International Journal of Remote Sensing and Earth Sciences Vol. 16 No.2 December 2019: 157-164

CLOUD IDENTIFICATION FROM MULTITEMPORAL LANDSAT-8 USING K-MEANS CLUSTERING

Wismu Sunarmodo[1,*], Anis Kamilah Hayati
[1]Remote Sensing Technology and Data Center, LAPAN
*e-mail: wismu.sunarmodo@lapan.go.id

Received: 31 December 2019; Revised: 18 February 2020; Approved: 18 February 2020

Abstract. In the processing and analysis of remote-sensing data, cloud that interferes with earth-surface data is still a challenge. Many methods have already been developed to identify cloud, and these can be classified into two categories: single-date and multi-date identification. Most of these methods also utilize the thresholding method which itself can be divided into two categories: local thresholding and global thresholding. Local thresholding works locally and is different for each pixel, while global thresholding works similarly for every pixel. To determine the global threshold, two approaches are commonly used: fixed value as threshold and adapted threshold. In this paper, we propose a cloud-identification method with an adapted threshold using K-means clustering. Each related multitemporal pixel is processed using K-means clustering to find the threshold. The threshold is then used to distinguish clouds from non-clouds. By using the L8 Biome cloud-cover assessment as a reference, the proposed method results in Kappa coefficient of above 0.9. Furthermore, the proposed method has lower levels of false negatives and omission errors than the FMask method.

Keywords: *cloud identification, Landsat-8, K-means clustering*

1 INTRODUCTION

Remote-sensing imagery is very useful for analysing and monitoring earth-surface phenomena. However, cloud often interferes with the processing and analysis of these images. Based on Landsat 8 metadata calculations acquired from September 2013 to August 2017, the average global cloud cover contained in Landsat 8 images is around 41.59%, with extremely high cloud cover observed in tropical rainforest regions (Zhu, Qui, He, & Deng, 2019).

Many methods have already been developed to identify cloud. In general, these methods fall into two major categories: single-date identification and multi-date identification. Most single-date identification algorithms utilize the physical characteristics of cloud such as brightness which can be identified from higher pixel value on visual bands. Another physical characteristic is cold temperatures which can be identified from thermal information such as Landsat 8 thermal band (Huang et al., 2010; Irish, Barker, Goward, & Arvidson, 2006; Lin, Tsai, Lai, & Chen, 2013; Zhu and Woodcock, 2012).

Li et al. (2017) developed a method which combines spectral, geometric and texture features to identify cloud in GaoFen-1 imagery. Sedano, Kempeneers, Strobl, and Kucera (2011) developed an algorithm to identify cloud in high-resolution data (SPOT4-HRVIR, SPOT5-HRG and IRS-LISS III) based on information obtained from data with a lower resolution (MODIS).

Multispectral data (e.g. Landsat) has an advantage in detecting cloud compared to data that only has visible bands, because bands such as NIR and SWIR which it provides can also be used

http://dx.doi.org/10.30536/j.ijreses.2019.v16.a3285

to detect cloud. Additionally, the thermal band is a primary band which can be used to detect cloud based on its temperature.

Multi-date identification methods utilize change detection between data. Most multi-date identification algorithms use reference data to identify cloud in other target data. Jin et al. (2013) use cloud-free data as reference data. Most of the multi-date methods utilize information such as sudden changes of reflectance on a pixel-by-pixel basis (Champion, 2012; Hagolle, Huc, Villa Pascual, & Dedieu 2010; Tang, Yu, Hagolle, & Jiang, 2013). Goodwin, Collett, Denham, and Flood (2013) use minimum and median values of the blue band as a reference.

There are studies which review and compare the cloud-detection algorithms. Foga et al. (2017) compared 13 cloud masking algorithms for Landsat-8 and found that FMask (Zhu and Woodcock, 2012) was the most accurate among the thermally based algorithms. Meanwhile, Zhu et al. (2019) conclude that most cloud-detection approaches for Landsat are based on single-date data and suggest that one of the disadvantages of using multi-date data is that those algorithms require large amounts of data and computation time. However, Goodwin et al. (2013) and Zhu and Woodcock (2014) show that approaches based on multi-date images could provide more accurate cloud identification.

In this paper, a multi-date approach with big-data tools is proposed. The cloud identification is performed using automatic thresholding as opposed to the static thresholding that is generally used. For this purpose, this study uses K-means clustering.

Twenty scenes of Landsat 8 path 113 row 063 data acquired during 2014 are used in this study. From the stacked data, K-means clustering generates classes for pixels at the same positions from 20 dates. Those thresholds are then applied to distinguish clouds from non-clouds.

Assessment is conducted visually and quantitatively by comparing the results from this study and FMask with a related scene from the L8 Biome cloud-cover assessment set (U.S. Geological Survey, 2016).

2 MATERIALS AND METHODOLOGY

2.1 Location and data

Data used in this experiment are 20 scenes of Landsat-8 from path 113 row 063 covering part of the South East Sulawesi area during 2014. The bands used in this paper are the visible bands (red, green and blue).

Table 2-1: List of acquisition dates and the cloud cover of the scenes used in this study.

Acquisition date	Cloud cover (%)
January 1, 2014	84.16
January 17, 2014	89.66
March 6, 2014	60.35
March 22, 2014	69.7
April 7, 2014	49.05
April 23, 2014	18.39
May 9, 2014	52.45
May 25, 2014	57.48
June 10, 2014	66.81
June 26, 2014	57.38
July 12, 2014	50.69
July 28, 2014	11.76
August 13, 2014	11.97
August 29, 2014	8.65
September 30, 2014	4.94
October 16, 2014	15.63
November 1, 2014	4.84
November 17, 2014	17.11
December 3, 2014	76.21
December 19, 2014	66.29

Visual bands are mostly available in optical remote-sensing data. In the future, a study which applies this methodology to other optical remote-

sensing data will be conducted. Although the proposed methodology only utilizes visual bands, the methodology technically could be applied to most optical remote-sensing data.

Table 2-1 shows the scenes that are used in this study and their cloud cover from metadata.

2.2 K-means clustering

K-means clustering is one of the popular cluster-analysis methods for unsupervised learning in data mining and machine learning. The aim of K-means clustering is to partition *n* observations into *k* clusters. Each observation will be included in the cluster with the nearest mean. This study uses K-means clustering developed in Scikit-learn (Pedregosa et al., 2011).

2.3 Cloud-identification methods

Figure 2-1 presents a flow chart of the method proposed in this paper. The scenes are stacked and then K-means clustering is applied to all sets of related pixels (pixels from the same position).

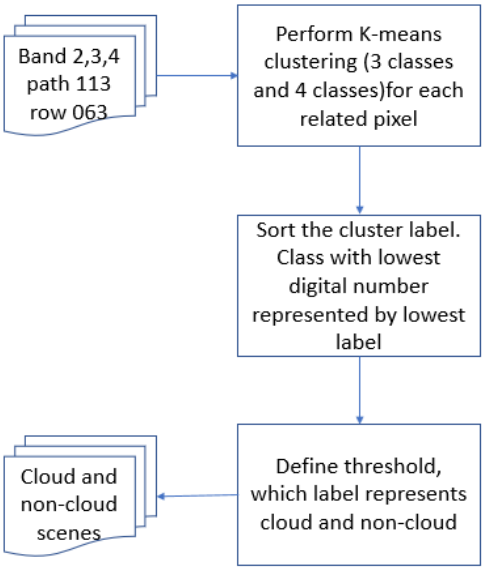

Figure 2-1: Flow chart of the cloud-identification method

Since the class labels from the clustering classes may not be in consecutive order, the labels need to be rearranged. Therefore, the smallest labels represent pixel groups that have the smallest digital numbers. Consequently, the higher the label, the higher the chance that the label represents a cloudy pixel.

In this study, three sets of data are generated. First, a set of data built using K-means clustering for three classes defines the first class label as 'non-cloud'. Second, K-means clustering of four classes defines the third and fourth class labels as 'cloud'. The last set of data, built of four classes from K-means clustering, defines the first class label as 'non-cloud'. Table 2-2. summarizes the three datasets generated in this study.

Table 2-2: Datasets generated in this study.

Dataset	K-means classes	Class labelled as non-cloud	Class labelled as cloud
1	3	1	2, 3
2	4	1, 2	3, 4
3	4	1	2, 3, 4

2.4 Assessment methods

Manual interpretation of cloud and cloud-shadow masks is an important data source for cloud-identification validation assessment (Foga et al., 2017; Irish et al., 2006; Zhu et al., 2019). In this study therefore, quantitative and qualitative assessments are conducted using cloud assessment data from L8 Biome (U.S. Geological Survey 2016) as a reference.

L8 Biome is a dataset of manually interpreted cloud and cloud-shadow masks which is publicly available. This dataset was developed by Foga et al. (2017) and is designed for Landsat-8 OLI/TIRS. The scenes used for L8 Biome were semi-randomly selected based on the biome in the scene itself, the path–row, and the approximate cloud cover. Furthermore, the digitization processes were performed by a single analyst to

reduce the probability of error due to different interpretations.

Currently, there are only two datasets of L8 Biome that cover Indonesia: path 113 row 063 (acquired on 29 August 2014) and path 104 row 062 (acquired on 7 March 2014). As the dataset from path 104 row 062 contains 95.95% cloud, the dataset from path 113 row 063 is used as a reference in this research.

The L8 Biome dataset includes identification of cloud, thin cloud and cloud shadow. However, since this study only covers cloud identification, the pixels that are used from the L8 Biome dataset are only the cloud and thin cloud pixels.

The FMask algorithm developed by Zhu and Woodcock (2012) is then applied to the same scene (path 113 row 063, acquired on 29 August 2014). The FMask algorithm is based on cloud and cloud-shadow physical properties such as brightness and low temperature utilizing bands 1 to 7 of Landsat-8 imagery. Figure 2-2 shows the steps used in the FMask algorithm.)

In this study, the FMask algorithm used is from the QGIS plugin named CloudMasking (Llano, 2019). The parameter for the cloud probability threshold in FMask is set to 22.5%, as this is the optimal global default threshold (Zhu et al., 2019). In addition, cloud buffer is set to 0 since the proposed method does not use a buffer.

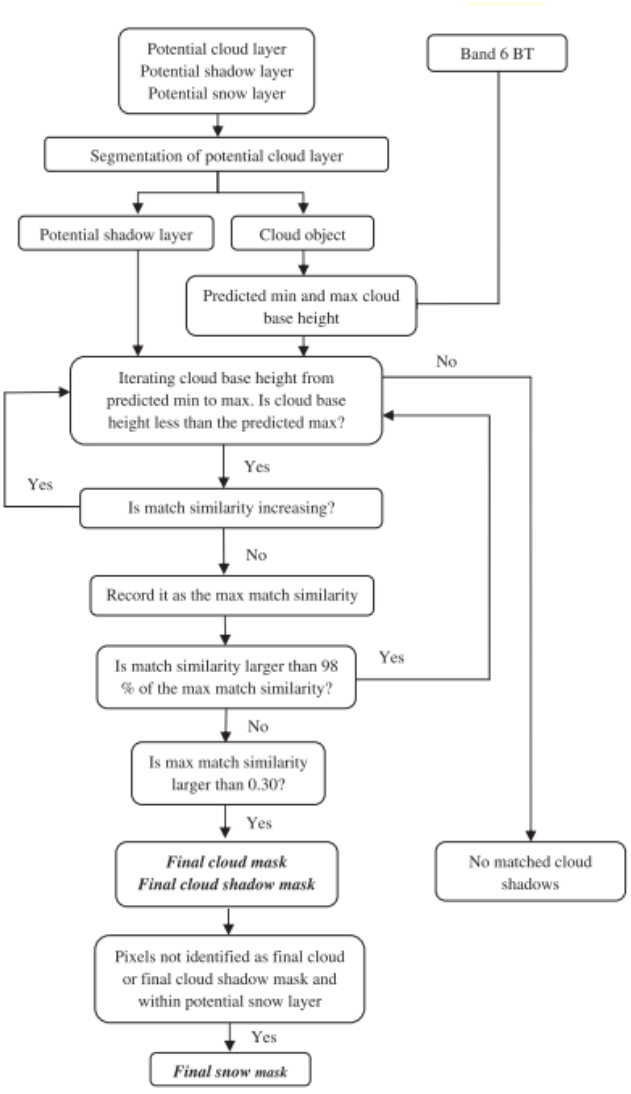

Figure 2-2: Flow chart of object-based cloud and cloud-shadow match algorithm (FMask). Source: Zhu and Woodcock (2012)

After every dataset is completed, Kappa coefficients for each dataset including FMask are calculated for use in qualitative assessment. Note that not all pixels in the scene are included in the calculation. Pixels included for consideration are only those that exist in all scenes listed in Table 2-1. This is because null data from one or more pixels could result in error results when performing K-means clustering.

3 RESULTS AND DISCUSSION

As can be seen in Figure 3-1, generally, classes generated using K-means clustering can be used to identify cloud. Visually, Dataset 1 (Figure 3-1 (b) and (f)) and Dataset 3 (Figure 3-1 (d) and (h)) provide results that are better than Dataset 2 (Figure 3-1(c) and (g)). Large, thin cloud areas are not identified as cloud in Dataset 2 (yellow circles), as seen in Figure 3-1(c). In addition, Dataset 3 outperforms Dataset 1 and Dataset 2 in identifying thin cloud, which reduces the omission error. However, as seen in Figure 3-1(h), thresholding in Dataset 3 includes some commission errors (red circles), such as land-cover change.

A comparison between the results of this study from Dataset 3, FMask and L8 Biome can be seen in Figure 3-2. Visually, the results from Dataset 3 (Figure 3-2(b) and (f)) are similar to the results from FMask (Figure 3-2(c) and (g)) and L8 Biome (Figure 3-2(d) and (h)).

However, some omission errors exist in the results from each. For example, in Figure 3-2 (b), a red circle shows thin cloud that is undetected in Dataset 3 but which is detected in FMask and L8 Biome. On the other hand, Figures 3-2(c) and (g) show omission errors (yellow circles) that are overcome by Dataset 3. Furthermore, although L8 Biome is widely used as a reference for cloud-identification algorithm validation, it still contains some omission errors (white circles).

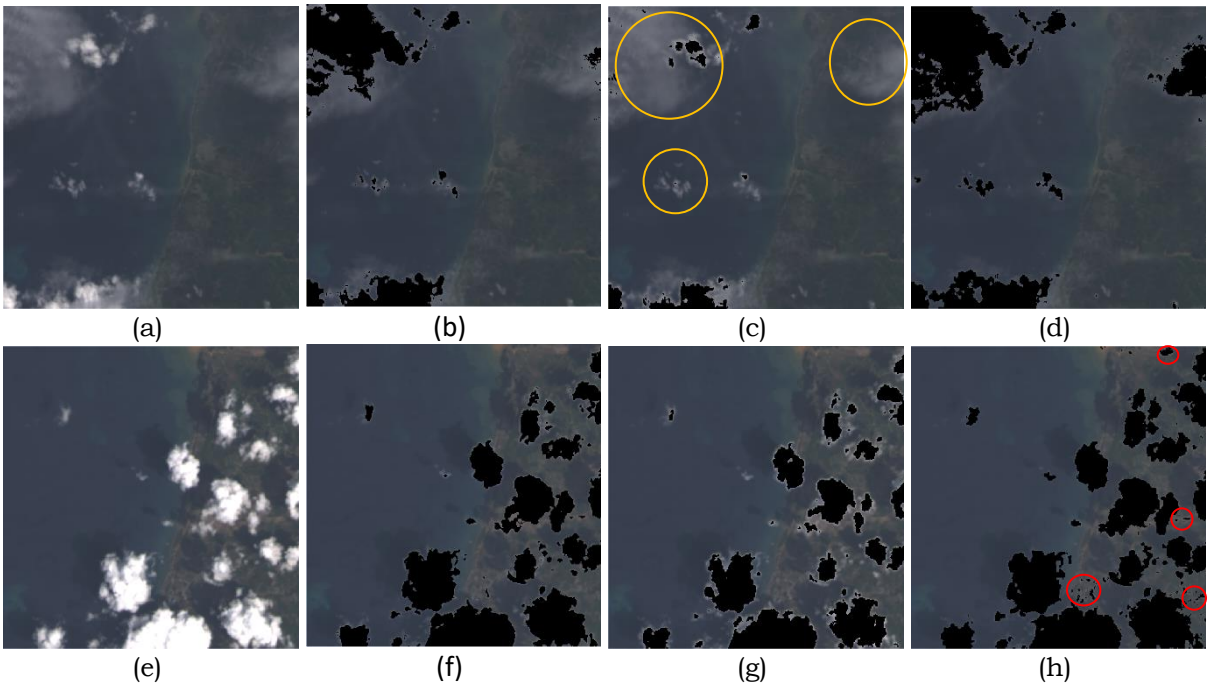

Figure 3-1: Comparisons of cloud identification. (a) The RGB composite tiles (acquisition date 10 June 2014). (e) The RGB composite tiles (acquisition date 17 November 2014). (b), (f) Cloud identification from Dataset 1. (c), (g) Cloud identification from Dataset 2. (d), (h) Cloud identification from Dataset 3

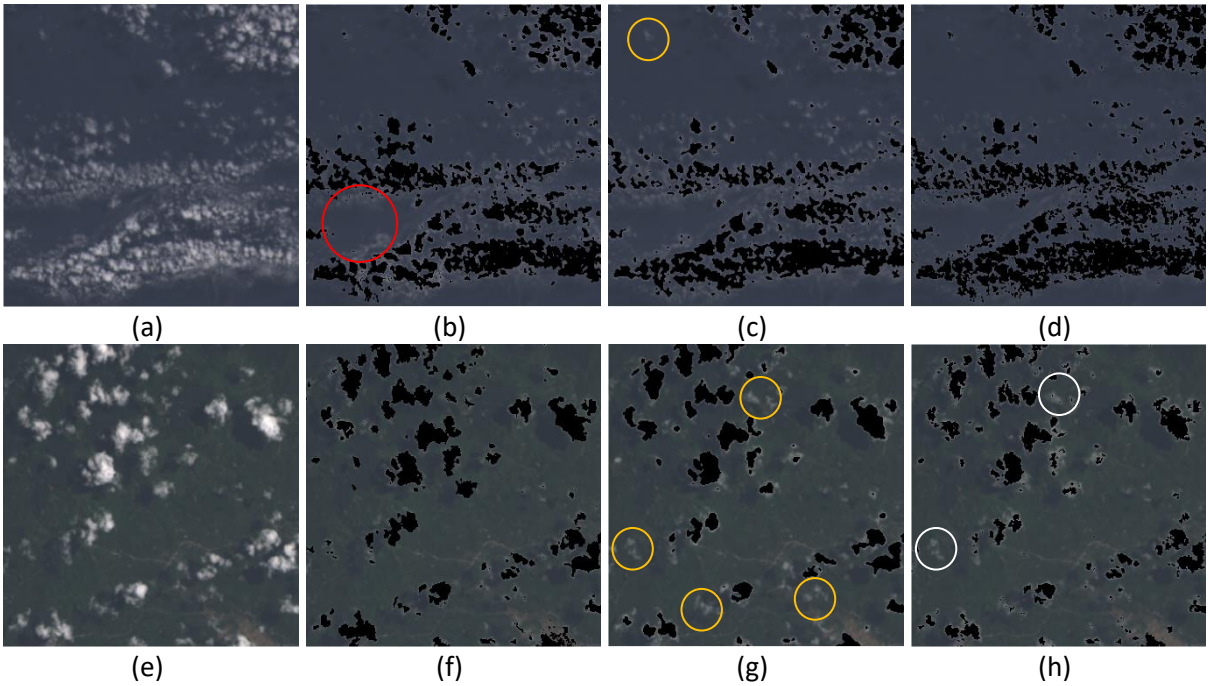

Figure 3-2: Comparison between this study's result, FMask and L8 Biome. (a), (e) RGB composite tiles (acquisition date 29 August 2014). (b), (f) Cloud identification from Dataset 3. (c), (g) Cloud identification using FMask algorithm. (d), (h) Cloud identification by L8 Biome.

The Kappa coefficients for Dataset 1 and Dataset 3 are slightly higher than for FMask, while for Dataset 2 is far lower (Table 3-1). In terms of error, while Dataset 1 and Dataset 3 have comparable Kappa coefficients, Dataset 1 has bigger false negative or omission error while Dataset 3 has bigger false positive or commission error. On the other hand, the result of this study shows that FMask has a bigger false negative and less false positives than Dataset 3.

Note that the input set has a big effect on the performance of the proposed method. If the input contains a small number of scenes, the chance that all related pixels are in the same group (clouds or non-clouds) is greater.

For example, if the scenes that are used are only from 1 January 2014; 17 January 2014; 22 March 2014; 10 June 2014; 3 December 2014; and 19 December 2014, which contain cloud cover greater than 65%, than most probably the proposed method will perform poorly. In this case, K-means clustering would not work properly in generating classes. Thus, the more scenes that are included in the processing, the better the K-means clustering is in generating classes. In this study, 20 scenes with variation of cloud-cover percentage proved to be sufficiently effective, resulting in Kappa coefficient of higher than 90%.

Table 3-1: Qualitative assessment

	Dataset 1	Dataset 2	Dataset 3	FMask
TN	30743113	30833686	30441191	30444433
FP	93213	2640	395135	391893
FN	332935	1177784	70628	111876
TP	2423059	1578210	2685366	2644118
Accuracy	0.987314	0.96486	0.986135	0.985003
Kappa coef.	0.912304	0.710499	0.912632	0.904839

The combination of the number of classes and choosing the right class as a threshold is also important. Dataset 2 and Dataset 3 have four classes, and choosing which classes are cloud and non-cloud affects the results from these classes. Dataset 2 has a higher false-negative level compared to Dataset 3. On the other hand, Dataset 3 has a higher false-positive level than Dataset 2. This is due to the second class being regarded as non-cloud in Dataset 2 but identified as cloud in Dataset 3. Generating more classes may result in better cloud identification. However, the trade-off is the processing performance, especially the K-means clustering, which will require more time to generate the classes.

4 CONCLUSIONS

A method using temporal data and K-means clustering to identify cloud was developed for Landsat-8 data. This study shows that generally K-means clustering could be used to identify cloud in multitemporal Landsat-8 data.

Qualitatively and quantitatively, Dataset 3 performs better than Dataset 1 and Dataset 2. Dataset 3, which is a dataset generated from K-means clustering with four classes and using the first class as a non-cloud threshold, has the highest Kappa coefficient among the datasets. However, Dataset 3 has a bigger level of commission error since it includes non-cloud pixels such as land-cover change.

Choosing the number of classes and which class will be used as a threshold are essential steps for this method. Different numbers of classes and different thresholds produce different cloud-identification results.

Using L8 Biome cloud assessment as a reference, the proposed method performs well, with Kappa coefficient of higher than 90% (Dataset 1 and Dataset 3). However, if the user wishes to minimize omission error, they are encouraged to choose Dataset 3 rather than Dataset 1.

Furthermore, qualitatively and quantitatively, the proposed method performs comparatively well with the FMask method. Dataset 3 has a qualitative result that is near to that of the FMask method.

ACKNOWLEDGEMENTS

This research was funded and facilitated by the Remote-Sensing Technology and Data Centre of LAPAN. We would like to thank everyone who has been involved in the preparation of this paper, particularly the acquisition and management team for providing access to the Landsat 8 data.

AUTHOR CONTRIBUTIONS

Cloud Identification from Multitemporal Landsat-8 using K-Means Clustering. Lead Author: Wismu Sunarmodo and Anis Kamilah Hayati.

REFERENCES

Champion, N. (2012). Automatic cloud detection from multi-temporal satellite images: Towards the use of Pléiades time series. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XXXIX-B3* (September), 559–564. doi:0.5194/isprsarchives-xxxix-b3-559-2012

Foga S. C., Scaramuzza, P., Guo, S., Zhu, Z., Diley, R., Beckman, T. … Laue, B. (2017). Cloud detection algorithm comparison and validation for operational Landsat data products. *Remote Sensing of Environment*, *194*, 379–390. doi:10.1016/j.rse.2017.03.026

Goodwin, N. R., Collett, L., Denham, R. J., & Flood, N. (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.

doi:10.1016/j.rse.2013.02.019

Hagolle, O., Huc, M., Villa Pascual, D., & Dedieu, *G.* (2010). A multi-temporal method for cloud detection, applied to FORMOSAT-2, VENμS, LANDSAT and SENTINEL-2 images. *Remote Sensing of Environment*, *114*(8), 1747–1755. doi:10.1016/j.rse.2010.03.002

Huang, C., Thomas, N., Goward, S. N., Masek, J. G., Zhu, Z., Townsend, J. R. G., & Vogelmann, J. E. (2010). Automated masking of cloud and cloud shadow for forest change analysis using Landsat images. *International Journal of Remote Sensing*, *31*(20), 5449–5464. doi:10.1080/01431160903369642

Irish, R. R., Barker, J. L., Goward, S. N., & Arvidson, T. (2006). Characterization of the Landsat-7 ETM+ automated cloud-cover assessment (ACCA) algorithm. *Photogrammetric Engineering and Remote Sensing*, *72*(10), 1179–1188. doi:10.14358/PERS.72.10.1179

Jin, S., Homer, C. G., 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. *International Journal of Remote Sensing*, *34*(5), 1540–1560. doi:10.1080/01431161.2012.720045

Li, Z., Shen, H., Li, H., Xia, G., Gamba, P., & Zhang, L. (2017). Multi-feature combined cloud and cloud shadow detection in GaoFen-1 wide field of view imagery. *Remote Sensing of Environment*, *191*(April 2013), 342–358. doi:10.1016/j.rse.2017.01.026

Lin, C., Tsai, P.-H., Lai, K.-H., & Chen, J.-Y. (2013). Cloud removal from multitemporal satellite images using information cloning. *IEEE Transactions on Geoscience and Remote Sensing*, *51*(1), 232–241.

Llano, X. C. (2019). CloudMasking. *GitHub repository*. Retreived from https://github.com/SMByC/CloudMasking

Pedregosa, F., Varoquaux, G. Gramfort, A., Michel, V., Thirion, B., Grisel, O., …Duchesnay, E. (2011), Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Sedano, F., Kempeneers, P., Strobl, P., & Kucera, J. (2011). A cloud mask methodology for high resolution remote sensing data combining information from high and medium resolution optical sensors. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(5), 588–596. doi:10.1016/j.isprsjprs.2011.03.005

Tang, H., Yu, K., Hagolle, O., & Jiang, K. (2013). A cloud detection method based on a time series of MODIS surface reflectance images. *International Journal of Digital Earth*, *6*(1), 157–171. doi:10.1080/17538947.2013.833313

U.S. Geological Survey (2016). L8 Biome cloud validation masks. *U.S. Geological Survey, data release.* doi:10.5066/F7251GDH

Zhu, Z., Qui, S., He, B., & Deng, C. (2019). Cloud and cloud shadow detection for Landsat images: The fundamental basis for analyzing Landsat time series. *Remote Sensing Time Series Image Processing*, (May), 3–23. https://doi.org/10.1201/9781315166636-1

Zhu, Z. & Woodcock, C. E. (2012). Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sensing of Environment*, *118*, 83–94. doi:10.1016/j.rse.2011.10.028

Zhu, Z. & Woodcock, C. E. (2014). Automated cloud, cloud shadow, and snow detection in multitemporal Landsat data: An algorithm designed specifically for monitoring land cover change. *Remote Sensing of Environment*, *152*, 217–234. doi:10.1016/j.rse.2014.06.01