A novel algorithm of cloud detection for water quality studies using 250 m downscaled MODIS imagery

Claudie Ratté-Fortin*, Karem Chokmani, Anas El Alem

Institut National de la Recherche Scientifique, Centre Eau Terre Environnement, 490 De la Couronne Street, G1K 9A9, Quebec city, Quebec, Canada.

*Corresponding author: claudie.ratte-fortin@ete.inrs.ca

Please cite this article as: Ratté-Fortin, C., Chokmani, K., and El-Alem, A. (2018). A novel algorithm of cloud detection for water quality studies using 250 m downscaled MODIS imagery. International Journal of Remote Sensing, 1-12.

**Keywords**: Cloud, mask, MODIS, chlorophyll-*a*, total suspended solids, dissolved organic matter, inland waters, lakes, algal blooms, cyanobacteria.

## Abstract

This study is part of a project aimed at developing an automated algorithm for algal bloom detection and quantification in inland water bodies using Moderate resolution imaging spectroradiometer (MODIS) imagery. An important step is to adequately detect and exclude clouds and haze because their presence affects chlorophyll-*a* (chl-*a*) estimations. Currently available cloud masking products appear to be ineffective in turbid coastal waters. The purpose of this study is to develop a cloud masking algorithm based on a probabilistic algorithm (Linear Discriminant Analysis) and designed for water bodies by using MODIS images downscaled at a 250 m spatial resolution (MODIS-D-250). Confusion matrix shows that the new cloud mask algorithm yields very satisfactory results, enabling water classification for heavy turbid conditions with a mean kappa coefficient ($\kappa$) (of 0.982 and a 95% confidence interval ranging from 0.979 to 0.986. The model also shows a very low commission error (sensitive to the presence of haze) which is essential for accurate water quality monitoring, knowing that the presence of clouds/haze/aerosols leads to major issues in the estimation of water quality parameters. The cloud mask model applied on MODIS-D-250 images improves the sensitivity to haze and the classification of turbid waters located at the edge of urban areas better than the operational MODIS products, and it clearly shows an improvement of the spatial resolution (250 m spatial resolution) compared to other cloud mask algorithms (500 m or 1 km spatial resolution) leading to an increase in exploitable data for water quality studies.

## 1. Introduction

Water colour satellite data are increasingly used to manage and monitor water quality for ocean and coastal waters. In water colour data processing, good cloud masking is an essential step in obtaining an accurate water colour signal. For that purpose, different cloud mask algorithms have been developed but all have certain issues, specifically in the processing of water colour data. In fact, a lot of these algorithms were developed specifically for turbid water colour data, which leads to classification errors or to the loss of valuable data (Chen & Zhang, 2015). Recently, efforts have been deployed to develop explicit algorithms for cloud masking over turbid water colour data, but most were applied on ocean and coastal waters (Wang & Shi, 2006; Banks & Mélin, 2015; Chen *et al.*, 2015). No cloud masking algorithm has been specifically designed for inland waters (lakes, rivers, and estuaries), where water contains a lot more optically active components such as chlorophyll-*a* (chl-*a*), total suspended solids (TSS), and coloured dissolved organic matter (CDOM).

In ocean water studies, cloud detection techniques are generally based on the hypothesis that the reflectance signal of water at near infrared (NIR) is almost null (Nagamani *et al.*, 2015). This approach becomes, however, less effective with the presence of optically active components in water, such as a high phytoplankton biomass, known to generate turbid waters, which significantly increase reflectance at red and NIR channels (Kahru *et al.*, 2004). Turbid waters can be mistaken as cloud pixels, even under clear skies. Moderate resolution imaging spectroradiometer (MODIS) Atmosphere Group developed the standard MODIS cloud product generated at a 1 km and 250 m spatial resolution. This product also uses a NIR threshold which is its principal weakness when applying the algorithm on turbid waters (Robinson *et al.*, 2003). Another 1 km-spatial resolution algorithm developed by Nordkvist *et al.* (2009) and based on spectral variability of visible and NIR often incorrectly mask intense phytoplankton blooms (Banks *et al.*, 2015). Considering the high spatial variability of clouds, there are also algorithms based on a spatial variability threshold of the MODIS green band (Martins *et al.*, 2002) and the MODIS NIR band (Nicolas *et al.*, 2005). Once again, the use of visible and NIR bands will identify turbid waters as clouds, due to their high spatial variability at these wavelengths (Lubac & Loisel, 2007). To avoid this problem, certain cloud detection algorithms use the MODIS shortwave infrared (SWIR) threshold such as that of Wang *et al.* (2006) and Chen *et al.* (2015) who proposed a spatial variability threshold at SWIR band. These cloud masks are generated at a spatial resolution of 1 km and 500 m respectively. These methods based on SWIR band threshold appear to show the best overall performance; however, they lack adequate spatial resolution for water studies in small to medium-sized lakes.

This study is part of a project aimed at monitoring and assessing past, present and future water quality in inland waters by using MODIS imagery downscaled to 250 m spatial resolution (MODIS-D-250). In fact, the Canadian Center for Remote Sensing has developed an approach allowing to downscale the spatial resolution of MODIS bands 3-7 from 500 m to 250 m (Trishchenko *et al.*, 2006). Annexe products are also generated with the downscaled images including a cloud mask at a spatial resolution of 250 m. However, this model generally doesn't perform well when detecting clouds and cloud shadows over water bodies (see figure 1, centre). Furthermore, the actual cloud masking product available for MODIS images is recorded at 250 m and 1 km spatial resolution (Ackerman *et al.*, 2010). The one generated at 1 km-spatial resolution is unsuitable for water quality monitoring in small to medium-sized inland waters and in addition, it appears to be ineffective in turbid coastal waters (see figure 1, right). The 250 m spatial resolution MODIS cloud mask (Platnick *et al.*, 2017) incorporates the results from the 1 km resolution tests to maintain consistency with the 1-km cloud mask, and so, it appears to show the same issues than the 1 km cloud mask in detecting thin clouds/haze and distinguishing turbid waters. The Linear Discriminant Analysis (LDA) appears to be an interesting alternative. This method, which is designed to highlight inland water bodies in remotely sensed imagery, has often been used for land cover classification (Friedl & Brodley, 1997; Xia *et al.*, 2014; Priedītis *et al.*, 2015) and for water index (Adrian Fisher & Danaher, 2013). Indeed, multivariate techniques provide much richer and more global information to the predictive model. The use of LDA is also preferred to threshold algorithms when finding an optimal discriminant model.

The objective of this paper is to develop a cloud mask for water bodies (inland, coastal, and open ocean) based on a LDA algorithm using MODIS-D-250 data. The present paper focuses on the application of a probabilistic method using 1-7 MODIS-D-250 bands to predict pixel classes, instead of actual parametric methods, as proposed in the literature (threshold algorithms).

## 2. Material and Methods

### 2.1. Data collection and pre-processing

Satellite data that cover the southern part of the province of Quebec, Canada (44º-50º N, 67º-80º W) were acquired from MODIS sensor aboard the Terra platform of NASA's

Earth Observation System (see figure 2). Characteristics of the MODIS bands used in this study are presented in table 1. The spatial resolution of bands 3-7 was downscaled from 500 m to 250 m by using an adaptive regression and radiometric normalization as described in Trishchenko *et al.* (2006). The approach used to downscale MODIS bands 3 to 7 from 500-m to 250 m spatial resolution (Trishchenko *et al.*, 2006) was validated using data at higher spatial resolution (Landsat ETM+ (30 m)). Results showed that the downscaling procedure does not alter the radiometric properties of a scene, and so, the higher resolution bands can be used to generate a reliable cloud mask at 250 m spatial resolution. Besides, the MODIS bands originally at 250 m spatial resolution (bands 1-2) and those downscaled (bands 3 to 7) are originally designed for aerosol, cloud and land applications. Images were then re-projected from the Sinusoidal to the Lambert Conformal Conic projection, and were corrected for atmospheric effects using the Simplified Model for Atmospheric Correction (SMAC). Image pre-processing, including downscaling, re-projection, and atmospheric correction was performed using an automatic tool developed by the Canadian Center for Remote Sensing (Trishchenko *et al.*, 2007). Finally, in order to better distinguish water pixels from mixed pixels (land-water), a land mask developed by El Alem (2014) was applied to the MODIS database.

**2.2. Model description**

This section briefly describes the linear discriminant analysis modelling framework, which was computed using Matlab software (R2016a). This method was proposed by Ronald Fisher (1936) and consists of finding a projection that minimizes the variance between classes while maximizing the distances between the projected means of the classes. A general description of LDA can be found in Xanthopoulos *et al.* (2013). We assume that we have a categorical dependent variable corresponding to the following classes water, haze (*a priori*), and cloud, and independent variables corresponding to the reflectance values of the 1-7 MODIS-D-250 bands. Independent variables are transformed for normality. LDA allows to determine a subspace of dimension inferior to that of the original data in which data are separable in terms of statistical measures of mean and variance values. First, the model discriminates the three classes (water, haze (*a priori*), and cloud), assuming that independent variables have a multivariate normal distribution and the same covariance matrix for each class (figure 3). Clear water is easy to distinguish from cloud and fog due to the low reflectance in visible and near-infrared. At the opposite, water containing optically active components such as TSS, CDOM and chl-*a* is more difficult to distinguish from cloud/fog pixels in this spectral region. For that reason, a second LDA is performed only on the pixels classified as fog to try to discriminate real fog from waters with moderate to high chl-*a* concentrations or turbid waters. The resulting data are further separated into three other classes: water (high turbidity), water (algal bloom), and haze. A chl-*a* concentration estimator designed to perform in optically complex inland waters (El-Alem *et al.,* 2014) was used to manually classify those three categories: fog, water (bloom), and water (turbidity). To classify these categories, the chl-*a* concentration estimator was applied to images taken during important algal blooms and on lakes known to have high turbidity.

## 2.3. Calibration and validation

A set of samples from twenty-six MODIS images were selected from the ice-free season (May to November) of the years 2000 to 2015, and used for model calibration and validation (table 2). We selected several free water samples (lakes, rivers, gulf, bay and estuaries) from each MODIS scene that are representative of trophic classes of waterbodies (oligotrophic, mesotrophic, eutrophic and hypereutrophic classes). Helped by visual inspection of the maps and the highly turbid lakes known in the literature, a chl-*a* concentration estimator designed to perform in optically complex inland waters (El-Alem *et al.*, 2014) was also used to distinguish clear water, algal blooms and turbid waters. The samples cover all the range of trophic classes based on very low chl-*a* concentrations (0,1 $\mu g\ l^{-1}$) to very high chl-*a* concentrations (more than 1000 $\mu g\ l^{-1}$).

The dataset was then partitioned into two sets: we saved some images for calibration, containing 70% (6186 pixels) of the data, and used the other for validation with 30% (2651 pixels) of the data. The performance of the statistical model is evaluated using a Monte-Carlo cross-validation: the random split of the original sample into calibration and validation data is repeated 10,000 times in order to obtain a distribution of the global success and the kappa coefficient ($\kappa$) values of the classification (see figure 4). To evaluate the performance of the cloud mask algorithm, the model was applied to several MODIS images (qualitative validation). These images were not used in the model calibration/validation steps. Scenes that include lakes and estuaries known to be highly turbid and lakes during a period when an algal bloom was occurring were selected. The algorithm estimating chl-*a* concentration in inland waters (El-Alem *et al.*, 2014) was also applied to the validation images, allowing us to detect algal blooms.

## 3. Results and Discussion

Table 3 presents the confusion matrix of the double discriminant analysis model over the three classes. Results show that the classification of cloud and water pixels is not problematic. The model adequately classifies water pixels with 0% false negative. The model underestimates cloud detection in 1% of cases (false negatives) but those pixels are classified as haze, which is not problematic for water colour data studies. Consequently, none of the water pixels are misclassified as cloud or haze, which is the major classification problem of actual cloud mask algorithms in presence of optically active components (chl-*a*, TSS or CDOM) in water (Banks *et al.*, 2015). Overall, the model's performance is very good with a $\kappa$ of 0.982 and a 95% confidence interval ranging from 0.979 to 0.986. Global success of the classification is 98.9% ranging from 99.0% to 99.2% (95% confidence interval). In order to compare our cloud mask algorithm with the 250 m and 1 km MODIS cloud masks, we also have generated the global success and $\kappa$ over two classes (cloud, no cloud) into one combined cloud class. Table 4 presents the results obtained with the three cloud masks applied on the same validation data set.

As a qualitative validation, the new cloud mask algorithm was applied to MODIS-D-250 images and compared to the current MODIS 1 km and 250 m cloud masks. Figures 5 and 6 present results for the Missisquoi Bay of Champlain Lake (during a period with moderate to high chl-*a* concentration), St-Lawrence river (moderate turbidity and moderate chl-*a* concentration) and Macamic Lake (high turbidity). MODIS cloud masks don't appear to be sensitive enough to haze, which leads to major issues in remote chl-*a* estimates. Figure 5 shows an example of that issue and the improvement of haze detection of our new algorithm. It presents the Missisquoi Bay during an algal bloom (at the top) and the Champlain Lake covered in part with cloud and haze (at the bottom). The three cloud masks are then presented (1km MODIS cloud mask, 250 m MODIS cloud mask, and the new 250 m cloud mask), and below, the chl-*a* concentration estimated with the remaining water pixels. The chl-*a* values were generated using an algorithm developed by El-Alem *et al.* (2014). Both MODIS cloud masks are not enough sensitive enough to haze, which yields some high estimates of chl-*a* concentration for pixels without a priori algal bloom.

MODIS cloud masks are also not suited to perform well in turbid waters. It happens that the masks falsely detect clouds in turbid waters. The St-Lawrence MODIS scene in figure 6 shows that the cloud/haze classification is highly improved with the new 250 m cloud mask compared to both MODIS cloud masks. Highly turbid waters located at the edge of an urban area, which are often problematic to cloud masking algorithms, are now much better classified as water pixels. It should be noted that the land mask which was developed and applied to the images covers transition zones from land to water (mixed pixels) up to 250 m of the edge of lakes. Also, on small to medium-sized lakes and particularly those with turbid waters, the false classification of MODIS cloud masks becomes a major issue in terms of exploitable data. Figure 6 (bottom) shows another MODIS scene on a smaller area, the Macamic Lake which has a surface area of 45 $km^2$. MODIS cloud masks falsely classify as cloud approximately 16 % of the lake area.

Figure 7 presents the cloud masks performance in thin haze and in cirrus conditions. The image of the Bay of Fundy from 24 August 2014 shows the very good performance of the algorithm in haze detection especially when compared to the MODIS cloud masks. The second scene taken on St-Lawrence river clearly shows a lack of performance in detecting cirrus clouds by the MODIS products. As we showed earlier, the lack of sensitivity to haze and thin clouds can lead to misinterpretation of the water quality parameters.

## 4. Conclusion

In conclusion, a cloud masking algorithm based on a double discriminant analysis at a resolution of 250 m for MODIS imagery was presented. Overall, the new cloud mask shows a better performance than the MODIS cloud mask when it is applied on turbid waters, and particularly on highly turbid waters located at the edge of an urban area. The new cloud mask presents an improved resolution of 250 m, leading to an increase of exploitable data in the context of water colour studies, and particularly for water quality monitoring in small to medium-sized inland waters. The new algorithm reduces potential commission errors more efficiently than the MODIS cloud mask, which is less sensitive to haze. The commission error reduction is essential for accurate algal blooms monitoring, because the presence of clouds and haze affects chl-*a* concentration estimations. Finally, the innovative aspect of this algorithm is the use of a probabilistic method to generate a cloud mask compared to current methods proposed in the literature based on threshold algorithms, leading to an optimal and accurate predictive model. Confusion matrix results highlight the very good concordance between observed and predicted classes using the algorithm on the downscaled MODIS bands, showing a global success average of 99.6% with a 95% confidence interval ranging from 99.4% to 99.8%, and a $\kappa$ average of 0.993 with a 95% confidence interval ranging from 0.990 to 0.997.

## References


Ackerman S, Strabala K, Menzel P, Frey R, Moeller C & Gumley L (2010) Discriminating clear-sky from cloud with MODIS algorithm theoretical basis document (MOD35. *MODIS Cloud Mask Team, Cooperative Institute for Meteorological Satellite Studies, University of Wisconsin.* Citeseer.

Banks AC & Mélin F (2015) An assessment of cloud masking schemes for satellite ocean colour data of marine optical extremes. *International Journal of Remote Sensing* 36(3):797-821.

Chen S & Zhang T (2015) An improved cloud masking algorithm for MODIS ocean colour data processing. *Remote Sensing Letters* 6(3):218-227.

El-Alem A, Chokmani K, Laurion I & El-Adlouni SE (2014) An adaptive model to monitor chlorophyll-a in inland waters in southern Quebec using downscaled MODIS imagery. *Remote Sensing* 6(7):6446-6471.

El Alem A (2014) *Développement d'une approche de suivi des fleurs d'eau d'algues à l'aide de l'imagerie désagrégée du capteur MODIS, adaptée aux lacs du Québec méridional*. (Université du Québec, Institut national de la recherche scientifique).

Fisher A & Danaher T (2013) A water index for SPOT5 HRG satellite imagery, New South Wales, Australia, determined by linear discriminant analysis. *Remote Sensing* 5(11):5907-5925.

Fisher R (1936) The use of multiple measurements in taxonomic problems. *Annals of eugenics* 7(2):179-188.

Friedl MA & Brodley CE (1997) Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment* 61(3):399-409.

Kahru M, Michell BG, Diaz A & Miura M (2004) MODIS detects a devastating algal bloom in Paracas Bay, Peru. *Eos, Transactions American Geophysical Union* 85(45):465-472.

Lubac B & Loisel H (2007) Variability and classification of remote sensing reflectance spectra in the eastern English Channel and southern North Sea. *Remote Sensing of Environment* 110(1):45-58.

Martins JV, Tanré D, Remer L, Kaufman Y, Mattoo S & Levy R (2002) MODIS cloud screening for remote sensing of aerosols over oceans using spatial variability. *Geophysical Research Letters* 29(12).

Nagamani P, Latha TP, Rao K, Suresh T, Choudhury S, Dutt C & Dadhwal V (2015) Setting of cloud albedo in the atmospheric correction procedure to generate the ocean colour data products from OCM-2. *Journal of the Indian Society of Remote Sensing* 43(2):439-444.

Nicolas J, Deschamps P, Loisel H & Moulin C (2005) Algorithm Theoretical Basis Document, POLDER-2/Ocean Color/Atmospheric corrections.).

Nordkvist K, Loisel H & Gaurier LD (2009) Cloud masking of SeaWiFS images over coastal waters using spectral variability. *Opt. Express* 17(15):12246-12258.

Platnick S, Meyer KG, King MD, Wind G, Amarasinghe N, Marchant B, Arnold GT, Zhang Z, Hubanks PA & Holz RE (2017) The MODIS cloud optical and microphysical products: Collection 6 updates and examples from Terra and Aqua. *IEEE Transactions on Geoscience and Remote Sensing* 55(1):502-525.

Priedītis G, Smits I, Dagis S, Paura L, Krumins J & Dubrovskis D (2015) Assessment of hyperspectral data analysis methods to classify tree species. *Research for Rural Development. International Scientific Conference Proceedings (Latvia).* Latvia University of Agriculture.

Robinson W, Franz B, Patt F, Bailey S & Werdell P (2003) Masks and flags updates. *Algorithm updates for the fourth Sea-WiFS data reprocessing, NASA Technical Memorandum* 206892:34-40.

Trishchenko A, Luo Y & Khlopenkov K (2006) A method for downscaling MODIS land channels to 250 m spatial resolution using adaptive regression and normalization. *Remote Sensing for Environmental Monitoring* 6366:36607-36607.

Trishchenko A, Luo Y, Khlopenkov K & Park W (2007) Multi-- spectral clear-- sky composites of MODIS/Terra Land Channels(B1-- B7) over Canada at 250m spatial resolution and 10-- day intervals since March, 2000: Top of the Atmosphere (TOA) data. *Enhancing Resilience in a Changing Climate. Earth Sciences Sector Canada Centre for Remote Sensing (CCRS). Natural Resources Canada*.

Wang M & Shi W (2006) Cloud masking for ocean color data processing in the coastal regions. *IEEE Transactions on Geoscience and Remote Sensing* 44(11):3196-3105.

Xanthopoulos P, Pardalos PM & Trafalis TB (2013) Linear discriminant analysis. *Robust Data Mining*, Springer. p 27-33.

Xia J, Du P, He X & Chanussot J (2014) Hyperspectral remote sensing image classification based on rotation forest. *IEEE Geoscience and Remote Sensing Letters* 11(1):239-243.

## Tables and Figures

Figure 1: (*a*) MODIS true color image, (*b*) corresponding cloud mask developed by the Canadian Center for Remote Sensing and (*c*) cloud mask developed by MODIS Atmosphere Group.

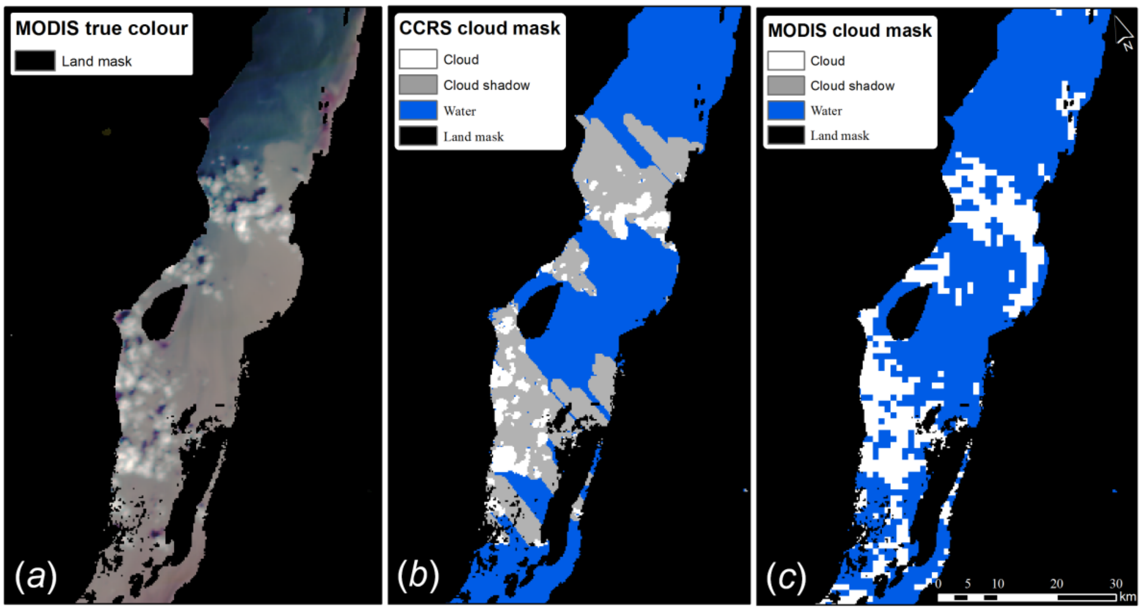

Figure 2: Geographic location of MODIS imagery historical database.

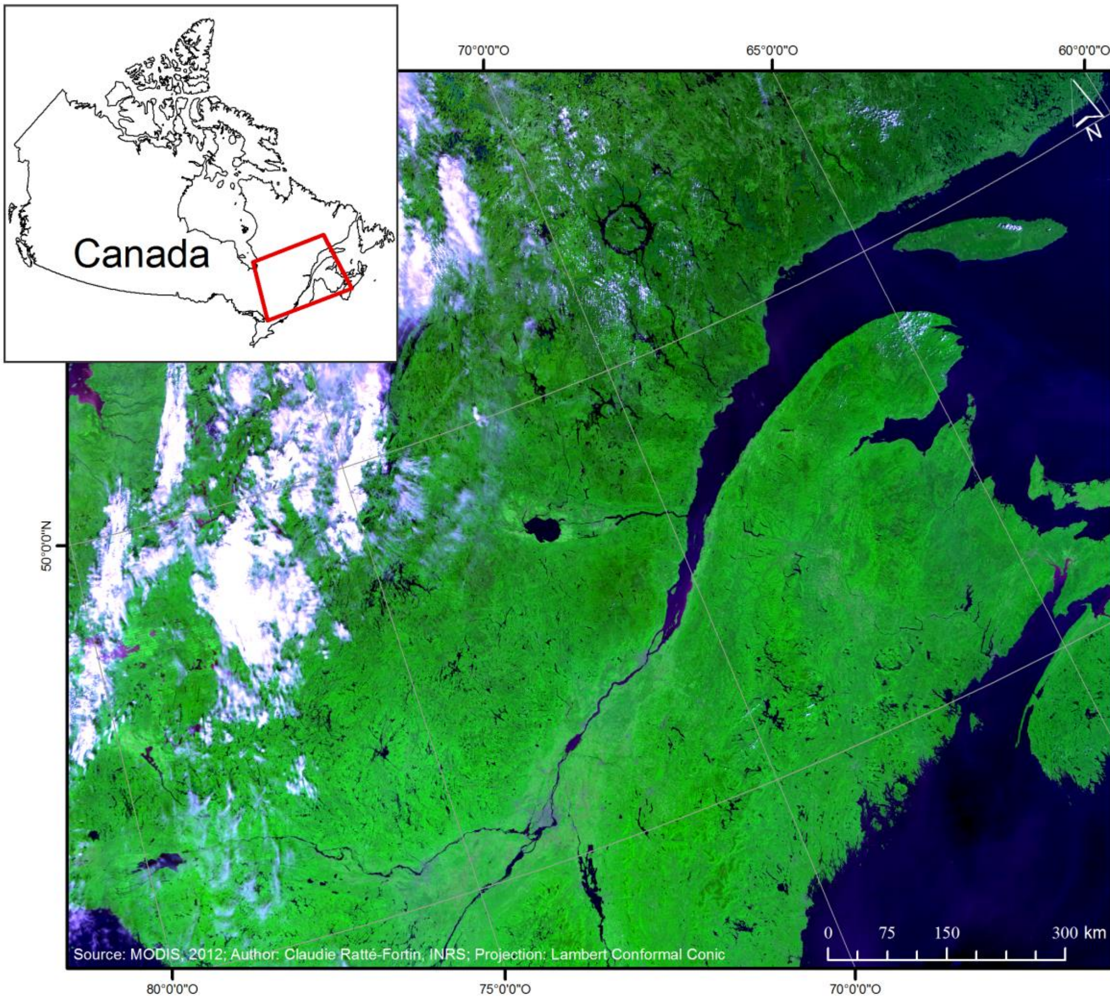

Figure 3: Detailed method used to distinguish between cloud and water classes using discriminant analysis.

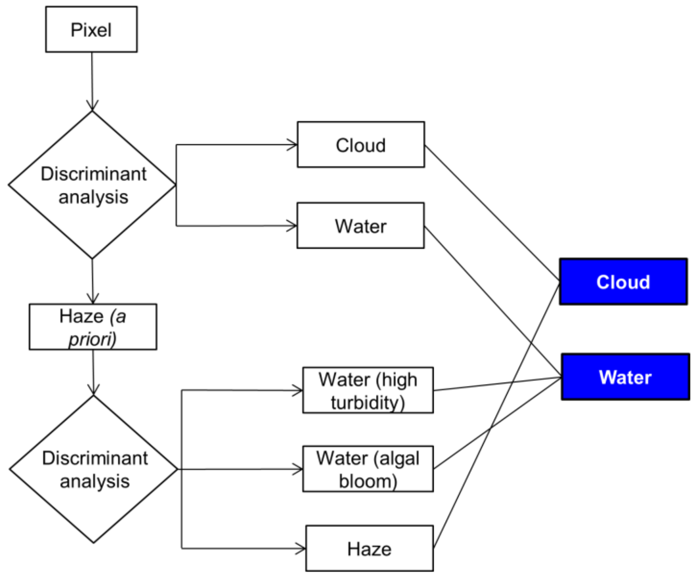

Figure 4: Details of the method used to estimate the distribution of the global success of the classification (%) and $\kappa$ using Monte-Carlo cross-validation.

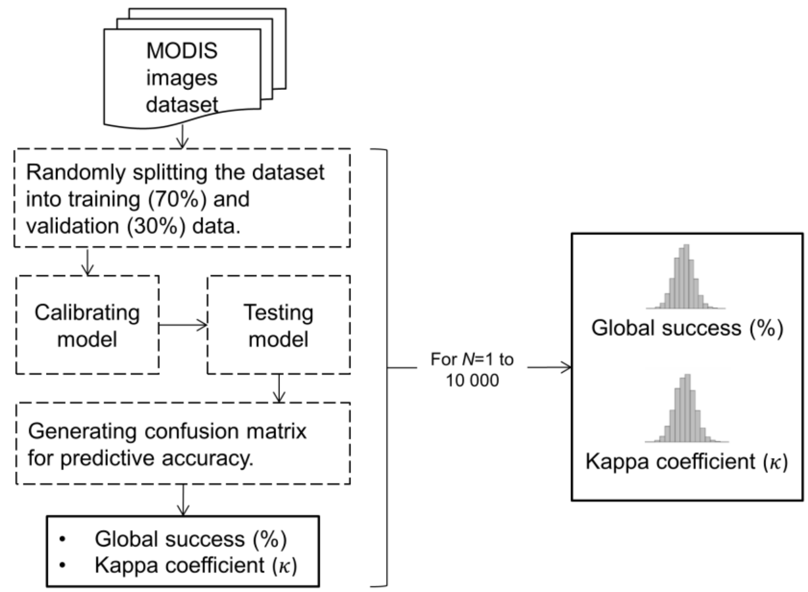

Figure 5 : (*a*) MODIS R-NIR-B color and R-G-B color of the Missisquoi Bay and the Champlain Lake, (*b*) the three cloud masks generated and (*c*) the corresponding chl-*a* concentration layers estimated with the remaining water pixels left (bottom-right). The red circles show high chl-*a* concentration values where there is *a priori* no bloom.

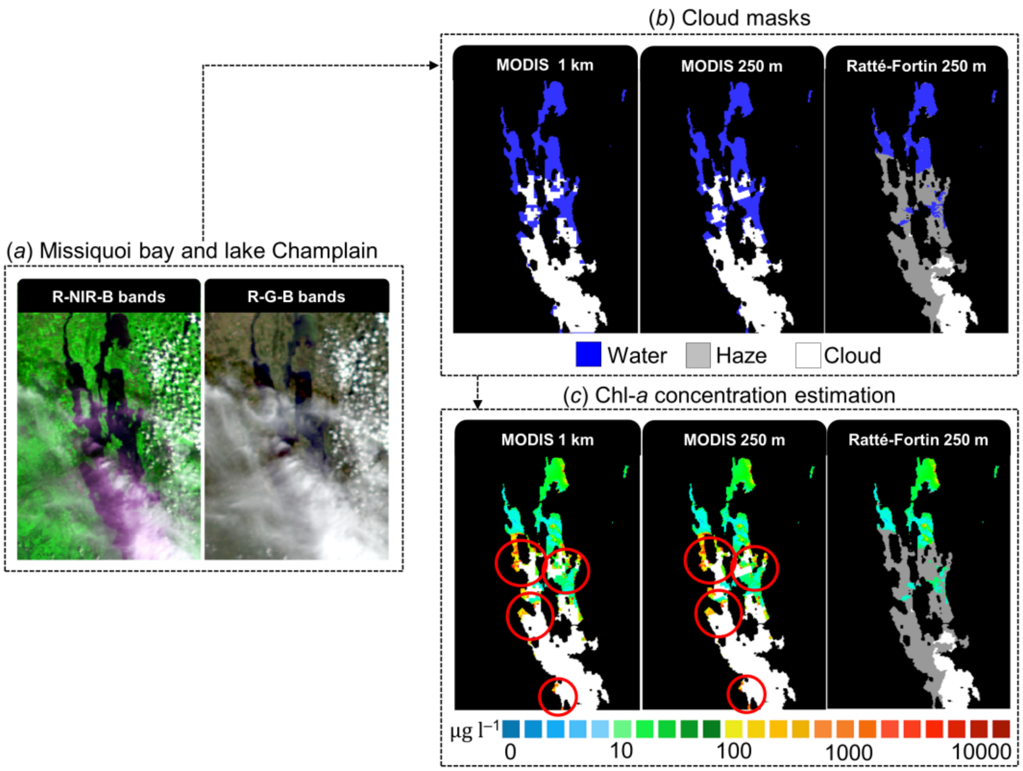

Figure 6 : MODIS R-NIR-B color and R-G-B color of the St-Lawrence river (*a*) and the lake Macamic (*c*), and the corresponding three cloud masks (*b*) and (*d*).

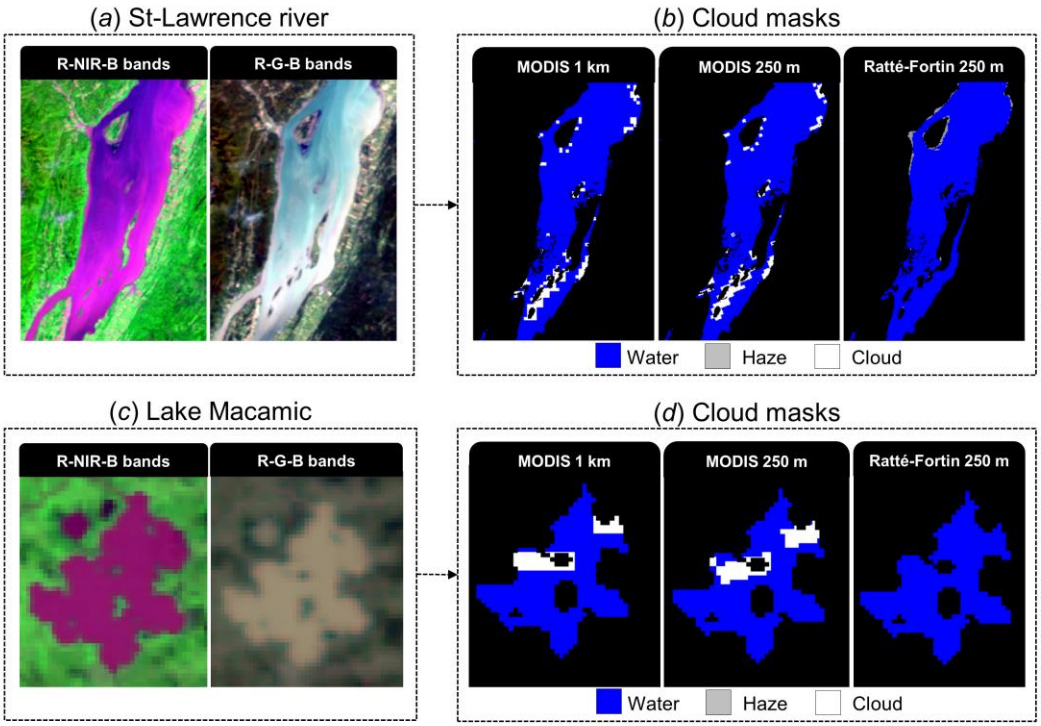

Figure 7 : MODIS R-NIR-B color and R-G-B color of the Bay of Fundy (*a*) and the St-Lawrence river (*c*), and the corresponding three cloud masks (*b*) and (*d*).

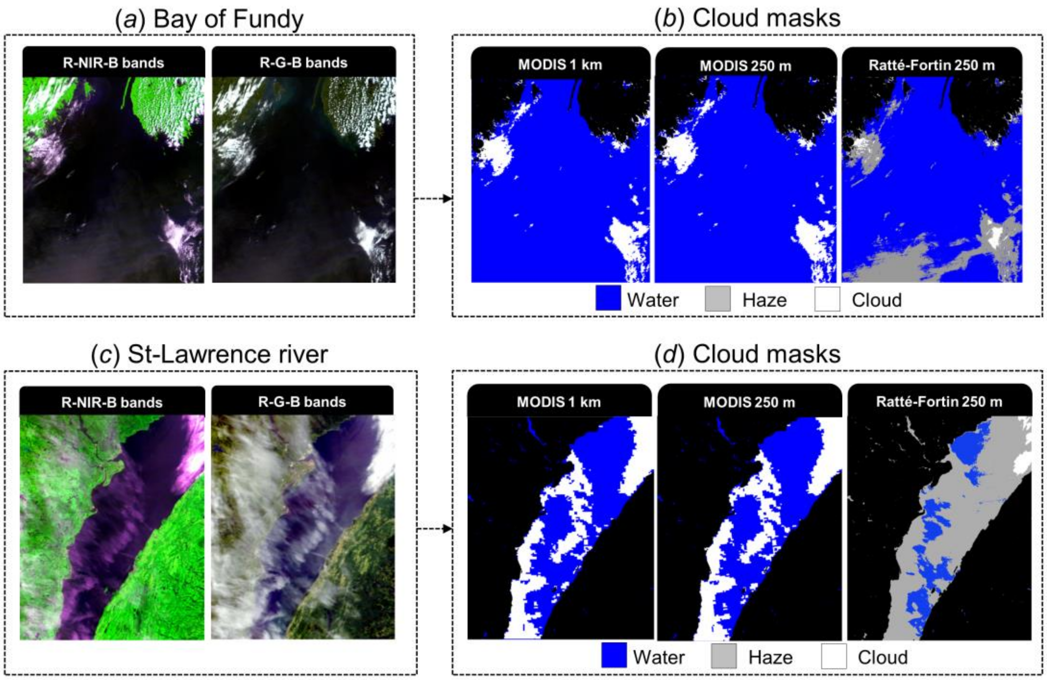

Table 1: Characteristics of the MODIS bands used in this study.

| MODIS sensor | | |
|---|---|---|
| Satellite | Terra (EOS AM-1) | |
| Operator | NASA | |
| Orbit | 705 km (ascending node) | |
| Temporal resolution | 1-2 days | |
| Quantization | 12 bits | |
| Swath | 2330 km | |
| MODIS bands | | |
| Band (resolution) | Wavelength (nm) | Description |
| 1 (250 m) | 620–670 | Red |
| 2 (250 m) | 841–876 | Near infrared |
| 3 (500 m) | 459–479 | Blue |
| 4 (500 m) | 545–565 | Green |
| 5 (500 m) | 1230–1250 | Short wave infrared |
| 6 (500 m) | 1628–1652 | Short wave infrared |
| 7 (500 m) | 2105–2155 | Short wave infrared |

Table 2: List of the MODIS images used for the model calibration and validation.

| Julian day | Year |
|---|---|
| 185-217-243-299 | 2000 |
| 262 | 2001 |
| 141-200-246-282 | 2002 |
| 133-195-231-267 | 2005 |
| 262 | 2007 |
| 136-189-234-293 | 2010 |
| 147-217-237-268 | 2013 |
| 170-201-234-266 | 2015 |
| **Number of images:** | **26** |

Table 3: Results of the double discriminant analysis confusion matrix with 95% confidence intervals (percentile 2.5 and 97.5 of the distribution) of global success and $\kappa$ means.

| | | **Observed** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Water | Fog | Cloud | Total | Commission error (%) | Success rate (%) | | |
| **Predicted** | Water | 1059 | 0 | 0 | 1059 | 0 | 100 | | |
| | Fog | 0 | 236 | 0 | 236 | 0 | 100 | | |
| | Cloud | 0 | 27 | 1329 | 1356 | 2 | 98 | | |
| | Total | 1059 | 263 | 1329 | 2651 | | | | |
| | Omission error (%) | 0 | 10.3 | 0 | | | | | |
| | Success rate (%) | 100 | 89.7 | 100 | | | 95% confidence interval of the mean | | |
| | **Global success (%)** | | | | | | 98.8 | **99.0** | 99.2 |
| | $\boldsymbol{\kappa}$ | | | | | | 0.979 | **0.982** | 0.986 |

Table 4 : Classification results of the two MODIS cloud products (1 km and 250 m) and the proposed approach.

| | **MODIS 1 km** | **MODIS 250 m** | **Ratte-Fortin 250 m** |
|---|---|---|---|
| Global success (%) | 91.3 | 95.3 | 99 |
| $\kappa$ | 0.827 | 0.905 | 0.982 |