

# In Situ Measurements of Organic Carbon in Soil Profiles Using vis-NIR Spectroscopy on the Qinghai–Tibet Plateau

Shuo Li,<sup>†</sup> Zhou Shi,<sup>\*,†</sup> Songchao Chen,<sup>†</sup> Wenjun Ji,<sup>†</sup> Lianqing Zhou,<sup>†</sup> Wu Yu,<sup>‡</sup> and Richard Webster<sup>§</sup>

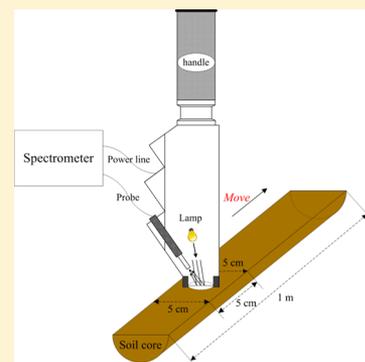
<sup>†</sup>Institute of Applied Remote Sensing and Information Technology, College of Environmental and Resource Sciences, Zhejiang University, 310058 Hangzhou, China

<sup>‡</sup>College of Resource and Environment, Tibet University, 860114 Nyingchi, China

<sup>§</sup>Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, United Kingdom

## S Supporting Information

**ABSTRACT:** We wish to estimate the amount of carbon (C) stored in the soil at high altitudes, for which there is little information. Collecting and transporting large numbers of soil samples from such terrain are difficult, and we have therefore evaluated the feasibility of scanning with visible near-infrared (vis-NIR) spectroscopy in situ for the rapid measurement of the soil in the field. We took 28 cores ( $\approx 1$  m depth and 5 cm diameter) of soil at altitudes from 2900 to 4500 m in the Sygera Mountains on the Qinghai–Tibet Plateau, China. Spectra were acquired from fresh, vertical faces  $5 \times 5$  cm in area from the centers of the cores to give 413 spectra in all. The raw spectra were pretreated by several methods to remove noise, and statistical models were built to predict of the organic C in the samples from the spectra by partial least-squares regression (PLSR) and least-squares support vector machine (LS-SVM). The bootstrap was used to assess the uncertainty of the predictions by the several combinations of pretreatment and models. The predictions by LS-SVM from the field spectra, for which  $R^2 = 0.81$ , the root-mean-square error RMSE = 8.40, and the ratio of the interquartile distance RPIQ = 2.66, were comparable to the PLSR predictions from the laboratory spectra ( $R^2 = 0.85$ , RMSE = 7.28, RPIQ = 3.09). We conclude that vis-NIR scanning in situ in the field is a sufficiently accurate rapid means of estimating the concentration of organic C in soil profiles in this high region and perhaps elsewhere.



## INTRODUCTION

The soil is estimated to contain 2344 Gt of organic carbon (C). That is roughly 4 times as much as in the biotic pool of C (estimated to be 560 Gt).<sup>1</sup> If the soil were to lose much of its C as CO<sub>2</sub>, then there would be a significant increase of CO<sub>2</sub> in the atmosphere and subsequently on global temperatures.<sup>2</sup> The ecosystems at high altitudes and high latitudes are especially sensitive to climate change, and they are likely to lose C from the soil as the climate warms. We ought to be able to monitor changes in the C content of the soil in those regions.<sup>3–5</sup>

The Qinghai–Tibet Plateau is the highest and largest plateau on earth. Except for cultivation and in its lower parts, it has hardly been disturbed by humans. Its vast expanse,  $2.5 \times 10^6$  km<sup>2</sup>, in a cold and fairly humid climate means that its soil contains a great deal of C. It also plays a role in moderating climate changes both in Asia and globally.<sup>6</sup> We know rather little about the soil's organic matter and its C content in the region, mainly because of its remoteness and the difficulties and expense of collecting soil material in that kind of terrain and transporting it to the laboratory.

In the last couple of decades, soil scientists have found that they can estimate the concentration of organic C in soil from reflectance in the visible (400–700 nm) and near-infrared (700–2500 nm) ranges (vis-NIR) of the electromagnetic spectrum. Spectroscopy in the vis-NIR range is rapid and has been found to be accurate enough in practice and cost-

effective;<sup>7–11</sup> it is a viable alternative to conventional chemical analysis in the laboratory. One reason is that the electronic transitions of atoms and overtones and combinations of chemical bonds in the organic components cause absorption of radiation at characteristic wavelengths or in characteristic wavebands in that range.<sup>10–12</sup>

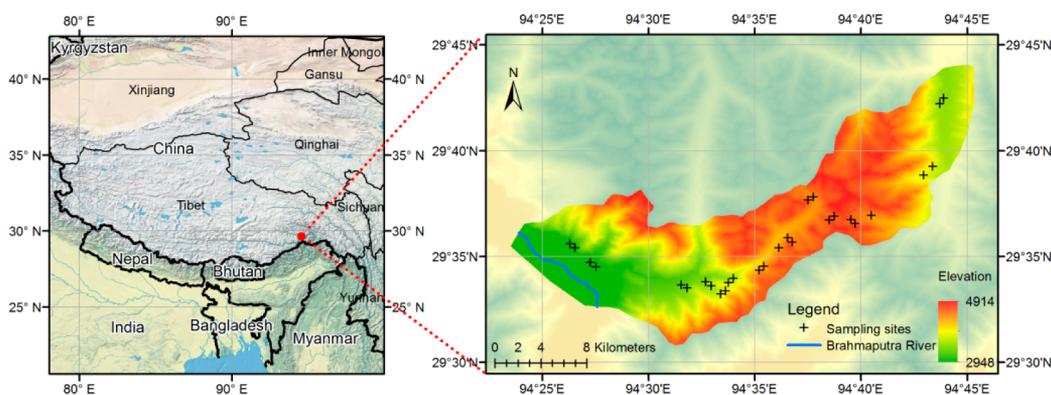
Measuring the spectra in the laboratory still involves the difficulties of transporting soil from the field and the tedium of drying, grinding, and sieving it. If one could measure the reflectance directly in situ with portable equipment, then much of the tedium and expense would be avoided. How reliable such measurements might be is a matter currently being investigated; we do not know in general how factors such as the soil's moisture content, its structure, color or texture and the ambient light affect the results. Variation in any of these soil properties might affect the accuracy of predictions from the spectra. Nevertheless, promising results have been reported in the literature, especially for the determination of organic C.<sup>13,14</sup> Most investigations on vis-NIR spectra in the field have been on topsoil, though a few have involved measurements down profiles.<sup>15–17</sup>

Received: August 31, 2014

Revised: February 5, 2015

Accepted: April 1, 2015

Published: April 1, 2015



**Figure 1.** Maps showing the location of the region sampled, and the positions of the sampling sites on the Sygera Mountains in the southeast of the Qinghai–Tibet Plateau.

**Table 1.** Statistics of Concentrations Organic Carbon (in  $\text{g kg}^{-1}$ ) Measured by Loss on Ignition in the Laboratory<sup>a</sup>

data set	N	min.	max.	mean	std dev.	CV (%)	skewness
calibration set	309	0.42	156.70	16.44	19.17	117	1.92
validation set	104	0.39	102.14	16.56	18.94	114	2.47
all data	413	0.39	156.70	16.48	19.09	116	2.33

<sup>a</sup>Note: N = sample size, min. = minimum, max. = maximum, std dev. = standard deviation, CV = coefficient of variation

Our aim in this study we report below was to determine the feasibility of scanning vis-NIR spectroscopy in the field to estimate the organic C content of soil to a depth of 1 m on the Qinghai–Tibet Plateau. We did so by (i) measuring the spectra reflected from cores of soil in the field and comparing the predictions from them with results from spectra measured on the same soil, air-dried, in the laboratory, (ii) evaluating the prediction accuracy of models of C content in soil at several depths down to 1 m from vis-NIR spectroscopy, and (iii) assessing the uncertainty of predictions using the bootstrap and validation against known values in a separate subset of samples of the soil.

## MATERIALS AND METHODS

**Soil Sampling.** The region in which we studied the soil is the Sygera Mountains (29.36° N, 94.36° E) in the southeast of the Qinghai–Tibet Plateau, China (Figure 1a). The mean annual precipitation there is 676 mm, most of which falls as rain in summer as the south Asian monsoon approaches the mountains along the valley of the Brahmaputra River. The average annual temperature is 15.8 °C. Arable land lies below 3500 m, with maize and barley as the main crops. The vegetation of the mountains is strongly zoned by altitude with coniferous forest between 3300 and 4300 m and alpine shrub and meadow above 4300 m.<sup>18</sup> The dominant species in the forest is Smith fir (*Abies georgei* var. *smithii*), locally accompanied by *Picea likiangensis* var. *linzhiensis*, *Quercus aquifolioides* and *Sabina saltuaria*. The alpine shrub and meadow zone is dominated by *Rhododendron* spp., *Kobresia pygmaea*, and *Kobresia prainii*.<sup>19</sup> In the World Reference Base (WRB), the soil belongs to the following classes: Cambisols, Luvisols, Phaeozems, and Umbrisols. The clay content ranges from 4.9% to 18.8% with an average of 10.3%. The average value of pH is 5.7 in a range of 4.2 to 7.3.

In August 2013, the soil of the region was sampled from 28 profiles from the foot of the mountains at about 2900 m altitude to the ridge at 4500 m along a main road (Figure 1b). At each site, a cylindrical core (5 cm in diameter) of soil was

taken down to 1 m, or to the limit of penetration if that was less.

**In Situ Spectral Measurement.** Spectra reflected from the fresh soil cores were measured in the field spectra with an ASD FieldSpec Pro FR spectrometer (Analytical Spectral Devices Inc., Boulder, CO, USA), which has a high-intensity contact probe attached to an external fiber-optic cable. The instrument has a spectral range of 350 to 2500 nm and a resolution of 3 nm at 700 and 10 nm at 1400 and 2100 nm. We used it with an independent light source to minimize measurement errors caused by stray light. We used also a Spectralon panel (Labsphere Inc., North Sutton, NH, USA) with 99% reflectance as a white reference before each measurement.

Each core was cut in half vertically, and the contact probe was placed on one of the plane surfaces to measure its spectral reflectance. At each position of the probe, the instrument made 10 internal scans for a satisfactory signal-to-noise ratio. Two or three spectra were recorded at randomly chosen positions in a 5 × 5 cm square at each depth, and averaged into one spectrum for that square. Measurements were made from the top to the base of each core to provide spectra for a maximum of 20 depths. In all, 413 spectra were recorded.

**Laboratory Spectral Measurement.** After the spectra had been obtained in the field, the cores of soil were sealed with plastic liners, labeled, and transported to the laboratory. There they were cut into sections to match the 5 × 5 cm faces sensed in the field. All the cut sections were then air-dried, ground, and sieved to pass 2 mm. A sample of each was placed in a Petri dish 10 cm in diameter and 1.5 cm deep. Its surface was flattened, and the spectrum from it was measured in the same way as in the field with the same instrument, though with the greater control that is possible in the laboratory. Again, 413 spectra were recorded.

**Measurement of Soil Organic Carbon.** The organic C content of the samples from the sections was measured by dry combustion at 1100 °C with a multi N/C 3100 instrument (Analytik Jena AG, Germany). Any inorganic C was removed before the determination by acidification of the samples with

0.1 mol L<sup>-1</sup> hydrochloric acid. The samples spanned a wide range of carbon content from 0.39 to 156.7 g kg<sup>-1</sup> soil; see Table 1.

The 413 samples were ranked in ascending order of their concentrations of organic C and separated into two groups: three-quarters of which were used to build prediction models (i.e., for calibration) and the remaining quarter for validation.

**Preprocessing of Spectra.** Some preprocessing of soil spectra was desirable to harmonize the spectra and diminish the differences arising from the differences in the optical environments. All the spectra were resampled to a uniform resolution of 1 nm. That gave us 2151 reflectance for each spectrum. The ends of the spectra in this range are noisy, and so we removed them to leave only values in the range 400–2450 nm. To reduce noise further or to enhance the signals, or both, we tested several spectroscopic preprocessing algorithms for both the laboratory and in situ spectra. The algorithms included Savitzky–Golay smoothing (SG),<sup>20</sup> multiplicative scatter correction (MSC),<sup>21</sup> standard normal variate transformation (SNV),<sup>22</sup> and the first derivative technology (1st D),<sup>23</sup> all of which are widely used. All preprocessing steps were programmed and implemented in MatLab version 7 (The MathWorks Inc., Natick, MA, USA).

**Comparison of Calibration Algorithms.** Estimating the content of organic C in the soil from the spectra requires calibration of the spectra against accurately measured C. Several fairly standard statistical techniques are used for this, and we wanted to know which might be best in our circumstances. We compared two, one partial least-squares regression (PLSR) and the other the least-squares support vector machine (LS-SVM).

**Partial Least-Squares Regression (PLSR).** Partial least-squares regression (PLSR) is one of the most popular multivariate techniques for spectral calibration and prediction.<sup>24</sup> It is closely related to principal component regression (PCR). Both methods reduce the number of variables in the prediction set. In PCR, the predictor variables are converted to principal components that are then ranked in order of the variance for which they account, regardless of their covariances with the predictands. One simply decides how many to retain and then uses them as the predictors in the subsequent regression. In PLSR, this arbitrary decision is largely avoided. Instead, the technique reduces the number of variables in the predictor set by selecting successive orthogonal factors from the variance–covariance matrix in a way that maximizes the covariance between the predictors and response variable or variables (there may be more than one variable being predicted). There is a risk of over-fitting or under-fitting, and to avoid this, we used the leave-one-out cross-validation method to find the optimal number of calibration factors that minimize the prediction error variance. For details, see Martens and Næs.<sup>23</sup>

**Least Squares Support Vector Machine (LS-SVM).** A support vector machine (SVM) is a kernel-based learning algorithm that is widely used for pattern classification and regression.<sup>9</sup> Such methods implicitly map input data in a high-dimensional feature space, a special type of hyperplane defined by a kernel function, in which a regression model is built. The least-squares SVM, the LS-SVM, is an optimizing algorithm based on the standard SVM. It uses a squared loss function instead of the e-insensitive loss function of the standard SVM. In doing so it avoids the complex calculations of the SVM, and the multivariate calibration problem can be solved relatively quickly. To avoid over-fitting, the structural risk was used. Suykens and Vandewalle provide details.<sup>25</sup>

**Uncertainty Analysis.** To assess the uncertainty of the model predictions, we used the bagging method in the bootstrap.<sup>26</sup> The method proceeds by repeated random sampling with replacement from the original  $N$  units of the data to obtain  $B$  bootstrap samples. Each bootstrap sample is of the same size as the original set of data but contains replicates of some units, while others are absent. In this study  $N = 309$ , i.e., three-quarters of the 413 spectra, and  $B = 100$ . The bagging estimate provides a robust predictor; the estimates are calculated from the mean of the bootstrap, and their uncertainty is assessed from the 95% confidence intervals in the bootstrap. Briefly, these improvements result from the aggregation of the  $B$  different bootstrapped models, each of which provides unique information.<sup>27</sup>

**Assessment of Statistics.** The coefficients of determination ( $R^2$ ), the root-mean-square error (RMSE), and the residual estimation deviation (RPD) may all be used to assess the accuracy of calibration models. The first,  $R^2$ , is familiar enough. In our case, it is simply the square of the Pearson correlation coefficient between the predictions and the measurements. We express the relation in this form to match convention in the subject. The RMSE is given by

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2}$$

in which  $y_i$  is the observed value,  $\hat{y}_i$  is the predicted value, and  $N$  is the number of units in the summation. The RPD, the ratio of the standard deviation of  $y$  to the RMSE, is widely used in vis-NIR spectroscopy to assess the goodness of a calibration model. For strongly skewed data, however, Bellon-Maurel et al. recommend that the RPD be replaced by the ratio of the interquartile range of  $y$  to the RMSE:

$$\text{RPIQ} = (Q_3 - Q_1) / \text{RMSE}$$

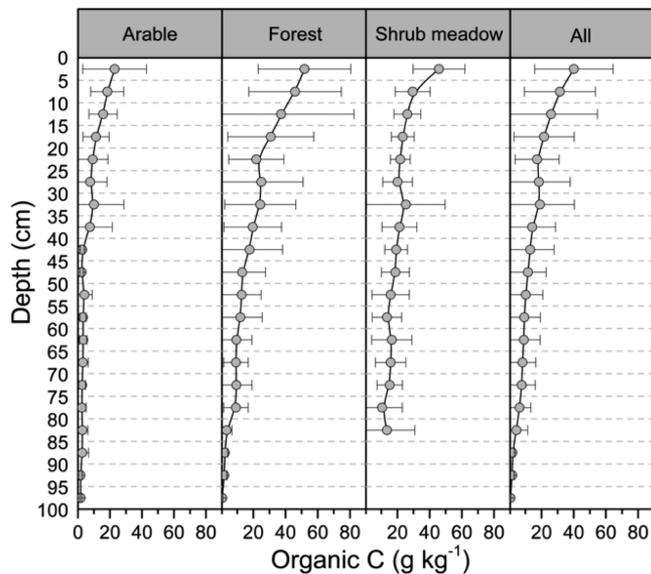
in which  $Q_3$  and  $Q_1$  are the upper and lower quartiles respectively of the data.<sup>28</sup> We have therefore used this quantity to compare our models. Generally, a model that performs well would have large values of  $R^2$  and RPIQ and a small RMSE, though they are not necessarily monotonically related.

## RESULTS

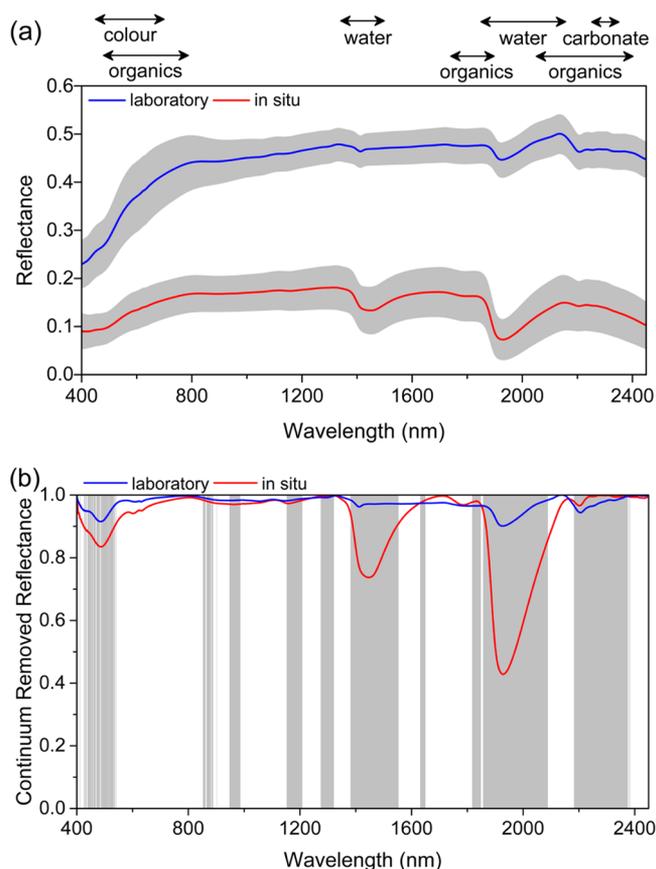
**Organic Carbon Content.** Figure 2 shows the average distributions of organic C in the soil profiles for the three types of land use. The concentration of organic C gradually decreases with increasing depth. At all depths, the mean content of C was less in arable soil than that under forest and meadow. The mean concentration of organic C in the uppermost 40 cm of the arable soil is substantially less than that in the forest soil. It also decreases more markedly with increasing depth. In the shrub meadow soil, the decrease in organic C with increasing depth was less marked below 10 cm.

Table 1 summarizes the measurements of organic C content of the soil in all 413 samples and in the calibration and validation sets separately. The C content for the whole set of data varied from 0.39 to 156.70 g kg<sup>-1</sup>, with an average of 16.48 g kg<sup>-1</sup>. The means in the three sets are similar as are the standard deviations, and we can therefore have confidence in the validation.

**Comparison of Field and Laboratory Spectra.** Figure 3a shows the average reflectance of laboratory and field measurements of all 413 samples and their standard deviations. The



**Figure 2.** Vertical distribution of organic C in soil profiles. The circles show the mean values and the horizontal bars are standard deviations.



**Figure 3.** Average laboratory and field vis-NIR spectra; (a) the reflectance spectra shown as curves with their standard deviations show as shaded gray bands about the means; and (b) the continuum removed spectra with wavelength at which there are significant differences at the 0.01 level between the laboratory and field spectra shown by the gray columns. The main positions of spectral absorption by water and chemical species are highlighted.

positions of constituents causing spectral absorption are also shown. Because the bands are broad and overlap, the spectra

are difficult to interpret. Nevertheless, this vis-NIR region contains useful information related to the organic C in the soil, due to the various chemical bonds such as C–C, C=C, C–H, C–N, and O–H.<sup>29,30</sup> The effect of moisture is strong; the water in the fresh soil strongly absorbs throughout the vis-NIR range and has caused the evident shift in the spectra from the lower curve for the moist soil in the field to the upper curve for the dry soil measured in the laboratory.

Figure 3b shows the average continuum-removed reflectance. The vertical gray bands show the wavelengths at which there were statistically significant differences between the laboratory and field spectra. The differences were assessed by Student's *t* test at probability 0.01 of the null hypothesis. The differences in absorption near 450 nm are probably related to color. The organic materials seem to absorb differentially near 2250 nm. The most obvious differences between the two types of spectra are in the two primary water absorption regions in the NIR region of the spectra, one near 1450 nm and the other near 1950 nm.

**Comparison of PLSR and LS-SVM Predictions.** The accuracy of a model depends strongly on the way the raw spectra are pretreated (Table 2). For laboratory spectra analyzed by PLSR, the best calibration was achieved with a combination of first derivatives and SG pretreatment ( $RMSE_c = 5.54 \text{ g kg}^{-1}$  and  $R_c^2 = 0.91$ ), which also produced the best validation results. For the field spectra, the best calibration was also achieved with that combination for predicting C contents ( $RMSE_c = 8.11 \text{ g kg}^{-1}$  and  $R_c^2 = 0.82$ ) and LS-SVM ( $RMSE_c = 4.88 \text{ g kg}^{-1}$  and  $R_c^2 = 0.94$ ).

The best calibrated PLSR and LS-SVM models were applied to predict the C contents of the data in validation sets. The results confirmed that these were the best methods. Comparisons of the prediction results of the three models with 1st derivatives plus SG pretreatment can be summarized as follows. The laboratory PLSR model ( $RMSE_p = 7.28$ ,  $R_p^2 = 0.85$ ,  $RPIQ = 3.09$ ) and field LS-SVM ( $RMSE_p = 8.40$ ,  $R_p^2 = 0.81$ ,  $RPIQ = 2.66$ ) performed similarly and better than the field PLSR model ( $RMSE_p = 9.32$ ,  $R_p^2 = 0.76$ ,  $RPIQ = 2.39$ ).

**Uncertainties of Predicting Models.** Figure 4a shows the prediction results of bagging-LS-SVM with the field spectra ( $RMSE_p = 8.50$ ,  $R_p^2 = 0.80$ ,  $RPIQ = 2.63$ ), and the uncertainties of the estimates, expressed as the standardized range of the 95% confidence intervals, were generally small where observations were dense (Figure 4b). Bagging can stabilize the LS-SVM model.

Figure 5 shows the prediction statistics for the field spectra of the whole profile obtained by leave-one-out cross-validation of the LS-SVM model. It shows that in general the correlation between the prediction and actuality gradually weakens ( $R^2$  decreased from a maximum of 0.85 to  $\approx 0.4$ ) with increasing depth, presumably because as the amount of organic matter decreases so the spectral reflectance from the soil becomes less effective as a predictor. The RMSE decreased sharply from  $\approx 14 \text{ g kg}^{-1}$  in the top 20 cm of soil, where the concentration of C is  $30.88 \text{ g kg}^{-1}$ , to  $\approx 6 \text{ g kg}^{-1}$  below 20 cm where the concentration is  $11.25 \text{ g kg}^{-1}$ . The RMSE eventually stabilized between 70 and 100 cm where the mean concentration of C is only  $5.39 \text{ g kg}^{-1}$ . The mean value of  $R^2 \approx 0.65$  and that of the  $RMSE \approx 7 \text{ g kg}^{-1}$  in the subsoil are similar to those reported by Kuang and Mouazen when they estimated the organic C content of fresh soil on European farms from vis-NIR spectra.<sup>31</sup>

**Table 2. Comparison of Prediction Accuracy for Laboratory PLSR, Field PLSR, and Field LS-SVM for the Calibration and Validation and Various Pretreatments<sup>a</sup>**

	model	pretreatment	calibration set		validation set		
			RMSE <sub>c</sub>	R <sub>c</sub> <sup>2</sup>	RMSE <sub>p</sub>	R <sub>p</sub> <sup>2</sup>	RPIQ
laboratory spectra	PLSR	none	7.32	0.85	7.68	0.83	2.93
		SG	6.52	0.88	7.72	0.83	2.91
		SNV + SG	7.91	0.83	7.78	0.83	2.89
		MSC + SG	7.73	0.84	8.04	0.82	2.79
		<b>1stD + SG</b>	<b>5.54</b>	<b>0.91</b>	<b>7.28</b>	<b>0.85</b>	<b>3.09</b>
field spectra	PLSR	none	9.85	0.74	10.14	0.71	2.20
		SG	10.12	0.72	10.11	0.71	2.21
		SNV + SG	10.09	0.72	9.53	0.74	2.34
		MSC + SG	9.44	0.76	9.67	0.74	2.31
		<b>1stD + SG</b>	<b>8.11</b>	<b>0.82</b>	<b>9.32</b>	<b>0.76</b>	<b>2.39</b>
	LS-SVM	none	5.34	0.92	8.62	0.79	2.59
		SG	6.03	0.90	8.45	0.80	2.64
		SNV + SG	5.01	0.93	8.79	0.78	2.54
		MSC + SG	4.93	0.94	8.95	0.78	2.49
		<b>1stD + SG</b>	<b>4.88</b>	<b>0.94</b>	<b>8.40</b>	<b>0.81</b>	<b>2.66</b>

<sup>a</sup>Note: The RMSEs are in g kg<sup>-1</sup>. R<sub>c</sub><sup>2</sup> and R<sub>p</sub><sup>2</sup> are the coefficients of determination and simply the squares of the Pearson correlation coefficients between the measured and predicted values for calibration and prediction, respectively; RMSE<sub>c</sub> and RMSE<sub>p</sub> are the root-mean-square errors of calibration and prediction, respectively; RPIQ is the ratio of performance to interquartile distance; SG is Savitzky–Golay smoothing; MSC is the multiplicative scatter correction; SNV is the standard normal variate transformation; 1stD is first derivatives of reflectance. The results of selected models are emphasized in bold.

## DISCUSSION

There are numerous other examples in the literature of strong correlations between predictions of the soil's organic C or organic matter content (equivalent to  $\approx 1.72 \times$  the organic C content) from spectra recorded in the field and laboratory measurements (Figure 6). As the size of samples,  $N$ , increases,  $R^2$  approaches 0.8. However, it varies widely between samples of size less than 1000. The reasons for this are not clear. They could include differences in modeling strategy (e.g., spectral range, preprocessing methods and calibration algorithms), local diversity (e.g., soil type, texture, structure, mineralogy, and land use type), and other variables such as lighting and laboratory equipment. All can strongly influence the reflectance spectra, and they will therefore affect the optimal model parameters. If we are to build libraries of soil spectra to cover large areas from numerous local surveys, we shall need to standardize the technology to take full advantage of it with a robust model.<sup>32,33</sup> The mean value of  $R^2$  from the field measurements ( $R^2 = 0.82$ ) is somewhat less than that of laboratory measurements ( $R^2 = 0.84$ ). We obtained similar values of  $R^2$  from both field (0.81) and laboratory (0.85) measurements. The stronger correlations in the laboratory were matched by greater accuracy as assessed by the RMSE: RMSE = 7.28 in the laboratory compared with RMSE = 8.40 in the field (Table 2). The accuracy of our predictions are better than those achieved by of Morgan et al. in a similar investigation ( $R^2 = 0.64$ , RMSE = 5.4, and RPD = 1.45).<sup>17</sup> This might be because we had only three kinds of land use. In other circumstances where there is a greater variety our technique might be less accurate. Testing this idea was beyond the scope of this project, however, and we plan to investigate in a future study.

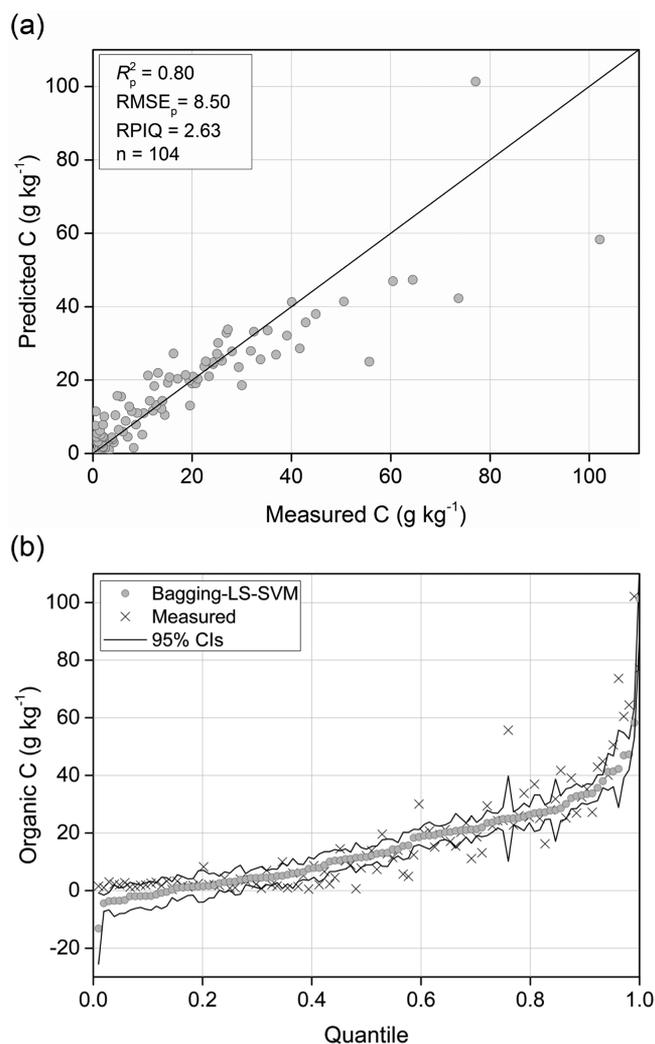
The differences are matched by the small but perceptible differences in the widths of the gray bands about the mean values in Figure 3a: the bands about the upper curve for the laboratory spectra are narrower than those for the lower curve for the field spectra. This is probably because we could standardize conditions in the laboratory whereas in the field we

had to contend with fluctuating weather, light and moisture and with variation in the soil's structure. Given such variation in the outdoor environment, we are well pleased with the results, and we think that the accuracy achieved in the field is acceptable. Notice, however, the best prediction in the field was by LS-SVM not PLSR; the latter was somewhat worse. The PLSR might suffer from the same fluctuations producing noise that could not be removed by the various pretreatment procedures.

The spectra measured in the laboratory were significantly different from those obtained in the field (Figure 3a). The overall decrease in reflectance is caused by water in the fresh moist soil.<sup>34</sup> The water replaces the air within soil pores, increases the forward scattering of light and increases the absorption at all wavelengths.<sup>35</sup> Viscarra Rossel et al. also compared spectra measured in the field with ones measured in the laboratory, and when they had eliminated the effects of water absorption they found no significant difference between the two in predictions of soil constituents from them,<sup>36</sup> which accords with our results shown in Figure 3b.

As Figure 6 shows, however, some others investigators have found relatively weak correlations. One reason is that the relation between water content and reflectance is not linear,<sup>37,38</sup> and so water in the fresh soil is a serious impediment to accurate prediction. Although we had to contend with variable soil water content in the field, by using the nonlinear LS-SVM algorithm, we could improve the accuracy of our predictions to the point at which it was comparable to the accuracy achieved in the laboratory with PLSR algorithm. Several other investigators have found the same.<sup>9,30,39</sup> Nevertheless, there is room for improvement.

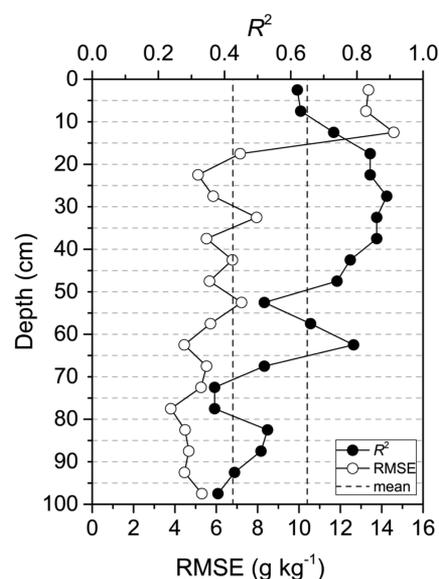
Final prediction results are bound to contain uncertainties, including parameter setting, measurement error, model input, and environmental conditions.<sup>40</sup> By standardizing our sampling protocol, using universal chemical analysis, preprocessing the spectra, and assessing statistics, we could diminish the sources of uncertainty to the best of our abilities. The results of our uncertainty analysis by bagging-LS-SVM (Figure 4b) showed



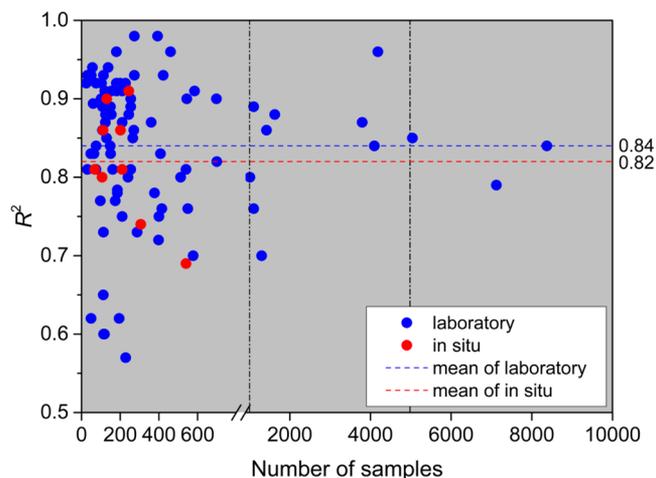
**Figure 4.** (a) Predicted C plotted against measured C in the validation set, showing the departures from the 1:1 line and for which the RMSE = 8.50; (b) cumulative distribution of the bagging predictions of C by LS-SVM from the field spectra (black discs) and the corresponding measured C (crosses); uncertainty is shown by their 95% confidence intervals (black lines).

that LS-SVM could create a robust model under field conditions. It seems that the bagging algorithm itself combined with the LS-SVM leads to models that reduce the prediction variances; this is an advantage.<sup>27</sup> Hence, we can recommend the combination of in situ vis-NIR spectroscopy in the field with LS-SVM for the rapid and cost-effective estimation of the soil's organic carbon content.

Despite our success, we should like to see more studies made with vis-NIR spectroscopy in the field for estimating concentrations of organic C in the soil, especially at high altitudes. One of our aims is to create a spectral library for the whole of the Qinghai–Tibet Plateau at an affordable price. The area is many times larger than the Sygera Mountains, and almost certainly more variable. We shall need for that a rigorous protocol for measuring spectra in the field and a robust standard combination of pretreatment, statistical modeling and uncertainty analysis of the spectra obtained. We might also refine technique by removing the effects of moisture on the vis-NIR spectra by, for example, the external parameter orthogonalization (EPO) and the direct standardization (DS)



**Figure 5.** Cross-validation results of the LS-SVM model with from field spectra from all 413 soil samples.



**Figure 6.** Coefficients of determination ( $R^2$ ) for organic C in soil estimated from vis-NIR spectra plotted vs sample size as reported in the literature.

algorithms.<sup>41,42</sup> We believe that the investigation we report above is a promising start, and we intend to build on it.

## ■ ASSOCIATED CONTENT

### 📄 Supporting Information

References of Figure 6. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## ■ AUTHOR INFORMATION

### Corresponding Author

\*Z. Shi. Tel: +86 571 88982831. Fax: +86 571 88982831. E-mail: [shizhou@zju.edu.cn](mailto:shizhou@zju.edu.cn).

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

We received funding for this research from the National Natural Science Foundation of China (No. 41271234), the

National High Technology Research and Development Program of China (No. 2013AA102301), and the Research Fund of State Key Laboratory of Soil and Sustainable Agriculture, Nanjing Institute of Soil Science, Chinese Academy of Science (No. Y412201430).

## REFERENCES

- (1) Jobbágy, E. G.; Jackson, R. B. The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecol. Appl.* **2000**, *10*, 423–436.
- (2) Davidson, E. A.; Janssens, I. A. Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature* **2006**, *440*, 165–173.
- (3) Yang, Y.; Fang, J.; Tang, Y.; Ji, C.; Zheng, C.; He, J.; Zhu, B. Storage, patterns and controls of soil organic carbon in the Tibetan grasslands. *Global Change Biol.* **2008**, *14*, 1592–1599.
- (4) Kato, T.; Tang, Y.; Gu, S.; Cui, X.; Hirota, M.; Du, M.; Li, Y.; Zhao, X.; Oikawa, T. Carbon dioxide exchange between the atmosphere and an alpine meadow ecosystem on the Qinghai–Tibetan Plateau, China. *Agr. Forest Meteorol.* **2004**, *124*, 121–134.
- (5) Zimov, S. A.; Schuur, E. A. G.; Chapin, F. S., III Permafrost and the global carbon budget. *Science* **2006**, *312*, 1612–1613.
- (6) Wang, G.; Qian, J.; Cheng, G.; Lai, Y. Soil organic carbon pool of grassland soils on the Qinghai–Tibetan Plateau and its global implication. *Sci. Total Environ.* **2002**, *291*, 207–217.
- (7) Ben-Dor, E.; Banin, A. Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. *Soil Sci. Soc. Am. J.* **1995**, *59*, 364–372.
- (8) Chang, C. W.; Laird, D. A.; Mausbach, M. J.; Hurburgh, C. R. Near-infrared reflectance spectroscopy—principal components regression analyses of soil properties. *Soil Sci. Soc. Am. J.* **2001**, *65*, 480–490.
- (9) Ji, W.; Shi, Z.; Huang, J.; Li, S. In situ measurement of some soil properties in paddy soils using vis-NIR spectroscopy. *PLoS One* **2014**, *9*, No. e105708.
- (10) Viscarra Rossel, R. A.; Walvoort, D. J. J.; McBratney, A. B.; Janik, L. J.; Skjemstad, J. O. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma* **2006**, *131*, 59–75.
- (11) Viscarra Rossel, R. A.; Webster, R. Predicting soil properties from the Australian soil visible–near infrared spectroscopic database. *Eur. J. Soil Sci.* **2012**, *63*, 848–860.
- (12) Stenberg, B.; Viscarra Rossel, R. A.; Mouazen, A. M. Chapter five-visible and near infrared spectroscopy in soil science. *Adv. Agron.* **2010**, *107*, 163–215.
- (13) Mouazen, A. M.; Maleki, M. R.; De Baerdemaeker, J.; Ramon, H. On-line measurement of some selected soil properties using a VIS–NIR sensor. *Soil Tillage Res.* **2007**, *93*, 13–27.
- (14) Nocita, M.; Kooistra, L.; Bachmann, M.; Müller, A.; Powell, M.; Weel, S. Predictions of soil surface and topsoil organic carbon content through the use of laboratory and field spectroscopy in the Albany Thicket Biome of Eastern Cape Province of South Africa. *Geoderma* **2011**, *167*, 295–302.
- (15) Kusumo, B. H.; Hedley, M. J.; Hedley, C. B.; Tuohy, M. P. Measuring carbon dynamics in field soils using soil spectral reflectance: Prediction of maize root density, soil organic carbon and nitrogen content. *Plant Soil* **2011**, *338*, 233–245.
- (16) Waiser, T. H.; Morgan, C. L. S.; Brown, D. J.; Hallmark, C. T. In situ characterization of soil clay content with visible near-infrared diffuse reflectance spectroscopy. *Soil Sci. Soc. Am. J.* **2007**, *71*, 389–396.
- (17) Morgan, C. L. S.; Waiser, T. H.; Brown, D. J.; Hallmark, C. T. Simulated in situ characterization of soil organic and inorganic carbon with visible near-infrared diffuse reflectance spectroscopy. *Geoderma* **2009**, *151*, 249–256.
- (18) Liang, E.; Wang, Y.; Xu, Y.; Liu, B.; Shao, X. Growth variation in *Abies georgei* var. *smithii* along altitudinal gradients in the Sygera Mountains, southeastern Tibetan Plateau. *Trees-Struct. Funct.* **2010**, *24*, 363–373.
- (19) Liang, E. Y.; Shao, X. M.; Xu, Y. Tree-ring evidence of recent abnormal warming on the southeast Tibetan Plateau. *Theor. Appl. Climatol.* **2009**, *98*, 9–18.
- (20) Savitzky, A.; Golay, M. J. E. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **1964**, *36*, 1627–1639.
- (21) Geladi, P.; MacDougall, D.; Martens, H. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Appl. Spectrosc.* **1985**, *39*, 491–500.
- (22) Barnes, R. J.; Dhanoa, M. S.; Lister, S. J. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc.* **1989**, *43*, 772–777.
- (23) Martens, H.; Næs, T. *Multivariate Calibration*; John Wiley & Sons: New York, 1989.
- (24) Wold, S.; Martens, H.; Wold, H. The multivariate calibration problem in chemistry solved by the PLS method. In *Matrix Pencils*; Kågström, B., Ruhe, A. H., Eds.; Springer: Berlin, 1983; pp 286–293.
- (25) Suykens, J. A. K.; Vandewalle, J. Least squares support vector machine classifiers. *Neural Process Lett.* **1999**, *9*, 293–300.
- (26) Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140.
- (27) Viscarra Rossel, R. A. Robust modelling of soil diffuse reflectance spectra by bagging-partial least squares regression. *J. Near Infrared Spec.* **2007**, *15*, 39–47.
- (28) Bellon-Maurel, V.; Fernandez-Ahumada, E.; Palagos, B.; Roger, J.; McBratney, A. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *Trend. Anal. Chem.* **2010**, *29*, 1073–1081.
- (29) Henderson, T. L.; Baumgardner, M. F.; Franzmeier, D. P.; Stott, D. E.; Coster, D. C. High dimensional reflectance analysis of soil organic matter. *Soil Sci. Soc. Am. J.* **1992**, *56*, 865–872.
- (30) Viscarra Rossel, R. A.; Behrens, T. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* **2010**, *158*, 46–54.
- (31) Kuang, B.; Mouazen, A. M. Calibration of visible and near infrared spectroscopy for soil analysis at the field scale on three European farms. *Eur. J. Soil. Sci.* **2011**, *62*, 629–636.
- (32) Brown, D. J. Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. *Geoderma* **2007**, *140*, 444–453.
- (33) Shi, Z.; Wang, Q. L.; Peng, J.; Ji, W. J.; Liu, H. J.; Li, X.; Viscarra Rossel, R. A. Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. *Sci. China Earth Sci.* **2014**, *57*, 1671–1680.
- (34) Biehl, L. L.; Stoner, E. Reflectance properties of soils. *Adv. Agron.* **1985**, *38*, 1–44.
- (35) Lobell, D. B.; Asner, G. P. Moisture effects on soil reflectance. *Soil Sci. Soc. Am. J.* **2002**, *66*, 722–727.
- (36) Viscarra Rossel, R. A.; Cattle, S. R.; Ortega, A.; Fouad, Y. In situ measurements of soil colour, mineral composition and clay content by vis–NIR spectroscopy. *Geoderma* **2009**, *150*, 253–266.
- (37) Brodský, L.; Vašát, R.; Klement, A.; Zádorová, T.; Jakšík, O. Uncertainty propagation in VNIR reflectance spectroscopy soil organic carbon mapping. *Geoderma* **2013**, *199*, 54–63.
- (38) Nocita, M.; Stevens, A.; Noon, C.; van Wesemael, B. Prediction of soil organic carbon for different levels of soil moisture using Vis–NIR spectroscopy. *Geoderma* **2013**, *199*, 37–42.
- (39) Stevens, A.; Udelhoven, T.; Denis, A.; Tychon, B.; Liroy, R.; Hoffmann, L.; van Wesemael, B. Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. *Geoderma* **2010**, *158*, 32–45.
- (40) Brown, J. D.; Heuvelink, G. B. M. Assessing uncertainty propagation through physically based models of soil water flow and solute transport. In *Encyclopedia of Hydrological Sciences*; Anderson, M. G., McDonnell, J. J., Eds.; John Wiley & Sons: Chichester, 2005; pp 1181–1196.
- (41) Ge, Y.; Morgan, C. L. S.; Ackerson, J. P. VisNIR spectra of dried ground soils predict properties of soils scanned moist and intact. *Geoderma* **2014**, *221*, 61–69.

(42) Ji, W.; Viscarra Rossel, R. A.; Shi, Z. Accounting for the effects of water and the environment on proximally sensed vis-NIR spectra and their calibrations. *Eur. J. Soil Sci.* **2015**, in press, DOI: 10.1111/ejss.12239.