AN ENSEMBLE BASED SYSTEM FOR CHLOROPHYLL-A ESTIMATION USING MODIS IMAGERY OVER SOUTHERN QUEBEC INLAND WATERS

Anas El-Alem [1,] Karem Chokmani [1], Isabelle Laurion [1] and Sallah E. El-Adlouni [2]*

[1] Institut National de la Recherche Scientifique Centre Eau-Terre-Environnement
[2] Moncton University

ABSTRACT

The purpose of this study was to enhance the performance of the Adaptive-model (AM) for Chlorophyll-a (Chl-a) concentration estimation using MODIS-imagery. The AM is based on the combination of spectral response classification and three semi-empirical *estimators*. Most critical step in Chl-a estimation when using the AM is the *estimator* selection. A wrong selection could lead to an over- or under-estimation and induces a staircase effect in Chl-a modelization. Ensemble based systems (EBS) development can be in interesting solution. One of the most used techniques to develop an EBS is *bagging*-algorithm. However, its main weakness is time consumption. The Gaussian-quadrature (GQ) formula has the potential to handle this kind of problems. The objective of this study is to develop a mixed ensemble based system for Chl-a estimation in Quebec inland waters using GQ formula. Statistical indexes evaluation was satisfied, relative-RMSE=15% and R^2=0.98 whereas matrix-confusion results were 0.92 and 0.8 for global-success and *Kappa*-index, respectively.

Index Terms— MODIS, HAB, Chl-a, Ensemble based systems, Gaussian quadrature.

1. INTRODUCTION

In reason of industrial, agricultural, and socio-economic development, the quality of many lakes has decreased because of the installation of Harmful algal blooms (HAB). On the other hand, with computer tools and satellite sensors development, remotely sensed data are increasingly used for monitoring HAB in inland waters [1]. HAB monitoring is possible due to the bio-optical activity of their principal pigment, the Chlorophyll-a (Chl-a), through bio-optical models linking inherent and apparent optical properties of water bodies. Thus, several models [2], indexes [3], and approaches [4] designed to estimate Chl-a concentrations in turbid water were recently developed, especially using medium-resolution spectroradiometers such as MODIS and MERIS sensors.

In this context, an Adaptive-model (AM) to estimate Chl-a concentration inland water bodies was recently developed based on the combination of water spectral response classification and three semi-empirical *estimators*. Discrimination between three trophic levels (waters poorly, moderately, and highly loaded in Chl-a), using a *classifier*, is firstly made before estimating the Chl-a with the corresponding *estimator* to the pre-identified trophic level (Figure 1 [5]). The *classifier* calibration was made using CART algorithm on the training data water spectral response. Once the AM *classifier* was calibrated, it was possible to sub-divide the same training data into three sub-groups allowing to calibrate three specific *estimators*.

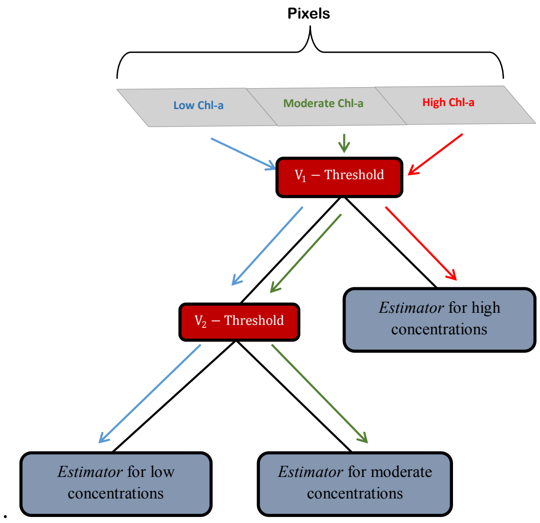
Pixels
Low Chl-a
Moderate Chl-a
High Chl-a
V1 – Threshold
V2 – Threshold
Estimator for high concentrations
Estimator for low concentrations
Estimator for moderate concentrations

Figure 1. Simplified scheme of the intern structure of the AM. V_1-Threshold and V_2-Threshold are the thresholds calculated by CART using, respectively, the Variable-1 and the Variable-2 on the training data.

Besides, most critical step in estimating Chl-a using the AM is the *estimator* selection. Indeed, Decision trees algorithms are known by their instability [6]. A wrong selection of a given *estimator* could then lead to a significant over- or under-estimation and induces a staircase effect to Chl-a estimations. Ensemble based classifiers (EBC) development seems to well manage this kind of classification problems [7]. On the other hand, because of modeling Chl-a concentration complexity in inland water bodies, many authors [8] suggest to seek for several *estimators* outputs

978-1-4799-5775-0/14/$31.00 ©2014 IEEE

before making a decision. The combination of all this outputs appears to be the most informative one. This technique, known under the name of an Ensemble based *estimators* (EBE), is highly used in hydrological [9] and financial [10] studies.

Among main keys to develop a strong EBC or EBE called also, in a general way, Ensemble based systems (BES) is to reach the highest diversity between the BES elements (*classifiers* or *estimators* [11]). One of the earliest, intuitive and, the simplest methods of an EBS development is the Breiman's *bagging*; short for *bootstrap aggregating* [12]. Diversity in *bagging* is obtained by using *bootstrapped* re-sampling on the training data; this allows covering almost all solution space. However, main weakness of this method is time consumption, particularly in imagery processing, due to the high number of both the EBS elements and images pixels.

The Gaussian-quadrature (GQ) formula could be an interesting method to solve this problematic. This method, frequently used in uncertainty propagation analysis [13], has the potential to convert re-sampling vector problem, which requires arduous calculations, to an integration problem where the numerical resolution techniques are simpler and are accurate and approved [14]. This technique is very suited for practical applications in hydrology and water studies by researchers not familiar with more complex statistical methods.

The objective of this study is then to develop a Mixed ensemble based system (MEBS) for Chl-a estimation in Quebec inland waters. The GQ application has permitted to develop an EBC composed of six *classifiers* and an EBE composed of twenty-one *estimators*. The MEBS has been evaluated by an adjusted k-fold *cross-validation,* confusion matrix, and the Coefficient of variance (CV).

2. METHODOLOGICAL APPRAOCH

2.1. Mixed Ensemble Based System calibration

The main key to enhance the AM performance is to avoid any mis-selection of the *estimator*. To handle this problem, it was necessary to determine an error variance on each AM thresholds (V_1- and V_2-threshold on Figure 1). This was possible by means of statistical moments (mean (μ_1) and variance (μ_2)) computation on the re-sampling vectors (v_1 and v_2) corresponding to each of the AM thresholds. The generation of v_1 and v_2 was made by re-computing the V_1- and V_2-threshold 25,000 times. Since the GQ requires the normality of the re-sampling vector, v_1 and v_2 were normalized using a *boxcox* transformation. Both vectors underwent a logarithmic transformation. The application of the GQ on v_1 and v_2 has allowed to compute an error margin (blue (upper-limit) and green (lower-limit) lines on the Figure 2) on either side of the two nominal thresholds (red lines on the Figure 2), which correspond to the MEBS-EBC.

By partitioning the [V_1, V_2] feature space with the developed EBC, it was possible to calibrate the MEBS-EBE composed of twenty-one *estimators*. Thirteen possible scenarios for estimating the Chl-a concentration were identified (due to paging limit, this section is not detailed at the present paper). Each one comprises a set of *estimators* combined by means of their corresponding weighting coefficients A*i* and A*j* (Table 1), using Eq (1). The GQ allows also to compute the variance on Chl-a estimates using Eq (2).

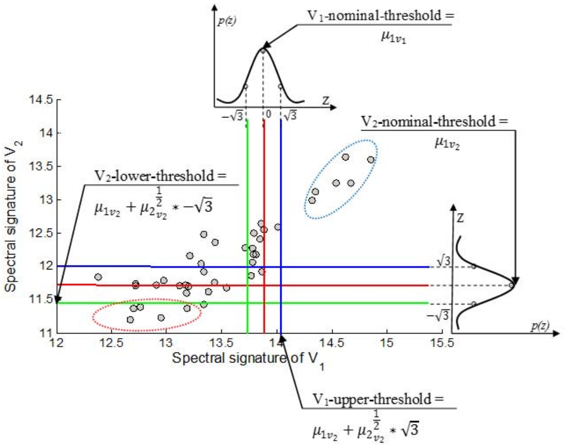

Figure 2. Scheme of the use of the Gaussian quadrature on 2-dimensions using two variables (V_1 and V_2) and its application to calibrate both *classifiers* and *estimators* of the mixed ensemble based system (MEBS). Red, blue and green lines represent, respectively, the nominal, upper and the lower thresholds of each variable (MEBS *classifiers*) and the sub-group of data surrounded by the blue ellipse represent the training data used to calibrate the *estimator* designed to estimate high chlorophyll-a concentrations whereas the sub-group of data surrounded by the red ellipse represent the training data used to calibrate the *estimator* designed to estimate low chlorophyll-a concentrations.

Table 1. Abscissas and weights for the standard normal distribution.

nodes	*z(i,j)*	*A(i,j)*
2	$-\sqrt{3}, 0, +\sqrt{3}$	$\frac{1}{6}, \frac{2}{3}, \frac{1}{6}$

$$[\mathrm{Chl-a}] = \sum_{i=0}^{n} A_i \sum_{j=0}^{n} A_j * \{f_i(Est), f_j(Est)\} \tag{1}$$

$$\mathrm{Var}[\mathrm{Chl-a}] = \sum_{i=0}^{n} A_i \sum_{j=0}^{n} A_j * \left[\{f_i(Est), f_j(Est)\} - \mu_1\right]^2 \tag{2}$$

Where *Est* is *estimator*, n is the number of each involved *Est*, for each pixel, at the finale estimation and A_i and A_j are their corresponding weighting coefficients.

2.2. Mixed ensemble based system validation

To assess the performance of the MEBS, it was evaluated by k-fold *cross-validation*, confusion matrix, and CV. The k-fold *cross-validation* consists in temporary removing a bloc of samples from the dataset and to use the remaining samples as training data to estimate the removed samples with the pre-calibrated sub-model. Once all Chl-a measurements are estimated, the model performance can be evaluated using statistical indexes such as the coefficient of determination (R^2), relative bias (BIASr), relative root mean square error (RMSEr), and relative NASH-Sutcliffe efficiency (Nr). Mathematical expressions of statistical indexes are as follow:

$$R^2 = \left[\frac{\sum_{i=1}^{n}(M_i - \overline{M})(Es - \overline{Es})}{\sqrt{\sum_{i=1}^{n}(M_i - \overline{M})^2}\sqrt{\sum_{i=1}^{n}(Es_i - \overline{Es})^2}}\right]^2 \quad (3)$$

$$BIASr = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Es_i - M_i}{M_i}\right) \quad (4)$$

$$RMSEr = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Es_i - M_i)}{M_i}\right)^2} \quad (5)$$

$$NASHr = 1 - \frac{\sum_{i=1}^{n}\left(\frac{M_i - Es_i}{M_i}\right)^2}{\sum_{i=1}^{n}\left(\frac{M_i - \overline{M}}{\overline{M}}\right)^2} \quad (6)$$

where n is the training data size and M, $\overline{M}$, ES and $\overline{Es}$ are, respectively, the measured and the estimated Chl-a concentrations with their corresponding averages.

A second data set, which is ordinal, was also used to assess the MEBS. This data set only indicates whether the cyanobacteria density was higher or lower than 20,000 cells per mL, equivalent to 10 mg Chl-a m^{-3}. Thus the evaluation time was made by confusion matrix.

A third evaluation by means of the Coefficient of variance (CV) was also made using Eq (7).

$$CV\ (\%) = \frac{\sqrt{\mu_2}}{\mu_1} * 100 \quad (7)$$

The CV measures the dispersion of individuals with respect to their mean. According to Martin and Gendron (2004), for a CV lower than 16% the mean can be considered as a reliable estimator, for a CV comprised between 16% and 33.3% the mean may comport some errors, and for a CV higher than 33.3% the dispersion is very important and the mean is no longer reliable [15].

3. RESULTS AND DISCUSSION

The k-fold *cross-validation* results were interesting as the coefficient of determination was about 0.98 showing that the MEBS can explain up to 98% of the variance of Chl-a concentration. The Nr, which is a severe evaluation index, indicates that the MEBS is robust with a success rate of 95%. The robustness of the model was also approved by the scatter-plot of *in situ* measurements versus their estimates (Figure 3) where all points are well distributed with respect to the line (1:1), confirming the accuracy of the MEBS estimations even at its extremities. The RMSEr was about 15% and the BIASr indicates that the MEBS underestimates Chl-a concentration by 2%.

The confusion matrix results (Table 2) indicate that globally the MEBS performance was satisfactory, especially for estimations higher than 10 mg Chl-a m^{-3} where the success rate reaches 96% and errors were lower than 8%. However, the accuracy was relatively less good for concentrations inferior than 10 mg Chl-a m^{-3} where omission and commission errors were about 9% and 19%, respectively. Nevertheless, the overall performance of the MEBS, when compared to the AM, has increased from 0.87 to 0.92 and from 0.69 to 0.80 for the global success and *Kappa*-index, respectively [5].

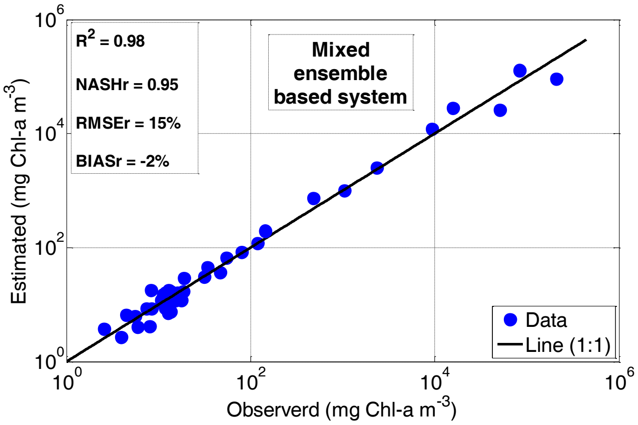

Figure 3. Chlorophyll-a concentration estimated from the ensemble based model compared to in situ measurements.

Table 2. The MEBS confusion matrix results.

		Measured				
		[Chla-]>10	[Chla-]<10	Total	Success Rate	Commission Error
Estimated	[Chla-]>10	**55**	2	57	96%	4%
	[Chla-]<10	5	**21**	26	81%	19%
	Total	60	23	**83**		
	Success Rate	92%	91%			
	Omission Error	8%	9%			
	Global Success					**92%**
	Kappa					**0.8**

In order to qualitatively evaluate the performance of the MEBS modeling to the original model, both models were applied on a bloom detected at the Bay Missisquoi of Champlain Lake (Figure 4). A clear correspondence between the bloom shape drown by the MEBS (Figure 4.B) and the HAB extension detected on the MODIS true color image (green hue (Figure 4.A)) can be seen. This concordance is somehow lost for the AM, especially for low-to-moderate concentrations where the model seems to overestimates the Chl-a (Figure 4.C).

The Figure 5 represents the spatial distribution of CV computed over the bloom detected on the Figure 4. Substantially all of Chl-a estimates are reliable, especially in high blooming condition (CVs≈0). For the rest of the Missisquoi Bay (moderate-to-low conditions), the CVs are generally less than 10%, except for some southern areas where it generally reaches 90%. The CV results are quite obvious since the highest dispersion coincides with low Chl-a (southern areas of the Missisquoi Bay (Figure 4.B)) where the highest commission and commission errors were obtained (Table 3).

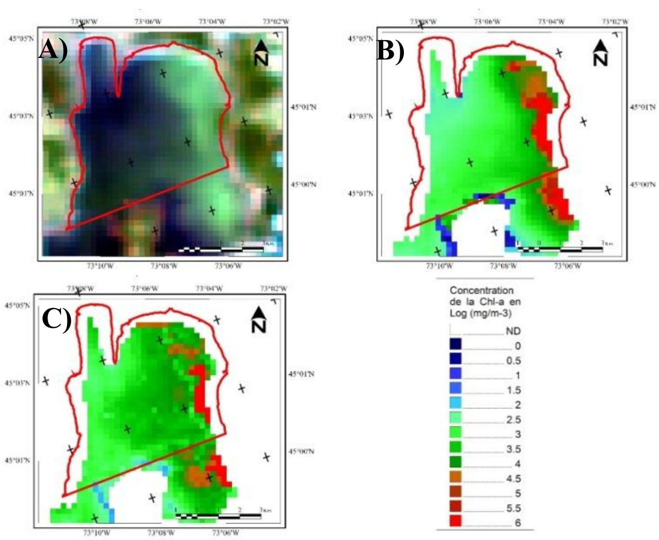

Figure 4. Comparison between the adaptive model and the mixed ensemble based system application on MODIS images. A) True color, B) OAM and C) AM. (19-Spet. 2001).

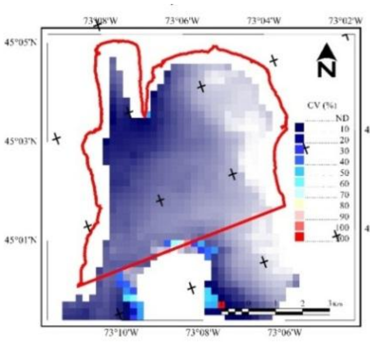

Figure 5. Spatial distribution of the coefficient of variance.

4. CONCLUSION

In general, when compared to the AM, the MEBS presented best results both quantitatively (RMSE=15%, R^2=0.98, and *Kappa*-index=0.80) and qualitatively (CV<10%) and showed a great potential to smooth the enormous contrast of Chl-a concentrations making the modelization more realistic, reliable, and accurate.

5. REFERENCES

[1] A. El-Alem, K. Chokmani, I. Laurion, and S. E. El-Adlouni, "Comparative Analysis of Four Models to Estimate Chlorophyll-a Concentration in Case-2 Waters Using MODerate Resolution Imaging Spectroradiometer (MODIS) Imagery," *Remote Sensing,* vol. 4, pp. 2373-2400, 2012.

[2] A. A. Gitelson, G. Dall'Olmo, W. Moses, D. C. Rundquist, T. Barrow, T. R. Fisher*, et al.*, "A simple semi-analytical model for remote estimation of chlorophyll-a in turbid waters: Validation," *Remote Sensing of Environment,* vol. 112, pp. 3582-3593, 2008.

[3] C. Hu, "A novel ocean color index to detect floating algae in the global oceans," *Remote Sensing of Environment,* vol. 113, pp. 2118-2129, 10// 2009.

[4] P. J. Baruah, M. Tamura, K. Oki, and H. Nishimura, "Neural network modeling of surface chlorophyll and sediment content in inland water from Landsat Thematic Mapper imagery using multidate spectrometer data," 2002, pp. 205-212.

[5] A. El-Alem, K. Chokmani, and I. Laurion, "Apport de la télédétection spatiale dans le suivi des épisodes de fleur d'eau d'algues dans les lacs du Québec méridional," Institut National de la Recherche Scientifique, Centre Eau-Terre-Environnementjuin-2013 2013.

[6] R.-H. Li and G. G. Belford, "Instability of decision tree classification algorithms," presented at the Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, Edmonton, Alberta, Canada, 2002.

[7] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review,* vol. 33, pp. 1-39, 2010/02/01 2010.

[8] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.,* vol. 3, pp. 79-87, 1991.

[9] H. L. Cloke and F. Pappenberger, "Ensemble flood forecasting: A review," *Journal of Hydrology,* vol. 375, pp. 613-626, 2009.

[10] C. Hung and J.-H. Chen, "A selective ensemble based on expected probabilities for bankruptcy prediction," *Expert Systems with Applications,* vol. 36, pp. 5297-5303, 2009.

[11] R. Polikar, "Ensemble based systems in decision making," *Circuits and Systems Magazine, IEEE,* vol. 6, pp. 21-45, 2006.

[12] L. Breiman, "Bagging Predictors," *Machine Learning,* vol. 24, pp. 123-140, 1996/08/01 1996.

[13] K. S. Kelly and R. Krzysztofowicz, "A bivariate meta-Gaussian density for use in hydrology," *Stochastic Hydrology and Hydraulics,* vol. 11, pp. 17-31, 1997/02/01 1997.

[14] H. Tørvi and T. Hertzberg, "Estimation of uncertainty in dynamic simulation results," *Computers & Chemical Engineering,* vol. 21, Supplement, pp. S181-S185, 5/20/ 1997.

[15] L. Martin and A. Gendron, *Méthodes statistiques appliquées à la psychologie: traitement de données avec Excel*, 2004.