Environmental Data Science (2025), 4: e27, 1–21
doi:10.1017/eds.2025.17

CAMBRIDGE UNIVERSITY PRESS

APPLICATION PAPER

Investigating reduced-dimensional variability in aircraft-observed aerosol–cloud parameters

Kayley M. Butler[1], Sam J. Silva[1,2], Armin Sorooshian[3,4], Richard H. Moore[5], Glenn S. Diskin[5], John B. Nowak[5], Luke Ziemba[5], Ewan Crosbie[5,6], Michael A. Shook[5], Joshua DiGangi[5], Edward Winstead[5,6], Claire Robinson[5,6,†] and Yonghoon Choi[5,6]

[1]Department of Civil and Environmental Engineering, University of Southern California, Los Angeles, CA, USA
[2]Department of Earth Sciences, University of Southern California, Los Angeles, CA, USA
[3]Department of Chemical and Environmental Engineering, The University of Arizona, Tucson, AZ, USA
[4]Department of Hydrology and Atmospheric Sciences, The University of Arizona, Tucson, AZ, USA
[5]NASA Langley Research Center, Hampton, VA, USA
[6]Analytical Mechanics Associates, Hampton, VA, USA
Corresponding author: Kayley M. Butler; Email: kayleybu@usc.edu

[†]C.R. is deceased.

Received: 09 July 2024; **Revised:** 19 February 2025; **Accepted:** 07 April 2025

Keywords: aerosol; atmospheric composition; clouds; dimensionality reduction; machine learning

Abstract

Aerosol-cloud interactions contribute significant uncertainty to modern climate model predictions. Analysis of complex observed aerosol-cloud parameter relationships is a crucial piece of reducing this uncertainty. Here, we apply two machine learning methods to explore variability in in-situ observations from the NASA ACTIVATE mission. These observations consist of flights over the Western North Atlantic Ocean, providing a large repository of data including aerosol, meteorological, and microphysical conditions in and out of clouds. We investigate this dataset using principal component analysis (PCA), a linear dimensionality reduction technique, and an autoencoder, a deep learning non-linear dimensionality reduction technique. We find that we can reduce the dimensionality of the parameter space by more than a factor of 2 and verify that the deep learning method outperforms a PCA baseline by two orders of magnitude. Analysis in the low dimensional space of both these techniques reveals two consistent physically interpretable regimes—a low pollution regime and an in-cloud regime. Through this work, we show that unsupervised machine learning techniques can learn useful information from in-situ atmospheric observations and provide interpretable results of low-dimensional variability.

Impact Statement

To eventually better understand aerosol-cloud interactions, we analyze the patterns in a large dataset of aircraft observations. We use a machine learning method to look at the underlying behavior within this dataset called an autoencoder and compare it to a more simple technique called principal component analysis as a baseline. We demonstrate that these dimensionality reduction techniques are useful for compressing in-situ atmospheric data with precision and that the autoencoder performs two orders of magnitude better than the principal component analysis baseline.

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

1. Introduction

Atmospheric aerosols, which are solid or liquid particles suspended in air, have a substantial impact on the radiative budget of Earth. These aerosols affect radiation in several ways: through direct absorption and scattering of radiation, and indirectly through modulating cloud formation, lifetime, and radiative properties (Twomey, 1959, 1974, 1977; Albrecht, 1989; Rosenfeld et al., 2008). The magnitude of these effects is large. The total effective radiative forcing associated with the direct effects of aerosols is -0.3 ± 0.3 W m^{-2} while aerosol interactions with clouds contribute -1.0 ± 0.7 W m^{-2} to the earth's radiative forcing (IPCC, 2021). The net effect of aerosol accounts for 18% of total anthropogenic radiative forcing (IPCC, 2021). Although this radiative impact is large, the interactions between aerosols and clouds contribute the largest uncertainty of any radiative forcing category to the physical climate system (National Academies of Sciences, Engineering, and Medicine, 2018; Bellouin et al., 2020; IPCC, 2021). Addressing this uncertainty is necessary for fine-tuning our predictive capability for the Earth System.

A vital step in reducing uncertainty in aerosol–cloud forcing is collecting and analyzing additional in-situ measurements of aerosol–cloud processes in varied atmospheric conditions (Seinfeld et al., 2016). The recent NASA ACTIVATE field campaign (Aerosol Cloud meTeorology Interactions oVer the western ATlantic Experiment) contributes to this with an immense dataset of measurements in and out of marine boundary layer (MBL) clouds (Sorooshian et al., 2023). A main goal of this campaign was to build a repository of in-situ and remotely sensed flight data with as many hours as possible. Over the course of three years (2020–2022), this campaign measured over 200 parameters across 574 flight hours. Despite the abundance of data available, it is challenging to discern which parameters are driving the variability of the aerosol-cloud processes in this region of the Atlantic. As is the case with many in-situ observational datasets, the substantial observed variability can be reduced through various analytical strategies to find key drivers and regimes.

Machine learning methods offer potential solutions to the analysis of large complex datasets like those from in-situ aircraft observations (e.g., Marais et al., 2020; Silva et al., 2022; Mohn et al., 2023; Mooers et al., 2023b). One such method is a class of techniques known as dimensionality reduction, which compresses datasets into a latent space of fewer dimensions than the original output. These latent spaces can often capture patterns of variability that are more human-interpretable. Here we investigate the performance of both linear and non-linear dimensionality reduction techniques on in-situ observations of aerosol–cloud parameters. The linear method explored here is Principal Component Analysis (PCA). PCA reduces dimensionality by mapping data onto new orthogonal axes and has been widely used in the atmospheric sciences (e.g., Achakulwisut et al., 2017; Fiore et al., 2022). For non-linear dimensionality reduction, we use an autoencoder. Autoencoders are a deep learning method where the information is encoded into a hidden bottleneck layer of a neural network. Recent work has demonstrated the utility of autoencoder methods for learning useful patterns and observed states in atmospheric science data (e.g., Behrens et al., 2022; Kurihana et al., 2022; Lerch and Polsterer, 2022). While a variety of studies have used reduced dimensional techniques to study aerosol and cloud processes (e.g., Jones and Christopher, 2010; Yorks et al., 2021), little work has been done on in-situ aircraft data.

In this paper, we use these dimensionality reduction techniques to investigate the variability in observed aerosol–cloud parameters. We use the NASA ACTIVATE dataset from the 2020 flight year and develop methods to specifically address challenges in applying machine learning to in-situ aircraft data—including a custom loss function for treating skewed data distributions. We evaluate the use of both PCA and autoencoders across a range of model parameters and observed atmospheric conditions. We find that the autoencoder substantially outperforms PCA in reconstruction accuracy, and both methods highlight the same two physically interpretable regimes. This work is a proof-of-concept solution to known challenges with using machine learning methods on in-situ aircraft observations (Dadashazar et al., 2021), and demonstrates the potential for dimensionality reduction techniques in this space.

2. Data and methods

We use data from the NASA ACTIVATE field campaign (Sorooshian et al., 2023). The ACTIVATE mission was an aircraft campaign focused on collecting extensive meteorological, microphysical, and

https://doi.org/10.1017/eds.2025.17 Published online by Cambridge University Press

chemical measurements to study aerosol–cloud interactions over the Western North Atlantic Ocean (WNAO) (Sorooshian et al., 2023). Aerosol–cloud processes in the WNAO region are influenced by a wide range of airmass sources, including anthropogenic emissions from the United States Eastern Seaboard, local oceanic processes, wildfire smoke from the Western United States, and dust from the Sahara Desert, making measurements and source attribution diverse (Sorooshian et al., 2019). The fieldwork conducted in this campaign presents valuable information for studying variability in aerosol–cloud processes and is extensive in its measurements, documenting over 200 observed parameters altogether. Our focus in this work is to use data-driven methods to investigate which of these parameters drive variability in the aerosol–cloud system. Here, we use the 2020 "merge files" (see sect 4.8 of Sorooshian et al., 2023) created by the ACTIVATE team at one second resolution and data specific to the HU-25 Falcon aircraft. The merge files have all the in-situ datasets on a similar time base using weighted time averages when needed.

Although ACTIVATE employed two spatially coordinated aircraft, the focus of this work is the lower-flying HU-25 Falcon. A total of 22 flights were conducted in the winter of 2020 (14 February—12 March) along with 18 more in the summer of 2020 (13 August—30 September). This amounted to approximately 73 and 60 total flight hours with the HU-25 Falcon for winter and summer deployments, respectively. Flights were based out of either NASA Langley Research Center or the Newport News/Williamsburg Airport, both of which are based near each other around the Hampton, Virginia area. These were out-and-back flights that focused on sampling areas offshore over the northwest Atlantic. The Falcon typically sampled below 3 km as its focus was sampling within and just above the marine boundary layer to characterize boundary layer clouds.

2.1. Data selection, filtering, and preprocessing

We select a subset of meteorological, chemical, and microphysical variables from the ACTIVATE campaign relevant to aerosol–cloud interactions. We balance including as many variables as possible for scientific discovery while including as much data as possible for data-driven machine learning. This tradeoff exists due to gaps in observational data records. Our subset selection was somewhat subjective but guided by literature as follows.

Initial variable selection was to limit analyses to warm clouds in the boundary layer, which was what was dominantly observed by the HU-25 Falcon. We removed aircraft-state observations (pitch, roll, heading, etc.), as well as local time and geolocation information (latitude and longitude). We further refined our variable list to focus on observations that were most widely studied in similar related papers (e.g., Fountoukis et al., 2007; McComiskey et al., 2009; Jones and Christopher, 2010). As a consequence of this literature search, we focused on using bulk aerosol and cloud droplet variables instead of their full observed size distributions. In accordance with guidance from instrument PIs on the ACTIVATE team, individual measurements marked as missing or otherwise flagged were removed.

Once a large subset of initial variables were identified, we evaluated the density of available data within each category. This is key because most modern machine learning methods require complete data records (i.e., complete rows in a tabular dataset). For the time period considered in this work, we calculate the percentage of data available for each of these variables. We observed a critical lack of data availability for some of the variables of interest, and so they were removed. This data density issue is a hallmark challenge of using observational data for machine learning instead of model data as in-situ measurements frequently have missing data. The final observational data we use is listed in Table 1.

Pressure altitude is measured by the Applanix 610 (Sorooshian et al., 2023). Static air temperature is measured with a Rosemount Model 102 non-deiced total air temperature sensor with a fast-response platinum sensing element (E102E4AL) (Thornhill et al., 2003). The NASA Langley Turbulent Air Motion Measurement System (TAMMS) supplies measurements of vertical velocity (Thornhill et al., 2003). Relative humidity was derived from water vapor mixing ratio measurements made by the Diode

https://doi.org/10.1017/eds.2025.17 Published online by Cambridge University Press

Table 1. *Variables used in the machine learning techniques*

Metric	Instrument	Unit
Altitude	Applanix 610	ft
Static air temperature	Rosemount Model 102	Celsius
Vertical velocity (w)	TAMMS	m s^{-1}
Relative humidity (RH)	DLH	%
Carbon monoxide concentration (CO)	PICARRO G2401-m	ppm
Ozone concentration (O_3)	2B Tech Model 205	ppbv
Cloud condensation nuclei concentration (CCN)	DMT Scanning Flow CCN Counter	# cm^{-3}
Liquid water content (LWC)	SPEC 2DS	kg m^{-3}
Effective radius (R_{eff})	SPEC FCDP	μm
Number concentration (N)	SPEC FCDP	# cm^{-3}

Laser Hygrometer (DLH) (Diskin et al., 2002). The DLH measures water vapor along an open external optical path using wavelength-modulated laser absorption. Chemical concentrations of carbon monoxide and ozone concentration were measured with the PICARRO G2401-m (DiGangi et al., 2021) and 2B Tech Model 205 (Wei et al., 2021), respectively. The Langley Aerosol Research Group (LARGE) instrument team attached probes to the aircraft to measure aerosol and cloud droplet size distributions (i.e., size-resolved concentrations). Measurements from the Fast Cloud Droplet Probe (FCDP) give aerosol and cloud number concentration and effective diameter (divided by 2 here to give effective radius) (Kirschler et al., 2022) while liquid water content (LWC) is measured by the SPEC two-dimensional stereo instrument (2DS) (Kirschler et al., 2023). Cloud condensation nuclei (CCN) number concentration was measured with a DMT Scanning Flow CCN Counter (Moore and Nenes, 2009). More details about the products provided in the ACTIVATE dataset can be found in Sorooshian et al. (2023).

With the list of variables listed in Table 1, we filter these observations following recommended data quality screening, including removing individual timestep measurements flagged by instrument PIs as missing (Sorooshian et al., 2023). We additionally filter data to remove periods marked as take-off or landing, see Supplementary Material (S1). To ensure a useful comparison of microphysical measurements, we filter and keep CCN measurements within a specific supersaturation range of 0.3–0.5%. We then removed rows where one or more of our chosen parameters had measurements that were missing. After filtering was complete, the clean dataset had 34,277 instances for further analysis.

All data are normalized before training following best practices for machine learning model training on atmospheric data (e.g., Geiss et al., 2022). Many of the features investigated here have log-normal or zero-inflated distributions, which cannot be trivially normalized following traditional standardization methods (e.g., z-score or min-max). Instead, we normalize each variable individually. For CCN concentration, we apply a box-cox transformation to achieve a more Gaussian distribution since the original distribution was heavily skewed. To aid with model training, we binarize LWC to serve as a cloud flag: 1 if greater than or equal to 0.2 x 10^{-5} kg m^{-3} representing in cloud and 0 if less than 0.2 × 10^{-5} kg m^{-3} representing out of cloud under the guidance of instrument PIs (Table 1). In the final normalization step, we subtract the minimum value of the parameter and divide it by the standard deviation of the parameter. The post-processing distributions of the parameters are shown in Figure 1 where air temperature, vertical velocity, CO, O_3, and CCN concentration have a roughly normal distribution, altitude, effective radius, and number concentration are slightly right skewed, and LWC is binary.

For model training, the data is randomly split into training (70%), validation (10%), and test (20%) sets. We use the training and validation sets to optimize the model parameters and hyperparameters, respectively, and withhold the test set for model evaluation.

https://doi.org/10.1017/eds.2025.17 Published online by Cambridge University Press

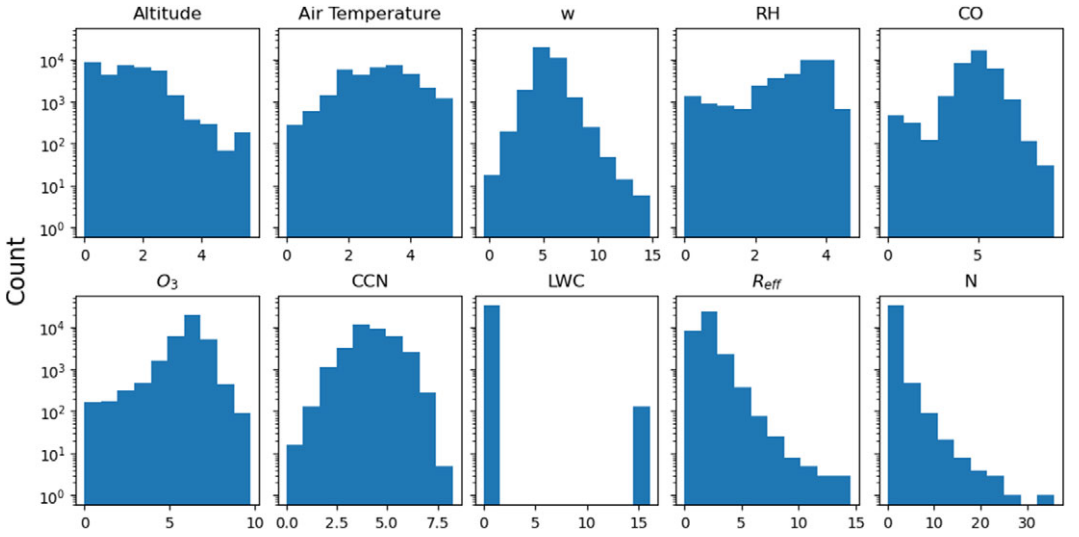

Figure 1. *Distribution of normalized parameters after post-processing, therefore values are unitless.*

2.2. Dimensionality reduction

2.2.1. Principal component analysis

Principal component analysis (PCA) is a linear dimensionality reduction technique that is widely used in atmospheric and environmental research (e.g., Jones and Christopher, 2010; Achakulwisut et al., 2017; Fiore et al., 2022; Benestad et al., 2023). Sometimes called empirical orthogonal functions (EOF) in the atmospheric science literature, PCA linearly transforms data input from n dimensions onto m new orthogonal axes. Each of these new axes, or principal components (PCs), explains some percentage of variability in the dataset. PC 1 explains the most, followed by PC 2, then PC 3, and so on. By restricting m to a value lower than the original n dimensions of the input, the data can be projected onto a reduced dimensional PC space. We use the scikit learn package to conduct PCA (Pedregosa et al., 2011). The main hyperparameter that must be selected is the m number of new orthogonal axes. Here, we use the Minka 'mle' method (built into the scikit-learn software) for optimizing m (Minka, 2000).

This dimensionality reduction technique has been used in many applications across the earth and atmospheric sciences to compress a large dataset into more interpretable dimensions. For example, PCA was used to determine factors controlling fine dust variability in the Western United States from a combination of ground observations and reanalysis products (Achakulwisut et al., 2017). By using PCA, 54–61% of spring fine dust variability can be attributed to only two latent space dimensions and specifically to key metrics such as regional precipitation, surface temperature, and soil moisture anomalies (Achakulwisut et al., 2017). PCA has also been used in assessing climate model output with the larger goal of discovering drivers in variability. PCs revealed distinct spatial regions which help in the analysis of pollution trends over decadal timescales over the eastern United States (Fiore et al., 2022).

Our use of PCA is two-fold. First, we use PCA as it is nominally used in dimensionality reduction analysis: optimize and analyze m principal components, which describe the variability in the dataset. We explore the organization of data in this reduced dimensional space by identifying regimes and relationships between variables and PCs. In this way, we are evaluating the linear decomposition of the data. Second, we use PCA as a methodological baseline to assess the performance of a more detailed deep-learning approach.

2.2.2. Autoencoder

We further explore dimensionality reduction with an autoencoder, a non-linear dimensionality reduction method. An autoencoder is a feed-forward neural network that is trained to predict its own inputs (i.e., as

https://doi.org/10.1017/eds.2025.17 Published online by Cambridge University Press

an approximation to the identity function). Autoencoders are constructed such that information flows through a low-dimensional "bottleneck" layer. We explore the information in that bottleneck space as a reduced-dimensional representation of the input data.

Autoencoders have been used in a variety of studies to explore low-dimensional variability in atmospheric datasets (e.g., Behrens et al., 2022; Kurihana et al., 2022). For example, Behrens et al. use a type of autoencoder called a variational autoencoder and reduce the dimensionality of climate model output by a factor of 12 (Behrens et al., 2022). They found that this low-dimensional latent space effectively represents large-scale climate variability and convective regimes (Behrens et al., 2022). Kurihana et al. use an autoencoder to capture the most relevant features of a large dataset of cloud satellite imagery. They find that an autoencoder can compress their dataset size by a factor of 15,000 and produces more descriptive categories than the nine categories previously established by the International Satellite Cloud Climatology Project (ISCCP) (Kurihana et al., 2022).

An example of neural network architecture for an autoencoder is shown in Figure 2. The autoencoder is composed of an input layer, followed by encoding layers, which bottleneck into a hidden layer, ultimately reversing the structure back to the size of the input layer through decoding layers. An autoencoder of this nature is trained by optimizing a set of weights and biases in each layer. These are commonly optimized by stochastic gradient descent adjusting to target a reduction in training loss calculated by a loss function (e.g., Chollet, 2021). For our purposes, we use the variables listed in Table 1 as input, where the data are then passed through the encoding layers, compressing the data into nodes within the hidden bottleneck layer. The hidden bottleneck layer feeds into a decoder, which expands the data back out and reconstructs the original input in the output layer. In this way, the autoencoder learns how to non-linearly encapsulate the data into the reduced dimensional space of the bottleneck layer during training.

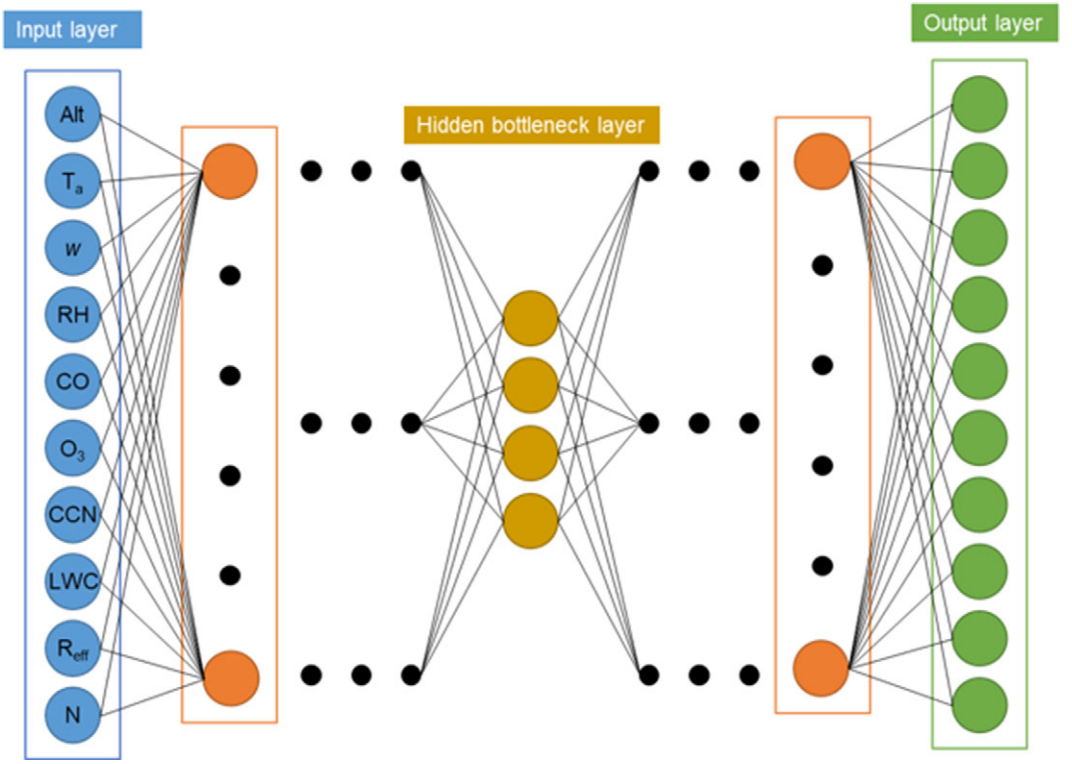

Figure 2. *Example of an autoencoder structure. The input layer is shown in blue, hidden layers are in orange, the bottleneck layer is in yellow, and the output layer is in green.*

https://doi.org/10.1017/eds.2025.17 Published online by Cambridge University Press

Table 2. *Comparison of PCA and Autoencoder technique features*

Feature	PCA	Autoencoder
Variability captured	Linear	Linear and non-linear
Method of dimensionality reduction	Finds best-fit new orthogonal axes across all variables	Subsequent encoding layers of descending node size feed into bottleneck layer with fewer nodes
New axes	Individual principal components	Individual nodes in hidden bottleneck layer

Autoencoders have a set of hyperparameters that are set prior to model training, which are described as follows. We use ReLU (rectified linear) activation function for all nodes. The learning rate is set to 0.001. The optimization method in this case is rmsprop (Hinton et al., 2012). We use a custom loss function, described in a later section, to assess the performance of the model. The batch size is also set to the default of 32. To help prevent overfitting, we use an early stopping of 25 epochs to halt training if validation loss does not improve.

To evaluate the autoencoder, we treat each node in the bottleneck layer as a new axis to plot the data and much like the PCA analysis, look for regimes and relationships between variables and nodes. A comparison of the two dimensionality reduction approaches in this study is included in the table below (Table 2).

3. Model training and performance

We complete model training for both PCA and the autoencoder after the pre-processing of the dataset and assess performance of both models on previously unseen test data.

3.1. Principal component analysis

Using the scikit-learn optimal number of principal components (Minka, 2000), we find nine principal components can explain 98.4% of the variance in the dataset. Figure 3 shows the associated scree plot,

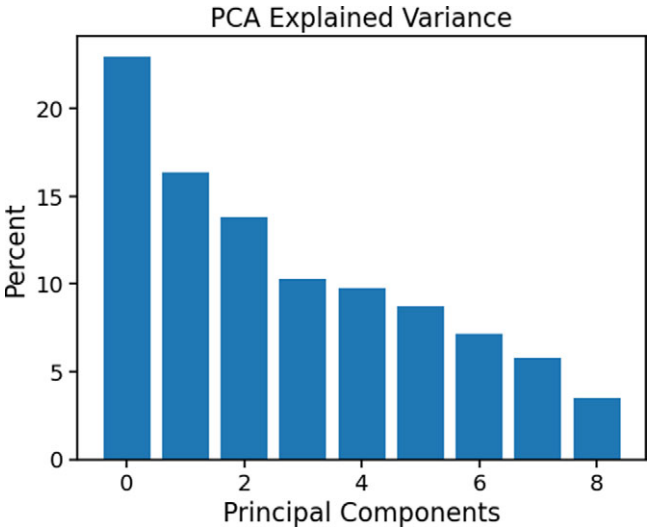

Figure 3. *Scree plot of principal component analysis (PCA) explained variance.*

https://doi.org/10.1017/eds.2025.17 Published online by Cambridge University Press

which shows how much variance is described by each principal component. From Figure 3, the majority (63.5%) of the variance in the entire dataset can be explained by the first four principal components. As a result, we use the first four principal components from this PCA to investigate the reduced dimensional space of our dataset given by this method.

3.2. Autoencoder

3.2.1. Network Architecture

We build an optimized autoencoder architecture specific to this learning task. This involves selecting an optimal number of layers and nodes in the encoder and decoder, as well as selecting an appropriate number of nodes in the hidden bottleneck layer. The encoding and decoding layers and the units in each layer are optimized using the KerasTuner Hyperband hyperparameter optimization method with a mean squared error loss (Li et al., 2018). We optimize the architecture for performance on our dataset by first using KerasTuner to find a candidate number of units for each layer, ranging from 1 to 100, and the number of layers in the encoder and decoder ranging from 1 to 5. We then fine-tune the KerasTuner hyperparameters to be sure that performance on each individual variable is good.

We separately optimize the number of units in the latent space by assessing the validation mean squared error for the autoencoder as a function of the size of the reduced-dimensional space. We compare this error against the validation loss from a PCA baseline, with results shown in Figure 4. For PCA, encoding dimensions represent the number of new orthogonal axes of the latent space. In the autoencoder, the encoding dimensions are the number of nodes in the hidden bottleneck layer. We ultimately chose 4 encoding dimensions since there is an "elbow" in the PCA plot at 3 encoding dimensions and a decent drop in loss between 3 and 5 dimensions for the autoencoder. In the end, this choice of encoding

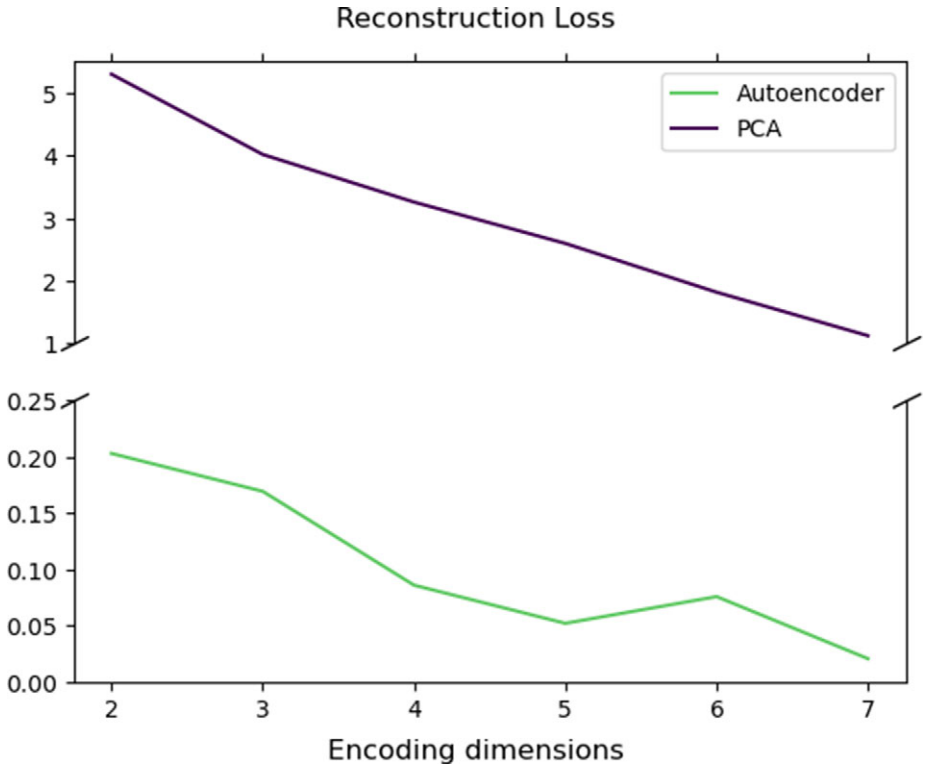

Figure 4. *Assessing autoencoder hidden layer units and principal component analysis (PCA) encoding dimensions. Note the break in the y-axis.*

https://doi.org/10.1017/eds.2025.17 Published online by Cambridge University Press

Table 3. *Autoencoder architecture*

Layer	Type	Units
Input_layer	Input	10
Encoding_1	Dense	81
Hidden_layer	Dense	4
Decoding_1	Dense	91
Decoding_2	Dense	21
Decoding_3	Dense	11
Output_layer	Dense	10

dimensions is subjective and balances keeping the number of dimensions low for understandability against the need for accuracy (which increases with the size of the number of encoding dimensions), similar to previous works (e.g., Behrens et al., 2022; Mooers et al., 2024).

The final architecture structure is illustrated in Table 3 and follows what is shown in Figure 2. The input layer of 10 units is followed by one dense encoding layer, the dense hidden layer of four units, three dense decoding layers, and one dense output layer of 10 units.

3.2.2. Custom loss

A few of the metrics included in our analysis are heavily zero-inflated, especially parameters like LWC, number concentration, and effective radius (see Figure 1). With zero-inflated data, the model can achieve high predictive accuracy by always predicting values of 0, with incorrect predictions for all non-zero values. To account for this during training, we build our own custom loss that applies a fractional penalty to values above a threshold.

The equation of our custom loss follows these conditions:

$$if\ y_{true} > threshold: \tag{1}$$

$$penalty \cdot \frac{1}{N}\sum_{i=1}^{n}\left(y_{predicted} - y_{true}\right)^2$$

$$else:$$

$$\frac{1}{N}\sum_{i=1}^{n}\left(y_{predicted} - y_{true}\right)^2$$

where y_{true} is the true value of a particular instance and $y_{predicted}$ is the associated predicted value from the autoencoder, n is the n^{th} instance in the dataset, and N is the total number of instances. If y_{true} is greater than the threshold, a penalty is applied to the mean-squared error.

For values below the threshold value, mean squared error is applied (Equation 1). In practice, this loss down-weights the relative importance of accurate prediction of low (near zero) values in the dataset and increases the relative importance of accurate prediction of high values. The value of the penalty and threshold is determined by a grid search ranging from 0 to 100 for the penalty and 0 to 10 for the threshold. With each combination of penalty and threshold in place, the mean squared error of each variable in reconstruction was calculated. The penalty and threshold values are chosen based on favorable values of mean squared error across all variables with a final threshold of 0.0 and a final penalty of 0.1.

3.2.3. Autoencoder training

For our autoencoder, we use the 70/20/10 split for training, testing, and validation, respectively, as described in the Data Filtering and Pre-processing section. Figure 5 shows the loss after each epoch for

https://doi.org/10.1017/eds.2025.17 Published online by Cambridge University Press

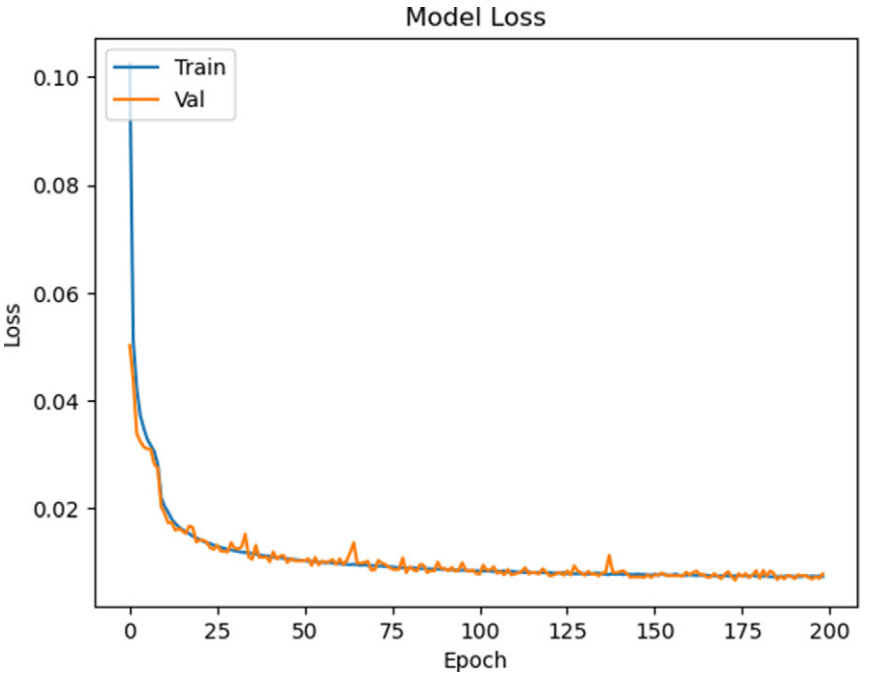

Figure 5. *Autoencoder model training and validation loss as a function of epoch.*

both the training and validation sets. This loss curve shows that there was no apparent overfitting of the validation data and that the model converges.

3.3. Performance assessment

We assess the quality of the dimensionality reduction methods by their respective errors in reconstructing the original data using the unseen test set. To compare the performance of the two different machine learning approaches, we evaluate the predictions and true values and calculate the mean squared error between the predicted and true values as shown in Equation 2:

$$\frac{1}{N}\sum_{i=1}^{n}\left(y_{predicted}-y_{true}\right)^2 \tag{2}$$

where y_{true} is the true value from the dataset and $y_{predicted}$ is the predicted value from the autoencoder, n is the n^{th} instance in the dataset, and N is the total number of instances.

The total reconstruction mean squared error for PCA is fairly small, being 1.67×10^{-1} on the test set for the normalized variables. Mean squared errors for individual features with normalized units are summarized in Table 4 and range from 1.83×10^{-1} to 6.73×10^{-1}. In terms of MSE, PCA reconstructs air temperature and carbon monoxide the best but number concentration and liquid water content poorly. The density plot results for each feature are shown for PCA in Figure 6 where the predicted value is plotted versus the true value. In general, there is a fair prediction of 6 of the 10 variables. The exceptions are vertical velocity, LWC, effective radius, and number concentration, where prediction errors are large in magnitude and skew off the 1:1 line. For vertical velocity, measurements with values in the extremes are not well predicted (above and below a value of 5) and there is a subset of values with 5 as the true value, which are predicted to have negative values in PCA transformation (Figure 6c). LWC does not have a good prediction for out-of-cloud data points (Figure 6h). For true values of R_{eff} and number concentration,

https://doi.org/10.1017/eds.2025.17 Published online by Cambridge University Press

Table 4. *Individual mean squared error (MSE) from PCA*

Metric	MSE
Altitude	2.07×10^{-1}
Air temperature	1.83×10^{-1}
Vertical velocity	4.08×10^{-1}
Relative humidity	2.68×10^{-1}
Carbon monoxide	1.92×10^{-1}
Ozone	3.94×10^{-1}
Cloud condensation nuclei	3.44×10^{-1}
Liquid water content	6.73×10^{-1}
Effective radius	4.18×10^{-1}
Number concentration	5.67×10^{-1}
Total reconstruction loss	1.67×10^{-1}

there is an underprediction by the PCA, suggesting that the PCA is not able to reconstruct the higher values of R_{eff} and number concentration (Figure 6i, j).

An analogous assessment of the autoencoder is shown in Figure 7. We find the total reconstruction mean squared error (MSE) for the autoencoder to be 8.67×10^{-3} for the test set, nearly two full orders of magnitude lower than the PCA approach. The MSE for each normalized variable is included in Table 5 and ranges from 1.65×10^{-2} to 1.36×10^{-1}. Compared to PCA, the error for each individual metric predicted by the autoencoder is less than the PCA approach by at least a factor of two. There is marked improvement in the autoencoder MSE for altitude, air temperature, vertical velocity, relative humidity, liquid water content, and number concentration over the PCA approach where the reconstruction error improves by an order of magnitude. Reconstruction error of carbon monoxide, ozone, cloud condensation nuclei, and effective radius also improves but to a lesser extent.

The density plots in Figure 7 demonstrate that there is generally a much better prediction of all variables for the autoencoder than PCA. Predicted values tend to cluster closer to the 1:1 line, with higher correlations and fewer outliers. In particular, the autoencoder better reproduces the extrema of the true values. This is particularly evident in R_{eff} and N (Figure 7i,j). Additionally, a subset of vertical velocity predictions near the mean of the true distribution are more accurately predicted along with a better prediction of LWC values of zero (Figure 7c,h). Overall, the autoencoder efficiently reconstructs the dataset over the range of their values, offering a robust and representative reduced dimensional space.

4. Results and discussion

We investigate the reduced dimensional space of a subset of in-situ observations collected through the ACTIVATE campaign. We focus on four encoding dimensions with PCA and an autoencoder and explore potential low-dimensional regimes present in the data, which may be indicative of particular observed states or unique modes of variability. For both types of dimensionality reduction, we plot all unseen test data in the latent spaces. We focus discussion on identified regimes and variables of scientific interest here, with additional figures in the Supplementary Material for completeness.

4.1. Principal component analysis

To investigate the data in the reduced dimensional space provided by PCA, we plot the test data on the first four principal components, shown in Figure 8. Some distinct regimes are found in the latent spaces as some of the principal components set groups of datapoints apart from the main data distribution. Although PC1 explains most of the variance in the dataset (Figure 3), we do not identify discernable structures in the first two panels (Figure 8a,b). However, we identify a specific regime, which is set apart by PC4

https://doi.org/10.1017/eds.2025.17 Published online by Cambridge University Press

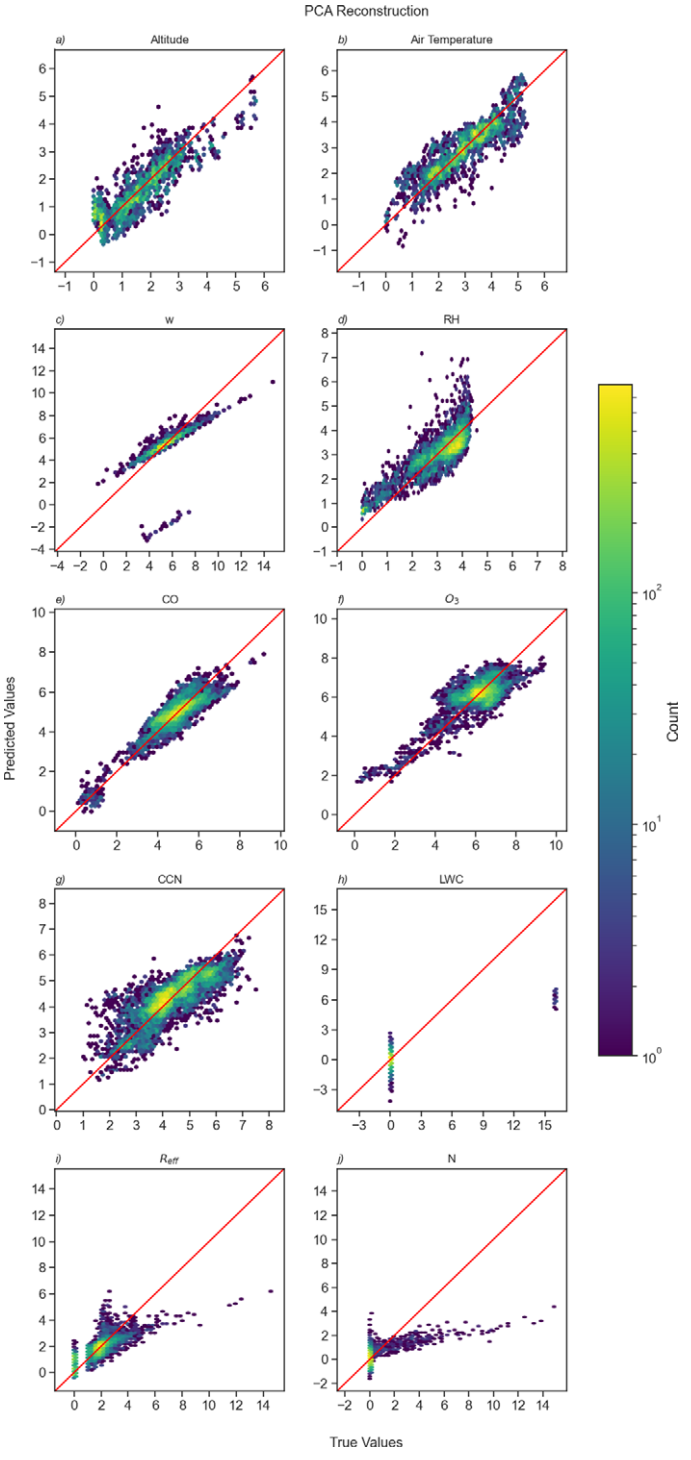

Figure 6. *Principal component analysis (PCA) performance for test data a) altitude, b) air temperature, c) vertical velocity (w), d) relative humidity (RH), e) carbon monoxide (CO), f) ozone (O_3), g) cloud condensation nuclei (CCN), h) liquid water content (LWC), i) effective radius (R_{eff}), j) number concentration (N). All units are normalized following the procedure outlined in Section 2.1.*

https://doi.org/10.1017/eds.2025.17 Published online by Cambridge University Press

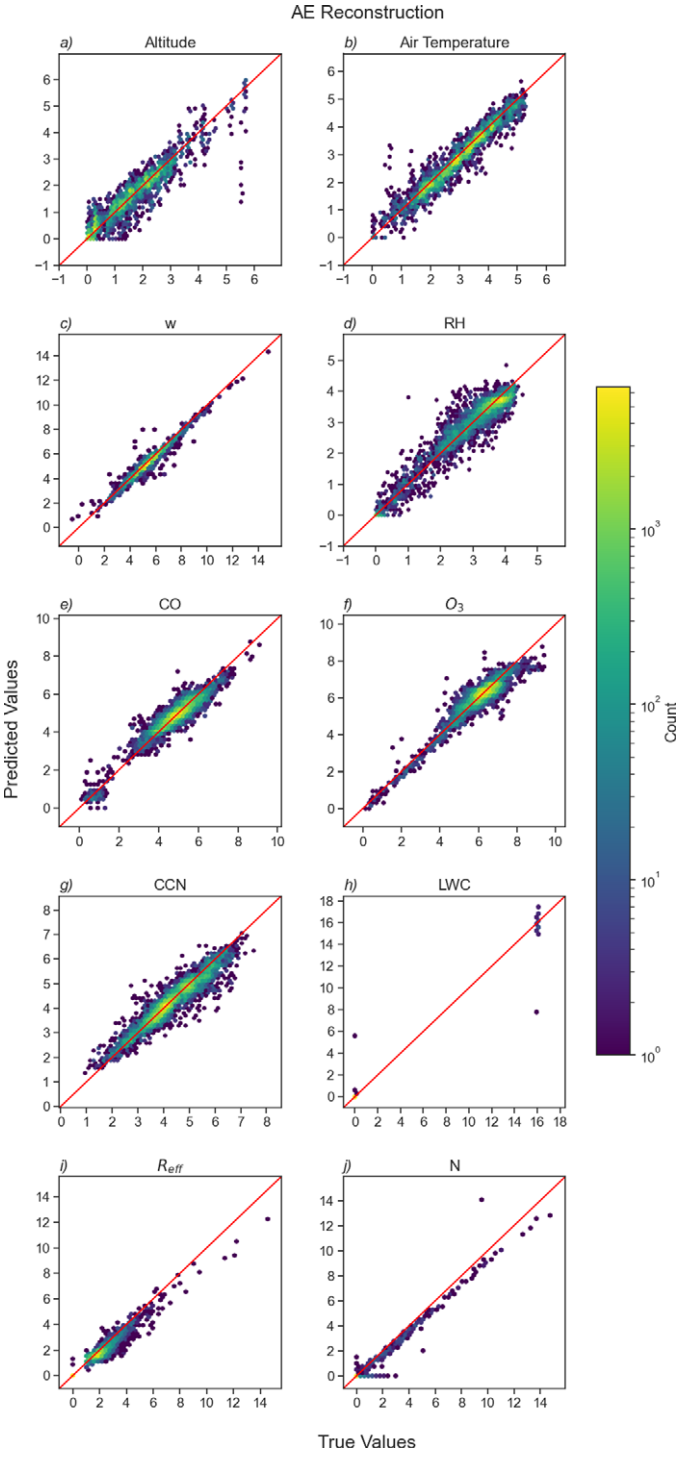

Figure 7. *Autoencoder performance for test data a) altitude, b) air temperature, c) vertical velocity (w), d) relative humidity (RH), e) carbon monoxide (CO), f) ozone (O_3), g) cloud condensation nuclei (CCN), h) liquid water content (LWC), i) effective radius (R_{eff}), j) number concentration (N). All units are normalized following the procedure outlined in Section 2.1.*

https://doi.org/10.1017/eds.2025.17 Published online by Cambridge University Press

Table 5. *Individual mean squared error (MSE) from autoencoder*

Metric	MSE
Altitude	9.88×10^{-2}
Air temperature	6.54×10^{-2}
Vertical velocity	5.66×10^{-2}
Relative humidity	7.97×10^{-2}
Carbon monoxide	1.14×10^{-1}
Ozone	1.24×10^{-1}
Cloud condensation nuclei	1.01×10^{-1}
Liquid water content	1.65×10^{-2}
Effective radius	1.36×10^{-1}
Number concentration	2.58×10^{-2}
Total reconstruction loss	8.67×10^{-3}

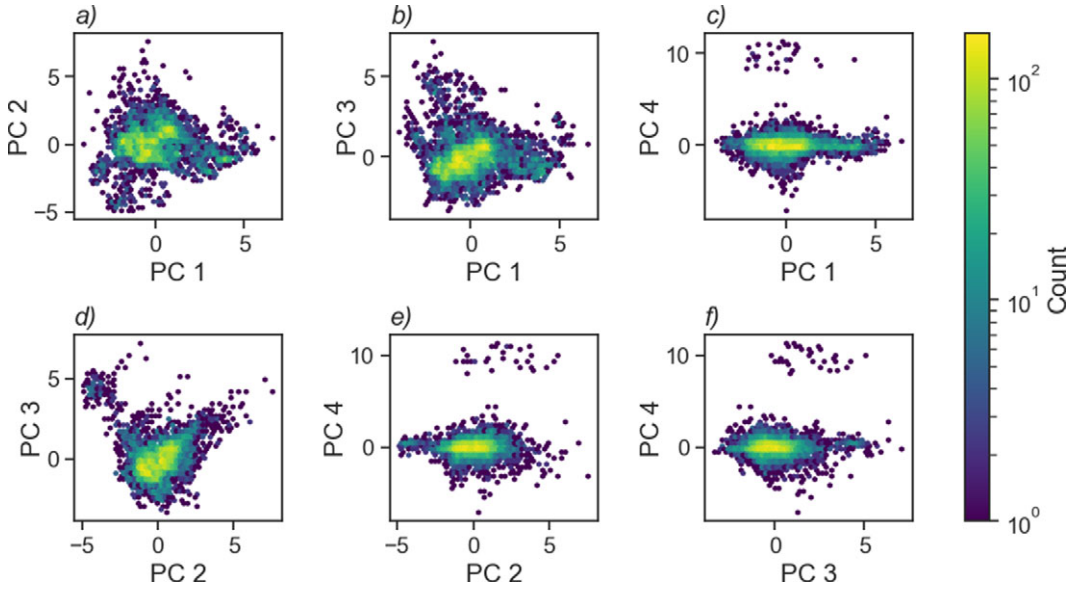

Figure 8. *Test data plotted in principal component (PC) space a) PC2 vs PC1, b) PC3 vs PC1, c) PC4 vs PC1, d) PC3 vs PC2, e) PC4 vs PC2, f) PC4 vs PC3.*

(Figure 8c,e,f). There is another cluster of points associated with low values of PC2 and high values of PC3 (Figure 8d), which will be explored in this section.

Ultimately, we find that two apparent regimes in Figure 8 can be attributed to "low pollution" (Figure 8d) and "in-cloud" (Figure 8c) conditions. We discuss the characteristics of these regimes below and illustrate the regimes through coloring points with the native measurements of metrics of interest.

The plot of PC3 vs PC2 (Figure 8d) is shown in Figure 9 with points colored by two gas-phase tracers of pollution, CO and O_3. The cluster of points present with low PC2 and high PC3 corresponds closely to low observed CO and O_3 concentrations, consistent with this regime being associated with lower pollution levels. There is some correlation between the regime highlighted and higher air temperatures, although this relationship is not as pronounced (Figure 9c).

In addition to the low pollution regime evident in Figure 8, PCA shows distinct in-cloud and out-of-cloud regimes. This in-cloud and out-of-cloud distinction is most apparent in PC4. To illustrate this, we color-test data points by metrics plotted in the PC4 vs PC1 latent space in Figure 10. In Figure 10a, we see

https://doi.org/10.1017/eds.2025.17 Published online by Cambridge University Press

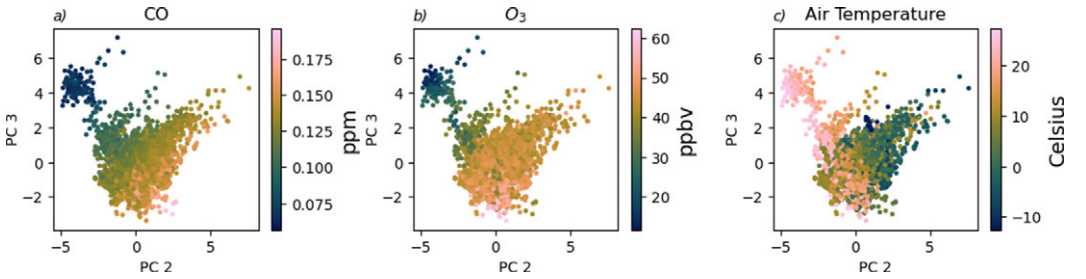

Figure 9. *Principal component analysis (PCA) for low pollution conditions. PC3 vs PC2 is colored by a) carbon monoxide concentration (CO), b) ozone concentration (O_3), and c) air temperature.*

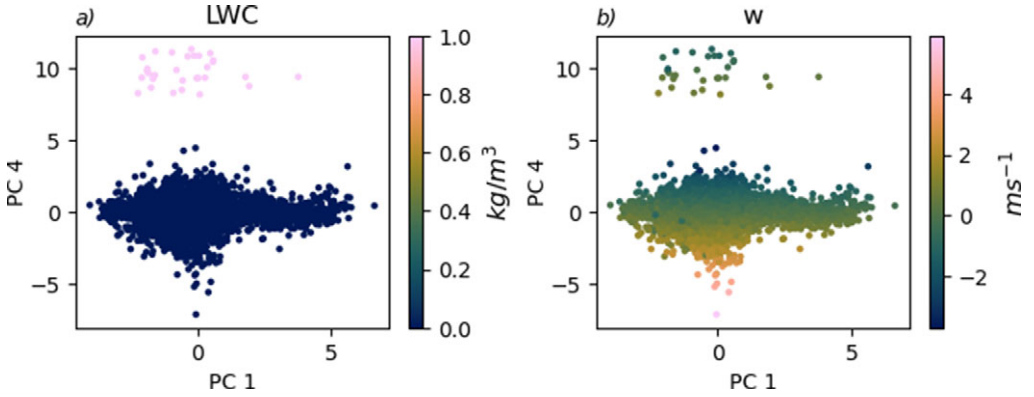

Figure 10. *Principal component analysis (PCA) in-cloud regime. PC4 versus PC1 is colored by a) liquid water content (LWC) and b) vertical velocity (w).*

that there is a clear separation of LWC binary values, which correspond with in-cloud (above or equal to 0.2 x 10^{-5} kg m^{-3}) and out-of-cloud (below 0.2 x 10^{-5} kg m^{-3}). Vertical velocity decreases with PC4, with the decrease separately occurring within the two regimes delineated by LWC.

While the PCA is able to successfully highlight low pollution and in-cloud regimes, it is important to recall that PCA performed poorly on reconstruction, and it is, therefore, useful to also evaluate a method that is better able to capture the nuances of this dataset.

4.2. Autoencoder

Next, we investigate test data in the latent space of the autoencoder. To do so, we treat each of the four encoding nodes of the hidden layer as a new axis on which to plot the test data, analogous to the projection onto the principal components in Section 4.1. The unseen test data plotted in log space on these new axes is illustrated in density plots in Figure 11. Here, we show all combinations of node relationships and identify regimes residing in this reduced dimensional space. Distinct structures are not identified in the first two nor fourth or sixth panels (Figure 11a,b or Figure 11d,f). We identify distinct regimes set apart in the third and fifth panels (Figure 11c,e). In Figure 11c, we see a distinct regime in high values of Node 4. In Figure 11e, there is a cluster of points set apart from the bulk of datapoints concentrated at high values of both Node 2 and Node 4. Overall, test data plotted in this latent space vary more continuously than the more obvious clustering evident in PCA latent spaces, but the latent spaces of the autoencoder represent specific states.

Similar to PCA, there is also a low pollution condition relationship in the autoencoder latent space with the log relationship of Node 1 and Node 4 as shown in Figure 12. At values of approximately 1 for the log

https://doi.org/10.1017/eds.2025.17 Published online by Cambridge University Press

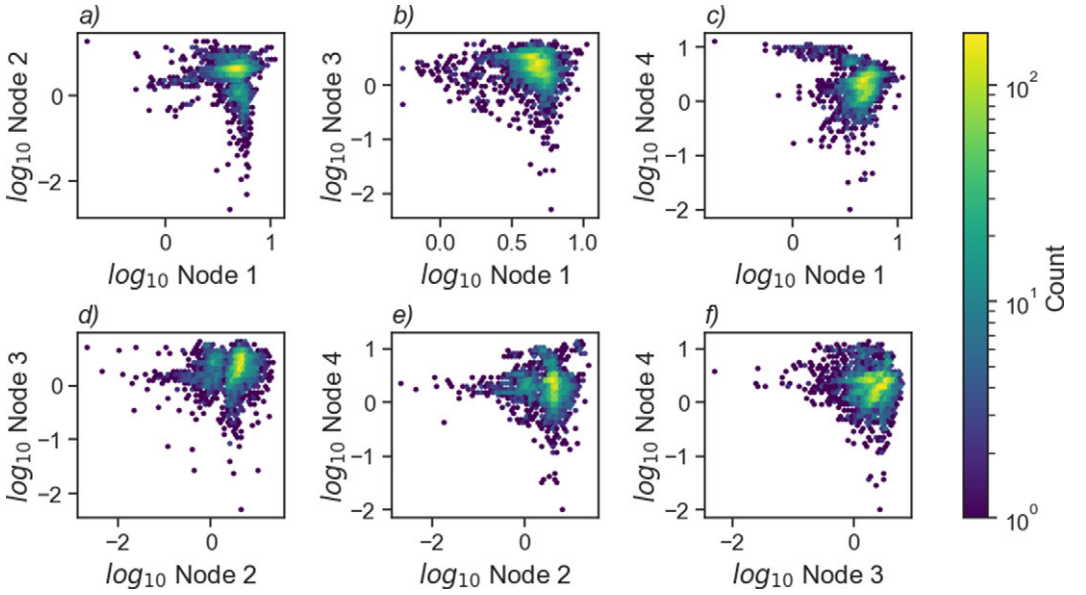

Figure 11. *Test data plotted in autoencoder node relationship plots a) Node 2 vs Node 1, b) Node 3 vs Node 1, c) Node 4 vs Node 1, d) Node 3 vs Node 2, e) Node 4 vs Node 2, f) Node 4 vs Node 3.*

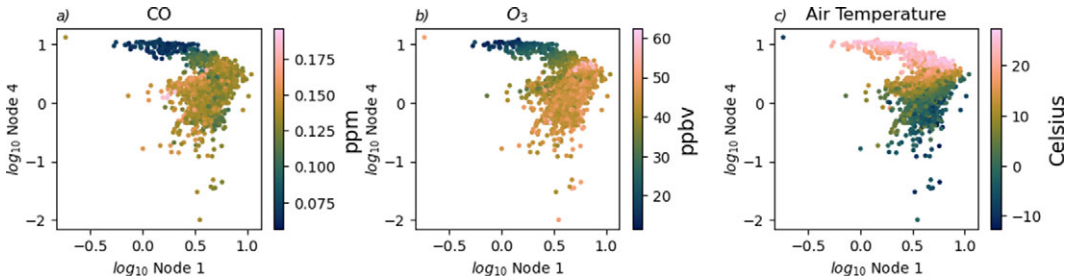

Figure 12. *Autoencoder low pollution conditions. Node 4 versus Node 1 is colored by a) carbon monoxide concentration (CO), b) ozone concentration (O_3), and c) air temperature.*

of Node 4 and 0 to 0.5 for the log of Node 1, we see the lowest values of CO and O_3 (Figure 12a and 12b). Whereas Figure 9c shows a strong relationship between PCs with CO and O_3 but a weak relationship between those same PCs and air temperature, we see that this low pollution regime is tightly delineated with a gradient in high temperatures as well as low pollution in this autoencoder regime (Figure 12c). This provides more insight into the meteorological conditions that go hand in hand with low levels of gas phase pollution in the ACTIVATE dataset.

As shown in Figure 13, the autoencoder highlights an in vs out of cloud regime with the relationship between Node 2 and Node 4, similar to PCA (Figure 10). Values of LWC indicating in-cloud measurements are set apart with high values of Node 2 and Node 4.

However, the autoencoder goes above a simple demarcation of LWC threshold the PCA showed (Figure 10). By focusing on the in-cloud regime located at high values of Node 4 and Node 2 (Figure 13), we find a correlation between this subset of datapoints arranged in this latent space and several variables shown in Figure 14. With PCA, we saw a linear relationship between vertical velocity and PC 4 (Figure 10). Here we see a correlation between Node 2 and Node 4 for both vertical velocity and R_{eff}

https://doi.org/10.1017/eds.2025.17 Published online by Cambridge University Press

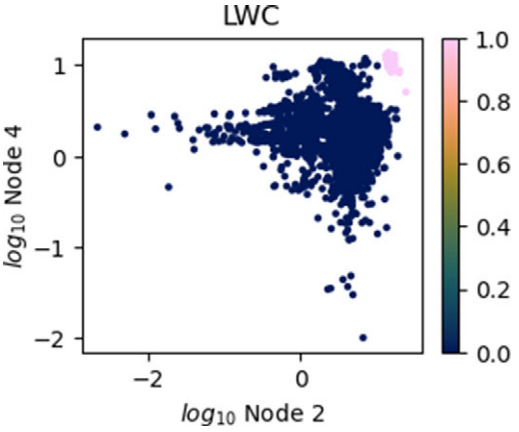

Figure 13. *Autoencoder in-cloud regime highlighted by Node 4 versus Node 2 colored by liquid water content (LWC).*

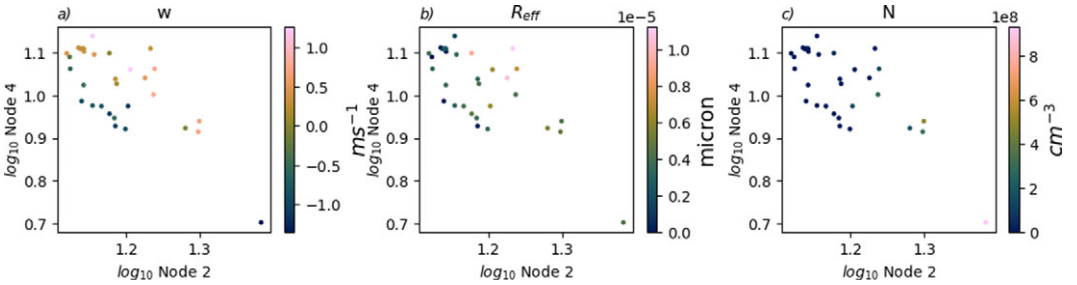

Figure 14. *Autoencoder in-cloud regime magnified. Node 4 versus Node 2 is colored by a) vertical velocity (w), b) effective radius (R_{eff}), and c) number concentration (N).*

(Figure 14a,b). There is also a potential correlation between Node 2 and Node 4 and number concentration, although this relationship is largely driven by a few instances of in-cloud datapoints (Figure 14c).

We also investigated the out-of-cloud regime to see if there are any relationships found in the latent space of log Node 4 versus log Node 2 in Figure 15. Here, we see that there is a positive correlation between Node 2 and number concentration but there is an additional strong relationship between effective radius and Node 2 (Figure 15c,b). The linear relationships between the metrics effective radius and number concentration with Node 2 are consistent in both in-cloud and out-of-cloud regimes. However, this is not the case with vertical velocity where the linear relationship shown in the in-cloud regime is not

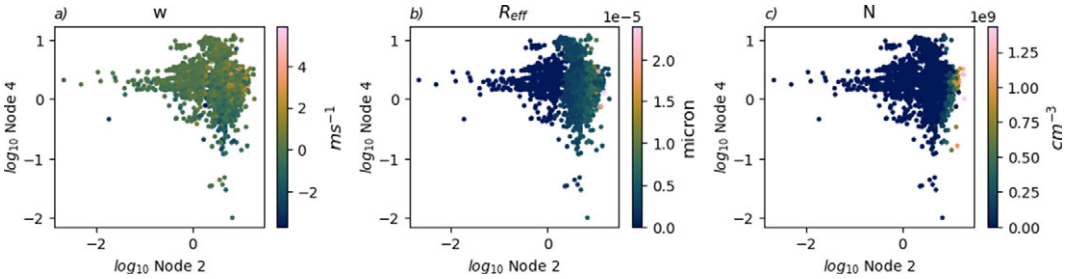

Figure 15. *Autoencoder out of cloud regime. Node 4 versus Node 2 is colored by a) vertical velocity (w), b) effective radius (R_{eff}), and c) number concentration (N).*

https://doi.org/10.1017/eds.2025.17 Published online by Cambridge University Press

nearly as strong in the out-of-cloud regime (Figure 14a and 15a). There also appears to be a potential relationship between Node 2 and the presence of particles as denoted by the measurement of effective radius and number concentration.

Ultimately, this relationship between nodes in- and out of cloud formations is consistent with our understanding that vertical velocity and aerosol size are important in aerosol-cloud processes (e.g., Twomey, 1959; Fountoukis et al., 2007).

There is utility in both the PCA and autoencoder approaches. The distinct regimes of cloudy and clean are consistent across both linear and non-linear methods. Both the encoded dimensions of PCA and the latent space of the autoencoder highlight low pollution conditions and in-cloud regimes. However, the autoencoder performs better than PCA by two orders of magnitude and represents the non-linearity of the data in the latent space. This is consistent even in applications of these tools in other applications (e.g., Hinton and Salakhutdinov, 2006; Kiran et al., 2018; Fournier and Aloise, 2019).

5. Conclusion

In our work, we investigate reduced dimensional variability in in-situ data collected from the NASA ACTIVATE campaign using two unsupervised machine learning methods—principal component analysis (PCA) and an autoencoder. We train, optimize, and apply an autoencoder to successfully reconstruct a dataset with zero-inflated parameters and real instrumental data missingness. Additionally, we analyze the compressed dimensions from the linear reduced space created by PCA and non-linear reduced space of an autoencoder. While both PCA and the autoencoder produce qualitatively similar results, the autoencoder is orders of magnitude more accurate and provides a low dimensional representation of the data that is slightly more consistent with domain knowledge. We highlight the utility of these dimensionality reduction approaches as a way to assess the variability in a dataset of aerosol–cloud parameters and identify regimes in observed aerosol-cloud parameter space. These techniques can be useful to those seeking to understand which relationships drive the variability in a dataset, especially where the number of input parameters is large.

This work finds key regimes through a purely data-driven approach that can be applied to theoretical predictions (e.g., box models and climate models). Future work could compare these results of observational dimensionality reduction to results of the same parameters based on climate model output, for example. In this way, we would be able to investigate discrepancies in how the two datasets behave in reduced dimensional space, which could provide complementary information to the univariate comparisons commonly used in model assessment. If we better understand the connections and relationships between aerosol–cloud parameters, we could better resolve the effects in climate models through parameterizations. The framework presented here could be extrapolated into any field in which large amounts of observational or simulation data are present as a first pass to understanding drivers of variability and key regimes that define data in reduced dimensional space.

Open peer review. To view the open peer review materials for this article, please visit http://doi.org/10.1017/eds.2025.17.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/eds.2025.17.

Acknowledgments. ACTIVATE is a NASA Earth Venture Suborbital-3 (EVS-3) investigation funded by NASA's Earth Science Division and managed through the Earth System Science Pathfinder Program Office. The authors acknowledge and thank Claire Robinson for her contributions to this study.

Author contribution. Conceptualization: S.J.S.; K.B. Methodology: S.J.S.; K.B. Data curation: A.S.; R.H.M.; G.S.D.; J.N.; L.Z.; E.C.J.D.; E.W.; M.S.; C.R.; Y.C. Data visualisation: K.B.; S.J.S. Writing original draft: K.B.; S.J.S. All authors approved the final submitted draft.

Competing interests. The authors declare none.

Data availability statement. Replication data can be found at https://asdc.larc.nasa.gov/project/ACTIVATE and code can be found on Zenodo: https://doi.org/10.5281/zenodo.11055899 (Butler, 2024).

Ethics statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Funding statement. This work was supported by a Wrigley Faculty Innovation Award and NASA Grant 80NSSC23K0322.

References

Achakulwisut P, **Shen L and Mickley LJ** (2017). What controls springtime fine dust variability in the Western United States? Investigating the 2002–2015 increase in fine dust in the U.S. Southwest. *Journal of Geophysical Research: Atmospheres*, *122*(22), 12449–12467. https://doi.org/10.1002/2017JD027208

Albrecht BA (1989) Aerosols, cloud microphysics, and fractional cloudiness. *Science 245*(4923), 1227–1231.

Behrens G, **Beucler T**, **Gentine P**, **Iglesias-Suarez F**, **Pritchard M and Eyring V** (2022) Non-linear dimensionality reduction with a variational encoder decoder to understand convective processes in climate models. *Journal of Advances in Modeling Earth Systems 14*(8), e2022MS003130. https://doi.org/10.1029/2022MS003130.

Bellouin N, **Quaas J**, **Gryspeerdt E**, **Kinne S**, **Stier P**, **Watson-Parris D**, **Boucher O**, **Carslaw KS**, **Christensen M**, **Daniau A-L**, **Dufresne J-L**, **Feingold G**, **Fiedler S**, **Forster P**, **Gettelman A**, **Haywood JM**, **Lohmann U**, **Malavelle F**, **Mauritsen T and Stevens B** (2020) Bounding global aerosol radiative forcing of climate change. *Reviews of Geophysics 58*(1), e2019RG000660. https://doi.org/10.1029/2019RG000660.

Benestad RE, **Mezghani A**, **Lutz J**, **Dobler A**, **Parding KM and Landgren OA** (2023) Various ways of using empirical orthogonal functions for climate model evaluation. *Geoscientific Model Development 16*(10), 2899–2913. https://doi.org/10.5194/gmd-16-2899-2023.

Butler K (2024) Investigating reduced-dimensional variability in aircraft-observed aerosol-cloud *parameters*. https://doi.org/10.5281/zenodo.11055899.

Chollet F (2021) *Deep Learning with Python*, 2nd Edition. Simon and Schuster.

Dadashazar H, **Painemal D**, **Alipanah M**, **Brunke M**, **Chellappan S**, **Corral AF**, **Crosbie E**, **Kirschler S**, **Liu H**, **Moore RH**, **Robinson C**, **Scarino AJ**, **Shook M**, **Sinclair K**, **Thornhill KL**, **Voigt C**, **Wang H**, **Winstead E**, **Zeng X and Sorooshian A** (2021) Cloud drop number concentrations over the western North Atlantic Ocean: seasonal cycle, aerosol interrelationships, and other influential factors. *Atmospheric Chemistry and Physics 21*(13), 10499–10526. https://doi.org/10.5194/acp-21-10499-2021.

DiGangi JP, **Choi Y**, **Nowak JB**, **Halliday HS**, **Diskin GS**, **Feng S**, **Barkley ZR**, **Lauvaux T**, **Pal S**, **Davis KJ**, **Baier BC and Sweeney C** (2021) Seasonal variability in local carbon dioxide biomass burning sources over central and eastern US using airborne in situ enhancement ratios. *Journal of Geophysical Research: Atmospheres 126*(24), e2020JD034525. https://doi.org/10.1029/2020JD034525.

Diskin GS, **Podolske JR**, **Sachse GW and Slate TA** (2002) Open-path airborne tunable diode laser hygrometer. *Diode Lasers and Applications in Atmospheric Sensing 4817*, 196–204. https://doi.org/10.1117/12.453736.

Fiore AM, **Milly GP**, **Hancock SE**, **Quiñones L**, **Bowden JH**, **Helstrom E**, **Lamarque J-F**, **Schnell J**, **West JJ and Xu Y** (2022) Characterizing changes in eastern U.S. pollution events in a warming world. *Journal of Geophysical Research: Atmospheres 127*(9), e2021JD035985. https://doi.org/10.1029/2021JD035985.

Fountoukis C, **Nenes A**, **Meskhidze N**, **Bahreini R**, **Conant WC**, **Jonsson H**, **Murphy S**, **Sorooshian A**, **Varutbangkul V**, **Brechtel F**, **Flagan RC and Seinfeld JH** (2007) Aerosol–cloud drop concentration closure for clouds sampled during the international consortium for atmospheric research on transport and transformation 2004 campaign. *Journal of Geophysical Research: Atmospheres 112*(D10). https://doi.org/10.1029/2006JD007272.

Fournier Q and Aloise D (2019) Empirical comparison between autoencoders and traditional dimensionality reduction methods. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. pp. 211–214. https://doi.org/10.1109/AIKE.2019.00044

Geiss A, **Silva SJ and Hardin JC** (2022) Downscaling atmospheric chemistry simulations with physically consistent deep learning. *Geoscientific Model Development 15*(17), 6677–6694. https://doi.org/10.5194/gmd-15-6677-2022.

Hinton GE and Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. *Science 313*(5786), 504–507. https://doi.org/10.1126/science.1127647.

Hinton GE, **Srivastava N and Swersky K** (2012) *rmsprop: divide the Gradient by a Running Average of Its Recent Magnitude.* Available at https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.

IPCC (2021) *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Jones TA and Christopher SA (2010) Statistical properties of aerosol-cloud-precipitation interactions in South America. *Atmospheric Chemistry and Physics 10*(5), 2287–2305. https://doi.org/10.5194/acp-10-2287-2010.

Kiran BR, **Thomas DM and Parakkal R** (2018) An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging 4*(2), 2. https://doi.org/10.3390/jimaging4020036.

Kirschler S, **Voigt C**, **Anderson B**, **Campos Braga R**, **Chen G**, **Corral AF**, **Crosbie E**, **Dadashazar H**, **Ferrare RA**, **Hahn V**, **Hendricks J**, **Kaufmann S**, **Moore R**, **Pöhlker ML**, **Robinson C**, **Scarino AJ**, **Schollmayer D**, **Shook MA**, **Thornhill KL and Sorooshian A** (2022) Seasonal updraft speeds change cloud droplet number concentrations in low-level clouds over the western North Atlantic. *Atmospheric Chemistry and Physics 22*(12), 8299–8319. https://doi.org/10.5194/acp-22-8299-2022.

Kirschler S, **Voigt C**, **Anderson BE**, **Chen G**, **Crosbie EC**, **Ferrare RA**, **Hahn V**, **Hair JW**, **Kaufmann S**, **Moore RH**, **Painemal D**, **Robinson CE**, **Sanchez KJ**, **Scarino AJ**, **Shingler TJ**, **Shook MA**, **Thornhill KL**, **Winstead EL**, **Ziemba LD and Sorooshian A** (2023) Overview and statistical analysis of boundary layer clouds and precipitation over the western North Atlantic Ocean. *Atmospheric Chemistry and Physics 23*(18), 10731–10750. https://doi.org/10.5194/acp-23-10731-2023.

https://doi.org/10.1017/eds.2025.17 Published online by Cambridge University Press

Kurihana T, **Moyer EJ and Foster IT** (2022) AICCA: AI-driven cloud classification atlas. *Remote Sensing*, *14*(22), 5690. https://doi.org/10.3390/rs14225690

Lerch S and Polsterer KL (2022) *Convolutional Autoencoders for Spatially-Informed Ensemble Post-Processing* (arXiv: 2204.05102). arXiv. Available at http://arxiv.org/abs/2204.05102.

Li L, **Jamieson K**, **DeSalvo G**, **Rostamizadeh A and Talwalkar A** (2018) Hyperband: a novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research 18*, 1–52.

Marais WJ, **Holz RE**, **Reid JS and Willett RM** (2020) Leveraging spatial textures, through machine learning, to identify aerosols and distinct cloud types from multispectral observations. *Atmospheric Measurement Techniques 13*(10), 5459–5480. https://doi.org/10.5194/amt-13-5459-2020.

McComiskey A, **Feingold G**, **Frisch AS**, **Turner DD**, **Miller MA**, **Chiu JC**, **Min Q and Ogren JA** (2009) An assessment of aerosol-cloud interactions in marine stratus clouds based on surface remote sensing. *Journal of Geophysical Research: Atmospheres 114*(D9). https://doi.org/10.1029/2008JD011006.

Minka T (2000) Automatic choice of dimensionality for PCA. *Advances in Neural Information Processing Systems 13*. Available at https://proceedings.neurips.cc/paper_files/paper/2000/hash/7503cfacd12053d309b6bed5c89de212-Abstract.html.

Mohn H, **Kreyling D**, **Wohltmann I**, **Lehmann R**, **Maass P and Rex M** (2023) Neural representation of the stratospheric ozone chemistry. *Environmental Data Science 2*, e41. https://doi.org/10.1017/eds.2023.35.

Mooers G, **Beucler T**, **Pritchard M and Mandt S** (2024) Understanding precipitation changes through unsupervised machine learning. *Environmental Data Science*, *3*. https://doi.org/10.1017/eds.2024.1.

Mooers G, **Pritchard M**, **Beucler T**, **Srivastava P**, **Mangipudi H**, **Peng L**, **Gentine P and Mandt S** (2023b) Comparing storm resolving models and climates via unsupervised machine learning. *Scientific Reports 13*(1), Article 1. https://doi.org/10.1038/s41598-023-49455-w

Moore RH and Nenes A (2009) Scanning flow CCN analysis—A method for fast measurements of CCN spectra. *Aerosol Science and Technology 43*(12), 1192–1207. https://doi.org/10.1080/02786820903289780.

National Academies of Sciences, Engineering and Medicine (2018) *Thriving on Our Changing Planet: A Decadal Strategy for Earth Observation From Space.* Available at https://www.nap.edu/catalog/24938/thriving-on-our-changing-planet-adecadal-strategy-for-earth.

Pedregosa F, **Varoquaux G**, **Gramfort A**, **Michel V**, **Thirion B**, **Grisel O**, **Blondel M**, **Prettenhofer P**, **Weiss R**, **Dubourg V**, **Vanderplas J**, **Passos A and Cournapeau D** (2011) Scikit-learn: machine learning in python. *Journal of Machine Learning Research 12*, 2825–2830.

Rosenfeld D, **Lohmann U**, **Raga GB**, **O'Dowd CD**, **Kulmala M**, **Fuzzi S**, **Reissell A and Andreae MO** (2008) Flood or drought: how do aerosols affect precipitation? *Science 321*(5894), 1309–1313. https://doi.org/10.1126/science.1160606.

Seinfeld JH, **Bretherton C**, **Carslaw KS**, **Coe H**, **DeMott PJ**, **Dunlea EJ**, **Feingold G**, **Ghan S**, **Guenther AB**, **Kahn R**, **Kraucunas I**, **Kreidenweis SM**, **Molina MJ**, **Nenes A**, **Penner JE**, **Prather KA**, **Ramanathan V**, **Ramaswamy V**, **Rasch PJ and Wood R** (2016) Improving our fundamental understanding of the role of aerosol−cloud interactions in the climate system. *Proceedings of the National Academy of Sciences 113*(21), 5781–5790. https://doi.org/10.1073/pnas.1514043113.

Silva SJ, **Keller CA and Hardin J** (2022) Using an explainable machine learning approach to characterize earth system model errors: application of SHAP analysis to Modeling lightning flash occurrence. *Journal of Advances in Modeling Earth Systems 14*(4), e2021MS002881. https://doi.org/10.1029/2021MS002881.

Sorooshian A, **Alexandrov MD**, **Bell AD**, **Bennett R**, **Betito G**, **Burton SP**, **Buzanowicz ME**, **Cairns B**, **Chemyakin EV**, **Chen G**, **Choi Y**, **Collister BL**, **Cook AL**, **Corral AF**, **Crosbie EC**, **van Diedenhoven B**, **DiGangi JP**, **Diskin GS**, **Dmitrovic S and Zuidema P** (2023) Spatially coordinated airborne data and complementary products for aerosol, gas, cloud, and meteorological studies: the NASA ACTIVATE dataset. *Earth System Science Data 15*(8), 3419–3472. https://doi.org/10.5194/essd-15-3419-2023.

Sorooshian A, **Anderson B**, **Bauer SE**, **Braun RA**, **Cairns B**, **Crosbie E**, **Dadashazar H**, **Diskin G**, **Ferrare R**, **Flagan RC**, **Hair J**, **Hostetler C**, **Jonsson HH**, **Kleb MM**, **Liu H**, **MacDonald AB**, **McComiskey A**, **Moore R**, **Painemal D and Zuidema P** (2019) Aerosol–cloud–meteorology interaction airborne field investigations: using lessons learned from the U.S. West Coast in the design of ACTIVATE off the U.S. East Coast. *Bulletin of the American Meteorological Society 100*(8), 1511–1528. https://doi.org/10.1175/BAMS-D-18-0100.1.

Thornhill KL, **Anderson BE**, **Barrick JDW**, **Bagwell DR**, **Friesen R and Lenschow DH** (2003) Air motion intercomparison flights during transport and chemical evolution in the Pacific (TRACE-P)/ACE-ASIA. *Journal of Geophysical Research: Atmospheres 108*(D20). https://doi.org/10.1029/2002JD003108.

Twomey S (1959) The nuclei of natural cloud formation part II: the supersaturation in natural clouds and the variation of cloud droplet concentration. *Geofisica Pura e Applicata 43*(1), 243–249. https://doi.org/10.1007/BF01993560.

Twomey, S. (1974). Pollution and the planetary albedo. *Atmospheric Environment (1967)*, *8*(12), 1251–1256. https://doi.org/10.1016/0004-6981(74)90004-3

Twomey S (1977) The influence of pollution on the shortwave albedo of clouds. *Journal of the Atmospheric Sciences 34*(7), 1149–1152. https://doi.org/10.1175/1520-0469(1977)034<1149:TIOPOT>2.0.CO;2.

https://doi.org/10.1017/eds.2025.17 Published online by Cambridge University Press

Wei Y, **Shrestha R**, **Pal S**, **Gerken T**, **Feng S**, **McNelis J**, **Singh D**, **Thornton MM**, **Boyer AG**, **Shook MA**, **Chen G**, **Baier BC**, **Barkley ZR**, **Barrick JD**, **Bennett JR**, **Browell EV**, **Campbell JF**, **Campbell LJ**, **Choi Y and Davis KJ** (2021) Atmospheric carbon and transport—America (ACT-America) data sets: description, management, and delivery. *Earth and Space Science 8*(7), e2020EA001634. https://doi.org/10.1029/2020EA001634.

Yorks JE, **Selmer PA**, **Kupchock A**, **Nowottnick EP**, **Christian KE**, **Rusinek D**, **Dacic N and McGill MJ** (2021) Aerosol and Cloud Detection Using Machine Learning Algorithms and Space-Based Lidar Data. *Atmosphere*, *12*(5), 606. https://doi.org/10.3390/atmos12050606

Cite this article: Butler KM, Silva SJ, Sorooshian A, Moore RH, Diskin GS, Nowak JB, Ziemba L, Crosbie E, Shook MA, DiGangi J, Winstead E, Robinson C and Choi Y (2025). Investigating reduced-dimensional variability in aircraft-observed aerosol–cloud parameters. *Environmental Data Science*, 4: e27. doi:10.1017/eds.2025.17