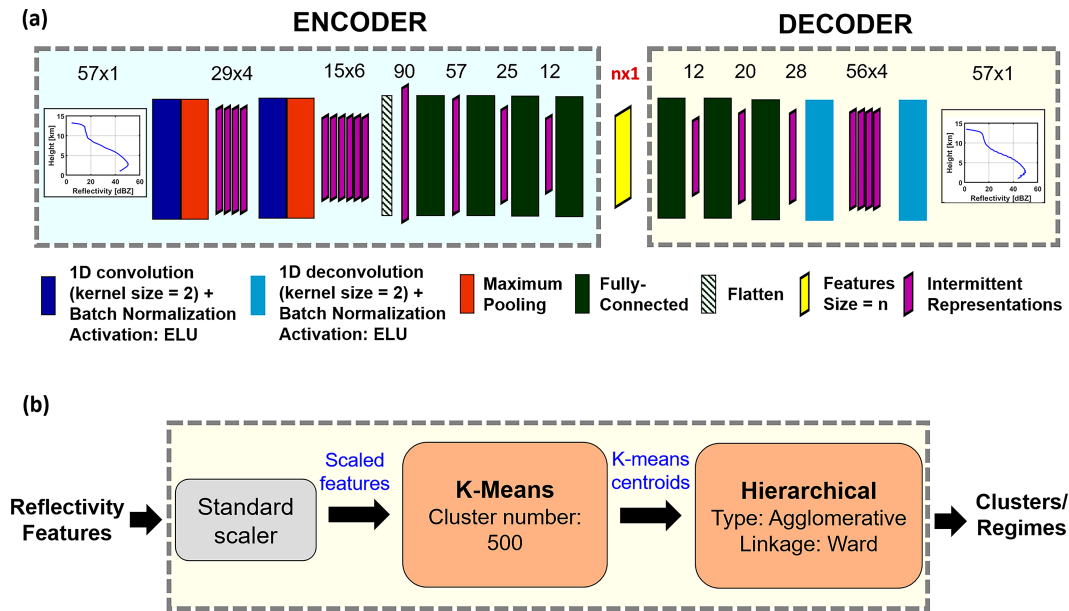
FIG. 2. (a) Schematic of the autoencoder model that extracts *n* features from input reflectivity profiles of size 57×1. (b) Schematic of the clustering framework used to identify precipitation regimes.

reflectivity profile and the reconstructed profile, which is the output of the decoder module:

$$L = \frac{1}{N} \sum_{i=1}^{N} (Z_i - \hat{Z}_i)^2, \quad (1)$$

where L (dBZ) is the loss, N is the total number of reflectivity profiles, Z (dBZ) is the input reflectivity factor, and $\hat{Z}$ (dBZ) is the output of the decoder module.

Feature extraction is performed for a different range of features using the same model architecture. As a result, the number of nodes in the final FC layer of the encoder is modified to accommodate the different number of extracted features. The model is trained for different number of features for at least 200 iterations until it converges. The best number of features is chosen based on a new metric called optimal model indicator (OMI) that optimizes the model complexity and model performance. OMI for a k-feature autoencoder model (OMI$_k$) is calculated as follows:

$$\text{OMI}_k = \frac{k}{\sum\limits_{i=\text{n}1}^{\text{n}2} i} + \frac{L_k}{\sum\limits_{i=\text{n}1}^{\text{n}2} L_i}, \quad (2)$$

where k is the number of features, n1 and n2 are the minimum and maximum number of features considered in the autoencoder model, and L_k is the MSE loss for the k-feature autoencoder model. The model complexity is defined by the number of features while the model performance is quantified by MSE loss (L). The autoencoder model with minimum OMI is considered as the optimal model and best number of features.

c. Clustering

In the final step, the output from the feature extraction framework is used as an input to group reflectivity profiles with similar structures. Here, the number of distinct profile patterns is undetermined. Thus, this step relies on unsupervised machine learning. Several unsupervised clustering algorithms have been proposed in the literature (described below), each very sensitive to the application and properties of the input data, and of challenging to evaluate performance.

The k-means algorithm (Anderberg 1973), the most commonly used method, clusters the data points into k distinct groups by minimizing the intracluster distances iteratively. The k-means algorithm is computationally inexpensive and can handle a large amount of data efficiently. However, the algorithm is sensitive to initial conditions, such as the choice of initial cluster centroids and the total number of clusters. At the same time, it is prone to gravitating toward local minima (Li and Wu 2012), leading to inefficient clustering. Besides, determining the number of clusters (especially for an unsupervised

TABLE 1. Hyperparameters of the CNN-based feature-extraction model. The number of features extracted from the encoder module is denoted by *n*.

Hyperparameters	Encoder	Decoder
Convolutional layers	2	—
Convolutional filters	4, 6	—
Convolutional kernel size	$1 \times 2, 1 \times 2$	—
Deconvolutional layers	—	2
Deconvolutional filters	—	4, 1
Deconvolutional kernel size	—	$1 \times 2, 1 \times 2$
Fully connected layers	4	3
Fully connected neurons	90, 57, 25, 12, n	12, 20, 28

scenario) is nontrivial, complicating clusters' organization and visualization for high-dimensional data.

Fred and Jain (2005) proposed that combining different clustering techniques has the capability for better performance and can help mitigate the disadvantages associated with a single algorithm. A coupled k-means and hierarchical clustering was tested on high-dimensional data and proved to perform better (Chen et al. 2005; Peterson et al. 2018). The k-means algorithm helps identify localized centers while the hierarchical clustering can improve the cluster visualization and understanding without the restriction on choosing the number of clusters a priori. A similar approach is adopted here to distinguish characteristic vertical reflectivity profiles.

The flowchart of the clustering framework used in this study is shown in Fig. 2b. The input data are the extracted features from the autoencoder model. As a first step, these features are normalized using a standard scaler to avoid potential bias toward higher values. The standard scaler removes the mean and normalizes the data using the variance. Next, the scaled features are provided as the input to the k-means algorithm. Here, we perform k-means clustering for a large value of k, and extract the corresponding cluster centroids. The extracted k-means centroids are the inputs to the agglomerative clustering algorithm (a type of hierarchical algorithm). Agglomerative clustering follows a bottom-to-top approach where every data point is an individual cluster initially. The algorithm combines similar instances following the minimizing criteria until all the points are grouped into a single cluster. Agglomerative clustering is computationally complex $[O(n^3)]$ as this method calculates pairwise distances for every data point. This study uses "ward" linkage as the minimizing criteria. Ward linkage follows Euclidean distance and aims to minimize intraclass variances similar to that of k-means algorithm. Typically, a dendrogram (tree diagram) is used to visualize the organization and relationship in hierarchical clustering and to determine the optimal number of clusters.

The best number of clusters can be derived based on the compactness within identified groups and high separations across the groups. Some of the commonly used metrics to estimate the best number of clusters are the Dunn index (Dunn 1974) and Davies–Bouldin (DB) score (Davies and Bouldin 1979). The Dunn index is defined as the ratio between the minimum intercluster distance and maximum intracluster distance. The DB score is the ratio between the average intracluster and intercluster distance. Higher values of the Dunn index and lower values of the DB score are usually preferred to obtain the best number of clusters. This study uses the DB score and inverse of the Dunn index to identify the best number of clusters.

4. Results

a. Feature extraction

The autoencoder model proposed in section 3b is used for feature extraction. The best model is expected to have the least reconstruction loss between the input data and reconstructed data. The model is trained for a different number of encoded features, ranging from 3 to 12, and the reconstruction loss is calculated. At least 200 epochs are executed for every experiment until the training and validation loss converges to prevent overfitting. The MSE loss for the training and validation samples from the autoencoder is shown in Fig. 3a. As the number of features increases, more information on the reflectivity profiles is retained by the encoder model, and thus, the reconstruction loss is reduced. However, when the number of features is increased to 8 or more, the computational complexity increases while the increase in performance is not significant. This is evident from OMI values shown in Fig. 3b. The minimum OMI is obtained for 8 features and thus, 8 features is recognized as the best number for preserving the information contained in the reflectivity profiles. Figure 3c shows an example profile from the MRMS data (solid black line) and the reconstructed profiles (dotted lines) from the autoencoder framework. The reconstructed profile from the 3-feature autoencoder model fails to accurately capture the entire structure, especially around 12.5 km AGL and below 2.5 km AGL. On the other hand, as the number of features increases, the overall shape is well captured. However, the improvement between the 8-feature and 12-feature reconstructed reflectivity profile is insignificant. Finally, Fig. 3d shows the convergence of the 8-feature autoencoder training and corresponding validation loss function. The training loss converges around 0.37, while the validation loss is approximately 0.39, confirming that the autoencoder model does not overfit and truly extracts the best features.

To further validate the ability to preserve the structure, correlation coefficient [CC; Eq. (3a)] and bias [Eq. (3b)] are calculated for the measured (Z) and reconstructed reflectivity profiles ($\hat{Z}$). These error assessments provide understanding of the biases in the feature extraction model that could lead to potential uncertainties. This analysis is performed on the test dataset, which includes more than 440 million profiles. Figure 4a shows the histogram of the correlation coefficient between the observed and the corresponding reconstructed reflectivity profiles. More than 93% of the reflectivity profiles have a correlation coefficient greater than 0.9 (magenta line in Fig. 4a). In particular, approximately 70% of the reflectivity profiles have CC around 0.99, suggesting that the reconstructed profiles could capture the structure of the measured profiles extremely well:

$$CC = \frac{1}{N} \frac{\sum_{i=1}^{N}[\hat{Z}_i - \text{mean}(\hat{Z})][Z_i - \text{mean}(Z)]}{\sqrt{\sum_{i=1}^{N}[\hat{Z}_i - \text{mean}(\hat{Z})]^2}\sqrt{\sum_{i=1}^{N}[Z_i - \text{mean}(Z)]^2}}, \quad (3a)$$

$$\text{bias} = \frac{\sum_{i=1}^{N}(\hat{Z}_i - Z_i)}{\sum_{i=1}^{N} Z_i}. \quad (3b)$$

The histogram of the bias between Z and $\hat{Z}$ is shown in Fig. 4b. Almost all the profiles (99%) exhibit less than 1% bias; 70% of
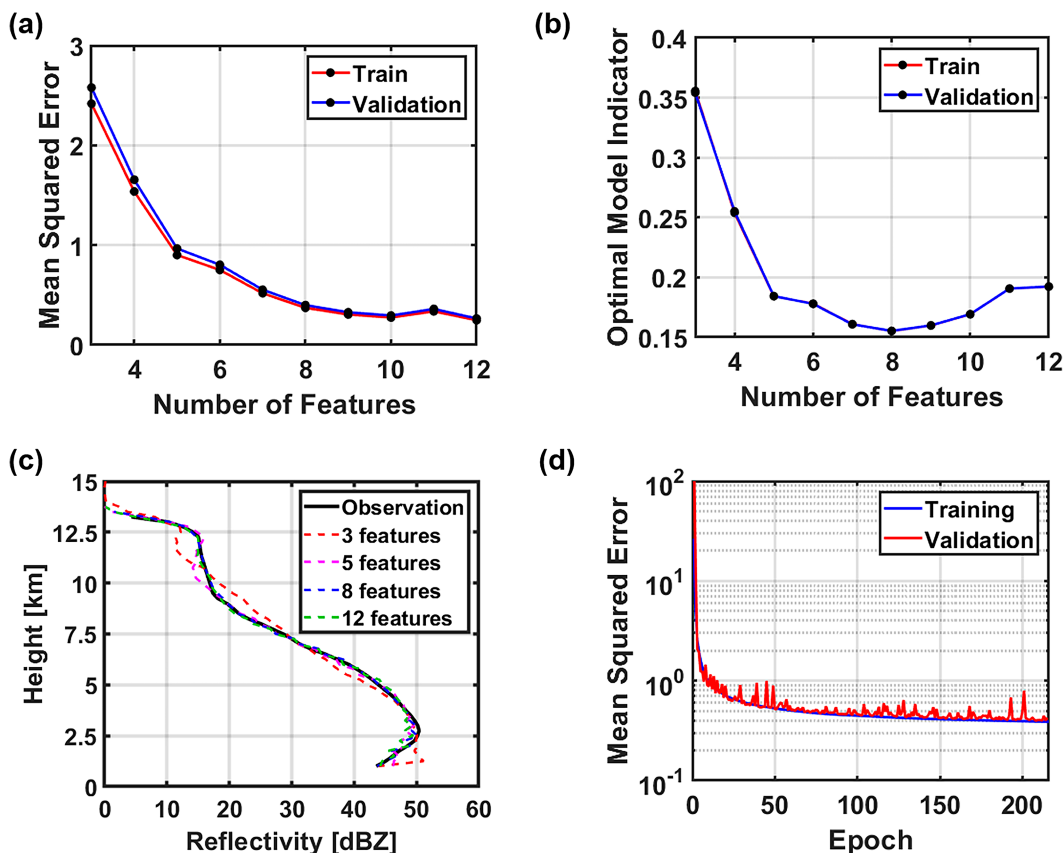
FIG. 3. (a) Reconstruction loss of the autoencoder model and (b) optimal model indicator (OMI) compared with the different number of features. The 8-feature model is identified as the best number of features. (c) An example MRMS-observed reflectivity profile (black curve) and its reconstruction using an autoencoder model with different dimensions. (d) MSE loss function for training and validation of the 8-feature autoencoder model.

the reflectivity profiles have less than 0.1% bias. Besides, overestimation seems to be slightly dominant compared to underestimation (61% of profiles have a positive bias). Overall, most of the profiles have higher CC and lower bias with no evidence of systematic errors in the model. Finally, the reconstructed profiles are evaluated to investigate the dependence of bias with height (Fig. 4c) and reflectivity factor value (Fig. 4e). Figure 4c shows lower bias in most vertical levels. The highest biases are observed around 14–15 km AGL, but other vertical levels have less than 5% bias. However, only 3% of the total profiles have reflectivity values above 13 km AGL (Fig. 4d). The bias analysis with respect to the reflectivity value (Fig. 4e) highlights a negative bias of approximately 6% only when reflectivity values are greater than 60 dBZ. All the other reflectivity values exhibited a bias of less than 2%. Therefore, the reconstructed profiles have significantly smaller biases except for the top 1 km (14–15 km) and at very high reflectivity values. However, the higher bias values are observed only for a limited sample size (3% profiles). This analysis confirms that the extracted features do not exhibit any systematic errors associated with the feature extraction model and can be used for precipitation regime identification.

b. Identification of precipitation regimes

The proposed clustering framework is demonstrated using the extracted features of the validation dataset. After scaling and normalization by the standard scaler, the features from the autoencoder model are provided as the input to the k-means clustering algorithm. The k-means clustering algorithm groups the data into 500 clusters, thereby extracting 500 cluster centroids. Even though this model is demonstrated with 500 k-means cluster centroids, note that the proposed framework is applicable for any large number of clusters, and the final interpretation of results is not sensitive to the number of k-means clusters chosen. Figure 5 shows the dendrogram of the agglomerative clustering algorithm applied to the 500 k-means cluster centers. Agglomerative clustering follows bottom to top approach. The full dendrogram starts from 500 individual clusters and extends to a single cluster. Here, the dendrogram shown is restricted to 20 clusters instead of all 500 cluster centers for easy visualization and understanding. The y axis of the dendrogram denotes a proxy to the distance at which the two different clusters are combined. The numbers in the individual boxes in Fig. 5 denote the number of k-means centroids in the corresponding level.
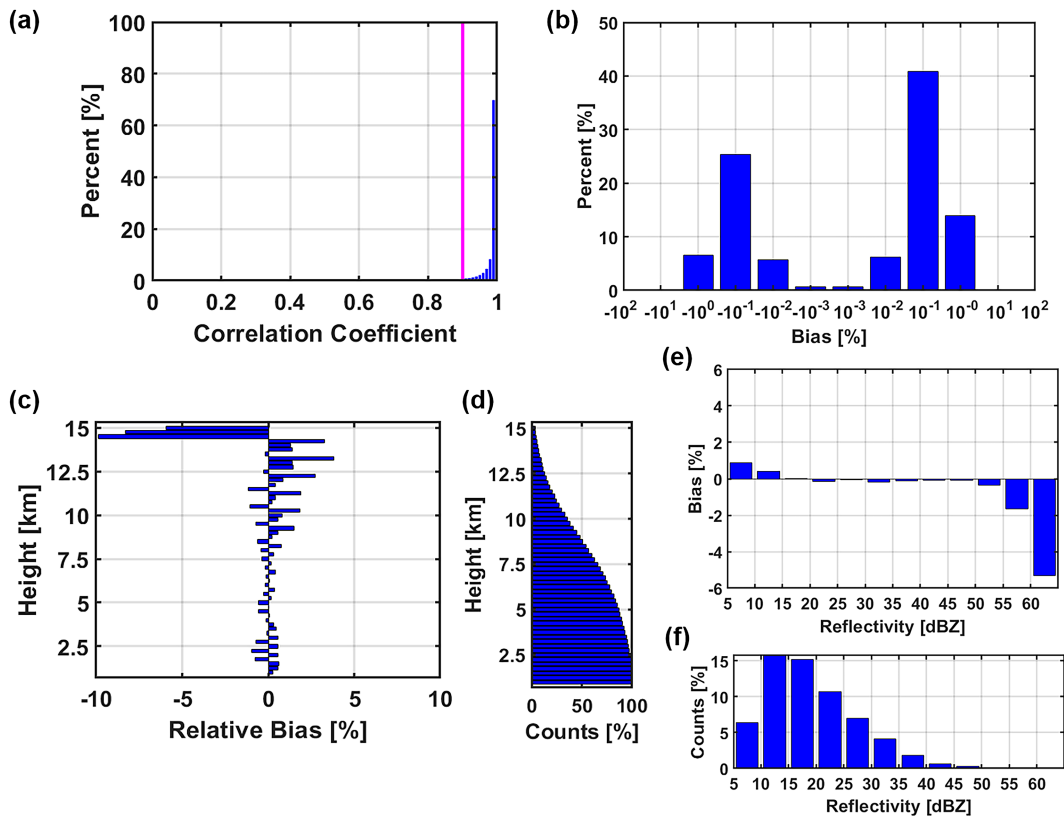
FIG. 4. Histograms of (a) CC and (b) bias for reflectivity profiles. Variation of bias with (c) vertical levels and (d) frequency of number of samples. Variation of bias with (e) reflectivity factor and (f) number of samples in corresponding reflectivity bins.

Next, the inverse Dunn index and DB score are calculated for 2–25 clusters for the agglomerative clustering to determine the best number of clusters (Fig. 6). The minima for the inverse Dunn index and DB score signify the best number of clusters with higher intercluster and lower intracluster distances. Both the inverse Dunn index and DB score point to two clusters as the optimal number of clusters. The two-cluster scenario separates reflectivity profiles into deep and shallow regime, with the results only based on the echo-top height. The highest variability of the reflectivity profiles is predominantly observed at the storm top height, where the values vanish to zero. When evaluating the contoured frequency by altitude diagrams (CFADs) for the scenarios with less than five clusters, the driving factor of the clustering algorithm is the top height. For example, the five-cluster case (red dotted line in Fig. 3) has clusters with echo-top heights around 6, 8, 10, 12, and more than 12 km AGL, respectively (Fig. 7). Beyond the five-cluster scenario, the other factors, such as near-surface reflectivity values, multilayer processes, bright band, and curvature of profiles, start playing a role in the classification. Thus, different vertical structure within similar echo-top height is identified when considering higher number of clusters.

The second local minima of the Dunn index and DB score (Fig. 6) curve is observed when the number of clusters is between 18 and 20. The higher number of clusters separated the

profiles based on the structural variations in addition to echo-top height. The lower number of clusters separations are primarily dependent on the top height of the storm. Even though Fig. 6 suggested 18 clusters as the best number of clusters based on the metrics, the 5-cluster groups were also chosen to identify the clusters within the 18-cluster group that has similar echo-top height for further analysis. Choosing the 5-cluster group in addition to the 18-cluster group enables to identify reflectivity profiles with similar echo-top height, but different overall vertical structure and near-surface environmental variables to highlight the importance of vertical structure variations and the associated precipitation regimes. The red and magenta dotted lines in Fig. 5 represent the distance thresholds considered to obtain 5 and 18 clusters, respectively. Note that the clusters in the 18-cluster scenario contribute to the 5-cluster groups, as seen in Fig. 5.

Hereinafter, we refer to the 5 clusters from the 5-cluster threshold as "macroclusters" and the source clusters from the 18-cluster threshold as "subclusters." The 5 macroclusters and their corresponding subclusters in the 18-cluster case are highlighted by the same color-shaded boxes in Fig. 5. For example, the first cluster of the 5-cluster scenario (C1) has five subclusters from the 18-cluster scheme. Thus, each cluster from the 5-cluster scenario can be mapped to an independent set of clusters in the 18-cluster case. The clusters in the
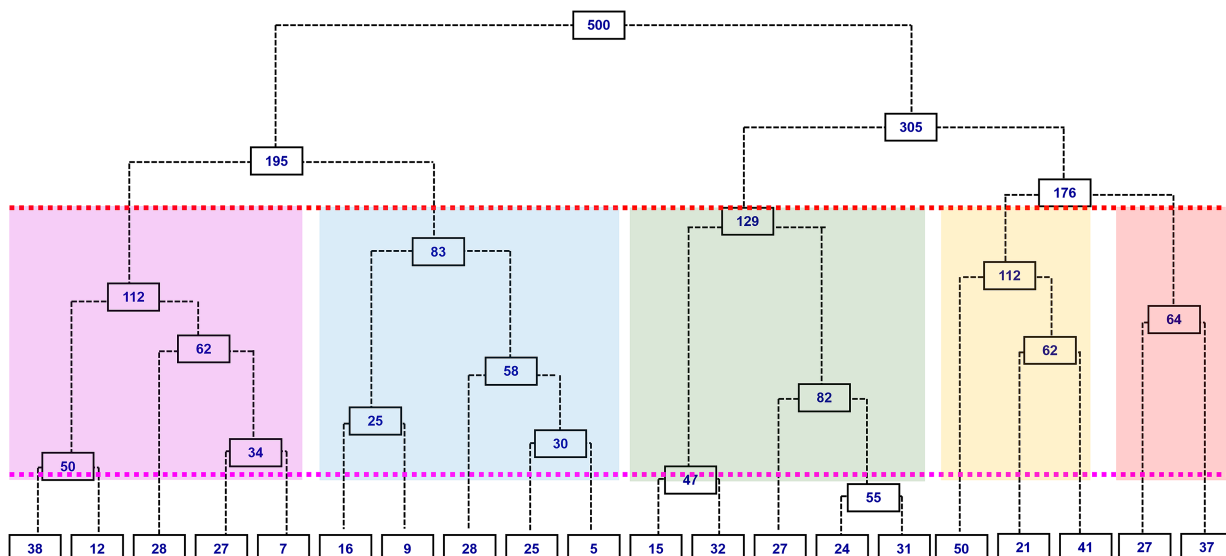
FIG. 5. Dendrogram constructed from agglomerative hierarchical clustering of 500 *k*-means cluster centroids. The numbers in each box denote the number of *k*-means centroids in that corresponding level. The red and magenta dotted lines denote distance thresholds for 5 and 18 clusters, respectively. The colored boxes represent each of the 5 clusters and the associated clusters in 18-cluster cases.

5-cluster scenario are denoted as C1, C2, C3, C4, and C5, with C1 pointing to the lowest echo-top cluster and C5 representing the group with the deepest reflectivity profiles. The subclusters of macrocluster C*i* are called C*i-j*, where *i* is the cluster number from the 5-cluster case, and *j* denotes the "cluster index"—the subcluster. Figure 8 shows the CFADs of reflectivity profiles from all 18 clusters. The panels of Fig. 8 are organized and named to highlight the subcluster nodes for the 5-cluster scenario. The CFADs confirm that the subclusters of the 5-cluster case could subclassify the reflectivity profiles beyond the storm top height.

c. Characterization of the precipitation regimes

The precipitation structure clusters are strictly based on the vertical structure of the reflectivity profiles. The structural features such as near-surface reflectivity factor, echo-top height, maximum reflectivity factor, and height at which maximum reflectivity occurs are derived from the reflectivity profiles. The echo-top height is calculated as the maximum height at which the reflectivity factor is greater than or equal to 10 dB*Z*. The precipitation type and rain rate are obtained from the MRMS product. The distribution of the structural features, precipitation rate and type, are computed to characterize the physical properties of the identified clusters. Figure 9 shows the boxplots of the reflectivity features for all 18 clusters. The naming convention for the clusters and color-shaded boxes is the same as seen in Fig. 8. The upper and lower limits of the box plots highlight the 25th and 75th percentile of the data points, and the solid red line depicts the median of the data. Figure 9a shows the distribution of echo-top height of the reflectivity
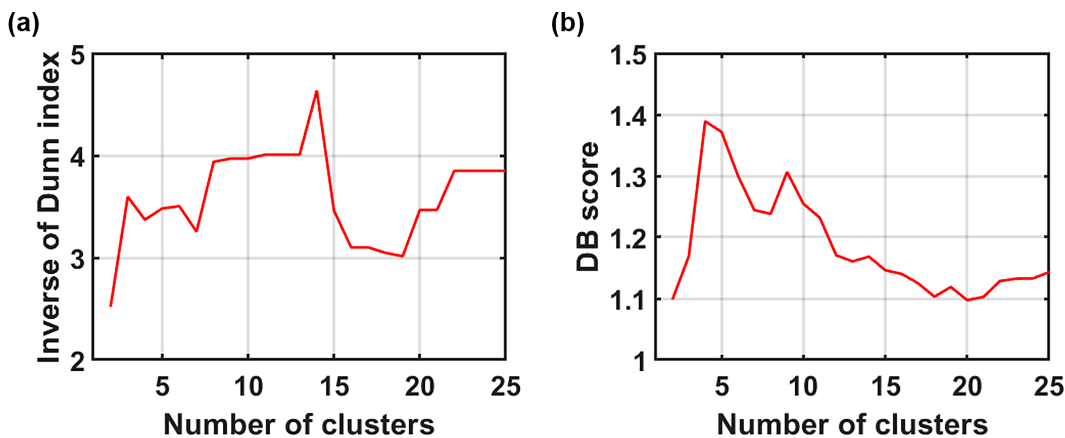

FIG. 6. (a) Inverse of the Dunn index and (b) DB score for different number of clusters applied to agglomerative clustering algorithm.
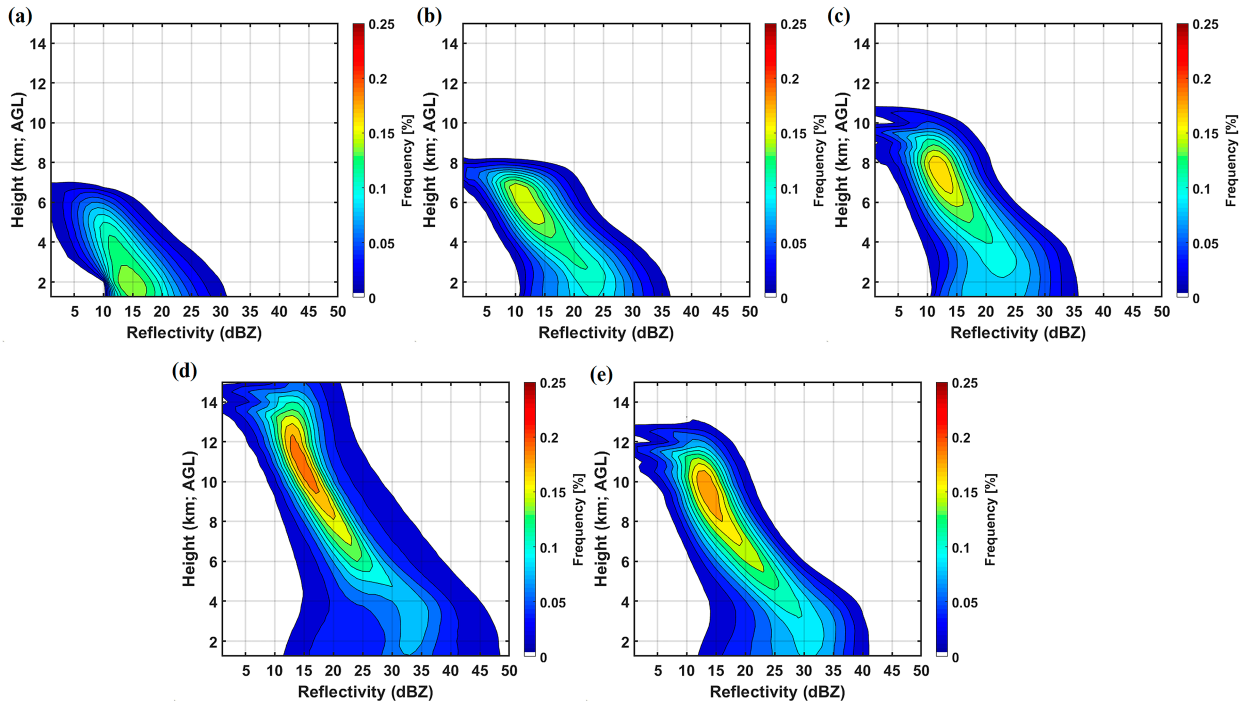
FIG. 7. Contoured frequency by altitude diagram of reflectivity profiles for clusters obtained from 5-cluster agglomerative clustering.

profiles. The maximum variation of echo-top height is noticed for the subclusters within "C1" clusters, where the height varies between 3 and 6 km AGL. The echo-top height is almost similar for all the other clusters and lies within a 1-km interval. This finding confirms that cluster C1 includes shallow profiles, and the echo-top height increases as the cluster number increases.

Cluster C1 consists of shallow profiles with a light precipitation rate. However, cluster C1-5 has a higher median near-surface and maximum reflectivity factor compared to the other subclusters of C1. The MRMS-derived rain rate (Fig. 9e) also supports this observation. Additionally, profiles in C1-5 seem to have the median maximum reflectivity around 3 km AGL. On the other hand, C1-1 to C1-4 exhibit maximum reflectivity value at 1 km AGL highlighting the vertical structure difference among cluster C1. Similarly, the profiles in cluster C2 have echo-top heights varying between 6 and 8 km AGL, denoting midlevel precipitating storms. All the parent clusters within the C2 group also exhibit a similar maximum reflectivity height of around 2 km. However, C2-3 includes profiles with lower near-surface precipitation rates than the other parent clusters of C2.

For profiles in the C3 cluster group, the median echo-top height is around 8–9 km AGL. Especially, C3-1 has a lower echo-top height, while C3-2 and C3-3 have similar echo-top heights around 9 km AGL. However, C3-1 and C3-3 have similar median maximum and near-surface reflectivity factors, while C3-2 has approximately five dBZ higher near-surface and maximum reflectivity factors. The distribution of height with maximum reflectivity factor is very similar for all three subclusters of the C3 group. The C4 group also has three subclusters, with the

median echo-top height varying between 10 and 12 km AGL. These profiles represent deep precipitation events with near-surface reflectivity values varying between 20 and 35 dBZ. The median rain rate observed by MRMS ranges between 1 and 5 mm h^{-1}. C4-3 consists of a lower rain rate than other subclusters, and the same pattern is seen in near-surface and maximum reflectivity factors. Maximum reflectivity height varies between 2 and 3 km AGL for the C4 scenario, with C4-3 showing a higher median height. These observations and the CFADS shown in Fig. 8p confirm that C4-3 has a clear brightband structure. Finally, C5 has two subclusters with very similar echo-top height of around 14 km AGL representing deep precipitation systems. The profiles in cluster C5-2 have a median rain rate of approximately 2 mm h^{-1} and a near-surface reflectivity factor of around 25 dBZ. In contrast, the profiles in the C5-1 group have a rain rate about 10 times higher than that of the C5-2. Also, profiles in C5-1 exhibit the highest median rain rate, maximum, and near-surface reflectivity compared to all the 18 clusters. Overall, the statistical distribution analysis of reflectivity features and near-surface rain rate confirms that the clustering algorithm could distinguish between profiles by echo-top height, presence of the bright band, near-surface enhancement, and rain rate.

The identified clusters are compared with the MRMS precipitation type. Figure 10 shows the distribution of the MRMS precipitation type with respect to the 18 clusters. Figure 10 does not show the frequency of stratiform precipitation (provided in the online supplemental material). Stratiform precipitation type dominates all but the C5-1 cluster. In addition to stratiform events, subclusters in C1 also include cool stratiform regimes. Convective and tropical stratiform precipitation
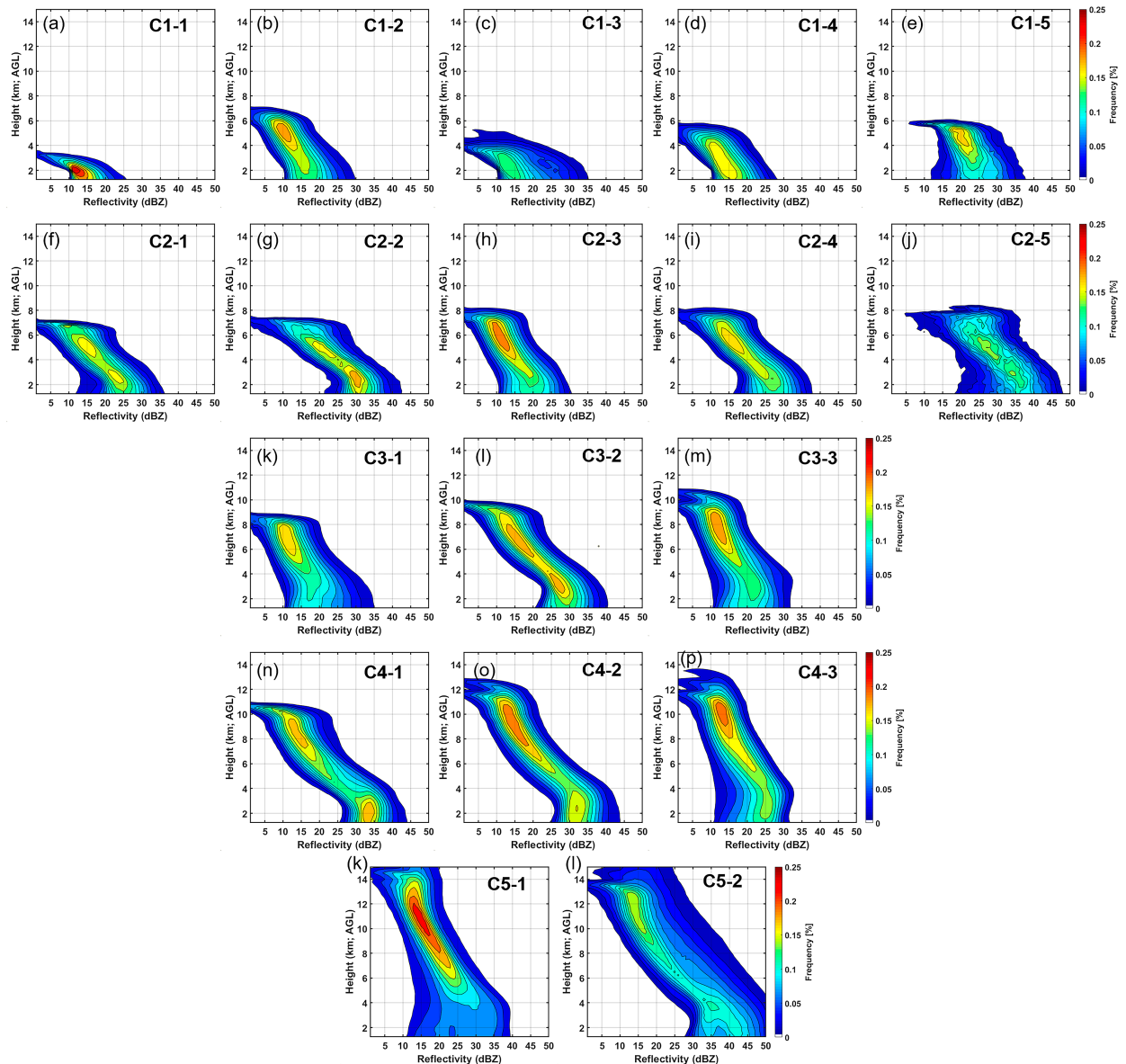
FIG. 8. Contoured frequency by altitude diagram (CFADs) of reflectivity profiles for clusters obtained from the 18-cluster case.

types are identified in clusters C1-5. Moreover, the profiles in C2, especially clusters C2-2 and C2-5, also have a higher frequency of convective and tropical stratiform precipitation. The rest of the C2 subclusters include tropical stratiform and cool stratiform cases. The C3 clusters have higher echo-top heights compared to the C1 and C2 clusters and are dominated by tropical stratiform cases. Finally, convective cases are found to be higher as the cluster number increase. Subclusters C4-1 and C4-2 have around 0.1% of reflectivity profiles classified as convective and tropical convective storms. The C4-3 cluster, even though it has a similar echo-top height as C4-2, it does not include dominant convective storm instances. Similarly, in the C5 cluster group, C5-1 includes most hail and convective storms, while C5-2 shows a higher frequency

of convective and tropical stratiform storm events. This analysis proves that the reflectivity structure classification reveals a more defined categorization associated with the structure compared to the traditional precipitation type classification. The methodology only depends on the reflectivity profiles enabling the identified clusters to offer a better understanding of precipitation regimes, which in turn could support the development of physically consistent precipitation retrievals for satellite- and ground-based sensors.

d. Linking regimes to environmental conditions

To understand the impact of environmental factors on precipitation vertical structure, we rely on surface variables derived from HRRR model forecasts, such as the surface
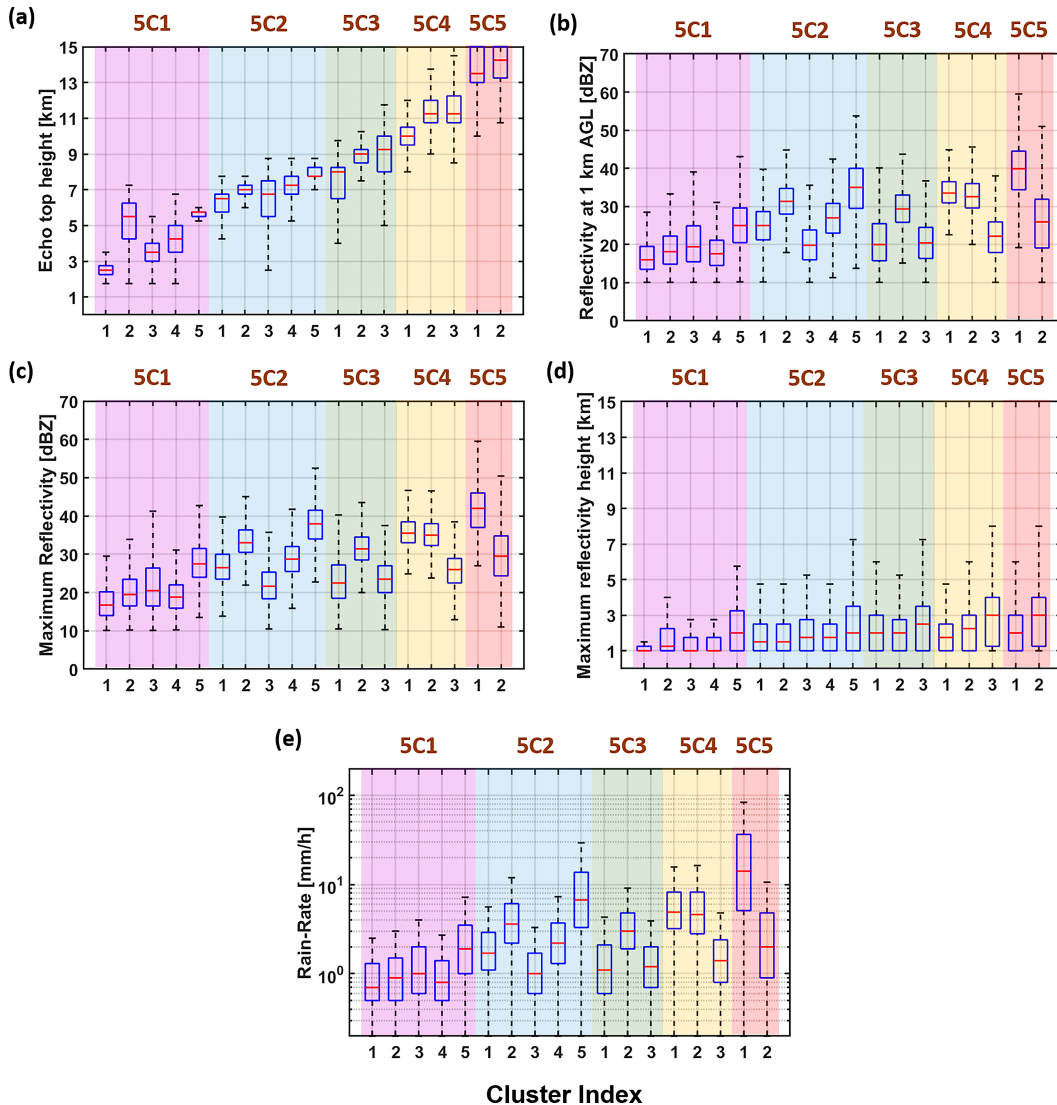
FIG. 9. Distributions of (a) echo-top height, (b) near-surface reflectivity factor, (c) maximum reflectivity, (d) height with maximum reflectivity, and (e) near-surface rain rate from MRMS. The whiskers in the boxes suggest the 25th and 75th percentile, and the red line shows the median. The color shades are consistent with Fig. 5.

pressure, relative humidity, convective available potential energy (CAPE), convective inhibition (CIN), 0° isotherm height, air temperature at 2 m above the surface, total precipitable water and total column integrated cloud water (TCW). To highlight the difference between the 18 identified clusters, Fig. 11 shows the distribution of the environmental parameters with respect to different clusters. The organization and color scheme is similar to that in Figs. 5 and 9. Cluster C1-1 shows approximately 3–4 K higher median 2-m air temperature and higher freezing level height than the C1-2, C1-3, and 5C1-4. This confirms that the profiles in C1-1 represent very shallow, light, and warm precipitation events. Clusters C1-2 and C1-5 have a similar echo-top height of 6 km AGL. However, cluster C1-2 has a lower 2 m temperature, 0° height, and CAPE compared to

cluster C1-5 suggesting that the profiles in C1-2 have a lower probability of developing into convective systems and mostly dominated belong to cool stratiform cases. C1-3 also has the lowest median 0° isotherm height suggesting the profiles in C1-3 are more prone to cold conditions than C1-1.

Reflectivity profiles in cluster C2 demonstrate the echo-top height of around 6–8 km but show higher variation with respect to rain rate. The C2-3 parent cluster has the lowest median rain rate, while C2-5 has the highest. Analysis with environmental variables for the profiles in C2-5 showed higher values for 2-m temperature, 0° isotherm height, precipitable water, and CAPE. This suggests the presence of updraft in cluster C2-5, with the events related to this cluster recognized as shallow convective systems. Cluster C2-2 also shows
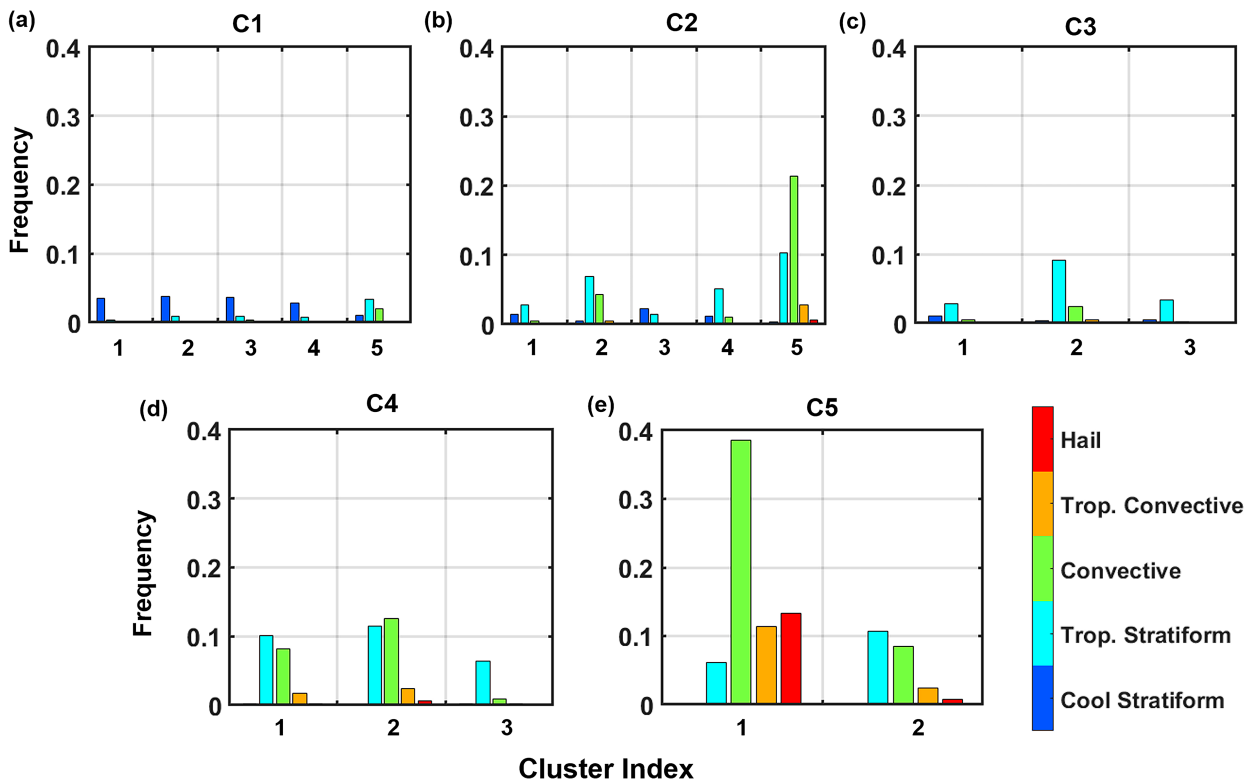
FIG. 10. Probability distribution function showing the variation of MRMS precipitation type within the macroclusters: (a) C1, (b) C2, (c) C3, (d) C4, and (e) C5.

higher CAPE variability and includes shallow convective and tropical stratiform cases. The profiles in cluster C3 have higher echo-top heights than C1 and C2, but have similar environmental variables values.

The last two clusters, C4 and C5, consist of deep precipitation events with echo-top heights greater than 8 km AGL. The total integrated cloud water values for C4-3 are lower than those of C4-2. Contrarily, C4-3 exhibits higher values of CAPE compared to that of C4-2. Both C4-2 and C4-3 show higher CIN values than the shallower profiles. Finally, C5 relates to deep precipitation events, with C5-1 having more convective cases than C5-2. This is evident from the higher CIN and column-integrated cloud water in C5-1. However, the CAPE values for C5-2 are higher, suggesting the events are in the beginning phase of convective events.

5. Discussion

The vertical profile clusters identified in this study depend only on the features derived from the autoencoder model shown in Fig. 2a. Linking these features to cloud physics demonstrates the value of deep learning algorithms. The correlation between the 8-feature autoencoder model output and the physical variables is examined. Figure 12a shows the correlation coefficient for the five structural variables shown in Fig. 9. All five variables, including echo-top height, near-surface reflectivity factor, maximum reflectivity factor, height of maximum reflectivity, and MRMS estimated rain rate, show a peak

for feature 3. The echo-top height and maximum reflectivity height also correlate highly with feature 7. This finding suggests that feature 3 consists of information regarding the reflectivity factor values, while feature 7 details the overall shape and structure of the reflectivity profiles. Other features such as feature 2, feature 5, and feature 6 also include information on maximum reflectivity height.

A similar analysis is conducted for the HRRR-based environmental variables and shown in Fig. 12b. Feature 3 demonstrates the highest correlation for most environmental variables; however, feature 1 captures the details in relative humidity, while feature 8 is highly correlated with CIN. Moreover, features 5–7 play a critical role in separating the reflectivity profiles based on 0°C height, precipitable water, and relative humidity. While some features, such as feature 3 and feature 7, capture the overall structural information, others capture the intrinsic structural differences in the reflectivity profiles, thus assisting in the physically explainable precipitation regime identification. In a nutshell, the extracted features from the autoencoder model include nonlinear information content on the structure of the reflectivity profiles that correlate well with the physical variables suggesting the dependence of precipitation regime on the vertical distribution of hydrometeors.

6. Summary and conclusions

A machine learning tool is developed to identify distinct precipitation regimes based on the vertical structure of their
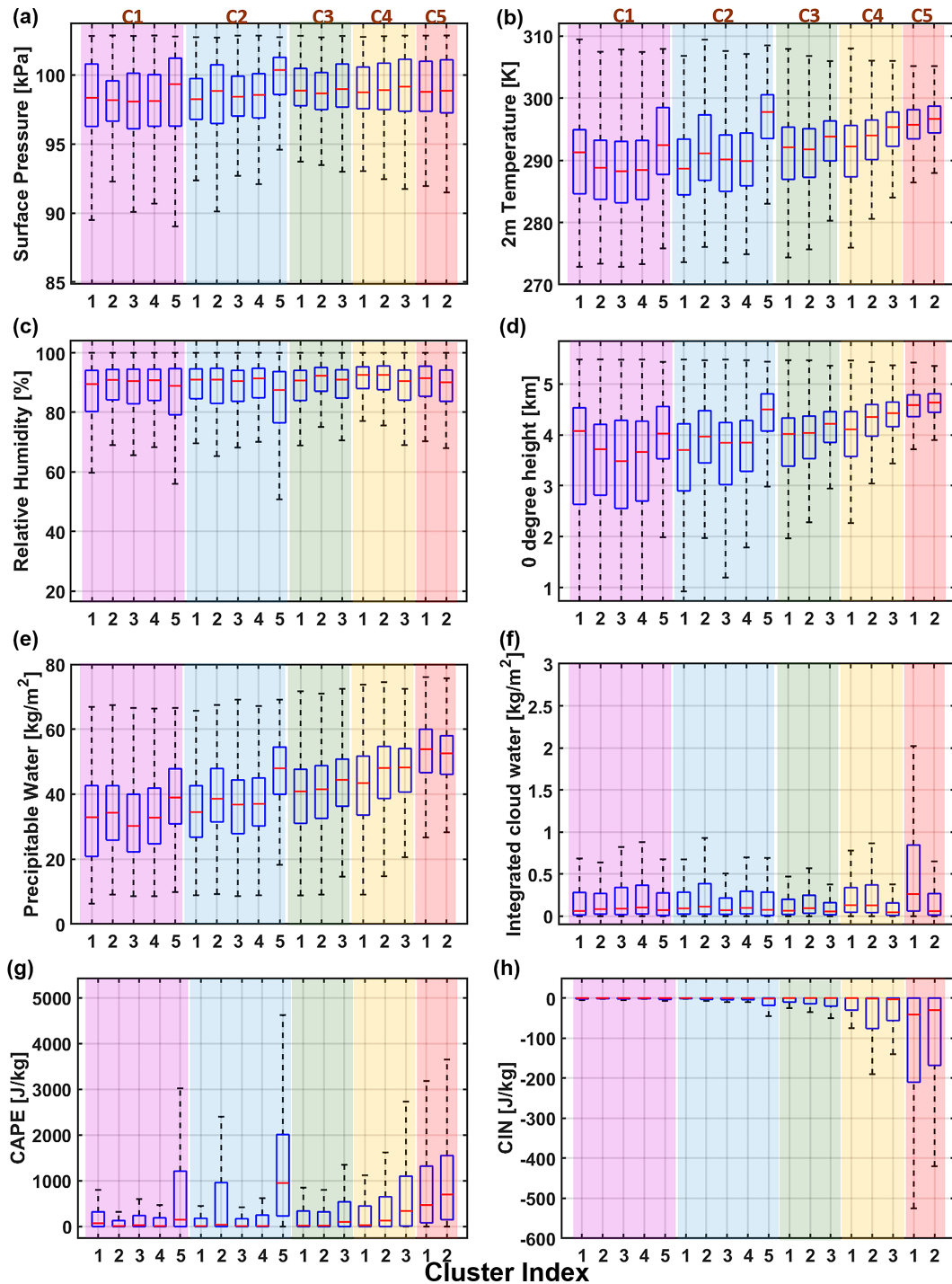
FIG. 11. Boxplots showing the variability of environmental variables: (a) surface pressure, (b) 2-m temperature, (c) relative humidity, (d) 0° isotherm height (e) precipitable water, (f) total integrated cloud water, (g) CAPE, and (h) CIN for the clusters.

reflectivity profiles. Demonstrated using the MRMS network observations, the tool employs an autoencoder and an unsupervised clustering technique constrained by the profile structure and reflectivity magnitude to define and identify different

precipitation states. The resulting regimes are linked to the precipitation type, rate, and environmental variables supporting the physical meaning of each cluster. The key findings are as follows:
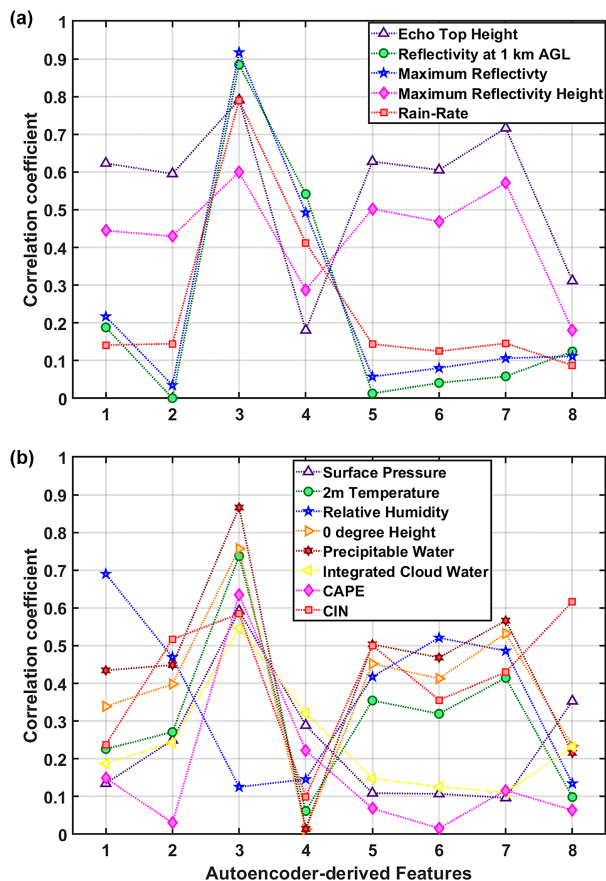
(a)

(b)

FIG. 12. Correlation between the features from the autoencoder model and (a) reflectivity-related parameters and (b) environmental variables from HRRR model.

- The proposed framework reduces the dimension of the reflectivity profiles by 85% with a mean squared error loss of approximately 0.4 dBZ. The bias for most of the profiles is within 0.1%. Moreover, the reconstructed profiles from the retrieved features are highly correlated with the observed ones.
- A coupled k-means and agglomerative clustering method identified 18 clusters that could be combined into five macrocluster groups (C1, very shallow; C2, shallow; C3, midlevel; C4, deep; and C5, very deep) based on the echo-top heights. The subclusters within the five macroclusters can capture the intrinsic structural difference.
- The first macrocluster (C1) captures very shallow precipitation profiles (echo-top height between 3 and 6 km AGL) and has five subclusters (C1-1–C1-5). The first subcluster (C1-1) is a warm shallow system, while subcluster 5 (C1-5) has higher CAPE suggesting updrafts and shallow convective systems. The other subclusters of very shallow group are identified as shallow stratiform precipitation systems.
- The second macrocluster (C2) includes profiles with echo-top heights of around 6–8 km AGL and has five subclusters (C2-1–C2-5). One of the subclusters (C2-5) in the shallow group has the highest median value of CAPE among all

clusters. C2-2 also shows higher variations with CAPE value suggesting the potential to develop into convective storms.

- The midlevel systems identified by the macrocluster, C3, show the least variability in the environment across its three subclusters (C3-1–C3-3) compared to the other macrocluster groups. The events in subcluster C3-3 have higher precipitable water content and 0° isotherm levels relative to the other subcluster groups. The CAPE and CIN values are low, suggesting the absence of convective activity within predominantly stratiform and tropical-stratiform systems characterized by brightband presence.
- The fourth macrocluster (C4) depicts the deep precipitation systems with three subcluster categories (C4-1–C4-3) and echo-top heights greater than 9 km AGL. Characteristic are higher CIN values and low variability of the CAPE across the all three subclusters. Exception is the total integrated cloud water content, with its minima found in C4-3 subgroup.
- The final macrocluster (C5) includes the deepest precipitation systems with similar echo-top heights within its two subcluster categories. C5-1 captures convective-dominated deep precipitation events of high intensity, while C5-2 includes significantly less intense precipitation events with lower amounts of integrated cloud water content and higher CAPE values.

The identified precipitation structure clusters align well with the environmental conditions and consist of more physical information than the traditional precipitation classes. Applicability of the presented tool is not limited to the region and observation sources used in this study. The method can easily be applied to any ground- and space-based observation reference, supporting local to global applications. The proposed tool can help to better understand and classify precipitation regimes, develop physics-guided retrievals and also serve as initial conditions for numerical weather models.

Data availability statement. The MRMS dataset used in this study is publicly available at https://mrms.ncep.noaa.gov/data/2D/. The archived MRMS data are accessed through IOWA Environmental Mesonet (https://mesonet.agron.iastate.edu/archive/). The HRRR environmental variables are accessed from Google Cloud storage (https://console.cloud.google.com/marketplace/details/noaa-public/hrrr).

REFERENCES

Anderberg, M. R., 1973: *Cluster Analysis for Applications*. Academic Press, 359 pp., https://doi.org/10.1016/C2013-0-06161-0.

Arulraj, M., and A. P. Barros, 2021: Automatic detection and classification of low-level orographic precipitation processes from space-borne radars using machine learning. *Remote Sens. Environ.*, **257**, 112355, https://doi.org/10.1016/j.rse.2021.112355.

Chen, T.-S., T.-H. Tsai, Y.-T. Chen, C.-C. Lin, R.-C. Chen, S.-Y. Li, and H.-Y. Chen, 2005: A combined K-means and hierarchical clustering method for improving the clustering efficiency of microarray. *2005 Int. Symp. on Intelligent Signal Processing and Communication Systems*, Hong Kong, China, Institute of Electrical and Electronics Engineers, 405–408, https://doi.org/10.1109/ISPACS.2005.1595432.

Davies, D. L., and D. W. Bouldin, 1979: A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, **1**, 224–227, https://doi.org/10.1109/TPAMI.1979.4766909.

Dolan, B., B. Fuchs, S. A. Rutledge, E. A. Barnes, and E. J. Thompson, 2018: Primary modes of global drop size distributions. *J. Atmos. Sci.*, **75**, 1453–1476, https://doi.org/10.1175/JAS-D-17-0242.1.

Dowell, D. C., and Coauthors, 2022: The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part I: Motivation and system description. *Wea. Forecasting*, **37**, 1371–1395, https://doi.org/10.1175/WAF-D-21-0151.1.

Dunn, J. C., 1974: Well-separated clusters and optimal fuzzy partitions. *J. Cybern.*, **4**, 95–104, https://doi.org/10.1080/01969727408546059.

Elsaesser, G. S., C. D. Kummerow, T. S. L'Ecuyer, Y. N. Takayabu, and S. Shige, 2010: Observed self-similarity of precipitation regimes over the tropical oceans. *J. Climate*, **23**, 2686–2698, https://doi.org/10.1175/2010JCLI3330.1.

Fred, A. L. N., and A. K. Jain, 2005: Combining multiple clusterings using evidence accumulation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**, 835–850, https://doi.org/10.1109/TPAMI.2005.113.

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.

Hence, D. A., and R. A. Houze Jr., 2011: Vertical structure of hurricane eyewalls as seen by the TRMM Precipitation Radar. *J. Atmos. Sci.*, **68**, 1637–1652, https://doi.org/10.1175/2011JAS3578.1.

Hong, Y., and J. J. Gourley, 2018: *Radar Hydrology: Principles, Models, and Applications*. CRC Press, 176 pp.

——, and Coauthors, 2019: Remote sensing precipitation: Sensors, retrievals, validations, and applications. *Observation and Measurement of Ecohydrological Processes*, X. Li and H. Vereecken, Eds., Ecohydrology, Vol. 2, Springer, 107–128, https://doi.org/10.1007/978-3-662-48297-1_4.

Hou, A. Y., and Coauthors, 2014: The Global Precipitation Measurement Mission. *Bull. Amer. Meteor. Soc.*, **95**, 701–722, https://doi.org/10.1175/BAMS-D-13-00164.1.

Houze, R. A., Jr., 1981: Structures of atmospheric precipitation systems: A global survey. *Radio Sci.*, **16**, 671–689, https://doi.org/10.1029/RS016i005p00671.

——, 2014: *Cloud Dynamics*. International Geophysics Series, Vol. 104, Academic Press, 496 pp.

Hu, J., and A. Ryzhkov, 2022: Climatology of the vertical profiles of polarimetric radar variables and retrieved microphysical parameters in continental/tropical MCSs and landfalling hurricanes. *J. Geophys. Res. Atmos.*, **127**, e2021JD035498, https://doi.org/10.1029/2021JD035498.

Iguchi, T., and Coauthors, 2010: GPM/DPR level-2. Algorithm Theoretical Basis Doc. Tech. Rep., 127 pp., https://gpm.nasa.gov/sites/default/files/document_files/ATBD_DPR_201811_with_Appendix3b_0.pdf.

Johnstone, I. M., and D. M. Titterington, 2009: Statistical challenges of high-dimensional data. *Philos. Trans. Roy. Soc.*, **A367**, 4237–4253, https://doi.org/10.1098/rsta.2009.0159.

Kirstetter, P.-E., J. J. Gourley, Y. Hong, J. Zhang, S. Moazamigoodarzi, C. Langston, and A. Arthur, 2015: Probabilistic precipitation rate estimates with ground-based radar networks. *Water Resour. Res.*, **51**, 1422–1442, https://doi.org/10.1002/2014WR015672.

Kummerow, C. D., D. L. Randel, M. Kulie, N.-Y. Wang, R. Ferraro, S. Joseph Munchak, and V. Petkovic, 2015: The evolution of the Goddard profiling algorithm to a fully parametric scheme. *J. Atmos. Oceanic Technol.*, **32**, 2265–2280, https://doi.org/10.1175/JTECH-D-15-0039.1.

Levizzani, V., and E. Cattani, 2019: Satellite remote sensing of precipitation and the terrestrial water cycle in a changing climate. *Remote Sens.*, **11**, 2301, https://doi.org/10.3390/rs11192301.

Li, Y., and H. Wu, 2012: A clustering method based on K-means algorithm. *Phys. Procedia*, **25**, 1104–1109, https://doi.org/10.1016/j.phpro.2012.03.206.

Liao, L., and R. Meneghini, 2022: GPM DPR retrievals: Algorithm, evaluation, and validation. *Remote Sens.*, **14**, 843, https://doi.org/10.3390/rs14040843.

Liu, C., D. J. Cecil, E. J. Zipser, K. Kronfeld, and R. Robertson, 2012: Relationships between lightning flash rates and radar reflectivity vertical structures in thunderstorms over the tropics and subtropics. *J. Geophys. Res.*, **117**, D06212, https://doi.org/10.1029/2011JD017123.

Luo, Z. J., R. C. Anderson, W. B. Rossow, and H. Takahashi, 2017: Tropical cloud and precipitation regimes as seen from near-simultaneous TRMM, CloudSat, and CALIPSO observations and comparison with ISCCP. *J. Geophys. Res. Atmos.*, **122**, 5988–6003, https://doi.org/10.1002/2017JD026569.

Mace, G. G., C. Jakob, and K. P. Moran, 1998: Validation of hydrometeor occurrence predicted by the ECMWF model using millimeter wave radar data. *Geophys. Res. Lett.*, **25**, 1645–1648, https://doi.org/10.1029/98GL00845.

Peterson, A. D., A. P. Ghosh, and R. Maitra, 2018: Merging K-means with hierarchical clustering for identifying general-shaped groups. *Stat*, **7**, e172, https://doi.org/10.1002/sta4.172.

Porcacchia, L., P.-E. Kirstetter, V. Maggioni, and S. Tanelli, 2019: Investigating the GPM dual-frequency precipitation radar signatures of low-level precipitation enhancement. *Quart. J. Roy. Meteor. Soc.*, **145**, 3161–3174, https://doi.org/10.1002/qj.3611.

Prat, O. P., and A. P. Barros, 2010: Ground observations to characterize the spatial gradients and vertical structure of orographic precipitation – Experiments in the inner region of the Great Smoky Mountains. *J. Hydrol.*, **391**, 141–156, https://doi.org/10.1016/j.jhydrol.2010.07.013.

Qi, Y., and J. Zhang, 2017: A physically based two-dimensional seamless reflectivity mosaic for radar QPE in the MRMS system. *J. Hydrometeor.*, **18**, 1327–1340, https://doi.org/10.1175/JHM-D-16-0197.1.

Schumacher, C., R. A. Houze Jr., and I. Kraucunas, 2004: The tropical dynamical response to latent heating estimates derived from the TRMM Precipitation Radar. *J. Atmos. Sci.*, **61**, 1341–1358, https://doi.org/10.1175/1520-0469(2004)061<1341:TTDRTL>2.0.CO;2.

Shi, L., Y. Qiu, J. Shi, and S. Zhao, 2015: Atmospheric influences analysis in passive microwave remote sensing. *2015 IEEE Int. Geoscience and Remote Sensing Symp. (IGARSS)*, Milan,

Italy, Institute of Electrical and Electronics Engineers, 2334–2337, https://doi.org/10.1109/IGARSS.2015.7326276.

Shupe, M. D., T. Uttal, S. Y. Matrosov, and A. S. Frisch, 2001: Cloud water contents and hydrometeor sizes during the FIRE Arctic Clouds Experiment. *J. Geophys. Res.*, **106**, 15 015–15 028, https://doi.org/10.1029/2000JD900476.

Singh, N., N. Garg, and J. Pant, 2014: A comprehensive study of challenges and approaches for clustering high dimensional data. *Int. J. Comput. Appl.*, **92**, 7–10, https://doi.org/10.5120/15995-4844.

Smedsmo, J. L., E. Foufoula-Georgiou, V. Vuruputur, F. Kong, and K. Droegemeier, 2005: On the vertical structure of modeled and observed deep convective storms: Insights for precipitation retrieval and microphysical parameterization. *J. Appl. Meteor.*, **44**, 1866–1884, https://doi.org/10.1175/JAM2306.1.

Tang, L., J. Zhang, M. Simpson, A. Arthur, H. Grams, Y. Wang, and C. Langston, 2020: Updates on the radar data quality control in the MRMS quantitative precipitation estimation system. *J. Atmos. Oceanic Technol.*, **37**, 1521–1537, https://doi.org/10.1175/JTECH-D-19-0165.1.

Verlinde, J., M. P. Rambukkange, E. E. Clothiaux, G. M. McFarquhar, and E. W. Eloranta, 2013: Arctic multilayered, mixed-phase cloud processes revealed in millimeter-wave cloud radar Doppler spectra. *J. Geophys. Res. Atmos.*, **118**, 13 199–13 213, https://doi.org/10.1002/2013JD020183.

Wen, Y., P. Kirstetter, J. J. Gourley, Y. Hong, A. Behrangi, and Z. Flamig, 2017: Evaluation of MRMS snowfall products over the western United States. *J. Hydrometeor.*, **18**, 1707–1713, https://doi.org/10.1175/JHM-D-16-0266.1.

Williams, C. R., A. B. White, K. S. Gage, and F. M. Ralph, 2007: Vertical structure of precipitation and related microphysics observed by NOAA profilers and TRMM during NAME 2004. *J. Climate*, **20**, 1693–1712, https://doi.org/10.1175/JCLI4102.1.

Wilson, A. M., and A. P. Barros, 2014: An investigation of warm rainfall microphysics in the southern Appalachians: Orographic enhancement via low-level seeder–feeder interactions. *J. Atmos. Sci.*, **71**, 1783–1805, https://doi.org/10.1175/JAS-D-13-0228.1.

You, Y., N. Wang, and R. Ferraro, 2015: A prototype precipitation retrieval algorithm over land using passive microwave observations stratified by surface condition and precipitation vertical structure. *J. Geophys. Res. Atmos.*, **120**, 5295–5315, https://doi.org/10.1002/2014JD022534.

Zhang, J., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 621–638, https://doi.org/10.1175/BAMS-D-14-00174.1.

Zhang, Y., S. Klein, G. G. Mace, and J. Boyle, 2007: Cluster analysis of tropical clouds using CloudSat data. *Geophys. Res. Lett.*, **34**, L12813, https://doi.org/10.1029/2007GL029336.