https://doi.org/10.1016/j.rse.2020.112005



# Cloud detection for Landsat imagery by combining the random forest and superpixels extracted via energy-driven sampling segmentation approaches

Jing Wei[1,2], Wei Huang[3], Zhanqing Li[2*], Lin Sun[4], Xiaolin Zhu[5], Qiangqiang Yuan[6], Lei Liu[7], Maureen Cribb[2]

1. State Key Laboratory of Remote Sensing Science, College of Global Change and Earth System Science, Beijing Normal University, Beijing, China
2. Department of Atmospheric and Oceanic Science, Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, USA
3. State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing, China
4. College of Geomatics, Shandong University of Science and Technology, Qingdao, China
5. Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong
6. School of Geodesy and Geomatics, Wuhan University, Wuhan, China
7. College of Earth and Environmental Sciences, Lanzhou University, Lanzhou, China

* Corresponding author. E-mail address: zli@atmos.umd.edu

23

**Abstract**

A primary challenge in cloud detection is associated with highly mixed scenes that are filled with broken and thin clouds over inhomogeneous land. To tackle with this challenge, we developed a new algorithm called the Random-Forest-based cloud mask (RFmask) which can improve the accuracy of cloud identification for Landsat Thematic Mapper (TM), Enhanced Thematic Mapper Plus (ETM+), and Operational Land Imager and Thermal Infrared Sensor (OLI/TIRS) images. For the development and validation of the algorithm, we first stratified select cloudy and clear-sky pixels to form a prior pixel database according to the land use cover around the world. Next, we select typical spectral channels and calculate spectral indices based on the spectral absorption or reflection characteristics of different land cover types using the top-of-atmosphere reflectance and brightness temperature. Then they are used as inputs to the Random Forest model for training and establishing the preliminary cloud detection model. Finally, the Super-pixels Extracted via Energy-Driven Sampling image segmentation approach is applied to process the preliminary classification results and to obtain the final cloud detection results. The RFmask results are evaluated against the globally distributed United States Geological Survey cloud-cover assessment validation products. The average overall accuracy for RFmask cloud results reaches up to 95.3% (Kappa coefficient = 0.85) with a small omission error of 8.3% and a commission error of 6.4%. The new model shows an excellent ability to capture broken and thin clouds, especially over bright surfaces. In general, the RFmask algorithm is accurate and efficient, and potentially useful for data preprocessing in related quantitative studies with Landsat images. The new model generally outperforms other methods, many being compared here, especially for identifying such challenging scenes as broken and thin clouds over bright surfaces. In general, the RFmask algorithm is accurate and efficient, and potentially useful for data preprocessing in conducting quantitative studies with Landsat images.

**Keywords:** Landsat, cloud detection, RFmask, random forest, Superpixels Extracted via Energy-Driven Sampling

## 1. Introduction

Clouds are ubiquitous in remote sensing images due to the influence of atmospheric conditions and imaging observations. The annual mean cloud cover can reach up to ~ 66%, especially in the tropics (Ju & Roy, 2008; Zhang et al., 2004). Clouds influence the atmospheric environment and global climate change by affecting the radiation budget balance by absorbing and reflecting surface and solar energy (Andreae & Rosenfeld, 2008; Z. Q. Li et al., 2016; Ramanathan et al., 1989; Stephens, 2005). The presence of clouds hinders quantitative extraction of surface and atmospheric parameters for such purposes as classification and monitoring of land-use and land-cover changes, and retrievals of vegetation coverage, surface temperature, aerosol optical depth, and particulate matters (Sun et al., 2016; Wei et al., 2018, 2019; Zhu & Woodcock, 2012; Zhu & Helmer, 2018; Wulder et al., 2019).

Other than for cloud studies, the presence of clouds is a major source of noise in quantitative remote sensing applications, and thus cloud detection has become an indispensable and essential preprocessing step (Arvidson et al., 2001; Irish, 2000). Grossly speaking, clouds may be classified as thick and thin clouds, and/or homogeneous and broken clouds. Thick clouds and homogenous clouds are usually easy to identify because of their distinct features. Due to their small size, tenuous features, and irregular shapes, broken and thin clouds are much more difficult to identify. Thin clouds, in particular, are usually translucent, revealing diverse underlying surfaces in images. It is usually very challenging to identify semi-transparent clouds as their spectral signals coming from both clouds and underlying surfaces (e.g., vegetation, soil, and water), especially over bright surfaces (Gao et al., 2002; Irish, 2000; Rossow & Dueñas, 2004; Sun et al., 2016; 2017).

Over the years, many cloud identification methods have been proposed for applications with various satellite imaging sensors such as the historical sensor of the Advanced Very High Resolution Radiometer (Saunders & Kriebel, 1988, Rossow & Dueñas, 2004), the MODerate-resolution Imaging Spectroradiometer (Ackerman et al., 2008; Frey et al., 2008; Lyapustin et al., 2008), and the Multi-angle Imaging SpectroRadiometer (Girolamo and Wilson, 2003; Yang et al., 2007). These traditional approaches are chiefly threshold-based applied to multi-spectral channels (e.g., thermal infrared, carbon dioxide, and water vapor absorption channels). For high spatial resolution sensors,

e.g., the US-Landsat, French-Sentinel, French-SPOT, and Chinese-HJ and GF, the spectral channels are usually much fewer, posing a lot more challenging for cloud identification.

Landsat satellite data have been most widely adopted for studying vegetation phenology, agriculture and forestry, surface temperature monitoring, and air pollution monitoring (Wei et al., 2015, 2017; Wu et al., 2019; Wulder et al., 2019) by virtue of its high spatial resolution, global coverage, and long-term data record of over 47 years. Currently, there are three popular widely used generations of Landsat sensors: the Thematic Mapper (TM) aboard Landsat 4/5 (launched in 1984), the Enhanced Thematic Mapper Plus (ETM+) aboard the Landsat 7 (launched in 1999), and the Operational Land Imager and Thermal Infrared Sensor (OLI/TIRS) aboard the Landsat 8 (launched in 2013). Table 1 provides detailed information about the Landsat 4-8 satellites.

Table 1. Detailed information about the Landsat 4–8 satellites.

| Landsat 4-5 TM | | | Landsat 7 ETM+ | | | Landsat 8 OLI/TIRS | | | |
|---|---|---|---|---|---|---|---|---|---|
| Band index | Wavelength (μm) | Spatial resolution | Band index | Wavelength (μm) | Spatial resolution | Band index | Wavelength (μm) | Spatial resolution | Band type |
| - | - | - | - | - | | 1 | 0.433–0.453 | 30 m | Coastal |
| 1 | 0.450–0.520 | 30 m | 1 | 0.450–0.515 | 30 m | 2 | 0.450–0.515 | 30 m | Blue |
| 2 | 0.520–0.600 | 30 m | 2 | 0.525–0.605 | 30 m | 3 | 0.525–0.600 | 30 m | Green |
| 3 | 0.630–0.690 | 30 m | 3 | 0.630–0.690 | 30 m | 4 | 0.630–0.680 | 30 m | Red |
| 4 | 0.760–0.900 | 30 m | 4 | 0.750–0.900 | 30 m | 5 | 0.845–0.885 | 30 m | NIR |
| 5 | 1.550–1.750 | 30 m | 5 | 1.550–1.750 | 30 m | 6 | 1.560–1.660 | 30 m | MIR |
| 6 | 10.40–12.50 | > 120 m | 6 | 10.40–12.50 | 60 m | 10 | 10.60–11.19 | 100 m | TIR-1 |
| 7 | 2.080–2.350 | 30 m | 7 | 2.090–2.350 | 30 m | 7 | 2.100–2.300 | 30 m | SWIR |
| - | - | - | 8 | 0.520–0.900 | 15 m | 8 | 0.500–0.680 | 15 m | Panchromatic |
| - | - | - | - | - | - | 9 | 1.360–1.390 | 30 m | Cirrus |
| - | - | - | - | - | | 11 | 11.50-12.50 | 100 m | TIR-2 |

NIR, MIR, SWIR, and TIR represent the near-infrared, mid-infrared, shortwave infrared, and thermal infrared bands, respectively.

Over the years, an increasing number of cloud detection algorithms have been developed for Landsat satellites. Irish (2000) proposed the Automated Cloud Cover Assessment (ACCA) algorithm for cloud screening from Landsat images based on multiple spectral-channel filters and TIR bands (Irish et al., 2006). Subsequently, Zhu and Woodcock (2012) proposed a Function of mask (Fmask) algorithm to identify the clouds for Landsat imagery through a series of spectral tests

and probabilities of normalized temperature, spectral variability, and brightness (Zhu et al., 2015). Sun et al. (2016) developed a Universal Dynamic Threshold Cloud Detection Algorithm (UDTCDA) to identify clouds based on a priori constructed surface reflectance database, which can minimize the effects of mixed surfaces and improve the overall accuracy of cloud recognition. Zhai et al. (2018) proposed a unified cloud detection algorithm with spectral indices (CSD-SI) according to the physical reflective characteristics of multiple optical remote sensing sensors. Moreover, several deep learning methods based on the convolutional neural network have been modified for detecting clouds in Landsat images, e.g., multi-scale convolutional feature fusion (MSCFF, Z. W. Li et al., 2019), SegNet (Chai et al., 2019), and U-Net (UNET, Wieland et al., 2019).

Despite some unique merits in these algorithms, due to complex and changeable of surface conditions, it is difficult to determine the appropriate cloud recognition thresholds using the few spectral channels, making the traditional threshold methods. Thus, traditional physical approaches still suffer from large errors in identifying broken clouds and thin clouds, especially over bright surfaces (Frantz et al., 2018; Irish et al., 2006; Oishi et al., 2018; Rossow & Dueñas, 2004; Sun et al., 2016). Deep learning approaches yield stronger data mining capability and can achieve more accurate cloud detection results. However, deep learning has more complex model parameters, and needs establish hundreds or thousands of internal network layers, thus model adjustment and training time increase dramatically (T. Li et al., 2017; Z. Li et al., 2019; Wei et al., 2019; Zhai et al., 2019). Meanwhile, model training and running are highly dependent on the computer configuration, making them difficult to be used in operational applications of data preprocessing in meteorological or environmental departments. Therefore, proposed here is an efficient and accurate cloud detection algorithm based on the tree-based ensemble learning approach, i.e., Random Forest (RF), for Landsat imagery.

Pixel database construction and spectral feature selection are first performed to provide adequate training samples. They are then used as inputs to the RF model for training and building cloud detection models for Landsat images. Finally, object-oriented segmentation technology is applied for postprocessing to reduce the effects of salt and pepper noise and obtain the final cloud detection results. Sections 2 and 3 introduce the data source and RF-based cloud mask (RFmask) algorithm.

Section 3 presents the qualitative and quantitative validations of the RFmask results. Sections 5 and 6 give a discussion and summary of this study.

## 2. Data source

In this study, two United States Geological Survey (USGS) cloud-cover assessment validation products, i.e., the L7 Irish Cloud Validation Masks and the L8 Biome Cloud Validation Masks, are selected and used for cloud detection experiments and validation (U.S. Geological Survey, 2016). The L7 Irish dataset includes a total of 206 Landsat 7 ETM+ (scan lines corrector on) Level-1G scenes which are evenly distributed in nine latitude zones around the world, including the austral, boreal, mid-latitude, polar, sub-tropical and tropical regions (Irish et al., 2006; Scaramuzza et al., 2012). The L8 Biome dataset includes a total of 96 Landsat 8 OLI/TIRS terrain-corrected Level-1T scenes which are evenly distributed globally, covering most land surface types, e.g., barren, forest, grass/crops, shrubland, snow/ice, water, and wetlands types (Foga et al., 2017). All of these selected Landsat images cover varying degrees of cloud amount and almost all types of underlying surfaces to ensure that these data are fully representative. However, due to the difficulty of identifying clouds at the pixel level, not all the assessment masks are accurate enough to be used for validation purposes (Foga et al., 2017). Therefore, 46 of the L7 Irish scenes and 6 of the L8 Biome scenes are excluded due to the low accuracy of the manual cloud masks or artifacts in the Landsat images. Therefore, the remaining 160 and 90 of the L7 Irish and L8 Biome scenes are used to conduct the cloud detection research, respectively. Figure 1 shows the geolocation of the L7 Irish and L8 Biome scenes around the world used in this study.

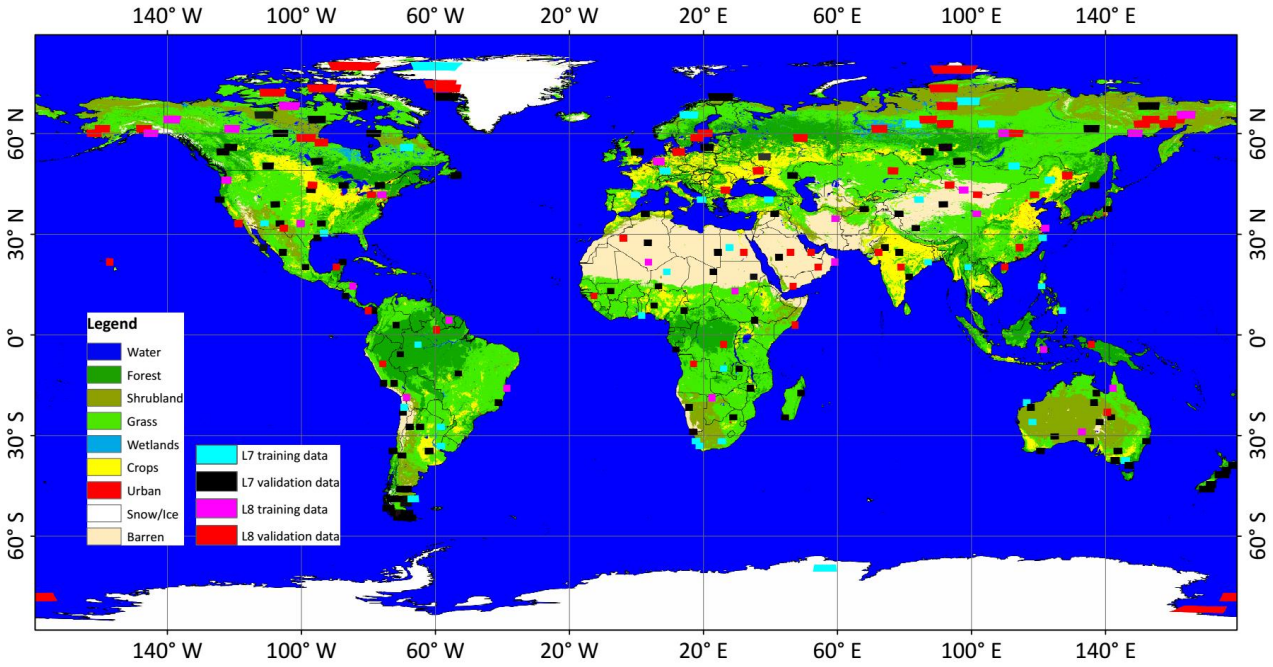



**Figure 1.** Spatial distributions of Landsat 7 Irish training (marked in cyan) and validation (marked in black) data, and Landsat 8 Biome training (marked in pink) and validation (marked in red) data used in this study. Background map is obtained from the MODIS global land use cover product in 2016 (https://search.earthdata.nasa.gov).

## 3. Methodology

Satellite-received signals are recorded as Digital Numbers (DN) from visible to thermal infrared channels in Landsat imagery. Therefore, before cloud detection, the DN values recorded in these channels are first translated into the top-of-atmosphere (TOA) reflectance or brightness temperature (BT) through radiometric calibration (Chander et al., 2009; https://www.usgs.gov/land-resources/nli/landsat/landsat-project-documents). In this study, a new cloud mask algorithm for Landsat imagery (named the RFmask algorithm) is proposed, which contains two key steps, including the pixel-based Random Forest (RF) classification and the object-oriented Super-pixels Extracted via Energy-Driven Sampling (SEEDS) segmentation, respectively.

### 3.1 RF classification

Random Forest (RF, Breiman, 2001) is a new and highly flexible machine learning algorithm, which has a wide range of application prospects. It has been successfully adopted in different research fields, e.g., marketing management and health insurance modeling (Bahnsen et al., 2015;

Khalilia et al., 2011; Mamyrova et al., 2014), risk assessment and prediction (Malekipirbazari & Aksakalli, 2015; Wang et al., 2015), and near-surface fine particles estimation (Hu et al., 2017; Wei et al., 2019). However, it has rarely been used for land use classification (Nitze et al., 2015; van Beijma et al., 2014), especially for cloud recognition. It is thus selected for use in this study. However, different from traditional machine learning methods, RF is a nonlinear algorithm that integrates multiple decision trees through the idea of ensemble learning. There are two key parts, one is "random", that is, random sampling is used to build a decision tree; the other is "forest", which is combined by hundreds of decision trees to generate forest. Then the classification results of several weak classifiers are voted to form a strong classifier, which is the idea of "Bagging". There are four key steps in RF classification:

(1) *n* samples are randomly selected from the original dataset (N) as a training set using the **B**ootstrap **agg**regat**ing** (Bagging) resampling algorithm;

(2) In each node generated, *D* features are selected randomly and unrepeatably, which are used to split the sample set, respectively, and the Gini index is used to calculate the criterion (Jiang et al., 2009; Wei et al., 2020) and determine the best split feature. Note that each tree can grow without pruning during the split process;

(3) Steps 1 to 2 are repeated for a total of *M* times, and *M* decision trees are built in the random forest. The Classification And Regression Tree (CART, Breiman et al., 1984) algorithm is used for tree building;

(4) The test samples are predicted by the random forest obtained from training, and the final classification results are determined by the majority voting the classification results of all week classifiers.

Two main factors characterize the performance of the RF model during the classification. The first one is that the RFs converge. This ensures that the model does not over-fit as the number of decision trees increases. The margin function ($m$) is used to measure the degree that the average number of votes of the right class at random vectors X and Y exceeds the average vote ($\mathrm{a}v_i$) of any other class,

$$\mathrm{m(X,Y)} = \mathrm{a}v_i f(h_i(X) = Y) - max_{j \neq Y}(\mathrm{a}v_i f(h_i(X) = j)) \quad , \quad (1)$$

where *f* represents the indicator function, and $h_i$ *(i = 1, 2, … , k)* represents an ensemble of classifiers. The greater the margin, the greater the confidence in the classification. The generalization error (GE*) refers to the error of the model on the test sample set,

$$\mathrm{GE}^* = E_{X,Y}(m(X,Y) < 0) \quad . \quad (2)$$

With the increase in decision trees, for almost all sequences ($\theta$), the GE* converges to,

$$\lim_{i\to\infty} \mathrm{GE}^* = E_{X,Y}\left(E_\theta(h_i(X,\theta) = Y) - max_{j\neq Y}(E_\theta(h_i(X,\theta) = j) < 0\right) \quad . \quad (3)$$

The second factor is strength ($s$) and correlation ($\bar{r}$), which are used to measure the accuracy and the dependence between individual classifiers, then the upper bound for the GE* can be derived,

$$\mathrm{GE}^* \leq \bar{r}(1 - s^2)/s^2 \quad , \quad (4)$$

The greater the strength and the smaller the correlation between decision trees, the more accurate the model is. Detailed information on the RF algorithm can be found in Breiman (2001).

### 3.1.1 Pixel database construction

In this study, a prior pixel database is first constructed for Landsat 7 and 8 satellites to provide abundant data samples for model training and validation. The pixel database contains a cloudy-pixel and a clear-sky-pixel part. For this purpose, the underlying surfaces are divided into nine main categories according to the MODIS global land cover product, i.e., barren, forest, grass, crops, shrubland, urban, wetlands, water, and snow/ice. First, about one-third of Landsat 7 Irish (~55 of 160) and Landsat 8 Biome (~30 of 90) images are stratified selected as training images which are globally evenly distributed according to the land use cover, and the remaining two-thirds are used as validation images (Figure 1). Meanwhile, the selected training images better have a moderate manual cloud amount between 35 and 65 percent to ensure that there are enough clouds and clear-sky pixels in the image.

Then, because the random forest classification is performed on the pixel level, instead of using the whole image, 60% of the total pixels are randomly selected from each image as training samples to build the pixel database to improve the training efficiency. This can ensure that the training samples can cover almost all kinds of clouds (e.g., thick, thin, and broken clouds) and clear sky over diverse land-use types. In general, there are a total number of approximately 5735 million (1258 million) clear-sky pixels and 5652 million (1279 million) cloudy pixels collected from Landsat 7 (8)

imagery in the pixel database over nine main categories. Table 2 provides detailed information of training dataset used in the pixel database construction for Landsat imagery.

**Table 2.** Statistics of the total number of clear-sky and cloudy pixels from the constructed pixel database for Landsat 7 and 8 imagery.

| Statistics | Landsat 7 imagery | | | Landsat 8 imagery | | |
|---|---|---|---|---|---|---|
| | $N_{Image}$ | $N_{Clear\text{-}sky}$ | $N_{Cloudy}$ | $N_{Image}$ | $N_{Clear\text{-}sky}$ | $N_{Cloudy}$ |
| | Scene | Million | Million | Scene | Million | Million |
| Barren | 8 | 780 | 795 | 3 | 149 | 162 |
| Forest | 7 | 871 | 844 | 4 | 200 | 216 |
| Crops | 8 | 472 | 446 | 3 | 118 | 114 |
| Grass | 6 | 768 | 801 | 3 | 133 | 120 |
| Shrubland | 6 | 588 | 607 | 4 | 161 | 153 |
| Urban | 5 | 693 | 667 | 4 | 177 | 187 |
| Wetlands | 5 | 734 | 709 | 3 | 136 | 145 |
| Water | 8 | 627 | 599 | 4 | 115 | 120 |
| Snow/ice | 2 | 202 | 185 | 2 | 69 | 62 |
| Total | 55 | 5735 | 5652 | 30 | 1258 | 1279 |

### 3.1.2 Feature attribute selection

The next important thing is to select the feature attributes of the data samples used in the RF classification. Previous studies have illustrated that the TOA reflectance of the clouds is much higher than most typical ground objects (e.g., water, soil, vegetation, artificial building, and rock) in short visible channels under ideal conditions. In addition, the near-infrared (NIR), mid-infrared (MIR), and short-wave infrared (SWIR) channels are also used to detect clouds due to noticeable differences between the reflectance of clouds and above ground objects. However, snow and ice exhibit very close spectral characteristics to clouds from short to medium wavelengths, thus the thermal infrared channels play an important role in separating them due to their large differences in brightness temperatures (Lin et al., 2012; Sun et al., 2016; Zhu and Woodcock, 2012). More importantly, an additional cirrus channel was designed for Landsat 8 satellite, which has been proved to be useful in detecting cirrus clouds (Gao et al., 2002; Gao and Li, 2000, 2017; Shen et al., 2015; Zhu et al., 2017). Therefore, seven and ten channels of the Landsat 7 (Bands 1–7 in Table 1) and Landsat 8 (Bands 1–11 exclude Band 9 in Table 1) are selected as basic spectral features.

Thick clouds can be easily departed from pure ground objects; however, on the one hand, thin and broken clouds are always covered above the underlying surfaces, and the mixed pixels formed by them are ubiquitous in remote sensing images. The spectral reflectance of different surface types can be changed a lot due to different cloud amounts. On the other hand, remote sensing images can become gradually blurred affected by the increasing air pollution, resulting in more complex spectral characteristics of ground objects. These two key factors largely increase the difficulties in separating clouds from changeable underlying surfaces through discrete spectral channels. This is also the main problem faced by traditional threshold-based methods (Sun et al., 2016; Zhu and Woodcock, 2012).

More importantly, different from other machine/deep learning approaches, random forest is one of the supervised classification algorithms, which is similar to the traditional decision tree classification but is a classifier composed of multiple decision trees. It is highly dependent on spectral features that are totally independent during the tree building. Therefore, the spectral absorption and reflection characteristics of these key land cover types mentioned previously are enhanced by introducing additional spectral indices.

For natural vegetation, four typical vegetation indices are considered: the widely used Normalized Difference Vegetation Index (NDVI, Eq. 5), which easily saturates in densely vegetated areas; the Ratio Vegetation Index (RVI, Eq. 6), which can enhance the radiation difference between vegetation and soil backgrounds; the Enhanced Vegetation Index (EVI, Eq. 7), which uses the blue channel to enhance the vegetation signal by correcting the effect of the soil background and aerosol scattering; and the SWIR-based NDVI ($NDVI_{swir}$, Eq. 8), which is not sensitive to aerosols:

$$NDVI = (\rho_{NIR} - \rho_{Red})/(\rho_{NIR} + \rho_{Red}) , \quad (5)$$

$$RVI = \rho_{NIR}/\rho_{Red} , \quad (6)$$

$$EVI = 2.5(\rho_{NIR} - \rho_{Red})/(\rho_{NIR} + 6\rho_{Red} - 7.5\rho_{Blue} + 1) , \quad (7)$$

$$NDVI_{swir} = (\rho_{SWIR} - \rho_{MIR})/(\rho_{SWIR} + \rho_{MIR}) . \quad (8)$$

For water, the Normalized Difference Water Index (NDWI, Eq. 9) is selected to highlight water bodies. However, NDWI is less effective in extracting water bodies when more buildings are in the background. Thus, a customized TOA reflectance ratio ($TR_{ng}$) involving NIR and green-channel reflectances is calculated simultaneously to help enhance the water information:

$$NDWI = (\rho_{Green} - \rho_{NIR})/(\rho_{Green} + \rho_{NIR}) \ . \quad (9)$$

Similarly, the Normalized Difference Building Index (NDBI, Eq. 10) is selected to enhance the impervious surface layers over urban areas. For barren surfaces, a customized TOA reflectance ratio ($TR_{nm}$) involving NIR and mid-infrared (MIR) channels is formulated to enhance the bright rock and desert information (Irish, 2000). The Normalized Difference Snow Index (NDSI, Eq. 11) is calculated to enhance the snow and ice information in Landsat images:

$$NDBI = (\rho_{MIR} - \rho_{NIR})/(\rho_{MIR} + \rho_{NIR}) \ , \quad (10)$$

$$NDSI = (\rho_{Green} - \rho_{MIR})/(\rho_{Green} + \rho_{MIR}) \ , \quad (11)$$

A "whiteness" index is also calculated to accentuate clouds since clouds look white and are highly reflective with relatively flat changes in the visible band. By contrast, other land cover types show more dramatic changes:

$$\bar{\rho} = (\rho_{Blue} + \rho_{Green} + \rho_{Red})/3 \ , \quad (12)$$

$$Whiteness = \sum_{i=1}^{3} \left| \frac{\rho_i - \bar{\rho}}{\bar{\rho}} \right|, (i = Blue, Green, Red) \ . \quad (13)$$

In summary, there are a total of 17 (20) spectral features, including seven (ten) spectral channels of TOA reflectance and BT, and ten common spectral indices for Landsat 7 (8) images.

### 3.1.3 Model training and validation

Random forest can process large amounts of data efficiently and handle numerous input variables without the need for data dimension reduction. Moreover, it not sensitive to sensitive to multivariate collinearity variables, and the results are relatively stable to missing and unbalanced data (Breiman, 2001; Wei et al., 2019). Therefore, all the above-mentioned spectral features of data samples are calculated and as inputs to the RF model for model training to construct the classification model of cloud detection for Landsat satellites.

In addition, random forest has an important advantage that it does not need cross-validation or a separate validation test because it can be evaluated internally, that is, an unbiased estimation of the error can be established during the generation process. During the model training, about 1/3 of the training samples (i.e., *oob* samples) did not participate in the generation of the decision tree in each round of bagging sampling, but they are used to calculate the out-of-bag (*oob*) error, which is an

unbiased estimation of the GE* of random forest, and it is similar to the *k*-fold cross validation which requires a lot of calculation. The *oob* scores (1 - *oob* error) is used to represent the generalization ability of the RF model. In the current study, the *oob* scores of the RF models for Landsat 7 and 8 imagery reach up to 0.963 and 0.989, suggesting strong classification models. Therefore, the constructed RF classification models are used to predict and generate preliminary cloud masks for Landsat imagery. Figure 1 shows the flowchart of the RF classification.

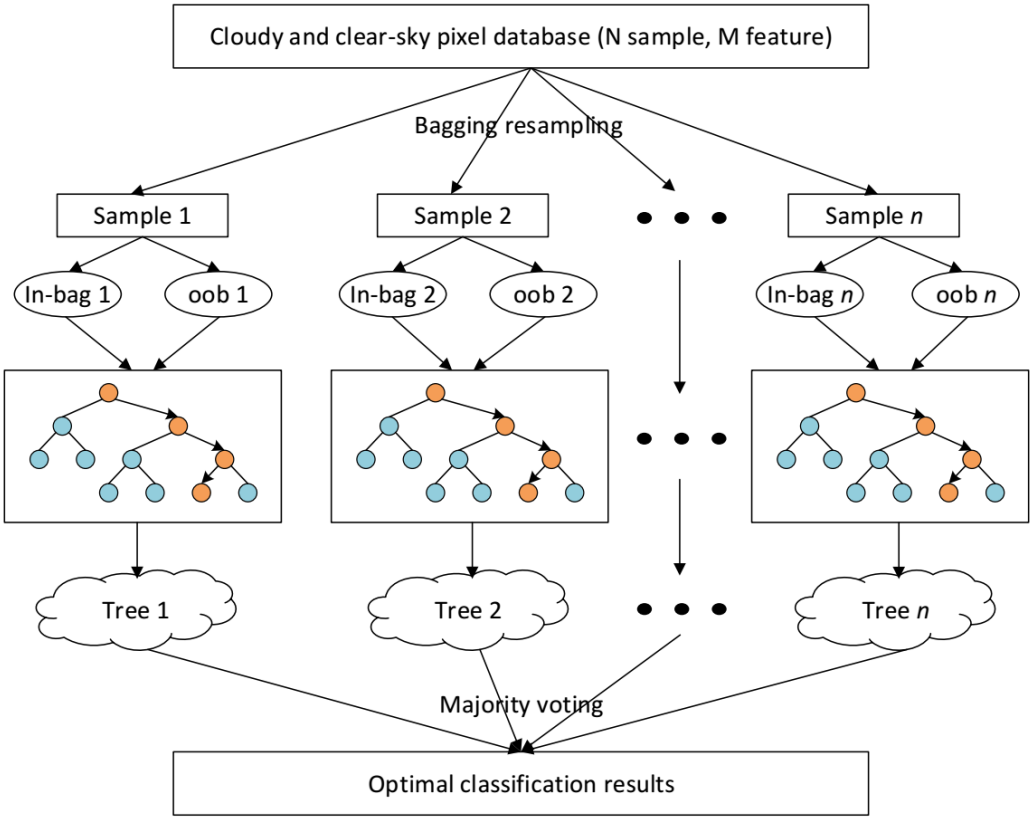



**Figure 2.** Flowchart of the Random Forest classification.

### 3.2 SEEDS segmentation

Moreover, an objected-oriented superpixel image segmentation technology, Superpixels Extracted via Energy-Driven Sampling (SEEDS, Bergh et al., 2012), is selected to post-process the cloud detection results to improve the overall accuracy of classification. It is based on the simple hill-climbing optimization to extract super-pixels, which starts with the initial super-pixel partition and continuously optimizes the super-pixel by modifying the boundary. The SEEDS superpixel segmentation and post-processing of classification mainly include the following four steps:

(1) The red, green, and blue channels of Landsat imagery are combined to an RGB image and used as the inputs of the SEEDS algorithm;

(2) Initialize the superpixel (Seed is the center of superpixel) at the same interval (St), and all the superpixels are rectangles of the same size fit the whole image;

(3) Select a pixel or a group of pixels (s) on the boundary and move them from superpixel *n* to superpixel *K*. If the partitioning $s \in S$ that maximizes the Energy function: E(s) > E(St), this pixel or a group of pixels can be regarded as a part of superpixels;

(4) Iterate step 3 until itself converges (default upper limit of times), and St is the final segmentation result. The total numbers of cloudy and clear-sky pixels in preliminary cloud masks within each superpixel are counted. Then the majority voting is used to determine the final class of all pixels of the entire superpixel.

In the SEEDS algorithm, the Energy function can be expressed as,

$$E(s) = \mathrm{H(s)} + \gamma \mathrm{G(s)}, \qquad (14)$$

where γ indicates the effect weight, and the term H(s) indicates the color distribution of the superpixels and is expressed as:

$$\mathrm{H(s)} = \sum_{k} \varphi(c_{A_k}) = \sum_{k} \sum_{H_j} (c_{A_k}(j))^2 \ , \qquad (15)$$

$$c_{A_k}(j) = \frac{1}{Z} \sum_{i \in A_k} \delta(\mathrm{I(i)} \in \mathrm{H}_j) \ , \qquad (16)$$

where $\varphi()$ denotes the quality measure of color distribution, $c_{Ak}(j)$ denotes the color histogram of the super-pixels ($A_k$) in the $j^{\text{th}}$ bin, $H_j$ denotes the colors in the $j^{\text{th}}$ bin of the histogram, $I$(i) denotes the color of the $i^{\text{th}}$ pixel, $Z$ denotes the normalization factor of the histogram, and $\delta$() is the indicator function. The term G(s) indicates the shape of the superpixels and is expressed as

$$\mathrm{G(s)} = \sum_{i} \sum_{k} (b_{N_i}(k))^2 \ , \qquad (17)$$

$$b_{N_i}(k) = \frac{1}{Z} \sum_{j \in N_i} \delta(\mathrm{j} \in \mathrm{A}_k) \ , \qquad (18)$$

where $N_i$ represents the N × N pixels around the $i^{th}$ pixel, and $b_{Ni}$ represents the histogram of super-pixel labels in the $N_i$ area. Figure 3 illustrates the brief flowchart of the RFmask cloud detection algorithm for Landsat imagery developed in this study.

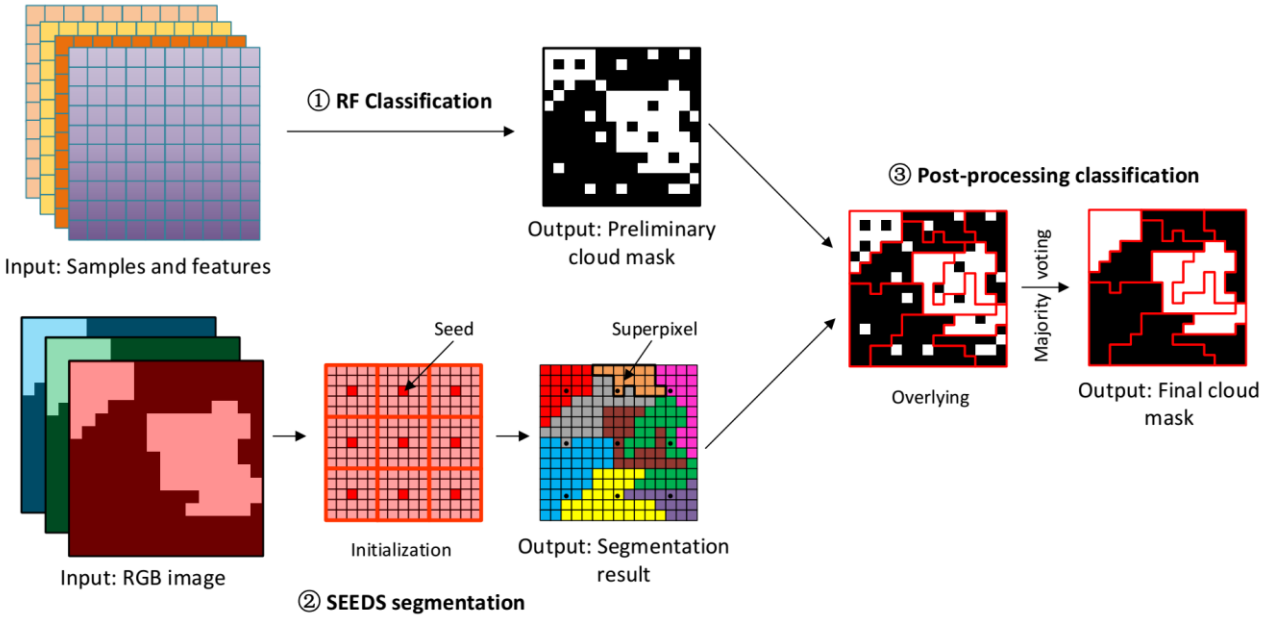



**Figure 3.** Brief flowchart of the RFmask cloud detection algorithm for Landsat imagery.

### 3.3 Evaluation approaches

The total cloud amounts (CAs) from both the cloud detection results and the validation masks for each Landsat image, and their cloud amount difference (CAD) are calculated (Sun et al., 2016). The CA is overestimated when CAD > 0 and underestimated when CAD < 0. The following metrics give a measure of the estimation uncertainty: the regression line, the coefficient of determination ($R^2$), the mean absolute error (MAE), and the root-mean-square error (RMSE). Moreover, the confusion matrix is also used to evaluate the overall accuracy and estimation error of RFmask cloud detection models for Landsat imagery based on six commonly used statistical indicators, i.e., the *Kappa* coefficient (Cohen, 1960), the overall accuracy ($A_O$), the producer's accuracy ($A_P$), the user's accuracy ($A_U$), the omission error (OE), and the commission error (CE):

$$A_O = \frac{TP + TN}{TP + TN + FP + FN}, \quad (19)$$

$$A_P = \frac{TP}{TP + FN}, \quad (20)$$

$$A_U = \frac{TP}{TP + FP}, \quad (21)$$

$$OE = \frac{FN}{TP + FN}, \quad (22)$$

$$CE = \frac{FP}{FP + TN}, \quad (23)$$

where TP (true positive) and TN (true negative) denote the total number pixels of correct prediction; FP (false positive) and FN (false negative) denote the total number pixels of incorrect outcome in cloud or clear-sky recognition, respectively (Sun et al., 2016; Li et al., 2019).

## 4. Results

### 4.1 Qualitative evaluation

Figure 4 illustrates four typical examples of standard-false-color (RGB: Bands 4-3-2) composite images (left two panels in each group of four panels) and binary RFmask cloud results (right two panels in each group of four panels) for Landsat ETM+ satellite data over different land surface types. To better compare the cloud detection results with visual interpretations, full-scene (~8050 × 7300 pixels, upper two panels in each group of four panels) and zoomed-in (~1000×1000 pixels, lower two panels in each group of four panels) images derived from the RFmask results are displayed. The RFmask algorithm appears to more accurately identify most clouds in the image that reveals a large amount of vegetation information (Fig. 4a). The spatial distributions in the cloud detection results are almost identical between the RFmask and the reference cloud masks. In addition, the RFmask algorithm still works well as the amount of vegetation information decreases. For these vegetation-dominated land surface types, e.g., forest (Fig. 4b), cropland (Fig. 4c), and mountains (Fig. 4d), the RFmask algorithm shows great performance with small differences in cloud spatial distribution compared with the reference masks. Furthermore, the RFmask algorithm detects most clouds for parts of the images with little vegetation, especially inland water (Fig. 4a), urban areas (Fig. 4c), and bare rock (Fig. 4b, d), suggesting acceptable classification results.

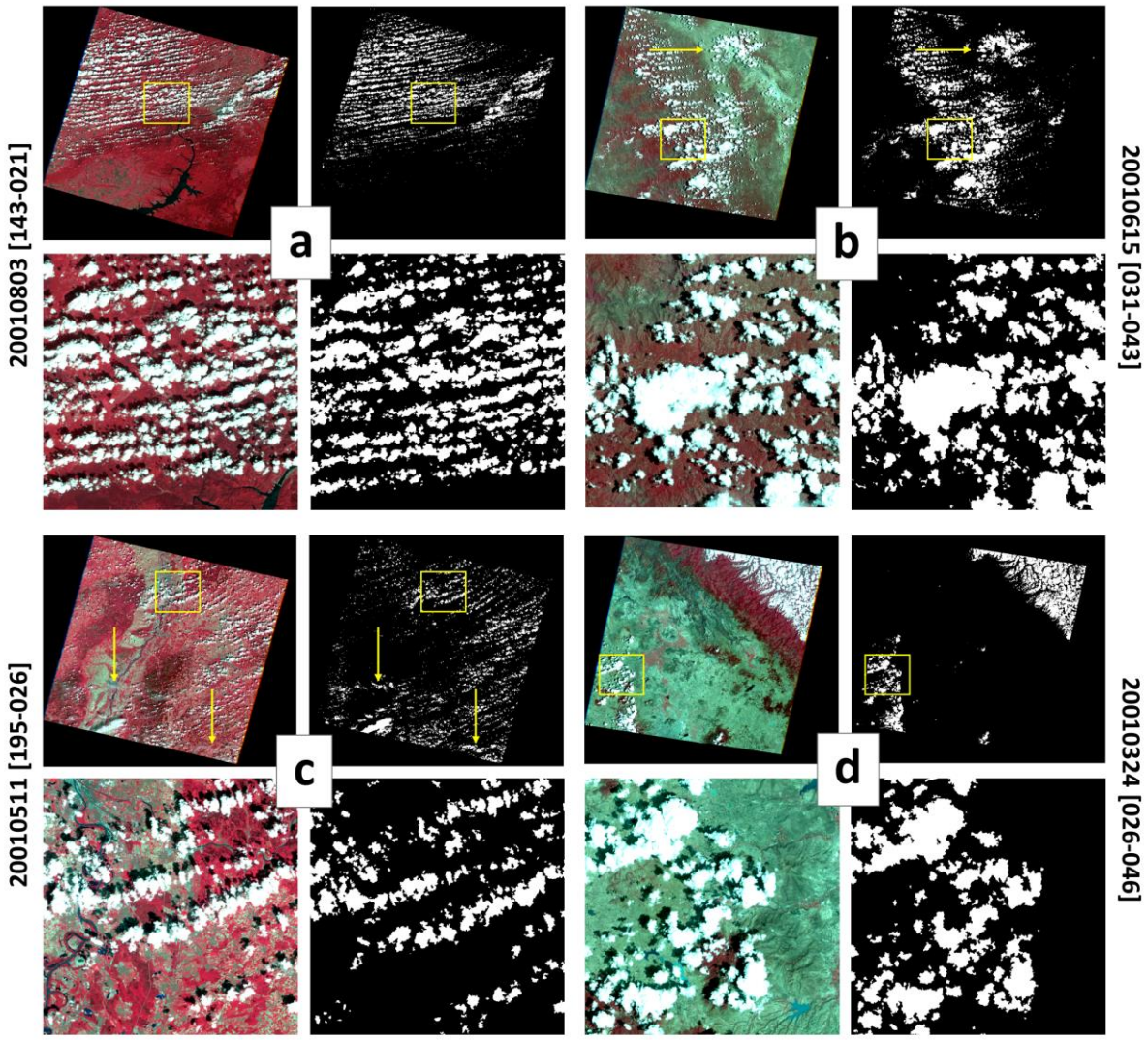
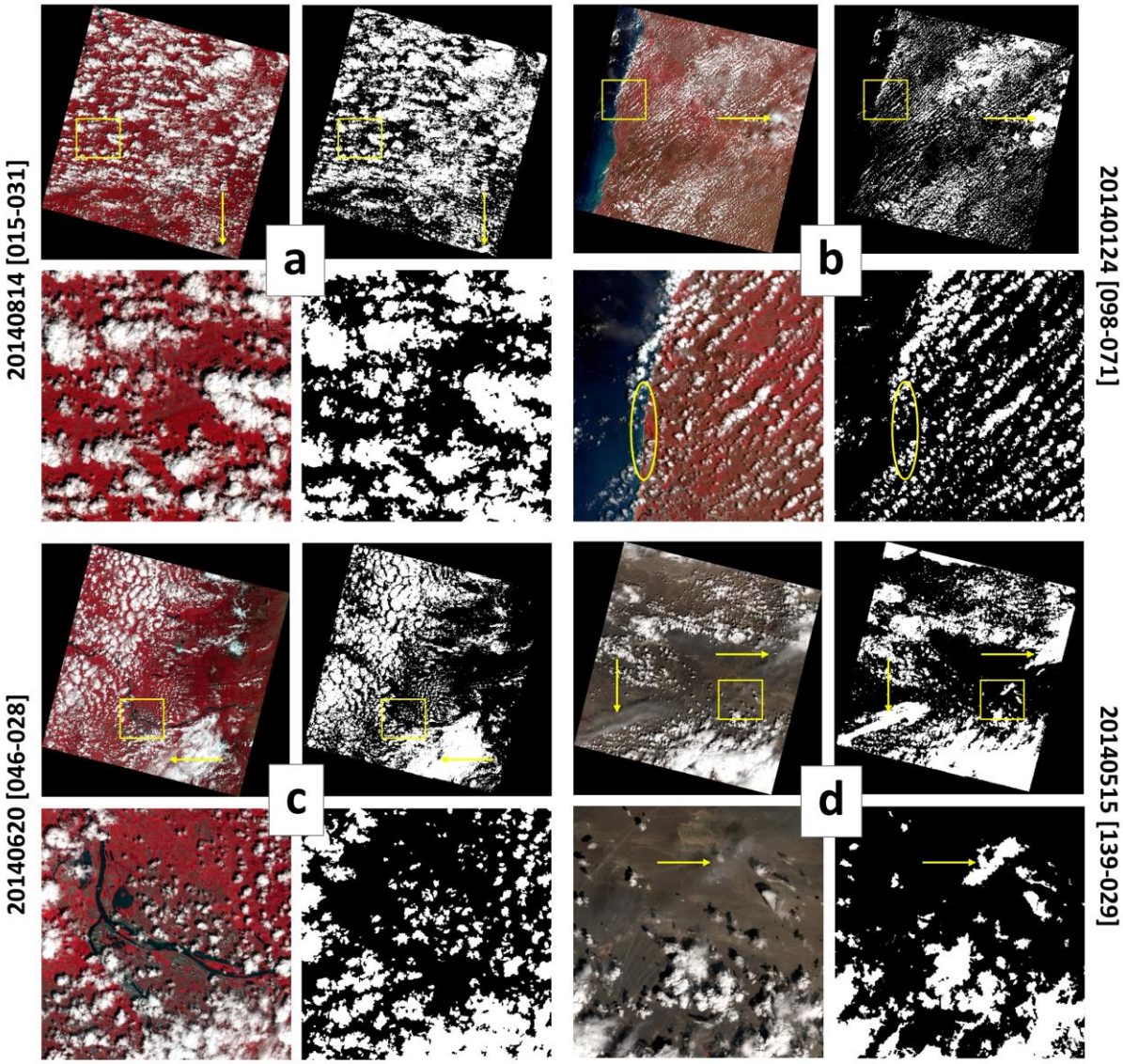



**Figure 4.** Examples of standard-false-color (RGB: Bands 4-3-2) composite images and cloud (masked as white) detection results for Landsat 7 full-scene and zoomed-in images (areas outlined by yellow boxes) over diverse underlying surfaces, where yellow arrows mark clouds and left- and right-side annotations indicate the acquisition time (yyyymmdd) and orbital record (Path-Row) of the Landsat 7 images.

Similarly, four typical examples of the RFmask cloud detection results over varying underlying surfaces for Landsat 8 satellite are also collected. Figure 5 shows the full-scene (~7750 × 7900 pixels) and zoomed-in (~1000 × 1000 pixels) standard-false-color (RGB: Bands 5-4-3) composite images (left two panels in each group of four panels) and RFmask results (right two panels in each group of four panels). The RFmask cloud detection results are substantially consistent with the true cloud distributions in the remote sensing images over densely vegetated areas (Fig. 5a-c). Clouds over darker surfaces, e.g., inland water and offshores, can also be accurately identified (Fig. 5a-c).

401 The RF mask algorithm also generates little incorrect classification results along coastlines, where
402 extreme bright-dark reflectance differences exist (yellow ellipse in Fig. 5b). In addition, clouds over
403 urban buildings and roads are more accurately identified (Fig. 5c). For barren land with little
404 vegetation coverage, the RFmask algorithm can still achieve better recognition results with few
405 missed or misjudged cases. Especially, clear skies are not misidentified as clouds by the RFmask
406 algorithm over bright bare surfaces deep inland (Fig. 5b). In general, the differences in cloud spatial
407 distributions between RFmask results and reference data are relatively small, and there are few
408 incorrect or missing cloud identification pixels, indicating good classification results.
409




410
411 **Figure 5.** Same as Figure 4 but showing Landsat 8 imagery.
412

In addition, Figure 6 illustrates eight typical examples of RFmask results for Landsat imagery over diverse underlying surfaces containing most types of bright surfaces. Bright surfaces have similar spectral characteristics as clouds due to their high surface reflectance, especially in visible and NIR bands. This presents a challenge for traditional cloud detection approaches because it is difficult to determine an appropriate threshold (Irish et al., 2006; Sun et al., 2016, 2017; Zhu and Woodcock, 2012). This can lead to the misidentification of bright surfaces as clouds and to difficulty in accurately detecting thin clouds. The results show that the RFmask algorithm appears to be able to detect most clouds over less vegetated areas (Fig. 6a), bare rocks (Fig. 6b), deserts (Fig. 6c), and plateau mountains (Fig. 6d) with few cloud omissions and false recognitions, especially for thin and broken clouds (pointed to by yellow arrows in Fig. 6). The RFmask algorithm is also capable of excluding very bright rocks (Fig. 6b, g), very bright snow/ice surfaces (Fig. 6d, e), rich mineral surfaces (Fig. 6f), and Gobi and rocky deserts (Fig. 6h) from cloud results. In particular, there is no misidentification of clouds over these cloud-free Landsat images (pointed to by red arrows in Fig. 6) over these typical bright surfaces (Fig. 6e-h).

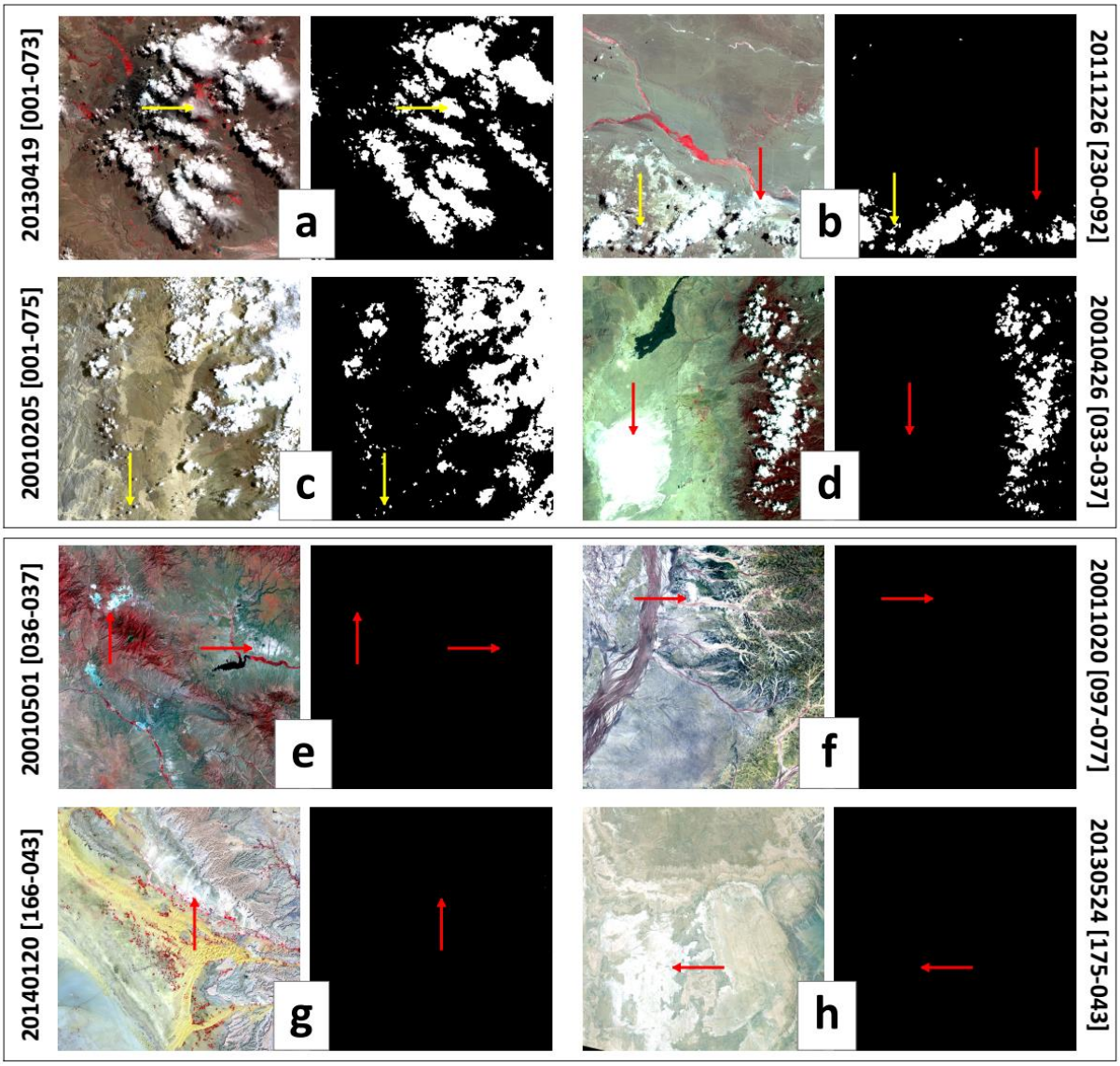



**Figure 6.** Examples of standard-false-color composite images and cloud (masked as white) detection results for Landsat 7-8 zoomed-in (~1000 × 1000 pixels) images over diverse underlying surfaces, where yellow and red arrows point to clouds and bright surfaces, respectively. Left- and right-side annotations indicate the acquisition time (yyyymmdd) and orbital record (Path-Row) of the Landsat images.

### 4.2 Quantitative evaluation

The above results mainly discuss the qualitative evaluation results of cloud detection results based on remote sensing visual interpretation. Therefore, in this section, the Landsat 7 Irish and Landsat 8 Biome validation data are selected to quantitatively evaluate the RFmask results. Table 4 summarizes the cloud amount and accuracy of RFmask-derived clouds with reference to USGS validation mask-determined clouds from all, Landsat 7, and Landsat 8 images, respectively. The

estimated percentages of cloud cover are highly consistent with the USGS manually determined percentages of cloud cover ($R^2$ = 0.99) with a relatively small MAE of 2.57% and RMSE of 4.29% for Landsat imagery. In general, most of the data points are evenly distributed along the 1:1 line with a slope of 1.01, a y-intercept of 0.54, and a mean bias of -0.83, respectively. More importantly, even when considering manual estimation uncertainties, more than 82% of the RFmask results differ from the reference results by less than 5%. Similar validations and comparisons were also done separately for Landsat 7 and 8 images. The estimated cloud cover percentages derived from both RFmask and the reference cloud cover percentages correlate well ($R^2$ = 0.99 and 0.98) with strong slopes close to 1 and small intercepts. The MAEs are 2.27% and 3.16%, and RMSEs are 3.75% and 5.20% for Landsat 7 and 8 imagery, respectively. This suggests that the RFmask algorithm can estimate more accurately the percentage of clouds per scene, an important part of Landsat data pre-screening.

**Table 3.** Statistics of evaluation results of cloud detection amount and accuracy for all, Landsat 7, and Landsat 8 imagery

| Cloud amount | Regression line | $R^2$ | CAD | MAE (%) | RMSE (%) |
|---|---|---|---|---|---|
| All | y = 1.01x + 0.54 | 0.99 | -0.83 | 2.57 | 4.29 |
| Landsat 7 | y = 1.00x + 0.44 | 0.99 | -0.44 | 2.27 | 3.75 |
| Landsat 8 | y = 1.01x + 1.44 | 0.98 | -1.62 | 3.16 | 5.20 |

| Accuracy | *Kappa* | $A_O$ (%) | $A_P$ (%) | $A_U$ (%) | OE (%) | CE (%) |
|---|---|---|---|---|---|---|
| All | 0.85 | 95.3 | 91.8 | 93.6 | 8.3 | 6.4 |
| Landsat 7 | 0.84 | 95.1 | 91.2 | 92.5 | 8.8 | 6.0 |
| Landsat 8 | 0.85 | 95.6 | 92.8 | 96.0 | 7.2 | 7.1 |

Figure 7 shows the frequency histograms of six accuracy indicators calculated from the confusion matrix for all Landsat RFmask results. The *Kappa* coefficient for the RFmask algorithm is 0.85, and the average $A_O$ reaches up to 95.3%. More than 86% and 88% of the RFmask results for Landsat imagery have $A_O$ and $K$ greater than 80% and 0.7, respectively. The average $A_P$ and $A_U$ are 91.8% and 93.6%, and in general, approximately 88% and 92% of the RFmask results have $A_P$ and $A_U$ values greater than 80%, respectively. The average OE and CE are 8.3% and 6.4%, and more than 66% and 80% of the RFmask results have OE and CE values < 10%, respectively. In addition, the RFmask algorithm works well with Landsat 7 imagery with average *Kappa* coefficient, $A_O$, $A_P$, and

$A_U$ values of 0.85, 95.4%, 91.2%, and 92.5%, respectively (Table 4), and small estimation errors, i.e., OE = 8.8% and CE = 6.0%. The RFmask algorithm also shows great performance in Landsat 8 imagery, i.e., *Kappa* = 0.85, $A_O$ = 95.0%, $A_P$ = 92.8%, and $A_U$ = 96.0%, and OE of 7.2% and CE of 7.1%. These results suggest that the RFmask algorithm can be applied to images from different Landsat satellites to detect clouds.

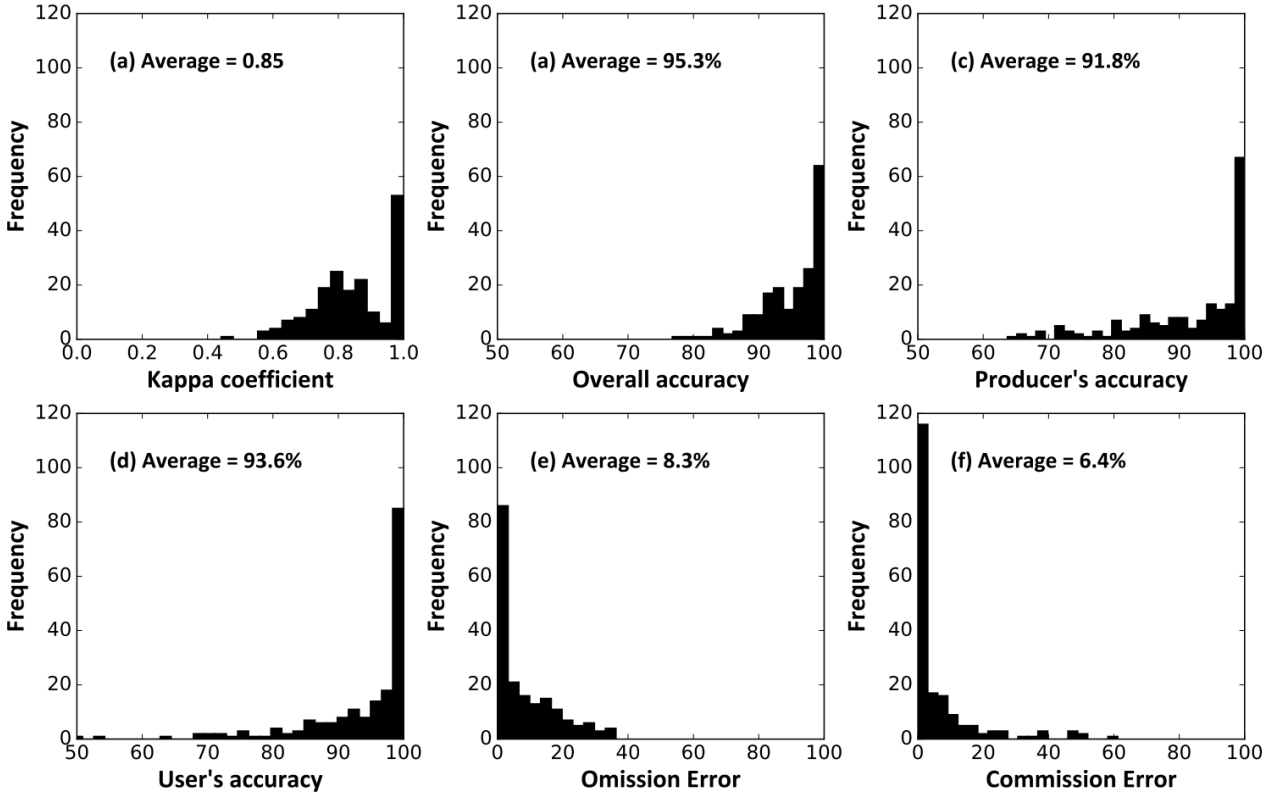



**Figure 7.** Frequency histograms of RFmask cloud results from Landsat images in terms of (a) overall accuracy, (b) the Kappa coefficient, (c) the producer's accuracy, (d) the user's accuracy, (e) the omission error, and (f) the commission error.

Furthermore, the performance of the RFmask algorithm also is validated over different land-use types for Landsat imagery (Figure 8). Results suggest that the RFmask algorithm shows great performance in detecting clouds over most dark surfaces with high overall accuracies > 95% and small omission errors < 10%, especially for these clouds over Shurbland and Wetlands. The main reason is that there are noticeable spectral differences between clouds and these land-use types with low surface reflectance, which are relatively easy to be distinguished, leading to fewer cloud recognition errors. The RFmask algorithm also shows high overall accuracies of 95–97% in cloud

identification over urban, because they are mainly scattered near the dark surfaces such as water and vegetation. However, the performance of the RFmask algorithm overall decreases over brighter surfaces (i.e., barren, snow and ice), possibly due to the significant reduction in spectral differences, making the clouds easy to be wrongly identified (CE > 12%). Nevertheless, the overall accuracy has reached 93% and 89% over barren and snow/ice, respectively.

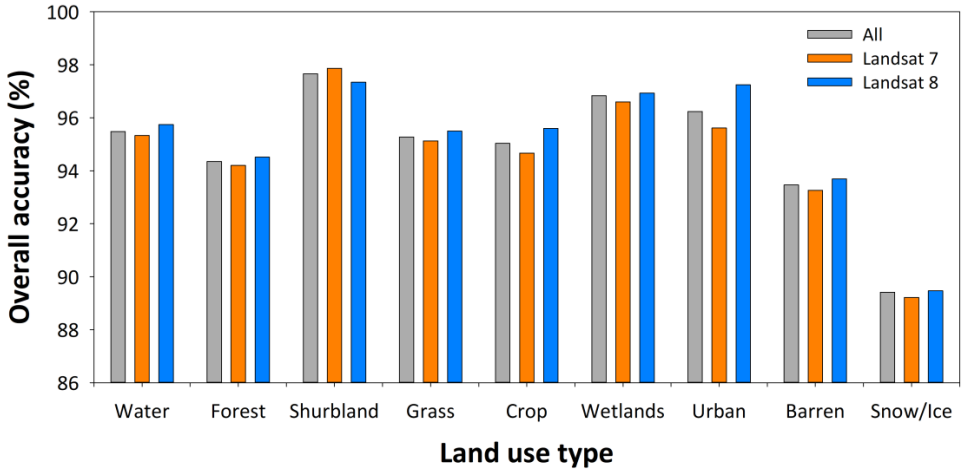



**Figure 8.** Overall accuracy (%) of the RFmask algorithm in cloud detection over diverse land use types for all, Landsat 7, and Landsat 8 imagery.

These results illustrate that our new RFmask algorithm is robust and can more accurately identify most clouds over complex and changeable underlying surfaces with few omission and commission errors, especially over bright surfaces. This is mainly due to the comprehensive inclusion of diverse mixed surfaces in the RFmask algorithm. Mixed cloudy- and clear-sky pixels are fully trained to learn and master their spectral characteristics and differences, so constructed are millions of decision trees to improve the overall cloud detection accuracy in Landsat images, especially for broken and thin clouds.

## 5. Discussion

### 5.1 Evaluation of feature importance

Random forest allows to evaluate the importance of each feature during the classification, i.e., importance score, which is calculated according to the GI index (Jiang et al., 2009; Calle and Urrea,

2011; Wei et al., 2020). Note that this score only indicates the importance of spectral features in splitting when building the decision tree, not the physical contribution. In addition, the importance of features may be varying when a new cluster is inputted; however, for a binary classification problem (e.g., cloud/clear sky), the importance score of the model will change slightly in numerical values with enough training samples. Figure 9 shows the multiple average importance score for each spectral feature in RF classification for Landsat 7 and 8 imagery.

Results show that most spectral features play similar roles in detecting clouds for two different sensors, where the discrete spectral channels are important, especially for thermal and shortwave bands. In addition, due to the lack of some channels, the important scores of visible bands of Landsat 7 increases, whereas the cirrus band of Landsat 8 also plays an important role in (cirrus) cloud detection, which is consistent with the conclusions reported by previous studies (Gao et al., 2002; Gao and Li, 2000, 2017; Shen et al., 2015; Zhu et al., 2015). Furthermore, spectral indices still show large effects on cloud identification, especially for those (i.e., NDBI, $TR_{nm}$, and NDSI) used to enhance bright surfaces, appearing to be more important than some discrete spectral channels. However, the "whiteness" is less important because it is mainly used to assist in identifying pixels that are not "white" enough to be clouds in physical models (Gomez-Chova et al., 2007; Zhu and Woodcock, et al., 2012). Moreover, an additional test is also performed, and the overall accuracy of the full model has been improved by approximately 5% compared to the model trained without considering spectral indices. These results illustrate that these spectral indices are also important in the tree-based ensemble learning approaches.

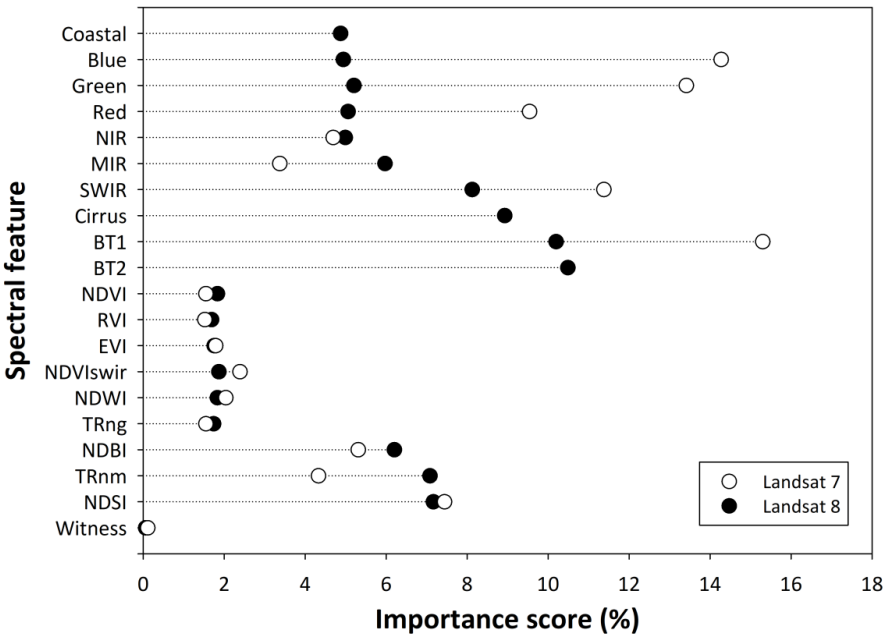



**Figure 9.** Multiple average importance score of each spectral feature in cloud detection during the RF classification for Landsat 7 and 8 imagery.

### 5.2 Importance of superpixel segmentation

Figure 10 compares one cloud detection result using the SEEDS segmentation and the one without the segmentation for Landsat imagery (Figure 10). The results show that without the SEEDS segmentation (Fig.10b), on the one hand, there are many scattered pixels in the cloudless places on the left and upper right corner of the image in the classification results, which are wrongly identified as cloud pixels (areas outlined by yellow circles), resulting in a lot of salt and pepper noises; on the other hand, there are also many pixels in the whole cloud layer that are incorrectly identified as clear-sky pixels (areas outlined by red circles). The main reason is that RF classification is performed at the pixel level, and the spatial autocorrelations and spatial texture information among the pixels are not considered, leading to inevitable noises in the classification results. However, super-pixel segmentation is object-oriented and is selected to address these issues. With the SEEDS segmentation (Fig.10c), these salt and pepper noises have been fully removed, and these patches of clouds have also been completely filled. Compared to the preliminary cloud detection result, the final result shows a higher consistency in cloud distribution with the true-color Landsat image. This suggests that the object-oriented segmentation technology can overall improve the classification accuracy and plays a very important role in pixel-based classification.

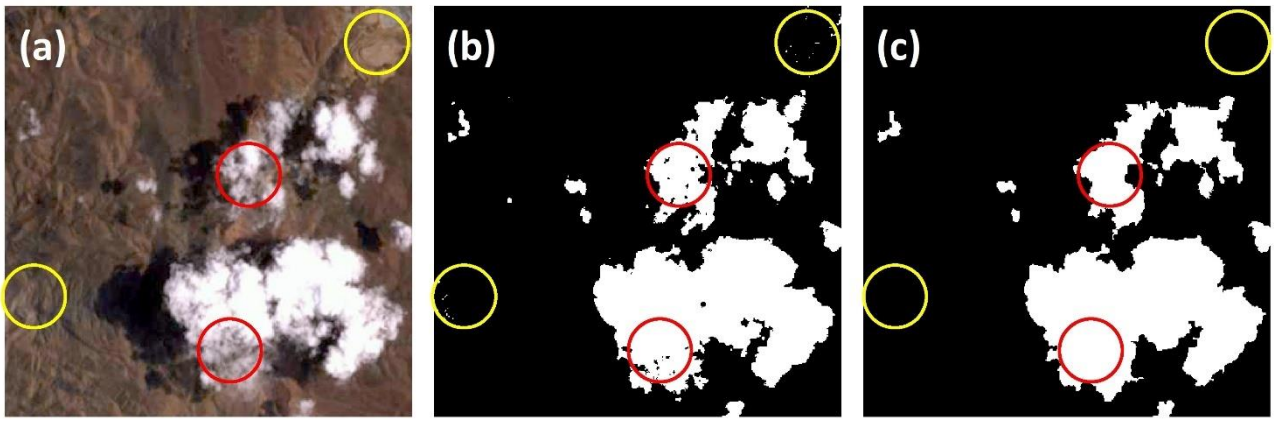



**Figure 10.** A typical example of (a) zoomed-in RGB combined image and (b) the preliminary cloud detection result without the SEEDS segmentation and (c) the final cloud detection using the SEEDS segmentation for Landsat imagery.

### 5.3 Comparison with related cloud studies

Over the years, many cloud detection algorithms have already been proposed. Most algorithms can correctly identify the thick clouds over all types of land surfaces because thick clouds have complete and regular shapes. However, for thin and broken clouds, they are irregular in shape and lesser in amount or usually occupy only a few pixels, even sub-pixels. Thus, due to the influence of mixed pixels formed by clouds and changeable underlying surfaces, traditional threshold-based cloud detection methods are difficult to set accurate cloud detection thresholds, and always fail to identify these clouds from Landsat images, especially over bright surfaces (Irish et al., 2006; Goodwin et al., 2013; Jin et al., 2013; Qiu et al., 2017; Oishi et al., 2018; Sun et al., 2018; Zhai et al., 2018; Zhu & Woodcock, 2012). However, the RFmask algorithm can well identify most thin and broken clouds in Landsat images (pointed to by colored arrows in Figs. 4–6), and show overall high accuracies in cloud detection over different land use covers (Figure 8).

Furthermore, in comparisons with the same validation sources of L7 Irish and L8 Biome reference masks (Table 4), the RFmask algorithm appears to outperform some traditional threshold-based models, e.g., the Artificial Thermal (AT)-ACCA and Fixed Temperature (FT)-ACCA, the C implementation of Function of Mask (CFmask), the Landsat 8 Surface Reflectance Code (LaSRC), and the See5 algorithm (Foga et al., 2017). In addition, the RFmask algorithm shows a comparable performance compared with recently developed deep learning algorithms based on convolutional

neural networks, e.g., MSCFF (Li et al., 2019), and SegNet (Chai et al., 2019). Note that such comparisons may not be entirely correct because the used validation images are not exactly the same. However, these deep learning models are hard to be reconstructed because of the unknown number of network layers and parameter settings in the model training. Therefore, more comprehensive comparisons in model accuracy and operating efficiency between our and previously developed algorithms will be carried out in our future study. Nevertheless, these results can illustrate that our study provided a new perspective in rapidly and efficiently automatic cloud detection for Landsat imagery.

**Table 4.** Comparison in cloud detection algorithm accuracies (Units: %) from previous studies with the same L7 Irish and L8 Biome reference masks.

| Study | Algorithm | $A_O$ (%) | $A_P$ (%) | $A_U$ (%) | *OE* (%) | *CE* (%) | Satellite | Reference |
|---|---|---|---|---|---|---|---|---|
| 1 | Fmask | 90.7 | 84.4 | 99.8 | - | - | Landsat 7 | Zhu et al., 2015 |
| | | 93.3 | 95.0 | 97.0 | - | - | Landsat 8 | |
| 2 | AT-ACCA | - | 87.5 | - | 12.4 | 9.8 | Landsat 7–8 | Foga et al., 2017 |
| | FT-ACCA | - | 74.2 | - | 8.07 | 3.8 | | |
| | CFmask | - | 89.3 | - | 2.7 | 12.0 | | |
| | LaSRC | - | 73.1 | - | 4.7 | 23.9 | | |
| | See5 | - | 85.8 | - | 14.8 | 5.7 | | |
| 3 | CDAL8 | - | 88.8 | - | 13.0 | 17.6 | Landsat 8 | Oishi et al., 2018 |
| 4 | SegNet | 94.3 | 86.5 | 91.3 | - | - | Landsat 7 | Chai et al., 2019 |
| | | 94.0 | 93.1 | 94.5 | - | - | Landsat 8 | |
| 5 | MSCFF | 94.5 | 93.6 | 92.5 | - | - | Landsat 7 | Li et al., 2019 |
| | | 95.0 | 95.1 | 93.9 | - | - | Landsat 8 | |

## 6. Summary and conclusion

There are currently many operational algorithms for Landsat satellites. However, due to the high spatial resolution and the smaller amount of spectral information from instruments onboard the Landsat satellites, traditional threshold-based methods still face great challenges in detecting broken and thin clouds, especially over bright surfaces. Therefore, in this study, we propose a new object-oriented Random-Forest-based cloud mask (RFmask) algorithm, which combines the pixel-based RF ensemble learning approach and object-oriented Super-pixels Extracted via Energy-Driven

Sampling (SEEDS) super-pixel segmentation technology, for the high-resolution imageries aboard the Landsat series of satellites. For this purpose, cloudy- and clear-sky pixels over diverse underlying surfaces were stratified collected from uniformly distributed Landsat images around the world, and a prior database is constructed. Then a variety of spectral features in distinguishing clouds from different land cover types are derived as inputs for model training and building. The preliminary cloud detection results are further processed using the super-pixel segmentation and validated against the USGS Landsat 7 and 8 cloud-cover assessment datasets.

The validation and comparison results show that the RFmask algorithm can accurately detect most clouds over diverse land surface types. The new algorithm works well in identifying broken clouds and thin clouds with few omissions. It can also more correctly distinguish most clouds from bright surfaces (e.g., urban, barren, and snow/ice) with few misjudgments. In general, the estimated cloud covers correlate well with the validation cloud masks ($R^2$ = 0.99), showing small estimation uncertainties (i.e., MAE = 2.57% and RMSE = 4.29%). The RFmask algorithm detects clouds well with an overall accuracy of 95.3%, small omission error of 8.3%, and commission error of 6.4%, respectively. The RFmask algorithm appears to outperform traditional threshold-based methods and be comparable to deep learning approaches presented in previous studies. This illustrates that the RFmask algorithm is robust and can significantly improve the detection of thin and broken clouds, which is of great importance for quantitative applications in the surface and atmospheric fields for Landsat missions. However, cloud shadow detection became more challenging than cloud and was not considered in the current study, and will be explored in our subsequent research. Furthermore, the RFmask algorithm will be considered to extend for other high-spatial-resolution sensors in our future studies.

**Acknowledgments**

This work was supported by the National Key R&D Program of China (2017YFC1501702), the National Science Foundation of China (91544217), and the U.S. National Science Foundation (AGS1534670 and AGS1837811). The USGS Landsat 7 and 8 imagery and cloud validation masks are available from https://landsat.usgs.gov/landsat-7-cloud-cover-assessment-validation-data and https://landsat.usgs.gov/landsat-8-cloud-cover-assessment-validation-data.

618

619 **References**

620 Ackerman, S., Holz, R., Frey, R., Eloranta, E., Maddux, B., McGill, M. (2008). Cloud detection
621 with MODIS. Part II: Validation. Journal of Atmospheric & Oceanic Technology, 25(7), 1073-
622 1086

623 Andreae, M., & Rosenfeld, D. (2008). Aerosol–cloud–precipitation interactions. Part 1. The nature
624 and sources of cloud-active aerosols. Earth Science Reviews, 89(1), 13–41.

625 Arvidson, T., Gasch, J., & Goward, S. N. (2001). Landsat's 7 long-term acquisition plane – an
626 innovative approach to building a global imagery archive. Remote Sensing of Environment,
627 78, 13–26.

628 Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015). Example-dependent cost-sensitive decision
629 trees. Expert Systems with Applications, 42(19), 6609–6619.

630 Bergh, M., Boix, X., Roig, G., Capitani, B., & Gool, L. (2012). SEEDS: superpixels extracted via
631 energy-driven sampling. International Journal of Computer Vision, 111(3), 298–314.

632 Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.

633 Breiman, L., Friedman, J., Olshen, R., Stone, C. (1984). Classification and Regression Trees.
634 Belmont, CA: Wadsworth.

635 Chai, D., Newsam, S., Zhang, H., Qiu, Y., & Huang, J. (2019). Cloud and cloud shadow detection in
636 Landsat imagery based on deep convolutional neural networks. Remote Sensing of
637 Environment, 225, 307–316.

638 Calle, M., and Urrea, V.: Letter to the editor: satiability of random forest importance measures,
639 Briefings Bioinformation, 12(1), 86–89, 2011.

640 Chander, G., Markham, B., & Helder, D. (2009). Summary of current radiometric calibration
641 coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors. Remote Sensing of
642 Environment, 113(5), 893–903.

643 Cohen, J. A. (1960). coefficient of agreement for nominal scales. Educational and Psychological
644 Measurement, 20, 37–46

645 Frantz, D., Haß, E., Uhl, A., Stoffels, J., Hill, J. (2018). Improvement of the Fmask algorithm for
646 Sentinel-2 images: Separating clouds from bright surfaces based on parallax effects, Remote

Sensing of Environment, 215, 471-481.

Foga, S., Scaramuzza, P., Guo, S., Zhu, Z., Dilley, R., Beckmann, T., … Laue, B. (2017). Cloud detection algorithm comparison and validation for operational Landsat data products. Remote Sensing of Environment, 194, 379–390.

Frey, R., Ackerman, S., Liu, Y., Strabala, K., Zhang, H., Key, J., Wang, X. (2008). Cloud detection with MODIS. Part I: Improvements in the MODIS cloud mask for Collection 5. Journal of Atmospheric & Oceanic Technology, 25(7), 1057-1072.

Gao, B., and Li, R. (2000). Quantitative Improvement in the Estimates of NDVI Values from Remotely Sensed Data by Correcting Thin Cirrus Scattering Effects. Remote Sensing of Environment, 74, 494–502.

Gao, B., and Li, R. (2017). Removal of Thin Cirrus Scattering Effects in Landsat 8 OLI Images Using the Cirrus Detecting Channel, Remote Sensing, 9, 834; doi:10.3390/rs9080834

Gao, B., Yang, P., Han, W., Li, R., & Wiscombe, W. (2002). An algorithm using visible and 1.38-μm channels to retrieve cirrus cloud reflectances from aircraft and satellite data. IEEE Transactions on Geoscience and Remote Sensing, 40(8), 1659–1668.

Girolamo, L., & Wilson, M. (2003). A first look at band-differenced angular signatures for cloud detection from MISR. IEEE Transactions on Geoscience & Remote Sensing, 41(7), 1730-1734.

Gomez-Chova, L., Camps-Valls, G., Galpe-Maravilla, J., Guanter, L., & Moreno, J. (2007). Cloud-screening algorithm for ENVISAT/MERIS multispectral images. Geoscience and Remote Sensing, 45(12), 4105–4118.

Goodwin, Nicholas R., Collett, Lisa J., Denham, Robert J., Flood, Neil, & Tindall, Daniel. (2013). Cloud and cloud shadow screening across Queensland, Australia: an automated method for Landsat TM/ETM+ time series. Remote Sensing of Environment, 134, 50-65.

Hu, X., Belle, J. H., Meng, X., Wildani, A., Waller, L., Strickland, M., & Liu, Y. (2017). Estimating $PM_{2.5}$ concentrations in the conterminous United States using the random forest approach. Environmental Science & Technology, 51(12), 6936–6944.

Irish, R. (2000). Landsat-7 automatic cloud cover assessment algorithms for multispectral, hyperspectral, and ultraspectral imagery. The International Society for Optical Engineering, 4049, 348–355.

Irish, R., Barker, J., Goward, S., & Arvidson, T. (2006). Characterization of the Landsat-7 ETM+ Automated Cloud-Cover Assessment (ACCA) algorithm. Photogrammetric Engineering and Remote Sensing, 72(10), 1179–1188.

Jiang, R., Tang, W., Wu, X., and Fu, W. (2009). A random forest approach to the detection of epistatic interactions in case-control studies, BMC Bioinformatics, 10(2), 135–135.

Jin, S., Homer, C., Yang, L., Xian, G., Fry, J., & Danielson, P., et al. (2013). Automated cloud and shadow detection and filling using two-date Landsat imagery in the USA. International Journal of Remote Sensing, 34(5), 1540-1560.

Ju, J., & Roy, D. (2008). The availability of cloud-free Landsat ETM+ data over the conterminous United States and globally. Remote Sensing of Environment, 112 (3), 1196–1211.

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. BMC Medical Informatics and Decision Making, 11(1). https://doi.org/10.1186/1472-6947-11-51

Li, T., Shen, H., Yuan, Q., Zhang, X., and Zhang, L. (2017). Estimating ground-level $PM_{2.5}$ by fusing satellite and station observations: A geo-intelligent deep learning approach. Geophysical Research Letters, 44(23), 11985–11993.

Li, Z. Q., Lau, W. K., Ramanathan, V., Wu, G., Ding, Y., & Manoj, M. G., et al. (2016). Aerosol and monsoon climate interactions over Asia. Reviews of Geophysics, 54(1-4).

Li, Z. W., Shen, H., Cheng, Q., Liu, Y., You, S., He, Z. (2019). Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors. ISPRS Journal of Photogrammetry and Remote Sensing 150, 197–212.

Lin, J., Feng, X., Xiao, P., Li, H., Wang, J., and Li, Y. (2012), Comparison of snow indexes in estimating snow cover fraction in a mountainous area in northwestern China, IEEE Geoscience and Remote Sensing Letters, 9, 725–729.

Lyapustin, A., Wang, Y., & Frey, R. (2008). An automatic cloud mask algorithm based on time series of MODIS measurements. Journal of Geophysical Research Atmospheres, 113, D16207.

https://doi.org/10.1029/2007JD009641

Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. Expert Systems with Applications, 42(10), 4621–4631.

Mamyrova, G., Terrance P. O'Hanlon, Monroe, J. B., Carrick, D. M., Malley, J. D., Adams, S., Reed, A. M., …Rider, L. G. (2014). Immunogenetic risk and protective factors for juvenile dermatomyositis in Caucasians. Arthritis and Rheumatology, 54(12), 3979–3987.

Nitze, I., Barrett, B., & Cawkwell, F. (2015). Temporal optimisation of image acquisition for land cover classification with random forest and MODIS time-series. International Journal of Applied Earth Observations and Geoinformation, 34(1), 136–146.

Oishi, Y., Ishida, H., & Nakamura, R. (2018). A new Landsat 8 cloud discrimination algorithm using thresholding tests. International Journal of Remote Sensing, 39(23), 9113–9133.

Qiu, S., He, B., Zhu, Z., Liao, Z., & Quan, X. (2017). Improving Fmask cloud and cloud shadow detection in mountainous area for Landsats 4–8 images. Remote Sensing of Environment, 199, 107–119.

Ramanathan, V., Cess, R., Harrison, E., Minnis, P., Barkstrom, B., Ahmad, E., & Hartmann, D. (1989). Cloud-radiative forcing and climate: results from the Earth Radiation Budget Experiment. Science, 243(4887), 57–63.

Rossow, W., & Dueñas, E. (2004). The International Satellite Cloud Climatology Project (ISCCP) web site. Bulletin of the American Meteorological Society, 85(2), 167–172.

Saunders, R., & Kriebel, K. (1988). An improved method for detecting clear sky and cloudy radiances from AVHRR data. International Journal of Remote Sensing, 9(1), 123–150.

Scaramuzza, P., Bouchard, M., & Dwyer, J. (2012). Development of the Landsat data continuity mission cloud-cover assessment algorithms. IEEE Transactions on Geoscience and Remote Sensing, 50(4), 1140–1154.

Shen, Y., Wang, Y., Lv, H., Qian, J. (2015). Removal of Thin Clouds in Landsat-8 OLI Data with Independent Component Analysis. Remote Sensing, 7, 11481–11500.

Stephens, G. L. (2005). Cloud feedbacks in the climate system: a critical review. Journal of Climate, 18(2), 237–273.

731 Sun, L., Mi, X., Wei, J., Wang, J., Tian, X., Yu, H., and Gan, P. (2017). A cloud detection algorithm-
732 generating method for remote sensing data at visible to short-wave infrared wavelengths.
733 ISPRS Journal of Photogrammetry and Remote Sensing, 124, 70-88.
734 Sun, L., Wei, J., Bilal, M., Tian, X., Jia, C., Guo, Y., & Mi, X. (2015). Aerosol optical depth
735 retrieval over bright areas using Landsat 8 OLI images. Remote Sensing, 8(1).
736 https://doi.org/10.3390/rs8010023
737 Sun, L., Wei, J., Wang, J., Mi, X., Guo, Y., Lv, Y., …, & Tian, X. (2016). A universal dynamic
738 threshold cloud detection algorithm (UDTCDA) supported by a prior surface reflectance
739 database. Journal of Geophysical Research Atmospheres, 121(12), 7172–7196.
740 Sun, L., Zhou, X., Wei, J., Wang, Q., Liu, X., Shu, M., Chen, T., and Chi, Y. (2018). A New Cloud
741 Detection Method Supported By GlobeLand30 Data set. IEEE Journal of Selected Topics in
742 Applied Earth Observations and Remote Sensing, 11(10), 3628-3645.
743 U.S. Geological Survey. (2016). L7 Irish Cloud Validation Masks. U.S. Geological Survey data
744 release. https://doi.org/10.5066/F7XD0ZWC
745 U.S. Geological Survey. (2016). L8 Biome Cloud Validation Masks. U.S. Geological Survey, data
746 release. https://doi.org/10.5066/F7251GDH
747 van Beijma, S. V., Comber, A., & Lamb, A. (2014). Random forest classification of salt marsh
748 vegetation habitats using quad-polarimetric airborne SAR, elevation and optical RS
749 data. Remote Sensing of Environment, 149, 118–129.
750 Wang, Z., Lai, C., Chen, X., Bing, Y., Zhao, S., & Bai, X. (2015). Flood hazard risk assessment
751 model based on random forest. Journal of Hydrology, 527, 1130–1141.
752 Wei, J., Huang, B., Sun, L., Zhang, Z., Wang, L., & Bilal, M. (2017). A simple and universal aerosol
753 retrieval algorithm for Landsat series images over complex surfaces. Journal of Geophysical
754 Research Atmospheres, 122, 13338–13355.
755 Wei, J., Sun, L., Peng, Y., Wang, L., Zhang, Z., Bilal, M., & Ma., Y. (2018). An improved high-
756 spatial-resolution aerosol retrieval algorithm for MODIS images over land. Journal of
757 Geophysical Research Atmospheres, 123(21), 12,291–12,307.
758 Wei, J., Huang, W., Li, Z., Xue, W., Peng, Y., Sun, L., & Cribb, M. (2019). Estimating 1-km-
759 resolution $PM_{2.5}$ concentrations across China using the space-time random forest approach.

760 Remote Sensing of Environment, 231, 111221. https://doi.org/10.1016/j.rse.2019.111221
761 Wei, J., Li, Z., Cribb, M., Huang, W., Xue, W., Sun, L., …, & Song, Y. (2020). Improved 1-km-
762 resolution $PM_{2.5}$ estimates across China using enhanced space-time extremely randomized
763 trees, Atmospheric Chemistry and Physics. https://doi.org/10.1016/j.rse.2019.05.022
764 Wieland, M., Li, Y., & Martinis, S. (2019). Multi-sensor cloud and cloud shadow segmentation with
765 a convolutional neural network. Remote Sensing of Environment, 230, 111203.
766 https://doi.org/10.1016/j.rse.2019.05.022
767 Wu, Z., Snyder, G., Vadnais, C., Arora, R., Babcock, M., Stensaas, G., … Newman, T. (2019). User
768 needs for future Landsat missions. Remote Sensing of Environment, 231.
769 Wulder, M., Loveland, T., Roy, D., Crawford, C., Masek, J., Woodcock, C., … Zhu, Z. (2019).
770 Current status of Landsat program, science, and applications. Remote Sensing of Environment,
771 225, 127–147.
772 Yang, Y., Girolamo, L. D., & Mazzoni, D. (2007). Selection of the automated thresholding
773 algorithm for the multi-angle imaging spectroradiometer radiometric camera-by-camera cloud
774 mask over land. Remote Sensing of Environment, 107(1-2), 159-171.
775 Zhai, H, Zhang, H, Zhang, L., & Li, P. (2018). Cloud/shadow detection based on spectral indices for
776 multi/hyperspectral optical remote sensing imagery. ISPRS Journal of Photogrammetry and
777 Remote Sensing, 144, 235–253.
778 Zhang, Y., Rossow, W. B., Lacis, A. A., Oinas, V., & Mishchenko, M. I. (2004). Calculation of
779 radiative fluxes from the surface to top of atmosphere based on ISCCP and other global data
780 sets: refinements of the radiative transfer model and the input data. Journal of Geophysical
781 Research Atmospheres, 109(19), 19105. https://doi.org/10.1029/2003JD004457
782 Zhu, X., & Helmer, E. (2018). An automatic method for screening clouds and cloud shadows in
783 optical satellite image time series in cloudy regions. Remote Sensing of Environment, 214,
784 135–153.
785 Zhu, Z., & Woodcock, C. (2012). Object-based cloud and cloud shadow detection in Landsat
786 imagery. Remote Sensing of Environment, 118, 83–94.
787 Zhu, Z., Wang, S., & Woodcock, C. (2015). Improvement and expansion of the Fmask algorithm:
788 cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images. Remote

789 Sensing of Environment, 159, 269–277.
790