1    **A Machine Learning-Based Cloud Detection and Thermodynamic**

2    **Phase Classification Algorithm using Passive Spectral Observations**

3    **Chenxi Wang[1,2], Steven Platnick[2], Kerry Meyer[2], Zhibo Zhang[3], Yaping Zhou[1,2]**

4

5    [1]Joint Center for Earth Systems Technology, University of Maryland Baltimore County,

6    Baltimore, MD, USA

7    [2]NASA Goddard Space Flight Center, Greenbelt, MD, USA.

8    [3]Department of Physics, University of Maryland Baltimore County, Baltimore, MD, USA.

9

10    **Abstract**

11    We trained two Random Forest (RF) machine-learning models for cloud mask and cloud

12    thermodynamic phase detection using spectral observations from VIIRS on Suomi-NPP (SNPP).

13    Observations from CALIOP were carefully selected to provide reference labels. The two RF

14    models were trained for all-day and daytime-only conditions using a 4-year collocated

15    VIIRS/CALIOP dataset from 2013 to 2016. Due to the orbit difference, the collocated CALIOP

16    and SNPP VIIRS training samples cover a broad viewing zenith angle range, which is a great

17    benefit to overall model performance. The all-day model uses 3 VIIRS infrared (IR) bands (8.6,

18    11, and 12 $\mu$m) and the daytime model uses 5 Near-IR (NIR) and Shortwave-IR (SWIR) bands

19    (0.86, 1.24, 1.38, 1.64 and 2.25 $\mu$m) together with the 3 IR bands to detect clear, liquid water, and

20    ice cloud pixels. Up to 7 surface types, namely, ocean/water, forest, cropland, grassland, snow/ice,

21    barren/desert, and shrubland, were considered separately to enhance performance for both models.

22    Detection of cloudy pixels and thermodynamic phase with the two RF models were compared

23    against collocated CALIOP products from 2017. It is shown that the two RF models have high

24    accuracy rates in comparison with the CALIOP reference for both cloud detection and

25    thermodynamic phase. For cloud detection, the accuracy rates of the daytime RF model are higher

26    than 92% for all surface types, while the accuracy rates of the all-day RF model decrease by 3~8%,

27    depending on surface type. For cloud thermodynamic phase, both RF models agree well with

28    CALIOP, except over barren/desert regions. Other existing SNPP VIIRS and Aqua MODIS cloud

29    mask and phase products are also evaluated, with results showing that the two RF models and the

30    MODIS MYD06 optical property phase product are the top 3 algorithms with respect to lidar

31    observations during the daytime. During the nighttime, the RF all-day model works best for both

32    cloud detection and phase, in particular for pixels over snow/ice surfaces. The present RF models

33    can be extended to other similar passive instruments if training samples can be collected from

34      CALIOP or other lidars. However, the quality of reference labels and potential sampling issues

35      that may impact model performance would need further attention.

36      **1. Introduction**

37      Detection and classification (DC) of atmospheric constituents using satellite observations is

38      often a critical initial step in many remote sensing algorithms. For example, a prerequisite for cloud

39      optical and microphysical property retrievals is identifying the presence of clouds, i.e., a

40      clear/cloudy classification [*Frey et al.*, 2008]. Additionally, characteristics such as cloud

41      thermodynamic phase are needed as they can strongly impact the scattering/absorption properties

42      of cloud droplets/particles [*Platnick et al.*, 2017]. Similarly, current operational aerosol algorithms

43      can only retrieve aerosol optical depth (AOD) for "non-cloudy" pixels since even slight cloud

44      contamination can result in erroneously high retrieved AOD [*Remer et al.*, 2005]. Therefore, errors

45      in detecting and classifying atmospheric components can significantly impact downstream

46      retrieval products and scientific analyses.

47      There are many examples of traditional DC algorithms designed for satellite instruments. For

48      example, the Moderate Resolution Imaging Spectroradiometer (MODIS) has algorithms

49      developed for cloud masking [*Frey et al.*, 2008; *Ackerman et al.*, 2008], cloud thermodynamic

50      phase [*Baum et al.*, 2012; *Marchant et al.*, 2016], aerosol type [*Levy et al.*, 2013; *Sayer et al.*,

51      2014], and snow coverage over land surfaces [*Hall and Riggs*, 2016]. Decision trees or voting

52      schemes involving multiple thresholds are typically used in these traditional algorithms. The

53      decision tree branches, tests, and thresholds are often determined empirically after a tedious hand

54      tuning/testing process based on the developer's experience and access to validation datasets.

55      Further, the branches and thresholds are often very sensitive to the specific instrument (e.g.,

56      spectral band pass, calibration, noise characteristics, view/solar geometry sampling). Therefore,

57      an obvious weakness of these traditional methods (e.g., decision trees/voting schemes/thresholds)

58    is that it is challenging and time consuming to develop algorithms across multiple instruments and

59    to maintain performance for a single instrument having radiometric stability issues. Meanwhile, a

60    well-designed traditional method may have remarkable performance in a specific region and

61    season yet have significant biases when applied globally and/or annually [*Cho et al.*, 2009; *Liu et*

62    *al.*, 2010; *Zhou et al.*, 2019]. Additional complexities arise when DC problems become more non-

63    linear across large spatial and temporal scales, and more variables need to be considered. It is

64    difficult to develop and apply a single or a few decision trees to complicated non-linear problems

65    that are controlled by dozens or more variables. As expected, a single decision tree can grow very

66    deep and tend to have a highly irregular structure in order to consider a large number of features

67    (variables) simultaneously, leading to a significant overfitting effect (i.e., an over-constrained

68    training that makes predictions too close to the training dataset but fails to predict future

69    observations reliably). For example, MODIS provides an all-day cloud phase product based only

70    on infrared (IR) observations (hereafter referred to as IR-Phase [*Baum et al.*, 2012]). Although it

71    can be expected that the tests and thresholds should vary with satellite viewing geometry [*Maddux*

72    *et al.*, 2010], full consideration of viewing geometries, together with the variations of many other

73    factors such as surface emission, geolocation, and cloud properties, is very challenging based on

74    manual tuning. As a consequence, it is found that the liquid water and ice cloud fractions from the

75    IR-Phase product exhibit noticeable view zenith angle (VZA) dependency (see Figure 12). This is

76    an undesirable but unavoidable artifact since cloud phase statistics should be independent from

77    solar/viewing geometry. Such VZA dependencies may strongly affect similar products from

78    geostationary instruments because of the fixed VZA-geolocation mapping. Similar artifacts may

79    also impact aerosol type and retrieval products [*Wu et al.*, 2016]. Finally, it is difficult to acquire

80    pixel-level classification uncertainties with traditional methods.

81      In contrast to traditional methods, Machine Learning (ML) based DC algorithms are designed

82    to autonomously find information (e.g., patterns of spectral, spatial, and/or time series) in one or

83    more given datasets and learn hidden signatures of different objects. An obvious advantage of ML

84    models is that the training process is efficient and highly flexible. Manually defined thresholds or

85    matching conditions to expected spectral patterns are no longer needed. In this paper, we developed

86    two ML-based DC algorithms for detecting cloud and cloud thermodynamic phase for different

87    local times (i.e., daytime and nighttime) with observations from the Visible Infrared Imaging

88    Radiometer Suite (VIIRS) on Suomi-NPP (SNPP). The ML models are trained with collocated

89    observations from SNPP VIIRS and Cloud-Aerosol Lidar with Orthogonal Polarization

90    (CALIOP), with CALIOP data used as reference. In Section 2, we give a brief discussion of the

91    ML models. Data generated for model training and validation will be introduced in Section 3.

92    Details of the model training and evaluation are shown in Section 4. Section 5 discusses the

93    advantages and potential limitations of the present ML models. Conclusions are given in Section

94    6.

95    **2. Traditional DC methods and Machine Learning Models**

96    **2.1 Traditional DC methods**

97       All DC algorithms with remote sensing observations are based on the underlying physics of

98    the spectral, spatial, and/or temporal structures of specified objects. In traditional DC algorithms,

99    all the physical rules and structures have to be explicitly defined as various tests and thresholds.

100    For example, the MODIS MOD35/MYD35 cloud mask algorithm uses more than 20 tests with

101    visible/near-infrared (VNIR), shortwave-infrared (SWIR), and infrared (IR) observations [*Frey et*

102    *al.*, 2008] that are carefully designed to consider numerous scenarios, including different surface

103    types (e.g., ocean, land, desert, snow, etc.) and local times (day/night). Similar algorithms are

104    designed for aerosol type and cloud thermodynamic phase classifications. As an example, Figure

105   1 illustrates spectral patterns of 5 typical daytime oceanic scenes (pixel types) observed by SNPP

106   VIIRS. Spectral pattern of each of the 5 scenes, namely, clear sky, liquid water cloud, ice cloud,

107   dust, and smoke, is averaged by using more than 1,000 pixels with the same type. It is clear that

108   the 5 scenes are different in either reflectance ratios between a given VNIR/SWIR band and the

109   0.86 $\mu$m band, or brightness temperature differences (BTD) between two IR window bands.

110   Consequently, such spectral features are frequently used to differentiate pixel types in DC

111   algorithms. In addition to spectral patterns, simple methods are developed to take into account

112   spatial information. For example, it is found that cloud reflectance usually has larger spatial

113   variability than aerosols [*Martins et al.*, 2002] and clear sky pixels [*Platnick et al.*, 2017].

114   Therefore, spatial variabilities of VNIR and SWIR reflectance bands are used to differentiate

115   clouds from non-cloudy pixels in the current MODIS clear sky restoral (CSR) algorithm [*Platnick

116   et al.*, 2017] and Dark Target aerosol retrieval algorithm [*Levy et al.*, 2013].

117   **2.2 Machine learning models**

118       Different from the traditional DC methods, ML algorithms are developed to autonomously

119   learn the hidden spectral/spatial/temporal patterns of different objects. Consequently, manually

120   defined thresholds or matching conditions to expected patterns are no longer needed. In image

121   recognition applications, numerous ML algorithms [e.g., *Joachims* 1998; *Breiman* 1999;

122   *Dieterich* 2000] have been developed for independent pixels using a single or small number of

123   decision trees in late 1990s. *Ho* [1998] and many other studies have demonstrated that although

124   these single or small number of decision trees can always provide maximum prediction accuracies

125   in training processes, significant overfitting effects cannot be avoided. Tremendous efforts have

126   been made to overcome the dilemma between maintenance of prediction accuracy and avoiding

127   overfitting.   Among these, the Random Forest (RF) and Gradient Boosting (GB) algorithm

128   [*Breiman* 1999; *Dieterich* 2000; *Friedman* 2001] provide a framework of using a large number of

Atmospheric
Measurement
Techniques
Discussions

129     decision trees (ensemble) but a subset of features in each tree to achieve optimization in the

130     performance. It has been proven that the ensemble-based algorithms can largely correct mistakes

131     made by individual trees [*Ji and Ma*, 1997; *Tumer and Ghosh*, 1996; *Latinne et al.*, 2001] and

132     avoid overfitting [*Freund et al.*, 2001]. Currently, the RF and GB algorithms are frequently used

133     in non-linear classification and regression problems. For example, RF models have been used in

134     several cloud/aerosol remote sensing applications, such as differentiating cloudy from clear

135     footprints for the Clouds and the Earth's Radiation Energy System (CERES) instrument [*Thampi*

136     *et al.*, 2017], estimating surface-level PM2.5 concentrations [*Hu et al.*, 2017], and detecting low

137     cloud using Advanced Baseline Imager (ABI) on Geostationary Operational Environmental

138     Satellites (GOES) [*Haynes et al.*, 2019]. In our study, we also choose the RF model based on its

139     proven record in earth science applications.

140     In the RF model, a final prediction is made based on majority vote computed from probability

141     ($P_i$) of each class ($i^{th}$):

142     $$P_i = \frac{w_i N_i}{\sum_{j=1}^{j=m} w_j N_j},$$     (1)

143     where $m$ is the total number of classes, $N_i$ and $N_j$ are the number of trees that predict the $i^{th}$ and $j^{th}$

144     classes, and $w_i$ and $w_j$ are weightings for the $i^{th}$ and $j^{th}$ classes, respectively. If all trees are equally

145     weighted, $w$ for individual classes are equal to 1. The two most important parameters for tuning

146     the RF algorithm are the number of decision trees ($N_{Tree}$) and the maximum tree depth ($N_{Depth}$).

147     However, an optimal definition of these two parameters is still an open question [*Latinne et al.*,

148     2001]. Larger $N_{Tree}$ and $N_{Depth}$ provide more accurate predictions, at the cost of significantly

149     increased computational resources. For many cases, larger $N_{Depth}$ may cause overfitting effects

150     [*Oshiro et al.*, 2012; *Scornet*, 2018]. Generally, the two parameters have to be large enough to let

151     the decision trees have a relatively wide diversity and capture the hidden patterns. For practical

152    purposes, however, the two parameters cannot be too large to prevent the models from overfitting

153    and to reduce computing burden [*Latinne et al.*, 2001; *Scornet* 2018].

154    In this study, we adopt a widely applied RF algorithm in the Scikit-learn Machine Learning

155    package [*Pedregosa et al.*, 2011]. We train two RF models for object DC using SNPP VIIRS

156    spectral observations at two observational times: an all-day RF model using three VIIRS thermal

157    IR observations (hereafter referred to as the RF all-day model) and a daytime-only RF model that

158    uses both VNIR/SWIR and thermal IR observations (hereafter the RF daytime model). The models

159    are trained to detect clear sky, liquid water cloud, and ice cloud pixels with single pixel level

160    information. Parameters of the two RF models will be tuned and tested carefully to achieve the

161    best accuracy and to avoid the overfitting effect. Details will be discussed in Section 4.

162    **3. Data**

163    **3.1 Reference label of pixels**

164    Space-borne active sensors, such as CALIOP onboard CALIPSO [*Winker et al.*, 2013], the

165    Cloud-Aerosol Transport System (CATS) [*McGill et al.*, 2015] onboard the International Space

166    Station (ISS), and CPR on board CloudSat [*Stephens et al.*, 2002], are frequently used to evaluate

167    the performance of traditional cloud/aerosol DC and property retrieval algorithms designed for

168    passive sensors [*Stubenrauch et al.*, 2013; *Wang et al.*, 2019]. Until its exit on 13 September 2018

169    (to join CloudSat in the C-Train), CALIPSO was a key member of the Afternoon Constellation of

170    satellites (A-Train), and began providing profiling observations of the atmosphere in 2006 [*Winker*

171    *et al.*, 2013]. The CALIPSO lidar CALIOP operates at wavelengths of 532 nm and 1064 nm,

172    measuring backscattering profiles at a 30-meter vertical and 333 m along-track resolution.

173    CALIOP also measures the perpendicular and parallel signals at 532 nm, along with the

174    depolarization ratio at 532 nm that is frequently used in cloud phase discrimination algorithms

Atmospheric
Measurement
Techniques
Discussions

175    because of its strong particle shape dependence. The CALIOP Version 4 Level 2 1 km Cloud Layer

176    product will be used to provide reference cloud phase labels in both model training and validation

177    stages.

178    While the CATS lidar and the CloudSat radar CPR also provide profiling information, both

179    have limitations that preclude their use here. CATS had a relatively short life time (from January

180    2015 to October 2017), and its low inclination angle (51°) orbit aboard the ISS excludes sampling

181    of high-latitude regions [*Noel et al.*, 2018]. CloudSat CPR observes reflectivity profiles at 94-GHz,

182    which are more sensitive to optically thicker clouds consisting of large particles but are blind to

183    aerosols and optically thin clouds. CloudSat also has difficulty in detecting clouds near the surface

184    due to the surface clutter effect [*Tanelli et al.*, 2008]. Therefore, only CALIOP data are used to

185    provide reference cloud phase labels in this study.

186    **3.2 RF model input**

187    It should be pointed out that ML models use similar input datasets as traditional methods. The

188    input variables (features) and reference labels of the present RF models are carefully selected based

189    on prior physical knowledge of the spectral characteristics of each object.

190    The SNPP VIIRS and the NOAA-20+ series provides spectral observations from 0.4 to 12 $\mu$m

191    at sub-kilometer spatial resolutions [*Lee et al.*, 2006]. Specifically, VIIRS has 16 moderate

192    resolution bands (M band) and 5 higher resolution imagery bands (I band) at 750 m and 375 m

193    nadir resolutions, respectively. The spectral capabilities of VIIRS allow for extracting abundant

194    information on the surface and atmospheric components, such as clouds [*Ackerman et al.*, 2019]

195    and aerosols [*Sayer et al.*, 2018]. It is also worth noting that VIIRS utilizes an on-board detector

196    aggregation scheme that minimizes pixel size growth in the across-track direction towards swath

197    edge [*Cao et al.*, 2013]. As an example, although the VIIRS M-bands and MODIS 1 km bands

198   have similar nadir spatial resolutions, the VIIRS across-track pixel size increases roughly only to

199   1.625 km at scan edge, which is much smaller than the MODIS pixel size of roughly 4.9 km at

200   scan edge [*Justice et al.*, 2011]. Another obvious advantage of using SNPP VIIRS rather than Aqua

201   MODIS data is that, due to the CALIPSO and SNPP orbit differences, the training samples cover

202   a broader viewing zenith angle range, which is a great benefit to overall model performance.

203   Consequently, Level-1B M-band observations from the SNPP VIIRS are used here.

204       Ancillary data, including the surface skin temperature, spectral surface emissivity, surface

205   types, and snow/ice coverage, are important in cloud DC related remote sensing applications [*Frey*

206   *et al.*, 2008; *Wolters et al.*, 2008; *Baum et al.*, 2012] and cloud/aerosol retrievals [*Levy et al.*, 2013;

207   *Wang et al.*, 2014; 2016a; 2016b; *Meyer et al.*, 2016; *Platnick et al.*, 2017]. The inst1_2d_asm_Nx

208   product (version 5.12.4) from the Modern-Era Retrospective Analysis for Research and

209   Applications, Version 2 (MERRA-2) [*Gelaro et al.*, 2017] is utilized to provide the hourly

210   instantaneous surface skin temperature and 10-meter surface wind speed. The UW-Madison

211   baseline fit land surface emissivity database [*Seemann et al.*, 2008] and the Terra/Aqua MODIS

212   combined Land surface product (MCD12C1 [*Sulla-Menashe and Friedl* 2018]) are used to provide

213   monthly mean land surface emissivities for the mid-wave to thermal IR bands (3.6 ~ 14.3 $\mu$m) and

214   surface white sky albedo for the VNIR bands (0.4 ~ 2.3 $\mu$m), respectively, at a 0.05×0.05° spatial

215   resolution. Surface types and snow/sea ice coverage data are from the International Geosphere-

216   Biosphere Programme (IGBP) and daily Near-real-time Ice and Snow Extent (NISE) data [*Brodzik*

217   *and Stewart*, 2016], respectively.

218   **3.3 Clear and cloud phase classifications from existing VIIRS and MODIS products**

219       Since the present RF models are trained with SNPP VIIRS observations, the first priority of

220   this study is evaluating and comparing the trained RF models with CALIOP and the existing VIIRS

221     cloud products. Due to the viewing geometry differences between VIIRS and MODIS, it is difficult

222     to simply attribute the mask/phase differences from the RF models and MODIS products to

223     algorithms. However, existing cloud mask and phase products from Aqua MODIS are still

224     compared as a reference in this work.

225     The Aqua MODIS and SNPP VIIRS CLDMSK (cloud mask) and CLDPROP (cloud top and

226     optical properties) [*Ackerman et al.*, 2019] products represent NASA's effort to establish a long-

227     term consistent cloud climate data record, including cloud detection and thermodynamic phase,

228     across the MODIS and VIIRS observational records. While the CLDMSK and CLDPROP

229     algorithms share heritage with the standard Collection 6.1 MODIS cloud mask (MYD35) and

230     cloud top and optical properties (MYD06) algorithms, both algorithms use only a subset of bands

231     common to both sensors to minimize differences in instrument spectral information content. The

232     initial Version 1.0 of the CLDMSK and CLDPROP products were publicly released in April 2019;

233     CLDPROP has since been reprocessed to Version 1.1, which includes a fix to the optical property

234     thermodynamic phase algorithm, with public release in October 2019.

235     The CLDMSK and MYD35 algorithms use a variety of band combinations and thresholds

236     depending on cloud and surface types [*Frey et al.*, 2008; *Ackerman et al.*, 2008]. Meanwhile, the

237     algorithms use different approaches for daytime (i.e., solar zenith angle less than 85°) and

238     nighttime pixels. In the CLDMSK and MYD35 algorithms, pixels are categorized into four

239     categories, namely confident clear, probably clear, probably cloudy, and cloudy. The CLDPROP

240     and MYD06 algorithms separate cloudy and probably cloudy pixels into liquid water, ice, and

241     unknown phase categories. Specifically, the MYD06 product includes two cloud phase algorithms:

242     an IR-Phase algorithm [*Baum et al.*, 2012] that uses observations in four MODIS IR bands for

243     daytime and nighttime phase classification (hereafter referred to as the MYD06 IR-Phase), and a

244     daytime-only algorithm designed for the cloud optical properties retrievals [*Marchant et al.*, 2016;

245    *Platnick et al.*, 2017] that uses VNIR/SWIR, and IR observations (hereafter referred to as the

246    MYD06 OP-Phase). A notable change for the VIIRS/MODIS CLDPROP algorithm with respect

247    to the standard MODIS MYD06 algorithm is the replacement of the MYD06 IR-Phase by a NOAA

248    operational algorithm originally developed for Clouds from AVHRR-Extended (CLAVR-x)

249    [*Heidinger et al.*, 2012] and now applied to VIIRS and GOES-16/17. This algorithm is used to

250    provide cloud top properties, including thermodynamic phase (hereafter CLDPROP CT-Phase), in

251    the absence of $CO_2$ IR gas absorption bands. IR bands are primarily used in the CLDPROP CT-

252    Phase algorithm, while complementary SWIR bands are used when available. The MYD06 OP-

253    Phase algorithm, applied to daytime pixels only, is included with only minor alteration (related to

254    cloud top properties changes) in the VIIRS/MODIS CLDPROP product (hereafter referred to as

255    the CLDPROP OP-Phase).

256    Although the MYD06 and CLDPROP OP-Phase products are developed for "cloudy" and

257    "probably cloudy" pixels from the MYD35 and CLDMSK products, a Clear Sky Restoral (CSR)

258    algorithm [*Platnick et al.*, 2017] is implemented to remove "false cloudy" pixels from the clear-

259    sky conservative MYD35 and CLDMSK products. Specifically, the CSR uses a set of spectral and

260    spatial reflectance variability tests to remove dust, smoke, and strong sunglint pixels that are

261    erroneously identified as "cloudy" or "probably cloudy" by the MYD35 and CLDMSK products

262    [*Platnick et al.*, 2017]. One should keep in mind that the CSR algorithm is only applied for the

263    optical property retrievals. Thus, the MYD35 and CLDMSK, and consequently the MYD06 IR-

264    Phase and CLDPROP CT-Phase, may have "false cloudy" pixels in comparison with CALIOP,

265    while the impact on the MYD06 and CLDPROP OP-Phase is reduced due to the CSR algorithm.

266    The cloud mask and thermodynamic phase products used in this study are summarized in Table 1.

267    **4. Model training and validation**

268    Here we discuss the training of the all-day and daytime RF models for different surface types.

269    Both shortwave (SW) and IR observations will be used in the daytime models while only IR

270    observations will be used in the all-day models.

271    **4.1 Surface Types**

272    RF models are trained for different surface types, defined here by the Collection 6 (C6) MODIS

273    annual IGBP surface type product (MCD12C1), to improve model performance over a single

274    general model for all surface types. Although the MCD12C1 product includes up to 18 surface

275    types, for this work we attempt to reduce the total number of surface types by combining surface

276    types with similar spectral white sky albedos and emissivities, as suggested by *Thampi et al.*

277    [2017]. An annual global IGBP surface type map and surface albedo data from the MODIS

278    MCD12C1 [*Sulla-Menashe and Friedl* 2018], and a UW-Madison monthly global land surface

279    emissivity database [*Seemann et al.*, 2008], are used to generate the climatology of land surface

280    white-sky albedo and IR emissivity spectra. The UW-Madison database is derived using input

281    from the MODIS operational land surface emissivity product MOD11 [*Wan et al.,* 2004] at six

282    wavelengths located at 3.8, 3.9, 4.0, 8.6, 11, and 12 $\mu$m.  A baseline fit method is applied to fill

283    the spectral gaps and provides a more comprehensive IR emissivity dataset at 10 wavelengths from

284    3.6 to 14.3 micron for global land surface with a 0.05° spatial resolution [*Seemann et al.*, 2008].

285    The MODIS MCD12C1 product also provides a white-sky albedo dataset at 0.47, 0.56, 0.66, 0.86,

286    1.24, 1.64, and 2.13 $\mu$m with a 0.05° spatial resolution [*Sulla-Menashe and Friedl* 2018]. The

287    means and standard deviations of surface emissivity and white-sky albedo spectra are shown in

288    Figures 2 a) and 3 a), respectively, for 16 different land surface types generated from the UW-

289    Madison and MCD12C1 data in 2015. Land surface types with similar IR emissivity and SW

290    white-sky albedo spectra are grouped to reduce to the total number of land surface types to 6

291    (forest, cropland, grassland, snow/ice, barren/desert, and shrubland), as shown in Figures 2 (b-f)

Atmospheric
Measurement
Techniques
Discussions
Open Access
EGU

292    and 3 (b-f). Figure 4 shows an example map of the reduced global surface type data generated

293    from the MCD12C1 product for 2015.

294    **4.2 Generating Training/Validation Datasets**

295        The training and validation data are obtained from a 5-year (2013-2017) SNPP VIIRS and

296    CALIOP collocated dataset. The collected dataset is generated with a strict collocation algorithm

297    that fully considers the spatial differences between the two instruments and parallax effects, as

298    described in *Holz et al.* [2008]. The SNPP VIIRS data include L1B calibrated reflectance and

299    brightness temperatures, and the CALIOP data include the L2 1 km cloud layer products. A strict

300    three-step quality control process is applied to all collocated pixels to ensure data quality in the

301    training process. First, VIIRS 750 m pixels that are potentially contaminated by aerosol are

302    excluded using a threshold of 0.1 column AOD at 532 nm from the CALIOP L2 5 km aerosol layer

303    product. Second, each aerosol-free pixel is labelled by one of four categories, namely, clear sky

304    and liquid, ice, and ambiguous clouds using the CALIOP L2 1 km layer product, where the liquid

305    and ice categories include both single-layered cloud scenes and multi-layered clouds with the same

306    thermodynamic phases (e.g., ice over ice, liquid over liquid). Furthermore, the pixels with

307    ambiguous types, including uncertain/unknown cloud phases from CALIOP and/or overlapping

308    objects belonging to different types (e.g., cirrus over liquid), are discarded. Finally, horizontally

309    inhomogeneous pixels, determined when the CALIOP 1 km label changes within 5 consecutive

310    VIIRS pixels, are discarded. Figure 5 shows the global distributions of the 5-year collocated clear

311    (first row) and cloudy pixels (second row) before and after applying the three-step quality control.

312    Globally, 50% of total clear pixels are excluded due to contamination of broken-cloud and/or

313    aerosol. In particular, a large fraction of clear pixels in central Africa, India, and southern China

314    (Figure 5c) are excluded due to relatively large aerosol optical thicknesses in those regions. About

315    40% of global cloudy pixels (Figure 5f) are excluded probably due to cloud heterogeneity and

316    aerosol contamination. Regions with complicated cloud structures (e.g., the Inter Tropical

317    Convergence Zone) have the minimum selection rates (20%). The remaining data are separated

318    into a training/testing population which consists of 33.4, 41.0 and 24.6 million pixels for clear sky,

319    liquid water cloud, and ice cloud from years 2013-2016, respectively, and a validation dataset that

320    consists of 6.9, 8.3 and 5.0 million pixels of clear-sky, liquid water cloud, ice cloud, respectively

321    from year 2017.

322    **4.3 RF model training and configuration**

323    RF model performance is determined by both its inputs (spectral or other information) and its

324    configuration ($N_{Tree}$ and $N_{Depth}$). Therefore, extensive testing must be conducted to find the optimal

325    inputs and configuration. The 4-year collocated VIIRS-CALIOP dataset from 2013 to 2016 after

326    quality control (see Section 4.2) is used for both training (75%) and testing (25%) purposes. The

327    testing set, also known as cross-validation set, is used to tune and optimize the RF model

328    parameters. Here we define an accuracy score to evaluate the overall model performance. The

329    accuracy score is the ratio of pixels (samples) that the both the CALIOP and RF model have the

330    same categories to total pixels.

331    Table 2 provides accuracy scores of the IR-based all-day model trained and tested with

332    different inputs. It shows that with a fixed RF model configuration ($N_{Tree}$ = 150 and $N_{Depth}$ = 15),

333    the RF all-day model with input #4 and #6 have the best overall accuracy scores for all surface

334    types. Generally, by including surface skin temperature ($T_s$) and geolocation (i.e., latitude and

335    longitude), the accuracy scores for all surface types increase by 2-3%. The surface emissivity

336    vector $\varepsilon_s$ is less important, likely because this information is highly correlated to surface type and

337    geolocation. In this study, input #4 is selected mainly because with similar performance, it requires

338    less memory and computing resources, and it is quite possible that more uncertainty is introduced

339    with the use of a surface emissivity vector $\varepsilon_s$ from another retrieval product.

340    A set of model configurations ($N_{Tree}$ and $N_{Depth}$) are also tested based on the selected input #4.

341    While the number of trees and the maximum depth of individual trees are important determinants

342    for RF model performance, the overall accuracy scores for all surface types are less sensitive to

343    these two model parameters when more than 100 trees and 10 maximum tree depths are used (not

344    shown here). Therefore, we trained the RF all-day models with input #4 and the model

345    configuration used in Table 2, i.e., $N_{Tree}$ = 150 and $N_{Depth}$ = 15.

346    Similar tests for the RF daytime model (IR plus NIR and SWIR observations) showed that the

347    optimal input includes reflectances in the 0.86, 1.24, 1.38, 1.64 and 2.25$\mu$m bands, BTs in the

348    same 3 IR bands in the all-day model, $T_s$, geolocation, and solar/satellite viewing zenith angles.

349    The same model configuration used in the all-day model, e.g., 150 trees with the maximum depth

350    15, is used in the daytime model. The accuracy scores of the RF daytime model are higher than

351    the RF all-day model by 2-3% over almost all surface types except high-latitude regions covered

352    by snow and ice, where the daytime model accuracy score is higher by up to 6% than the all-day

353    model due to the inclusion of the 1.38, 1.64 and 2.25µm SWIR bands.

354    **4.4 Evaluating the RF Models**

355    The trained RF all-day and daytime models are validated using collocated CALIOP data in

356    2017. Existing VIIRS cloud products CLDMSK and CLDPROP (see Table 1) are included for

357    direct comparison with the RF models and CALIOP reference. Several other products, such as the

358    MODIS CLDMSK and CLDPROP and standard MYD35 and MYD06, are also included for

359    comparison although they could be different from the RF models due to other non-algorithm

360    reasons, such as the VZA and pixel size differences mentioned before.

361    *4.5.1 Cloud mask*

362    Cloud mask from the two RF models and VIIRS/MODIS products are first compared with

363    CALIOP lidar observations. For the two models, a cloudy pixel indicates a predicted label "liquid"

364    or "ice". Here we define cloudy and clear pixels as "positive" and "negative" events, respectively.

365    A true positive rate (TPR) and false positive rate (FPR) can then be used to evaluate model

366    performance. The TPR and FPR are defined as:

367    $$\text{TPR} = \frac{TP}{TP+FN}, \tag{2}$$

368    $$\text{FPR} = \frac{FP}{FP+TN}, \tag{3}$$

369    where TP (True Positive) and TN (True Negative) are the number of lidar-labeled "cloudy" and

370    "clear" pixels, respectively, that are correctly detected by the models; whereas FN (False Negative)

371    and FP (False Positive) are the number of lidar-labeled "cloudy" and "clear" pixels incorrectly

372    identified by the models. Therefore, TPR, also called model sensitivity, indicates the fraction of

373    all positive events (i.e., lidar cloudy pixels) that are correctly detected by the models. Similarly,

374    FPR, also called false alarm rate, indicates the fraction of all negative events (i.e., lidar clear pixels)

375    that are incorrectly detected as positive (cloudy). TPR and FPR are two critical parameters in

376    model evaluation. A perfect model is associated with a high TPR (close to 1) and a low FPR (close

377    to 0).

378    Figure 6 shows daytime cloud mask TPR-FPR plots from the two RF models and the other

379    products listed in Table 1. Globally, all products agree well with lidar observations (Figure 6a).

380    The overall TPRs are higher than 0.94 and FPRs are lower than 0.08. The RF daytime model (red

381    circle), with a TPR of 0.97 and an FPR of 0.05, is slightly better than the RF all-day model (yellow

382    circle) and other products. Figure 6b-6h show comparisons over different surface types. It is clear

383    that the RF daytime model has a robust performance for all surface types. The MODIS MYD35

384    cloud mask algorithm (black circle) performs best over ocean but has a relatively high FPR (0.22)

Atmospheric
Measurement
Techniques
Discussions

Open Access

EGU

385 over forest and low TPR over snow/ice and barren (0.85) regions. As mentioned in Section 3, the

386 "false cloudy" pixels from MYD35 and CLDMSK may increase the FPRs correspondingly.

387 The RF all-day model works fairly well and is comparable to other products for all surface

388 types regardless of the fact that it only uses three IR window channels from VIIRS while all other

389 products in the daytime models use VNIR observations. Nighttime (SZA > 85°) cloud mask

390 comparisons are shown in Figure 7. The overall performances of all operational products decrease

391 in particular for snow/ice regions. For example, the VIIRS/MODIS CLDMSK products over

392 snow/ice surface have large fractions of missing "cloudy" pixels (e.g., TPRs < 0.7) and false alarm

393 rates (FPRs > 0.2) over snow/ice surface. The decrease is more likely explained by the lack of

394 SWIR bands and the small cloud-snow/ice surface temperature contrast during the nighttime of

395 summer polar regions. However, the RF all-day model has the best performance for nighttime

396 pixels, indicating the strong capability of ML based algorithm in capturing hidden spectral features

397 and optimizing dynamic thresholds of clear and cloudy pixels.

398 *4.5.2 Cloud thermodynamic phase*

399 The RF model derived cloud thermodynamic phase products are also compared with CALIOP

400 lidar and existing VIIRS and MODIS products. For consistent nomenclature, we arbitrarily define

401 ice clouds and liquid water clouds as "positive" and "negative" events, respectively. To focus on

402 cloud thermodynamic phase classification, pixels detected as "clear" by either the lidar reference

403 labels or by the RF models and existing products are excluded. The OP-Phase from both MYD06

404 and CLDPROP, and the IR-Phase from MYD06, have an "unknown phase" category, which is not

405 included in the TPR-FPR analysis.

406 Figure 8 shows daytime cloud phase TPR-FPR plots from the two RF models and the

407 MODIS/VIIRS products. The two RF models and the MODIS MYD06 OP-Phase are the top 3

408    phase algorithms for all surface types. The MODIS MYD06 IR-Phase, MODIS/VIIRS CLDPROP

409    OP-Phase, and CT-Phase have either relatively lower TPRs or higher FPRs over particular surface

410    types, such as shrubland, snow/ice, and barren regions. Comparisons between nighttime phase

411    algorithms are shown in Figure 9. For nighttime clouds, the RF all-day model works better than

412    both CT-Phase and IR-Phase algorithms for all surface types.

413        Figure 10 shows monthly mean daytime cloud and phase fractions from the VIIRS CLDMSK

414    and CLDPROP OP-Phase products (top row), and those from the RF daytime model (second row),

415    in January 2017. For the cloud mask comparison, cloud fractions (CF) from the two products have

416    similar spatial patterns, while it is also clear that the VIIRS CLDMSK CFs are higher over tropical

417    oceans by approximately 10% and lower over land by 5% (Figure 10 c). This is consistent with

418    the cloud mask TPR-FPR analysis shown in Figure 6. Over the tropical ocean, the VIIRS

419    CLDMSK is more "cloudy", probably due to a fraction of sunglint pixels that are detected as liquid

420    clouds, leading to a large FPR rate. Another reason for the relatively large cloud fraction (or liquid

421    water cloud fraction) difference is that in regions covered by "broken" cumulus clouds, and or

422    clouds with more complicated structures, the inherent viewing geometry differences in the training

423    datasets may adversely affect the performance of the RF models. For example, CALIOP, with a

424    nadir viewing geometry may observe clear gaps between two small cloud pieces, while VIIRS,

425    with an oblique viewing angle, detects broken liquid clouds nearby or high clouds along its long

426    line-of sight. Comparison between the VIIRS product and the RF daytime model shows more ice

427    clouds from the RF daytime models over land, which is consistent with the cloud phase TPR-FPR

428    plots as shown in Figure 8. The RF daytime model may have better performance due to the

429    consideration of surface type. However, it is also important to notice that due to the lack of

430    "aerosol" types in current training, in central Africa, the RF models may misidentify elevated

431     smoke as ice cloudy pixels. For most land surface types except snow/ice, the CLDPROP OP-Phase

432     has lower TPR rates than the RF daytime models by 0.1, in comparison with the CALIOP.

433     In addition to the higher CFs over low latitude ocean from the VIIRS CLDMSK product, more

434     pronounced CF (liquid) differences can be found in northeast and northwest China. Cloud

435     differences in the two regions are spatially correlated with locations that have heavy aerosol

436     loadings or snow coverage. For example, heavy aerosol loadings due to pollution in Northeast

437     China, and a wide land snow coverage in Northwest China are frequently observed in the winter.

438     The VIIRS CLDMSK may identify pixels with white surface and heavy aerosol loadings as

439     "cloudy". Some of these pixels are expected to be restored to clear-sky category in the CLDPROP

440     OP-Phase product (Figure 10 f and 10 i). As evidence, Figure 11 shows comparisons between the

441     VIIRS products and the RF daytime model in July 2017. The large cloud (liquid) fraction

442     differences over North China vanish in the summer. This indicates that the RF models might be

443     able to handle complicated (or unexpected) surface type and strong aerosol events better than the

444     traditional VIIRS algorithm. However, further investigation is required to understand the

445     performances of both the VIIRS products and the RF models.

446     **5. Discussion**

447     In this Section, we will review the strengths and potential limitations and weaknesses of the

448     RF models.

449     **5.1 Advantages**

450     The above results show that the two RF models have better and more consistent performance

451     over different regions and surface types in comparison with the MODIS and VIIRS products. In

452     addition to better performance, it is convenient and efficient to apply the present RF models or

453     other similar ML-based models to other instruments similar to VIIRS, such as the geostationary

Atmospheric
Measurement
Techniques
Discussions

454    imagers Advanced Himawari Imager (AHI) on Himawari-8/9, the ABI on GOES-16/17, and the

455    Spinning Enhanced Visible and Infrared Imager (SEVIRI) on Meteosat Second Generation, as

456    long as reliable reference pixel labels are available. The RF models can be trained and tested for

457    different surface types and using different input variables in a few hours. In contrast, traditional

458    methods may suffer from the change of instrument, solar/viewing geometries, and surface

459    conditions. For example, although the MODIS MYD06 OP-Phase and CLDPROP OP-Phase use

460    similar input and strategies, cloud thermodynamic phases from the two products are different by

461    up to 5% for all surface pixels, and by up to 20% over surfaces covered by snow/ice (see Figure 8

462    black and light blue circles). Besides providing a discrete category for each pixel, the RF models

463    provide an ensemble of predictions and probabilities of individual categories, which are useful

464    diagnostic variables in evaluating models in complicated scenarios.

465    **5.2 Limitations and possible caveats**

466    Although the evaluation demonstrates that the current RF models are highly consistent with

467    CALIOP, the models may suffer some artifacts due to the quality of the training data and due to

468    sampling issues.

469    *5.2.1 Quality of the training data*

470    The RF models learn spectral structures of cloud/clear pixels according to the reference labels.

471    As a consequence, the present model performance relies heavily on the quality of CALIOP Level-

472    2 data. It is already known that the lidar signal has limitations in detecting the bottom of an

473    optically thick cloud or lower level clouds underneath an opaque cloud [*Sassen and Cho*, 1992].

474    Cloud phase classification from the RF daytime model, using both SW and IR observations, may

475    be different from the CALIOP at multi-layer scenes if the top cloud layer is optically thick enough

476    for lidar (e.g., optical thickness greater than 3). Using combined CALIOP and CloudSat data as

477　reference and introducing a "multiple layer clouds" category could be a way to mitigate this

478　impact.

479　　　Additional uncertainties may come from the inconsistency in view angles between the

480　collocated CALIOP labels and VIIRS spectral observations. For instance, CALIOP always has a

481　quasi-nadir viewing angle (e.g., 3°) whereas the collocated VIIRS observations have a wide VZA

482　range (e.g., 0° to 50°). A wide VIIRS VZA range in the training dataset improves model

483　performance, especially for predicting VIIRS pixels with large VZAs. However, the difference

484　between the CALIOP and VIIRS viewing geometry could create undesirable artifacts in the

485　training process. As shown in Figure 10, in the descending areas of the Hadley cell over low-

486　latitude ocean, where marine boundary layer clouds are dominant, there are relatively large CF

487　differences between the CLDMSK and the RF models. A reason for the large liquid cloud fraction

488　differences is that the quality of training datasets decreases in regions covered by "broken"

489　cumulus clouds, and or clouds with more complicated structures. Further investigation is required

490　to check if the data screening process introduces sampling bias into the training dataset.

491　*5.2.2 Sampling issue*

492　　　Uneven sampling may also influence the training of RF models. Figure 12 shows the cloud

493　fraction as a function of viewing geometry. Quasi-constant fractions of both liquid and ice clouds

494　are found for all operational products and the RF models when VZAs are smaller than 45°, except

495　the MODIS MYD06 IR-Phase, which has a strong VZA dependency. However, liquid (ice) cloud

496　fractions from the two RF models increase (decrease) rapidly at high VZAs (greater than 50°),

497　which is likely caused by the sampling issue. A significant fraction of the training data (greater

498　than 98%) is located in the region with VZA less than 50° (see the gray dashed distributions in

499　Figure 12). It is difficult to mitigate this issue using collocated VIIRS-CALIOP data or

500    observations from other similar instruments in the training process. One possible way is using

501    model- generated synthetic training data and labels with reliable radiative transfer models. Results

502    from the RF daytime model are not shown in Figure 12 since they are highly consistent with the

503    RF all-day model.

504    *5.2.3 Labeling strategy*

505    For RF or other ML models, each pixel's classification is determined by prediction

506    probabilities ($P$) of all potential types. Here we selected a regular strategy that labels a pixel using

507    the class with the highest probability (see Eq. 1). This strategy is logical for problems with two

508    categories (e.g., cloud mask only). For problems including 3 or more classes, however, the present

509    strategy is not the only way to label pixels. For example, a pixel is labeled as "clear" if $P_{clear}$ is

510    larger than both $P_{liquid}$ and $P_{ice}$ according to the current labeling strategy. It is also possible that,

511    for the same pixel (less than 0.5% for the two RF models), $P_{clear}$ is lower than the sum of $P_{liquid}$

512    and $P_{ice}$, making a "cloudy" label more appropriate. For the cloud mask and phase problem

513    discussed in this paper, in addition to pixel labels, users must be aware of probabilities of the three

514    types. Another possible way to avoid the ambiguous labeling is using two RF models, one for

515    cloud masking and one for phase, such that a "clear" or "cloudy" label is given first by the cloud

516    mask model, while a corresponding "liquid" or "ice" label is assigned to "cloudy" pixels in the

517    cloud phase model. However, two RF models double the training process and require more

518    computing resources in operational applications. Using multiple RF models becomes more

519    impractical if both aerosol and cloudy pixels are considered.

520    **6. Conclusions**

521    Two Machine-Learning Random Forest (RF) models were trained to provide pixel types (i.e.,

522    clear, liquid water cloud, and ice cloud) using VIIRS 750-meter spectral observations. A daytime

523    model that uses NIR, SWIR, and IR bands and an all-day model that only uses IR bands were

524    trained separately. In the training processes, reference pixel labels are from collocated CALIOP

525    Level 2 1 km cloud layer and 5 km aerosol layer products from 2013 to 2016. Careful tests were

526    conducted to optimize model input and configuration. The two RF models were trained for 7

527    different surface types (i.e., ocean/water, forest, cropland, grassland, snow/ice, barren/desert, and

528    shrubland) to improve model performance. In addition to geolocation and solar/satellite geometry

529    information, we found that using 5 NIR and SWIR bands (0.86, 1.24, 1.38, 1.64 and 2.25 $\mu$m) and

530    three IR bands (8.6, 11, and 12$\mu$m) in the daytime RF model and using the three IR bands and

531    surface temperatures in the all-day RF model can achieve the best performances for all surface

532    types.

533    The cloud mask and thermodynamic phase classifications from the two RF models were

534    validated using the collocated CALIOP products in 2017. For daytime cloud mask comparisons

535    over all surface types, the RF daytime model, with a high TPR (0.93 and higher) and low FPR

536    (0.07 and lower), performs best among all models evaluated, including MODIS MYD35 and

537    MODIS/VIIRS CLDMSK products. The RF all-day model works fairly well and is comparable to

538    other products for all surface types, even in daytime when all other products use shortwave

539    observations and it does not. For the nighttime cloud mask, the RF all-day model has the best

540    performance over all products, demonstrating the strong capability of ML-based algorithms for

541    capturing hidden spectral features of clear and cloudy pixels. All nighttime products perform

542    slightly weaker at snow/ice regions. The decline is likely explained by the lack of SWIR bands

543    and the small thermal contrast between the clouds and the surface during the summer nighttime in

544    polar regions. In this case, the ML-based algorithms are not able to compensate for the missing

545    physical signatures.

546     For the daytime cloud thermodynamic phase comparison, we showed that the two RF models

547     are comparable with the MODIS MYD06 OP-Phase product, and are among the top 3 phase

548     algorithms for all surface types. The MODIS MYD06 IR-Phase, VIIRS/MODIS CLDPROP OP-

549     Phase, and CT-Phase have either relatively lower TPRs or higher FPRs over certain surface types,

550     such as shrubland, snow/ice, and barren regions. For nighttime clouds, the RF all-day model works

551     better than both CLDPROP CT-Phase and MYD06 IR-Phase for all surface types.

552     In this study, we have demonstrated the advantages of using ML-based (specifically, RF)

553     models in cloud masking and thermodynamic phase detections. In contrast to traditional methods,

554     manually-defined thresholds and matching conditions are no longer needed. The RF models can

555     be trained and tested for different surface types and using different input variables in a few hours.

556     Meanwhile, the two RF models show better and more consistent performance over different

557     regions and surface types in comparison with existing VIIRS and MODIS products. In the future,

558     more spectral bands and/or spatial patterns can be used to improve pixel classification skills, such

559     as including more pixel types (e.g., dust and smoke). It is convenient to apply RF models or other

560     similar ML-based models to other instruments similar to VIIRS with the help of active instruments.

561     Most importantly, cloud mask and thermodynamic phase products from well-trained RF models

562     can be used to train other instruments in the absence of active sensors. For example, the current

563     RF model based VIIRS cloud mask/phase data can be used as reference to train ML-based models

564     for other instruments, such as MODIS, ABI/AHI, SEVIRI, and airborne instruments.

565     It is also important to emphasize that the model performance is highly reliant on the quality of

566     the training samples and reference labels. For example, in this study, more than 98% of the training

567     data have a VZA less than 50°, leading to an unavoidable bias of cloud phase fractions at large

568     VZAs. Using synthetic training data generated with reliable radiative transfer models could be a

569     possible way to mitigate this artifact.

Atmospheric
Measurement
Techniques
Discussions

583

584

Atmospheric
Measurement
Techniques
Discussions

# Reference:

Ackerman, S. A., Holz, R. E., Frey, R., Eloranta, E. W., Maddux, B. C., and McGill, M., Cloud detection with MODIS. Part II: Validation, *J. Atmos. Oceanic Technol.*, **25,** 1073–1086, doi: 10.1175/2007JTECHA1053.1, 2008.

Ackerman, S. A., Frey, R., Heidinger, A., Li, Y., Walther, A., Platnick, S., Meyer, K., Wind, G., Amarasinghe, N., Wang, C., Marchant, B., Holz, R. E., Dutcher, S., Hubanks, P., EOS MODIS and SNPP VIIRS Cloud Properties: User guide for climate data record continuity Level-2 cloud top and optical properties product (CLDPROP), version 1, 2019.

Baum, B. A., Menzel, W. P., Frey, R. A., Tobin, D. C., Holz, R. E., Ackerman, S. A., Heidinger, A. K., and Yang, P., MODIS cloud-top property refinements for Collection 6, *J. Appl. Meteor. Climatol.*, **51,** 1145-1163, doi: 10.1175/JAMC-D-11-0203.1, 2012.

Brodzik M. J., and Stewart J. S., Near-Real-Time SSM/I-SSMIS EASE-Grid Daily Global Ice Concentration and Snow Extent, Version 5, doi:10.5067/3KB2JPLFPK3R, 2016.

Cao, C., Xiong, J., Blonski, S., Liu, Q., Uprety, S., Shao, X., Bai, Y., and Weng, F., Suomi NPP VIIRS sensor data record verification, validation, and long-term performance monitoring, *J. Geophys. Res. Atmos.*, **118,** 11,664-11,678, doi:10.1002/2013JD020418, 2013.

Cho, H., Nasiri, S. L., and Yang, P., Application of CALIOP Measurements to the Evaluation of Cloud Phase Derived from MODIS Infrared Channels, *J. Appl. Meteor. Climatol.*, **48,** 2169-2180, doi:10.1175/2009JAMC2238.1, 2009.

Dietterich, T. G., Ensemble methods in machine learning. International Workshop on Multiple Classifier Systems, MCS 2000, Lecture Notes in Computer Science, **vol. 1857**, Springer, Berlin, Heidelberg, 2000.

Erwan Scornet, Tuning parameters in random forests. ESAIM: Procs, 60: 144–162, 2018.

Freund, Y., An Adaptive Version of the Boost by Majority Algorithm, in Machine Learning, **43,** 293-318, 2001.

Frey, R. A., Ackerman, S. A., Liu, Y., Strabala, K. I., Zhang, H., Key, J. R., and Wang, X.: Cloud detection with MODIS. Part I: Improvements in the MODIS cloud mask for Collection 5, *J. Atmos. Oceanic Technol.,* **25,** 1057–1072, doi:10.1175/2008JTECHA1052.1, 2008.

Friedman, J. H., Greedy function approximation: a gradient boosting machine, *Ann. Stat.*, **29,** 1189–1232, 2001.

Gelaro, R., et al., The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), *J. Climate*, **30,** 5419–5454, doi:10.1175/JCLI-D-16-0758.1, 2017.

Hall, D. K., and Riggs, G. A., MODIS/Aqua Snow Cover Daily L3 Global 500m SIN Grid, Version 6. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center, doi:10.5067/MODIS/MYD10A1.006, 2016.

Haynes, J. M., Noh, Y. J., Miller, S. D., Heidinger, A., and Forsythe, J. M., Cloud geometric thickness and improved cloud boundary detection with GEOS ABI, 15th Annual Symposium on New Generation Operational Environment Satellite Systems, Phoenix, AZ, 6 - 10 January, 2019.

Atmospheric
Measurement
Techniques
Discussions

624 Heidinger, A. K., Evan, A. T., Foster, M. J., and Walther, A., A naive bayesian cloud-detection
625      scheme derived from CALIPSO and applied within PATMOS-x, *J. Appl. Meteor. Climatol.*,
626      **51**, 1129–1144, doi:10.1175/JAMC-D-11-02.1, 2012.

627 Ho, T. K, The random subspace method for constructing decision forests, *IEEE Trans. Pattern*
628      *Anal. Mach. Intell*. **20,** 832–844, 1998.

629 Holz, R. E., Ackerman, S. A., Nagle, F. W., Frey, R., Dutcher, S., Kuehn, R. E., Vaughan, M. A.,
630      and Baum, B., Global Moderate Resolution Imaging Spectroradiometer (MODIS) cloud
631      detection and height evaluation using CALIOP, *J. Geophys. Res.*, **113,** D00A19,
632      doi:10.1029/2008JD009837, 2008.

633 Hu, X. F., Belle, J. H., Meng, X., Wildani, A., Waller, L. A., Strickland, M. J., and Liu, Y.,
634      Estimating PM2.5 concentrations in the conterminous United States using the random forest
635      approach, *Environmental Science & Technology,* **51,** 6936–6944,
636      doi:10.1021/acs.est.7b01210, 2017.

637 Ji, C. and Ma, S., Combinations of weak classifiers, *IEEE Transactions on Neural Networks*, **8**,
638      32–42, 1997.

639 Joachims, T., Text categorization with support vector machines: Learning with many relevant
640      features. In Proceedings of the 10th European Conference on Machine Learning, 137–142,
641      Springer-Verlag, 1998.

642 Justice C. O., Vermote, E., Privette J., and Sei, A., The Evolution of U.S. Moderate Resolution
643      Optical Land Remote Sensing from AVHRR to VIIRS. Land Remote Sensing and Global
644      Environmental Change, B. Ramachandran, C. Justice, and M. Abrams, Eds., Remote Sensing
645      and Digital Image Processing, **11,** Springer, New York, NY., 781-806, 2011.

646 Latinne, P., Debeir, O., Decaestecker, C., Limiting the number of trees in random forests, in
647      Multiple Classifier Systems, Manchester, U.K. IEEE, **2013**, 178-187, 2001.

648 Lee, T. E., Miller, S. D., Turk, F. J., Schueler, C., Julian, R., Deyo, S., Dills, P., and Wang, S., The
649      NPOESS VIIRS Day/Night Visible Sensor, *Bull. Amer. Meteor. Soc.*, **87,** 191–200,
650      https://doi.org/10.1175/BAMS-87-2-191, 2006.

651 Leo Breiman, Random forests - random features. Technical report, University of California at
652      Berkeley, Berkeley, California, 1999.

653 Levy, R. C., Mattoo, S., Munchak, L. A., Remer, L. A., Sayer, A. M., Patadia, F., and Hsu, N. C.,
654      The Collection 6 MODIS aerosol products over land and ocean, *Atmos. Meas. Tech.*, **6,** 2989–
655      3034, doi:10.5194/amt-6-2989-2013, 2013.

656 Liu, Y., Ackerman, S. A., Maddux, B. C., Key, J. R., and Frey, R. A., Errors in cloud detection
657      over the Arctic using a satellite imager and implications for observing feedback mechanisms,
658      *J. Climate*, **23,** 1894–1907, doi:10.1175/2009JCLI3386.1, 2010.

659 Maddux, B. C., Ackerman, S. A., and Platnick, S., Viewing geometry dependencies in MODIS
660      cloud products, *J. Atmos. Oceanic Technol.*, **27,** 1519–1528,
661      doi:10.1175/2010JTECHA1432.1, 2010.

662 Martins, J. V., Tanré, D., Remer, L., Kaufman, Y., Mattoo, S., and Levy, R., MODIS cloud
663      screening for remote sensing of aerosols over oceans using spatial variability, *Geophys. Res.*
664      *Lett.*, **29**, doi:10.1029/2001GL013252, 2002.

665 Marchant, B., Platnick, S., Meyer, K. G., Arnold, G. T., and Riedi, J., MODIS Collection 6
666    shortwave-derived cloud phase classification algorithm and comparisons with CALIOP,
667    *Atmos. Meas. Tech.*, **9,** 1587–1599, doi:10.5194/amt-9-1587-2016, 2016.

668 McGill, M. J., Yorks, J. E., Scott, V. S., Kupchock, A. W., and Selmer, P. A., The Cloud-Aerosol
669    Transport System (CATS): A technology demonstration on the *International Space Station,*
670    *Proc. SPIE* **9612**, Lidar Remote Sensing for Environmental Monitoring XV, 96120A,
671    doi:10.1117/12.2190841, 2015.

672 Meyer, K. G., Platnick, S., Arnold, G. T., Holz, R. E., Veglio, P., Yorks, J. E., and Wang, C.,
673    Cirrus cloud optical and microphysical property retrievals from eMAS during SEAC4RS using
674    bi-spectral reflectance measurements within the 1.88 μm water vapor absorption band,
675    *Atmospheric Measurement Techniques*, **9** (4), 1743-1753, doi:10.5194/amt-9-1743-2016,
676    2016.

677 Noel, V., Chepfer, H., Chiriaco, M., and Yorks, J.: The diurnal cycle of cloud profiles over land
678    and ocean between 51° S and 51° N, seen by the CATS spaceborne lidar from the International
679    Space Station, *Atmos. Chem. Phys.*, **18,** 9457–9473, doi:10.5194/acp-18-9457-2018, 2018.

680 Oshiro T. M., Perez P. S., Baranauskas J. A., How many trees in a random forest, in Machine
681    Learning and Data Mining in Pattern Recognition. MLDM 2012. Lecture Notes in Computer
682    Science, **7376,** Springer, Berlin, Heidelberg, 2012.

683 Pedregosa, F. et al., Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830,
684    2011.

685 Platnick, S., Meyer, K. G., King, M. D., Wind, G., Amarasinghe, N., Marchant, B., Arnold, G. T.,
686    Zhang, Z., Hubanks, P. A., Holz, R. E., Yang, P., Ridgway, W. L., Riedi, J.: The MODIS cloud
687    optical and microphysical products: Collection 6 updates and examples from Terra and Aqua,
688    *IEEE Transactions on Geoscience and Remote Sensing*, **55,** 502-525, doi:
689    10.1109/TGRS.2016.2610522, 2017.

690 Remer, L. A., Kaufman, Y. J., Tanré, D., Mattoo, S., Chu, D. A., Martins, J. V., Li, R., Ichoku, C.,
691    Levy, R. C., Kleidman, R. G., Eck, T. F., Vermote, E., and Holben, B. N., The MODIS aerosol
692    algorithm, products, and validation, *J. Atmos. Sci.*, **62**, 947-973, doi:10.1175/JAS3385.1, 2005.

693 Sassen, K., and Cho, B. S., Subvisual-thin cirrus lidar dataset for satellite verification and
694    climatological research, *American Meteorological Society*, **31,** 1275–1285.
695    http://doi.org/10.1175/1520-0450(1992)031<1275:STCLDF>2.0.CO;2, 1992.

696 Sayer, A. M., Munchak, L. A., Hsu, N. C., Levy, R. C., Bettenhausen, C., and Jeong, M.-J., MODIS
697    Collection 6 aerosol products: Comparison between Aqua's e-Deep Blue, Dark Target, and
698    "merged" data sets, and usage recommendations, *J. Geophys. Res. Atmos.*, **119,** 13,965-13,989,
699    doi:10.1002/2014JD022453, 2014.

700 Sayer, A. M., Hsu, N. C., Lee, J., Bettenhausen, C., Kim, W. V., and Smirnov, A., Satellite Ocean
701    Aerosol Retrieval (SOAR) algorithm extension to S-NPP VIIRS as part of the "Deep Blue"
702    aerosol project, *J. Geophys. Res. Atmos.*, **123,** doi:10.1002/2017JD027412, 2017.

703 Seemann, S. W., Borbas, E. E., Knuteson, R. O., Stephenson, G. R., and Huang, H., Development
704    of a global infrared land surface emissivity database for application to clear sky sounding
705    retrievals from multispectral satellite radiance measurements, *J. Appl. Meteor. Climatol.*, **47,**
706    108–123, 2008.

Atmospheric
Measurement
Techniques
Discussions

707 Stephens, G. L., et al., The CloudSat mission and the A-Train: A new dimension of space-based
708    observations of clouds and precipitation, *Bull. Amer. Meteorol. Soc.*, **83,** 1771-1790,
709    doi:10.1175/BAMS-83-12-1771, 2002.

710 Stubenrauch, C. J., Rossow, W. B., Kinne, S., Ackerman, S., Cesana, G., Chepfer, H., Di
711    Girolamo, L., Getzewich, B., Guignard, A., Heidinger, A., Maddux, B. C., Menzel, W. P.,
712    Minnis, P., Pearl, C., Platnick, S., Poulsen, C., Riedi, J., Sun-Mack, S., Walther, A., Winker,
713    D., Zeng, S., and Zhao, G., Assessment of Global Cloud Datasets from Satellites: Project and
714    Database Initiated by the GEWEX Radiation Panel, *Bull. Amer. Meteor. Soc.*, **94,** 1031–1049,
715    doi:10.1175/BAMS-D-12-00117.1, 2013.

716 Sulla-Menashe, D., and Friedl, M. A., User Guide to Collection 6 MODIS Land Cover (MCD12Q1
717    and MCD12C1) Product; USGS: Reston, VA, USA, 2018.

718 Tanelli, S., Durden, S. L., Im, E., Pak, K., Reinke, D., Partain, P., Haynes, J., and Marchand, R.,
719    CloudSat's cloud profiling radar after two years in orbit: Performance, calibration, and
720    processing,    *IEEE    Trans.    Geosci.    Remote    Sens.*,    **46,**    3560–3573,
721    doi:10.1109/TGRS.2008.2002030, 2008.

722 Thampi, B. V., Wong, T., Lukashin, C., and Loeb, N. G., Determination of CERES TOA fluxes
723    using machine learning algorithms. Part I: Classification and retrieval of CERES cloudy and
724    clear scenes, *J. Atmos. Oceanic Technol.*, **34,** 2329–2345, doi:10.1175/JTECH-D-16-0183.1,
725    2017.

726 Tumer, K., and Ghosh, J., Error correlation and error reduction in ensemble classifiers, *Connection
727    Science,* **8**, 385-403, doi:10.1080/095400996116839, 1996.

728 Wan, Z., Zhang, Y., Zhang, Q., and Li, Z.-L., Quality assessment and validation of the MODIS
729    global    land    surface    temperature,    *Int.    J.    Remote    Sens.*,    **25**,    261–274,
730    doi:10.1080/0143116031000116417, 2004.

731 Wang, C., Yang, P., Dessler, A., Baum, B. A., and Hu, Y., Estimation of the cirrus cloud scattering
732    phase function from satellite observations, *Journal of Quantitative Spectroscopy and Radiative
733    Transfer*, **138**, 36-49 doi:10.1016/j.jqsrt.2014.02.001, 2014.

734 Wang, C., Platnick, S., Zhang, Z., Meyer, K., and Yang, P., Retrieval of ice cloud properties using
735    an optimal estimation algorithm and MODIS infrared observations: 1. Forward model, error
736    analysis,   and   information   content,   *J.   Geophys.   Res.   Atmos.*,   **121,**   5809-5826
737    doi:10.1002/2015jd024526, 2016a.

738 Wang, C., Platnick, S., Zhang, Z., Meyer, K., Wind, G., and Yang, P., Retrieval of ice cloud
739    properties using an optimal estimation algorithm and MODIS infrared observations: 2.
740    Retrieval evaluation, *J. Geophys. Res. Atmos.*, **121**, doi:10.1002/2015jd024528, 2016b.

741 Wang, C., Platnick, S., Fauchez, T., Meyer, K., Zhang, Z., Iwabuchi, H., and Kahn, B. H., An
742    assessment of the impacts of cloud vertical heterogeneity on global ice cloud data records from
743    passive satellite retrievals, *Journal of Geophysical Research: Atmospheres*, **124,** 1578-1595.
744    doi:10.1029/2018JD029681, 2019.

745 Winker, D. M., Tackett, J. L., Getzewich, B. J., Liu, Z., Vaughan, M. A., and Rogers, R. R., The
746    global 3-D distribution of tropospheric aerosols as characterized by CALIOP, *Atmos. Chem.
747    Phys.*, **13,** 3345-3361, doi:10.5194/acp-13-3345-2013, 2013.

Atmospheric
Measurement
Techniques

Open Access

EGU

Discussions

748  Wolters, E. L., Roebeling, R. A., and Feijt, A. J., Evaluation of cloud-phase retrieval methods for
749      SEVIRI on Meteosat-8 using ground-based lidar and cloud radar data, *J. Appl. Meteor.*
750      *Climatol.*, **47,** 1723–1738, doi:10.1175/2007JAMC1591.1, 2008.

751  Wu, Y., de Graaf, M., and Menenti, M., Improved MODIS Dark Target aerosol optical depth
752      algorithm over land: angular effect correction, *Atmos. Meas. Tech.*, **9,** 5575-5589,
753      doi:10.5194/amt-9-5575-2016, 2016.

754  Zhou, Y., Levy, R., Remer, L., Mattoo, S., Espinosa, R., Dust detection and dust aerosol retrieval
755      with non-spherical aerosol models over Oceans within MODIS Dark-Target algorithm, *Atmos.*
756      *Meas. Tech, to be submitted*, 2019.

757

758

759

760

761

762

763

764

765

766

767

768

769

770

Atmospheric
Measurement
Techniques
Discussions
Open Access
EGU

771     Table 1. Existing VIIRS and MODIS cloud mask and phase products used for comparison. Note
772     that MYD35 and MYD06 are the standard MODIS Aqua products, and CLDMSK and CLDPROP
773     are the MODIS Aqua and VIIRS common algorithm continuity products.

774

| Instrument | Cloud Mask | Cloud Phase |
|---|---|---|
| **MODIS** | MYD35 V6.1 | MYD06 IR-Phase V6.1 |
| | | MYD06 OP-Phase V6.1 |
| | CLDMSK V1.0 | CLDPROP CT-Phase V1.0 |
| | | CLDPROP OP-Phase V1.1 |
| **VIIRS** | CLDMSK V1.0 | CLDPROP CT-Phase V1.0 |
| | | CLDPROP OP-Phase V1.1 |

775

776

Atmospheric
Measurement
Techniques
Discussions

777    Table 2: Accuracy scores of RF all-day models based on testing pixels with different inputs and a fixed model
778    configuration (N_Trees = 150 and Max_TreeDepths = 15).

| # Input | Model Input | Ocean | Forest | Shrubland | Crop | Grassland | Barren | Snow/Ice | All Surface |
|---|---|---|---|---|---|---|---|---|---|
| 1 | $BT_{8.6}$, $BT_{11}$, $BT_{12}$, and VZA | 0.90 | 0.89 | 0.88 | 0.88 | 0.88 | 0.88 | 0.87 | 0.89 |
| 2 | $BT_{8.6}$, $BT_{11}$, $BT_{12}$, VZA, and Lat/Lon | 0.92 | 0.90 | 0.90 | 0.91 | 0.90 | 0.90 | 0.88 | 0.91 |
| 3 | $BT_{8.6}$, $BT_{11}$, $BT_{12}$, VZA, and $T_S$ | 0.93 | 0.91 | 0.90 | 0.91 | 0.90 | 0.90 | 0.89 | 0.92 |
| 4 | $BT_{8.6}$, $BT_{11}$, $BT_{12}$, VZA, Lat/Lon, and $T_S$ | 0.93 | 0.92 | 0.90 | 0.92 | 0.91 | 0.91 | 0.89 | 0.92 |
| 5 | $BT_{8.6}$, $BT_{11}$, $BT_{12}$, VZA, $T_S$, and $\varepsilon_S$ | 0.93 | 0.91 | 0.90 | 0.91 | 0.90 | 0.90 | 0.89 | 0.92 |
| 6 | $BT_{8.6}$, $BT_{11}$, $BT_{12}$, VZA, Lat/Lon, $T_S$, and $\varepsilon_S$ | 0.93 | 0.92 | 0.90 | 0.92 | 0.91 | 0.91 | 0.89 | 0.92 |

779

780
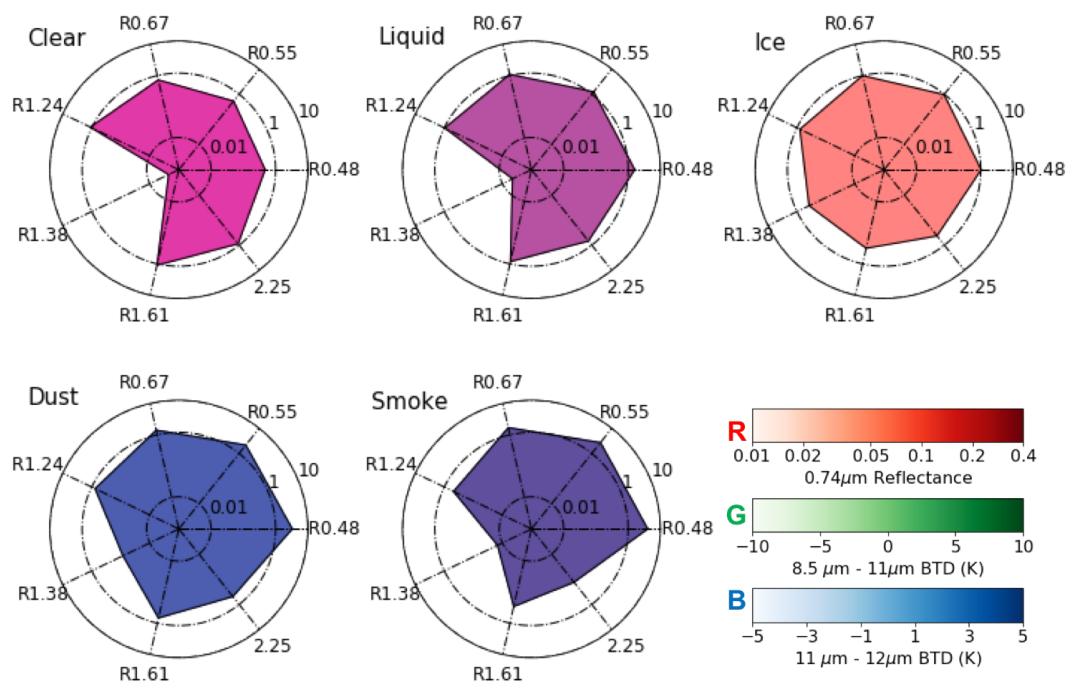
Figure 1. Spectral patterns of the five different pixel types (averaged over 1,000 pixels for each type). For each plot, an apex indicates reflectance ratio between a given VNIR/SWIR band and the 0.86-$\mu$m band, and the spread is filled by false RGB composite (Red: 0.74-$\mu$m reflectance; Green: 8.5-11$\mu$m brightness temperature difference (BTD); Blue: 11-12$\mu$m BTD). The spectral patterns are used in the machine learning algorithms.

786
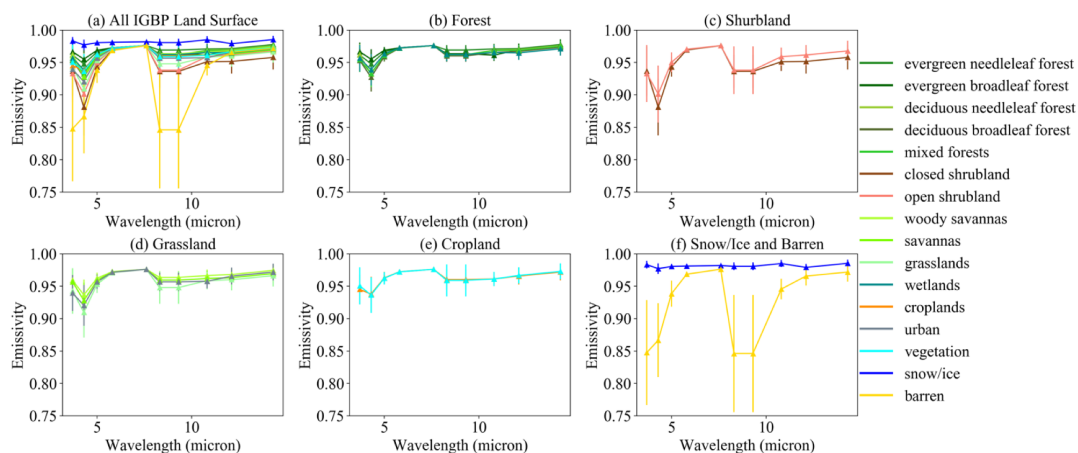
787

788   Figure 2. Climatology of the spectral surface emissivity data from the UW-Madison baseline fit
789   land surface emissivity database [*Seemann et al.*, 2008] for different IGBP surface types. Error
790   bars indicate the emissivity standard deviations at given wavelengths.
791
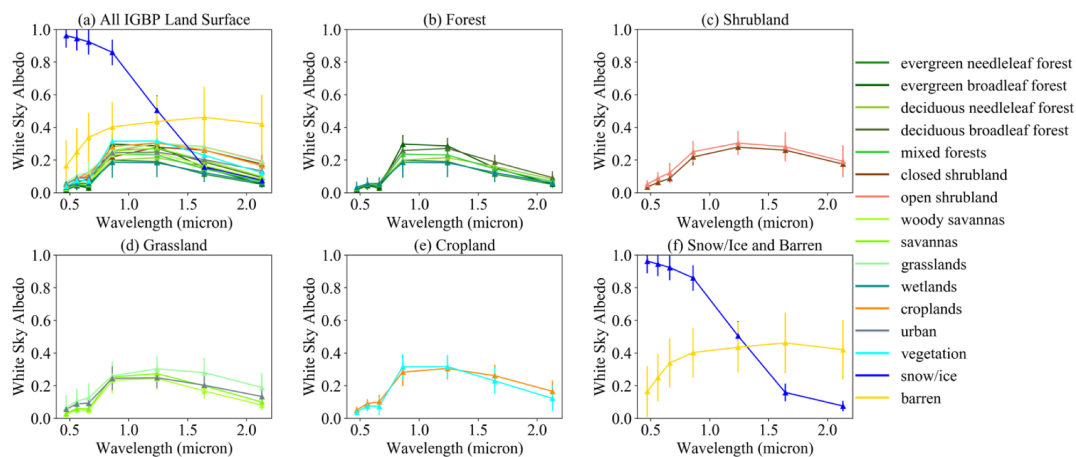
Figure 3. Climatology of the spectral surface white sky surface albedo data from MCD12C1 [*Sulla-Menashe and Friedl* 2018] for different IGBP surface types. Error bars indicate the albedo standard deviations at given wavelengths.
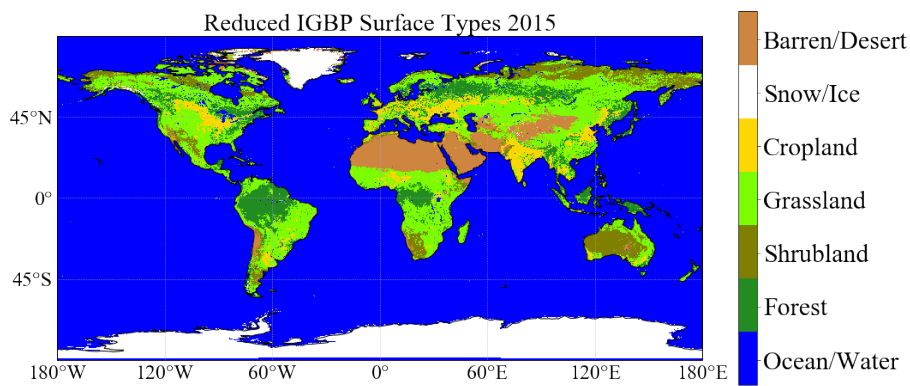
797



798    Figure 4. A global map of the seven reduced surface types chosen for the RF model training.

799
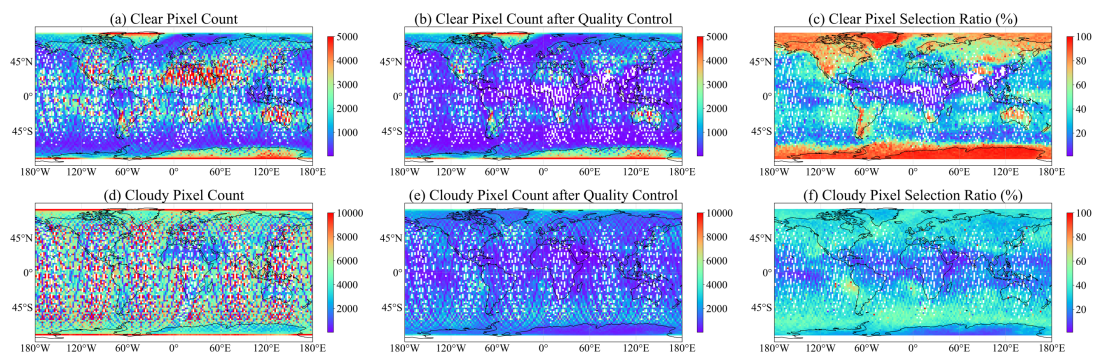
Figure 5. Global distributions of the of clear and cloudy pixels from collocated VIIRS and CALIOP data from 2013 to 2017. Panels a) and d) show the total clear and cloudy pixel counts, respectively. Panels b) and d) show the pixel counts after applying the quality control. The corresponding selection ratios are shown in panels c) and f).

806
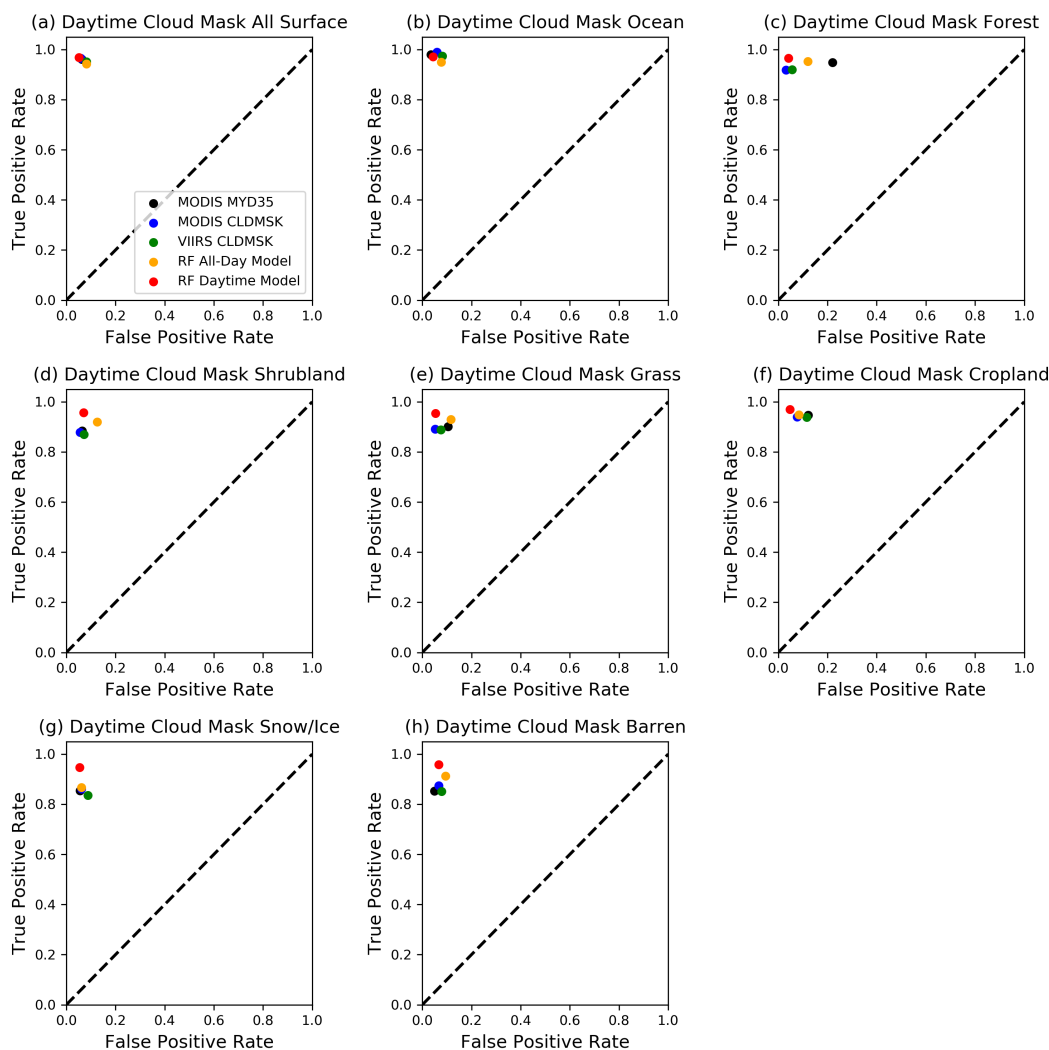
Figure 6. False Positive Rate (FPR) versus True Positive Rate (TPR) plots of daytime cloud mask
from the two RF models and operational algorithms. Collocated CALIOP Level 2 products in 2017
are used as reference. Global comparisons are shown in panel (a), while panels (b) through (h)
show comparisons for difference surface types.
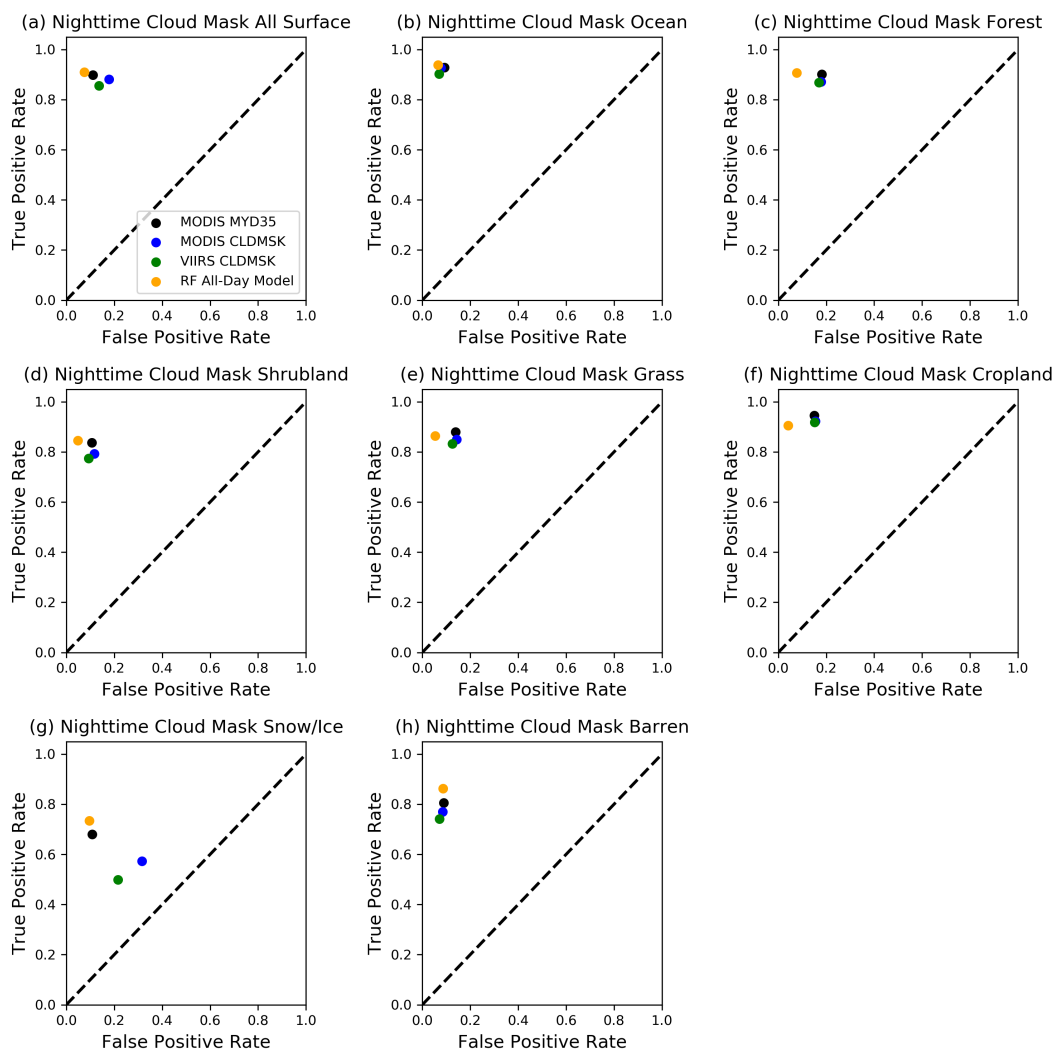
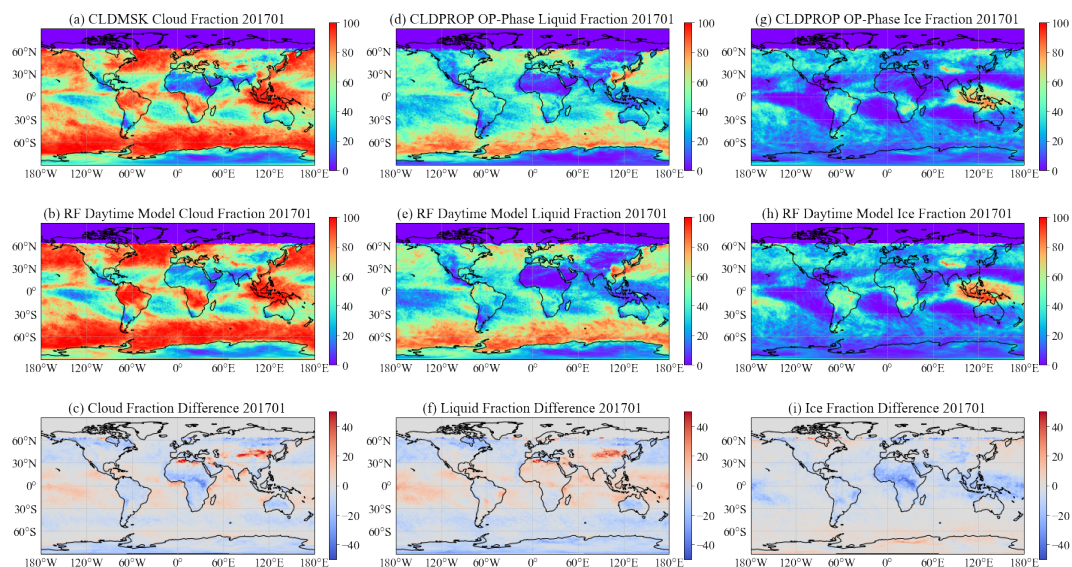Figure 7. Similar to Figure 6, but for nighttime cloud mask comparisons.
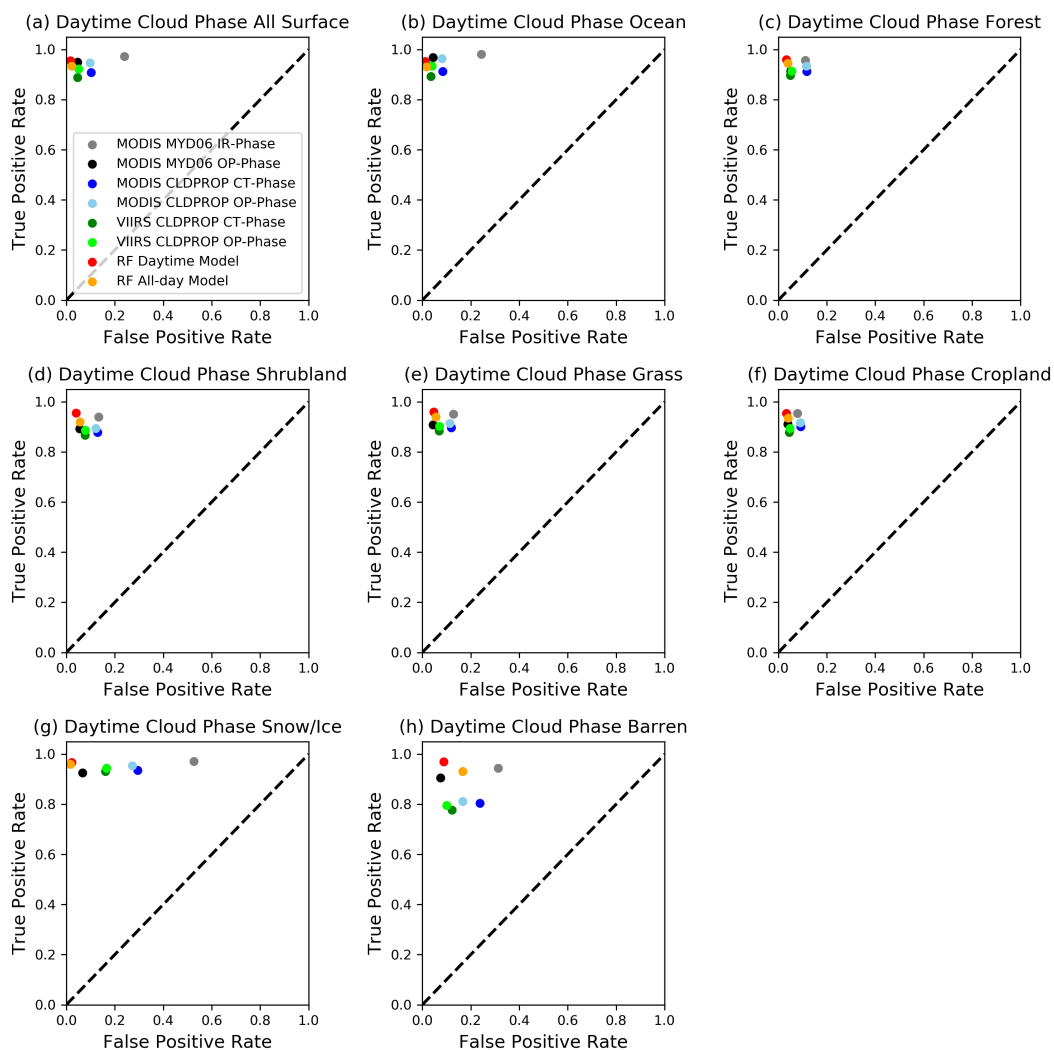
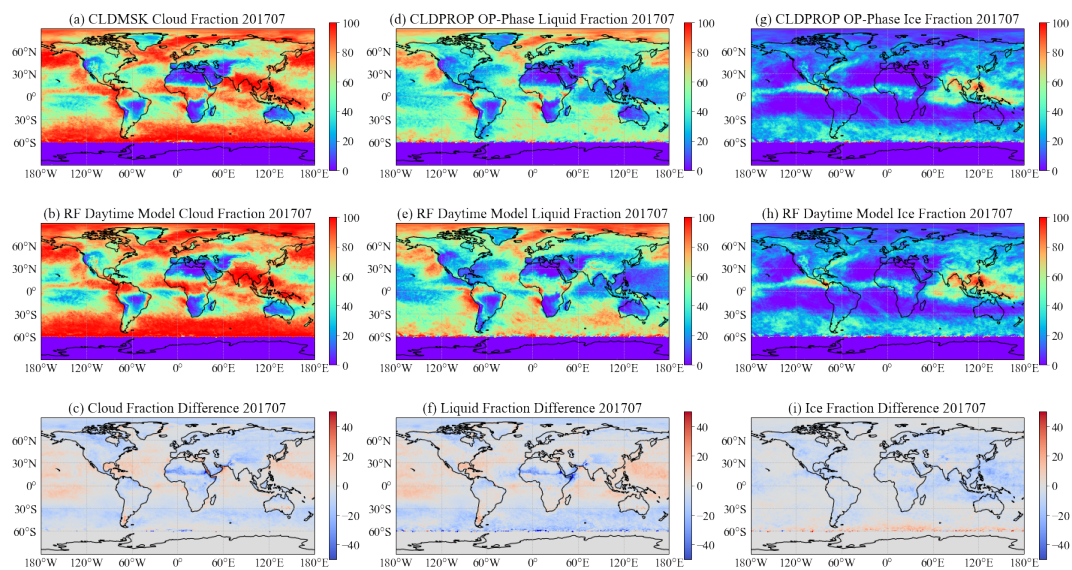Figure 8. Similar to Figure 6, but for daytime cloud thermodynamic phase comparisons.
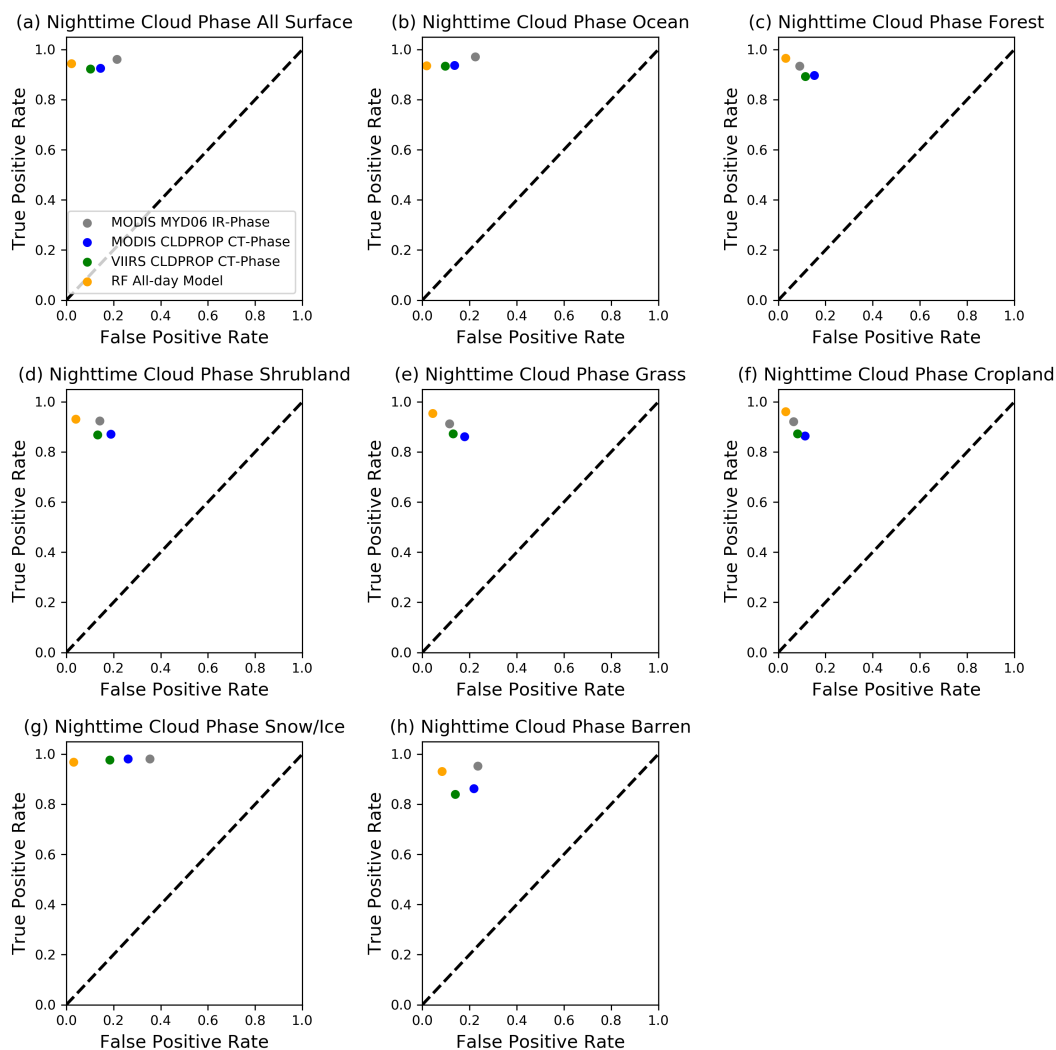
Figure 9. Similar to Figure 6, but for nighttime cloud thermodynamic phase comparisons.

Figure 10. Comparisons between one-month daytime cloud mask and thermodynamic phase products from the VIIRS CLDMSK and CLDPROP OP-Phase (top row) and the RF daytime model (second row), and their differences (VIIRS – RF daytime, bottom row) in January, 2017.

Figure 11. Similar to Figure 10, but for comparisons in July, 2017.

Atmospheric
Measurement
Techniques

Open Access

EGU

Discussions

829
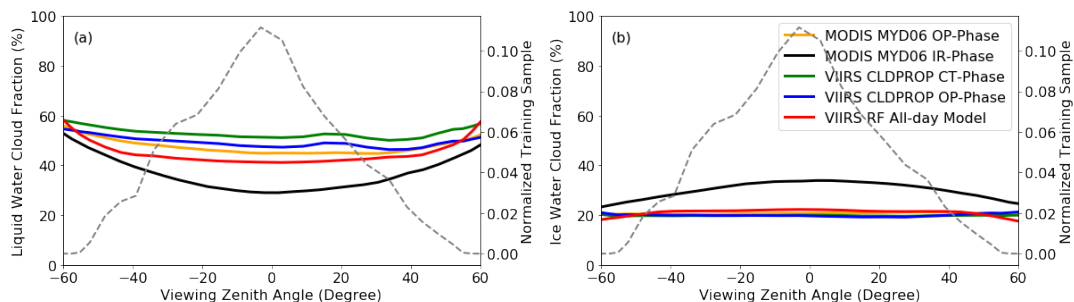
Figure 12. Liquid water (a) and ice (b) cloud fractions as a function of viewing zenith angle from
the one-month daytime cloud mask/phase products in January 2017. The gray dashed curve is the
probability density function of the 4-year VIIRS/CALIOP training samples (2013-2016).

833