https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021

Atmospheric Measurement Techniques Discussions

Improved cloud detection for the Aura Microwave Limb Sounder: Training an artificial neural network on colocated MLS and Aqua-MODIS data

Frank Werner[1], Nathaniel J. Livesey[1], Michael J. Schwartz[1], William J. Read[1], Michelle L. Santee[1], and Galina Wind[2,3]

[1]Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA
[2]NASA Goddard Space Flight Center, Greenbelt, Maryland, 20771, USA
[3]SSAI Inc., Lanham, Maryland, 20706, USA

Correspondence: Frank Werner (frank.werner@jpl.nasa.gov)

Abstract. An improved cloud detection algorithm for the Aura Microwave Limb Sounder (MLS) is presented. This new algorithm is based on a feedforward artificial neural network and uses as input, for each MLS limb scan, a vector consisting of 1,710 brightness temperatures provided by MLS observations from 15 different tangent altitudes and up to 13 spectral channels in each of 10 different MLS bands. The model has been trained on global cloud properties reported by Aqua's Moderate Resolution Imaging Spectroradiometer (MODIS). In total, the colocated MLS-MODIS data set consists of $162,117$ combined scenes sampled on 208 days over 2005–2020. We show that the algorithm can correctly classify $> 96\%$ of cloudy and clear instances for previously unseen MLS scans. A comparison to the current MLS cloudiness flag used in "Level 2" processing reveals a huge improvement in classification performance. For all profiles in the colocated MLS-MODIS data set, the algorithm successfully detects 97.8% of profiles affected by clouds, up from 15.8% for the Level 2 flagging. Meanwhile, false positives reported for actually clear profiles are reduced to 1.7%, down from 6.2% in Level 2. The classification performance is not dependent on geolocation. The new cloudiness flag is applied to determine average global cloud cover between 2015 and 2019, successfully reproducing the spatial patterns of mid-level to high clouds reported in previous studies. It is also applied to four example cloud fields to illustrate the reliable performance for different cloud structures with varying degrees of complexity. Training a similar model on MODIS-retrieved cloud top pressure yields reliable predictions with correlation coefficients greater than 0.99. The combination of cloudiness flag and predicted cloud top pressure provides the means to identify MLS profiles in the presence of high-reaching convection.

1 Introduction

The impact of clouds on Earth's hydrological, chemical, and radiative budget is well established (e.g., Warren et al., 1988; Ramanathan et al., 1989; Stephens, 2005). With the introduction of satellite imagery, the first studies of cloud observations from

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021


Atmospheric Measurement Techniques Discussions

space concentrated on the determination of cloud amount (e.g., Arking, 1964; Clapp, 1964). After the advent of multispectral satellite radiometry, retrievals of increasingly comprehensive suites of cloud macrophysical, microphysical, and optical characteristics were developed (e.g., Rossow et al., 1983; Arking and Childs, 1985; Minnis et al., 1992; Kaufman and Nakajima, 1993; Han et al., 1994; Platnick and Twomey, 1994). Such efforts require a reliable cloud detection prior to the actual retrieval process. Conversely, there are remote sensing applications where clouds, rather than being the subject of interest, are a source of artifacts that negatively impact the observation of desired geophysical variables. For land and water classifications, clouds and cloud shadows represent unusable data points that need to be detected accurately and discarded (e.g., Ratté-Fortin et al., 2018; Wang et al., 2019). Because of the similar spectral behavior of aerosols and clouds, and their complicated interactions, deriving reliable aerosol properties from space requires careful cloud detection with high spatial resolution (e.g., Varnai and Marshak, 2018). Meanwhile, instruments operating in the ultraviolet to infrared spectral wavelength ranges cannot penetrate any but the optically thinnest clouds. As a result, retrievals of atmospheric composition in the presence of clouds are severely limited.

Approaches to cloud detection from satellite-based imagers are characterized by varying levels of complexity, from simple thresholding and contrast methods to multi-level decision trees (e.g., Ackerman et al., 1998; Ackerman et al., 2008; Zhao and Di Girolamo, 2007; Saponaro et al., 2013; Werner et al., 2016). In recent years fast machine learning algorithms have been employed to detect cloudiness based on observed spatial and spectral patterns (e.g., Saponaro et al., 2013; Jeppesen et al., 2019; Sun et al., 2020). Regardless of the technique, each algorithm must be designed purposefully and with the respective application in mind, as discussed in Yang and Di Girolamo (2008).

The Aura Microwave Limb Sounder (MLS), which has provided global retrievals of atmospheric constituent profiles from $\sim$10 km to $\sim$90 km since 2004, operates at frequencies from 118 GHz to 2.5 THz. In this spectral range clouds are much more transparent than at shorter wavelengths, and the impact on the measured radiances is low. Only clouds with high liquid and/or ice water content reaching altitudes of $\sim$9 km and higher can significantly impact the sampled radiances. The current MLS "Level 2" cloud detection algorithm is based on the computation of cloud induced radiances (T_{cir}), which represent the difference between individual observations and calculated clear sky radiances (Wu et al., 2006). The latter are derived after the retrieval of the other MLS data products. To first order, scattering from thick clouds diverts a mix of large upwelling radiances from lower in the atmosphere and smaller downwelling radiances from above into the MLS raypath. Accordingly, for sufficiently thick clouds within the MLS field of view, T_{cir} will be positive for limb pointings above an altitude of $\sim$ 9 km, where non-scattered limb views are characterized by low radiances. Conversely, T_{cir} will be negative below $\sim$ 9 km, where non-scattered signals would otherwise be large. In the MLS Level 2 processing, if the absolute value of T_{cir} exceeds predefined detection thresholds, then the respective profile is flagged as being influenced by high or low clouds, respectively. The thresholds are set for individual retrieval phases and spectral bands; e.g., for MLS bands 7–9, around a center frequency of 240 GHz, radiances are flagged where $T_{\mathrm{cir}} > 30$ K or $T_{\mathrm{cir}} < -20$ K. Subsequently, separate retrieval algorithms deduce ice water content and path from the T_{cir} information (Wu et al., 2008). Note that in earlier phases of the MLS Level 2 processing, a similar scheme, computing clear sky radiances based on preliminary retrievals of temperature and composition, is used to

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021


Atmospheric Measurement Techniques Discussions

identify MLS radiances that have been significantly affected by clouds and discount them in the final atmospheric composition retrievals.

The focus for the Level 2 flagging is on identifying cases where clouds impact the MLS signals sufficiently to potentially affect the MLS composition retrievals. However, the reliance on estimated clear sky radiances and the use of predefined thresholds induces uncertainties in the current algorithm. For optically thinner clouds, where T_{cir} values are close to but do not exceed the prescribed thresholds, the current cloud flag will provide a false clear classification. Improvements to the current cloud detection scheme could allow: (i) a comprehensive uncertainty analysis of the retrieval bias induced by clouds, (ii) more reliable MLS retrievals in the presence of clouds, where a potential future correction of MLS radiances could account for the cloud influence, (iii) identification of composition profiles that can be confidently considered to be completely clear sky, and (iv) the reliable identification of profiles in the presence of high-reaching convection. Points (iii) and (iv) have the potential to enable new science studies. For example, a reliable cloud mask for individual MLS profiles would enable more comprehensive analysis of lower-stratospheric water vapor enhancements associated with overshooting convection. Currently, studies of these events rely on computationally expensive colocation of water vapor profiles with cloud properties from different observational sources (e.g., Tinney and Homeyer, 2020; Werner et al., 2020; Yu et al., 2020).

This study describes the training and validation of an improved MLS cloud detection scheme employing a feedforward artificial neural network ("ANN" hereinafter). This algorithm is designed to classify clear and cloudy conditions for individual MLS profiles, based purely on the sampled MLS radiances. Two specific goals are set for the new algorithm: (i) detection of both high and mid-level clouds (e.g., stratocumulus and altostratus), and (ii) detection of less opaque clouds containing lower amounts of liquid or ice water. Observed cloud conditions, used to train the ANN, are provided by the cloud products reported by the Moderate Resolution Imaging Spectroradiometer (MODIS) aboard NASA's Aqua platform. Of the major satellite instruments, Aqua MODIS observations are ideal for this study, as they provide operational cloud products on a global scale that are essentially coincident and concurrent with the MLS observations.

The manuscript is structured as follows: section 2 describes both the MLS and MODIS data used in this study. Then a short introduction to the general setup of a feedforward ANN is given in section 3.1, followed by specifics on the output (section 3.2), input (section 3.3), and the training and validation procedure (section 3.4) of the developed models. Results of applying the new algorithm to MLS data are given in section 4, which includes a statistical comparison of the prediction performance between the Level 2 and new cloud detection schemes (section 4.1), a discussion about ANN performance for uncertain cases (section 4.2), a global performance evaluation and cloud cover analysis (section 4.3), and four examples scenes contrasting the performance of the Level 2 flag and the new algorithm for different cloud fields in section 4.4. Subsequent training on MODIS-retrieved cloud top pressure, an evaluation of the prediction performance, and three example scenes are presented in section 5. The main conclusions and a brief summary are given in section 6.

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021


Atmospheric Measurement Techniques Discussions

2 Data

Aura MLS samples brightness temperatures (T_{B}) in five spectral frequency ranges around 118, 190, 240, 640, and 2,500 GHz (Waters et al., 2006) (the latter, measured with separate, independent optics, was deactivated in 2010 and is not considered here). Multiple bands, consisting of 4–25 spectral channels, cover each of these frequency ranges. The exact position of the respective bands is dictated by the different absorption features of the various atmospheric constituents that MLS observes. MLS makes ≈ 3500 daily vertical limb scans (called major frames; MAFs), each consisting of 125 minor frames (MIFs) that can be associated with tangent pressures ($\mathrm{p_{tan}}$) at different altitudes in the atmosphere. These observations provide the input for profile retrievals of a wide-ranging set of atmospheric trace gas concentrations including water vapor, ozone, and nitric acid. The respective Level 2 Geophysical Product (L2GP) files also report a status diagnostic for every MLS profile, which includes flags indicating high and low cloud influence. The most recent MLS dataset is version 5; however, at the time the ANN was being developed, reprocessing of the entire 16-year MLS record with the v5 software had not yet been completed. Accordingly, L2GP cloudiness flags in this study are provided by the version 4.2x data products (Livesey et al., 2020), and v4.2x is also the source for the Level 1 radiance measurements used herein. The spatial resolution of MLS Level 2 products varies from species to species, but typical values are 3 km in the vertical and 5×500 km in the cross-track and along-track dimensions. The distance along the orbit track between adjacent sampled profiles is ≈ 165 km.

Global cloud variables used in this study are provided by retrievals from the Aqua-MODIS instrument, which precedes the Aura overpass by about 15 minutes. However, because of the differences in their viewing geometries, the true time separation between MLS and MODIS measurements is substantially smaller than 15 minutes (see section 3.2). MODIS collects radiance data from 36 spectral bands in the wavelength range between 0.415–14.235 μm. For a majority of the channel observations and subsequently retrieved cloud properties, the spatial resolution at nadir is $1,000$ m, although the pixel dimensions increase towards the edges of a MODIS granule. Each granule has a viewing swath width of $2,330$ km, enabling MODIS to provide global coverage every two days. More information on MODIS and its cloud product algorithms (the current version is Data Collection 6.1) is given in Ardanuy et al. (1992); Barnes et al. (1998); Platnick et al. (2017). Each pixel, j, within a MODIS granule reports a value for the cloud flag, a cloud top pressure ($p^{\mathrm{j}}_{\mathrm{CT}}$), cloud optical thickness (τ^{j}), and effective droplet radius ($r^{\mathrm{j}}_{\mathrm{eff}}$). These last two variables are used to derive the total water path ($Q^{\mathrm{j}}_{\mathrm{T}}$), which contains both the liquid and ice water path and characterizes the amount of water in a remotely sensed cloud column. It can be calculated following the discussions in Brenguier et al. (2000); Miller et al. (2016):

$$Q^{\mathrm{j}}_{\mathrm{T}} = \Gamma \cdot \rho^{\mathrm{j}} \cdot \tau^{\mathrm{j}} \cdot r^{\mathrm{j}}_{\mathrm{eff}}, \tag{1}$$

where ρ^{j} is the bulk density of water in either the liquid or ice phase (following the cloud phase retrieval for pixel j), and the factor Γ accounts for the vertical cloud structure. For vertically homogeneous clouds it can be shown that $\Gamma = 2/3$.

Table 1 lists the 208 days that comprise the global data set used in this study. It consists of eleven random days from each year between 2005 and 2020, as well as a pair of two consecutive days to bring the yearly coverage to thirteen days. Particular attention was paid to ensure that each month is represented (close to) equally in the final data set.

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021

Atmospheric Measurement Techniques Discussions

3 Artificial neural network

This section provides details about the ANN setup and training. Here, we constructed and trained a multilayer perceptron, which is a subcategory of feedforward ANNs that sequentially connects neurons between different layers. An introduction to multilayer perceptrons is given in section 3.1. The output vector containing the labels (i.e., the binary cloud classifications) based on a colocated MLS-MODIS data set, and the input matrices, which consist of MLS T_{B} observations, are described in sections 3.2 and 3.3, respectively. The choice of hyperparameters, the training setup, and the validation results from the algorithm are provided in section 3.4.

The weights that connect the input to the output data are determined by the "Keras" library for Python (version 2.2.4; Chollet et al., 2015) with "TensorFlow" (version 1.13.1) as the backend (Abadi et al., 2016).

3.1 Algorithm description

Figure 1 illustrates the general setup of a simplified multilayer perceptron that contains four layers. The input layer (shown in blue) consists of $m = 3$ vectors that contain selected MLS brightness temperatures $\mathbf{T}_{\mathrm{B1}}$, $\mathbf{T}_{\mathrm{B2}}$, and $\mathbf{T}_{\mathrm{B3}}$. The input layer is succeeded by two hidden layers (shown in green) with two neurons each ($\mathrm{N}_{\mathrm{h1}-1}$ and $\mathrm{N}_{\mathrm{h1}-2}$, as well as $\mathrm{N}_{\mathrm{h2}-1}$ and $\mathrm{N}_{\mathrm{h2}-2}$) and the respective bias vectors ($\mathbf{B}_1$ and $\mathbf{B}_2$). The following output layer (shown in orange) consists of a single vector ($\mathbf{L}$; containing the predicted labels) and a corresponding bias (B_{L}). The brightness temperature vectors ($\mathbf{T}_{\mathrm{Bi}}$; $i = 1, 2, 3$) used as input for the ANN are provided by T_{B} observations in selected channels, bands, and minor frames. They are of length n, which describes the number of scalar MLS observations ($T^{\mathrm{j}}_{\mathrm{Bi}}$). This means, that $i = 1, 2, 3$ brightness temperatures were sampled by MLS at $j = 1, \ldots, n$ major frames. Similarly, there is a scalar label L^{j} for each MAF, so $\mathbf{L}$ is also of length n.

At each neuron $N_{\mathrm{h1}-k}$, k =1–2 in the first hidden layer a scalar value $\gamma^{\mathrm{j}}_{1-1}$ and $\gamma^{\mathrm{j}}_{1-2}$ for each of the j MAFs is calculated:

$$\gamma^{\mathrm{j}}_{1-1} = B_{1-1} \cdot \omega_{0,1} + T^{\mathrm{j}}_{\mathrm{B1}} \cdot \omega_{1,1} + T^{\mathrm{j}}_{\mathrm{B2}} \cdot \omega_{2,1} + T^{\mathrm{j}}_{\mathrm{B3}} \cdot \omega_{3,1} \quad (2)$$

$$\gamma^{\mathrm{j}}_{1-2} = B_{1-2} \cdot \omega_{0,2} + T^{\mathrm{j}}_{\mathrm{B1}} \cdot \omega_{1,2} + T^{\mathrm{j}}_{\mathrm{B2}} \cdot \omega_{2,2} + T^{\mathrm{j}}_{\mathrm{B3}} \cdot \omega_{3,2}. \quad (3)$$

These values are subsequently modified by an activation function, which introduces non–linearity into the neuron output. The hyperbolic tangent activation function is applied, which is shown to be very efficient during training because of its steep gradients (e.g., LeCun et al., 1989; LeCun et al., 1998)) and yields new values $\Gamma^{\mathrm{j}}_{1-1}$ and $\Gamma^{\mathrm{j}}_{1-2}$. For the second hidden layer, the scalar neuron values at $N_{\mathrm{h2}-k}$, k =1–2 for each MAF j are derived as:

$$\gamma^{\mathrm{j}}_{2-1} = B_{2-1} \cdot \varpi_{0,1} + \Gamma^{\mathrm{j}}_{1-1} \cdot \varpi_{1,1} + \Gamma^{\mathrm{j}}_{1-2} \cdot \varpi_{2,1} \quad (4)$$

$$\gamma^{\mathrm{j}}_{2-2} = B_{2-2} \cdot \varpi_{0,2} + \Gamma^{\mathrm{j}}_{1-1} \cdot \varpi_{1,2} + \Gamma^{\mathrm{j}}_{1-2} \cdot \varpi_{2,2}. \quad (5)$$

As before, these values are transformed by the hyperbolic tangent activation function, which yields the transformed neuron values $\Gamma^{\mathrm{j}}_{2-1}$ and $\Gamma^{\mathrm{j}}_{2-2}$.

Finally, the neuron output from $\mathrm{N}_{\mathrm{h2}-1}$ and $\mathrm{N}_{\mathrm{h2}-2}$ is connected to the single vector $\mathbf{L}$ in the output layer. For each MAF j the respective scalar value λ^{j} is calculated as:

$$\lambda^{\mathrm{j}} = B_{\mathrm{L}} \cdot \Omega_0 + \Gamma^{\mathrm{j}}_{2-1} \cdot \Omega_1 + \Gamma^{\mathrm{j}}_{2-2} \cdot \Omega_2. \quad (6)$$

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021

Atmospheric Measurement Techniques Discussions

We aim for a binary, two–class classification setup (i.e., either cloudy or clear designations). As a result, the softmax function normalizes the λ^{j} results at the output layer. The softmax activation function is identical to the logistic sigmoid function for a binary, two–class classification setup. This means that the predicted neuron output in the output layer is calculated as:

$$\hat{L}^{\mathrm{j}} = \frac{1}{1+\exp\left(-\lambda^{\mathrm{j}}\right)}. \tag{7}$$

The ideal weights in Eqs. (2), (3), (4), (5) and (6) need to be derived iteratively by evaluating a loss function (χ), which is the log–loss function (or cross-entropy) in the classification setup. If L^{j} and $\hat{L}^{\mathrm{j}}$ are the individual elements of the two output vectors $\mathbf{L}$ and $\hat{\mathbf{L}}$ (i.e., the prescribed and currently predicted labels), χ for two classes is defined as:

$$\chi = -\sum_{j=1}^{n} L^{\mathrm{j}} \cdot \ln\left(\hat{L}^{\mathrm{j}}\right) + \left(1-L^{\mathrm{j}}\right) \cdot \ln\left(1-\hat{L}^{\mathrm{j}}\right). \tag{8}$$

Note that in case of $L^{\mathrm{j}} = 0$ or $\hat{L}^{\mathrm{j}} = 0$ an infinitesimal quantity $\epsilon \approx 0$ is added to the respective label to avoid the undefined $\ln 0$. The "Keras" algorithm includes multiple optimizers to solve Eq. (8) in a numerically efficient way. The exact setup and choice of hyperparameters need to be determined carefully via cross-validation during the training process (see section 3.4).

3.2 The labels from colocated MLS-MODIS cloud data

Training data for the output vector $\mathbf{L}$, which contains the prescribed labels for Eq. (8), is provided by the MODIS C6.1 data set described in section 2. The reported MODIS cloud products are first colocated with individual MLS profiles. The illustration in Figure 2a depicts how colocation is performed. If n_{per} is the number of MODIS pixels (gray shaded squares) within a $1^\circ \times 1^\circ$ box (in latitude and longitude; blue box) around an MLS profile (blue "x"), then each of the n_{per} pixels reports a cloudiness flag, as well as a total water path ($Q_{\mathrm{T}}^{\mathrm{j}}$) and a cloud top pressure ($p_{\mathrm{CT}}^{\mathrm{j}}$), with $j = 1, 2, ..., n_{\mathrm{per}}$ denoting the individual pixels within the $1^\circ \times 1^\circ$ box. Note that for legibility the cloud properties of only three MODIS pixels are shown. For the respective MLS profile, these parameters are aggregated to more general cloud statistics consisting of the cloud cover (C) within the $1^\circ \times 1^\circ$ box, as well as the median total water path (Q_{T}) and median cloud top pressure (p_{CT}).

Figure 2b shows the global distribution of sample frequencies for the colocated MLS-MODIS training data set within grid boxes of length $60^\circ \times 60^\circ$ (latitude and longitude). While not every grid box contains the same number of profiles, each area contains at least 5,000 samples. Apart from a single grid box over Africa, the higher latitudes tend to contain more samples, because both Aqua and Aura are polar-orbiting satellites. A majority of grid boxes contain more than 8,000 samples.

The aggregated profile-level cloud statistics are used to define the observed clear sky and cloudy conditions. All profiles that are characterized by $C \geq 2/3$, $p_{\mathrm{CT}} < 700\,\mathrm{hPa}$, and $Q_{\mathrm{T}} > 50\,\mathrm{g\,m^{-2}}$ are labeled as cloudy, while profiles with $C < 1/3$ and $Q_{\mathrm{T}} < 25\,\mathrm{g\,m^{-2}}$ are considered to be associated with clear sky samples. While the cloud cover threshold is somewhat arbitrary, the p_{CT} limit for cloudy observations and the Q_{T} thresholds are carefully selected. The large opacity of the atmosphere for longer path lengths means that MLS shows almost no sensitivity towards clouds with $p_{\mathrm{CT}} > 700\,\mathrm{hPa}$ (see section 3.3). This upper pressure limit, which in the 1976 US Standard Atmosphere (COESA, 1976) is located at an altitude of $\sim 3\,\mathrm{km}$, is around the lower limit of observed cloud tops of mid-level cloud types (e.g., altostratus, altocumulus). Meanwhile, the 10^{th} and 25^{th}

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021


Atmospheric Measurement Techniques Discussions

percentiles of all profiles containing clouds within the 1° perimeter, regardless of C, are $Q_\mathrm{T} \approx 25\,\mathrm{g\,m^{-2}}$ and $Q_\mathrm{T} \approx 50\,\mathrm{g\,m^{-2}}$, respectively. These definitions have an additional benefit: they almost evenly split the data set into cloudy and clear sky profiles (52.0% and 48.0%, respectively), which improves the reliability of the trained weights for the cloud classification. Naturally, these definitions leave some profiles undefined (e.g., those with C in the range 1/3–2/3). These profiles (about the number of the combined cloudy and clear classes) cannot be included in the training of the ANN, as they lack a prescribed label. The discussion in section 4.1 provides an analysis of the ANN performance for a redefined classification based on a simple threshold of $C = 0.5$ (in addition to a positive Q_T) to distinguish between cloudy and clear sky profiles.

It is important to note that the difference in viewing geometry between MLS and MODIS (i.e., limb geometry versus nadir viewing) induces a considerable degree of uncertainty in the colocation. While it is reasonable to assume that the majority of a potential cloud signal (or lack thereof) will come from the $1^\circ \times 1^\circ$ box around the respective MLS profile, there are certain scenarios that will lead to a false classification. The most likely such scenario consists of an MLS line-of-sight that passes through a high-altitude cloud before a clear sky $1^\circ \times 1^\circ$ box. Here, MLS will detect a strong cloud signal, even though the nadir-viewing MODIS instrument does not record any cloudiness at the location of the respective MLS profile. Less likely is the scenario of a very low-altitude cloud located right after (in terms of an MLS line-of-sight) a clear sky $1^\circ \times 1^\circ$ box. This would also result in a false cloud classification (if the MODIS observations are taken as reference). However, because of the increase in atmospheric opacity, the sensitivity of the MLS instrument towards signals further along the line-of-sight decreases, and it is less likely that MLS would detect these cloud signals in any case. One contributor to the overall uncertainty that is of less concern is the time difference between the Aqua and Aura orbits (≈ 15 minutes). Because MLS looks forward in the limb, the temporal discrepancy between the sampling of individual MLS profiles and the colocated MODIS pixels is in the range of 0.6–1.4 minutes. The results presented in section 3.4 illustrate that by training the ANN with a large data set, as well as cross-validating the training results against a large number of random validation data, the contributions of uncertainties associated with colocation (both in space and time) can be considered small and do not overly impact the reliability of the cloud detection algorithm.

3.3 The input matrix from MLS brightness temperature observations

Figures 3a-c show the spectral behavior of T_B sampled in MLS bands 2, 33, and 14 at MIF=15, which on average corresponds to $p_\mathrm{tan} \sim 576\,\mathrm{hPa}$ (at an altitude of $\sim 4.5\,\mathrm{km}$ in the 1976 US Standard Atmosphere). In this section we mostly omit the superscript "j" to indicate the statistical analysis of all T_B^j in the respective band ($j = 1, 2, \cdots, n$). The median T_B for profiles associated with clear sky (orange) and cloudy conditions (blue), based on the classifications from the colocated MLS-MODIS data set described in section 3.2, are shown by the solid lines and circles. The shaded orange and blue areas indicate the interquartile range (IQR; 75^th-25^th percentile of data points) of clear and cloudy profiles. Data are from profiles sampled in the latitudinal range of -30° to $+30^\circ$.

Median clear sky profiles exhibit consistently larger T_B than cloudy observations, with differences of up to 10 K. This behavior confirms the findings in Wu et al. (2006), where ice clouds at an altitude of 4.7 km reduce band 33 T_B at the lower

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021


Atmospheric Measurement Techniques Discussions

minor frames (i.e., larger p_{tan}). The IQR ranges of the two different data sets are very close for band 2 observations (i.e., within 1–2 K), while there is overlap for the T_{B} sampled in bands 33 and 14.

To illustrate the reduced sensitivity of MLS to signals from very low clouds, the median T_{B} from profiles with $p_{\text{CT}} > 700\,\text{hPa}$ is shown in green (for clarity the corresponding IQR is omitted). These profiles behave similarly to clear sky observations, and the difference in median T_{B} is less than 1 K.

Figures 3d-f illustrate the spectral behavior of T_{B} sampled at MIF=33, which corresponds to an average p_{tan} of $\sim 200\,\text{hPa}$ (at an altitude of $\sim 12\,\text{km}$ in the 1976 US Standard Atmosphere). Similar to the results for the lower MIF, a clear separation between median T_{B} from clear sky and cloudy ($100\,\text{hPa} \leq p_{\text{CT}} < 700\,\text{hPa}$) profiles is observed, while those profiles associated with low clouds ($p_{\text{CT}} > 700\,\text{hPa}$) again behave similarly to clear samples. For bands 2 and 33 observations the cloudy profiles show significantly higher T_{B}. Again, this confirms the reported behavior in Wu et al. (2006), who found an increase in band 33 T_{B} for cloudy conditions compared to the clear background. Conversely, band 14 observations behave similarly to those sampled at MIF=15 and the cloudy profiles exhibit lower T_{B}.

The significant contrast in median T_{B} between clear sky and cloudy profiles, especially for band 2 and partly for band 33, might suggest the possibility of a simple cloud detection approach via thresholds. However, the respective IQR ranges often overlap, which indicates that a simple T_{B} threshold would miss about 25% of the clear and cloudy data, respectively. Moreover, the behavior illustrated in Figure 3 is specific to the latitudinal range of $-30°$ to $+30°$. For higher latitudes, changes in atmospheric temperature and composition yield a noticeable decrease in the observed contrast, while close to the poles the clear sky profiles almost always have lower T_{B} than the cloudy observations (even at the lower MIFs). A more sophisticated classification approach, with T_{B} samples from additional MLS bands and minor frames, is necessary to derive a more reliable global cloud detection.

Table 2 details the MLS bands, as well as the respective channels and MIFs, that comprise the $m \times n$ input matrix for the ANN. The input matrix consists of m different T_{B}^{j}, sampled in individual channels (within the respective MLS bands) and MIFs, at n different times. To reduce the computational costs during the training of the model, not all MLS observations are considered. Instead, three different bands are chosen from the 190, 240, and 640 GHz regions, respectively. Those are bands 2, 3, 6, bands 7, 8, 33, and bands 10, 14, 28 for the three receivers. These bands were carefully selected after a statistical analysis of the altitude-dependent contrast in observed T_{B} between clear and cloudy profiles. This contrast is generally low (in the range of 1 K) for the observations from the 118 GHz region, so only band 1 from this receiver is included in the model input. For most bands, every second channel is included in the input (except for band 33, which only has 4 channels in total), while considering every third MIF in the range 7–49 yields a decent vertical resolution between 15 hPa (for the highest altitudes) and 150 hPa (at the lowest altitudes). Overall, the input matrix for the training and validation of the ANN is of shape $1,710 \times 162,117$, i.e., it consists of $m = 1,710$ different features (T_{B}^{j} at different frequencies and altitudes) from $n = 162,117$ MAFs (either classified as clear sky or cloudy).

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021


Atmospheric Measurement Techniques Discussions

3.4 Training and validation

The "Keras" python library provides convenient ways to manage the setup, training, and validation of ANN models. The optimal weights for Eqs. (2), (3), (4), (5) and (6) are derived in three steps: (i) determining the most appropriate hyperparameters via k-fold cross-validation, (ii) training and validating a number of different models with the best set of hyperparameters on multiple, random splits between training and validation data sets, and (iii) comparing the performance scores for the different model runs to evaluate the stability of the approach and pick the best set of weights. Each model is set up with two hidden layers. The number of neurons per hidden layer is set to 856, which corresponds to the average between the number of nodes in the input and output layers (i.e., $1,710$ and 1, respectively).

The hyperparameters to be determined are (i) the optimizer for the cloud classification, (ii) the learning rate, (iii) the mini-batch size, and (iv) the value for the weight decay (i.e., the L2 regularization parameter). While the optimizer choice and learning rate control how quickly and accurately the minimum of the cost function in Eq. (8) is determined, the values for mini-batches and the L2 regularization characterize the level of noise and degree of freedom in the models, which have a noticeable impact on model performance for new, previously unseen data. More information about ANN hyperparameters and their impact on the reliability of model predictions can be found in, e.g., Reed and Markslll (1999); Goodfellow et al. (2016). Note that the number of epochs (i.e., the number of iterations during the training process) is not considered an important hyperparameter for this study. Instead, the models are run with a large number of epochs, and the lowest validation loss is recorded, so an increase in validation loss during the training (i.e., cases where the model is overfitting the training data at some point) has no impact on the overall performance evaluation.

At first, the data set is randomly shuffled and split into $k = 4$ parts. Subsequently, one of the four parts is used as the validation data set, and the other three are used to train the ANN with a certain set of hyperparameters. Here, each of the $1,710$ features is individually standardized, i.e., each feature is transformed to have a mean value of 0 and unit variance. This step is essential for a successful ANN training, as the individual features are characterized by different dynamic ranges. Meanwhile, the labels for clear and cloudy profiles are simply set to 0 and 1, respectively. After model convergence and determination of a set of performance scores, the model is discarded and a different set of three parts is used for training (the remaining fourth part is again used for validation). After cycling through each of the four parts (and recording four sets of performance scores), the set of hyperparameters is changed and the process begins anew. An evaluation of each set of performance scores, for each set of hyperparameters, reveals the ideal setup for the ANN.

The performance scores employed in this study are three commonly used binary classification metrics, based on the calculation of a confusion matrix $\mathbf{M}$ for the two classes (i.e, clear sky and cloudy profiles). If tp and tn are the number of true positives and negatives, respectively, and fp and fn are the number of false positives and negatives, respectively, then the confusion matrix is defined as:

$$\mathbf{M} = \begin{pmatrix} tp & fp \\ fn & tn \end{pmatrix}. \tag{9}$$

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021


Atmospheric Measurement Techniques Discussions

From $\mathbf{M}$ the accuracy (Ac), F1 score ($F1$) and Matthews correlation coefficient (Mcc) can be derived as:

$$Ac = \frac{tp + tn}{tp + tn + fp + fn} \tag{10}$$

$$F1 = \frac{2 \cdot tp}{2 \cdot tp + fp + fn} \tag{11}$$

$$Mcc = \frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp + fp) \cdot (tp + fn) \cdot (tn + fp) \cdot (tn + fn)}}. \tag{12}$$

While Ac quantifies the proportion of correctly classified samples, $F1$ describes the harmonic mean value between precision (proportion of true positives in the positively predicted ensemble, i.e., the ratio of tp to tp+fp) and recall (proportion of correctly predicted true positives, i.e., the ratio of tp to tp+fn). Generally, $F1$ assigns more relevance to false predictions and is more suitable for imbalanced classes. Meanwhile, all elements of the confusion matrix are important in determining the Mcc, which yields values between -1 and 1 and thus is analogous to a correlation coefficient.

This analysis revealed that the stochastic gradient descent optimizer, using a learning rate of 0.001, and a Nesterov momentum value of 0.9 yielded the overall best validation scores. The best weight decay and mini-batch size values were found to be 5×10^{-4} and 1024 (i.e., 0.8% of the training data), respectively.

Due to randomness during the assignment of individual observations to either the training or validation data set, developing a single model might result in evaluation scores that are overly optimistic or pessimistic. By chance, the most obvious cloud cases (e.g., $C = 1$ and very large Q_{T} values) might have ended up in the validation data set, or vice versa, and the trained weights might be inappropriate. Moreover, a large disparity in validation scores for multiple models might be indicative of an ill-posed problem, where the MLS observations do not provide a reasonable answer to the cloud classification problem. Therefore, developing multiple models with a reasonable split of training and validation data, as well as careful monitoring of the spread in validation scores, is imperative. In this study, 100 different models are developed. Before each model run, the data set is randomly shuffled and split into 75% training and 25% validation data. As mentioned earlier, each model is run with a large number of epochs, and the weights associated with the lowest validation loss are recorded.

The output of each ANN model is a cloudiness probability (P) between 0 (clear) and 1 (cloudy). Note that throughout this study we simply group each prediction in either the clear or cloudy class, i.e., MAFs with predicted probabilities $0 \leq P < 0.5$ are considered to be sampled under clear sky conditions, while MAFs with $0.5 \leq P \leq 1$ are considered to be cloudy. The one exception is the discussion in section 4.2, where the actually predicted P are employed to study the ANN performance for undefined cloud conditions (with respect to the clear sky and cloudy definitions presented in section 3.2).

A summary of the derived prediction statistics is shown in Figure 4a. Each histogram shows the average percentage of correctly predicted clear sky (i.e., tn, orange shading) and cloudy (i.e., tp, blue shading) labels for all 100 validation data sets. Also shown are the percentages of false classifications (the blue and orange lines for fp and fn, respectively). The gray shaded horizontal areas at the top of each histogram illustrate the standard deviation for each class, calculated from the 100 validation data sets. The average percentage of correct clear sky and cloudy predictions is 97.4% and 96.8%, respectively, while a false cloudy or clear sky prediction occurs for 2.6% and 3.2% of profiles in the validation data. The standard deviation for all four groups is 0.3%.

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021


Atmospheric Measurement Techniques Discussions

Figure 4b shows a scatter plot of all Mcc values as a function of the respective $F1$. Even though the Mcc penalizes false classifications more severely than $F1$, a high Pearson's product-moment correlation coefficient of > 0.99 is observed. Moreover, there is little variability in the 100 derived binary statistical metrics, with average Ac, $F1$, and Mcc values of 0.971 ± 0.001, 0.971 ± 0.001, and 0.942 ± 0.003. These results illustrate that the derived models are well suited to predict cloudiness for new MLS data (i.e., measurements not involved in the training of the models) and that the trained weights are very stable (i.e., all models exhibit very similar binary statistics, regardless of the training or validation data set).

Given the statistical robustness of the results, the model with the highest Mcc provides the weights for cloud classification going forward.

4 Results and examples

This section includes a detailed comparison between the cloud classifications from the current MLS v4.2x and the new ANN-based algorithms in section 4.1, followed in section 4.2 by a discussion of predicted cloudiness probabilities that illustrates the performance of the new ANN cloud flag for less confident cases (i.e., those outside of the training and validation data sets). This section also presents an analysis of the latitudinal dependence of the ANN performance and derived global cloud cover statistics in section 4.3, as well as a close-up look at cloudiness predictions for some example scenes over both the North American and Asian monsoon regions (section 4.4).

4.1 Prediction performance of current L2GP and new ANN cloud flag

The analysis in section 3.4 indicates that the ANN setup can reliably identify MLS profiles that have been influenced by substantial cloudiness within a $1^\circ \times 1^\circ$ box. Figure 5 provides a closer look at the performance of the new ANN-based and v4.2x cloud flags for all $n = 162,117$ profiles associated with either the clear sky or cloudy class. Note that these results include scenes taken from both the training and the validation data sets. This yields a slight increase in classification performance compared to the results shown in Figure 4, because the comparison includes data the models have been trained on. This increase in classification performance is in the range of $\sim 1\%$ and is expected to decrease if future versions of the ANN are trained on even larger, more comprehensive data sets.

Figures 5a and c present the percentage of correctly classified (blue) and falsely classified (orange) cloudy profiles, as determined by the cloudiness definition for the colocated MLS-MODIS data set described in section 3.2. The frequency of predicted labels from the (a) new ANN-based algorithm and (c) v4.2x cloud flag are shown as a function of Q_{T}. Note that because of the general cloudiness definition, only those profiles with $Q_{\mathrm{T}} > 50\,\mathrm{g\ m^{-2}}$ are considered (see section 3.2). The flags predicted by the ANN correctly classify $82,503$ (97.8%) of the cloudy profiles. In particular, the thickest clouds, those with $Q_{\mathrm{T}} \geq 1,000\,\mathrm{g\ m^{-2}}$, are detected in 95.8% of cases. Conversely, the current v4.2x status flag only detects $13,338$ of the $84,323$ cloudy profiles (i.e., 15.8%). A peak of 15.2% of clouds are missed for low Q_{T}, where the ANN performs significantly better. This is understandable, as the current v4.2x status flags for high and low cloud influences should only be set for profiles

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021


Atmospheric Measurement Techniques Discussions

where the extinction along the line-of-sight is large enough to be attributed to a fairly thick cloud. However, even for very large $Q_{\mathrm{T}} \geq 1{,}000\,\mathrm{g\ m^{-2}}$, only 187 of 645 (29.0%) of the cloudy profiles are detected.

Histograms for clear sky observations as a function of C are presented in Figures 5b and d. Only 1.7% of clear profiles are falsely classified as cloudy by the new ANN algorithm, while the current v4.2x status flag mislabels 6.2% of these profiles. Most of the clear observations occur for very low values of $C < 0.05$, of which the ANN and v4.2x flags detect 51.6% and 48.1%, respectively. Note that the slightly larger fraction of false positives from the v4.2x flag is not necessarily incorrect, i.e., there might actually be clouds in the line-of-sight of one or more MLS scans associated with the respective profiles. They might, however, be well before (very high clouds) or past (very low clouds) the tangent point and outside of the $1^\circ \times 1^\circ$ box defined in section 3.2.

Table 3 gives an overview of the confusion matrix elements for each cloud flagging scheme, as well as metrics to evaluate binary statistics. The new ANN algorithm yields values of $Ac = 0.98$, $F1 = 0.98$, and $Mcc = 0.96$, confirming the reliable classification performance shown in Figure 5. The v4.2x flag yields low binary performance scores of $Ac = 0.53$, $F1 = 0.26$, and $Mcc = 0.15$, mainly due to the low fraction of true positives.

4.2 Probabilities for different cloud conditions

The clear sky and cloudy classes defined in section 3.2 leave a number of profiles unaccounted for (i.e., neither clear sky nor cloudy), such as those with $1/3 \leq C < 2/3$ or $p_{\mathrm{CT}} > 700\,\mathrm{hPa}$. While it is reasonable to only train the model on the confidently clear and cloudy conditions, it is essential to understand the ANN performance for the undefined, in-between cases.

Figure 6a shows average ANN-predicted cloudiness probabilities as a function of C and Q_{T} with no restrictions on p_{CT}. P values are distributed into four groups: confidently clear ("Conf. Clr."; $P < 0.25$), probably clear ("Prob. Clr."; $0.25 \leq P < 0.5$), probably cloudy ("Prob. Cld."; $0.5 \leq P < 0.75$), and confidently cloudy ("Conf. Cld."; $P \geq 0.75$). The previously defined clear sky and cloudy regions, which comprise the training and validation data sets, are indicated by the white and black dashed lines, respectively. Profiles with low $C < 1/3$ and $Q_{\mathrm{T}} < 25\,\mathrm{g\ m^{-2}}$, regardless of p_{CT}, are characterized by the lowest P values, reliably reproducing the clear sky class defined in section 3.2. Meanwhile, almost all profiles with $C > 0.8$ are flagged to be probably cloudy ($P > 0.5$). However, only profiles that also have $Q_{\mathrm{T}} > 100\,\mathrm{g\ m^{-2}}$ are reliably predicted to have $P > 0.75$. The less-confident identification of the $Q_{\mathrm{T}} > 100\,\mathrm{g\ m^{-2}}$ cases reflects the fact that many of them have low cloud tops, $p_{\mathrm{CT}} > 700\,\mathrm{hPa}$, and are thus not readily observed by MLS. As noted in section 3.3, these profiles exhibit similar spectral behavior to clear ones and the ANN is expected to miss most of these clouds. With increasing Q_{T} even profiles with smaller cloud fractions (as little as $C = 0.25$) are flagged as cloudy. Note that the P results become noisy for very large $Q_{\mathrm{T}} > 500\,\mathrm{g\ m^{-2}}$, conditions that are only observed for less than 4% of the total samples (< 1% for $Q_{\mathrm{T}} > 1000\,\mathrm{g\ m^{-2}}$).

The behavior of predicted P for observations with $p_{\mathrm{CT}} < 700\,\mathrm{hPa}$ is shown in Figure 6b. Here, the confidently clear predictions remain largely unchanged. However, probably cloudy predictions are observed for $C > 0.5$, even for low Q_{T}. Confidently cloudy predictions dominate the previously defined cloudy class ($C > 2/3$ and $Q_{\mathrm{T}} > 50\,\mathrm{g\ m^{-2}}$).

In order to evaluate the ANN performance when more of these uncertain cases are encompassed in the validation, we included in Table 3 a comparison of the binary performance scores for a redefined set of the cases classified as clear and

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021


Atmospheric Measurement Techniques Discussions

cloudy according to less conservative thresholds for the cloud cover and the total water path ($C < 0.5$ and $Q_{\mathrm{T}} < 25\,\mathrm{g\ m^{-2}}$ for clear sky profiles, $C \geq 0.5$ and $Q_{\mathrm{T}} \geq 25\,\mathrm{g\ m^{-2}}$ for cloudy profiles). These changes increase the validation data set from $n = 162,117$ to $n = 328,286$ profiles. Due to the looser definitions, there is a significant drop in performance scores, which can mostly be attributed to a lower true positive rate (i.e., cloud detection) of 0.69 and 0.08 for the ANN classification and v4.2x, respectively. The fraction of false positives (i.e., false prediction of cloudiness for actually clear profiles) remains basically unchanged (increases of 0.02 and 0.00 for the ANN and v4.2x flags, respectively). This means that even with a looser cloudiness definition, the ANN does not yield a multitude of false cloud classifications; rather, the algorithm fails to detect a larger fraction of cloudy profiles. This mostly applies to lower-level clouds and those with small Q_{T}. The ANN still detects 73.0% of cloudy profiles with $Q_{\mathrm{T}} \geq 1,000\,\mathrm{g\ m^{-2}}$ (compared to 17.3% for the v4.2x flag). As a consequence of the reduced true positive rates for the redefined class definitions, the derived $F1$ for the ANN score is reduced to 0.81 (from 0.98), while $F1$ for the current v4.2x flag drops from 0.26 to 0.14.

4.3 Geolocation-dependent performance and global cloud cover distribution

The spectral behavior for clear sky and cloudy profiles shown in Figure 3 only applies for observations made in the latitudinal range of $-30°$ to $+30°$. As mentioned in section 3.3, the contrast between the two classes of data decreases for increasing latitude. While the analysis in section 4.1 illustrates that the new ANN-based cloud classification can reliable identify cloudy profiles, it is important to make sure that there is no latitudinal bias in the cloud detection, i.e., assuring that the algorithm performance is good for MLS observations at all latitude bands.

Calculated $F1$ determined from the ANN model setup is shown in Figure 7a for different regions of the globe. Each grid box covers an area of $60° \times 60°$ (latitude and longitude) and includes on average $9,007$ profiles. A negligible dependence on geographical region is observed and the binary performance metrics exhibit high values throughout. Derived $F1$ values vary between 0.972–0.986, with an average and standard deviation of 0.981 ± 0.003. The range in Mcc (not shown) is 0.931–0.972, with an average and standard deviation of 0.957 ± 0.01.

In contrast to the results for the ANN algorithm, there is a clear latitudinal dependence for the performance of the v4.2x algorithm, illustrated in Figure 7b. $F1$ values are in the range of 0.39–0.48 in the tropics and < 0.24 everywhere else (note the different $F1$ scales in panels a and b)

As there is no significant geographical dependence in prediction performance, the ANN algorithm is applied to derive global cloud cover maps, based solely on the MLS observed T_{B} and the calculated model weights. A map of cloudiness from all MLS profiles sampled over 2015–2019, averaged within $3° \times 5°$ (latitude and longitude) grid boxes, is shown in Figure 7c. Note that this data set includes more than 6 million MLS profiles, while only 65 days in the 5-year span were part of the training data. Profiles are considered to be cloudy when predicted $P \geq 0.5$. Three large-scale regions close to the equator show the largest average cloud covers with $C > 80\%$ (dark orange colors): (i) an area over the northern part of South America, (ii) central Africa, and (iii) a large band encompassing the Maritime Continent. Large zonal bands of $C \approx 60\%$ are observed in the mid-latitudes of both hemispheres. Conversely, large areas of low $C < 20\%$ are observed west of the North American, South American, and African continents, as well as over Australia, northern Africa, and Antarctica. The derived cloud covers,

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021


Atmospheric Measurement Techniques Discussions

as well as the observed spatial patterns of mid to high clouds, agree well with those reported in King et al. (2013); Lacagnina and Selten (2014).

As before, we are interested in comparing the results of the new ANN classification to the ones from the current v4.2x cloud flag. Therefore, a similar map of derived global cloud cover from the current v4.2x cloud flag is shown in Figure 7d. In contrast to the ANN results, calculated $C < 32\%$ almost everywhere. This behavior is consistent with the focus of the v4.2x classification, where only very opaque clouds around ∼300 hPa are flagged. The global patterns identified by the new ANN flag are reproduced, albeit with much lower results for C. However, the v4.2x flag yields a global maximum of $C > 72\%$ over Antarctica. Here, the new ANN flag reports C as low as 3%. This behavior in the v4.2x cloud flag is a well-understood feature caused by misinterpretation of radiances that are reflected by the surface (the unique combination of high topography and low optical depth makes Antarctica one of the few places where MLS can observe the Earth's surface).

Figures 7e-f show similar cloud cover maps generated from Aqua-MODIS observations. Due to the size of that data set and the high computational costs, only samples from 2019 are included here. The cloud cover maps were generated considering cloud mask flag values of 0 and 1 (confident cloudy and probably cloudy) as defined in Menzel et al. (2008). All available 1 km-resolution MODIS cloud mask data was considered. The aggregation used the high-resolution cloud top pressure product, not generally available as a global aggregation. This cloud top pressure product however is the one utilized by retrievals of MODIS cloud optical properties. Such custom aggregation thus ensures the maximum dataset consistency across variables. While all clouds are considered in the map in panel e, only clouds with $p_{\mathrm{CT}} < 700$ hPa are included to derive C in panel f. It is obvious that including clouds with $p_{\mathrm{CT}} > 700$ hPa dramatically increases the derived cloud covers. Due to the reduced sensitivity towards such clouds (see the discussion in section 3.3), the cloud covers predicted by the ANN are much closer to the MODIS results that do not include low clouds. Nonetheless, the ANN-derived C are, on average, $\sim 9\%$ higher than the MODIS results, suggesting that MLS is able to detect some of the lower clouds with $p_{\mathrm{CT}} > 700$ hPa. This behavior is also illustrated in the example scenes in Figure 8–9 in section 4.4. In comparison, there is a much lower agreement between the MODIS and v4.2x results, which are on average $\sim 26\%$ lower.

This analysis indicates that the new ANN algorithm can produce considerably more reliable cloud classifications, on a global scale.

4.4 Example scenes

The analysis in the previous sections centered on statistical metrics and the reproduction of large-scale, global cloud patterns. There, the cloud flag based on the new ANN algorithm yields reliable results, both in comparison to the current v4.2x status flag and as a standalone product. However, a more qualitative assessment of the model performance for individual cloud scenes provides additional confidence in the technique, as well as insights into the classification performance for different cloud types. Again, profiles are flagged as cloudy when $P \geq 0.5$.

Figure 8 shows two example cloud fields over the North American monsoon region. During the summer months of July and August, this area is characterized by the regular occurrence of mesoscale convective systems that can occasionally overshoot into the lowermost stratosphere, where the sublimation of ice particles can lead to local humidity enhancements (Anderson

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021


Atmospheric Measurement Techniques Discussions

et al., 2012; Schwartz et al., 2013; Werner et al., 2020). Observed p_{CT} and Q_{T} derived from Aqua MODIS observations over the first example scene, sampled on 31 August 2017, are shown in Figures 8a and 8b, respectively. The MLS overpass is illustrated in gray transparent circles. A cloud system with $p_{\mathrm{CT}} < 500\,\mathrm{hPa}$ exists in the northern part of the scene, with the lowest $p_{\mathrm{CT}} \sim 200\,\mathrm{hPa}$. The MLS track passes some smaller cloud clusters characterized by large Q_{T}, which are indicated in yellow. In the south, low clouds with $Q_{\mathrm{T}} = 50 - 450\,\mathrm{g\ m^{-2}}$ are observed. The new ANN and current v4.2x cloud flags are shown in Figures 8c-d. The ANN algorithm flags every profile in the northern part of the scene as cloudy, while also detecting the very low clouds in the south. Conversely, the classifications from the current v4.2x flag identify a cloud influence for a single MLS profile in the north, which happens to actually be over an area with low Q_{T}. A second example cloud field is shown in Figures 8e-h. This scene consists of clouds all along the MLS track, and large areas with elevated Q_{T} up to $1,000\,\mathrm{g\ m^{-2}}$. Note that there is a gap in the MLS track, where the level 2 products are screened out, according to the rules in the MLS quality document (Livesey et al., 2020). The ANN algorithm correctly determines that every profile along the path was sampled under cloudy conditions. However, even for the very high clouds that contain large water abundances, the v4.2x algorithm only occasionally flags the respective profiles as cloudy. In the northern part of the track, the flag actually alternates between clear sky and cloudy classifications.

Similarly, Figure 9 shows two example cloud fields over the Asian summer monsoon region, which also regularly contains overshooting convection from mesoscale cloud systems. The first scene, shown in Figures 9a-d, displays a mix of different cloud conditions. There are high clouds with $p_{\mathrm{CT}} < 350\,\mathrm{hPa}$ and $Q_{\mathrm{T}} = 50 - 450\,\mathrm{g\ m^{-2}}$ in the northern part, a large clear sky area in the middle, and then a mix of very high and low-level clouds in the south that exhibits low Q_{T} and likely represents a multi-layer cloud structure with thin cirrus above boundary layer clouds. The new ANN-based flag successfully detects both the northern and southern cloud fields, while the current v4.2x flag only detects a single profile with cloud influence. The last example scene, illustrated in Figures 9e-h, displays the kind of clouds that are hardest to detect by means of MLS observations: very low clouds with $Q_{\mathrm{T}} < 150\,\mathrm{g\ m^{-2}}$. As expected, no MLS profile is considered to be influenced by clouds according to the current v4.2x status flag. However, the ANN algorithm detects those boundary layer clouds in the south of the scene, while correctly identifying the clear sky region along the rest of the track.

Note that the two example scenes in Figure 9 represent previously unseen data for the ANN, i.e., the models were not trained on these MLS observations.

5 Predicting cloud top pressure

The results in section 4 illustrate that the proposed ANN algorithm can successfully detect the subtle cloud signatures in the spectral T_{B} profiles shown in Figure 3. For many MLS bands, the differences between cloudy and clear sky T_{B} are usually in the range of just a few Kelvin, and the spectral behavior heavily depends on the respective MIF (i.e., pressure level at the tangent point of each scan). This section demonstrates how this behavior can be used in a similar ANN setup to infer the MODIS-retrieved p_{CT}. Here, our goal is to reliably differentiate between mid- to low-level clouds and high-reaching convection

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021


Atmospheric Measurement Techniques Discussions

with $p_{\mathrm{CT}} < 300$ hPa. As mentioned in the introduction, not only can these high clouds impact the MLS retrieval of atmospheric constituents, but they can also breach the tropopause and inject ice particles into the lowermost stratosphere.

Only slight changes to the ANN algorithm are required to predict p_{CT}, while the development and testing procedures remain identical to the steps described in section 3.4. The input layer and the two hidden layers remain unchanged from the cloud classification setup. The labels in the output layer, instead of being set to either "0" or "1" (i.e., clear sky or cloudy), now contain the respective p_{CT} reported by the colocated MLS-MODIS data set. A simple linear function replaces the "softmax" activation in the output layer, i.e., $\hat{L}^{\mathrm{j}} = \lambda^{\mathrm{j}}$ in Eq. (7). Similarly, the model optimizer, learning rate and mini-batch size reported in section 3.4 for the cloud classification ANN provide the best set of hyperparameters; here the only change concerns the weight decay parameter, which is turned off. As before, the model with the best validation loss provides the weights for the following evaluation.

Joint histograms of true (in the sense that they are the prescribed labels to train the ANN) and predicted p_{CT} for all cloudy profiles in the colocated MLS-MODIS data set are presented in Figure 10. Here, panels a and b show the results for the $63,242$ and $21,081$ samples in the training and validation data sets, respectively. High values of Pearson's product-moment correlation coefficient of $r > 0.99$ are observed for both data sets, and a majority of data are aligned along the 1:1 line (indicated by the yellow colors). The white dashed line illustrates the envelope defined by the 1^{st} and 99^{th} percentiles of predicted p_{CT} for each true p_{CT}-bin.

Similar to the discussion in section 4.4, comparisons between maps of predicted and MODIS-retrieved p_{CT} for individual cloud fields provide a qualitative assessment of the model performance. Three example scenes with the MODIS p_{CT} are shown in Figures 11a, c, and e, while the respective predictions from the ANN algorithm are shown in Figures 11b, d, and f. Each scene was sampled over the North American monsoon anticyclone, where large convective systems are prevalent over the continental United States during summer. The first example (panels a-b) is characterized by high clouds in the northern part with p_{CT} as low as ∼200 hPa, while at the southern tip there are low clouds with $p_{\mathrm{CT}} > 600$ hPa. The ANN can reliably reproduce the low cloud top pressures in the north; however, the p_{CT} values for the low clouds in the south are slightly underestimated, with predicted $p_{\mathrm{CT}} \approx 500 - 550$ hPa. The second example scene (panels c-d) consists of a mix of high and mid-level clouds with $p_{\mathrm{CT}} < 450$ hPa, which the ANN predictions correctly reproduce. The last example scene (panels e-f) shows a complicated mix of low, mid-level, and high clouds that basically covers the full p_{CT} range, as well as some clear sky areas in between. The ANN algorithm detects the small mid-level convection in the north, followed by very low clouds and the cloud gap over the center of the scene. It also correctly detects the subsequent large band of high clouds with $p_{\mathrm{CT}} < 350$ hPa over the southern continental United States.

Similar to the cloud detection algorithm, the prediction performance for p_{CT} appears to decline with an increase in cloud top pressure, consistent with the reduced contrast between clear sky and cloudy T_{B} around $p_{\mathrm{CT}} \sim 700$ hPa, as shown in Figure 3. However, the ANN can reliably distinguish between high-reaching convection with $p_{\mathrm{CT}} < 300$ hPa and mid- to low-level clouds.

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021


Atmospheric Measurement Techniques Discussions

6 Summary and conclusions

The current MLS cloud flags, reported in the Level 2 Geophysical Product of version 4.2x, are designed to identify profiles that are influenced by significantly opaque clouds, with the main goal being to identify cases where retrieved composition profiles may have been adversely affected either by the clouds or by the steps taken in the retrieval to exclude cloud-affected radiances. In this study, we present an improved cloud detection scheme based on the popular "Keras" Python library for setting up, testing, and validating feedforward artificial neural networks (ANNs). This new algorithm is shown to not only reliably detect high and mid-level convection containing even small amounts of cloud water, but also to distinguish between high-reaching and mid- to low-level convection.

To train the ANN models we colocated global MLS brightness temperatures (T_{B}), sampled on 208 days between 2005 and 2020, with nadir-viewing MODIS-retrieved cloud properties aggregated within a $1^{\circ} \times 1^{\circ}$ box (in latitude and longitude) around each MLS profile. This yielded a median cloud cover (C), cloud top pressure (p_{CT}), and cloud water path (Q_{T}) associated with each of the $162,117$ MLS scans in the colocated data set. These variables are used to discriminate clear sky ($C < 1/3$ and $Q_{\mathrm{T}} < 25\,\mathrm{g\ m^{-2}}$) from cloudy ($C \geq 2/3$, $100\,\mathrm{hPa} \leq p_{\mathrm{CT}} < 700\,\mathrm{hPa}$, and $Q_{\mathrm{T}} > 50\,\mathrm{g\ m^{-2}}$) profiles. Overall, the input variables for the ANN consist of $1,710$ MLS-observed T_{B} from different spectral bands, channels, and minor frames (i.e., views at different altitudes in the atmosphere). Comprehensive testing and cross-validation procedures are conducted to identify the right set of hyperparameters (i.e., model settings). The ideal model parameters are used to train 100 different ANN models, where the colocated data are randomly shuffled and split into 75% training and 25% validation data. Three binary classification metrics are calculated for every model run to evaluate the respective prediction performance for unseen data: the accuracy (Ac), F1 Score ($F1$), and Matthew's correlation coefficient (Mcc). Average values and standard deviations from the 100 different model runs are $Ac = 0.971 \pm 0.001$, $F1 = 0.971 \pm 0.002$, and $Mcc = 0.942 \pm 0.003$, and, on average, the models correctly classify $\sim 97.0\%$ of the clear sky and cloudy profiles. The high statistical scores and low variability in the results illustrate that the ANN algorithm yields reliable cloud classifications for previously unseen MLS observations.

A comparison with the current v4.2x status flags reveals that for the complete data set in this study the new ANN results provide a significant improvement in cloud classification. The ANN algorithm correctly identifies 97.8% of cloudy profiles, while only 1.7% of the clear profiles are falsely flagged. No significant dependence on geolocation is observed, indicating that the ANN flag yields reliable classification results on a global scale. In contrast, the current v4.2x flag detects only 15.8% of cloudy profiles, and even though it is designed to identify sufficiently opaque clouds, it only correctly classifies 29.0% of cloudy profiles with $Q_{\mathrm{T}} > 1,000\,\mathrm{g\ m^{-2}}$. The fraction of falsely flagged clear profiles (6.2%) is also higher compared to the ANN results. A global cloud cover map for data collected between 2005 and 2019 is presented, generated solely from MLS-sampled T_{B} and the determined ANN weights. Typically observed cloud patterns and reported cloud fractions are reproduced by the ANN algorithm. Moreover, detailed examination of four examples scenes from the North American and Asian summer monsoon regions reveals that the ANN can reliably identify diverse cloud fields, including those characterized by low-level clouds and low Q_{T}. Together with the consistently large statistical agreement, these global and regional examples of successful cloud detection illustrate that the predefined cloudiness conditions (following thresholds for C, p_{CT}, and Q_{T}) are reasonable.

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021


Atmospheric Measurement Techniques Discussions

Moreover, the uncertainties arising from associating MLS observations in the limb with nadir MODIS images do not seem to substantially impact the reliability of the ANN algorithm.

This study demonstrates that the ANN algorithm can not only detect cloud influences for individual MLS profiles, but also that it can reliably predict MODIS-retrieved p_{CT}. This is illustrated by high correlation coefficients of > 0.99 and objectively good model performance for three example cloud fields of varying degrees of complexity.

This new cloud classification scheme, which will be included in future versions of the MLS v4.2x, provides the means to reliably identify profiles with potential cloud influence. As mention in the introduction, this new algorithm will facilitate future research on reducing uncertainties in the retrieval of atmospheric constituents in the presence of clouds. Moreover, studies on convective moistening of the lowermost stratosphere, as well cloud scavenging of atmospheric pollutants, will benefit from these new capabilities.

Data availability. MLS brightness temperatures and L2GP data, including status flags, are available at https://mls.jpl.nasa.gov. Aqua-MODIS data are obtained from the LAADS-DAAC at https://ladsweb.modaps.eosdis.nasa.gov/search/order/1/MODIS:Aqua

Author contributions. All authors have shaped the concept of this study and refined the approach during extensive discussions. FW carried out the data analysis and wrote the initial draft of the manuscript, which was subsequently refined by all authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. ©2020. California Institute of Technology. Government sponsorship acknowledged. The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004).

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021


Atmospheric Measurement Techniques Discussions

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, arXiv [preprint], arxiv:1603.04467v2, 2016.

Ackerman, S. A., Strabala, K. I., Menzel, W. P., Frey, R. A., Moeller, C. C., and Gumley, L. E.: Discriminating clear sky from clouds with MODIS, J. Geophys. Res., 103, 32 141–32 157, 1998.

Ackerman, S. A., Holz, R. E., Frey, R., Eloranta, E. W., Maddux, B. C., and McGill, M.: Cloud Detection with MODIS. Part II: Validation, Journal of Atmospheric and Oceanic Technology, 25, 1073–1086, https://doi.org/http://dx.doi.org/10.1175/2007JTECHA1053.1, 2008.

Anderson, J. G., Wilmouth, D. M., Smith, J. B., and Sayres, D. S.: UV Dosage Levels in Summer: Increased Risk of Ozone Loss from Convectively Injected Water Vapor, Science, 337, 835–839, https://doi.org/10.1126/science.1222978, 2012.

Ardanuy, P. A., Han, D., and Salomonson, V. V.: The Moderate Resolution Imaging Spectrometer (MODIS), IEEE Trans. Geosci. Remote Sensing, 30, 2–27, 1992.

Arking, A.: Latitudinal distribution of cloud cover from TIROS III photographs, Science, 143, 569–572, 1964.

Arking, A. and Childs, J. D.: Retrieval of cloud cover parameters from multispectral satellite images, J. Climate Appl. Meteor., 24, 322–333, 1985.

Barnes, W. L., Pagano, T. S., and Salomonson, V. V.: Prelaunch characteristics of the 'Moderate Resolution Imaging Spectroradiometer' (MODIS) on EOS–AM1, IEEE Trans. Geosci. Remote Sensing, 36, 1088–1100, 1998.

Brenguier, J.-L., Pawlowska, H., Schüller, L., Preusker, R., Fischer, J., and Fouquart, Y.: Radiative properties of boundary layer clouds: Droplet effective radius versus number concentration, J. Atmos. Sci., 57, 803–821, 2000.

Chollet, F. et al.: Keras, https://keras.io, 2015.

Clapp, P. F.: Global cloud cover for seasons using TIROS nephanalyses, Mon. Wea. Rev., 92, 495–507, 1964.

COESA: U.S. Standard Atmosphere, 1976, Technical report, U.S. Government Printing Office, United States, 1976.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning (Adaptive Computation and Machine Learning series), The MIT Press, Cambridge, MA, 2016.

Han, Q., Rossow, W., and Lacis, A.: Near-global survey of effective droplet radii in liquid water clouds using ISCCP data, J. Climate, 7, 465–497, 1994.

Jeppesen, J. H., H., J. R., Inceoglu, F., and Toftegaard, T. S.: A cloud detection algorithm for satellite imagery based on deep learning, Remote Sens. Environ., 229, 247–259, https://doi.org/10.1016/j.rse.2019.03.039, 2019.

Kaufman, Y. J. and Nakajima, T.: Effect of Amazon smoke on cloud microphysics and albedo – Analysis from satellite imagery, J. Appl. Meteor., 32, 729–744, 1993.

King, M. D., Platnick, S., Wenzel, W. P., Ackerman, S. A., and Hubanks, P. A.: Spatial and Temporal Distribution of Clouds Observed by MODIS Onboard the Terra and Aqua Satellites, IEEE Trans. Geosci. Remote Sens., 51, 3826–3852, https://doi.org/10.1109/TGRS.2012.2227333, 2013.

Lacagnina, C. and Selten, F.: Evaluation of clouds and radiative fluxes in the EC-Earth general circulation model, Climate Dynamics, 43, 2777–2796, https://doi.org/10.1007/s00382-014-2093-9, 2014.

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D.: Backpropagation Applied to Handwritten Zip Code Recognition, Neural Computation, 1, 541–551, 1989.

LeCun, Y., Bottou, L., Orr, G., and Müller, K. R.: Neural Networks: Tricks of the trade, Efficient BackProp, Springer, 1998.

Livesey, N. J., Read, W. G., Wagner, P. A., Froidevaux, L., Lambert, A., Manney, G. L., Valle, L. F. M., Pumphrey, H. C., Santee, M. L., Schwartz, M. J., Wang, S., Fuller, R. A., Jarnot, R. F., Knosp, B. W., Martinez, E., and Lay, R. R.: Version 4.2x Level 2 and 3 data quality and description document., Tech. Rep. JPL D-33509 Rev. E, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, 91109-8099, 2020.

Menzel, W. P., Frey, R. A., Zhang, H., Wylie, D. P., Moeller, C. C., Holz, R. E., Maddux, B., Baum, B. A., Strabala, K. I., and Gumley, L. E.: MODIS global cloud-top pressure and amount estimation: Algorithm description and results, J. Appl. Meteor. Clim., 47, 1175–1198, 2008.

Miller, D. J., Zhang, Z., Ackerman, A. S., Platnick, S., and Baum, B. A.: The impact of cloud vertical profile on liquid water path retrieval based on the bispectral method: A theoretical study based on large-eddy simulations of shallow marine boundary layer clouds, J. Geophys. Res., 121, 4122–4141, https://doi.org/10.1002/2015JD024322, 2016.

Minnis, P., Heck, P., Young, D., Fairall, C., and Snider, J.: Stratocumulus cloud properties derived from simultaneous satellite and island-based instrumentation during FIRE, J. Appl. Meteorol., 31, 317–339, 1992.

Platnick, S. and Twomey, S.: Determining the susceptibility of cloud albedo to changes in droplet concentration with the Advanced Very High Resolution Radiometer, J. Appl. Meteorol., 33, 334–347, 1994.

Platnick, S., Meyer, K. G., King, M. D., Wind, G., Amarasinghe, N., Marchant, B., Arnold, G. T., Zhang, Z., Hubanks, P. A., Holz, R. E., Yang, P., Ridgway, W. L., and Riedi, J.: The MODIS Cloud Optical and Microphysical Products: Collection 6 Updates and Examples From Terra and Aqua, IEEE Transactions on Geoscience and Remote Sensing, 55, 502–525, https://doi.org/10.1109/TGRS.2016.2610522, 2017.

Ramanathan, V., CESS, R. D., Harrison, E. F., Minnis, P., Barkstrom, B. R., Ahmad, E., and Hartmann, D. L.: Cloud–Radiative Forcing and Climate: Results from the Earth Radiation Budget Experiment, Science, 243, 57–63, 1989.

Ratté-Fortin, C., Chokmani, K., and El-Alem, A.: A novel algorithm of cloud detection for water quality studies using 250 m downscaled MODIS imagery, International Journal of Remote Sensing, 39, 6429–6439, https://doi.org/10.1080/01431161.2018.1460506, 2018.

Reed, R. and Marksll, R. J.: Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks, A Bradford Book, 1999.

Rossow, W. B., Kinsella, E., and Garder, L.: Seasonal and global cloud variations deduced from polar orbiting satellite radiance measurements, Proc. Fifth Conf. on Atmospheric Radiation, Baltimore, Amer. Meteor. Soc., pp. 195–198, 1983.

Saponaro, G., Kolmonen, P., Karhunen, J., Tamminen, J., and de Leeuw, G.: A neural network algorithm for cloud fraction estimation using NASA-Aura OMI VIS radiance measurements, Atmospheric Measurement Techniques, 6, 2301–2309, https://doi.org/10.5194/amt-6-2301-2013, 2013.

Schwartz, M. J., Read, W. G., Santee, M. L., Livesey, N. J., Froidevaux, L., Lambert, A., and Manney, G. L.: Convectively injected water vapor in the North American summer lowermost stratosphere, Geophysical Research Letters, 40, 2316–2321, https://doi.org/10.1002/grl.50421, 2013.

Stephens, G.: Cloud feedbacks in the climate system: A critical review, J. Climate, 18, 237–273, 2005.

Sun, L., Yang, X., Jia, S., Jia, C., Wang, Q., Liu, X., Wei, J., and Zhou, X.: Satellite data cloud detection using deep learning supported by hyperspectral data, International Journal of Remote Sensing, 41, 1349–1371, https://doi.org/10.1080/01431161.2019.1667548, 2020.

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021


Atmospheric Measurement Techniques Discussions

Tinney, E. N. and Homeyer, C. R.: A 13-year Trajectory-Based Analysis of Convection-Driven Changes in Upper Troposphere Lower Stratosphere Composition over the United States, Journal of Geophysical Research: Atmospheres, n/a, e2020JD033 657, https://doi.org/https://doi.org/10.1029/2020JD033657, 2020.

Varnai, T. and Marshak, A.: Satellite Observations of Cloud-Related Variations in Aerosol Properties, Atmosphere, 9, 430, https://doi.org/10.3390/atmos9110430, 2018.

Wang, T., Shi, J., Letu, H., Ma, Y., Li, X., and Zheng, Y.: Detection and Removal of Clouds and Associated Shadows in Satellite Imagery Based on Simulated Radiance Fields, Journal of Geophysical Research: Atmospheres, 124, 7207–7225, https://doi.org/10.1029/2018JD029960, 2019.

Warren, S., Hahn, C., London, J., Chervin, R., and Jenne, R.: Global distribution of total cloud cover and cloud type amounts over the ocean, Tech. rep., Boulder, CO, 1988.

Waters, J. W., Froidevaux, L., Harwood, R. S., Jarnot, R. F., Pickett, H. M., Read, W. G., Siegel, P. H., Cofield, R. E., Filipiak, M. J., Flower, D. A., Holden, J. R., Lau, G. K., Livesey, N. J., Manney, G. L., Pumphrey, H. C., Santee, M. L., Wu, D. L., Cuddy, D. T., Lay, R. R., Loo, M. S., Perun, V. S., Schwartz, M. J., Stek, P. C., Thurstans, R. P., Boyles, M. A., Chandra, K. M., Chavez, M. C., Gun-Shing Chen, Chudasama, B. V., Dodge, R., Fuller, R. A., Girard, M. A., Jiang, J. H., Yibo Jiang, Knosp, B. W., LaBelle, R. C., Lam, J. C., Lee, K. A., Miller, D., Oswald, J. E., Patel, N. C., Pukala, D. M., Quintero, O., Scaff, D. M., Van Snyder, W., Tope, M. C., Wagner, P. A., and Walch, M. J.: The Earth observing system microwave limb sounder (EOS MLS) on the aura Satellite, IEEE Transactions on Geoscience and Remote Sensing, 44, 1075–1092, https://doi.org/10.1109/TGRS.2006.873771, 2006.

Werner, F., Wind, G., Zhang, Z., Platnick, S., Di Girolamo, L., Zhao, G., Amarasinghe, N., and Meyer, K.: Marine boundary layer cloud property retrievals from high-resolution ASTER observations: case studies and comparison with Terra MODIS, Atmos. Meas. Tech, 9, 5869–5894, https://doi.org/10.5194/amt-9-5869-2016, http://www.atmos-meas-tech.net/9/5869/2016/, 2016.

Werner, F., Schwartz, M. J., Livesey, N. J., Read, W. G., and Santee, M. L.: Extreme Outliers in Lower Stratospheric Water Vapor Over North America Observed by MLS: Relation to Overshooting Convection Diagnosed From Colocated Aqua-MODIS Data, Geophysical Research Letters, 47, e2020GL090 131, https://doi.org/https://doi.org/10.1029/2020GL090131, 2020.

Wu, D. L., Jiang, J. H., and Davis, C. P.: EOS MLS cloud ice measurements and cloudy-sky radiative transfer model, IEEE Transactions on Geoscience and Remote Sensing, 44, 1156–1165, https://doi.org/10.1109/TGRS.2006.869994, 2006.

Wu, D. L., Jiang, J. H., Read, W. G., Austin, R. T., Davis, C. P., Lambert, A., Stephens, G. L., Vane, D. G., and Waters, J. W.: Validation of the Aura MLS cloud ice water content measurements, Journal of Geophysical Research: Atmospheres, 113, https://doi.org/10.1029/2007JD008931, 2008.

Yang, Y. and Di Girolamo, L.: Impacts of 3-D radiative effects on satellite cloud detection and their consequences on cloud fraction and aerosol optical depth retrievals, J. Geophys. Res., 113, 2008.

Yu, W., Dessler, A. E., Park, M., and Jensen, E. J.: Influence of convection on stratospheric water vapor in the North American monsoon region, Atmospheric Chemistry and Physics, 20, 12 153–12 161, https://doi.org/10.5194/acp-20-12153-2020, 2020.

Zhao, G. and Di Girolamo, L.: Statistics on the macrophysical properties of trade wind cumuli over the tropical western Atlantic, J. Geophys. Res., 112, 2007.

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021

Atmospheric Measurement Techniques Discussions

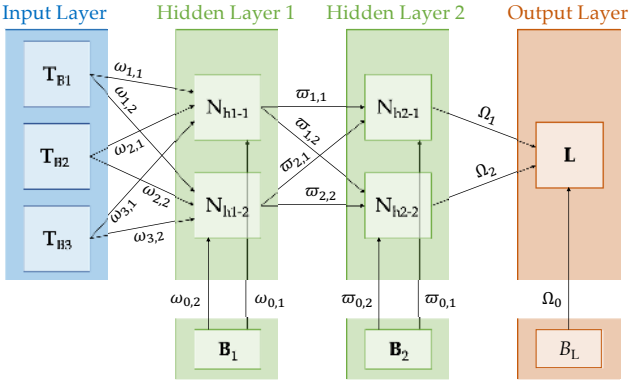


Figure 1. Simplified sketch of the algorithm setup, including three vectors in the input layer (blue) that contain MLS brightness temperatures ($\mathbf{T}_i$; i =1–3), two hidden layers (green) with two neurons ($N_{\mathrm{h1}-k}$ and $N_{\mathrm{h2}-k}$; k =1–2) and one "bias" node each ($\mathbf{B}_k$; k =1–2), and output layer (orange) with the labels vector ($\mathbf{L}$) and "bias" node ($\mathbf{B}_\mathrm{L}$). Also shown are the input weights ($\omega_{\mathrm{i,k}}$; i =0–3, k =1–2), connecting weights ($\varpi_{\mathrm{k,l}}$; k =0–2, l =1–2), and output weights (Ω_l; l =0–2) that connect the input variables to the neurons in the first hidden layer, the neurons from the two hidden layers, and the neurons from the second hidden layer to the labels vector, respectively.

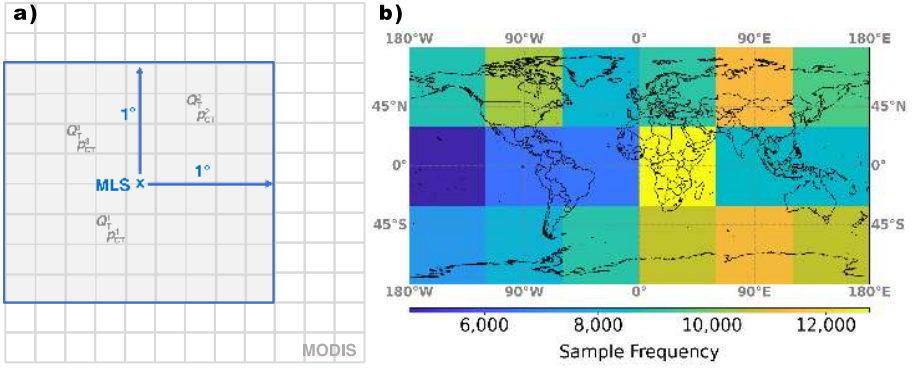
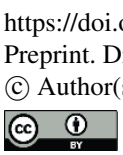


Figure 2. (a) Illustration of the colocation of MLS and MODIS data. (b) Global map of sample frequencies for the colocated MLS-MODIS data set used in the training of the ANN.

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021



Atmospheric Measurement Techniques Discussions

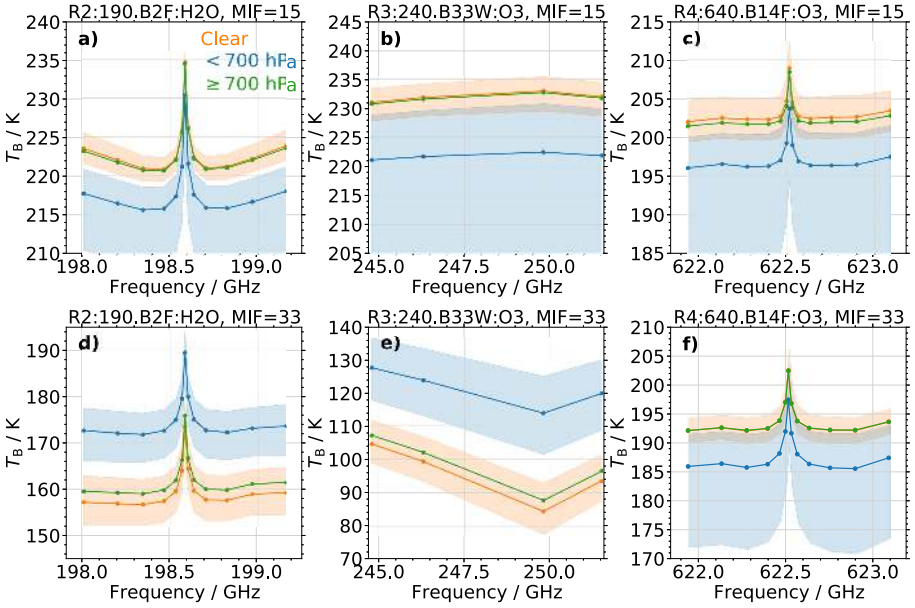


Figure 3. (a) Statistic of the brightness temperature (T_B) from MLS observations sampled in band 2 of receiver 2 at minor frame (MIF) 15 (at an altitude of $\approx 4.5\,\mathrm{km}$) in the latitudinal range of $-30°$ to $+30°$ as a function of frequency. The orange, blue, and green curves show the median T_B associated with clear sky conditions, clouds with a cloud top pressure $p_{CT} < 700\,\mathrm{hPa}$, and clouds with $p_{CT} \geq 700\,\mathrm{hPa}$, respectively. The shaded orange and blue areas indicate the interquartile range of the respective T_B (omitted for low clouds to enhance legibility). Samples are provided by the colocated MLS-MODIS data set. (b) Same as (a), but for band 33 of radiometer 3. (c) Same as (a), but for band 14 of radiometer 4. (d)–(f) Same as (a)–(c), but at MIF=33 (at an altitude of $\approx 12\,\mathrm{km}$).

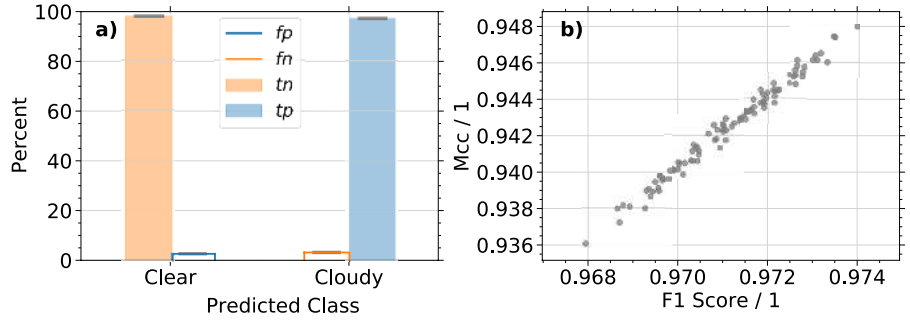


Figure 4. (a) Histograms of classifications from the ANN algorithm for 100 random combinations of training and validation data sets. Orange and blue shading depicts the percent of correctly predicted clear (i.e., true negatives, tn) and cloudy (i.e., true positives, tp) labels for actually observed clear and cloudy profiles, respectively. Blue and orange lines depict the percent of falsely predicted cloudy (i.e., false positives, fp) and clear (i.e., false negatives, tn) labels for actually observed clear and cloudy profiles, respectively. The vertical extent of the gray horizontal bars on top of each histogram indicates the standard deviation derived from all 100 predictions (the horizontal extent is arbitrary). (b) Scatter plot of Matthews correlation coefficient (Mcc) and F1 score for the same 100 random combinations of training and validation data sets shown in (a).

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021

Atmospheric Measurement Techniques Discussions

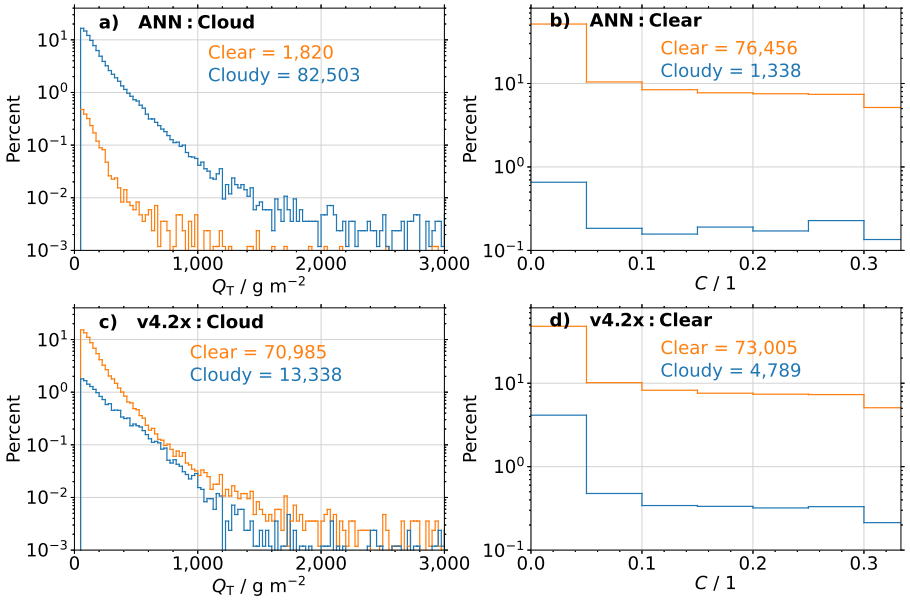


Figure 5. (a) Histograms of classifications from the new ANN-based cloud flag for actually observed cloudy profiles as a function of total water path (Q_T). Orange and blue colors depict the distributions of predicted clear and cloudy labels, respectively. The number of clear and cloudy predictions is also given. (b) Similar to (a), but for actually observed clear profiles as a function of cloud cover (C). (c)-(d) Same as (a)-(b), but for classifications from the current v4.2x cloud flag.

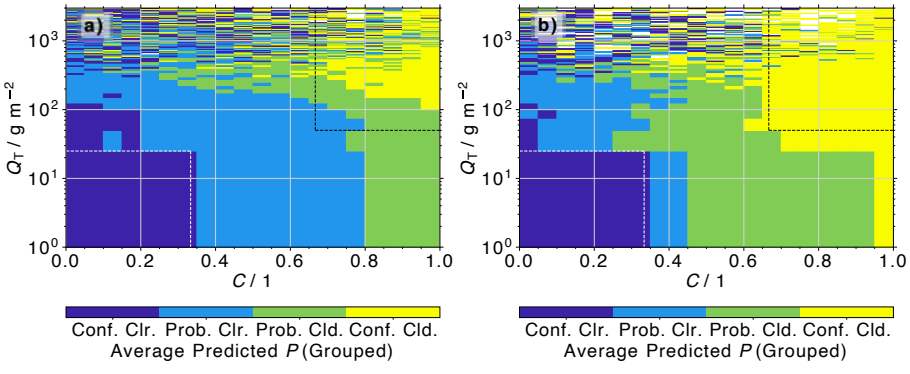


Figure 6. (a) Average probability of cloudiness (P) predicted by the ANN as a function of C and Q_T. P are grouped into four classes: confidently clear ("Conf. Clr."; $P < 0.25$), probably clear ("Prob. Clr."; $0.25 \leq P < 0.5$), probably cloudy ("Prob. Cld."; $0.5 \leq P < 0.75$), and confidently cloudy ("Conf. Cld."; $P \geq 0.75$). No restrictions on cloud top pressure (p_{CT}) are imposed. (b) Same as (a), but for profiles with $p_{CT} < 700$ hPa.

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021

Atmospheric Measurement Techniques Discussions

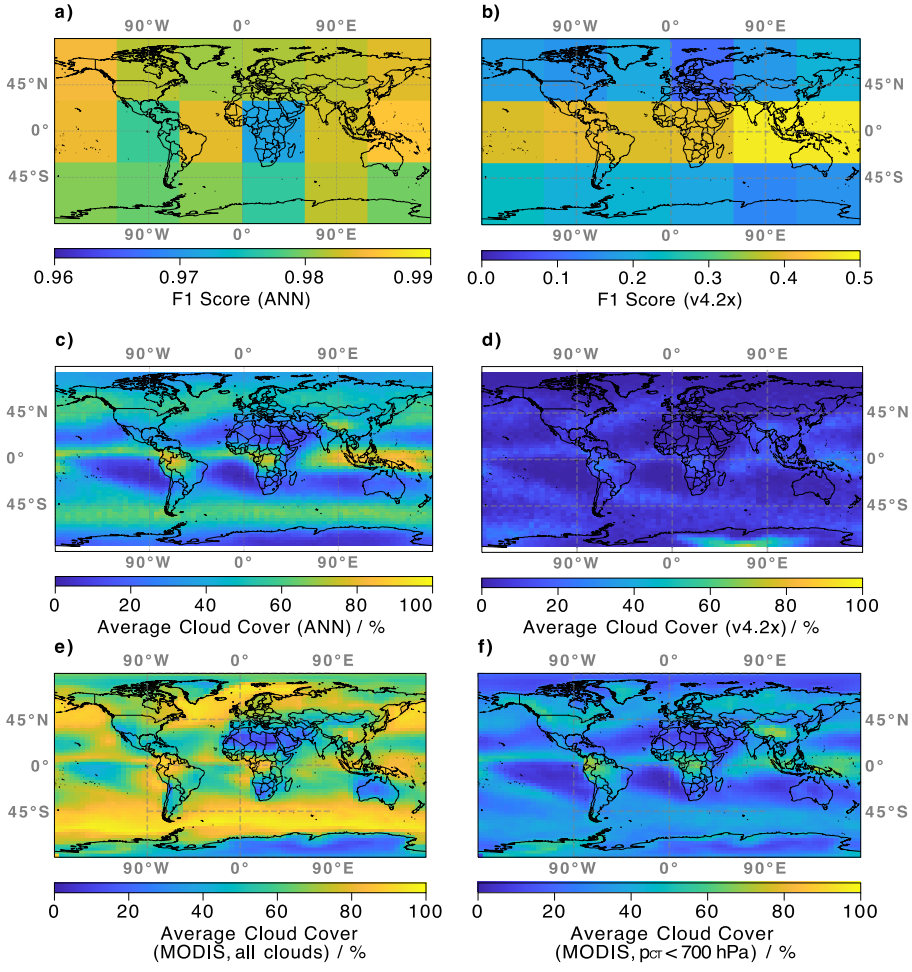


Figure 7. (a) Latitudinal and longitudinal dependence of the performance of the ANN algorithm, determined by the F1 score for binary classifications. Observations and actual cloudiness flags are provided by the colocated MLS-MODIS data set. (b) Same as (a), but for the current v4.2x cloud flag. (c) Average global cloud cover derived from MLS brightness temperature observations and the weights determined from the trained ANN. All MLS observations sampled between 2015 and 2019 are represented. (d) Same as (c), but for the current v4.2x cloud flag. (e) Same as (c), but from Aqua-Modis observations sampled in 2019. (f) Same as (e), but with retrieved cloud top pressure < 700 hPa.

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021

Atmospheric Measurement Techniques Discussions

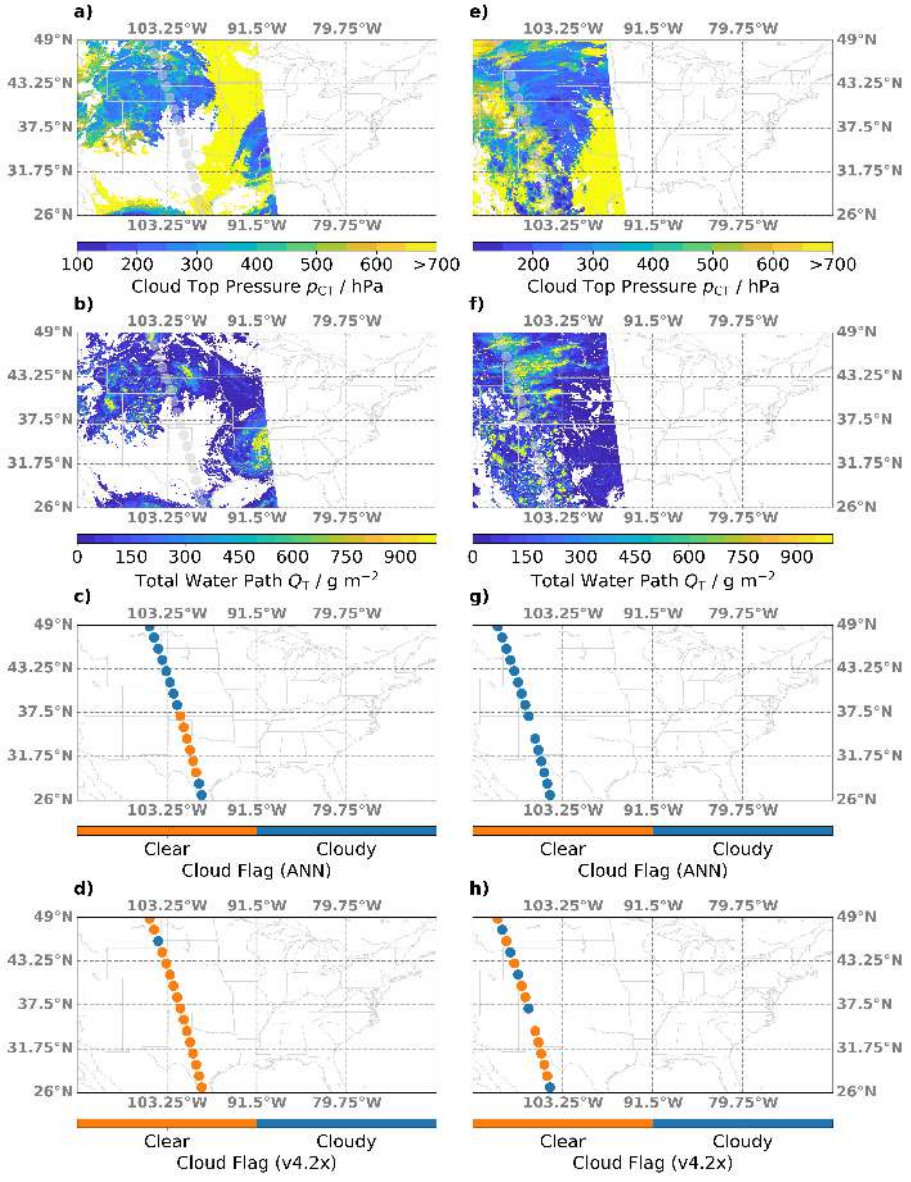


Figure 8. (a) Map of cloud top pressure (p_{CT}) retrieved from MODIS observations on 31 August 2017 over North America. Transparent circles indicate the MLS orbit. (b) Same as (a), but for the total water path (Q_t). (c) Clear (orange) and cloudy (blue) profiles as determined from the new ANN algorithm. (d) Same as (c), but determined from the current v4.2x status flags. (e)-(h) Same as (a)-(d), but for MLS and MODIS observations on 5 July 2015.

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021

Atmospheric Measurement Techniques Discussions

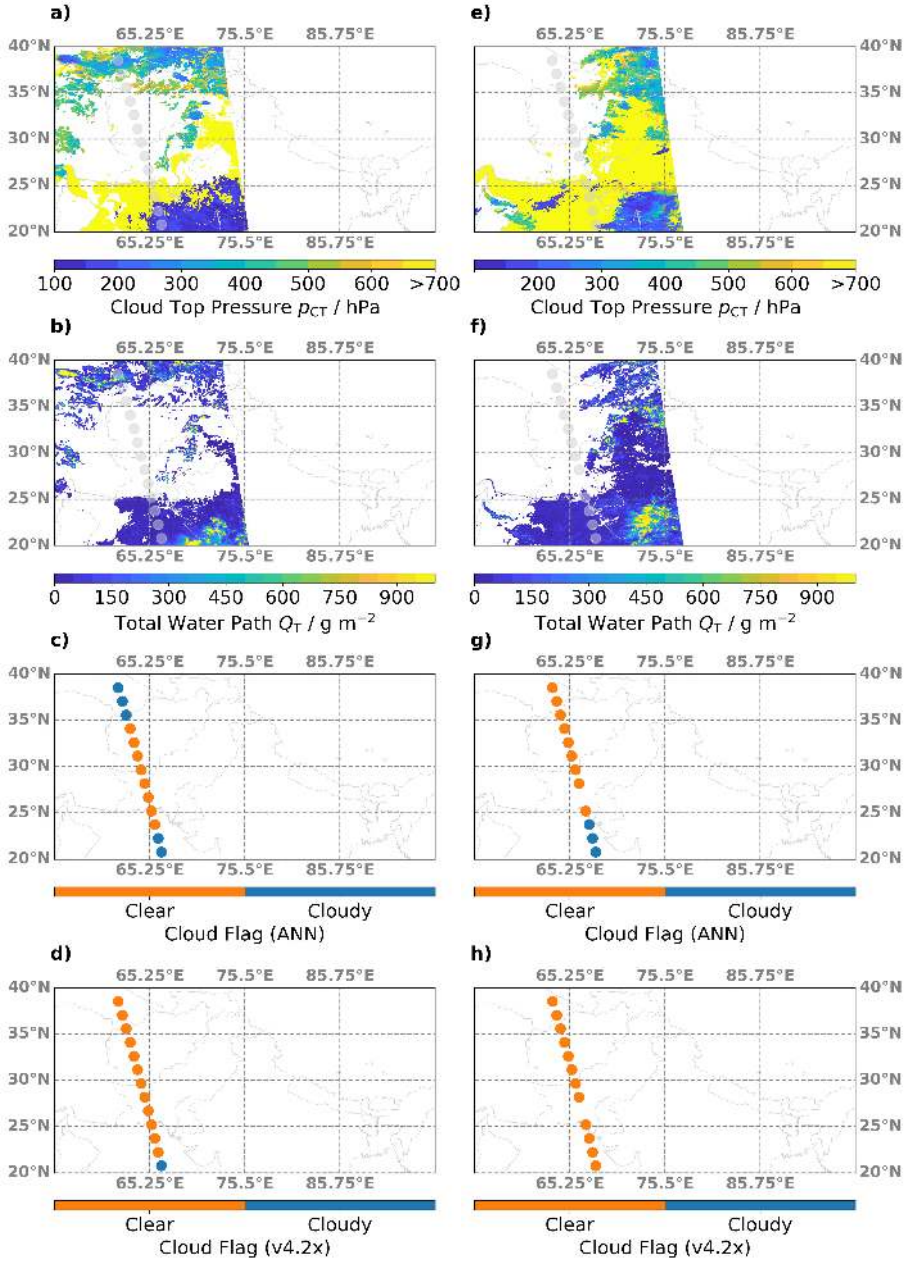


Figure 9. Similar to Figure 8, but for MLS and MODIS observations on (a)-(d) 28 June 2019 and (e)-(h) 4 July 2018, respectively, over South Asia. These scenes were captured over the Asian summer monsoon region.

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021

Atmospheric Measurement Techniques Discussions

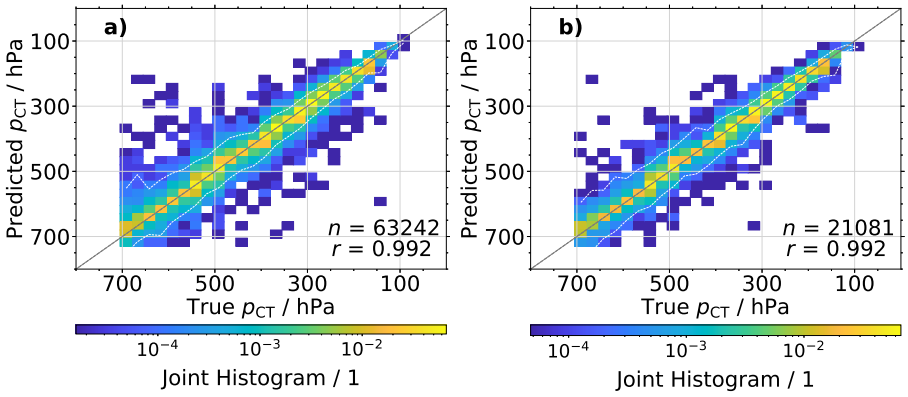
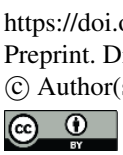


Figure 10. (a) Normalized joint histograms of true and predicted cloud top pressure (p_{CT}). Data are from the training data set. (b) Same as (a), but for the validation data set.

Table 1. Details about the colocated MLS-MODIS data set, which contains observations from thirteen random Julian days (d01-d13) for each year over 2005–2020.

	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
d01	001	005	006	018	031	021	010	004	014	002	015	010	021	008	010	008
d02	034	041	041	055	055	041	040	036	037	052	016	031	053	043	041	048
d03	064	045	066	064	056	064	076	061	079	065	033	044	069	075	083	087
d04	117	066	092	068	079	092	077	101	113	101	064	076	106	109	109	181
d05	121	098	134	076	092	126	112	134	121	139	116	110	107	124	139	192
d06	122	099	173	119	145	164	142	169	154	167	124	140	143	163	181	196
d07	158	140	197	122	164	201	177	187	206	202	159	172	169	183	189	214
d08	206	161	230	158	212	240	185	224	238	238	186	212	205	223	239	215
d09	223	194	265	197	221	264	224	225	273	255	216	213	243	269	256	247
d10	244	242	302	198	251	275	261	246	274	256	248	233	270	280	257	251
d11	286	259	322	224	286	276	293	300	313	290	284	266	283	281	286	258
d12	319	301	323	292	314	308	317	332	337	330	309	299	332	324	308	284
d13	349	323	360	328	353	356	359	350	338	354	342	323	358	358	341	310

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021



Atmospheric Measurement Techniques Discussions

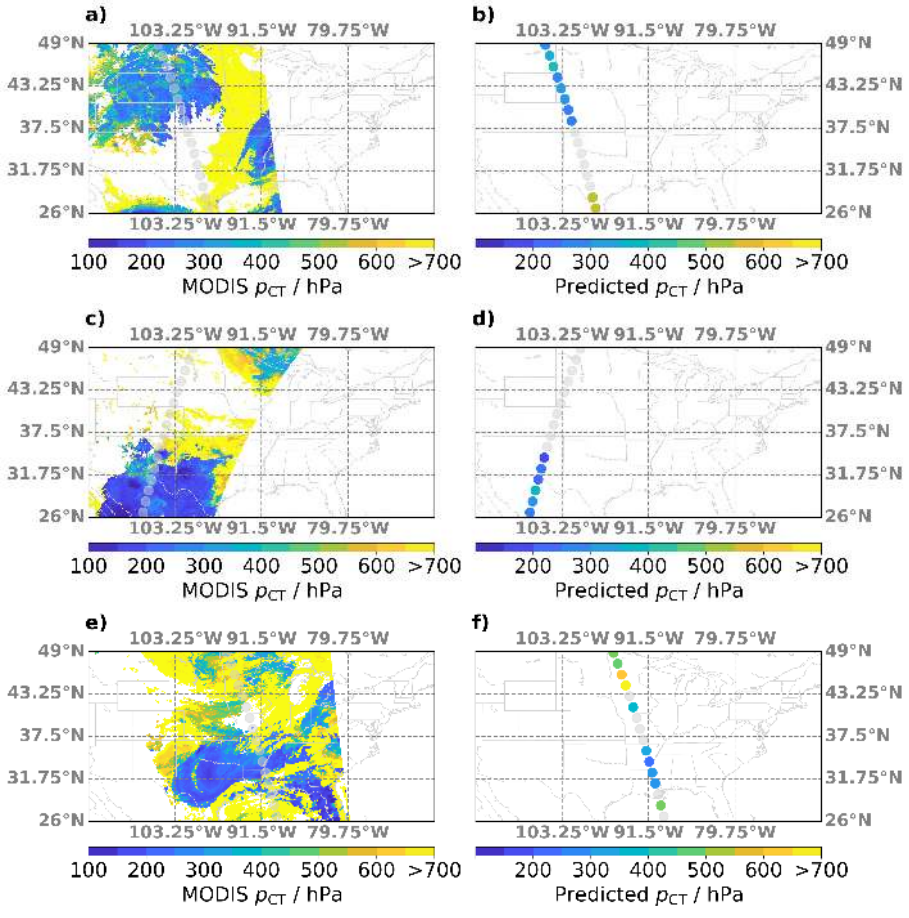


Figure 11. (a) Map of cloud top pressure (p_{CT}) retrieved from MODIS observations on 30 August 2017 over North America. Transparent circles indicate the MLS orbit. (b) Same as (a), but for the predicted p_{CT} based on the ANN algorithm. (c)-(d) Same as (a)-(b), but for MLS and MODIS observations on 28 August 2019. (e)-(f) Same as (a)-(b), but for MLS and MODIS observations on 27 August 2019.

https://doi.org/10.5194/amt-2021-146
Preprint. Discussion started: 25 May 2021

Atmospheric Measurement Techniques Discussions

Table 2. Details of the input variables for the ANN algorithm, which consist of MLS brightness temperature observations in 10 different bands from 4 radiometers. Besides the official radiometer and band designations, the local oscillator (LO) and primary species of interest in the respective band are given, as well as the ranges of minor frames (MIFs) and channels used as input for the ANN.

Spectrometer	Band	LO (GHz)	Species	MIF	Channel
R1A	B1F	118	p_{tan}	[7, 10, 13, …, 49]	[1, 3, 5 , …, 25]
R2	B2F	190	H_2O	[7, 10, 13, …, 49]	[1, 3, 5 , …, 25]
R2	B3F	190	N_2O	[7, 10, 13, …, 49]	[1, 3, 5 , …, 25]
R2	B6F	190	O_3	[7, 10, 13, …, 49]	[1, 3, 5 , …, 25]
R3	B7F	240	O_3	[7, 10, 13, …, 49]	[1, 3, 5 , …, 25]
R3	B8F	240	p_{tan}	[7, 10, 13, …, 49]	[1, 3, 5 , …, 25]
R3	B33W	240	O_3	[7, 10, 13, …, 49]	[1, 2, 3, 4]
R4	B10F	640	ClO	[7, 10, 13, …, 49]	[1, 3, 5 , …, 25]
R4	B14F	640	O_3	[7, 10, 13, …, 49]	[1, 3, 5 , …, 25]
R4	B28M	640	HO_2	[7, 10, 13, …, 49]	[1, 3, 5 , …, 11]

Table 3. Binary classification statistics for the new ANN algorithm, as well as the classification provided by the current MLS v4.2x status flag. Prescribed labels (i.e., clear sky or cloudy) are provided by the standard definitions presented in section 3.2, as well as a redefined classification based on looser thresholds. The fraction of true positives and negatives (tp and tn), as well as false positives and negatives (fp and fn), are given. Finally, three measures for the evaluation of binary statistics are listed: the accuracy (Ac), the F1 score ($F1$), and the Matthews correlation coefficient (Mcc).

	tp	tn	fp	fn	Ac	$F1$	Mcc
ANN	0.98	0.98	0.02	0.02	0.98	0.98	0.96
v4.2x	0.16	0.94	0.06	0.84	0.53	0.26	0.15
ANN (redefined)	0.69	0.96	0.04	0.31	0.77	0.81	0.60
v4.2x (redefined)	0.08	0.94	0.06	0.92	0.34	0.14	0.04