U.S. DEPARTMENT OF ENERGY | Office of Science

DOE/SC-ARM-TR-274

Micropulse Lidar Cloud Mask Machine-Learning Value-Added Product Report

D Flynn
D Zhang
E Cromwell

March 2023

ARM

DOE/SC-ARM-TR-274

Micropulse Lidar Cloud Mask Machine-Learning Value-Added Product Report

D Flynn
E Cromwell
D Zhang
All at Pacific Northwest National Laboratory

March 2023

How to cite this document:

Flynn, D, E Cromwell, and D Zhang. 2023. Micropulse Lidar Cloud Mask Machine-Learning Value-Added Product Report. U.S. Department of Energy, Atmospheric Radiation Measurement user facility, Richland, Washington. DOE/SC-ARM-TR-274.

Work supported by the U.S. Department of Energy,
Office of Science, Office of Biological and Environmental Research

Acronyms and Abbreviations

AGL	above ground level
ARM	Atmospheric Radiation Measurement
AWR	ARM West Antarctica Radiation Experiment
CACTI	Cloud, Aerosol, and Complex Terrain Interactions
COR	Córdoba, Argentina (CACTI campaign)
LDR	linear depolarization ratio
MPL	micropulse lidar
MPLCMASKML	Micropulse Lidar Cloud Mask Machine-Learning
NaN	not a number
NRB	normalized relative backscatter
OLI	Oliktok Point
SGP	Southern Great Plains
UTC	Coordinated Universal Time
VAP	value-added product

Contents

Figures

Tables

1.0 Introduction

Cloud detection algorithms of various techniques have been developed and applied to atmospheric ground-based lidar data to identify cloud boundaries and produce clouds masks. While these algorithms are able to identify a wide variety of cloud types and conditions, it is often observed that the algorithms can still fail to accurately detect clouds that are readily discernible when inspecting the lidar imagery. Based on this observation, an alternative approach for cloud detection is to take advantage of machine-learning capabilities and the trained human eye as an interpreter of lidar images, and in turn, to train a neural network to recognize the desired features in the lidar data.

This approach has been applied to data from the Atmospheric Radiation Measurement (ARM) user facility's micropulse lidar (MPL) systems to develop a machine-learning model that can segment images semantically and produce pixel-to-pixel predictions of cloud in lidar images (Cromwell et al. 2019). In addition to these cloud predictions, the model can provide a confidence rating of each pixel prediction. The cloud prediction output from the machine-learning model developed is available in the value-added product (VAP) Micropulse Lidar Cloud Mask Machine-Learning (MPLCMASKML). This VAP provides the cloud mask generated from the prediction output as well as the number of cloud layers and the cloud-layer boundaries.

The cloud detection model developed for the MPLCMASKML VAP was trained exclusively with data from the MPL located at the ARM Southern Great Plains (SGP) observatory between January and March of 2015 plus isolated days with heavy aerosol loading. Although trained with site-, date-, and instrument-specific data, the performance of the model has been quite good when applied to other ARM MPL data sets. The transfer-learning abilities of the model from one data set to the next are discussed further in Summary and Future Works.

2.0 Overview of Model

The machine-learning model in the MPLCMASKML has been trained to identify cloud pixels from MPL lidar images. The images provided to the model for both input and training are two-channel, with the MPL normalized relative backscatter (NRB) as one channel and the corresponding linear depolarization ratio (LDR) calculation as the other. This permits the simultaneous assessment of the lidar backscatter and the lidar attenuation behavior from the NRB with the cloud phase, clear sky, or aerosol indication from the LDR. Additionally, because a cloud mask is the primary output of the machine-learning model, a cloud mask corresponding to the two-channel lidar images is also required for the training and testing process. The cloud mask product from the ARM Micropulse Lidar Cloud Mask (MPLCMASK) VAP (Flynn 2020) is used for pre-training steps while "hand-labeled" cloud masks are used during the final training. The hand-labeled cloud masks are created from visual inspection of the two-channel images. Further details on the input and training images and cloud masks are provided in sections 3.0 and 5.0 of this report.

The method used to train the MPLCMASKML cloud detection model was a three-stage process that included (1) cloud or no-cloud classification, (2) pre-training for cloud location, and (3) final training (fine tuning) for cloud location. This machine-learning approach can be described as a semi-supervised learning method based on the labeling of the training data used in the multiple stages. In the first stage, 890 images with cloud and 890 images without cloud were used to train a classification model to determine whether an image contains any cloud or not. In the second stage, the model was trained ("pre-trained") with cloud mask from the MPLCMASK VAP. These VAP cloud masks were not the desired output but were close enough to be useful to initialize the model. In this case, 4200 images were used for this step. For the final training, over 100 days of cloud masks were created for fine tuning as well as assessing the performance of the end model. These "hand-labeled" cloud masks were flipped to create a mirror image, effectively doubling the amount of data for the fine-tuning stage. Hold-out data were randomly selected from the hand-labeled cloud masks and were not used for training but were reserved for assessing the model performance. This assessment included calculating the model precision (percentage of predicted clouds that are actual clouds) and recall (percentage of clouds that are predicted as clouds). From these statistics a third model performance metric, the F1-score (harmonic mean of the precision and recall), was also calculated.

3.0 Model Pre-Training and Training Data

3.1 Pre-Training

The pre-training data were from ARM fast-switching polarized SGP MPL measurements that have been processed by the MPLCMASK VAP. The VAP output used included NRB, LDR, and cloud mask with 30 seconds averaged for time resolution, and 30 meters averaged for vertical height resolution. To allow more efficient training of the model, each 24-hour day of data was divided into quarters. Each quarter had a small amount of overlap with adjacent quarters in the same day and was therefore just over six hours long.

3.1.1 Classification Pre-Training Data

For the pre-training classification stage of the model, a set of 1780 quarter days was selected. Each quarter of data has image-level annotations identifying whether the quarter day contains clouds. A quarter day is marked as containing clouds if the MPLCMASK cloud mask product contains clouds. The data set was divided evenly between quarter days with clouds (890) and without clouds (890).

3.1.2 Cloud Location Pre-Training Data

The pre-training cloud location data consists of more than 4200 quarter days. These training data were selected between the years 2010 and 2015 and included NRB and LDR as well as the MPLCMASK-derived cloud masks.

3.2 Final Training for Cloud Location

The final training data consists of over 100 days of hand-labeled cloud masks for MPL measurements from the SGP site. Most of the data were from between January 1 and March 30, 2015, with some supplementary data being added between April and December of the same year. The hand-labeled data were created using Matlab tools developed specifically for this work. During the labeling process the user can view and inspect the NRB and LDR imagery. After visually assessing the data, the user has different options on how to select which pixels in the MPL imagery are cloud. Much of the labeling can employ basic image-processing techniques, but portions of the labeling do require pixel-by-pixel discrimination. The labeled cloud masks are flipped to effectively double the final training data available.

The definition of what is cloud for cloud masks varies in published descriptions and is particularly influenced by what type of study is using the retrieved cloud location data. The cloud mask hand-labeling for the final training data closely followed the definition adopted in Wang et al. (2001). The labeling attempts to exclude liquid precipitation and virga while including all other liquid droplet, ice particles, and mixed-phase hydrometeors that can be considered radiatively important suspended cloud particles.

4.0 VAP Basic Steps and Flow Chart

The following steps briefly describe the logic used in the VAP:

1. Input Pre-Processing of input data from MPLCMASK VAP.
   a) Real log base 10 of NRB (log[NRB]) to improve visualization of the lidar imagery.
   b) Missing values, infinite values, or not-a-number (NaN) values are set to the daily minimum.
   c) For each day, log(NRB) is then zero-centered by subtracting the mean from the log(NRB) and normalized by dividing by the standard deviation of the zero-centered log(NRB).
   d) The LDR valid data range is between 0 and 1. Missing values and NaNs are set to 0, any values greater than 1 are set to 1, and values less than 0 are set to 0.

2. Run trained model on each quarter day to estimate cloud mask for each quarter day.

   The 24-hour day is divided into the following fourths (quarter-days):

   First quarter: time bins 0 to 800 (00:00:00 to 06:40:00 UTC)

   Second quarter: time bins 680 to 1480 (05:40:00 to 12:20:00 UTC)

   Third quarter: time bins 1400 to 2200 (11:40:00 to 18:20:00 UTC)

   Fourth quarter: time bins 2080 to 2880 (17:20:00 to 24:00:00 UTC).

3. Merge quarter-day cloud masks into full-day cloud mask.

   Three overlapping time periods of cloud mask from the quarter days need to be merged:

   05:40:00 to 06:40:00 UTC

   11:40:00 to 12:20:00 UTC

   17:20:00 to 18:20:00 UTC.

   For a given time and height location in the overlap period, if at least one quarter-day cloud mask has the location marked as a cloud, then the location in the merged mask is marked as a cloud.

4. Apply clustering algorithm and merge layers in cloud mask.

   a) Run cluster test from MPLCMASK VAP on all bins identified as cloud above 10 km to detect and attempt to eliminate any small false clouds or clutter (Flynn et al. 2020).

b) Merge cloud bins time-wise if clouds are less than 2 time bins apart (1 minute).

c) Merge cloud bins height-wise if clouds are less than 4 bins apart (120m).

5. Calculate cloud base, cloud top, and number of cloud layers.

5.0 Input Data

The primary inputs required for MPLCMASKML are NRB and LDR calculations from polarized ARM MPL systems. Historically, the ARM user facility operated single-wavelength non-polarized MPLs at all observatories and mobile facilities. In 2004, polarized MPL systems were installed at all observatories and mobile facilities. In addition, polarized MPL systems were upgraded starting in June 2010 to fast-switching systems to allow improved switching capability between the linear and circular polarization channels. While alternative versions of the VAP can be developed to handle older MPL data sets, the current version of MPLCMASKML is only processing fast-switching polarized MPL data. The processed measurements are provided by MPLCMASK.

The MPLCMASK VAP produces one output file. The name of the file is:

SSS30smplcmask1zwangXX.c1.YYYYMMDD.hhmmss

Where:

- SSS is the site of the instrument (e.g., sgp)
- XX is the facility (i.e., C1, C2, etc.)
- YYYY is the year
- MM is the month of the year
- DD is the day of the month
- hh is the hour of the day
- mm is the minute of the hour
- ss is the second of the minute of data start.

6.0 Output Data

The MPLCMASKML VAP produces one output file. The name of the output file is:

SSSmplcmaskmlXX.c1.YYYYMMDD.hhmmss

Where:

- SSS is the site of the instrument (e.g., sgp)
- XX is the facility (i.e., C1, C2, etc.)
- YYYY is the year

- MM is the month of the year
- DD is the day of the month
- hh is the hour of the day
- mm is the minute of the hour
- ss is the second of the minute of data start.

Table 1. Primary variable in output data set.

Variable (dimensions)	Long Name
backscatter (time, height)	Total attenuated backscatter
range_corrected_backscatter (time, height)	Range-corrected total attenuated backscatter
preprocess_backscatter (time, height)	Preprocessed range-corrected total attenuated backscatter for model
linear_depol_ratio (time, height)	Linear depolarization ratio
preprocess_linear_depol_ratio (time, height)	Preprocessed linear depolarization ratio for model
cloud_mask_dl (time, height)	Cloud mask from deep learning model
cloud_mask_confidence (time, height)	Model confidence in cloud prediction, where 1 is highly confident point is cloud, 0 is highly confident point is not cloud
cloud_mask (time, height)	Cloud mask from deep-learning model with clustering and merging applied
cloud_base (time)	Lowest cloud base height above ground level (AGL)
cloud_top (time)	Highest cloud top height AGL
num_cloud_layers (time)	Number of cloud layers

7.0 Example Plots

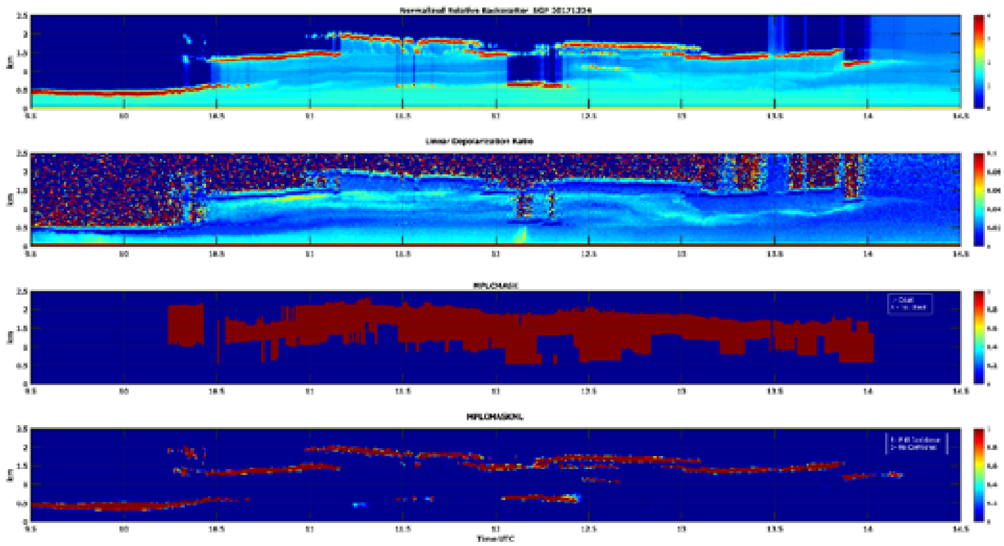

Figure 1. Example plots from the SGP show the log of the normalized relative backscatter (log10 NRB), linear depolarization ratio (LDR), and cloud masks for MPLCMASK and MPLCMASKML for December 24, 2017. The strength of the machine-learning approach is apparent on this day when moderate aerosol loading is likely interfering with accurate cloud boundary identification for the MPLCMASK algorithm in the boundary layer between 10-14 UTC. Additionally, the MPLCMASKML is able to identify the attenuating cloud below 500 m between 10-11 UTC.

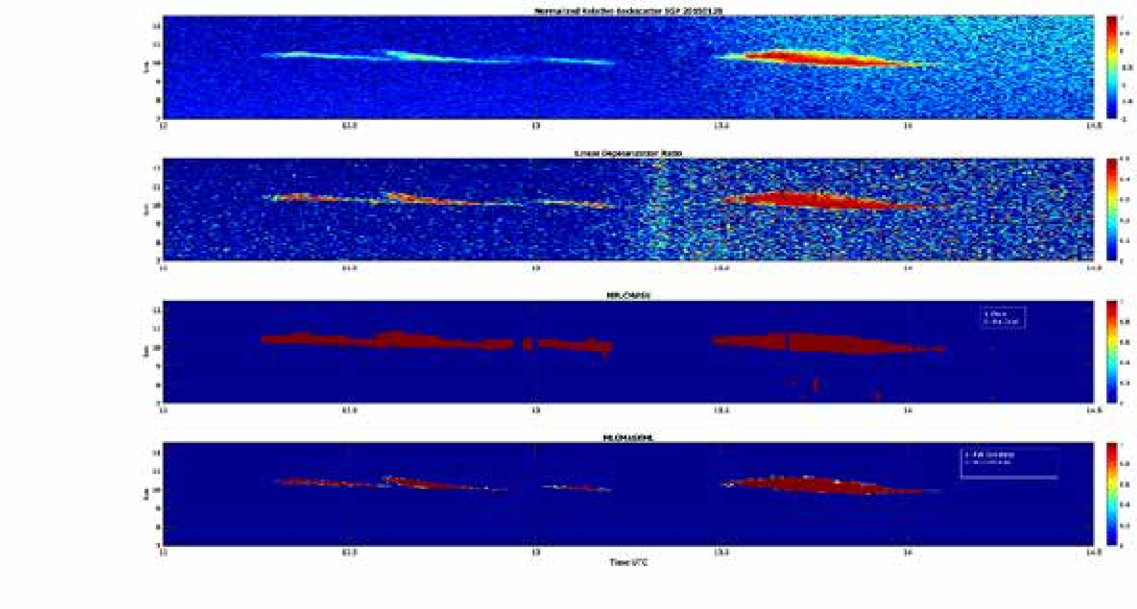

Figure 2. Sample MPLCMASKML plots from the SGP show the log of the normalized relative backscatter (log10 NRB), linear depolarization ratio (LDR), and cloud masks for the MPLCMASKML and MPLCMASK for January 28, 2015. The time series for this example begins close to sunrise with the resulting background (noise) increasing with time. Cirrus clouds are present and centered around 10 km. Because the clouds do not fully attenuate the lidar signal, this is a good case to illustrate the cloud mask's performance when the NRB and the LDR signal-to-noise ratios are degraded. MPLCMASKML identifies most of the cirrus cloud but does not detect some of the thinner cloud edges. The MPLCMASK detects some edges of the cirrus clouds that the machine-learning version misses, but then tends to oversample both the cloud top and bottom boundaries.

8.0 Summary and Assessment

The MPLCMASKML is an alternative approach to traditional cloud-detection algorithms for atmospheric ground-based lidar data. This VAP applies a machine-learning model to data from the ARM fast-switching polarized MPL systems to produce a daily cloud mask with a corresponding confidence rating and identifies the number of cloud layers as well as the cloud-layer boundaries. The model was trained using MPLCMASK cloud masks and hand-labeled ground truth correlated with calculations of NRB and LDR from the MPL.

8.1 Model Assessment

The example plot provided in this technical report (section 7) demonstrates the overall observed advantages when compared to the VAP predecessor as (1) less exaggerated cloud boundaries, (2) improved cloud layer separation, and (3) cloud identification below 500 m.

The model performance can also be evaluated by comparing the model and ground truth cloud masks and calculating the precision, recall, and the F1 score, defined as the harmonic mean of the precision and recall. As mentioned previously, these three model performance metrics are calculated from ground truth hold-out data (66 quarter days) that were randomly selected and not used for the model training. In addition, 27 days (108 quarter-days) from March 2015 at SGP C1 are used to further evaluate the model. The metrics for the MPLCMASKML VAP cloud masks produced from 2015 SGP MPL data are shown below in Table 2. We also include the metrics for the MPLCMASK algorithm on the same data sets in Table 3 for comparison.

Table 2. Performance metrics of the cloud mask detection model (MPLCMASKML) on the ground truth hold-out data set ("Hold-Out") and the March 2015 data set ("March").

	F1-Score	**Precision**	**Recall**
Hold-Out	0.8790	0.8505	0.9094
March	0.8626	0.8432	0.8829

Table 3. Performance metrics of the MPLCMASK cloud mask on the ground truth hold-out data set ("Hold-Out") and the March 2015 data set ("March").

	F1-Score	**Precision**	**Recall**
Hold-Out	0.5892	0.4423	0.8795
March	0.65	0.5072	0.9049

Overall, the MPLCMASKML model outperforms the MPLCMASK algorithm, achieving higher F1-Scores and double the precision.

8.2 Transfer Learning

Transfer learning is a machine-learning technique that includes applying a model that was trained with one data set to different but related data sets. In the case of the MPLCMASKML VAP, the dominant training data for the model is from an ARM MPL that was operating at the SGP site during the winter and early spring of 2015. That specific MPL instrument, together with the location and the time period of this data, characterize the model that is trained with this data.

Two features of the specific MPL can set its lidar data output apart from others: calibration and polarization. The MPL calibration is fortunately not an issue because it is effectively removed or normalized during the pre-processing of the model input data. Changes in an instrument's polarization behavior, however, can be an issue. If the polarization degrades significantly, the LDR values increase, and the model struggles to accurately distinguish clouds from clear air or aerosol.

The cloud types and atmospheric conditions present in the MPL data vary by location and time period. For example, a climatology examination of cloud types at the mid-latitude SGP site (Lim 2019) found cirrus clouds were the dominant cloud type, while in the Arctic, stratiform clouds prevail (Serreze 2005, Przybylak 2003). Seasonal variations are also observed in these two contrasting environments. The same study of clouds at the SGP site found low clouds were more than twice as frequent during February and March than July and August. Shupe (2011) also illustrated a distinction between summer and winter arctic clouds in terms of both height and phase. It was found that the cloud height increased significantly in summer. Likewise, mixed-phases cloud occurrence had a minimum in the winter season but increased in summer.

Because the training data is representative of a specific data set, a model trained on data from a MPL operating at a mid-latitude site may not necessarily perform as well on data from a different MPL operating at an arctic location.

8.3 Application of Transfer Learning

To test the transfer learning ability of the MPLCMASKML model, the VAP was applied to data from the ARM arctic site at Oliktok Point (OLI), North Slope of Alaska. For OLI, the model was evaluated on 14 days (56 quarter-days) of hand-labeled data from May 2016. The performance is shown below in Table 4.

Table 4. Performance metrics of the cloud mask detection model (MPLCMASKML) and MPLCMASK algorithm on the OLI data set.

	F1-Score	**Precision**	**Recall**
MPLCMASKML	0.7530	0.7438	0.7643
MPLCMASK	0.4185	0.371	0.48

Like the SGP data set, the MPLCMASKML achieves better overall performance compared to the MPLCMASK algorithm. One reason for the improvement in performance is that the MPLCMASKML can identify clouds below 500m, which are very common at the OLI site (see Figure 3). These results demonstrate the favorable transfer-learning ability of the MPLCMASKML model and that from a single model it can produce significantly improved cloud masks at different observation sites.

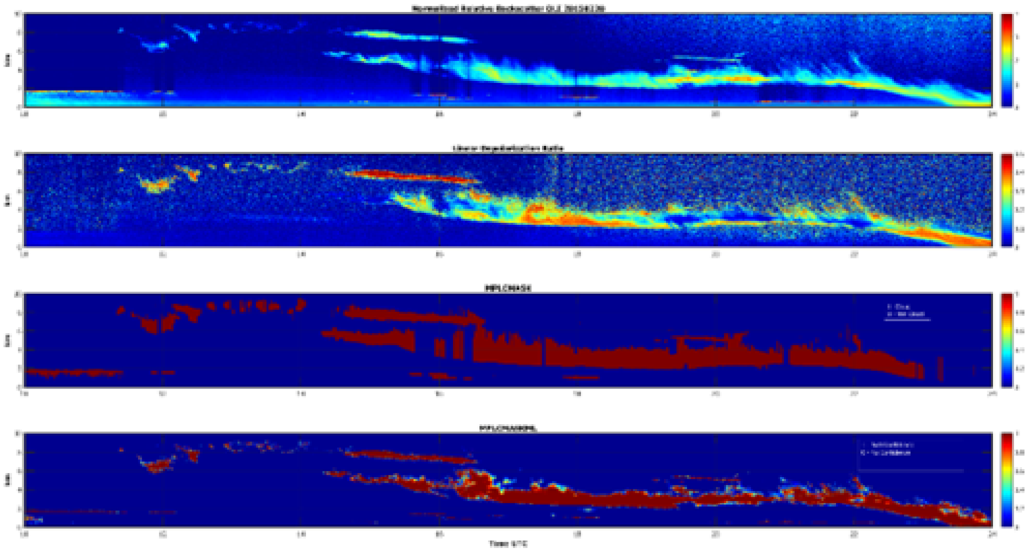

Figure 3. Example plots from Oliktok Point (OLI) show the log of the normalized relative backscatter (log10 NRB), linear depolarization ratio (LDR), and cloud masks for MPLCMASK and MPLCMASKML for February 20, 2015. Mixed-phase stratus clouds prevail 15-24 UTC with boundary-layer clouds also present below 2 km.

9.0 Production Data

Production data is currently available for the following sites and date ranges (inclusive):

Table 5. Available evaluation data.

Site and Facility	Start Date	End Date
SGP C1	2010-07-10	Present day
ENA C1	2013-10-03	Present day
NSA C1	2010-10-09	Present day
OLI M1	2013-09-13	2021-06-13
COR M1	2018-09-27	2019-04-30
AWR M1	2015-12-01	2017-01-02

10.0 References

Cromwell E, and DM Flynn. 2019. "Lidar Cloud Detection with Fully Convolutional Networks." *IEEE Winter Conference on Applications of Computer Vision 2019*: 619-627. PNNL-SA-138402. https://doi.org/10.1109/WACV.2019.00071

Flynn, D, C Sivaraman, J Comstock, and D Zhang. 2020. Micropulse Lidar Cloud Mask (MPLCMASK) Value-Added Product for the Fast-Switching Polarized Micropulse Lidar Technical Report. ARM user facility. DOE/SC-ARM/TR-098.

Lim K-S S, LD Riihimaki, Y Shi, D Flynn, JM Kleiss, LK Berg, WI Gustafson Jr., Y Zhang, and KL Johnson. 2019. “Long-Term Retrievals of Cloud Type and Fair-Weather Shallow Cumulus Events at the ARM SGP Site.” *Journal of Atmospheric and Oceanic Technology* 36(10): 2031–2043. https://doi.org/10.1175/JTECH-D-18-0215.1

Przybylak, R. 2003. *The Climate of the Arctic*. 1st ed. Kluwer Academic.

Serreze, MC, and RG Barry. 2005. *The Arctic Climate System*. Cambridge University Press.

Wang, Z, and K Sassen. 2001. “Cloud Type and Macrophysical Property Retrieval Using Multiple Remote Sensors.” *Journal of Applied Meteorology* 40(10): 1665–1682, https://doi.org/10.1175/1520-0450(2001)040<1665:CTAMPR>2.0.CO;2

www.arm.gov

U.S. DEPARTMENT OF ENERGY

Office of Science