

1 **Retrieval of cloud fraction using random forest based on FY4A AGRI**
2 **observations.**

3 Jinyi Xia¹ Li Guan¹

4 ¹China Meteorological Administration Aerosol-Cloud and Precipitation Key
5 Laboratory, Nanjing University of Information Science and Technology, Nanjing
6 210044, China

7 Correspondence to: Li Guan liguan@nuist.edu.cn

8

9 **Abstract**

10 Cloud fraction as a vital component of meteorological satellite products plays an
11 essential role in environmental monitoring, disaster detection, climate analysis, and
12 other research areas. A random forest machine learning algorithm is used in this paper
13 to retrieve the cloud fraction of AGRI (Advanced Geosynchronous Radiation Imager)
14 onboard FY-4A satellite based on its full-disc level-1 radiance observation. Corrections
15 has been made subsequently to the retrieved cloud fraction in areas where solar glint
16 occurs using a correction curve fitted with sun-glint angle as weight. The algorithm
17 includes two steps: the cloud detection is conducted firstly for each AGRI field of view
18 to identify whether it is clear sky, partly cloudy or overcast within the observation field.
19 Then the cloud fraction is retrieved for the scene identified as partly cloudy. The 2B-
20 CLDCLASS-LIDAR cloud fraction product from Cloudsat& CALIPSO active remote
21 sensing satellite is employed as the truth to assess the accuracy of the retrieval algorithm.

22 Comparison with the operational AGRI level 2 cloud fraction product is also conducted
23 at the same time. During daytime, the probability of detection (POD) for clear sky,
24 partly cloudy, and overcast scenes in the operational cloud detection product were
25 0.5359, 0.7041, and 0.7826, respectively. The POD for cloud detection using the
26 random forest algorithm were 0.6984, 0.8971, and 0.8613. While the operational
27 product often misclassified clear sky scenes as cloudy, the random forest algorithm
28 improved the discrimination of clear sky scenes. For partly cloudy scenes, the mean
29 error (ME) and root-mean-square error (RMSE) of the operational product were 0.2374
30 and 0.3269. The random forest algorithm exhibited lower ME (0.1457) and RMSE
31 (0.2022) than the operational product. The large reflectance in the sun-glint region
32 resulted in significant cloud fraction retrieval errors using the random forest algorithm.
33 However, after applying the correction, the accuracy of cloud cover retrieval in this
34 region gets greatly improved. During nighttime, the random forest model demonstrated
35 improved POD for clear sky and partly cloudy scenes compared to the operational
36 product, while maintaining a similar POD value for overcast scenes and a lower FAR.
37 For partly cloudy scenes at night, the operational product exhibited a positive mean
38 error, indicating an overestimation of cloud cover, whereas the random forest model
39 showed a negative mean error, indicating an underestimation of cloud cover. The
40 random forest model also exhibited a lower RMSE compared to the operational product.
41 **Key words:** Cloud detection, cloud fraction, FY-4A AGRI, Random Forest.

42 **Introduction**

43 Clouds occupy a significant proportion within satellite remote sensing data
44 acquired for Earth observation. According to the statistics from the International
45 Satellite Cloud Climatology Project (ISCCP), the annual average global cloud coverage
46 within satellite remote sensing data is around 66% with even higher cloud coverage in
47 specific regions (such as the tropics) (Zhang, et al., 2004). The impact of clouds on the
48 radiation balance of the Earth's atmospheric system is influenced by the optical
49 properties of clouds. Cloud detection, as a vital component of remote sensing image
50 data processing, is considered a critical step for the subsequent identification, analysis,
51 and interpretation of remote sensing images. Therefore, accurately determining cloud
52 coverage is essential in various research domains, such as environmental monitoring,
53 disaster surveillance and climate analysis.

54 Fengyun-4A (FY-4A) is a comprehensive atmospheric observation satellite
55 launched by China in 2016. The uploaded AGRI (Advanced Geosynchronous Radiation
56 Imager) has 14 channels and captures full-disk observation every 15 minutes. In
57 addition to observing clouds, water vapor, vegetation and the Earth's surface, it also
58 possesses the capability to capture aerosols and snow. Moreover, it can clearly
59 distinguish different phases and particle size of clouds and obtain high- to mid-level
60 water vapor content. It is particularly suitable for cloud detection due to its
61 simultaneous use of visible, near-infrared, and long-wave infrared channels for

62 observation with 4km spatial resolution.

63 Numerous cloud detection algorithms have been provided based on observations
64 from satellite-borne imagers. The threshold method has been widely employed by
65 researchers, including the early ISCCP (International Satellite Cloud Climatology
66 Project) method (Rossow, 1993) and the proposed threshold methods based on different
67 spectral features or underlying surfaces (Kegelmeyer, 1994; Solvsteen, 1995; Baum and
68 Trepte, 1996). However, there is a significant subjectivity in selection of thresholds
69 whether it is the single and fixed threshold in the early days, multiple thresholds,
70 dynamic thresholds, or adaptive thresholds. The selection of thresholds is influenced
71 by season and climate. Surface reflectance varies significantly between different
72 seasons, such as increased reflectance from snow in winter and vegetation flourishing
73 in summer affecting reflectance. As a result, changes in surface features during different
74 seasons lead to variations in the distribution of grayscale values in images, requiring
75 adjustments to thresholds based on seasonal characteristics. Climate conditions like
76 cloud cover, atmospheric humidity, etc., impact the distinguishability of clouds and
77 other features. For instance, in humid or cloudy climates, the reflectance of the surface
78 and clouds may be similar, necessitating stricter thresholds for differentiation.
79 Therefore, climate conditions also influence threshold selection.

80 The other category of cloud detection algorithms is based on statistical probability
81 theory. For example the principal component discriminant analysis and quadratic
82 discriminant analysis methods were used to SEVIRI (Spinning Enhanced Visible and

83 Infrared Imager) cloud detection (Amato et al., 2008). The cloud detection algorithm
84 for Thermal Infrared (TIR) sensor was based on the Bayesian theory of total probability
85 (Merchant et al., 2010) and the naive Bayes algorithm for AGRI (Qu , et al., 2022). The
86 unsupervised clustering cloud detection algorithms for MERIS (Medium Resolution
87 Imaging Spectrometer) (GomezChova , et al., 2007) and the fuzzy C-means clustering
88 algorithms for MODIS (Pan, et al., 2009) all have achieved high accuracy in cloud
89 detection.

90 More and more machine learning algorithms are being utilized by researchers in
91 cloud detection studies with the development of machine learning. For instance, the
92 probabilistic neural networks, especially radial basis function networks was used for
93 AVHRR cloud detection (Zhang, et al., 2001). The utilization of convolutional neural
94 network methods (Hu, et al., 2020) offers important perspectives for cloud detection
95 research.

96 Currently, there is limited research literature on cloud detection and cloud fraction
97 retrieval algorithms for FY-4A/4B AGRI. The operational cloud fraction product of FY-
98 4A AGRI utilized a threshold method with 4 km spatial resolution. Differences in
99 climatic and environmental factors lead to varying albedo and brightness temperature
100 observations for the instrument at different times and locations. Therefore, the choice
101 of thresholds is easily influenced by factors such as season, latitude and land surface
102 type (Gao and Jing, 2019). Using multiple sets of thresholds for discrimination would
103 significantly slow down the cloud detection process. Moreover, most algorithms focus

104 solely on cloud detection, which classified the observed scenes into cloud or clear-sky
105 without providing the specific cloud fraction information for the scenes. The use of
106 active remote sensing instruments carried by Cloudsat & Calypso is not influenced by
107 thresholds when retrieving cloud fraction, enabling a more accurate cloud fraction
108 retrieval. However, due to Cloudsat & Calypso being polar-orbiting satellites, the cloud
109 fraction over the full disk cannot be obtained. Utilizing the Cloudsat & Calypso Level
110 2 product 2B-CLDCLASS-LIDAR as the reference truth, a random forest model trained
111 based on FY4A AGRI full disk radiation data can address the shortcomings of threshold
112 methods and achieve a high accuracy of cloud fraction over the full disk. Moreover, the
113 parallel processing during training, randomness in feature selection, and random
114 sampling of samples in random forest make it have a faster training speed compared to
115 other algorithms with similar performance.

116 In summary, a random forest machine learning algorithm for cloud fraction
117 retrieval was established using level-1 radiation observations from FY-4A AGRI full-
118 disk scanning in this paper. The cloud fraction of the level-2 product 2B-CLDCLASS-
119 LIDAR from Cloudsat&CALIPSO was used as the reference label. The retrievals were
120 compared against with the cloud fraction of 2B-CLDCLASS-LIDAR and the AGRI
121 operational products to verify the algorithm accuracy.

122 **1 Research Data and Preprocessing**

123 **1.1 FY-4A data**

124 FY-4A was successfully launched on December 11, 2016. Starting from May 25, 2017,
125 FY-4A drifted to a position near the main business location of the Fengyun
126 geostationary satellite at 104.7 degrees east longitude on the equator. Its successful
127 launch marked the beginning of a new era for China's next-generation geostationary
128 meteorological satellites as an advanced comprehensive atmospheric observation
129 satellite. The Advanced Geosynchronous Radiation Imager (AGRI), one of the main
130 payloads of the Fengyun-4 series geostationary meteorological satellites, can perform
131 large-disk scans and rapid regional scans at a minute level. It has 14 observation
132 channels in total with the main task of acquiring cloud images. The channel parameters
133 and main uses of AGRI are detailed in Table 1
134 (<https://www.nsmc.org.cn/nsmc/cn/instrument/AGRI.html>). FY-4A AGRI data was
135 downloaded from the official website of the China national satellite meteorological
136 center (<http://satellite.nsmc.org.cn>), including level-1 full disk radiation observation
137 data preprocessed through quality control, geolocation and radiation calibration as well
138 as level-2 cloud fraction product (CFR). The spatial resolution of these data is all 4 km
139 at nadir and the temporal resolution is 15 minutes.

140 **Table 1** FY-4A AGRI channel parameters

Channel Number	Band Range / μm	Central Wavelength / μm	Spatial resolution/km	Main Applications
1	0.45 ~ 0.49	0.47	1	clouds, dust, aerosols

2	0.55 ~ 0.75	0.65	0.5	clouds, sand dust, snow
3	0.75 ~ 0.90	0.825	1	vegetation
4	1.36 ~ 1.39	1.375	2	cirrus
5	1.58 ~ 1.64	1.61	2	clouds、 snow
6	2.10 ~ 2.35	2.225	2	cirrus、 aerosols
7	3.50 ~ 4.00	3.75H	2	fire point, the intense solar reflection signal
8	3.50 ~ 4.00	3.75L	4	low clouds, fog
9	5.80 ~ 6.70	6.25	4	upper-level water vapor
10	6.90 ~ 7.30	7.1	4	mid-level water vapor
11	8.00 ~ 9.00	8.5	4	subsurface water vapor
12	10.30 ~ 11.30	10.8	4	surface and cloud-top temperatures
13	11.5 0~ 12.50	12.0	4	surface and cloud-top temperatures
14	13.2 ~ 13.8	13.5	4	cloud-top height

141

142 1.2 CloudSat & Calipso Cloud Product

143 CALIPSO (Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations)

144 is a satellite jointly launched by NASA and CNES (the French National Center for
145 Space Studies) in 2006. It is a member of the A-Train satellite observation system.

146 CALIPSO is equipped with three payloads, among which CALIOP (the Cloud and
147 Aerosol Lidar with Orthogonal Polarization) is a primary observational instrument.

148 Observing with dual wavelengths (532 nm and 1064 nm) CALIOP can provide high-
149 resolution vertical profiles of clouds and aerosols with 30 m vertical resolution. As the

150 first satellite designed to observe global cloud characteristics in a sun-synchronous orbit

151 CloudSat is also among NASA's A-Train series satellites. The CPR (Cloud Profile

152 Radar) installed on it operates at 94 GHz millimeter-wave and is capable of detecting

153 the vertical structure of clouds and providing vertical profiles of cloud parameters. The
154 scanning wavelengths of CPR and CALIOP are different. CALIOP is capable of
155 observing the top of mid-to-high level clouds, whereas CPR can penetrate optically
156 thick clouds. Combining the strengths of these two instruments enables the acquisition
157 of precise and detailed information on cloud layers and cloud fraction.

158 The joint level 2 product 2B-CLDCLASS-LIDAR is mainly utilizing in this study.
159 It provides the cloud fraction at different heights with horizontal resolution 2.5 km
160 (along-track) \times 1.4 km (cross-track) through combining the observations from CPR and
161 CALIOP. Since the two instruments have different spatial domain such as vertical
162 resolution, spatial resolution and spatial frequency, the spatial domain of the output
163 products is defined in terms of the spatial grid of the CPR. In the algorithm, the cloud
164 fraction is calculated using a weighted scheme based on the spatial probability of
165 overlap between the radar and lidar observations. The calculation of the lidar cloud
166 fraction within a radar footprint is represented by the equation 1 (Mace, G. G., et al,
167 2007):

$$168 \quad C_l = \frac{\sum_{i=1}^{\# \text{ of lidar obs}} w_i \delta_i}{\sum_{i=1}^{\# \text{ of lidar obs}} w_i} \quad (1)$$

169 Where:

170 C_l represents the lidar cloud fraction within a radar footprint.

171 w_i is the spatial probability of overlap for a particular lidar observation.

172 δ_i indicates the lidar hydrometeor occurrence, where a value of 1 signifies the
173 presence of hydrometeor and 0 indicates the absence.

174 i counts the lidar profile in a specific radar observational domain.

175 This calculation considers the contributions of multiple lidar observations within
176 a radar resolution volume to determine the cloud fraction within that volume. The
177 CloudSat product manual (Wang, 2019) can be referred for more detailed information
178 on 2B-CLDCLASS-LIDAR. The data used is available to download from the ICARE
179 data and services center ([https://www.icare.univ-lille.fr/data-access/data-archive-
180 access/](https://www.icare.univ-lille.fr/data-access/data-archive-access/)).

181 **1.3 Establishment of Training Data**

182 The crucial aspect of establishing a training data in machine learning algorithms
183 is how to obtain the cloud fraction values (ground truth) as labels. The error in cloud
184 fraction retrieved solely from passive remote sensing instruments is significant. Using
185 active remote sensing data can provide more accurate cloud fraction information in the
186 vertical direction. Therefore, the spatiotemporally matched 2B-CLDCLASS-LIDAR
187 cloud fraction are utilized as output labels in this paper.

188 The FY-4A AGRI and 2B-CLDCLASS-LIDAR data with a spatial difference
189 between fields of view within 1.5 km and a time difference within 15 minutes are
190 spatiotemporal matched. To make the 2B-CLDCLASS-LIDAR cloud fraction data
191 collocated within AGRI pixels more effective, at least two 2B-CLDCLASS-LIDAR
192 pixels are required within each AGRI field of view. The cloud fraction average of these
193 pixels is used as the cloud fraction for that AGRI pixel.

194 Cloud detection and cloud fraction label generation for 2B-CLDCLASS-LIDAR
195 are as follows. There may be multiple layers of clouds in each field of view. If there is
196 at least one layer cloud with cloud fraction of 1 in the 2B-CLDCLASS-LIDAR profile,
197 then the scene is labeled as overcast with a cloud fraction of 1. If all layers in the profile
198 are cloud-free, the scene is labeled as clear sky. The scene between the above two
199 situations is labeled as partly cloudy and the cloud fraction is the average of cloud
200 fractions at different layers.

201 The algorithm includes two steps: the cloud detection is conducted firstly for each
202 AGRI field of view to identify whether it is clear sky, partly cloudy or overcast within
203 the observation field. Then the cloud fraction is retrieved for the scene identified as
204 partly cloudy. So the training data include A dataset used for cloud detection and B
205 dataset for cloud fraction retrieval. The input variables in A dataset are the FY-4A
206 AGRI level-1 radiative observations from 14 channels and the output variable is the
207 temporally and spatially matched 2B-CLDCLASS-LIDAR cloud detection label. The
208 output is categorized into three types: overcast, partly cloudy and clear sky with values
209 1, 2 and 3 respectively. To ensure diversity and representativeness of the samples, the
210 three conditions of overcast, partly cloudy, and clear sky each account for one-third of
211 the sample size in dataset A. Regarding the samples for partly cloudy type in dataset A,
212 the collocated 2B-CLDCLASS-LIDAR cloud fraction products serve as output labels
213 for cloud fraction retrieval model B. The input of training dataset B remains the FY-4A
214 AGRI level-1 radiative observations.

215 Due to the instrument's limited lifespan, only 2B-CLDCLASS-LIDAR data up to
216 August 2019 can be obtained. Additionally, the latitude range for a single observation
217 of FY-4A AGRI is $-83.3\sim 83.3$. Within this latitude range, data from different seasons,
218 climates, and surface types are included. In the training samples matched in space-time
219 with 2B-CLDCLASS-LIDAR, seasons and climates vary with latitude. Therefore, there
220 is no need to include data from a larger time range as training samples. The FY-4A
221 AGRI observations and 2B-CLDLASS-LIDAR matched in time and space in May 2019
222 are used as training samples to build the algorithm model. The paired samples of whole
223 June 2019 are served as the testing samples to assess the model's retrieval accuracy. The
224 number of training samples in May are 12,420 for dataset A and 4140 for B. Testing
225 samples in June are 15,459 for A and 5,153 for B.

226 Although the retrieval model was trained and tested using 2019 data, the algorithm
227 was also applied to real-time observations of FY-4A and FY-4B AGRI in 2023 to verify
228 its universality.

229

230 **2. Random Forest Algorithm**

231 The random forest algorithm integrates multiple trees based on the Bagging idea
232 of ensemble learning, with the basic element being the decision tree (Breiman, 1999).
233 When building a decision tree, N sets of independent and dependent variables are

234 randomly sampled with replacement from the original training samples to create a new
235 training sample set; m variables are randomly sampled without replacement from all
236 independent variables, the dependent variable data is split into two parts using the
237 selected variables, and the purity of the subsets is calculated for each split method. The
238 variable utilized by the split method with the highest purity is used to partition the data,
239 completing the decision at that node. This process of binary splitting continues to grow
240 the decision tree until stopping criteria are met, completing the construction of a single
241 decision tree. These steps are repeated N_{tree} times to build a random forest model
242 consisting of N_{tree} decision trees (Quesada-Ruiz et al., 2022). Random Forest adopts
243 ensemble algorithms, with the advantage of high accuracy. It can handle both discrete
244 and continuous data, without the need for normalization, making it more efficient
245 compared to other algorithms.

246 In this study, when using the trained model for prediction, observations from 14
247 channels are inputted into the model. Each decision tree independently predicts the
248 outcome, with a majority vote determining the final classification category of overcast,
249 partly cloudy, or clear sky. For regression tree models, the average of all tree outputs is
250 taken as the final output, representing the specific cloud fraction.

251 Two crucial parameters in the random forest model are the node splitting
252 frequency M_{try} and the number of decision trees N_{tree} , which directly impact the
253 model's performance. A high M_{try} value can increase model complexity, leading to
254 overfitting; conversely, a low M_{try} can result in a model that is too simple and underfits

255 the data. A small Ntree value can result in underfitting, while a large Ntree significantly
256 increases computational load, with minimal performance improvement beyond a
257 certain threshold. Typically, setting Mtry to \sqrt{M} , where M represents the number of
258 input variables, results in the lowest model error. For daytime models, M is 14, while
259 for nighttime, it is 8. Mtry is set at 3 for daytime cloud detection and cloud fraction
260 retrieval models, and at 2 for nighttime models. When determining the size of Ntree, it
261 is necessary to do so through cross-validation. The dataset is divided into training and
262 validation sets, using a different number of trees in each training iteration, and then
263 evaluating the model's performance on the validation set. The best number of trees is
264 selected by comparing the performance of the model with different numbers of trees.
265 Both daytime and nighttime cloud detection models are configured with Ntree set to
266 380, while cloud fraction retrieval models have Ntree set to 300 for both daytime and
267 nighttime scenarios.

268 **3. Results and Analysis**

269 To assess the accuracy and stability of the retrieval model, two types of validation
270 methods are utilized. One way involves a direct comparison from images, qualitatively
271 comparing the model's retrieval results and official cloud fraction products with AGRI
272 observed cloud images. Another way is quantitative comparison using 2B-
273 CLDCLASS-LIDAR as the true value. Four quantitative parameters, including

274 possibility of detection (POD), false alarm rate (FAR), mean error (ME), and root mean
275 square error (RMSE) are introduced. The POD is calculated using the formula
276 $POD = TP / (TP + FN)$, and the FAR is calculated using the formula $FAR = FP / (TP + FP)$.
277 Taking the overcast scenes as an example, TP represents the number of correctly
278 identified overcast, FN represents the number of overcast scenes wrongly identified as
279 partly cloudy or clear sky, and FP represents the number of clear sky or partly cloudy
280 scenes wrongly identified as overcast. The ME (mean error) and RMSE (root mean
281 square error) are utilized to assess the accuracy of the random forest cloud fraction
282 model in retrieving cloud fractions for partly cloudy scenes.

283 **3.1 Objective Analysis of Cloud Fraction Retrievals**

284 The test samples from dataset A (i.e., June data) are used to perform cloud
285 detection experiments based on the cloud detection model mentioned above. The
286 temporally and spatially matched 2B CLDCLASS-LIDAR cloud mask products are
287 used as reference to evaluate the accuracy of cloud detection. The POD and FAR for
288 different view field classifications are shown in Table 2. Columns 2 and 4 represent the
289 operational cloud detection products for daytime and nighttime respectively, for the
290 same time and pixel. Columns 3 and 5 represent the random forest cloud detection
291 results for daytime and nighttime respectively. The table indicates that during daytime,
292 operational cloud detection products have a relatively low possibility of detection for
293 clear sky view fields. However, the random forest model increases the possibility of

294 detection for clear sky from 0.54 to 0.70. Moreover, for partly cloudy and overcast view
 295 fields, the POD is higher than operational cloud detection products. During nighttime,
 296 compared to operational cloud detection products, the random forest model increases
 297 the POD for clear sky from 0.51 to 0.67, with higher POD for partly cloudy view fields
 298 compared to the operational products, while the POD for overcast view fields is lower.
 299 During the day, the Operational product has a lower FAR for clear sky compared to the
 300 random forest model, while the random forest model has a lower FAR for partly cloudy
 301 and overcast conditions compared to the operational product. At night, the random
 302 forest model significantly reduces the FAR for overcast conditions compared to the
 303 Operational product.

304 **Table 2** POD and FAR of Cloud Detection

	Sky Classification	Daytime Operational Product	Daytime RF Results	Nighttime Operational Product	Nighttime RF Results
PO D	Clear Sky	0.5359	0.6984	0.5136	0.6733
	Partly cloudy	0.7041	0.8971	0.6957	0.7438
	Overcast	0.7826	0.8613	0.7984	0.7979
FAR	Clear Sky	0.2174	0.2431	0.1789	0.2016
	Partly cloudy	0.2959	0.1754	0.3107	0.2847
	Overcast	0.4641	0.2766	0.5543	0.3331

305

306 For the field identified as partly cloudy by the previous model, the random forest
 307 cloud fraction model established in the preceding text is used to retrieve the cloud
 308 fraction in the AGRI field. For samples classified as partly cloudy by the model, and

309 operational products, and 2B-CLDCLASS-LIDAR cloud fraction products, the mean
310 error and root mean square error (RMSE) of the cloud fraction retrieval were calculated
311 based on the matched 2B-CLDCLASS-LIDAR cloud fraction product as ground truth,
312 separately for daytime and nighttime operational cloud fraction products (columns 2
313 and 4) and the random forest- retrieved cloud fraction (columns 3 and 5), as shown in
314 Table 3. It can be observed that during daytime, compared to the FY-4A operational
315 cloud fraction product, the random forest cloud fraction retrieval model shows
316 significant improvement in both ME and RMSE. The ME decreases from 0.23 to 0.11,
317 and the RMSE decreases from 0.32 to 0.15, indicating that the random forest cloud
318 fraction retrieval model provides more accurate estimates of cloud fraction. For
319 nighttime, the ME of the operational cloud fraction product is positive, indicating an
320 overall overestimation of cloud fraction. In contrast, the ME of the random forest model
321 is negative, indicating an overall underestimation of cloud fraction. The RMSE of the
322 random forest model retrieval results during nighttime is lower than that of the
323 operational cloud fraction product.

324 **Table 3** Errors in cloud fraction retrieval

	Daytime Operational Product	Daytime RF Results	Nighttime Operational Product	Nighttime RF Results
ME	0.2374	0.1457	0.2488	-0.1984
RMSE	0.3269	0.2022	0.3374	0.2434

325 3.2 Cloud fraction correction in sun glint regions

326 Sun glint refers to the bright areas created by the reflection of sunlight to the
327 sensors of observation systems (satellites or aircrafts). This phenomenon usually occurs
328 on extensive water surfaces, such as oceans lakes or rivers. This specular reflection of
329 sunlight will cause an increase in the reflected solar radiation received by onboard
330 sensors, manifested as an enhancement of white brightness in visible images. The
331 increase in visible channel observation albedo will affect various subsequent
332 applications of data, including cloud detection and cloud cover retrieval, etc.

333 The position of Sun glint area can be determined using the SunGlintAngle value
334 in the FY-4A GEO file. SunGlintAngle is defined as the angle between the satellite
335 observation direction or reflected radiation direction and the mirror reflection direction
336 on a calm surface (horizontal plane). It is generally accepted that the range of
337 SunGlintAngle $< 15^\circ$ is easily affected by sun glint (Kay S, et al., 2009). The positions
338 of the SunGlintAngle contour lines at 5 and 15° are marked in Figure 1(a). It can be
339 observed that the edge of sun glint in Figure 1(a) essentially overlaps with the position
340 of SunGlintAngle = 15° . Thus, the region where SunGlintAngle $< 15^\circ$ is defined as the
341 sun glint range in this paper and only the cloud fraction within this range will be
342 adjusted in the subsequent correction.

343 To correct the cloud fraction in the sun glint region, we initially identified 672
344 fields of view where sun glint occurred in the FY-4A AGRI observations between 1

345 June and 31 July 2019. Subsequently, a direct least squares fitting was conducted
346 between the retrieved cloud fraction and the collocated 2B-CLDCLASS-LIDAR cloud
347 fraction (ground truth). The scatter plot is illustrated in Figure 1(b), where x-axis is the
348 2B-CLDCLASS-LIDAR cloud fraction and y-axis is the model-retrieved cloud fraction.
349 The blue line represents the curve (namely Eq.2) fitted by the least squares method
350 between the retrievals and the truths. The thin dash line is the $x=y$ line. It is evident that
351 the retrieved cloud fraction is generally slightly overestimated.

352 Taking observations at 04:00 on 5 June 2019 as an example, Figure 1(c) presents
353 the distribution of SunGlintAngle and the flight trajectory of the Cloudsat&Calypso
354 satellite. White circles denote the sun glint region with SunGlintAngle $< 15^\circ$ and the
355 white line represents the satellite flight track. As depicted in the figure, the majority of
356 Cloudsat&Calypso flight trajectories do not pass through the central position of sun
357 glint area but instead traverse locations with larger SunGlintAngle values. The
358 intensity of sun glint effect decreases with the increase of SunGlintAngle. This
359 suggests that the true values for spatial and temporal matching mostly do not fall within
360 the strongest sun glint region. From Figure 1(d), it can be seen that the impact of sun
361 glint becomes stronger as SunGlintAngle decreasing, which results in a higher
362 observation albedo. This further leads to the overestimated cloud fraction values in the
363 retrieval. It is evident that the cloud fraction error is related to the value of
364 SunGlintAngle and this influence is not considered in Eq. (2). Directly applying
365 equation (2) to correct the cloud fraction retrievals would result in a too small correction

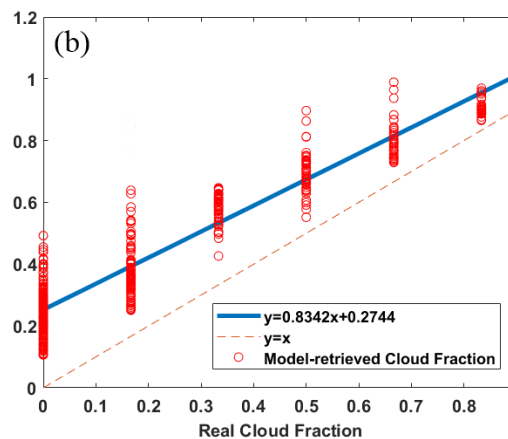
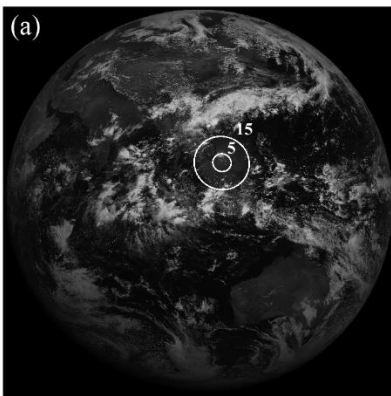
366 intensity for the FOVs near the center of sun glint and an excessively large correction
 367 intensity for the FOVs in the Sun-glint edge region (even erroneous clear sky may
 368 appear). Considering this, a correction formula (3)-(4) using SunGlintAngle as weight
 369 is introduced, where W_i represents the angle weight for a certain pixel i in the sun glint
 370 region, n is the number of pixels within the SunGlintAngle $< 15^\circ$ range, y_i is the initial
 371 model retrieval of cloud cover for the field of view i and x_i is the final corrected cloud
 372 fraction.

$$373 \quad x = (y - 0.2744)/0.8342 \quad (2)$$

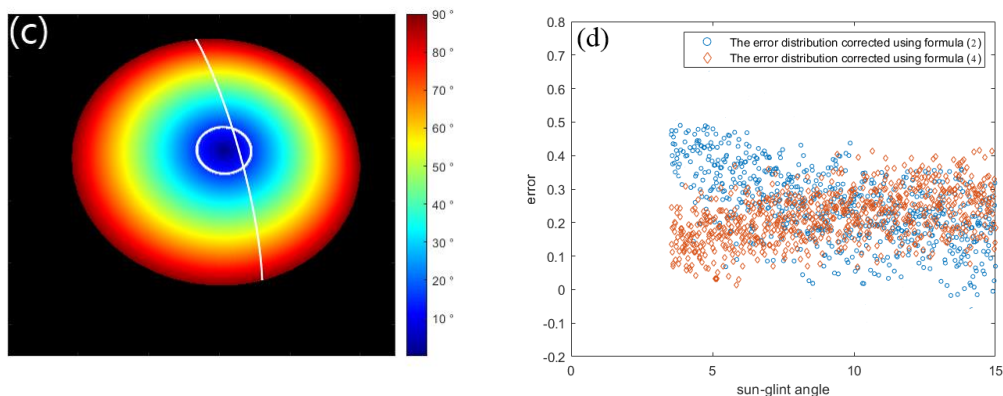
$$374 \quad W_i = \frac{glintangle_i}{\frac{1}{n}\sum_{i=0}^n glintangle_i} \quad (3)$$

$$375 \quad x_i = W_i \left(\frac{y_i - 0.2744}{0.8342} \right) \quad (4)$$

376 Figure 1(d) shows the distribution of errors with respect to SunGlintAngle,
 377 where the blue dots represent the error distribution corrected using formula
 378 (2), and the orange dots represent the error distribution corrected using
 379 formula (4). It can be seen from Figure 1(d) that after correction by formula
 380 (4), the errors in the smaller range of SunGlintAngle are significantly reduced.



381



382

383 **Figure 1** (a) albedo image of 0.67μm channel (the circles are the contours of the sun-

384 glint angle), (b) Scatter plot of cloud fraction in sun glint region (The blue line

385 represents the curve (namely Eq.2) fitted by the least squares method between the

386 retrievals and the truths.), (c) Distribution of SunGlintAngle and satellite flight track of

387 CloudSat & Calypso at 4:00 on June 5, 2019, (d) Distribution of cloud fraction retrieval

388 error with sun-glint angle.

389 3.3 Algorithm universal applicability testing

390 Although the retrieval model in this article was built based on data from May 2019

391 due to the limited lifespan of the instrument, how effective is it in real-time FY-4A

392 AGRI observations and even subsequent FY-4B AGRI applications? The algorithm's

393 universal applicability was tested using real-time observations from FY-4A and FY-4B

394 AGRI in 2023.

395 Taking the full-disk observation of FY-4A AGRI at 04:00 (UTC, the same below)

396 on 1 June 2023 as an example, the radiance observations from 14 channels are initially

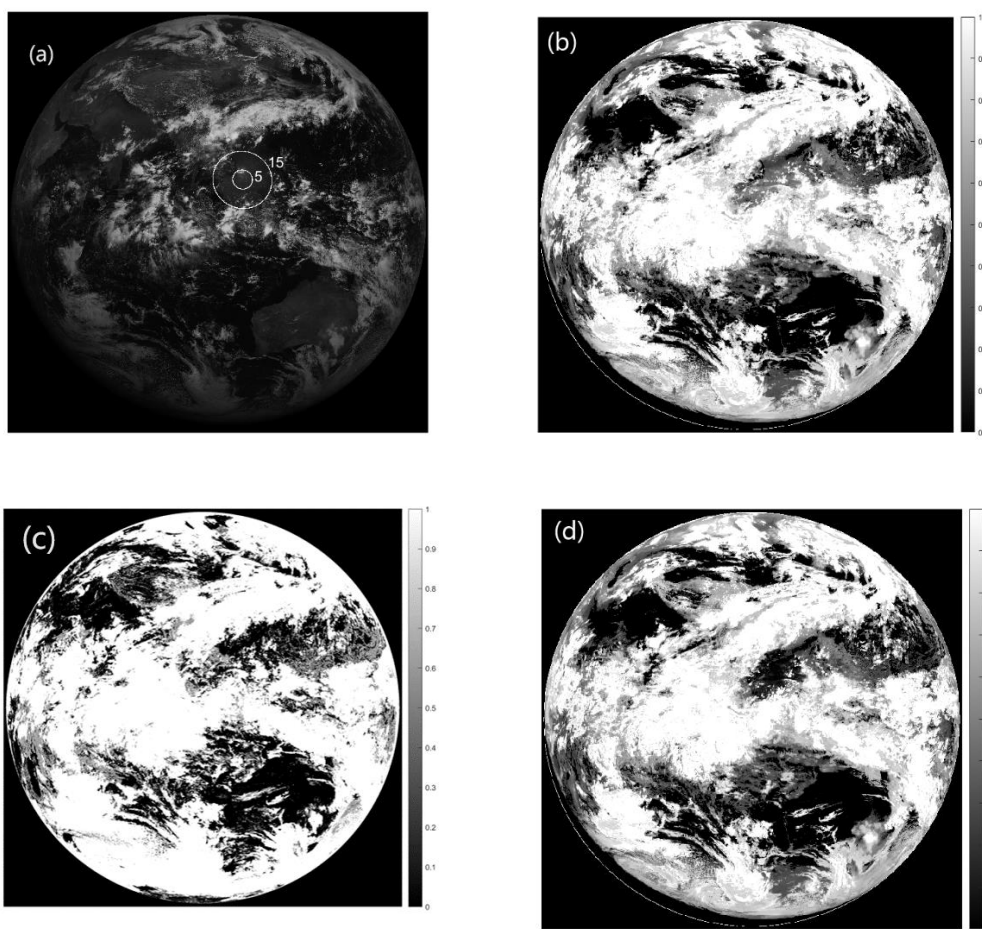
397 fed into the random forest cloud detection model to determine the sky classification

398 (overcast, partly cloudy or clear sky) in each AGRI field. The random forest cloud
399 fraction retrieval model is utilized to retrieve the cloud fraction in scenes identified as
400 partly cloudy. Figure 2(a) is the observed albedo at $0.67\ \mu\text{m}$, where the circles represent
401 the contours of the sun glint angle, (b) is the cloud fraction retrievals from random forest
402 algorithm, (c) is the official operational cloud fraction product and (d) is random forest
403 cloud fraction retrievals with sun-glnt correction. It can be seen from Figure 2 that
404 many clear-sky scenes are erroneously identified as cloudy by the operational product
405 and the cloud fraction is generally overestimated with many scenes having a cloud
406 fraction of 1. The random forest algorithm identifies more regions as clear skies or
407 partly cloudy than the operational products, matching better with the observations in
408 the $0.67\ \mu\text{m}$ albedo image. Brighter regions in the visible image correspond to cloud
409 cover areas and darker areas represent clear sky conditions. The sun glint region in the
410 central South China Sea (the circled area in Figure 2(a)) is depicted in Figure 2(b),
411 where the clear-sky scenes over the ocean are misidentified as partly cloudy by random
412 forest algorithm due to the increase in observed albedo. Although operational product
413 in this area also suffers from the impact of unremoved sun glint, it identifies more clear-
414 sky scenes and the cloud fraction is relatively low. Thus, it is evident that the random
415 forest algorithm exhibits significant cloud detection and cloud fraction errors in these
416 sun glint regions. Correction is necessary for the cloud fraction retrievals in the sun
417 glint region.

418 Figure 2(d) shows the cloud fraction distribution after correction using equation

419 (9) in the sun glint region., The correction eliminates the influence of sun glint
420 comparing to the cloud fraction in sun glint area before correction in Figure 2(b). The
421 scenes misjudged as partly cloudy are corrected to clear sky and match well with the
422 actual albedo observations in 2(a), which accurately restores the true cloud coverage
423 over the South China Sea.

424



425

426

427 **Figure 2** FY-4A AGRI at 04:00 on 1 June 2023 (a) albedo image of 0.67 μm channel
428 (the circles are the contours of the sun-glint angle), (b) random forest cloud fraction
429 retrieval without sun-glint correction, (c) operational cloud fraction product, (d)

430 random forest cloud fraction retrieval with sun-glnt correction.

431 Statistical analysis was conducted on the correction effect using samples with sun
 432 glint in the training data. The POD and FAR in sun glint area is listed in table 4 and the
 433 error is in table 5. The POD for clear skies has increased from 0.11 to 0.84. The FAR
 434 for partly cloudy has decreased from 0.9 to 0.2. The mean error of cloud fraction
 435 retrievals decreased from 0.398 to 0.136. These all indicate that the positive effect of
 436 the sun glint correction.

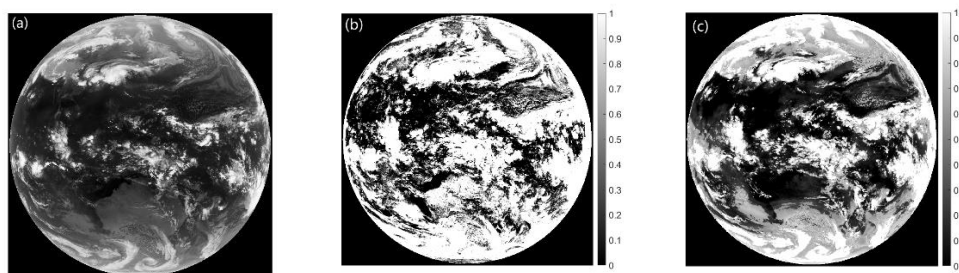
437 **Table 4** POD and FAR of Cloud Detection in sun glint area

	Sky Classification	Operational Product	RF	RF after Correction
POD	Clear Sky	0.5535	0.1137	0.8443
	Partly cloudy	0.6738	0.8342	0.7677
	Overcast	0.8505	0.9498	0.9498
FAR	Clear Sky	0.1437	0.0120	0.2354
	Partly cloudy	0.3742	0.9077	0.2019
	Overcast	0.5545	0.0745	0.0745

438 **Table 5** cloud fraction Errors in sun glint area

	Operational Product	RF Retrievals	RF after Correction
ME	0.2691	0.3987	0.1365
RMSE	0.3458	0.3774	0.1639

440 FY-4B launched in 2021 has a total of 15 channels with an additional low-level
441 water vapor channel at 7.42 μm compared to FY-4A. Taking the full-disk observation
442 of FY-4B AGRI at 17:00 on April 18, 2023, as an example, The radiance observation
443 data of the remaining eight channels (near-infrared and infrared channels) except for
444 the 7.42 μm channel and the visible light channels were input into the random forest
445 cloud detection model. Figure 3 (a) shows the brightness temperature distribution
446 observed in the 10.8 μm channel of FY-4B AGRI, (b) represents the operational cloud
447 fraction product for FY-4B AGRI and (c) shows the cloud fraction retrieved by this
448 algorithm. Figure 3 illustrates that the random forest algorithm identifies more regions
449 as clear skies or partly cloudy than the operational products, aligning better with the
450 brightness temperature observations in 10.8 μm . Especially in high latitude regions of
451 the southern hemisphere and areas with strong convection near the equator, the cloud
452 cover provided by operational products is too high and even misjudged. It can be seen
453 that the random forest algorithm is also suitable for cloud fraction retrieval of FY-4B
454 AGRI.



455
456 **Figure 3** FY-4B AGRI at 17:00 on 18 April 2023, (a) brightness temperature of
457 10.8 μm channel, (b) operational cloud fraction product, (c) random forest cloud

458 fraction retrieval.

459

460 **4 Conclusion**

461 The random forest machine learning algorithm based on FY-4A AGRI full-disc
462 level-1 radiance observations is developed to retrieve the cloud fraction for each field
463 of view in this paper. The accuracy of the algorithm is validated using the 2B
464 CLDCLASS-LIDAR cloud fraction product from the Cloudsat&Calypso active remote
465 sensing satellite and FY-4A AGRI level 2 operational product. The following
466 conclusions are drawn:

467 (1) Not only the cloud detection but also the cloud fraction within each FY-4A
468 AGRI field of view can be retrieved by the random forest machine learning
469 algorithm.

470 (2) The operational product has a relatively low POD for clear sky scenes, while
471 the random forest algorithm improves the POD for clear sky scenes during the
472 daytime from 0.54 to 0.69. The POD for clear sky scenes at night increases
473 from 0.51 to 0.67, and the POD for partly cloudy and overcast scenes is
474 comparable to the operational product.

475 (3) For partly cloudy fields, during the day, the ME and RMSE of the operational
476 product are 0.2374 and 0.3269, respectively, while this algorithm exhibits

477 lower ME (0.1475) and RMSE (0.2022) compared to the operational product.
478 At night, the operational product tends to overestimate cloud cover, while this
479 algorithm underestimates cloud cover, with a lower RMSE compared to the
480 operational product.

481 (4) The cloud fraction correction curve for sun glint region fitted with
482 SunGlintAngle as weight significantly improves the accuracy of the random
483 forest cloud fraction retrievals. It reduces the misjudgment rate where increased
484 albedo leads to the identification of clear-sky scene as partly cloudy or overcast.

485

486 *Data availability*

487 FY-4A AGRI data is available at <http://satellite.nsmc.org.cn> and the 2B-CLDCLASS-
488 LIDAR data at <https://www.icare.univ-lille.fr/data-access/data-archive-access/>

489

490 *Author contributions*

491 JX: Formal analysis, Methodology, Software, Visualization and Writing – original draft
492 preparation. LG: Conceptualization, Data curation, Funding acquisition, Supervision,
493 Validation and Writing – review & editing.

494

495 *Competing interests*

496 The contact author has declared that none of the authors has any competing interests.

497

498 **Disclaimer**

499 **Acknowledgements**

500 Funding: This work was supported by the National Natural Science Foundation of
501 China under grant no. 41975028.

502 *We acknowledge the High Performance Computing Center of Nanjing University of*
503 *Information Science & Technology for their support of this work.*

504 **References**

- 505 Baum, B., Trepte Q.: A Grouped Threshold Approach for Scene Identification in
506 AVHRR Imagery, *Journal of Atmospheric & Oceanic Technology*, 16, 793-800,
507 [https://doi.org/10.1175/1520-0426\(1999\)016<0793:AGTAFS>2.0.CO;2](https://doi.org/10.1175/1520-0426(1999)016<0793:AGTAFS>2.0.CO;2), 1999.
- 508 Breiman L.1999. Random Forests-Random Features [J]. *Machine Learning*.45(1): 5-32.
- 509 Merchant, C.J., Harris, A.R., Maturi, E., Maccallum S.: Probabilistic physically based
510 cloud screening of satellite infrared imagery for operational sea surface temperature
511 retrieval, *Quarterly Journal of the Royal Meteorological Society*, 131, 2735-2755,
512 <https://doi.org/10.1256/qj.05.15>, 2005.
- 513 Gao, J., Jing, Y.: Satellite Remote Sensing Cloud Detection Method Based on Fully
514 Convolutional Neural Network,*Infrared Technology*, 41, 607-615, 2019.
- 515 Gomez-Chova, L., Camps-Valls, G., Amoros-Lopez, J., Guanter, L., Alonso, L.,
516 Calpe, J., Moreno, J.: New Cloud Detection Algorithm for Multispectral and
517 Hyperspectral Images: Application to ENVISAT/MERIS and PROBA/CHRIS

518 Sensors, *IEEE International Symposium on Geoscience and Remote Sensing*, 2757–
519 2760, doi:10.1109/igarss.2006.709, 2006.

520 Hu, J.: Research on Cloud Detection Algorithm of Remote Sensing Image Based on
521 Convolution Neural Network, *Nanjing University of Information Science and
522 Technology*, doi:10.27248/d.cnki.gnjqc, 2020.

523 Kay, S., Hedley, J., Lavender, S.: Sun Glint Correction of High and Low Spatial
524 Resolution Images of Aquatic Scenes: a Review of Methods for Visible and Near-
525 Infrared Wavelengths, *Remote Sensing*, 1, 697-730,
526 <https://doi.org/10.3390/rs1040697>, 2009.

527 Kegelmeyer, W.P.J.: Extraction of cloud statistics from whole sky imaging
528 cameras, 1994.

529 Kong, Y.-L., Huang, Q., Wang, C., Chen, J., Chen, J., & He, D. (2018). Long Short-
530 Term Memory Neural Networks for Online Disturbance Detection in Satellite
531 Image Time Series. *Remote Sensing*, 10(3), 452. doi:10.3390/rs10030452

532 Mace, G. G., R. Marchand, Q. Zhang, et al. (2007). CloudSat Project: Level 2 Radar-
533 Lidar GEOPROF product process description and interface control document. Jet
534 Propulsion Laboratory.

535 Pan, C., Xia B., Chen, Y.: Research on MODIS Cloud Detection Algorithms Based on
536 Fuzzy Clustering, *Microcomputer Information*, 25, 124-125+131, 2009.

537 Quesada-Ruiz L C, Rodriguez-Galiano V F, Zurita-Milla R, et al. 2022. Area and
538 Feature Guided Regularised Random Forest: a novel method for predictive

539 modelling of binary phenomena. The case of illegal landfill in Canary Island [J].
540 International Journal of Geographical Information Science, 36(12): 2473-2495.

541 Rossow, W. B., Leonid, C.G.: Cloud detection using satellite measurements of
542 infrared and visible radiances for ISCCP. *Journal of Climate*, 12, 2341-2369,
543 [https://doi.org/10.1175/1520-0442\(1993\)006<2341:CDUSMO>2.0.CO;2](https://doi.org/10.1175/1520-0442(1993)006<2341:CDUSMO>2.0.CO;2), 1993.

544 Solvsteen, C.: Correlation based cloud-detection and an examination of the split-
545 window method, *Proceedings of SPIE - The International Society for Optical*
546 *Engineering*, 86-97, 1995.

547 Amato, U., Antoniadis, A., Cuomo, V., Cutillo, L., Franzese, M., Murino, L., Serio,
548 C.: Statistical cloud detection from SEVIRI multispectral images, *Remote Sensing*
549 *of Environment*, 112, 750–766, <https://doi.org/10.1016/j.rse.2007.06.004>, 2008.

550 Wang, Z.: CloudSat Project: CloudSat 2B-CLDCLASS-LIDAR product process
551 description and interface control document, *Jet Propulsion Laboratory*, 2019.

552 Yan, J., Guo, X., Qu, J.: An FY-4A/AGRI cloud detection model based on the naive
553 Bayes algorithm, *Remote Sensing for Natural Resources*, 34, 33-42, 2022.

554 Zhang, W., He, M., Mak, M.W.: Cloud detection using probabilistic neural networks,
555 *Geoscience and Remote Sensing Symposium*, IEEE 2373-2375, 2001.

556 Zhang, Y., William, B. R., Andrew, A. L., Valdar, O., Michael, I. M.: Calculation of
557 radiative fluxes from the surface to the top of atmo- sphere based on ISCCP and
558 other global data sets: Refine- ments of the radiative transfer model and the input
559 data, *Journal of Geophysical Research Atmospheres*, 109, 1-27,

560 <https://doi.org/10.1029/2003JD004457>, 2004.

561