Remote Sensing of Environment 248 (2020) 112005

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

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

Jing Wei[a,b], Wei Huang[c], Zhanqing Li[b,*], Lin Sun[d], Xiaolin Zhu[e], Qiangqiang Yuan[f], Lei Liu[g], Maureen Cribb[b]

[a] State Key Laboratory of Remote Sensing Science, College of Global Change and Earth System Science, Beijing Normal University, Beijing, China
[b] Department of Atmospheric and Oceanic Science, Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, USA
[c] State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing, China
[d] College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao, China
[e] Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong
[f] School of Geodesy and Geomatics, Wuhan University, Wuhan, China
[g] College of Earth and Environmental Sciences, Lanzhou University, Lanzhou, China

ARTICLE INFO

Keywords:
Landsat
Cloud detection
RFmask
Random forest
SEEDS
Superpixel segmentation

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 this challenge, we developed a new algorithm called the Random-Forest-based cloud mask (RFmask), which can improve the accuracy of cloud identification from 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 chose the stratified sampling method to pre-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 reflection characteristics of different land cover types using the top-of-atmosphere reflectance and brightness temperature. These are then used as inputs to the RF model for training and establishing a preliminary cloud detection model. Finally, the Super-pixels Extracted via Energy-Driven Sampling (SEEDS) segmentation approach is applied to re-process the preliminary classification results in order to obtain the final cloud detection results. The RFmask detection results are evaluated against the globally distributed United States Geological Survey (USGS) cloud-cover assessment validation products. The average overall accuracy for RFmask cloud detection reaches 93.8% (*Kappa* coefficient = 0.77) with an omission error of 12.0% and a commission error of 7.4%. The RFmask algorithm is able to identify broken and thin clouds over both dark and bright surfaces. The new model generally outperforms other methods that are compared here, especially over these challenging scenes. The RFmask algorithm is not only accurate but also computationally efficient. It is potentially useful for a variety of applications in using Landsat data, especially for monitoring land cover and land-use changes.

1. Introduction

Clouds are ubiquitous with an annual and global mean amount of ~66%, especially in the tropics (IPCC, 2013; Ju and 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 and Rosenfeld, 2008; Li et al., 2016; Ramanathan et al., 1989; Stephens, 2005; Guo et al., 2017). The presence of clouds hinders the quantitative extraction of surface and atmospheric parameters for such purposes as classification and monitoring of land-use and land-cover changes (Sun et al., 2016a; Wulder et al., 2019; Zhu and Woodcock, 2012), and retrievals of aerosol optical properties (Li et al., 2009; Wei et al., 2017, 2018; Su et al., 2018, 2020), and fine particulate matters (Wei et al., 2019a, 2019b).

Cloud detection is an essential step for many remote sensing applications (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

* Corresponding author.
E-mail address: zli@atmos.umd.edu (Z. Li).

https://doi.org/10.1016/j.rse.2020.112005
Received 2 September 2019; Received in revised form 13 July 2020; Accepted 14 July 2020

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 because their spectral signals come from both clouds and underlying surfaces (e.g., vegetation, soil, and water), especially over bright surfaces (Gao et al., 2002; Irish, 2000; Rossow and Dueñas, 2004; Sun et al., 2016a, 2017).

Over the years, many cloud identification methods have been proposed for applications with various satellite imaging sensors such as the Advanced Very High Resolution Radiometer (AVHRR, Rossow and Dueñas, 2004; Saunders and Kriebel, 1988), the MODerate-resolution Imaging Spectroradiometer (MODIS; Ackerman et al., 2008; Frey et al., 2008; Lyapustin et al., 2008), and the Multi-angle Imaging SpectroRadiometer (MISR, 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 onboard different satellite platforms, e.g., the U.S. Landsat, the French Sentinel, and Satellite pour l'Observation de la Terre (SPOT), and the Chinese Huan Jing (HJ) and Gaofen (GF) satellites, the spectral channels are usually much fewer, posing more challenges 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 (Sun et al., 2016b; Wei et al., 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 widely used generations of Landsat sensors: the Thematic Mapper (TM) onboard Landsat 4/5 (launched in 1984), the Enhanced Thematic Mapper Plus (ETM+) onboard the Landsat 7 (launched in 1999), and the Operational Land Imager and Thermal Infrared Sensor (OLI/TIRS) onboard the Landsat 8 (launched in 2013). Table 1 provides detailed information about Landsat 4, 5, 7, and 8 satellites.

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 thermal-infrared bands (Irish et al., 2006). Subsequently, Zhu and Woodcock (2012) proposed a Function of mask (Fmask) algorithm to identify clouds from Landsat imagery through a series of spectral tests and probabilities of normalized temperature, spectral variability, and brightness. Sun et al. (2016a) developed a Universal Dynamic Threshold Cloud Detection (UDTCDA) algorithm to identify clouds based on a priori constructed surface reflectance database, minimizing the effects of mixed surfaces and improving the overall accuracy of cloud recognition. Orishi et al. (2008) proposed a propose a Cloud Discrimination Algorithm for Landsat 8 (CDAL8) according to multiple judgment tests. 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. Beside traditional threshold-based methods, nowadays, artificial intelligence methods become hot, and several deep learning methods based on the convolutional neural network have been modified for detecting clouds, e.g., multi-scale convolutional feature fusion (MSCFF; Li et al., 2019), SegNet (Chai et al., 2019), and U-Net (Wieland et al., 2019).

Despite some unique merits of these algorithms, due to complex and changeable surface conditions, it is difficult to determine appropriate cloud recognition thresholds using a few spectral channels. Although traditional threshold-based methods are simple and easy to implement, they still suffer from large errors in identifying broken and thin clouds, especially over bright surfaces (Frantz et al., 2018; Irish et al., 2006; Oishi et al., 2018; Rossow and Dueñas, 2004). Deep learning approaches yield stronger data mining capabilities and can achieve more accurate cloud detection results. However, deep learning has more complex model parameters and needs to establish hundreds or thousands of internal network layers. As such, model adjustment and training time increase dramatically (Li et al., 2017, 2019; Wei et al., 2020; Zhai et al., 2018). In addition, model training and running are highly dependent on the computer configuration, making them difficult to be used operationally in data preprocessing for meteorological or environmental applications.

Therefore, a new, efficient, and accurate cloud detection algorithm is proposed here, combining the tree-based ensemble learning approach, i.e., Random Forest (RF, Breiman, 2001), and a superpixel image segmentation technology, namely, the Superpixels Extracted via Energy-Driven Sampling (SEEDS, Bergh et al., 2012), for application to Landsat imagery. Clear-sky and cloudy pixel database construction and spectral feature selection are first performed to provide adequate training samples. They are then used as inputs to the pixel-based RF model. Finally, the object-oriented SEEDS segmentation algorithm is applied for post-processing to reduce the noise and obtain the final cloud detection results. Sections 2 and 3 introduce the data source and RF-based cloud mask (RFmask) algorithm. Section 4 presents qualitative and quantitative validations of the RFmask results, and Section 5 and Section 6 give a discussion, summary and conclusion of this study.

2. Data source

In this study, two United States Geological Survey (USGS) cloud-cover assessment validation products, i.e., L7 Irish Cloud Validation Masks and L8 Biome Cloud Validation Masks, are selected and used for cloud detection experiments and validation (U.S. Geological Survey, 2016a, 2016b). The L7 Irish dataset includes a total of 206 Landsat 7 ETM+ (scan line corrector on) Level-1G scenes, evenly distributed in nine latitude zones around the world, including the austral, boreal, mid-

Table 1
Detailed information about the Landsat 4, 5, 7, and 8 satellites.

Landsat 4 and 5 TM			Landsat 7 ETM+			Landsat 8 OLI/TIRS			Band type
Band index	Wavelength (μm)	Spatial resolution	Band index	Wavelength (μm)	Spatial resolution	Band index	Wavelength (μm)	Spatial resolution	
–	–	–	–	–		1	0.435–0.451	30 m	Coastal
1	0.45–0.52	30 m	1	0.441–0.514	30 m	2	0.452–0.512	30 m	Blue
2	0.52–0.60	30 m	2	0.519–0.601	30 m	3	0.533–0.590	30 m	Green
3	0.63–0.69	30 m	3	0.631–0.692	30 m	4	0.636–0.673	30 m	Red
4	0.76–0.90	30 m	4	0.772–0.898	30 m	5	0.851–0.879	30 m	NIR
5	1.55–1.75	30 m	5	1.547–1.749	30 m	6	1.566–1.651	30 m	MIR
6	10.40–12.50	120 m	6	10.31–12.36	60 m	10	10.60–11.19	100 m	TIR-1
7	2.08–2.35	30 m	7	2.064–2.345	30 m	7	2.107–2.294	30 m	SWIR
–	–	–	8	0.515–0.896	15 m	8	0.503–0.676	15 m	Panchromatic
–	–	–	–	–	–	9	1.363–1.384	30 m	Cirrus
–	–	–	–	–		11	11.50–12.51	100 m	TIR-2

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

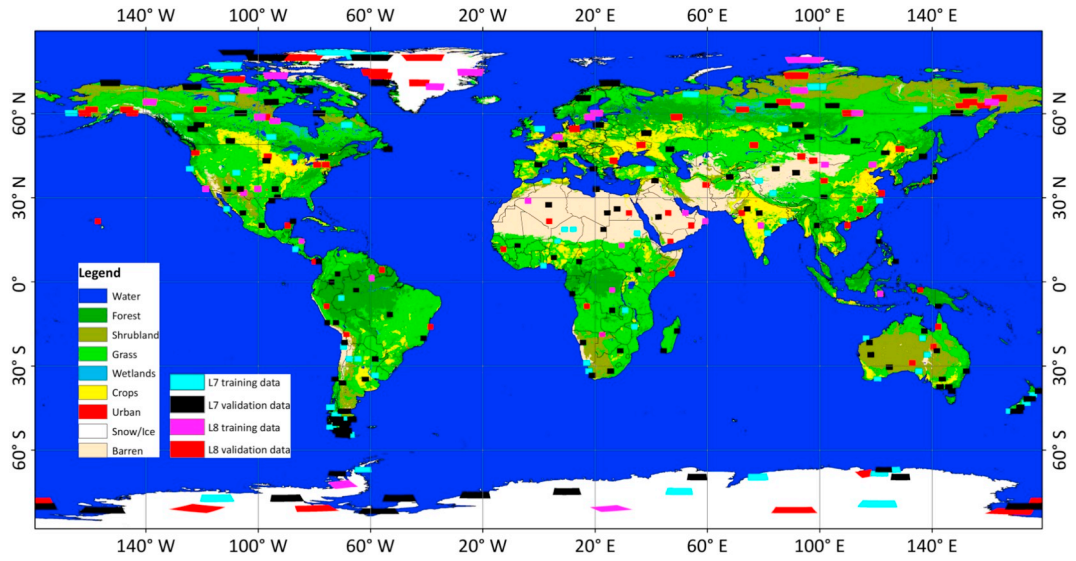

Fig. 1. The spatial distributions of all the Landsat 7 Irish training (marked in cyan) and validation (marked in black) images, and Landsat 8 Biome training (marked in pink) and validation (marked in red) images used in this study. The background map is obtained from the MODIS global land use cover product in 2016 (https://search.earthdata.nasa.gov). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

latitude, polar, subtropical, 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, evenly distributed globally and 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. Therefore, all the L7 Irish (206) and L8 Biome (96) scenes are employed in this study, as shown in Fig. 1.

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; Sun et al., 2016b). In this study, proposed is a new cloud mask algorithm for Landsat imagery (named the RFmask algorithm) is proposed, containing three key steps, namely, the pixel-based RF classification, the object-oriented SEEDS segmentation, and the post-classification processing.

3.1. RF classification

RF (Breiman, 2001) is a new and highly flexible machine learning algorithm with a wide range of applications. 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 and Aksakalli, 2015; Wang et al., 2015), and near-surface fine-particle estimation (Hu et al., 2017; Wei et al., 2019a). 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.

Different from traditional machine learning methods, RF is a non-linear algorithm that integrates multiple decision trees through the idea of ensemble learning. There are two key parts: one is “random”, which refers to building a decision tree by random sampling; the other is “forest”, which consists of hundreds of decision trees. Last, each tree acts as a weak classifier, and all the weak classifiers are majority voted to form a strong classifier. 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 Bootstrap aggregating (Bagging) resampling algorithm;
(2) In each node generated, D features are selected randomly and repeatably, used to split the sample set, respectively. 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 RF. The Classification And Regression Tree (CART, Breiman et al., 1984) algorithm is used for tree building;
(4) Test samples are predicted by the RF obtained from training, and the final classification results are determined by majority voting the classification results of all week classifiers.

Three main properties 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 (av_i) of any other class:

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

where f represents the indicator function, and $h_i(i = 1, 2, \ldots, k)$ 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,

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

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

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

The other properties are strength (s) and correlation ($\bar{\mathrm{r}}$), used to measure the accuracy and the dependence between individual classifiers and to derive the upper bound for the GE*:

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

The greater the strength and the smaller the correlation between decision trees, the more accurate the model is. Breiman (2001) provides detailed information about the RF algorithm.

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 first 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. Then the training images are stratified by evenly selecting images of various cloud amounts, < 35% (Clear), 35–65% (MidClouds), and > 65% (Cloudy), from both the Landsat 7 Irish and Landsat 8 Biome Cloud Cover Assessment Validation Database (U.S. Geological Survey, 2016a, 2016b) according to the land use cover. This can ensure that there are enough cloudy and clear-sky pixels in both model training and validation images with different cloud amounts.

In our study, all the cloudy and clear-sky pixels from the whole image are selected as training samples to build the pixel database. This ensures that the training samples include almost all kinds of clouds (e.g., thick, thin, and broken clouds) and clear skies over diverse land-use types. However, because the RF classification is performed at the pixel level, it is also feasible to randomly select an appropriate proportion from 60% to 100% of training samples from the whole image to improve the training efficiency. This is highly dependent on user demand and computer performance. Here, about one-third of the Landsat 7 Irish (~68 of 206 images) and Landsat 8 (~32 of 96 images) Biome images are stratified as training images. The remaining two-thirds are used as validation images, which are evenly distributed throughout the world (Fig. 1). Then all the clear-sky pixels and cloudy pixels collected from all the selected Landsat 7 Irish and Landsat 8 training images are used to construct the prior pixel database encompassing the nine main categories.

3.1.2. Feature attribute selection

The next important step is to select the feature attributes of the data samples used in the RF classification. The TOA reflectance of clouds is much higher than most typical ground objects (e.g., water, soil, vegetation, artificial buildings, and rocks) in visible channels under ideal conditions. Also used to detect clouds are the near-infrared (NIR), mid-infrared (MIR), and short-wave infrared (SWIR) channels due to noticeable differences between the reflectance of clouds and above-ground objects. However, snow and ice have similar spectral characteristics as clouds from short to medium wavelengths, so thermal infrared channels play an important role in distinguishing them due to their large differences in brightness temperature (Lin et al., 2012; Sun et al., 2016a; Zhu and Woodcock, 2012). More importantly, an additional cirrus channel was designed for Landsat 8 satellite, which has proven useful in detecting cirrus clouds (Gao and Li, 2000, 2017; Gao et al., 2002; Shen et al., 2015; Zhu et al., 2015). Thus Landsat 7 bands 1–7 and Landsat 8 bands 1–11 (excluding band 8) are selected as basic spectral features.

Thick clouds can be easily distinguished from pure ground objects. However, thin and broken clouds are less distinguishable from underlying surfaces, and the mixed land/cloud pixels formed by them are ubiquitous in remote sensing images. The spectral reflectance of different surface types can change according to the cloud amount. Remote sensing images can also become gradually blurred, affected by increasing air pollution, resulting in more complex spectral characteristics of ground objects. This largely increases the difficulties in separating clouds from different underlying surfaces through discrete spectral channels. This is also the main problem faced by traditional threshold-based methods (Sun et al., 2016a; Zhu and Woodcock, 2012).

More importantly, unlike other image-based machines or deep learning methods, RF is a supervised classification method, and its basic unit is the decision tree. It is composed of multiple decision trees whose construction highly depends on the input attributes. When the node is splitting in one decision tree construction, all input attributes are independently selected, and no other combination operation is performed internally. Therefore, the spectral absorption and reflection characteristics of these key land cover types previously mentioned 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 for the effect of the soil background and aerosol scattering; and the SWIR-based NDVI ($NDVI_{swir}$, Eq. (8)), which is not sensitive to aerosols:

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

$$RVI = \rho_{NIR}/\rho_{Red} \tag{6}$$

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

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

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:

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

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 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:

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

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

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:

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

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

In summary, there are a total of 17 (20) spectral features, including 7 (10) spectral channels of TOA reflectance and BT, and 10 common spectral indices for Landsat 7 (8) images.

3.1.3. Model training and validation

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

RF also has the important advantage of not needing cross-validation or a separate validation test because it can be evaluated internally, i.e.,

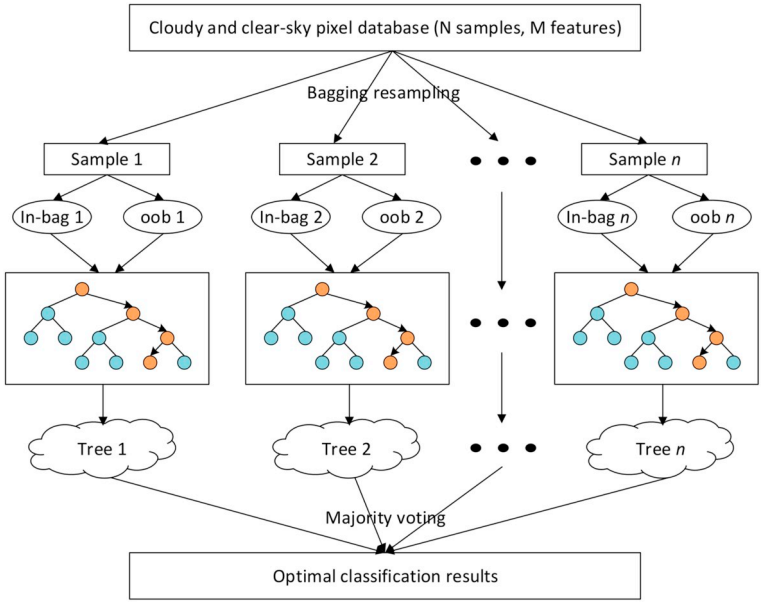

Fig. 2. Flowchart of the Random Forest classification.

an unbiased estimation of the error can be established during the generation process. During the model training, about one-third of the training samples (i.e., out-of-bag, or *oob*, samples) did not participate in the generation of the decision tree in each round of bagging sampling but were used to calculate the *oob* error, an unbiased estimation of the GE* of the RF. This is similar to the *k*-fold cross-validation, a calculation-intensive procedure. The *oob* score (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. Fig. 2 shows the flowchart of the RF classification.

Furthermore, RF allows for the evaluation of the importance of each feature during the classification, i.e., the importance score, calculated using the Gini index (Calle and Urrea, 2011; Jiang et al., 2009; Wei et al., 2020). Fig. 3 shows the importance score for each spectral feature in RF classification for Landsat imagery. Results show that most spectral features from the two different sensors play similar roles in detecting clouds, where discrete spectral channels are important, especially for thermal and shortwave bands. Due to the lack of some channels, the important scores of the visible bands of Landsat 7 are higher than those of Landsat 8. The cirrus band of Landsat 8 also plays an important role in (cirrus) cloud detection, consistent with conclusions reported in previous studies (Gao and Li, 2000, 2017; Gao et al., 2002; Shen et al., 2015; Zhu et al., 2015). Furthermore, spectral indices still impact cloud identification, especially those used to enhance bright surfaces (i.e., NDBI, TR_{nm}, and NDSI), appearing to be more important than some discrete spectral channels. However, "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, 2012). These results illustrate that these spectral indices are also important in tree-based ensemble learning approaches.

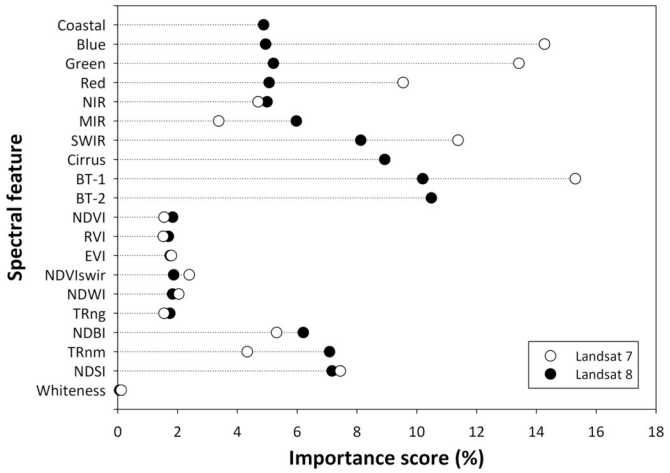

Fig. 3. Importance score of each spectral feature in cloud detection during the RF classification for Landsat 7 and 8 imagery, where BT-1 and BT-2 represent two brightness temperature bands for Landsat 8 satellite.

3.2. SEEDS segmentation

The RF clarification is performed on the pixel level, and did not consider the spatial characteristics of clouds. Therefore, the object-based SEEDS algorithm (Bergh et al., 2012) is selected to segment the remote sensing image by superpixel here. It is based on the simple hill-climbing optimization to extract superpixels, which starts with the initial superpixel partition and continuously optimizes the superpixel by modifying the boundary. The SEEDS superpixel segmentation mainly includes the following four steps:

(1) The red, green, and blue channels of Landsat imagery are combined into a red/green/blue (RGB) image and used as input to the SEEDS algorithm;
(2) Initialize the superpixel (Seed is the center of the superpixel) at the same interval (St), and all superpixels are rectangles of the same size fitting 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 $s \in S$ maximizes the Energy function (E), $E(s) > E(St)$, this pixel or a group of pixels can be regarded as a part of the superpixels;
(4) Iterate step 3 until it converges (default upper limit of times), and St is the final segmentation result.

In the SEEDS algorithm, E is expressed as

$$E(s) = H(s) + \gamma G(s) \quad (14)$$

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

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

$$c_{A_k}(j) = \frac{1}{Z} \sum_{i \in A_k} \delta(I(i) \in H_j) \quad (16)$$

where $\varphi()$ denotes the quality measure of the color distribution, $c_{Ak}(j)$ denotes the color histogram of the superpixels (A_k) in the jth bin, H_j denotes the colors in the jth bin of the histogram, $I(i)$ denotes the color of the ith 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

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

$$b_{N_i}(k) = \frac{1}{Z} \sum_{j \in N_i} \delta(j \in A_k) \quad (18)$$

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.

3.3. Post-classification processing

The RF classification and SEEDS segmentation are worked parallel to obtain the primarily cloud detection and image segmentation results, respectively. Last, the post-classification processing is performed to obtain the final result by overlying and combining these two results to improve the overall accuracy of cloud detection for Landsat imagery. For this purpose, the total number of cloudy and clear-sky pixels in preliminary cloud results within each superpixel is counted, and then the majority voting is used by adjusting to an appropriate decision threshold to determine the final class of all pixels of each entire superpixel. Fig. 4 illustrates a condensed flowchart of the RFmask cloud detection algorithm for Landsat imagery developed in this study.

3.4. Evaluation approaches

Calculated are total cloud amount (CA) from both the cloud detection results and the validation masks for each Landsat image, and their cloud amount difference (CAD; Sun et al., 2016a). 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 correlation coefficient (R^2), the mean absolute error (MAE), and the root-mean-square error (RMSE). 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 (*K*; 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 of pixels correctly predicted, and FP (false positive) and FN (false negative) denote the total number of pixels with an incorrect outcome from the cloud and clear-sky recognition, respectively (Li et al., 2019; Sun et al., 2016a).

4. Results

4.1. RFmask cloud detection results

4.1.1. RFmask results for Landsat 7 imagery

Fig. 5 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, upper two panels in each group of

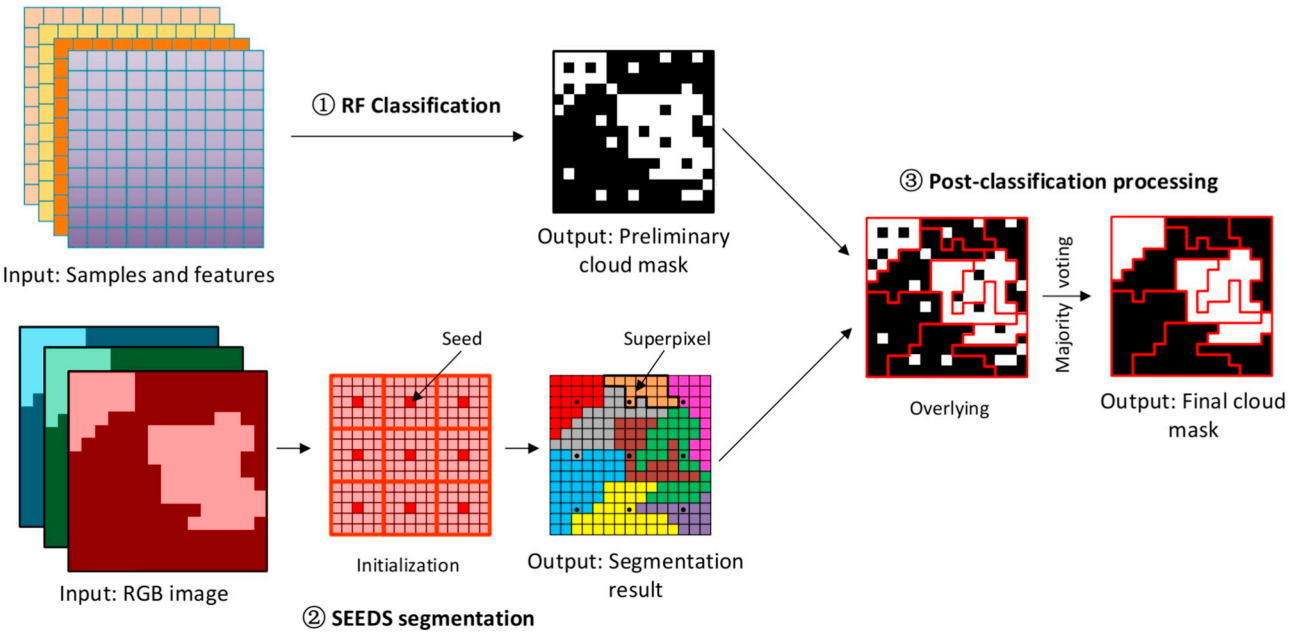

Fig. 4. Condensed flowchart of the RFmask cloud detection algorithm for Landsat imagery.

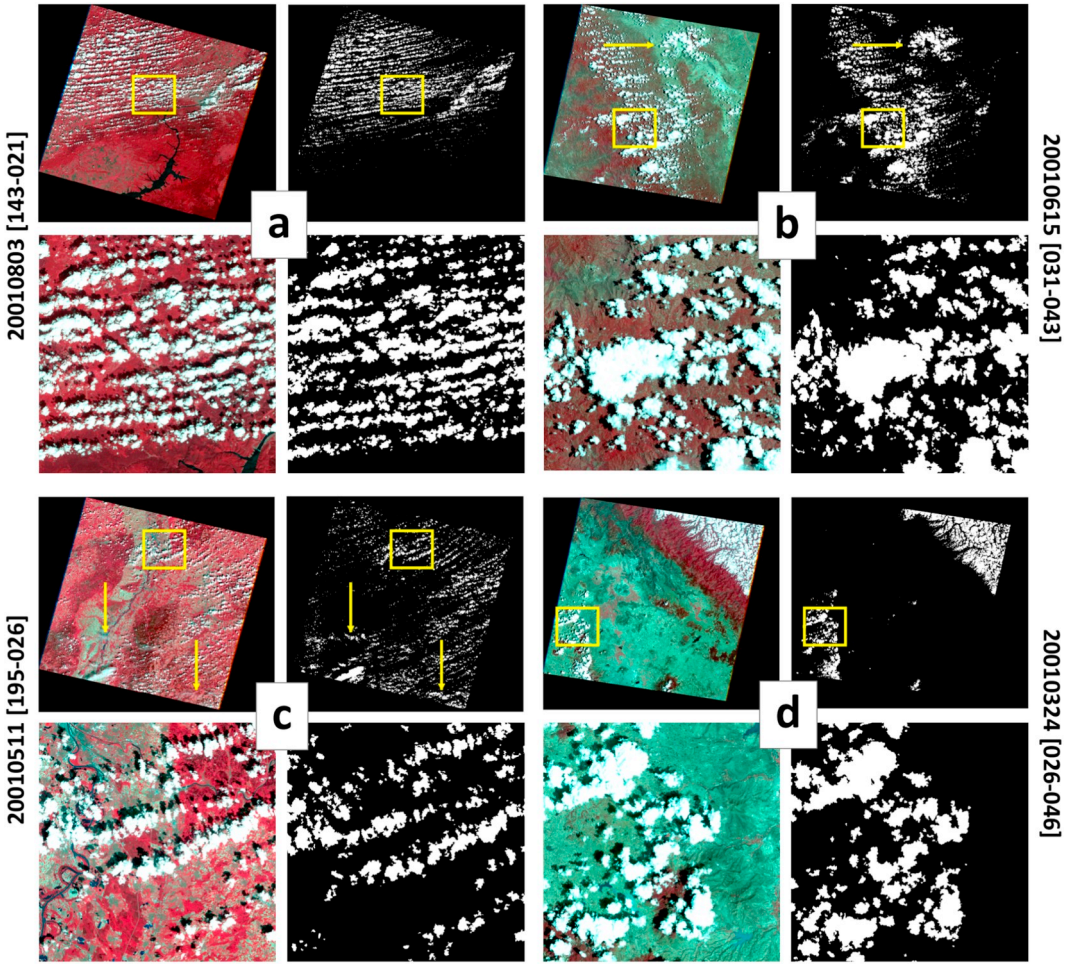

Fig. 5. Four examples of standard-false-color (RGB: Bands 4–3–2) composite images and cloud (masked as white) detection results for Landsat 7 full-scene and zoomed-in images (areas outlined by yellow boxes) over diverse underlying surfaces, where the right two images in each group show identified clouds in white. Yellow arrows point to clouds, and left- and right-side annotations indicate the acquisition times (yyyymmdd, where yyyy = year, mm = month, and dd = day) and orbital records (path-row) of the Landsat 7 images. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

four panels) and zoomed-in (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. 5a). The spatial distributions are almost identical between the RFmask cloud detection results and color composite images. The RFmask algorithm still works well as the amount of vegetation information decreases. For these vegetation-dominated land surface types, e.g., forest (Fig. 5b), cropland (Fig. 5c), and mountains (Fig. 5d), the RFmask algorithm performs well with small differences in cloud spatial distributions compared with the color composite images. Furthermore, the RFmask algorithm detects most clouds over those parts of the images with little vegetation, especially inland water (Fig. 5a), urban areas (Fig. 5c), and bare rock (Fig. 5b, d), suggesting acceptable classification results (pointed by yellow arrows in Fig. 5).

4.1.2. RFmask results for Landsat 8 imagery

Fig. 6 shows the full-scene and zoomed-in standard-false-color (RGB: Bands 5–4–3) composite images and RFmask results for Landsat 8 imagery. The RFmask cloud detection results are consistent with the true cloud distributions seen in the remote sensing images over densely vegetated areas (Fig. 6a-c). Clouds over darker surfaces, e.g., inland water and offshore areas, can also be accurately identified (Fig. 6a-c). The RFmask algorithm also performs well in coastline areas where extreme bright-dark reflectance differences exist (yellow ellipse in Fig. 6b). Clouds over urban buildings and roads are also more accurately identified (Fig. 6c). For barren land with little vegetation coverage, the RFmask algorithm still achieves better recognition results with few missed or misjudged cases. Notably, clear skies are not misidentified as clouds by the RFmask algorithm over bright bare surfaces deep inland (Fig. 6d). In general, the differences in cloud spatial distributions between RFmask results and color composite images are relatively small, and there are few incorrect or missing cloud identification pixels, indicating good classification results (pointed by yellow arrows in Fig. 6).

4.1.3. RFmask results over bright surfaces

Fig. 7 illustrates eight typical examples of the standard-false-color composite images (left panels in each group of eight panels) and RFmask results (right panels in each group of eight panels) 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 the 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; Oishi et al., 2018; Sun et al., 2016a; Zhu and Woodcock, 2012). This can lead to the misidentification of bright surfaces as clouds and to difficulties in accurately detecting thin clouds. The RFmask algorithm appears to be able to detect most clouds over less vegetated areas (Fig. 7a), bare rocks (Fig. 7b), deserts (Fig. 7c), and

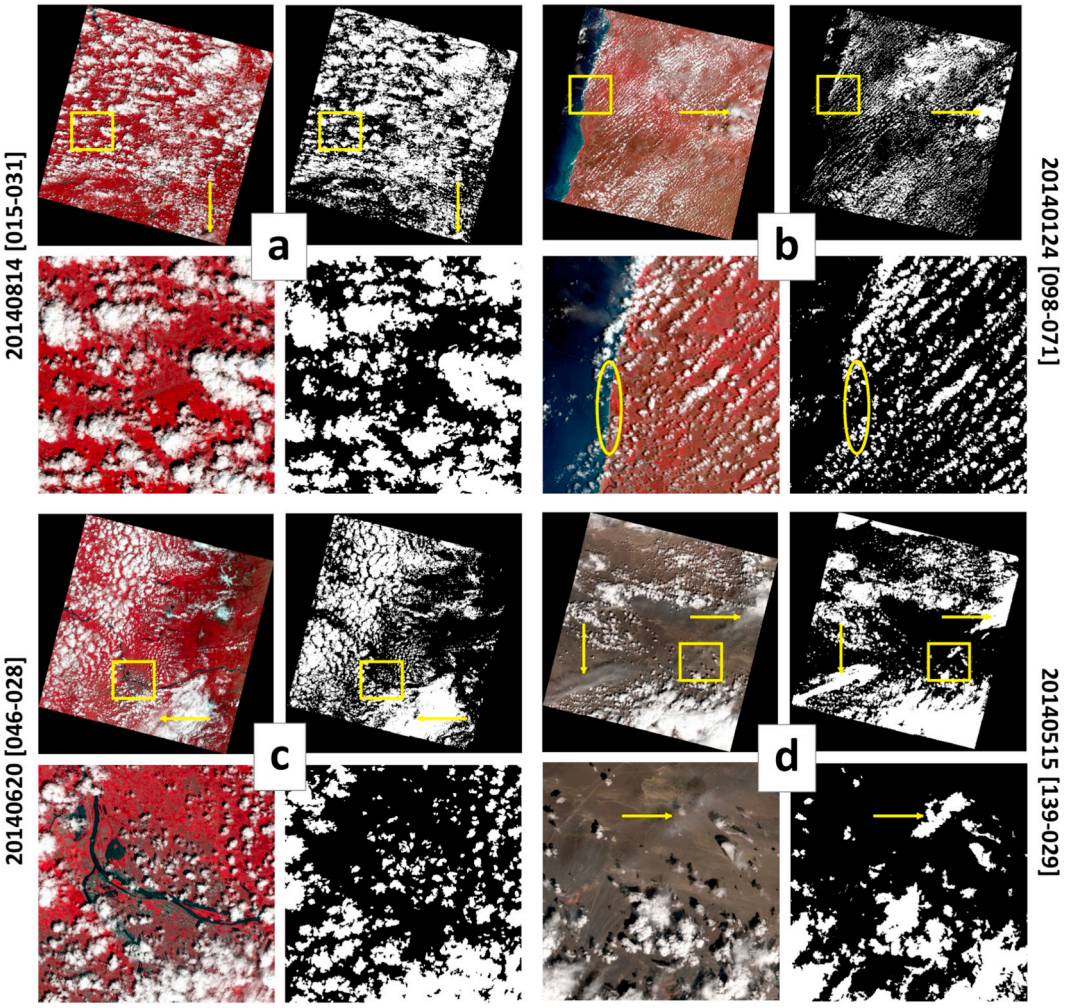

Fig. 6. Same as Fig. 5 but showing Landsat 8 imagery, while the left two images in each group are composited of bands 5, 4, and 3.

plateau mountains (Fig. 7d), with few cloud omissions and false recognitions, especially for thin and broken clouds (pointed by yellow arrows in Fig. 7). Furthermore, the RFmask algorithm is also capable of excluding very bright rocks (Fig. 7b, g), high-altitude snow/ice (Fig. 7d, e), rich mineral (Fig. 7f), and Gobi or rocky deserts (Fig. 7h) from cloud results. In particular, there is no misidentification of clouds over typical bright surfaces in the cloud-free Landsat images (Fig. 7e-h, pointed by red arrows in Fig. 7).

The RFmask algorithm is also tested for the most challenging scenes over permanent snow/ice surfaces in the polar regions. Fig. 8 presents the full-scene and zoomed-in false-color composite image scenes and clouds identified. For the sake of visual interpretation and comparison, the spectral Landsat data in MIR, NIR, and red channels are composited, which can better differentiate snow/ice from cloudy scenes. The RFmask algorithm can detect thick clouds more accurately, and most broken clouds (Fig. 8a, b, d). Moreover, it also works well in identifying thin clouds covering a large area as pointed by the yellow arrows (Fig. 8a-c). In addition, it can also differentiate bright sea ice and offshore reefs from clouds as pointed by the red arrows (Fig. 8d). Nevertheless, there are a considerable number of cloudy pixels misidentified or missed.

4.2. Accuracy assessment and analysis

4.2.1. Overall performance evaluation

Section 4.1 qualitatively examines the cloud detection results based on visual interpretations. Here, Landsat 7 Irish and Landsat 8 Biome validation data are selected to quantitatively evaluate the RFmask results. Table 2 summarizes the comparison between RFmask-derived cloud amount against USGS validation mask-determined cloud amount for all Landsat images, and Landsat 7 and 8 images, respectively. The estimated percentages of cloud cover are consistent with the USGS manually determined percentages of cloud cover (R^2 = 0.97) with a slope of 0.98, a y-intercept of 0.55, and a mean bias of −0.09, and the average MAE and RMSE values are 3.54% and 6.27% for Landsat imagery, respectively. More importantly, even when considering manual estimation uncertainties, approximately 90% of the RFmask results differ from the reference results by less than 5%. Similar validations and comparisons were also made separately for Landsat 7 and 8 images. The estimated cloud cover percentages derived from the RFmask algorithm and the USGS reference database correlate well (R^2 = 0.97 and 0.96 for Landsat 7 and 8, respectively), with strong slopes close to 1 and small intercepts. The MAEs are 3.03% and 4.64%, and RMSEs are 5.06% and 8.31% 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 prescreening.

Fig. 9 shows the frequency histograms of six accuracy indicators calculated from the confusion matrix for all Landsat RFmask results. The K for the RFmask algorithm is 0.77, and the average A_O reaches up to 93.8%. More than 79% and 72% of the RFmask results for Landsat imagery have A_O and K greater than 90% and 0.7, respectively. The average A_P and A_U are 88.0% and 89.1%, and in general, approximately 82% and 89% of the RFmask results have A_P and A_U values greater than

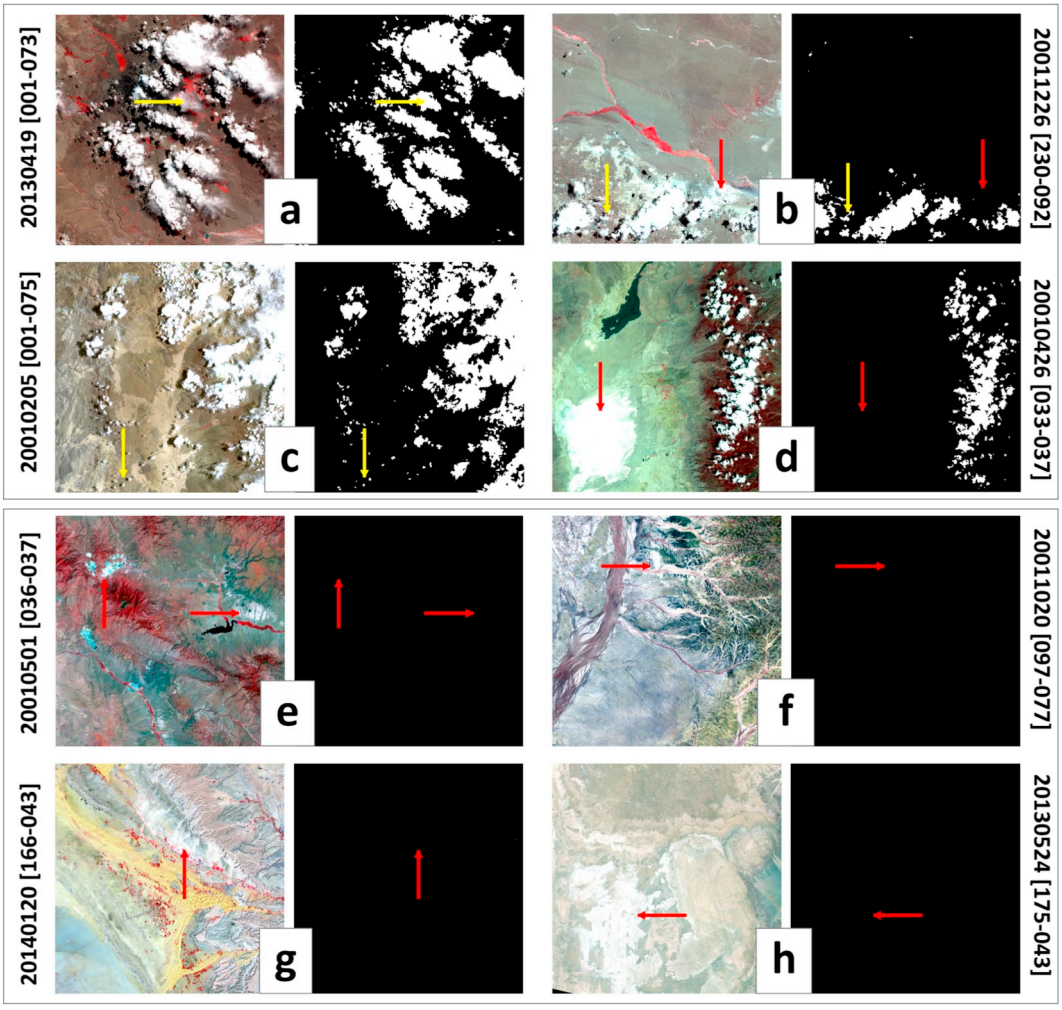

Fig. 7. Eight examples of standard-false-color composite images and cloud (masked as white) detection results for Landsat 7 and 8 zoomed-in images over diverse underlying surfaces. Yellow and red arrows point to clouds and bright surfaces, respectively. Left- and right-side annotations indicate the acquisition times (yyyymmdd, where yyyy = year, mm = month, and dd = day) and orbital records (path-row) of the Landsat images. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

80%, respectively. The average OE and CE are 12.0% and 7.4%, and more than 70% and 76% of the RFmask results have OE values < 15% and CE values < 10%, respectively. The RFmask algorithm also works well with Landsat 7 imagery with average K, A_O, A_P, and A_U values of 0.80, 93.9%, 88.1%, and 89.1%, respectively (Table 2), and small estimation errors, i.e., OE = 11.9% and CE = 6.8%. For Landsat 8 imagery, i.e., K = 0.72, A_O = 93.7%, A_P = 87.6%, and A_U = 89.0%, with an OE of 12.4% and CE of 8.8%. These results suggest that the RFmask algorithm can be applied to images from different Landsat satellites to detect clouds.

4.2.2. Evaluation on different land-use types

Next validated is the performance of the RFmask algorithm over different land-use types for Landsat imagery (Table 3). Results suggest that the RFmask algorithm performs well in detecting clouds over most dark surfaces, with overall high *Kappa* coefficients > 0.73, high overall accuracies > 92% and small commission errors < 9%, especially wetlands (e.g., *Kappa* = 0.80, A_O = 96.9%, OE = 9.7% and CE = 4.9%) and shrubland (e.g., *Kappa* = 0.79, A_O = 96.5%, OE = 9.4% and CE = 5.2%). The main reason is that there are noticeable spectral differences between clouds and these land-use types with low surface reflectances, which are relatively easy to distinguish, leading to fewer cloud recognition errors. The RFmask algorithm also shows a good ability in identifying clouds over urban areas (e.g., *Kappa* = 0.82, A_O = 96.1%, OE = 10.6% and CE = 3.8%) because they are mainly scattered near dark surfaces such as natural vegetation areas.

However, the performance of the RFmask algorithm overall decreases over brighter surfaces (e.g., *Kappa* < 0.7), due to the significant reduction in spectral differences, making clouds easier to be missed (OE > 10%) or misidentified (CE > 10%). For snow/ice surfaces, the overall accuracy of the RFmask algorithm is about 88%. Bright snow/ice pixels are more likely misidentified as clouds, leading to larger commission errors of approximately 21% (Table 3). This has long been recognized as a major challenge in cloud identification due to similar spectral characteristics in both visible and infrared channels, except some differences in near-IR and mid-IR regions (Li and Leighton, 1991). Given the inherent limitation of the Landsat spectral signals, the identification accuracy is reasonably high, especially in comparison with others (Chai et al., 2019; Li et al., 2019; Wieland et al., 2019), as the majority of clouds are identified for both extensive cloud decks, as well as for broken cloud cells. It is also worth noting that the manual visual interpretation is also subject to larger uncertainties, that may undermine our training and validation.

In general, the thick clouds are easy to be correctly identified over all types of land surfaces because thick clouds have strong signals. However, thin and broken clouds are irregular in shape and lesser in amount, usually occupying only a few pixels or even sub-pixels, and cirrus can be widespread with varying optical depths. Thus due to the

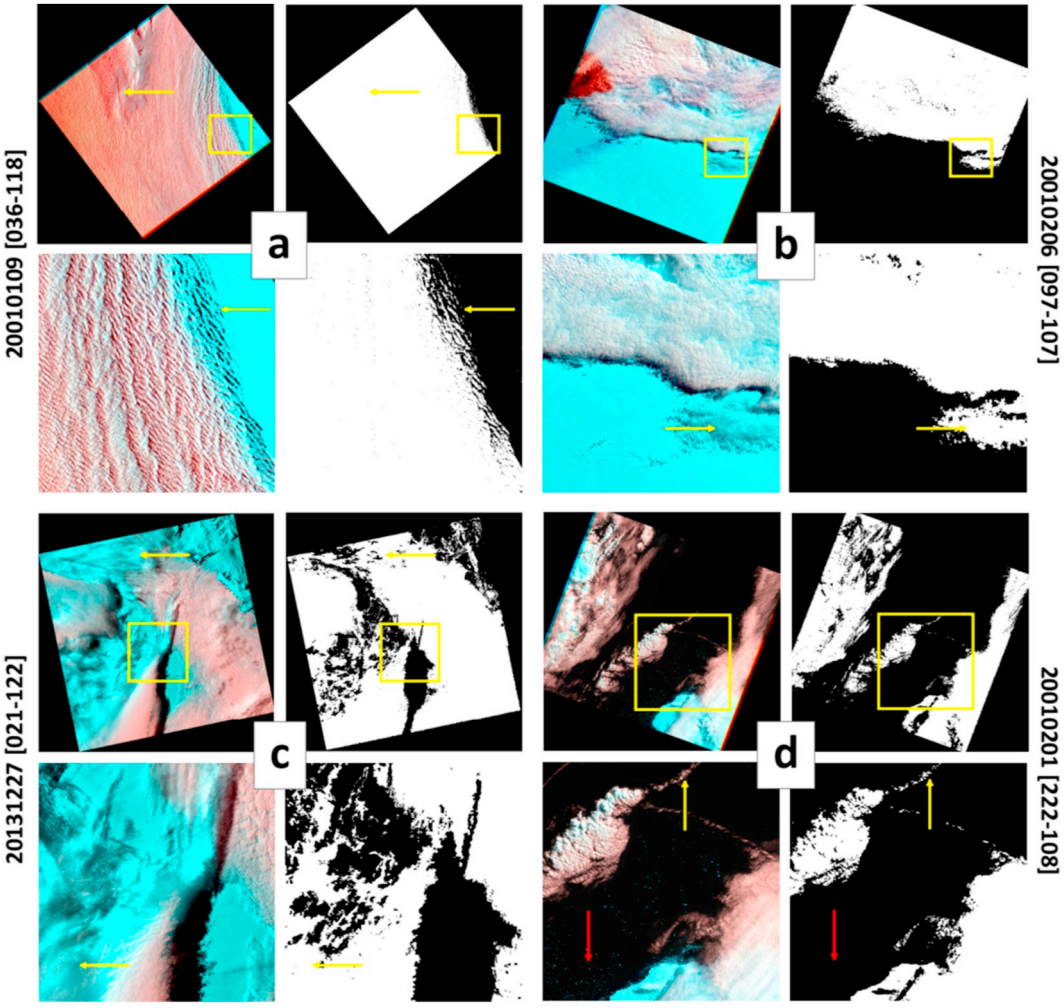

Fig. 8. Same as Fig. 5 but for snow/ice surfaces in the polar regions, while the left two images in each group are composited of bands 5, 4, and 3 for Landsat 7, and bands 6, 5, and 4 for Landsat 8, respectively.

Table 2
Statistics describing the evaluation results of cloud detection amount and accuracy for all images, and Landsat 7 and 8 imagery separately.

Cloud amount	N	R^2	Regression line		MAE (%)	RMSE (%)	CAD (%)
			Slope	Intercept			
All	202	0.97	0.98	0.55	3.54	6.27	−0.09
Landsat 7	138	0.97	0.99	0.06	3.03	5.06	−0.19
Landsat 8	64	0.96	0.97	1.81	4.64	8.31	0.14

Accuracy	N	*Kappa*	A_O (%)	A_P (%)	A_U (%)	OE (%)	CE (%)
All	202	0.77	93.8	88.0	89.1	12.0	7.4
Landsat 7	138	0.80	93.9	88.1	89.1	11.9	6.8
Landsat 8	64	0.72	93.7	87.6	89.0	12.4	8.8

influence of mixed pixels formed by clouds and different underlying surfaces, traditional threshold-based methods have difficulty setting up accurate thresholds and always fail to identify these clouds from Landsat images, especially over bright surfaces (Goodwin et al., 2013; Jin et al., 2013; Oishi et al., 2018; Sun et al., 2018; Zhu and Woodcock, 2012). 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. Importance of superpixel segmentation

Fig. 10 compares one cloud detection result using the SEEDS segmentation and one without the segmentation for Landsat imagery. Without the SEEDS segmentation (Fig.10b), there are many scattered pixels in cloudless places on the left and in the upper right corner of the image in the classification results, wrongly identified as cloud pixels (areas outlined by yellow circles), resulting in a lot of “salt-and-pepper” noise. There are also many pixels in the 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, where spatial autocorrelations and spatial texture information among the pixels are not considered, leading to inevitable noise in the classification results. However, superpixel segmentation is object-oriented and selected to address these issues. With the SEEDS segmentation (Fig.10c), the noise is removed, and patchy clouds are completely filled. In general, the final result appears to be more consistent in cloud distribution with the true-color Landsat image compared to the preliminary result via the visual interpretation (Fig. 10a). However, it

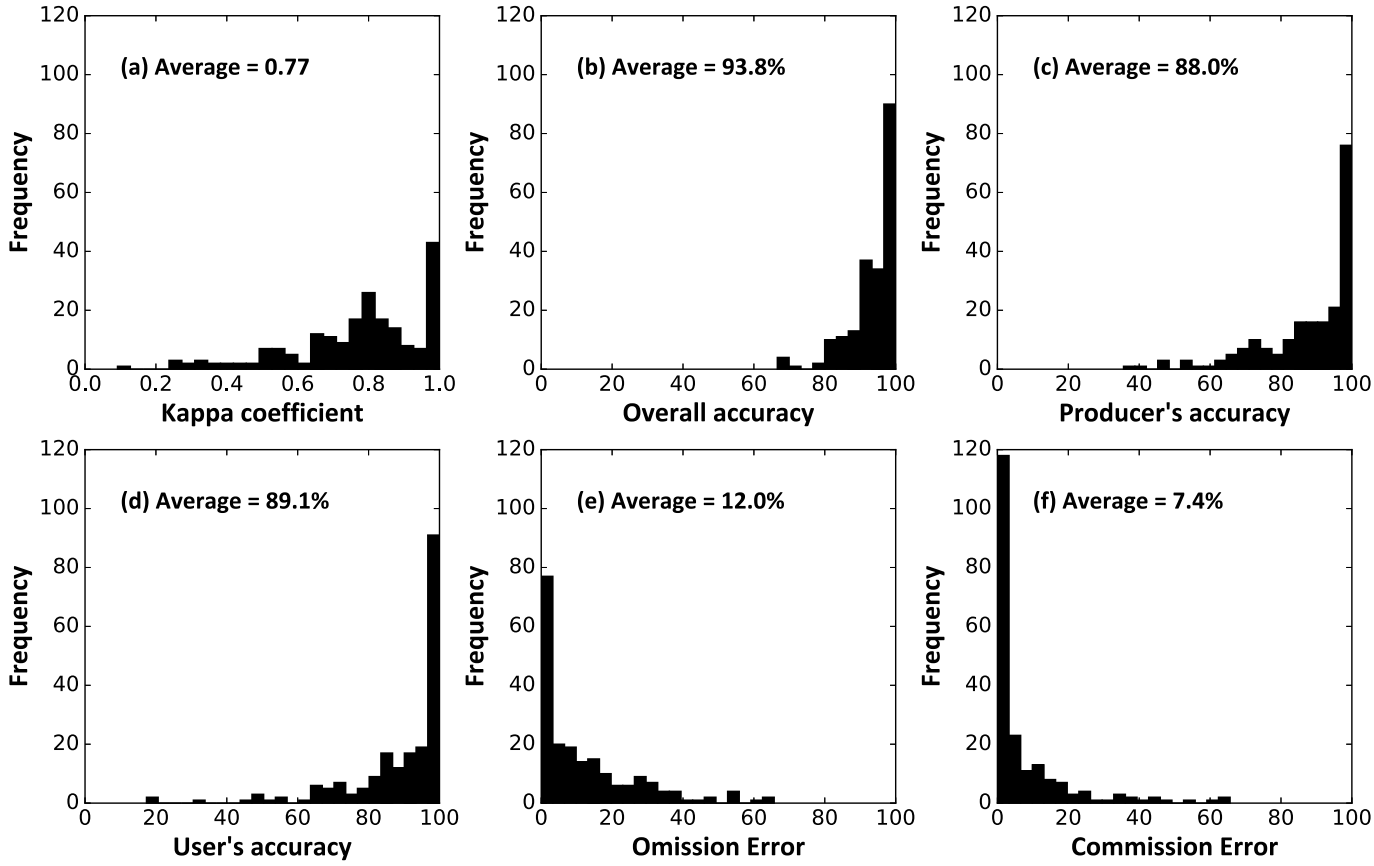

Fig. 9. Frequency histograms of RFmask cloud results from Landsat images in terms of (a) the *Kappa* coefficient, (b) the overall accuracy, (c) the producer's accuracy, (d) the user's accuracy, (e) the omission error, and (f) the commission error.

Table 3
Accuracy and error statistics of the RFmask algorithm in cloud detection over diverse land-use types for Landsat imagery.

Land-use type	*Kappa*	A_O (%)	A_P (%)	A_U (%)	OE (%)	CE (%)
Water	0.76	92.7	85.7	89.1	14.3	8.3
Forest	0.75	92.9	87.4	92.6	12.6	5.2
Shrubland	0.79	96.5	90.6	85.5	9.4	5.2
Grass	0.73	93.3	85.4	89.4	14.6	6.6
Wetlands	0.80	96.9	90.3	94.9	9.7	4.9
Crops	0.83	96.9	87.3	86.7	12.7	2.6
Urban	0.82	96.1	89.4	93.0	10.6	3.8
Barren	0.69	92.9	87.2	83.7	12.8	11.6
Snow/ice	0.67	87.6	89.7	79.4	10.3	20.9

should be noted that some small clouds containing a small number of pixels can also be excluded from the final result due to the defined decision threshold in merging the RF identification and image segmentation results in each superpixel during the post-classification processing. Nevertheless, it has little impact, and the quantitative comparison results show that after superpixel segmentation, all the accuracy evaluation metrics have been overall improved, and the omission and commission errors have been overall reduced for the final results with reference to the preliminary results. This suggests that the object-oriented SEEDS segmentation technology plays an important role in pixel-based classification by post-processing the preliminary cloud detection results, which benefited in improving the overall accuracy of cloud detection.

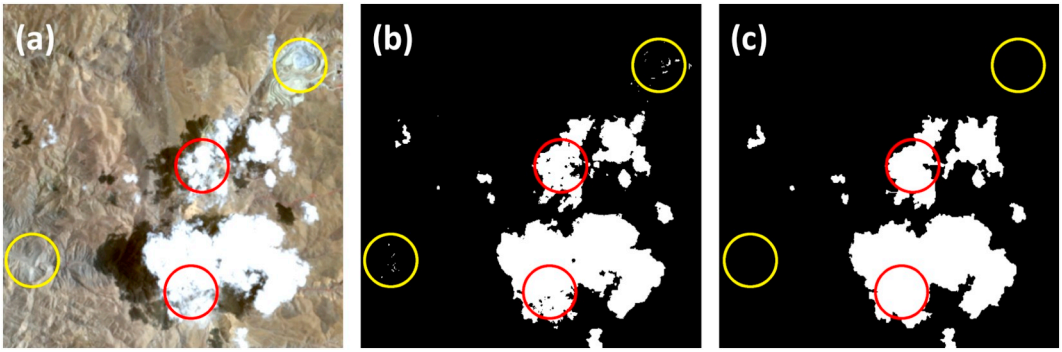

Fig. 10. An example of (a) a zoomed-in RGB combined image, (b) the preliminary cloud detection result without the SEEDS segmentation, and (c) the final cloud detection using the SEEDS segmentation for Landsat imagery.

Table 4
Comparison of cloud detection algorithm accuracies (unit: %) from previous studies and this study using 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	ACCA	83.8	–	–	6.66	5.9	Landsat 7–8	Foga et al., 2017
	AT-ACCA	87.5	–	–	12.4	9.8		
	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	
	CFmask	89.9	83.6	83.2			Landsat 7	
		84.6	82.7	74.3			Landsat 8	
5	MSCFF	94.5	93.6	92.5	–	–	Landsat 7	Li et al., 2019
		95.0	95.1	93.9	–	–	Landsat 8	
	Fmask	91.7	89.7	89.8			Landsat 7	
		89.6	85.8	93.0			Landsat 8	
	DeepLab	90.2	88.3	87.3			Landsat 7	
		87.7	91.3	81.4			Landsat 8	
	DCN	92.2	93.9	86.2			Landsat 7	
		92.4	95.6	87.3			Landsat 8	
6	RFmask	93.9	88.1	89.1	11.9	6.8	Landsat 7	This study
		93.7	87.6	89.0	12.4	8.8	Landsat 8	
		93.8	88.0	89.1	12.0	7.4	Landsat 7–8	

5.2. Comparison with related cloud studies

Here, we perform a simple comparison with existing cloud detection algorithms using the same validation sources of L7 Irish and L8 Biome reference masks (Table 4). Note that the reference images used are not totally the same, which could make the accuracy comparison not particularly fair. The result shows that the RFmask algorithm appears to outperform some traditional threshold-based models, e.g., the Fmask algorithm (Li et al., 2019; Zhu et al., 2015), the ACCA, Artificial Thermal (AT)-ACCA and Fixed Temperature (FT)-ACCA algorithms, the C implementation of Function of Mask (CFmask) algorithm, the Landsat 8 Surface Reflectance Code (LaSRC) algorithm (Chai et al., 2019; Foga et al., 2017), and the CDAL8 algorithm (Oishi et al., 2018). The RFmask algorithm also shows a comparable performance with recently developed machining learning algorithms, e.g., and the See5 algorithm (Foga et al., 2017) or deep learning algorithms, e.g., SegNet (Chai et al., 2019), MSCFF, DeepLab, and Deep Convolutional Network (DCN) (Li et al., 2019). In general, although the performance of our RFmask algorithm is not superior to some previous studies in all aspects, it is new, rapid, and automatic for cloud detection of Landsat imagery, especially in comparison to the time-consuming deep learning approaches.

6. Summary and conclusions

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

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 = 0.97), showing small estimation uncertainties (i.e., MAE = 3.54% and RMSE = 6.27%). The RFmask algorithm detects clouds well with an overall accuracy of 93.8%, a small omission error of 12.0%, and a commission error of 7.4%. 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.

Although the RFmask algorithm can achieve a high accuracy in cloud detection, there are still some deficiencies that need to be further explored. The separation of snow/ice from clouds is still a difficult task which warrants further investigations, and so is another challenging task of detecting cloud shadow. Due to the difficulty in the model reconstructing, a more comprehensive comparison of model accuracy and operating efficiency between our and previously developed algorithms will be considered in a future study. In addition, due to the lack of validation data for different types of clouds, comparisons of the performance among different algorithms are not feasible at present. Last, the RFmask algorithm will also be applied to other high-spatial-resolution sensors in future studies.

Credit author statement

Jing Wei: Conceptualization, Data curation, Formal analysis, Methodology, Writing – original draft. **Wei Huang**: Software, Validation. **Zhanqing Li**: Funding acquisition, Methodology, Supervision, Writing - review & editing. **Lin Sun**, **Xiaolin Zhu**, **Qiangqiang Yuan**, **Lei Liu**, and **Maureen Cribb**: Writing - review & editing.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgments

This work was supported by the National Key R&D Program of China (2017YFC1501702), and the National Natural Science Foundation of China (91544217). 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.

References

Ackerman, S., Holz, R., Frey, R., Eloranta, E., Maddux, B., McGill, M., 2008. Cloud detection with MODIS. Part II: validation. J. Atmos. Ocean. Technol. 25 (7), 1073–1086. https://doi.org/10.1175/2007JTECHA1053.1.

Andreae, M., Rosenfeld, D., 2008. Aerosol–cloud–precipitation interactions. Part 1. The nature and sources of cloud-active aerosols. Earth-Sci. Rev. 89 (1), 13–41. https://doi.org/10.1016/j.earscirev.2008.03.001.

Arvidson, T., Gasch, J., Goward, S.N., 2001. Landsat's 7 long-term acquisition plane – an innovative approach to building a global imagery archive. Remote Sens. Environ. 78, 13–26. https://doi.org/10.1016/S0034-4257(01)00263-2.

Bahnsen, A.C., Aouada, D., Ottersten, B., 2015. Example-dependent cost-sensitive decision trees. Expert Syst. Appl. 42 (19), 6609–6619. https://doi.org/10.1016/j.eswa.2015.04.042.

van Beijma, S.V., Comber, A., Lamb, A., 2014. Random forest classification of salt marsh vegetation habitats using quad-polarimetric airborne SAR, elevation and optical RS data. Remote Sens. Environ. 149, 118–129. https://doi.org/10.1016/j.rse.2014.04.010.

Bergh, M., Boix, X., Roig, G., Capitani, B., Gool, L., 2012. SEEDS: superpixels extracted via energy-driven sampling. Int. J. Comput. Vis. 111 (3), 298–314. https://doi.org/10.1007/s11263-014-0744-2.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth, Belmont, CA. https://doi.org/10.1002/widm.8.

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

Chai, D., Newsam, S., Zhang, H., Qiu, Y., Huang, J., 2019. Cloud and cloud shadow detection in Landsat imagery based on deep convolutional neural networks. Remote Sens. Environ. 225, 307–316. https://doi.org/10.1016/j.rse.2019.03.007.

Chander, G., Markham, B., Helder, D., 2009. Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors. Remote Sens. Environ. 113 (5), 893–903. https://doi.org/10.1016/j.rse.2009.01.007.

Cohen, J.A., 1960. Coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46. https://doi.org/10.1177/001316446002000104.

Foga, S., Scaramuzza, P., Guo, S., Zhu, Z., Dilley, R., Beckmann, T., et al., 2017. Cloud detection algorithm comparison and validation for operational Landsat data products. Remote Sens. Environ. 194, 379–390. https://doi.org/10.1016/j.rse.2017.03.026.

Frantz, D., Haß, E., Uhl, A., Stoffels, J., Hill, J., 2018. Improvement of the Fmask algorithm for Sentinel-2 images: separating clouds from bright surfaces based on parallax effects. Remote Sens. Environ. 215, 471–481. https://doi.org/10.1016/j.rse.2018.04.046.

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. J. Atmos. Ocean. Technol. 25 (7), 1057–1072. https://doi.org/10.1175/2008JTECHA1052.1.

Gao, B., Li, R., 2000. Quantitative improvement in the estimates of NDVI values from remotely sensed data by correcting thin cirrus scattering effects. Remote Sens. Environ. 74, 494–502. https://doi.org/10.1016/S0034-4257(00)00141-3.

Gao, B., Li, R., 2017. Removal of thin cirrus scattering effects in Landsat 8 OLI images using the cirrus detecting channel. Remote Sens. 9, 834. https://doi.org/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 Trans. Geosci. Remote Sens. 40 (8), 1659–1668. https://doi.org/10.1109/TGRS.2002.802454.

Girolamo, L., Wilson, M., 2003. A first look at band-differenced angular signatures for cloud detection from MISR. IEEE Trans. Geosci. Remote Sens. 41 (7), 1730–1734. https://doi.org/10.1109/TGRS.2003.815659.

Gomez-Chova, L., Camps-Valls, G., Galpe-Maravilla, J., Guanter, L., Moreno, J., 2007. Cloud-screening algorithm for ENVISAT/MERIS multispectral images. Geosci. Remote Sens. 45 (12), 4105–4118. https://doi.org/10.1109/TGRS.2007.905312.

Goodwin, R., Collett, L.J., Denham, R.J., Flood, N., Tindall, D., 2013. Cloud and cloud shadow screening across Queensland, Australia: an automated method for Landsat TM/ETM+ time series. Remote Sens. Environ. 134, 50–65. https://doi.org/10.1016/j.rse.2013.02.019.

Guo, J., Su, T., Li, Z., Miao, Y., Li, J., Liu, H., Xu, H., Cribb, M., Zhai, P., 2017. Declining frequency of summertime local-scale precipitation over eastern China from 1970 to 2010 and its potential link to aerosols. Geophys. Res. Lett. 44 (11), 5700–5708. https://doi.org/10.1002/2017GL073533.

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. Environ. Sci. Technol. 51 (12), 6936–6944. https://doi.org/10.1021/acs.est.7b01210.

IPCC, 2013. IPCC Third Assessment Report Chapter 7. Physical Climate Processes and Feedbacks (Atmospheric Processes and Feedbacks 7.2) (Report). International Panel on Climate Change. (Archived from the original on August 5, 2013. Retrieved August 24, 2013).

Irish, R., 2000. Landsat 7 automatic cloud cover assessment, proc. SPIE 4049, algorithms for multispectral, hyperspectral, and ultraspectral imagery VI. Int. Soc. Opt. Eng. 4049, 348–355. https://doi.org/10.1117/12.410358.

Irish, R., Barker, J., Goward, S., Arvidson, T., 2006. Characterization of the Landsat-7 ETM+ automated cloud-cover assessment (ACCA) algorithm. Photogram. Eng. Remote Sens. 72 (10), 1179–1188. https://doi.org/10.14358/PERS.72.10.1179.

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

Jin, S., Homer, C., Yang, L., Xian, G., Fry, J., Danielson, P., Townsend, P.A., 2013. Automated cloud and shadow detection and filling using two-date Landsat imagery in the USA. Int. J. Remote Sens. 34 (5), 1540–1560. https://doi.org/10.1080/01431161.2012.720045.

Ju, J., Roy, D., 2008. The availability of cloud-free Landsat ETM+ data over the conterminous United States and globally. Remote Sens. Environ. 112 (3), 1196–1211. https://doi.org/10.1016/j.rse.2007.08.011.

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

Li, T., Shen, H., Yuan, Q., Zhang, X., Zhang, L., 2017. Estimating ground-level $PM_{2.5}$ by fusing satellite and station observations: a geo-intelligent deep learning approach. Geophys. Res. Lett. 44 (23), 11,985–11,993. https://doi.org/10.1002/2017GL075710.

Li, Z., Leighton, H.G., 1991. Scene identification and its effect on cloud radiative forcing in the Arctic. J. Geophys. Res. Atmos. 96, 9175–9188. https://doi.org/10.1029/91JD00529.

Li, Z., Zhao, X., Kahn, R., Mishchenko, M., Remer, L., Lee, K.H., et al., 2009. Uncertainties in satellite remote sensing of aerosols and impact on monitoring its long-term trend: a review and perspective. Ann. Geophys. 27 (7), 2755–2770. https://doi.org/10.5194/angeo-27-2755-2009.

Li, Z., Lau, W., Ramanathan, V., Wu, G., Ding, Y., Manoj, M.G., et al., 2016. Aerosol and monsoon climate interactions over Asia. Rev. Geophys. 54 (4), 866–929. https://doi.org/10.1002/2015RG000500.

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 J. Photogramm. Remote Sens. 150, 197–212. https://doi.org/10.1016/j.isprsjprs.2019.02.017.

Lin, J., Feng, X., Xiao, P., Li, H., Wang, J., Li, Y., 2012. Comparison of snow indexes in estimating snow cover fraction in a mountainous area in northwestern China. IEEE Geosci. Remote Sens. Lett. 9, 725–729. https://doi.org/10.1109/LGRS.2011.2179634.

Lyapustin, A., Wang, Y., Frey, R., 2008. An automatic cloud mask algorithm based on time series of MODIS measurements. J. Geophys. Res. Atmos. 113, D16207. https://doi.org/10.1029/2007JD009641.

Malekipirbazari, M., Aksakalli, V., 2015. Risk assessment in social lending via random forests. Expert Syst. Appl. 42 (10), 4621–4631. https://doi.org/10.1016/j.eswa.2015.02.001.

Mamyrova, G., O'Hanlon, T.P., Monroe, J.B., Carrick, D.M., Malley, J.D., Adams, S., Reed, A.M., et al., 2014. Immunogenetic risk and protective factors for juvenile dermatomyositis in Caucasians. Arthritis Rheumatol. 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. Int. J. Appl. Earth Obs. Geoinform. 34 (1), 136–146. https://doi.org/10.1016/j.jag.2014.08.001.

Oishi, Y., Ishida, H., Nakamura, R., 2018. A new Landsat 8 cloud discrimination algorithm using thresholding tests. Int. J. Remote Sens. 39 (23), 9113–9133. https://doi.org/10.1080/01431161.2018.1506183.

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. https://doi.org/10.1126/science.243.4887.57.

Rossow, W., Dueñas, E., 2004. The international satellite cloud climatology project (ISCCP) website. Bull. Am. Meteorol. Soc. 85 (2), 167–172.

Saunders, R., Kriebel, K., 1988. An improved method for detecting clear sky and cloudy radiances from AVHRR data. Int. J. Remote Sens. 9 (1), 123–150. https://doi.org/10.1080/01431168808954841.

Scaramuzza, P., Bouchard, M., Dwyer, J., 2012. Development of the Landsat data continuity mission cloud-cover assessment algorithms. IEEE Trans. Geosci. Remote Sens. 50 (4), 1140–1154. https://doi.org/10.1109/TGRS.2011.2164087.

Shen, Y., Wang, Y., Lv, H., Qian, J., 2015. Removal of thin clouds in Landsat-8 OLI data with independent component analysis. Remote Sens. 7, 11,481–11,500. https://doi.org/10.3390/rs70911481.

Stephens, G.L., 2005. Cloud feedbacks in the climate system: a critical review. J. Clim. 18 (2), 237–273. https://doi.org/10.1175/JCLI-3243.1.

Su, T., Li, Z., Kahn, R., 2018. Relationships between the planetary boundary layer height and surface pollutants derived from lidar observations over China: regional pattern and influencing factors. Atmos. Chem. Phys. 18 (21), 15921–15935. https://doi.org/

10.5194/acp-18-15921-2018.
Su, T., Li, Z., Kahn, R., 2020. A new method to retrieve the diurnal variability of planetary boundary layer height from lidar under different thermodynamic stability conditions. Remote Sens. Environ. 237, 111519. https://doi.org/10.1016/j.rse.2019.111519.
Sun, L., Wei, J., Wang, J., Mi, X., Guo, Y., Lv, Y., et al., 2016a. A universal dynamic threshold cloud detection algorithm (UDTCDA) supported by a prior surface reflectance database. J. Geophys. Res. Atmos. 121 (12), 7172–7196. https://doi.org/10.1002/2015JD024722.
Sun, L., Wei, J., Bilal, M., Tian, X., Jia, C., Guo, Y., Mi, X., 2016b. Aerosol optical depth retrieval over bright areas using Landsat 8 OLI images. Remote Sens. 8 (1). https://doi.org/10.3390/rs8010023.
Sun, L., Mi, X., Wei, J., Wang, J., Tian, X., Yu, H., Gan, P., 2017. A cloud detection algorithm-generating method for remote sensing data at visible to short-wave infrared wavelengths. ISPRS J. Photogramm. Remote Sens. 124, 70–88. https://doi.org/10.1016/j.isprsjprs.2016.12.005.
Sun, L., Zhou, X., Wei, J., Wang, Q., Liu, X., Shu, M., Chen, T., et al., 2018. A new cloud detection method supported by GlobeLand30 data set. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11 (10), 3628–3645. https://doi.org/10.1109/JSTARS.2018.2861755.
U.S. Geological Survey, 2016a. L7 Irish Cloud Validation Masks. U.S. Geological Survey data releasehttps://doi.org/10.5066/F7XD0ZWC.
U.S. Geological Survey, 2016b. L8 Biome Cloud Validation Masks. U.S. Geological Survey, data release.
Wang, Z., Lai, C., Chen, X., Bing, Y., Zhao, S., Bai, X., 2015. Flood hazard risk assessment model based on random forest. J. Hydrol. 527, 1130–1141. https://doi.org/10.1016/j.jhydrol.2015.06.008.
Wei, J., Huang, B., Sun, L., Zhang, Z., Wang, L., Bilal, M., 2017. A simple and universal aerosol retrieval algorithm for Landsat series images over complex surfaces. J. Geophys. Res. Atmos. 122, 13,338–13,355. https://doi.org/10.1002/2017JD026922.
Wei, J., Sun, L., Peng, Y., Wang, L., Zhang, Z., Bilal, M., Ma, Y., 2018. An improved high-spatial-resolution aerosol retrieval algorithm for MODIS images over land. J. Geophys. Res. Atmos. 123 (21), 12,291–12,307. https://doi.org/10.1029/2017JD027795.
Wei, J., Huang, W., Li, Z., Xue, W., Peng, Y., Sun, L., Cribb, M., 2019a. Estimating 1-km-resolution $PM_{2.5}$ concentrations across China using the space-time random forest approach. Remote Sens. Environ. 231, 111221. https://doi.org/10.1016/j.rse.2019.111221.
Wei, J., Li, Z., Guo, J., Sun, L., Huang, W., Xue, W., Fan, T., Cribb, M., 2019b. Satellite-derived 1-km-resolution PM_1 concentrations from 2014 to 2018 across China. Environ. Sci. Technol. 53 (22), 13265–13274. https://doi.org/10.1021/acs.est.9b03258.
Wei, J., Li, Z., Cribb, M., Huang, W., Xue, W., Sun, L., Guo, J., Peng, Y., Li, J., Lyapustin, A., Liu, L., Wu, H., Song, Y., 2020. Improved 1 km resolution $PM_{2.5}$ estimates across China using enhanced space-time extremely randomized trees. Atmos. Chem. Phys. 20 (6), 3273–3289. https://doi.org/10.5194/acp-20-3273-2020.
Wieland, M., Li, Y., Martinis, S., 2019. Multi-sensor cloud and cloud shadow segmentation with a convolutional neural network. Remote Sens. Environ. 230, 111203. https://doi.org/10.1016/j.rse.2019.05.022.
Wu, Z., Snyder, G., Vadnais, C., Arora, R., Babcock, M., Stensaas, G., et al., 2019. User needs for future Landsat missions. Remote Sens. Environ. 231. https://doi.org/10.1016/j.rse.2019.111214.
Wulder, M., Loveland, T., Roy, D., Crawford, C., Masek, J., Woodcock, C., et al., 2019. Current status of Landsat program, science, and applications. Remote Sens. Environ. 225, 127–147. https://doi.org/10.1016/j.rse.2019.02.015.
Yang, Y., Girolamo, L.D., Mazzoni, D., 2007. Selection of the automated thresholding algorithm for the multi-angle imaging spectroradiometer radiometric camera-by-camera cloud mask over land. Remote Sens. Environ. 107 (1–2), 159–171. https://doi.org/10.1016/j.rse.2006.05.020.
Zhai, H., Zhang, H., Zhang, L., Li, P., 2018. Cloud/shadow detection based on spectral indices for multi/hyperspectral optical remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 144, 235–253. https://doi.org/10.1016/j.isprsjprs.2018.07.006.
Zhang, Y., Rossow, W.B., Lacis, A.A., Oinas, V., Mishchenko, M.I., 2004. Calculation of radiative fluxes from the surface to top of atmosphere based on ISCCP and other global data sets: refinements of the radiative transfer model and the input data. J. Geophys. Res. Atmos. 109 (19), 19105. https://doi.org/10.1029/2003JD004457.
Zhu, Z., Woodcock, C., 2012. Object-based cloud and cloud shadow detection in Landsat imagery. Remote Sens. Environ. 118, 83–94. https://doi.org/10.1016/j.rse.2011.10.028.
Zhu, Z., Wang, S., Woodcock, C., 2015. Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and sentinel 2 images. Remote Sens. Environ. 159, 269–277. https://doi.org/10.1016/j.rse.2014.12.014.