High-resolution prediction of organic matter concentration with hyperspectral imaging on a sediment core

Kevin Jacq, Yves Perrette, Bernard Fanget, Pierre Sabatier, Didier Coquin, Ruth Martinez Lamas, Maxime Debret, Fabien Arnaud

High-resolution prediction of organic matter concentration with hyperspectral imaging on a sediment core

Kévin Jacq[a,b], Yves Perrette[a], Bernard Fanget[a], Pierre Sabatier[a], Didier Coquin[b], Ruth Martinez-Lamas[c,d], Maxime Debret[c], Fabien Arnaud[a]

a Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, EDYTEM, 73000 Chambéry, France

b Laboratoire d'Informatique, Systèmes, Traitement de l'Information et de la Connaissance (LISTIC), Université Savoie Mont-Blanc, 74944 Annecy Le Vieux Cedex, France

c Laboratoire de Morphodynamique Continentale et Côtière, Université de Rouen, UMR CNRS 6143, 76821 Mont-Saint-Aignan, France, Université de Caen, UMR CNRS 6143, 14000 Caen, France

d IFREMER, UR Géosciences Marines, Laboratoire Géophysique et Enregistrements Sédimentaires, BP70, 29280 Plouzané, France

Corresponding Author:

Kévin Jacq

Environnements, Dynamiques et Territoires de la Montagne (EDYTEM)

Université Savoie Mont Blanc, Campus Scientifique,

73376 Le Bourget du Lac Cedex, France

kevin.jacq@univ-smb.fr

1 **Keywords**: Hyperspectral Imaging, Chemometrics, Near-Infrared Spectroscopy, High

2 Resolution Analysis, Organic Matter

3

Abstract

In the case of environmental samples, the use of a chemometrics-based prediction model is highly challenging because of the difficulty in experimentally creating a well-ranged reference sample set. In this study, we present a methodology using short wave infrared hyperspectral imaging to create a partial least squares regression model on a cored sediment sample. It was applied to a sediment core of the well-known Lake Bourget (Western Alps, France) to develop and validate a model for downcore high resolution LOI550 measurements used as a proxy of the organic matter. In lake and marine sediment, the organic matter content is widely used, for example, to reconstruct carbon flux variations through time. Organic matter analysis through routine analysis methods is time- and material-consuming, as well as not spatially resolved. A new instrument based on hyperspectral imaging allows high spatial and spectral resolutions to be acquired all along a sediment core. In this study, we obtain a model characterized by a 0.95 r prediction, with 0.77 wt% of model uncertainty based on 27 relevant wavelengths. The concentration map shows the variation inside each laminae and flood deposit. LOI550 reference values obtained with the loss on ignition are highly correlated to the inc/coh ratio used as a proxy of the organic matter in X-ray fluorescence with a correlation coefficient of 0.81. This ratio is also correlated with the averaged subsampled hyperspectral prediction with a r of 0.65.

1. Introduction

The organic matter is usually used for sediment studies. For example, it can be used to estimate the carbon stocks over time. Among different methods, the organic matter can be approximated through Loss of Ignition at 550°C (Heiri et al., 2001), which are a widely used method in paleo-environmental studies. All these methods are time-consuming, relatively expensive, destructive and have low spatial resolution (0.5-1 cm sampling).

Hyperspectral imaging is a method at the interface of spectroscopy and imaging. The ability to predict the organic matter has been demonstrated based on visible (Vis, 400-800 nm) and near-infrared (NIR, 800-2500 nm) spectroscopy, known as Vis-NIR spectroscopy (Li et al., 2015; Nawar and Mouazen, 2017; Van Exem et al., 2018; Viscarra Rossel and Behrens, 2010). Predictions are still efficient with only NIR spectroscopy (Clairotte et al., 2016; Leach et al., 2008; Zornoza et al., 2008). Mid-infrared (MIR, 2500-25000 nm) spectroscopy also has the same ability (Clairotte et al., 2016; Rosén et al., 2011; Vohland et al., 2014). X-ray fluorescence (XRF) spectroscopy is also used as an indirect qualitative proxy of the organic matter (Bajard et al., 2016; Chawchai et al., 2016; Croudace and Rothwell, 2015; Lintern et al., 2016).

Chemometrics methods can be used to extract relevant information from spectral bands by comparison with discrete sample measurements and can be used downstream to predict the organic matter or other variables. Partial least squares regression (PLSR) has probably been the most used method to extract organic matter information from spectra (Clairotte et al., 2016; Dhawale et al., 2015; Nawar and Mouazen, 2017; Viscarra Rossel and Behrens, 2010; Vohland et al., 2014). However, other methods can be used with their

specific conditions. For instance, the artificial neural network (ANN) and the support vector machine (SVM) require a large dataset. A comparison of different chemometrics methods show that advanced methods such as ANN, SVM, and Multivariate Adaptive Regression Splines (MARS) can slightly improve the performance (Kuang et al., 2015; Li et al., 2015; Viscarra Rossel and Behrens, 2010). Transferring a model between several samples can be very interesting, and PLSR is better than ANN to model local variations, while they are quite similar for global variations (Wijewardane et al., 2016). In order to improve model performance and robustness, spectral variable selection is achieved to reduce the spectral collinearity. As a side effect, this reduction makes chemical interpretation easier (Peng et al., 2014; Viscarra Rossel and Behrens, 2010; Vohland et al., 2014). Furthermore, coupling spectroscopic datasets have been shown to improve model performance (Clairotte et al., 2016) by the increase of the spectral and thus chemical information.

Chemometrics methods have been widely used in the biochemistry and the pharmacology domains. These methods are usually based on an experimentally built sample set in which the target concentrations are known or spiked. A direct chemometrics approach is promising to predict concentrations in environmental solids such as sediments, soils, and biological tissues. The challenge is here to propose a procedure designed to build a robust model based on a natural dataset only. Indeed, we can expect that the range and spread of real target values of a learning dataset is less robust than a model generated with an experimentally designed dataset.

The aim of this study is to propose a way to build a PLSR model with hyperspectral data on a natural environmental sediment using the opportunity of its heterogeneity to generate robust models without the constitution of an experimental set. We applied this method to predict the organic matter content in a sediment core by the way of a PLSR

model. A sample from the Lake Bourget was used and analyzed with short-wave infrared (SWIR, 900-2500 nm) hyperspectral imaging. The loss on ignition (LOI) was the reference method for the organic matter values. LOI is predicted from a hyperspectral image and a variable selection applied to increase the robustness of the model. Predictions were validated with LOI550 measurements and compared with the inc/coh ratio calculated with XRF as a qualitative proxy of the organic matter. Finally, the pixel prediction validity with the scale law (micrometer and millimeter) is discussed.

2. Materials and methods

2.1. Site descriptions

Lake Bourget (231.5 masl, 18 km long, and 2.8 km wide) is a hard-water lake at the northwestern edge of the French Alps (figure 1). The lake was formed between the Pre-Alpine and Jura Mountain ranges within the Molasse Basin by the retreat of Wurmian glaciers. Two small rivers (Leysse and Sierroz) usually flow into Lake Bourget, which then flows into the Rhone River by the Saviere Channel. However, during flooding of the Rhone River, the water-current of this channel is reversed, and river water flows into Lake Bourget.

In the northern deep basin, flood deposits from the Rhone contribute to a variable sediment fraction. During such flood events, the Lake Bourget catchment area is 4,600 km^2, including tributaries draining part of the Jura Mountains and the Inner Alps. Otherwise, excluding the Rhone, it is 580 km^2. Over the last 10,000 years, the river-borne silicate fraction has ranged between 10 and 40% of the bulk sediment, with the remaining amount

composed of carbonate depending on past climate conditions (Arnaud et al., 2012, 2005; Debret et al., 2010; Giguet-Covex et al., 2010; Jenny et al., 2014).

In 2009, sediment (LDB09-P101, length = 54 cm, width = 9 cm) was cored in the northern basin of Lake Bourget (N45 45.334, E5 51.332, 145 m water depth) in the frame of the IPER-RETRO program (ANR-08-VUL 005, (Perga et al., 2015)). This core was selected for this study because it contains both seasonally paced bioinduced millimeter lamina as well as interbedded deposits brought by Rhone River flood events.

2.2. Sample preparation and LOI550 measurement

The core was sampled every ca. 2 cm in 5 mm slices cautiously avoiding the mixing of different sediment facies (i.e., lamina and flood deposits for example). However, in this 5 mm, there can be several lamina which are around 2 mm thick. Discrete samples were dried at 60 °C for 72 h then crushed (Basma et al., 1994).

The LOI was measured following the methodology detailed in (Heiri et al., 2001). Briefly, the protocol is as follows: (1) heat to 550 °C over 4 h to estimate the organic content and (2) heat to 950 °C over 2 h to estimate the mineral carbon content. At each step, the sample was weighed to calculate the loss of weight (wt%) with equations (1) and (2).

$$LOI550\ (wt\%) = \frac{m_0 - m_1}{m_0} * 100 \quad (1)$$

$$LOI950\ (wt\%) = \frac{(m_1 - m_2)}{m_0} * 100 \quad (2)$$

where m_0 is the initial weight of dried sediment, m_1 is the weight after the first step, and m_2 is the weight at the end.

The uncertainty of this method is 0.14 wt% ($\alpha = 0.05$, $n = 63$) for LOI550 values and 0.04 wt% ($\alpha = 0.05$, $n = 63$) for LOI950 values. It was estimated with 7 samples and 9 replicates each from several cores.

2.3. SWIR hyperspectral imaging acquisition

Hyperspectral imaging consists of acquiring an image with high spatial resolution, in which each pixel contains spectral information with a continuous spectral resolution.

The core was analyzed in less than 15 minutes with the SWIR hyperspectral camera (Specim Ltd., Finland) with the lens OLES22,5 at the M2C lab, University of Normandie-Rouen. It covers the SWIR range between 968 nm and 2574 nm with a reflectance spectral resolution of 12 nm (144 wavelengths) and a spatial theoretical resolution of 200 µm in both directions all along the core. The spectral unit is reflectance by 10,000 (R x 10,000).

The protocol followed to acquire the hyperspectral image is detailed in (Butz et al., 2015). The core was cleaned before acquisition. Then, the camera was calibrated with a spectralon reference (white), and the shutter was closed (black) for determining the spectral dimensions. Spatial calibration was achieved by the way of imaging a known object for its shape (squared pixels) and color (intensity). Deviation was checked at the end of the acquisition in the same manner, and no deviation was observed.

2.4. X-ray fluorescence spectroscopy

The relative contents of major elements were analyzed by X-ray fluorescence (XRF) at a 200 µm resolution on the surface of the sediment core with an ITRAX XRF Core Scanner (Cox Analytical Systems) at the CEREGE laboratory in 2010. The split core surface was first covered with a 4-µm-thick Ultralene film to avoid contamination and desiccation of the sediment. The X-ray beam was generated with a molybdenum tube at 35 kV and 30 mA, with a runtime of 15 s. Compton (incoherent) and Rayleigh (coherent) scattering data were extracted from it, and the inc/coh ratio was calculated. It has been used in many studies as a qualitative proxy of organic matter (Bajard et al., 2016; Chawchai et al., 2016; Croudace and Rothwell, 2015; Lintern et al., 2016).

2.5. Data analysis

Data acquisition was performed using Specim hardware ENVI 4.8. The data was then converted, processed and analyzed with MATLAB (R2017a, MathWorks). Several free and open MATLAB toolboxes were used and are detailed afterwards.

A key point to predict the LOI550 from hyperspectral imaging is the coupling of the volume of the sample at 5 mm resolution with surface pixels of approximately 0.2 mm. The proposed method is composed of four main steps. (1) A bootstrapping selection in the hyperspectral spectra of the LOI sampling area was performed to construct several datasets. (2) Then, PLSR was used to model LOI550 thanks to these spectra and the discrete LOI550 measurements for all the datasets. (3) The 10 optimal models were retained based on their performances. (4) Wavelengths highly correlated with LOI550 were identified with variable

selection on the retained models in (3). A final reduced model was estimate with these wavelengths. The complete workflow followed in this study is presented in Figure 2.

2.5.1. Spectral preprocessing

Spectral bands between 968-1127 nm (15 wavelengths) and 2418-2574 nm (15 wavelengths) were excluded from the processing because of their very low signal-to-noise ratio. In fact, noise is known to artificially increase the prediction of a chemometrics model (Tetko et al., 1995). 39 wavelengths of H_2O absorption bands (between 1094 and 1176 nm, 1339 and 1465 nm, and 1773 and 2005 nm) were removed as proposed by (Gomez et al., 2015). Consequently, 75/144 wavelengths were retained (figure 3).

In order to produce a robust model, some data in the spectra and LOI550 values have been removed based on two standard deviations of the Mahalanobis distance in a principal component analysis space (Mark and Tunnell, 1985). These data were due to spectral artifacts (specular reflection, not enough signal, too noisy) or aberrant values. Pixels that correspond to high surface variations (gap areas for example) were removed by setting a lower limit for the standard deviation of a spectrum. This threshold must be set for each image and based on the spectrum regions without noisy bands. For Lake Bourget images, a standard deviation limit is set at 250 R x 10,000 (reflectance per 10,000).

Several standard spectral preprocessing (detrend, standard normal variate (SNV), Savitzky-Golay derivatives) were tested to correct spectra from the scattering effect and/or to normalize spectra (Vidal and Amigo, 2012). Finally, in order to optimize the prediction model, the data were preprocessed by the way of an autoscaling by mean-centering and standardizing each wavelength (Z-score).

184

185 2.5.2. Spectral sub-sampling

186

187 Our goal is to build a regression model between a set of 20 data of LOI and a set of 20
188 regions of thousands of pixels on the image which correspond to the sampling area.
189 Different techniques could be used to generate a representative value of the spectrum used
190 for regression with LOI data, (mean, bootstrap, min or max, mode or median spectrum). The
191 limit of a mean spectrum for each sampling area leads to smooth spectral variabilities;
192 resulting model were not relevant. We applied a bootstrapping procedure rather than a
193 mean spectrum assuming that LOI could be better explained by a specific spectrum of the
194 region of interest than by the mean spectrum of this region. Bootstrapping was used to
195 randomly select 100 spectra in each sampling area to generate 100 datasets of 20 spectra.

196

197 2.5.3. Partial least squares regression

198

199 PLSR was used to establish a relationship between the LOI550 values and SWIR
200 hyperspectral imaging. This method is based on the extraction of orthogonal predictors (also
201 called latent variables, LV) corresponding to the maximum variability in the spectral bands
202 used as predictors linked to one or several predicted variable(s) (Wold et al., 1984).
203 Each of the 100 bootstrapped datasets was divided in a calibration set (13 data) and a
204 validation set (the 7 unused data). For each of the 100 bootstrapped datasets, a PLSR model
205 is generated based on the calibration set, and then applied to the validation set in order to
206 assess the performances of the prediction.

The optimal number of latent variables was estimated by the use of the Durbin Watson test (Durbin and Watson, 1950) which is similar to a signal-to-noise ratio and was used to improve the robustness of the model.

Performances of the prediction model were estimated using the coefficient of correlation, r, and the uncertainty of both data sets (calibration and validation sets). They are calculated by equations (3) and (4).

$$r = \sqrt{1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}} \quad (3)$$

$$Model\ uncertainty = 2 \times \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (4)$$

where y_i is a LOI550 reference value, $\hat{y}_i$ is a LOI550 predicted value, $\bar{y}_i$ is the mean LOI550 reference value, and n is the number of values.

2.5.4. Variable selection

Variable selection algorithms were used to decrease the redundancy between neighboring wavelengths and remove non-informative ones to make the models more robust and easier to interpret chemically. The algorithms that were used can be found in MathWorks and are detailed in (Leardi and Lupiáñez, 1998; Li et al., 2014). The genetic algorithm was the method that gives the optimal model performance, as it aims is to solve an optimization problem by finding the lowest number of wavelengths to have the optimal prediction performance. It is inspired by natural selection, where wavelengths that are highly correlated to the LOI550 are first selected and then the optimal combination is chosen.

The GA-PLSR of the 10 best performing bootstrapping models were used to increase the certainty for the selected wavelengths. Then, the wavelengths are ranked by their number of occurrences. PLSR models are calculated by increasing the number of ranked selected wavelengths. Finally, a model is chosen with the optimal r and model uncertainty. At last, a geochemical map is estimated by the prediction of all the hyperspectral pixels to observe the distribution down and across the core. Once an optimal model is created, it can be used theoretically without other sampling on other cores sampled in the similar sedimentation environment.

3. Results

3.1. Lithology and OM content

The Lake Bourget sediment core presents three main units, which were described in detail in (Giguet-Covex et al., 2010). The uppermost one, between 0 cm and 18 cm, presents millimeter-thick biochemical varve and corresponds to anoxic conditions at the lake-sediment interface, with LOI550 values greater than 6 wt% (Giguet-Covex et al., 2010; Jenny, 2013). The second unit, between 18 cm and 26 cm, presents disturbed varve and LOI550 values that are stable at 6 wt%. In the lowermost unit, between 26 cm and 54 cm, there are fuzzy laminated structures, and the LOI550 values are quite stable at 5 wt%. Flood-triggered deposits (Jenny et al., 2014) are also present in the two first units, with a major one between 13.5 cm and 14.5 cm. Twenty measurements of the LOI were performed with a range

between 4.97 wt% and 8.60 wt% and a mean value of 6.41 wt% along with a standard deviation of 1.37 wt%. LOI550 variations according to the depth can be seen in figure 4.d.

3.2. Organic matter predictive models

Four PLSR models are presented in table 1 to illustrate the importance of removing water bands and wavelength selection.

Having a high number of LVs can increase the r of calibration but decrease the model robustness because more LVs increase the calibration noise, which is not present in the validation set. This is shown between models with and without water bands (table 1 blue and green). When the number of LVs is higher, the calibration r is too, but the validation r is lower.

Wavelength selection leads to increase the performance of the prediction model. This is visible with a number of latent variables (LV) that is smaller (6-8) and an uncertainty that decreases when less wavelengths are used.

Finally, the best compromise was found with a model with 27 wavelengths, validated with 6 latent variables, and a prediction r of 0.95 ($p < 0.05$). A model uncertainty of 0.77 wt% ($\alpha = 0.05$, $n = 20$) for LOI550 prediction by hyperspectral imaging was found, whereas it is 0.14 wt% ($\alpha = 0.05$, $n = 20$) for the LOI550 analysis. This is in agreement with PLSR models developed with spectroscopic devices in the literature (Clairotte et al., 2016; Leach et al., 2008; Li et al., 2015). This uncertainty is quite important because of the low number of LOI550 values and their low dispersion. We expect that a greater and spreader set of LOI values should lead to a more accurate PLSR model.

3.3. Wavelength correlation with organic matter

Variable selection algorithms are dependent on the spectral preprocessing (Peng et al., 2014). First models with O-H bands show that they were selected as relevant, whereas some organic bands were missing. These models, however, were less robust in terms of prediction. These results could depend on O-H water bands that disturb the model. Removing O-H wavelengths can increase the performance and robustness if other bands can add relevant information to replace the missing information of organic matter O-H bands.

The genetic algorithm allows the number of wavelengths to be reduced to 27 (figure 3). They can be divided into three main classes, organic, carbonate and clay bands according to (Peng et al., 2014; Viscarra Rossel and Behrens, 2010). Organic bands correspond to C-H bonds (1205, 1217, 1622, 1735, 1746, 1757, 2250, 2261, 2272, 2306, 2317, and 2328 nm), C-O bonds (2127 and 2138 nm), and N-H bonds (2093, 2105, and 2116 nm). Bands associated with C=O bonds (1487 and 1510 nm) corresponds to organic or carbonate bands, and these may discriminate the lamina type (dark or light). For clay bands, 5 bands were selected (2149, 2160, 2183, 2228, and 2239 nm). The three other bands are not directly associated with chemical bonds (1273, 1555, and 1566 nm). Some bands are spectrally close (for example, 1205-1217 nm and 1746-1757 nm), which could reveal some band shape changes due to organic mineral interaction.

3.4. Range, prediction and confidence intervals

The PLSR model was made with LOI550 measurements that have a range of 4.97 wt% to 8.60 wt%. We assume that LOI550 model intercept at the origin (there is no organic matter spectroscopic signal when there is no organic matter in the sediment). Then we consider that we can predict LOI550 in the range 0 wt% to 8.60 wt%. 93 % of the data are included in it and have a minimal prediction error of 2.25 wt% ($\alpha = 0.05$) in the center of the calibration (6.26 wt%).

Prediction and confidence intervals were calculated thanks to the complete dataset (figure 5.c). The prediction error map (figure 5.b) shows side effects (right and left), which can be due to surface variation or illumination. This observation is also visible in the concentration map. Obviously, areas with high surface variations (gap, fissures) will not be used in the LOI550 profile, and their prediction errors are more important than other predicted values at the same depth. Figure 5.e shows that the prediction error is high in the middle of light lamina (near 2.90 wt%), whereas it is low for the dark lamina (near 2.30 wt%). Predictions of these light lamina concentrations have a range between 9-12 wt% (figure 5.d). These predictions, out of the calibration range, must be interpreted cautiously.

3.5. Prediction distribution

The LOI550 concentration map in figure 4.c shows the three lithology units. A horizontal merging is used to estimate a time series of LOI550 (figure 4.e) prediction to be compared with LOI550 reference values (figure 4.d). The first unit presents LOI550 prediction values that oscillate with a maximum near 11 wt% and a minimum of 4 wt%, and the global trend is a decrease from 11 wt% to 6 wt%, which is similar to the reference values. The second unit is characterized by a lower variation between 5-7 wt% and a steady trend at approximately 6

wt%. The last unit is quite homogeneous at 5 wt%. Floods can be characterized with a rapid decrease in the prediction value, for example, the one between 13.5 and 14.5 cm.

Focusing on unit 1, the variations of the average LOI550 predicted values (figure 6.a, blue curve) appear to closely match the varve laminations, with high values for light lamina and low values for black lamina. This is also in agreement with the LOI550 reference values (green line) by averaging a 5 mm thickness. In the LOI550 concentration map (figure 6.b), the sequence of varve is observable along the core with 200 µm resolution. Even if values are out of the range, a qualitative study can be made. Variations inside light lamina show that the middle has a higher concentration than the edges, and for dark lamina, an opposite trend exists, with the lowest values in the middle. There can be mixing between lamina inside the sediment or inside the spectra. Additionally, it is heterogeneous across lamina with LOI550 variations. In a flood, the average LOI550 predicted values (figure 6.c) present an upward decreasing trend from 5 wt% to 6.5 wt%. The LOI550 concentration map (figure 6.d) shows that values vary a lot, approximately 3 wt% across the core.

4. Discussion

4.1. Comparison of LOI550 with the inc/coh ratio by XRF

In order to discuss this prediction, we compare this prediction with another proxy of organic matter which can be measured at a similar spatial resolution. The incoherent to coherent ratio was calculated with an XRF core scanner on this core. This ratio is used as an indirect qualitative proxy of the organic matter, whereas the LOI550 is quantitative. To ensure the

pertinency of the comparison of PLSR model and inc/coh proxy, we compare first with LOI550 with this proxy for this sample.

This ratio has been subsampled, the 25 pixels corresponding to 5 mm of LOI550 were averaged. LOI550 measurements and the inc/coh are correlated (r = 0.81; $p < 0.05$) and can thus be compared. Figure 4.d,e show that differences are mainly in unit 2 with inc/coh values similar to unit 1, whereas LOI550 measurements between these two units are decreasing. This could be due either to the surface state (oxidation) or to the difference between a surface (inc/coh) and a volume (LOI550) measurement.

The hyperspectral quantitative prediction can also be compared with the inc/coh ratio when resampled at the same resolution. Figure 4.f shows the ratio variations along the core, and the three units are characterized with high, medium, and low oscillations in the hyperspectral prediction as well as variation in the flood deposit. The correlation coefficient between the LOI550 prediction and inc/coh is 0.65 ($p < 0.05$). With the high-resolution image on varved lamina (figure 6.a, c), it is easier to show and compare both variations. The LOI550 quantitative prediction and inc/coh qualitative ratio are in agreement for varve lamina and flood deposits. When the varve are not parallel or mixed, XRF cannot detect them, whereas hyperspectral imaging can, and that can be seen at a 4 cm depth. Between depths of 3.2 cm and 3.6 cm, values of inc/coh and predicted LOI550 are steady because of the presence of irregularly shaped varve. For hyperspectral imaging, it is possible to see them on the concentration map, and with image processing, it will be possible to adjust the varve sequence. For the flood (figure 6.c), the inc/coh ratio increases faster down the core than the hyperspectral prediction, which may indicate that one of the two methods is biased by grain size variations.

4.2. Micrometric surface prediction and volume analysis

The analysis of the LOI550 *vs* inc/coh leads to suspect a difference between surface *vs* volume measurement. This volume-surface correlation is noted $r_{V/S}$. LOI550 measurements and the average predicted values are correlated with a $r_{V/S}$ of 0.73 ($p < 0.05$) and a root mean square error (RMSE) of 1.72 wt%.

Some points between depths of 6 cm and 12 cm are affected by a surface variations such as the gap on the top right of the core (figure 4.a, c), and without this area, the correlation $r_{V/S}$ increased to 0.81 ($p < 0.05$). This induces a loss of spectral intensity and thus weaker LOI550 prediction values. Cores need to have as large as possible of a plane surface to have a correct prediction.

The $r_{V/S}$ is weaker than that of the model due mainly to chemical properties. The surface state can be the main explanation, as the oxidation of some chemical compounds can induce intensity variations on the selected bands in some pixels. That is why the core needs to be cleaned before beginning, and the acquisition needs to be fast for no reappearance of surface states between the first and the last spectra. The other differences can mean that the surface and volume sediment analysis don't evaluate the same property for the organic matter content.

4.3. Proposed methodology

The proposed methodology seems to be relevant with the validation of PLSR models by XRF comparison and by the characteristics of the chemical bonds selected with the variable

selection algorithm. However, based on the concentration map and selected spectral bands, we can wonder if a unique PLSR model is the optimal method in a sediment characterized by three different mineralogical patterns: the two types of lamina (light and dark) and the homogeneous part. If PLSR models were created for each of these three patterns independently, prediction could be more accurate with different selected wavelengths. With our LOI measurements, it is not possible to estimate a model by the type of lamina because the sampling resolution of an operator does not allow the ability to distinguish between them. Improvement of such a model could occur by the way of a higher sampling resolution that is closer to the lithology of the sediment core.

4.4. Outlook for PLSR application on sediment cores and other environmental matrices

Quantitatively predicting proxies at high resolution along and across the core in less than an hour, from the acquisition to the prediction, is very effective and is not done by any other methods at this time. This is an important advance for the study of sediment cores to infer the paleo-environment and paleo-climate. The high spatial resolution allows us to look inside the laminae, and the spectral dimensions allow chemical variations to be studied with further work to identify the patterns of organic matter. The high-resolution approach is relevant in some type of sedimentary environments where there are successive deposits as the lamina or the varve and even in some homogenous sediment. In the case of a very homogeneous sediment such as those subject to bioturbation or redistribution, the high

411 resolution loses some interests, but the detection of specific compounds may still keep an
412 interest for a global characterization (without chronology).

413 The use of other proxies may help to improve the quality of the predictions. With this
414 application on the Lake Bourget core, grain size effects seem to skew the LOI550 modeling.
415 Predicting several proxies at the same time can make the model more robust. For example, if
416 the grain size has a real effect on the LOI550 prediction, creating a unique model could
417 increase the robustness. Furthermore, the spatial and the spectral dimensions of the
418 hyperspectral images may be combined in the data processing to improve our understanding
419 of the sediment records. For instance, thanks to lamina, chemical, physical and biological
420 variations could be estimated at the seasonal scale. In the Lake Bourget core, organic carbon
421 fluxes could be estimated for each type of laminae (dark or light) and compared in eutrophic
422 and non-eutrophic parts of the core (Jenny, 2013).

423 Creating a universal model that can be applied to several sedimentation sites (lake,
424 marine, etc.) and types (eutrophic, detrital, etc.) needs standardized parameters. Obviously,
425 the hyperspectral acquisition is a critical step that need to follow a strict protocol (Butz et al.,
426 2015). Even if a model seems to be applied to a core which was not used on the model
427 generation, some few chemical verifications may be achieved in order to avoid unexpected
428 bias. As a first step, the model could be theoretically transferred on samples cored in similar
429 sedimentary environments to keep the same type of matrix effects.

430 Obviously, with all these considerations, this methodology can be transferred to
431 other natural heterogeneous samples, for example speleothems, soils, ice cores, and trees,
432 to infer the paleo-environment, paleo-climate, soil health, and pollution.

433

434 ## 5. Conclusion

435

436 Hyperspectral imaging is a high resolution (200 µm), nondestructive and fast analysis
437 method (15 minutes). Coupling hyperspectral imaging and partial least squares regression
438 shows great possibilities for the creation of quantitative predictive models and the
439 prediction for any kind of natural sample at high resolution. In this study, a methodology was
440 proposed to estimate a robust PLSR model using the heterogeneity of the sample without
441 the constitution of a specific calibrated concentration range.

442 It was applied to LOI550 prediction and it was validated with LOI550 measurements
443 with selected near-infrared wavelengths that correspond to relevant chemical bonds. It was
444 also successfully compared with the inc/coh XRF qualitative ratio used as a proxy of the
445 organic matter. Therefore, this methodology seems to be relevant and allows for the organic
446 matter content to be quantitatively inferred at high resolution along and across the core.
447 This proxy for laminated or varve sediment could be precise at the seasonal scale with the
448 200 µm pixel size.

449 This methodology must be tested on other cores or heterogeneous samples to verify
450 its relevance. The “universality” of the model generated from this dataset, should be applied
451 on different lake systems in order to test the ability to predict organic matter in other
452 environments and to set the limitations of this. It may be used on other proxies that are
453 characteristic in the basin catchment. Predicting several proxies at the same time may also
454 be studied to create a robust model that is not disturbed by some chemical or physical
455 variabilities. The proposed methodology applied to sediment cores has great possibilities to
456 more precisely infer the paleo-climate and paleo-environment that are recorded in the core
457 inside each sediment structure.

458

Acknowledgements

The core used in this study is stored in the EDYTEM laboratory that also performed the LOI analysis in 2009 during the IPER-RETRO program (ANR-08-VUL 005). Hyperspectral imaging was processed at the University of Normandie-Rouen and was funded by the Region Normandie, which supports the scientific consortium SCALE UMR CNRS 3730.

References

Arnaud, F., Revel, M., Chapron, E., Desmet, M., Tribovillard, N., 2005. 7200 years of Rhône river flooding activity in Lake Le Bourget, France: a high-resolution sediment record of NW Alps hydrology. Holocene 15, 420–428. https://doi.org/10.1191/0959683605hl801rp>

Arnaud, F., Révillon, S., Debret, M., Revel, M., Chapron, E., Jacob, J., Giguet-Covex, C., Poulenard, J., Magny, M., 2012. Lake Bourget regional erosion patterns reconstruction reveals Holocene NW European Alps soil evolution and paleohydrology. Quat. Sci. Rev. 51, 81–92. https://doi.org/10.1016/j.quascirev.2012.07.025

Bajard, M., Sabatier, P., David, F., Develle, A.-L., Reyss, J.-L., Fanget, B., Malet, E., Arnaud, D., Augustin, L., Crouzet, C., Poulenard, J., Arnaud, F., 2016. Erosion record in Lake La Thuile sediments (Prealps, France): Evidence of montane landscape dynamics throughout the Holocene. The Holocene 26, 350–364. https://doi.org/10.1177/0959683615609750

Basma, A.A., Al-Homoud, A.S., Al-Tabari, E.Y., 1994. Effects of methods of drying on the engineering behavior of clays. Appl. Clay Sci. 9, 151–164. https://doi.org/10.1016/0169-1317(94)90017-5

Butz, C., Grosjean, M., Fischer, D., Wunderle, S., Tylmann, W., Rein, B., 2015. Hyperspectral imaging spectroscopy: a promising method for the biogeochemical analysis of lake sediments. J. Appl. Remote Sens. 9, 1–20. https://doi.org/10.1117/1.JRS.9.096031

Chawchai, S., Kylander, M.E., Chabangborn, A., Löwemark, L., Wohlfarth, B., 2016. Testing commonly used X-ray fluorescence core scanning-based proxies for organic-rich lake sediments and peat. Boreas 45, 180–189. https://doi.org/10.1111/bor.12145

488 Clairotte, M., Grinand, C., Kouakoua, E., Thébault, A., Saby, N.P.A., Bernoux, M., Barthès,
489 B.G., 2016. National calibration of soil organic carbon concentration using diffuse
490 infrared reflectance spectroscopy. Geoderma 276, 41–52.
491 https://doi.org/10.1016/j.geoderma.2016.04.021
492 Croudace, I.W., Rothwell, R.G., 2015. Micro-XRF studies of sediment cores : applications of a
493 non-destructive tool for the environmental sciences. https://doi.org/10.1007/978-94-
494 017-9849-5
495 Debret, M., Chapron, E., Desmet, M., Rolland-Revel, M., Magand, O., Trentesaux, A., Bout-
496 Roumazeille, V., Nomade, J., Arnaud, F., 2010. North western Alps Holocene
497 paleohydrology recorded by flooding activity in Lake Le Bourget, France. Quat. Sci. Rev.
498 29, 2185–2200. https://doi.org/10.1016/J.QUASCIREV.2010.05.016
499 Dhawale, N.M., Adamchuk, V., Prasher, S.O., Viscarra Rossel, R.A., Ismail, A.A., Kaur, J., 2015.
500 Proximal soil sensing of soil texture and organic matter with a prototype portable mid-
501 infrared spectrometer. Eur. J. Soil Sci. 66, 661–669. https://doi.org/10.1111/ejss.12265
502 Durbin, J., Watson, G.S., 1950. Testing for Serial Correlation in Least Squares Regression: I.
503 Biometrika 37, 409–428. https://doi.org/10.2307/2332391
504 Giguet-Covex, C., Arnaud, F., Poulenard, J., Enters, D., Reyss, J.-L., Millet, L., Lazzaroto, J.,
505 Vidal, O., 2010. Sedimentological and geochemical records of past trophic state and
506 hypolimnetic anoxia in large, hard-water Lake Bourget, French Alps. J. Paleolimnol. 43,
507 171–190. https://doi.org/10.1007/s10933-009-9324-9
508 Gomez, C., Drost, A.P.A., Roger, J.M., 2015. Analysis of the uncertainties affecting predictions
509 of clay contents from VNIR/SWIR hyperspectral data. Remote Sens. Environ. 156, 58–
510 70. https://doi.org/10.1016/j.rse.2014.09.032
511 Heiri, O., Lotter, A.F., Lemcke, G., 2001. Loss on ignition as a method for estimating organic

512 and carbonate content in sediments: reproducibility and comparability of results. J.
513 Paleolimnol. 25, 101–110. https://doi.org/10.1023/A:1008119611481
514 Jenny, J.-P., 2013. Réponses des grands lacs périalpins aux pressions anthropiques et
515 climatiques récentes : reconstitutions spatio-temporelles à partir d'archives
516 sédimentaires. HAL. Université Grenoble Alpes.
517 Jenny, J.-P., Wilhelm, B., Arnaud, F., Sabatier, P., Giguet Covex, C., Mélo, A., Fanget, B.,
518 Malet, E., Ployon, E., Perga, M.E., 2014. A 4D sedimentological approach to
519 reconstructing the flood frequency and intensity of the Rhône River (Lake Bourget, NW
520 European Alps). J. Paleolimnol. 51, 469–483. https://doi.org/10.1007/s10933-014-9768-
521 4
522 Kuang, B., Tekin, Y., Mouazen, A.M., 2015. Comparison between artificial neural network and
523 partial least squares for on-line visible and near infrared spectroscopy measurement of
524 soil organic carbon, pH and clay content. Soil Tillage Res. 146, 243–252.
525 https://doi.org/10.1016/j.still.2014.11.002
526 Leach, C.J., Wagner, T., Jones, M., Juggins, S., Stevenson, A.C., 2008. Rapid determination of
527 total organic carbon concentration in marine sediments using Fourier transform near-
528 infrared spectroscopy (FT-NIRS). Org. Geochem. 39, 910–914.
529 https://doi.org/10.1016/j.orggeochem.2008.04.012
530 Leardi, R., Lupiáñez, A., 1998. Genetic algorithms applied to feature selection in PLS
531 regression: how and when to use them. Chemom. Intell. Lab. Syst. 41, 195–207.
532 https://doi.org/10.1016/S0169-7439(98)00051-3
533 Li, H., Xu, Q., Liang, Y., 2014. libPLS: An Integrated Library for Partial Least Squares
534 Regression and Discriminant Analysis. PeerJ Prepr.
535 https://doi.org/10.7287/peerj.preprints.190v1

536 Li, S., Shi, Z., Chen, S., Ji, W., Zhou, L., Yu, W., Webster, R., 2015. In Situ Measurements of
537 Organic Carbon in Soil Profiles Using vis-NIR Spectroscopy on the Qinghai-Tibet Plateau.
538 Environ. Sci. Technol. 49, 4980–4987. https://doi.org/10.1021/es504272x
539 Lintern, A., Leahy, P.J., Zawadzki, A., Gadd, P., Heijnis, H., Jacobsen, G., Connor, S., Deletic,
540 A., McCarthy, D.T., 2016. Sediment cores as archives of historical changes in floodplain
541 lake hydrology. Sci. Total Environ. 544, 1008–1019.
542 https://doi.org/10.1016/j.scitotenv.2015.11.153
543 Mark, H.L., Tunnell, D., 1985. Qualitative near-infrared reflectance analysis using
544 Mahalanobis distances. Anal. Chem. 57, 1449–1456.
545 https://doi.org/10.1021/ac00284a061
546 Nawar, S., Mouazen, A.M., 2017. Predictive performance of mobile vis-near infrared
547 spectroscopy for key soil properties at different geographical scales by using spiking and
548 data mining techniques. CATENA 151, 118–129.
549 https://doi.org/10.1016/j.catena.2016.12.014
550 Peng, X., Shi, T., Song, A., Chen, Y., Gao, W., 2014. Estimating soil organic carbon using
551 VIS/NIR spectroscopy with SVMR and SPA methods. Remote Sens. 6, 2699–2717.
552 https://doi.org/10.3390/rs6042699
553 Perga, M.-E., Frossard, V., Jenny, J.-P., Alric, B., Arnaud, F., Berthon, V., Black, J.L., Domaizon,
554 I., Giguet-Covex, C., Kirkham, A., Magny, M., Manca, M., Marchetto, A., Millet, L.,
555 Paillès, C., Pignol, C., Poulenard, J., Reyss, J.-L., Rimet, F., Sabatier, P., Savichtcheva, O.,
556 Sylvestre, F., Verneaux, V., 2015. High-resolution paleolimnology opens new
557 management perspectives for lakes adaptation to climate warming. Front. Ecol. Evol. 3,
558 72. https://doi.org/10.3389/fevo.2015.00072
559 Rosén, P., Vogel, H., Cunningham, L., Hahn, A., Hausmann, S., Pienitz, R., Zolitschka, B.,

560 Wagner, B., Persson, P., 2011. Universally Applicable Model for the Quantitative
561 Determination of Lake Sediment Composition Using Fourier Transform Infrared
562 Spectroscopy. Environ. Sci. Technol. 45, 8858–8865.
563 Tetko, I. V., Livingstone, D.J., Luik, A.I., 1995. Neural network studies. 1. Comparison of
564 overfitting and overtraining. J. Chem. Inf. Model. 35, 826–833.
565 https://doi.org/10.1021/ci00027a006
566 Van Exem, A., Debret, M., Copard, Y., Vannière, B., Sabatier, P., Marcotte, S., Laignel, B.,
567 Reyss, J.-L., Desmet, M., 2018. Hyperspectral core logging for fire reconstruction
568 studies. J. Paleolimnol. 59, 297–308. https://doi.org/10.1007/s10933-017-0009-5
569 Vidal, M., Amigo, J.M., 2012. Pre-processing of hyperspectral images. Essential steps before
570 image analysis. Chemom. Intell. Lab. Syst. 117, 138–148.
571 https://doi.org/10.1016/J.CHEMOLAB.2012.05.009
572 Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse
573 reflectance spectra. Geoderma 158, 46–54.
574 https://doi.org/10.1016/j.geoderma.2009.12.025
575 Vohland, M., Ludwig, M., Thiele-Bruhn, S., Ludwig, B., 2014. Determination of soil properties
576 with visible to near- and mid-infrared spectroscopy: Effects of spectral variable
577 selection. Geoderma 223, 88–96. https://doi.org/10.1016/j.geoderma.2014.01.013
578 Wijewardane, N.K., Ge, Y., Wills, S., Loecke, T., 2016. Prediction of Soil Carbon in the
579 Conterminous United States: Visible and Near Infrared Reflectance Spectroscopy
580 Analysis of the Rapid Carbon Assessment Project. Soil Sci. Soc. Am. J. 80, 973–982.
581 https://doi.org/10.2136/sssaj2016.02.0052
582 Wold, S., Ruhe, A., Wold, H., Dunn, III, W.J., 1984. The Collinearity Problem in Linear
583 Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. SIAM J.

584 Sci. Stat. Comput. 5, 735–743. https://doi.org/10.1137/0905052

585 Zornoza, R., Guerrero, C., Mataix-Solera, J., Scow, K.M., Arcenegui, V., Mataix-Beneyto, J.,

586 2008. Near infrared spectroscopy for determination of various physical, chemical and

587 biochemical properties in Mediterranean soils. Soil Biol. Biochem. 40, 1923–1930.

588 https://doi.org/10.1016/j.soilbio.2008.04.003

589

590 ## Tables

591

592 *Table 1: Performance of complete and reduced PLSR models for the LOI550 prediction without removing water bands in blue*

593 *and after removing them in green.*

Wavelengths	Latent Variable	r calibration	r prediction	Model uncertainty (wt%)
121	8	0.99	0.80	1.40
21	7	0.96	0.88	1.20
75	7	0.99	0.92	1.02
27	**6**	**0.98**	**0.95**	**0.77**

594

595

Figure captions

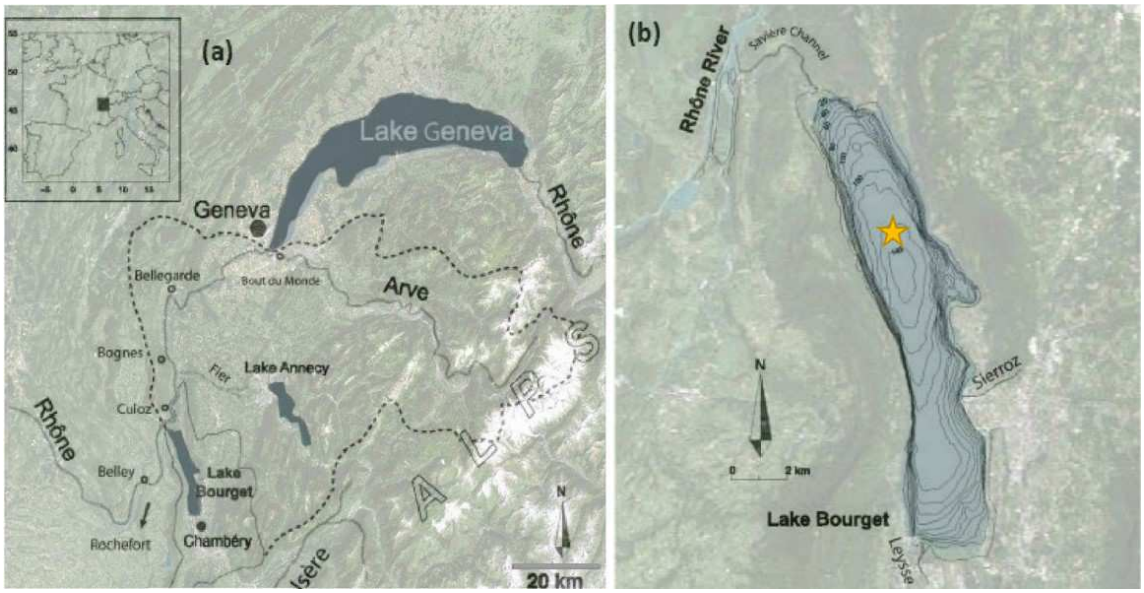


Figure 1: (a) Location and catchment area of Lake Bourget. (b) Bathymetry, tributaries and effluents of the lake.

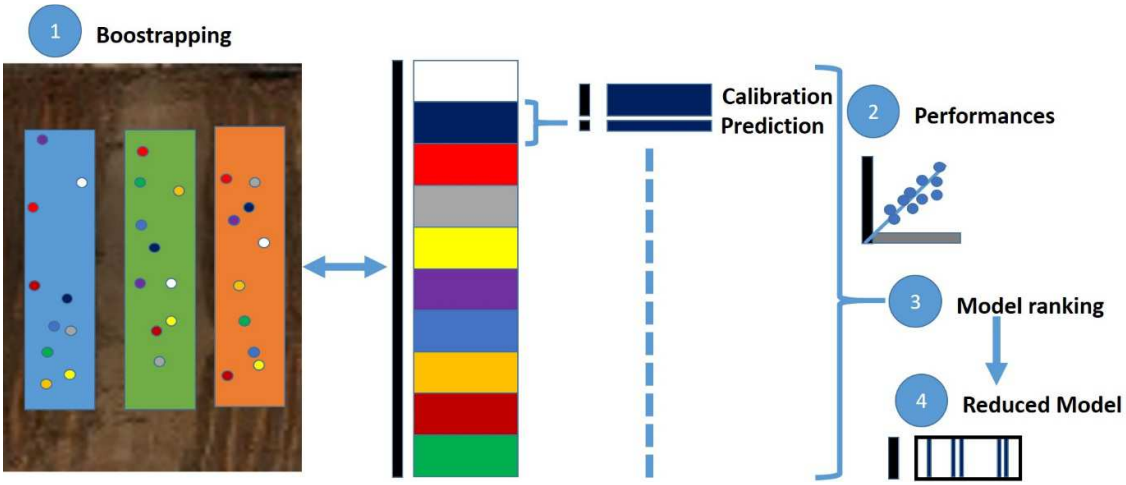
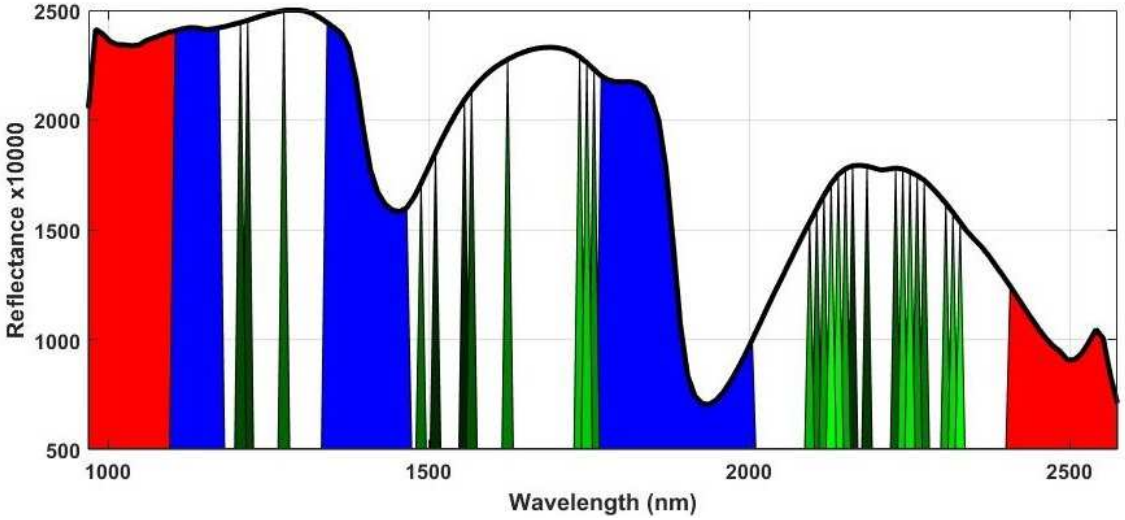


Figure 2: The four principal steps followed to create a model: (1) dataset creations with bootstrapping in the sampling area that are associated with the corresponding LOI550 value, (2) create models, estimate performances and estimate the regression coefficients, (3) rank the models based on their performances, and (4) reduce the number of wavelengths with 10 optimal models and create the reduced model.

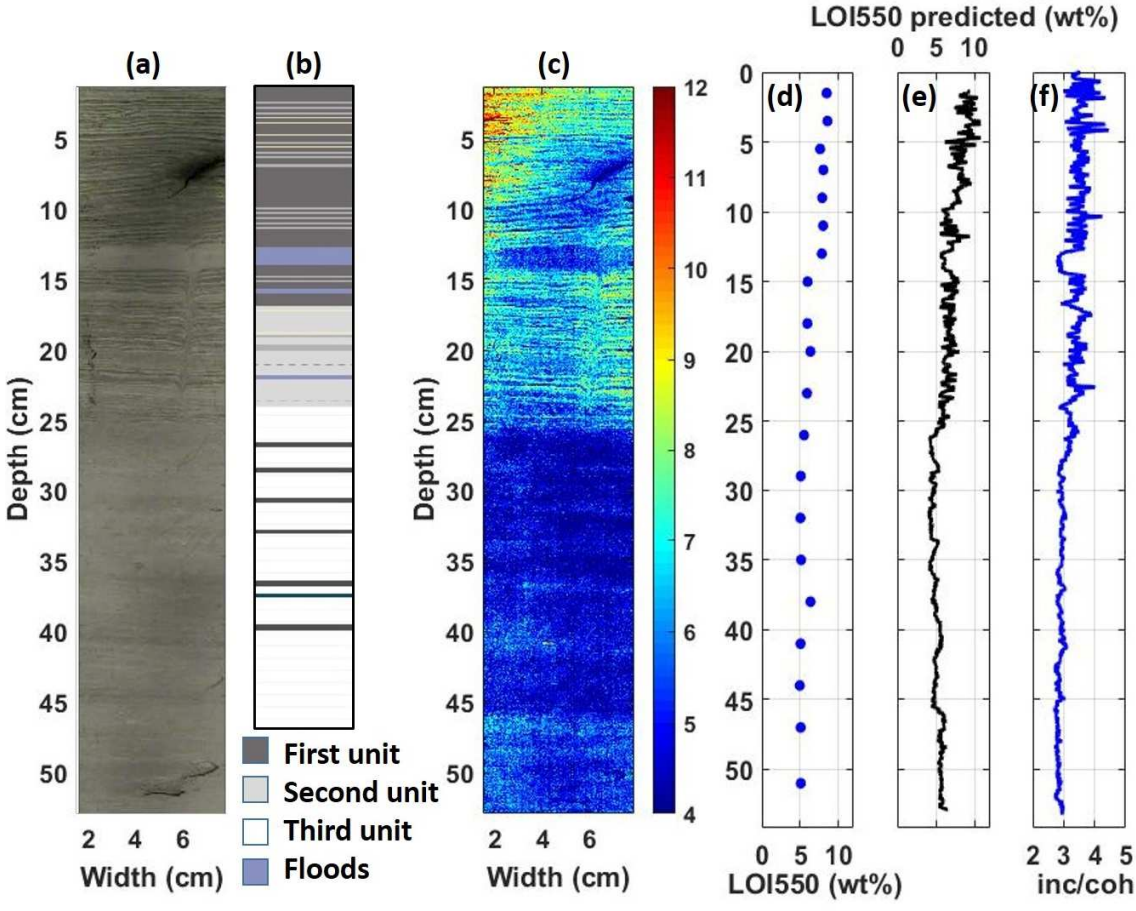


604

605 *Figure 3: 27 most correlated wavelengths to LOI550 are highlighted with a green gradient (light green: most), the water*

606 *bands are in blue, and the noise bands are in red on the mean spectrum in black.*

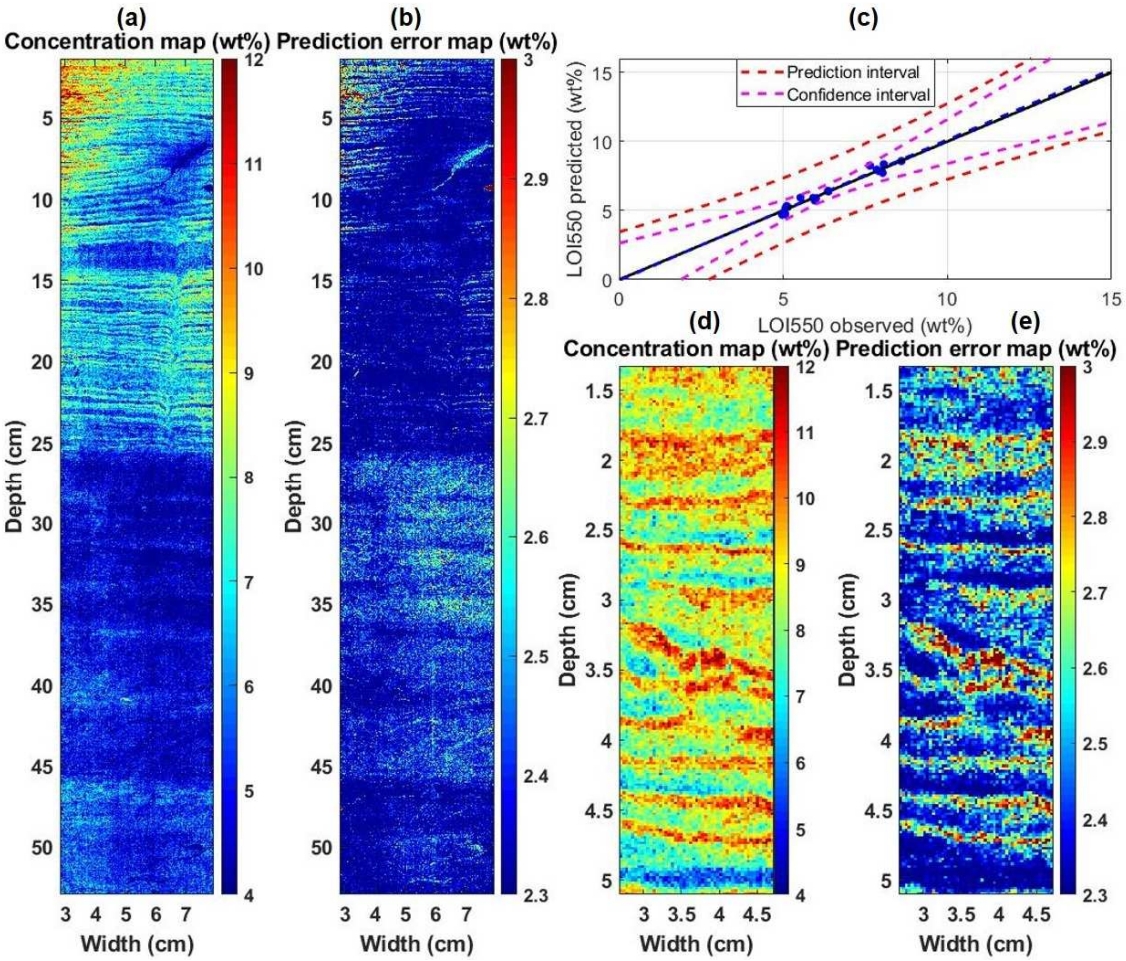
607



608

609 *Figure 4: Lake Bourget (a) RGB image, (b) lithology units, (c) LOI550 predicted concentration map with the reduce model, (d)*

610 *LOI550 reference values, (e) average LOI550 predicted values, and (f) inc/coh ratios.*



611

612 *Figure 5: (a) Concentration map of the core with its prediction error map (b). (c) Prediction and confidence intervals of the*

613 *LOI550 model. (d) Concentration map of a varved area and his prediction error map (e).*

614

615

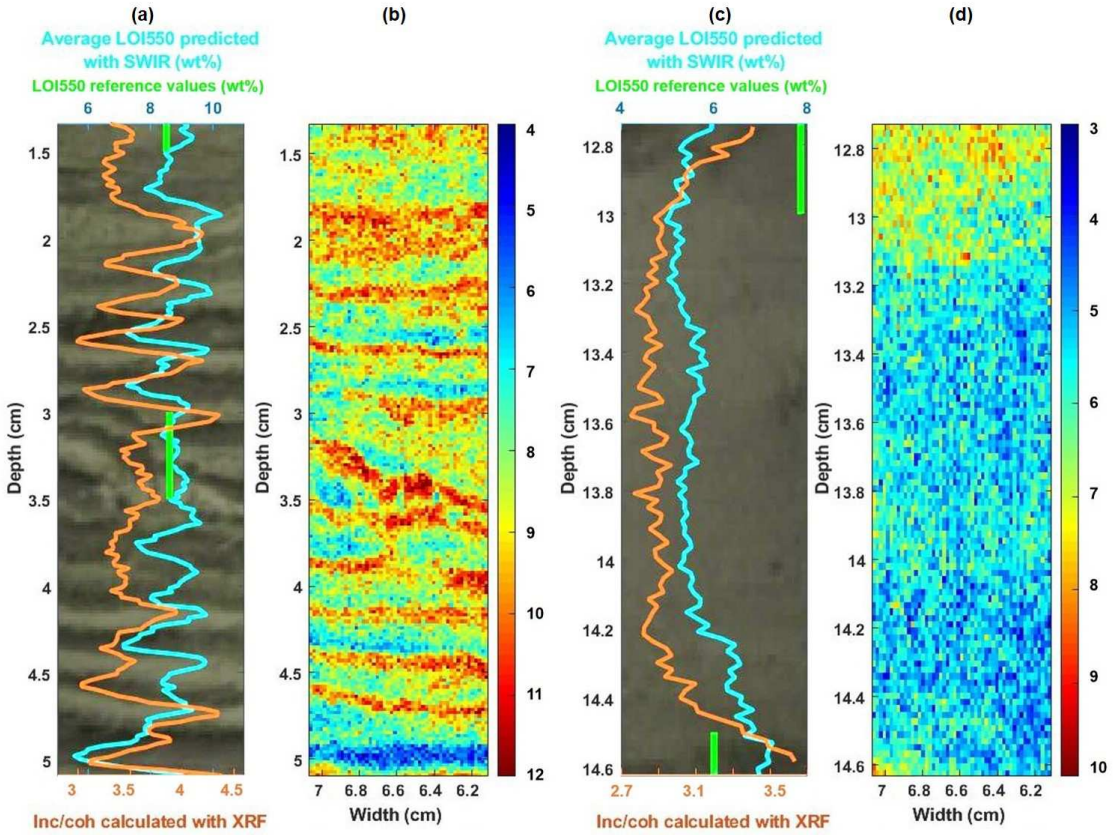


Figure 6: Zoom in for two areas on unit 1: varve (a-b) and flood (c-d). RGB image (a-c) with LOI550 reference values (green line) predicted by SWIR data (blue curve) and the inc/coh ratio (orange curve) as well as LOI550 concentration maps (b-d).