UC Davis
UC Davis Previously Published Works

Title

Regional Scale Dryland Vegetation Classification with an Integrated Lidar-Hyperspectral Approach

Permalink

Journal

ISSN

Authors

Dashti, Hamid
Poley, Andrew
Glenn, Nancy F
et al.

Publication Date

DOI

Empirical Methods for Remote Sensing of Nitrogen in Drylands May Lead to Unreliable Interpretation of Ecosystem Function

Hamid Dashti, Nancy F. Glenn, Susan Ustin, Jessica J. Mitchell, Yi Qi, Nayani T. Ilangakoon, Alejandro N. Flores, José Luis Silván-Cárdenas, Kaiguang Zhao, Lucas P. Spaete, and Marie-Anne de Graaff

Abstract—Nitrogen (N) has been linked to different ecosystem processes, and retrieving this important foliar biochemical constituent from remote sensing observations is of widespread interest. Since N is not explicitly represented in physically based radiative transfer models, empirical methods have been used as an alternative. The spectral bands selected during the calibration of empirical methods have been interpreted in the context of light-N interactions and, consequently, ecosystem function. The low amount of leaves on shrubs and their sparse distribution in drylands create an environment, in which the canopy structure and the bright background soil play an important role in the canopy total radiation budget. In this paper, we examine the assumption that removing the impact of canopy structure and soil will result in improved N retrieval using the empirical methods. We report the inconsistencies in the selection of spectral bands among the empirical approaches. Moreover, these methods are sensitive to the scale of the study and band transformations. Using the generalized theory of canopy spectral invariants, we found that at the canopy scale, a combination of canopy structure and soil dominates the total canopy radiation. At the plot scale, soil contributes up to 95% of the total reflectance. Correction for these two confounding factors leads to no correlation between N and vegetation reflectance at both scales. We conclude that while cross-validated predictive models may be statistically achievable, caution should be taken when interpreting their results in the context of N-light interactions and ecosystem function. Our approach can be extended to all terrestrial ecosystems with multiple layers of canopy and understory.

Index Terms—Drylands, ecosystem function, empirical methods, radiation budget.

H. Dashti, N. F. Glenn, N. T. Ilangakoon, and A. N. Flores are with the Department of Geosciences, Boise State University, Boise, ID 83725 USA (e-mail: hamiddashti@u.boisestate.edu; nancyglenn@boisestate.edu; nayaniilangakoon@u.boisestate.edu; lejoflores@boisestate.edu; marie-annedegraaff@boisestate.edu).

M.-A. de Graaff is with the Department of Biology, Boise State University, Boise, ID 83725 USA (e-mail: marie-annedegraaff@boisestate.edu).

S. Ustin is with the Department of Land, Air and Water Resources, University of California at Davis, Davis, CA 95616 USA (e-mail: slustin@ucdavis.edu).

J. J. Mitchell is with the Montana Natural Heritage Program, University of Montana, Missoula, MT 59812 USA (e-mail: jessica.mitchell@mso.umt.edu).

Y. Qi is with the School of Natural Resources, University of Nebraska-Lincoln, Lincoln, NE 68583 USA (e-mail: yi.qi@unl.edu).

J. L. Silván-Cárdenas is with the Centro de Investigación en Ciencias de Información Geoespacial, Mexico City 14240, Mexico (e-mail: jsilvan@centrogeo.edu.mx).

K. Zhao is with the School of Environment and Natural Resources, Ohio State University, Columbus, OH 43210 USA (e-mail: zhao.1423osu@gmail.com).

L. P. Spaete is with Superior Geospatial, Superior, WI 54880 USA (e-mail: lucasspaete@boisestate.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2018.2889318

I. INTRODUCTION

TERRESTRIAL ecosystem processes have been interpreted from the remote sensing estimates of foliar nitrogen and other leaf biochemicals. Canopy nitrogen (N) has been related to forest albedo and linked to climate change [1], nutrient limitation [2], the Amazonian functional biodiversity [3], and the role of the plant community in controlling canopy biochemistry [4].

There are two general approaches for remote sensing of canopy chemistry: physical methods based on the concept of radiative transfer models (RTMs) and empirical/statistical methods based on regression analysis. A combination of these two approaches, known as the hybrid methods, can also be used [5]. Since there are no reliable RTMs that include leaf N, this foliar variable is mostly identified by empirical methods. Acceptable estimates of N have been reported using a range of empircal methods (average R^2 and root-mean-square error (RMSE) of 0.72 and $\pm$ 0.16, respectively, [6]). These include multiple linear regression [7], [8], partial least squares (PLSs) regression [7], [9]–[11], stepwise multiple linear regression [12], and, more recently, popular methods, such as neural networks [13], [14], support vector machines (SVMs) [15], [16], Bayesian regression (BR) [17], and random forest (RF) [18].

The goal of statistical analysis is to fit a model between N and the feature space (i.e., spectral bands) or a transformation of the feature space. The developed model is then tested for its predictive power using cross validation. The most influential variables on the model fit are then discussed in the context of light-N interactions. There are three issues related to this type of study. First, it is known that multiple chemical, physical, and structural properties of vegetation and

background soil control the spectral signal received at leaf, canopy, and plot scales. In many cases, such as in sparse vegetation cover, N does not dominate the spectra [19], [20]. Ideal empirical studies usually include relevant predictor and response variables. It is thus essential to consider whether the empirical relationships between reflectance spectra and N are suitable. Another consideration is that there is a limited consistency between empirical studies in the selected bands for N prediction [6], [21]. More importantly, in some cases, the selected bands are not consistent with the concepts of radiative transfer of N absorption but rather driven by the canopy structure or other factors that may or may not covary with N. For example, in dense forests with the assumption of a zero canopy background contribution (i.e., black soil) to the total canopy bidirectional reflectance factor (BRF), it has been shown that the canopy structure derives positive correlation between the near-infrared (NIR, 800–850 nm) and N [22]. In fact, multiple studies have identified NIR as the important predictors of N [1], [23], [24]. Finally, further investigation on the generalizability of empirical studies with cross validation is needed. The number of successful N retrieval studies using statistical methods has been used as the affirmation for the replicability of these models [25]. There is an urgent need to examine the interpretability of empirical methods and their fundamental meaning to the remote sensing and ecology community.

One way to study the interpretation of empirical methods is to investigate their link to underlying light-canopy physical processes. Knyazikhin *et al.* [22] used the theory of spectral invariants [26], which is based on radiative transfer. [27], and introduced a directional area scattering factor (DASF) as a new measure of canopy structure. DASF, in concept, is an estimate of the ratio of the leaf area that forms the canopy boundary, as seen along a given direction, to the total leaf area. Normalization of BRF to DASF results in canopy scattering coefficients (W) that are insensitive to canopy structure. In contrast to empirical findings, W exhibits either negative or no correlation with N [22], [28]. A complicating factor is that while the DASF approach assumes a black soil background, in many ecosystems, this assumption is violated, and indeed, impacts of soil can be larger than those from the canopy structure itself.

While empirical methods are widely used for canopy N retrieval, comprehensive studies linking these results to physical processes, such as canopy radiative transfer, are lacking. Our goal is to examine the empirical methods used for more than two decades in the remote sensing community to answer the fundamental question of whether we can rely on these methods to predict N in the context of the confounding factors of canopy structure and soil. Our null hypothesis is that correcting for confounding factors will improve the N predictions using empirical methods. To test our hypothesis, we implement a range of empirical models and physical analyses based on the generalized theory of canopy spectral invariants [22], [29]. Our study advances the community discussion of the light-N interactions beyond dense forests to include ecosystems with multiple layers of canopy and understory.

II. MATERIALS AND METHODS

Five sites were selected across western USA in the semiarid ecosystem, known as the Great Basin (GB), for the field study and data collection (see Fig. S1 in the Supplementary Material). The Reynolds Creek Experimental Watershed, Birds of Prey, and Hollister sites are located in Idaho, and the Big Pine and Lone Pine sites are located in California on the eastern side of the Sierra Mountains. The dominant vegetation cover in the GB is the sparsely distributed shrubs. These dryland study sites provide the opportunity to study the impact of canopy structure and soil on remote sensing of N and extend the previous work in dense forests to xeric ecosystems. Most of the ecosystems follow the same pattern, in which an understory layer (e.g., soil, grass, and so on) contributes to the total pixel radiation budget. Field data sampling was conducted during 2014–2015 (see Table S1 and data set S1 in the Supplementary Material). Considering the dominance of sagebrush (*Artemisia tridentata*) and bitterbrush (*Purshia tridentata*) in the study sites, plots were selected based on the dominant cover of one of these two species. We define three scales for the study: at the leaf scale, the spectra were collected from dry leaves; at the canopy scale, the spectrometer was held above the top of the canopy; and at plot scale (10 m × 10 m), the spectra are acquired from the Airborne Visible/Infrared Imaging Spectrometer—Next Generation (see data set S2 in the Supplementary Material) airborne hyperspectral sensor. An extended version of the data collection can be found in Text1 in the Supplementary Material. Four different types of spectral transformations were applied to the spectra, the smoothed data set using the Savitzky–Golay filter [30] with a frame size of 11, second-degree polynomial, logarithmic transformation, first derivative of smoothed data set, and logarithm of the first derivative. These transformations are widely used in remote sensing of canopy chemistry and are known to enhance the absorption features [31]–[33]. For the statistical analysis, we implemented PLS [34], SVM [35], RF [36], and BR regression [17] methods and a newly developed multimethod ensemble variable selection (VS) based on the integration of PLS, SVM, and RF [37]. In this Ensemble approach, a spectral band is important if it is identified as important by all three methods. Each method returns band importance, which will be weighted by the explained variance of selected model for each method according to the following equation:

$$I_{iw} = (I_i * R^2)/\sigma_I \tag{1}$$

where I_{iw} is the weighted importance of band i, I_i is the raw measure of the band importance of the regression method, R^2 is the explained variance of the model in cross validation or out-of-the-bag testing, and σ_I is the standard deviation across the raw measures of importance of each model. The product of the three weighted importance values is considered as the Ensemble importance. As a base method, we calculate the variable importance in projection (VIP) by developing 1000 PLS models [37], [38]. VIP is the most common approach for VS based on PLS outputs. In our approach, a band was considered

important when its average VIP value along with one standard deviation (derived from 1000 iterations) was above 1. We refer to the mean of the 1000 PLS models as the PLS_{Ref} model. The reported R^2 and the coefficient of variation (CV) for PLS_{ref} is the mean of all 1000 model runs. K-fold cross validation and also leave one data set out validation has been used to assess the model's performance. In the leave one data set out approach, a complete data set that has been collected in a given year and site was kept out of the calibration step. A synthetic data set was created to test the performance of all methods in an ideal case. This data set contained 200 observations with 500 correlated predictors; 10 of the predictors were set to have a coefficient value of 1 (i.e., relevant predictors), and all other predictors are 0 (unrelated predictors). A complete description of a similar data set construction is presented in [17]. The purpose of the statistical analysis is not to test a comprehensive list of the algorithms but to implement the most common ones.

The physical analysis is based on the theory of canopy spectral invariants. According to this theory, under the assumption of black soil, canopy scattering $s(\lambda)$ and absorption $a(\lambda)$ are expressed in the following equations:

$$s(\lambda) = \frac{(1-p)\omega_l(\lambda)}{1 - p\omega_l(\lambda)} i_0 \qquad (2)$$

$$a(\lambda) = \frac{1 - \omega_l(\lambda)}{1 - p\omega_l(\lambda)} i_0 \qquad (3)$$

where i_0 is the probability of canopy interceptance, ω_l is the single-scattering albedo of an average phytoelement at any wavelength, and p is the recollision probability [39]. The recollision probability can be interpreted as the probability that a photon interacted with canopy elements will interact within the canopy again [40]. This theory can be generalized to the situation with multiple end-members, and the interaction between photons and end-members can be treated as an infinite-state Markov chain [29]. Then, (2) and (3) have the following form:

$$s(\lambda) = \mathbf{q}^T \mathbf{\Omega}(\lambda) \mathbf{i}(\lambda) \qquad (4)$$
$$a(\lambda) = \propto^T \mathbf{i}(\lambda) \qquad (5)$$

where $\mathbf{q} = (\mathbf{I} - \mathbf{P_n})\mathbf{1}$ is the vector of escape probability after n interactions, $\mathbf{\Omega}(\lambda) = \mathbf{diag}[\omega_1(\lambda)\ldots\omega_m(\lambda)]$ is a diagonal matrix of single-scattering albedo associated with end-members, and $\propto (\lambda) = (\mathbf{I} - \mathbf{\Omega}(\lambda))\mathbf{1}$ is the vector of the end-member's absorptance. The quantity $\mathbf{q}$ defines the probability that a photon scattered by phytoelements will escape vegetation toward a given direction [41]. This generalization includes both photons from the canopy end-members r_c and background end-members r_B or $s = r_c + r_B$. Furthermore, based on the principle of energy conservation, we can calculate the canopy radiation budget (CRB) and quantify canopy and surface contribution to the CRB. We assume that there are two layers consisting of a canopy layer on top of a flat soil layer. The solution of CRB with reflective surface is built by integrating the black soil solution (BS) problem, in which soil impact is negligible, and the second solution where soil is considered as the source of illumination (S problem). Then,

CRB can be expressed in the following equations:

$$r_c(\lambda) = r_{\text{BS}}(\lambda) + \frac{\rho_s(\lambda)t_{\text{BS}}(\lambda)}{1 - \rho_s(\lambda)r_s(\lambda)}(t_s(\lambda) - q_s) \qquad (6)$$

$$a_c(\lambda) = a_{\text{BS}}(\lambda) + \frac{\rho_s(\lambda)t_{\text{BS}}(\lambda)}{1 - \rho_s(\lambda)r_s(\lambda)}a_s(\lambda) \qquad (7)$$

$$t_c(\lambda) = t_{\text{BS}}(\lambda) + \frac{\rho_s(\lambda)t_{\text{BS}}(\lambda)}{1 - \rho_s(\lambda)r_s(\lambda)}(r_s(\lambda) - p_{sl}) \qquad (8)$$

where r_c, a_c, t_c, ρ_s, and q_s are canopy reflectance, canopy absorptance, canopy transmittance, soil reflectance, and photon escape probability from soil, respectively. The BS' and S' reflectance, absorptance, and transmittance are defined in the following equations:

$$r_{\text{BS}}(\lambda) = \frac{q_l\omega_l(\lambda)}{1 - p\_\{LL\}\omega_l(\lambda)} i_0 \qquad (9)$$

$$a_{\text{BS}}(\lambda) = \frac{1 - \omega_l(\lambda)}{1 - p\_\{LL\}\omega_l(\lambda)} i_0 \qquad (10)$$

$$t_{\text{BS}}(\lambda) = t_0 + \frac{p\_\{LS\} - \omega_l(\lambda)}{1 - p\_\{LL\}\omega_l(\lambda)} i_0 \qquad (11)$$

and

$$r_s = \frac{(t_{\text{BS}} - t_0)p\_\{SL\}}{i_0} \qquad (12)$$

$$a_s = \frac{a_{\text{BS}} p\_\{SL\}}{i_0} \qquad (13)$$

$$t_s = \frac{r_{\text{BS}} p\_\{SL\}}{i_0} + q_s \qquad (14)$$

where q_l, i_0, and t_0 are canopy escape probability, canopy interceptance, and uncollided transmittance of the BS problem, respectively. $p\_\{LL\}$, $p\_\{LS\}$, $p\_\{SL\}$, and $p\_\{SS\}$ are the leaf–leaf, leaf–soil, soil–leaf, and soil–soil recollision probabilities. The second term in (6) accounts for the influence of soil to the CRB. Photons that escape directly from the surface are not part of the canopy reflectance even if they interact with canopy before reaching the surface. Subtracting this term from r_c leave us with r_{BS}, which has the form noted in (9). Not surprisingly, r_{BS} is similar to the model developed for the black soil problem [39, eq. (2)] with recollision probability $p = p\_\{LL\}$. Thus, one should note that the contribution of soil to the total reflectance is the sum of the photons that escape from soil (r_B) and reach the sensor and the portion of photons that escape from soil and influence the canopy reflectance (r_c). We refer to the first component as soil contribution to the total reflectance and second component as the soil contribution to the reflectance of CRB. A complete derivation and description of the terms in (4)–(14) are provided in [29] and [42]. With the assumption that soil is flat (no interaction between soil end-members), $p\_\{SS\}$ is much lower than the other recollision probabilities and can be neglected ($p\_\{SS\} = 0$). In order to estimate $p\_\{LL\}$, $p\_\{LS\}$, $p\_\{SL\}$, and i_0, we fit the spectra at each scale to (4) using the covariance matrix adaptation evolution strategy (CMAES) optimization approach [43]. The mean RMSE between simulated and measured total canopy spectra at the canopy scale is 0.01 and at the plot scale is 0.02. CMAES is a state-of-the-art evolutionary algorithm developed for nonlinear, nonconvex black-box optimization problems.
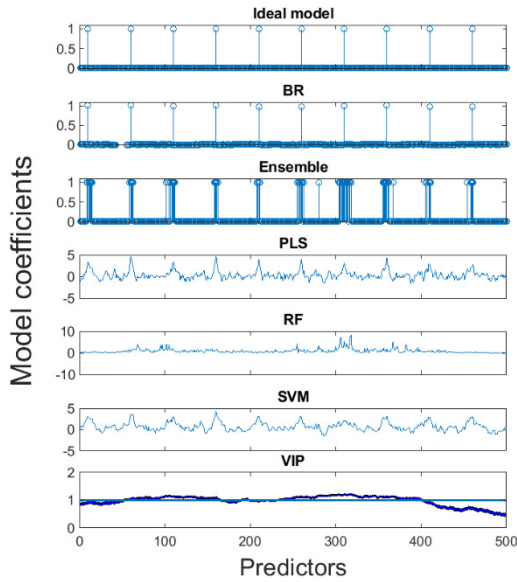
Fig. 1. Performance of different empirical models on an ideal synthetic data set.

Fig. 2. VS using different methods at the leaf scale for four different spectral transformations for nitrogen estimation.

We added 4% Gaussian noise to the spectrums and performed inversion. Our preliminary analysis (not shown in this paper) using 100 different random initial points and a noisy objective function showed that CMAES is not sensitive to the initial points or noise and consistently returns the global minima. Using the recollision probabilities and (6)–(8), CRB was calculated. For the retrieval of DASF, r_{BS} was used. The DASF was calculated using the algorithm developed in [22]; the only modification was instead of using a reference leaf albedo, we used measured green leaf reflectance of sagebrush and bitterbrush [44]. In order to calculate the leaf albedo, leaf transmittance was acquired by inversion of PROSPECT-5 leaf model [45] using the measured leaves reflectance. In summary, the steps of our physical analysis are as follows.

1) Fit the spectrum to (4) to estimate the recollision probabilities.
2) Use estimated recollision probabilities to calculate r_{BS} using (6)–(14).
3) Use r_{BS} to calculate DASF and canopy scattering.
4) Canopy scattering is then used for statistical analysis.

III. RESULTS AND DISCUSSION

A. Variable Selection Is Sensitive to Transformation and Scale

Using an ideal data set (materials and method), Fig. 1 shows the performance of an Ensemble model and its submodels (PLS, SVM, and RF), VIP, and BR VS techniques. We start by discussing the VIP results. Between methods, VIP performs the worst in VS by identifying numerous unrelated bands as important. In cases where there is high covariation between variables (e.g., spectral bands) but small correlation to the target (e.g., N), the standardization used by PLS can address the latter but at the cost of increasing the weight of minor variables with low signal-to-noise ratios [46]. This directly impacts the V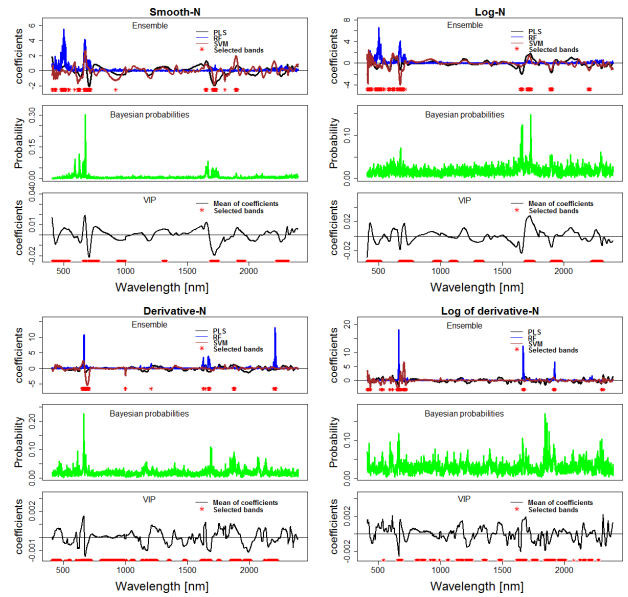IP as it is a filter method and uses the PLS output with no further postprocessing. This behavior can also be seen at the leaf scale (see Fig. 2), though to a lesser degree than the canopy and plot scales. Our first observation shows that the VIP is sensitive to the type of transformation. Across all transformations, the VIP identified the NIR region (∼800–1350 nm) as important predictors of leaf N. This is controversial since most of the incident radiation is reflected and transmitted (∼50% each) in the NIR region by the leaf mesophyll [47], and it is extremely difficult to identify weak N absorption bands (e.g., 910 and 1020 nm [48]) using VS techniques. While many hyperspectral studies use VIP for band selection [32], [49], [50] and, specifically, in foliar N estimation [10], [24], [51], [52], the selection of unrelated bands may require further examination of VIP.

Fig. 1 shows that the BR produced the closest results to the ideal model for band selection, and the Ensemble approach follows with less ideal but comparable results. In the Ensemble approach, a spectral band is important if it is considered important by all three regression methods (PLS, RF, and SVM). The peaks in the PLS match the coefficients of the ideal model, but PLS has assigned weights to many unrelated variables. RF is more restrictive, and consequently, many unrelated bands are close to zero. This restrictive behavior, however, caused RF to miss some of the important variables. The SVM results are similar to PLS, which has been observed in other studies [37]. The improved performance of the BR and Ensemble methods can be attributed to the fact that both methods are ensembles of competing models. BR is based on the Bayesian model averaging and has theoretical advantages over standard regression analysis [17]. The Ensemble method, on the other hand, uses the relative merits of PLS, SVM, and RF. For example, the restrictive VS method of RF is balanced with the more inclusive PLS and SVM in the Ensemble approach.

When we used field data at the leaf scale, both the Ensemble and RF methods show less sensitivity to the transformations. However, the individual models of the Ensemble in isolation show sensitivity to the transformation. These transformations have been accepted as a standard preprocessing method for remote sensing of foliar biochemistry. Thus, two studies with the same data set, but different spectral transformations for predicting N can suggest different spectral bands as important predictors. At the leaf scale, our findings show that the log-transformed data set leads to more meaningful bands when compared with known N absorption regions (see Table S2 in the Supplementary Material). If assuming the leaf reflectance spectrum has the same shape as leaf transmittance, then the log transformation is an approximation of the foliar absorption spectra and is consistent with the Beer's law absorption coefficient [53].

Using the log transformation and the Ensemble approach for interpretation, most of the selected bands at the leaf scale are in the visible and mid-infrared (>1350 nm) with no bands selected in the NIR region. Selection of the visible spectrum can be attributed more explicitly to the chlorophyll content [54]. Moderate correlation is reported between chlorophyll and N [6]. In the shortwave-infrared region identified bands centered near 1655, 1715, 1900, and 2200 nm can be attributed to N absorption at 1690, 1940, and 2240 nm. Among them, the bands near 2240 and 2300 nm are more directly related to the N or proteins that carry N, while other bands can be associated with absorption by biochemicals, such as lignin, cellulose, and starch. The complex interactions between N and these biochemicals may result in misinterpretation of the selected bands and, consequently, unreliable statistical estimates of N. The Ensemble method is a VS method and does not provide predictions. Since the VS performance of BR and Ensemble is close, BR can be used for both VS and model fitting at the leaf scale. We conclude that at the leaf scale, with the assumption of no confounding factors, robust empirical methods, such as BR, are likely to provide a meaningful model for N predictions. However, it is important to choose the appropriate spectral transformation (e.g., logarithmic transformation), and results of the VS should be checked with known absorption regions.

Moving from the leaf to canopy and plot scales, the selected bands for N differ, regardless of the method used. Fig. 3 shows VS for the log-transformed data set. Other transformations can be found in Fig. S3 in the Supplementary Material. The fundamental assumption of empirical predictions is foliar N controls the reflectance [55], [56], and thus, we would expect consistency in the bands identified as important across scales. However, Fig. 3 indicates a lack of this consistency across scales. This is the first evidence that there are other elements that may have more dominant role in controlling the canopy radiation.

Using log-transformed data and focusing on the Ensemble or BR methods, the spectral regions identified as important for N are similar to the regions identified as important for a similar analysis of leaf area index (LAI) at the canopy and plot scales (see Fig. 4, and Fig. S4 in the Supplementary Material). Most of these bands lie in the red edge and the NIR region. The
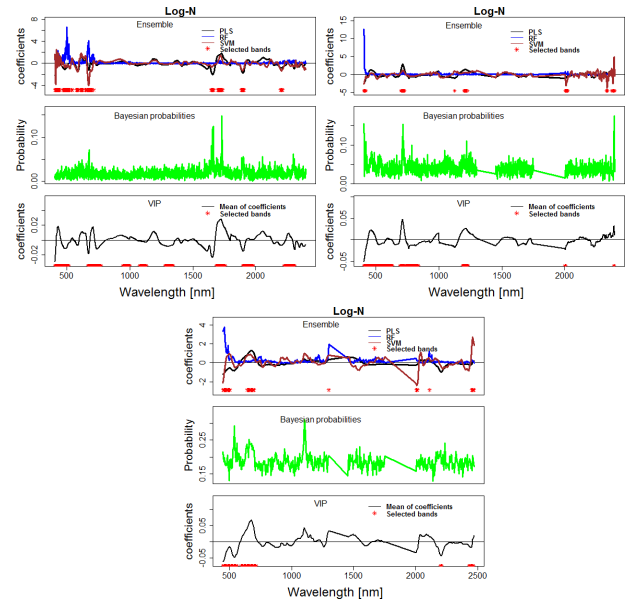

Fig. 3. Scaling up band selection for N from (Top Left) leaf to (Top Right) canopy and (Bottom) plot. Data are shown for log transformation.
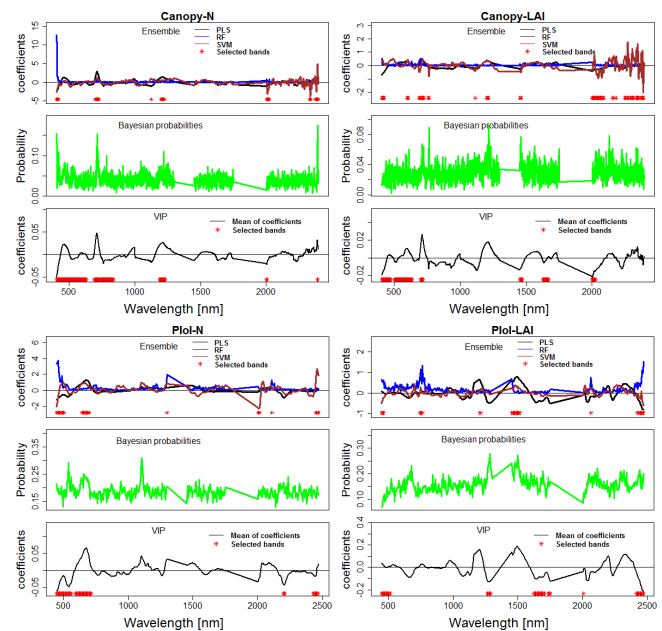

Fig. 4. VS for LAI and N at (Top) canopy and (Bottom) plot scales (data shown for log transformation).

NIR region in particular is known to be attributed to canopy structure (i.e., LAI). However, the predictive power of these bands for LAI is much less than N (see Tables S3 and S4 in the Supplementary Material). For example, the cross-validated mean R^2 and CV of all methods for the log-transformed prediction of N at the canopy scale are 0.61 and 16.67, respectively, while for LAI, they are 0.26 and 35.39, respectively. It is tempting to discuss this in terms of ecology and the association between canopy N and LAI; however, it is more likely a statistical problem. An explanation is that statistically significant explanatory variables (e.g., spectral bands) that

have an association with a target variable might not necessarily carry the most predictive power, and the most predictive variables are not necessarily the most significant ones [57], [58]. A key distinction that makes a variable significant or predictive lies in the properties of their underlying distribution. This issue has been observed in different disciplines from genome-wide association studies [59] for disease predictions to social studies predicting civil wars [60]. Thus, the problem can be statistically framed by asking if the research goal is to find *highly significant* or *highly predictive* variables, whereas searching for both significant and predictive variables can lead to conflicting directions [57]. The notion of predictability and significance of variables has not been explored in remote sensing of canopy biochemistry. In Section III-B, we discuss the problem associated with cross validation and demonstrate that these predictive variables for N lose their predictive power when applied to an external data set, of which time or location is different from the calibration data set.

B. Cross Validation Is Overoptimistic

Generalization of a statistical method is a key concept and refers to applying a model based on a particular target population to other populations [61]. As discussed in the classical paper by Verstraete *et al.* [55] (1996), "a relation, obtained by statistically correlating remote sensing measurements and field observations, is *useful* only for those locations and times other than those used to establish the correlation." Otherwise, remote sensing provides no new information. Cross validation is accepted as the *de facto* standard method used in remote sensing communities to test for the generalization of the developed model. Table S5 in the Supplementary Material shows results for normal cross-validation versus validation results for the data set that was kept out of calibration (see Section II). The data sets are different in either time (year of data collection) or spatial location. While cross validation shows *well enough* performance of different methods, when applying the same models to an independent data set, R^2 of at least one of the methods is close to 0. Similar results are reported in other disciplines [62], [63]. Even strict cross validations may still be overoptimistic due to heterogeneity between data sets [64]. Thus, our results indicate that indeed these statistical methods are not replicable.

Changes in time and location change the distribution of feature space [65], [66] due to differences in measurements or state variables that control the soil–plant-reflectance interactions. The state variables are those that are clearly represented in RTMs [55]. The simplest statistical fix for the replicability of statistical methods is to seek models that perform well enough in the context of the leave one data set out test. None of the methods used in this paper show strong overall performance with this test. More sophisticated solutions are based on the methods that can compensate (or correct) for distributional shifts that may also be referred to as "domain adaptation" [67], [68]. The recent framework developed in [65] would enable us to identify and correct the distributional shift. The important point here is even if we correct for distribution shifting and ignore the fact that we do
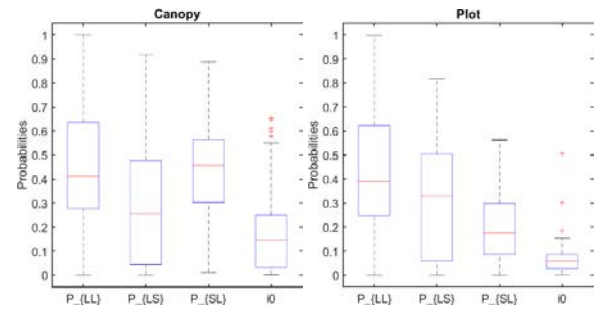

Fig. 5. Spectral invariant parameters at (Left) canopy and (Right) plot scales.

not have enough information about the vegetation to derive N (see Section III-C), our model is still, at best, a predictive model and is not correlative. Because we are changing the distribution of features, as a consequence, we might obtain a new set of features that might have good predictive power, even for external data sets, but with limited correlative relation with N. Our conclusion is that if we ignore the impacts of canopy structure and soil on the total canopy reflectance and if the question at hand is just prediction not interpretation, then with some statistical manipulations, we are able to produce predictive models. However, caution must be observed when interpreting these models and, in particular, at scales larger than the leaf scale.

C. Canopy Structure and Soil Dominate the Total Canopy Reflectance

Fig. 5 shows the box plot of the estimated spectral invariants at the canopy and plot scales. Here, we assume a two-layer system, in which a layer of a canopy is on the top of a layer of soil, an assumption similar to many open canopy ecosystems. However, this approach can be extended to the layers of multiple canopies from different species and understories. The probabilities of a photon intercepted by the canopy and soil are i_0 and $1 - i_0$, respectively. $P\_\{LL\}$, $P\_\{LS\}$, and $P\_\{SL\}$ are the probabilities of photon interactions between canopy–canopy, canopy–soil, and soil–canopy, respectively. Fig. 5 shows that changing scale from canopy to plot should affect $P\_\{SL\}$ and i_0 but not $P\_\{LL\}$ and $P\_\{LS\}$. This is because, the probability that a photon escapes from the canopy $(1 - P\_\{LL\} - P\_\{LS\})$ should remain independent of the soil condition (e.g., reflective versus nonreflective). At the canopy scale, the mean of i_0 is 0.17, and at the plot scale, it is 0.05. This low number simply shows the large impact of the soil on the total reflectance at both canopy and plot scales. For example, if we assume no additional interaction between photons from vegetation and soil, the total canopy reflectance is composed of 17% information from the shrub and 83% from the soil. At the plot scale, the contribution of soil can be represented in two forms: photons leaving the soil toward the sensor and photons leaving the soil and contributing to the CRB. Thus, the impacts of the soil at the plot scale are more than the canopy. Fig. 6 compares the simulation for two shrubs. The contribution of canopy reflectance to total reflectance is much higher in the green shrub. If we remove the contribution
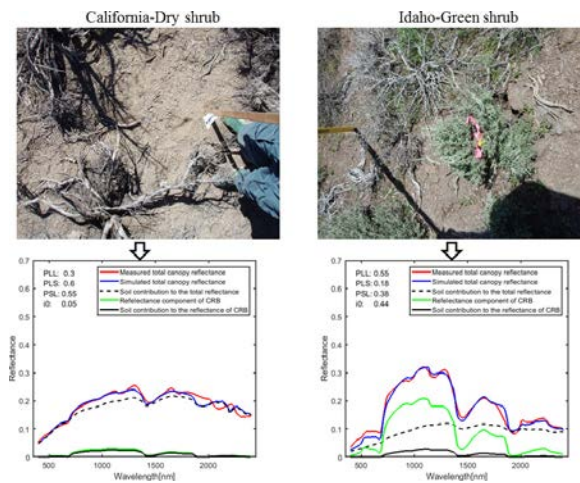
Fig. 6. Comparison of simulations between (Left) dry shrub located in California and (Right) green shrub located in Idaho.
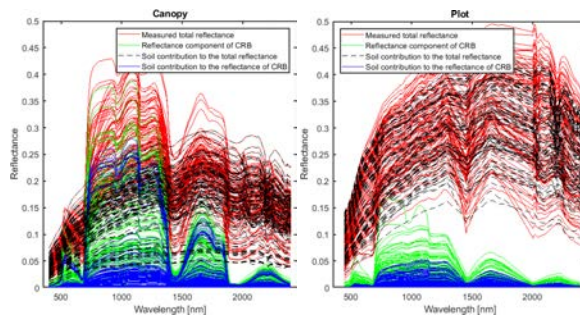

Fig. 7. Simulation of different components of total reflectance of (Left) canopy and (Right) plot samples.

Fig. 8. Relationship between DASF and BRF in the NIR region.
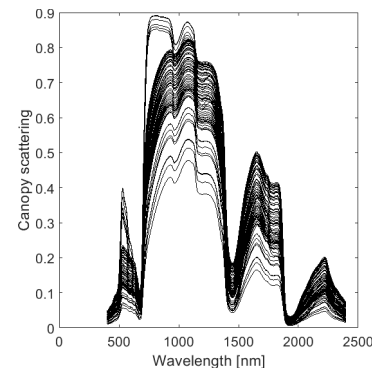
Fig. 9. Canopy scattering correction for soil and canopy structure.

of soil in the CRB for the dry shrub, then the residual (which is reflectance of the canopy itself) is close to 0. This is expected since there is no leaf on the canopy. Finally, Fig. 7 shows the simulations of different components of the total canopy and plot reflectance for all samples. Not surprisingly, the greater impact of soil at the plot scale, compared to the canopy scale, is observable.

The next step is to estimate the impact of canopy structure on the total BRF. We removed the soil contribution in total BRF, and the residual was used to calculate the DASF (see Section II). Among 151 samples at the canopy scale, we could not calculate DASF for 49 samples. These shrubs are mostly located in California sites, suffering drought, that have few small leaves [e.g., Fig. 6 (Left)]. Thus, after removing the impact of soil, the residual is close to 0, and calculating DASF is meaningless. As expected, this problem is worse at the plot scale. In fact, we could not calculate DASF for any of the plot samples after removing soil. This is another piece of evidence that supports the rejection of using a cross-validated statistical N estimate at the plot scale (see Table S3 in the Supplementary Material). In conclusion, the majority of the information contained in the plot reflectance is from the soil (up to 95%), and after removing it, there might not be enough information to infer canopy structure or N. Thus, our statistical prediction of foliar N at the plot scale is unreliable.
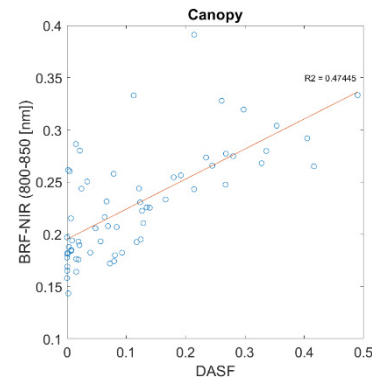
R^2 of the BRF/leaf albedo ratio versus BRF relationship is an estimate of the DASF quality retrieval. A note of caution is that, in theory, it is still possible to estimate DASF values with an R^2 close to 1 for small BRF. However, the estimated DASF is very small. Normalizing canopy BRF to a small DASF will result in large canopy scattering (W more than 1), which is impossible. This is important since filtering sparse vegetation based on this R^2 has been recently suggested [69], which may lead to incorrect canopy scattering. For green shrubs, DASF has a positive correlation (see Fig. 8) with the shoulder of the NIR region (800–850 nm). This is in line with the radiative transfer process and the findings by Knyazikhin *et al.* [22] (2013).

Fig. 9 shows canopy scattering for the shrubs, for which we were able to calculate DASF. Canopy scattering mimics a typical leaf albedo and is insensitive to canopy structure [22], demonstrating success with the DASF approach. In Section III-D, we experiment with using canopy scattering across all wavelengths to predict foliar N.

D. Correlation and Causation and the Concept of Counterfactual

Table I shows the results of corrected canopy samples versus their counterpart samples, which were not corrected for soil and canopy structure. It can be seen that after correction for canopy structure and soil, none of the methods or transformations predict N for canopy-scale data. Since we have selected the log transformation for our statistical analysis in

TABLE I

EMPIRICAL ESTIMATION OF CANOPY N FOR SHRUBS CORRECTED FOR THE IMPACT OF SOIL AND CANOPY STRUCTURE

	Before correction for structure and soil					After correction for structure and soil				
	Ensemble			BR	PLS_ref	Ensemble			BR	PLS_ref
	PLS	SVM	RF			PLS	SVM	RF		
Smoothed										
R2	0.61	0.49	0.37	0.37	0.51	0.19	0.18	0.16	0.18	0.08
CV	16.87	21.90	22.66	18.32	19.17	26.54	26.75	30.32	23.73	27.38
Log transformation										
R2	0.60	0.62	0.37	0.47	0.52	0.18	0.19	0.16	0	0.08
CV	18.74	19.63	22.34	16.49	19.42	26.57	26.91	30.54	26.9	27.36
First derivative										
R2	0.57	0.54	0.61	0.35	0.42	0.17	0.16	0.15	0	0.07
CV	19.79	19.46	16.24	18.34	21.68	26.58	26.71	30.1	26.3	27.46
Log transformation of the first derivative										
R2	0.58	0.74	0.67	0.36	0.52	0.12	0.16	0.17	0	0.05
CV	18.27	14.21	15.45	16.13	19.28	26.52	26.59	30.33	27.06	27.41

Section III-A, we also provided its 1:1 plots between measured and predicted N before and after corrections (SI, Fig. S5 in the Supplementary Material). The correlation between N and reflectance implies association (dependence) rather than causation. Association can exist between two variables in the presence or absence of causality [70]. It is common to deduce a causal relationship from a correlation. For any causal claim to be verified, one should consider the *might be* condition. For example, what *might be* the case for the N-light relationship if canopy structure and soil did not exist? This is known as the theory of *counterfactual* [71]. Causality is the fundamental property of a system, which means a causal relationship would be *invariant* to the changes of the system. Our physical analysis provides the basis to test for the counterfactual outcomes. Since the results of the N prediction before and after correction for soil and canopy structure have changed, a causal relationship between reflectance and N cannot be concluded. Consequently, we reject our null hypothesis that removing confounding factors improves predictions.

One argument is that N-reflectance correlation implies a functional association [72], which is consistent with ecological understanding (i.e., plant physiology). From our analysis, correction for canopy structure and soil leads to no correlation. This does not, however, invalidate the functional association. Undoubtedly, N plays an important role in different canopy processes; however, not all associations lead to correlation [70]. The functional association can be translated into *a priori* information. Currently, remote sensing alone is not able to incorporate such *a priori* information into the predictions of N. Dynamic vegetation models (DVMs) can be used to reconcile the theory of remote sensing and ecology. For example, these models incorporate both ecological processes (e.g., photosynthesis) and light-canopy interactions (e.g., RTMs).

To account for the role of canopy structure—and to some extent soil, two solutions have been proposed. First, empirical models are applied to adjusted spectra that have been filtered with a normalized difference vegetation index and height and, then, adjusted with a brightness normalization [73]. A second approach is to add light detection and ranging (LiDAR)-derived canopy structural parameters, such as canopy height or fractional cover, to the feature space to construct the statistical model [52]. In both approaches, multiple scattering is not explicitly solved. The impact of canopy geometry, such as orientation and arrangement of leaves and branches, as well as multiple scatterings between the canopy and different layers of understory, including soil, confound the N signal. Adding LiDAR variables to the feature space makes the final model more complicated to interpret rather than simpler.

IV. CONCLUSION

Due to the lack of inclusion of N in leaf RTMs, N, historically, has been estimated with remote sensing data using statistical methods [6]. The interpretation of the statistical models depends on the spectral bands selected during the process of model fitting. We have shown that different models can identify different important bands (see Fig. 1). Moreover, each model is sensitive to the type of transformation applied to the spectra before model fitting. These experiments show that common VS routines for foliar biochemistry studies at scales coarser than the leaf may be misleading. Strong prediction rates reported in remote sensing studies are often based on cross validation, which may be overoptimistic. None of our empirical methods could reproduce cross-validation results when applied to an external data set. Thus, we concluded that these methods are not replicable.

We extended the physical work of [22] to cases, where vegetation is sparse, and soil cannot be ignored. At the plot scale, the impact of soil is a dominant confounding factor, and in more extreme cases, such as drylands, there might not be enough information to retrieve biochemistry or some canopy structure variables, such as DASF. Recent developments using full-waveform LiDAR may solve the problem of canopy structure [74]. Removing confounding factors at the canopy and plot scales leads to different statistical models that might or might not have prediction power. We discussed this against the theory of counterfactuals, leading to rejection of our null hypothesis that removing confounding factors will improve empirical predictions. The idea of functional association, which is used to justify statistical methods, is best suited for remote sensing coupled with DVMs.

Our overall conclusion is that if we are interested in predicting N with remote sensing, then we might be able to produce such empirical models, in particular with the growing body of machine learning algorithms. However, one must be cautious in interpreting these models, particularly in complex ecosystems, because they may be affected by the canopy structure, soil, spectral transformation, and the type of model implemented. With advancements in space-borne hyperspectral and full-waveform LiDAR observations and, thus, sophisticated measurements of ecological processes

(e.g., photosynthesis, solar-induced fluorescence, and so on), there is strong potential to gain new insights of changing ecosystems [75]. Empirical methods serve an important role in analyzing the products of these newly developed sensors. Given the issues related to empirical methods, the remote sensing community should be cautious about the statistical tools they use. We encourage deeper discussions of these methods and the need to explore new fields in statistics for remote sensing, such as a rigorous investigation of causality and predictivity versus significance of variables.

Our results might imply that radiative transfer modeling is the best approach to estimate N. However, currently there is no reliable leaf scale RTM that incorporates N because of the lack of absorption coefficients and other information due to the many different types of bonds N carries [76]. Moreover, RTMs at both the leaf and canopy scales carry multiple implicit and explicit assumptions that might cause relative error up to 70% [77]. For example, the assumption of considering canopy as turbid media in 1-D RTMs leads to the definition of *effective* LAI, which is different from true LAI measured [78]. Regardless, these deficiencies do not make RTMs useless. Most DVMs require RTMs to simulate the CRB. Thus, a potentially better approach to solve the problem of N retrieval is to incorporate it with the DVMs (i.e., through data assimilation) rather than using remote sensing data in insolation. For example, ED2 [79], [80] and CLM [81] DVMs use a simple two-stream canopy RTM [82]. Incorporation of more sophisticated 3-D RTMs into these models might improve their performance [83] and, consequently, facilitate better N predictions. In addition, the canopy spectral invariants theory has produced results close to 3-D RTMs [84] with much faster performance, and thus, a logical next step is to incorporate the generalized canopy spectral invariants theory into DVMs.

V. Acknowledgment

References

[1] S. V. Ollinger *et al.*, "Canopy nitrogen, carbon assimilation, and albedo in temperate and boreal forests: Functional relations and potential climate feedbacks," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 49, pp. 19336–19341, Dec. 2008.

[2] G. P. Asner, C. B. Anderson, R. E. Martin, R. Tupayachi, D. E. Knapp, and F. Sinca, "Landscape biogeochemistry reflected in shifting distributions of chemical traits in the Amazon forest canopy," *Nature Geosci.*, vol. 8, pp. 567–573, May 2015.

[3] G. P. Asner *et al.*, "Amazonian functional diversity from forest canopy chemical assembly," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 15, pp. 5604–5609, Apr. 2014.

[4] K. M. Dahlin, G. P. Asner, and C. B. Field, "Environmental and community controls on plant canopy chemistry in a Mediterranean-type ecosystem," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 17, pp. 6895–6900, Apr. 2013.

[5] S. Liang, "Estimation of land surface biophysical variables," in *Quantitative Remote Sensing of Land Surfaces*. Hoboken, NJ, USA: Wiley, 2003, pp. 246–309.

[6] L. Homolová, Z. Malenovský, J. G. P. W. Clevers, G. García-Santos, and M. E. Schaepman, "Review of optical-based remote sensing for plant trait mapping," *Ecol. Complex.*, vol. 15, pp. 1–16, Sep. 2013.

[7] W. Feng *et al.*, "Measuring leaf nitrogen concentration in winter wheat using double-peak spectral reflection remote sensing data," *Field Crops Res.*, vol. 159, pp. 43–52, Mar. 2014.

[8] J. Pacheco-Labrador, R. González-Cascón, M. P. Martín, and D. Riaño, "Understanding the optical responses of leaf nitrogen in Mediterranean Holm oak *(Quercus ilex)* using field spectroscopy," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 26, pp. 105–118, Feb. 2014.

[9] A. Ramoelo, A. K. Skidmore, M. A. Cho, M. Schlerf, R. Mathieu, and I. M. A. Heitkönig, "Regional estimation of savanna grass nitrogen using the red-edge band of the spaceborne RapidEye sensor," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 19, pp. 151–162, Oct. 2012.

[10] L. C. Lepine, S. V. Ollinger, A. P. Ouimette, and M. E. Martin, "Examining spectral reflectance features related to foliar nitrogen in forests: Implications for broad-scale nitrogen mapping," *Remote Sens. Environ.*, vol. 173, pp. 174–186, Feb. 2016.

[11] J. J. Mitchell, N. F. Glenn, T. T. Sankey, D. R. Derryberry, and M. J. Germino, "Remote sensing of sagebrush canopy nitrogen," *Remote Sens. Environ.*, vol. 124, pp. 217–223, Sep. 2012.

[12] P. Miphokasap, K. Honda, C. Vaiphasa, M. Souris, and M. Nagai, "Estimating canopy nitrogen concentration in sugarcane using field imaging spectroscopy," *Remote Sens.*, vol. 4, no. 6, pp. 1651–1670, 2012.

[13] M. Kalacska, M. Lalonde, and T. R. Moore, "Estimation of foliar chlorophyll and nitrogen content in an ombrotrophic bog from hyperspectral data: Scaling from leaf to image," *Remote Sens. Environ.*, vol. 169, pp. 270–279, Nov. 2015.

[14] Z. Huang, B. J. Turner, S. J. Dury, I. R. Wallis, and W. J. Foley, "Estimating foliage nitrogen concentration from HYMAP data using continuum removal analysis," *Remote Sens. Environ.*, vol. 93, nos. 1–2, pp. 18–29, Oct. 2004.

[15] J. Sun *et al.*, "Evaluation of hyperspectral LiDAR for monitoring rice leaf nitrogen by comparison with multispectral LiDAR and passive spectrometer," *Sci. Rep.*, vol. 7, Jan. 2017, Art. no. 40362.

[16] Y. Karimi, S. O. Prasher, R. M. Patel, and S. H. Kim, "Application of support vector machine technology for weed and nitrogen stress detection in corn," *Comput. Electron. Agricult.*, vol. 51, nos. 1–2, pp. 99–109, 2006.

[17] K. Zhao, D. Valle, S. Popescu, X. Zhang, and B. Mallick, "Hyperspectral remote sensing of plant biochemistry using Bayesian model averaging with variable and band selection," *Remote Sens. Environ.*, vol. 132, pp. 102–119, May 2013.

[18] X. Li, X. Liu, M. Liu, and L. Wu, "Random forest algorithm and regional applications of spectral inversion model for estimating canopy nitrogen concentration in rice," *J. Remote Sens.*, vol. 18, no. 4, pp. 934–945, 2014.

[19] G. P. Asner, "Biophysical and biochemical sources of variability in canopy reflectance," *Remote Sens. Environ.*, vol. 64, no. 3, pp. 234–253, Jun. 1998.

[20] G. P. Asner, C. A. Wessman, C. A. Bateson, and J. L. Privette, "Impact of tissue, canopy, and landscape factors on the hyperspectral reflectance variability of arid ecosystems," *Remote Sens. Environ.*, vol. 74, no. 1, pp. 69–84, Oct. 2000.

[21] Y. L. Grossman, S. L. Ustin, S. Jacquemoud, E. W. Sanderson, G. Schmuck, and J. Verdebout, "Critique of stepwise multiple linear regression for the extraction of leaf biochemistry information from leaf reflectance data," *Remote Sens. Environ.*, vol. 56, no. 3, pp. 182–193, 1996.

[22] Y. Knyazikhin *et al.*, "Hyperspectral remote sensing of foliar nitrogen content," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 3, pp. E185–E192, Jan. 2013.

[23] M. E. Martin, L. C. Plourde, S. V. Ollinger, M.-L. Smith, and B. E. McNeil, "A generalizable method for remote sensing of canopy nitrogen across a wide range of forest ecosystems," *Remote Sens. Environ.*, vol. 112, no. 9, pp. 3511–3519, 2008.

[24] Z. Wang *et al.*, "Vegetation indices for mapping canopy foliar nitrogen in a mixed temperate forest," *Remote Sens.*, vol. 8, no. 6, p. 491, 2016.

[25] P. A. Townsend, S. P. Serbin, E. L. Kruger, and J. A. Gamon, "Disentangling the contribution of biological and physical properties of leaves and canopies in imaging spectroscopy data," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 12, p. E1074, Mar. 2013.

[26] D. Huang *et al.*, "Canopy spectral invariants for remote sensing and model applications," *Remote Sens. Environ.*, vol. 106, no. 1, pp. 106–122, Jan. 2007.

[27] Y. Knyazikhin, J. V. Martonchik, R. B. Myneni, D. J. Diner, and S. W. Running, "Synergistic algorithm for estimating vegetation canopy leaf area index and fraction of absorbed photosynthetically active radiation from MODIS and MISR data," *J. Geophys. Res., Atmos.*, vol. 103, no. D24, pp. 32257–32275, 1998.

[28] P. Latorre-Carmona, Y. Knyazikhin, L. Alonso, J. F. Moreno, and F. Pla, "On hyperspectral remote sensing of leaf biophysical constituents: Decoupling vegetation structure and leaf optics using CHRIS–PROBA data over crops in Barrax," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 9, pp. 1579–1583, Sep. 2014.

[29] J. L. Silván-Cárdenas and N. Corona-Romero, "Radiation budget of vegetation canopies with reflective surface: A generalization using the Markovian approach," *Remote Sens. Environ.*, vol. 189, pp. 118–131, Feb. 2017.

[30] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, Jul. 1964.

[31] C. Axelsson, A. K. Skidmore, M. Schlerf, A. Fauzi, and W. Verhoef, "Hyperspectral analysis of mangrove foliar chemistry using PLSR and support vector regression," *Int. J. Remote Sens.*, vol. 34, no. 5, pp. 1724–1743, 2013.

[32] X. Li *et al.*, "Exploring the best hyperspectral features for LAI estimation using partial least squares regression," *Remote Sens.*, vol. 6, no. 7, pp. 6221–6241, Jul. 2014.

[33] P. J. Curran, J. L. Dungan, and D. L. Peterson, "Estimating the foliar biochemical concentration of leaves with reflectance spectrometry: Testing the Kokaly and Clark methodologies," *Remote Sens. Environ.*, vol. 76, no. 3, pp. 349–359, Jun. 2001.

[34] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: A basic tool of chemometrics," *Chemometrics Intell. Lab. Syst.*, vol. 58, no. 2, pp. 109–130, 2001.

[35] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, Aug. 2004.

[36] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[37] H. Feilhauer, G. P. Asner, and R. E. Martin, "Multi-method ensemble selection of spectral bands related to leaf biochemistry," *Remote Sens. Environ.*, vol. 164, pp. 57–65, Jul. 2015.

[38] A. Singh, S. P. Serbin, B. E. McNeil, C. C. Kingdon, and P. A. Townsend, "Imaging spectroscopy algorithms for mapping canopy foliar chemical and morphological traits and their uncertainties," *Ecol. Appl.*, vol. 25, no. 8, pp. 2180–2197, Dec. 2015.

[39] S. Smolander and P. Stenberg, "Simple parameterizations of the radiation budget of uniform broadleaved and coniferous canopies," *Remote Sens. Environ.*, vol. 94, no. 3, pp. 355–363, Feb. 2005.

[40] K. M. Vanhatalo, M. Rautiainen, and P. Stenberg, "Monitoring the broadleaf fraction and canopy cover of boreal forests using spectral invariants," *J. Quant. Spectrosc. Radiat. Transf.*, vol. 133, pp. 482–488, Jan. 2014.

[41] Y. Knyazikhin, M. A. Schull, L. Xu, R. B. Myneni, and A. Samanta, "Canopy spectral invariants. Part 1: A new concept in remote sensing of vegetation," *J. Quant. Spectrosc. Radiat. Transf.*, vol. 112, no. 4, pp. 727–735, Mar. 2011.

[42] J. L. Silván-Cárdenas and L. Wang, "Retrieval of subpixel *Tamarix* canopy cover from Landsat data along the Forgotten River using linear and nonlinear spectral mixture models," *Remote Sens. Environ.*, vol. 114, no. 8, pp. 1777–1790, Aug. 2010.

[43] N. Hansen. (Apr. 2016). "The CMA evolution strategy: A tutorial." [Online]. Available: https://arxiv.org/abs/1604.00772

[44] D. A. Roberts, M. Gardner, R. Church, S. Ustin, G. Scheer, and R. O. Green, "Mapping chaparral in the Santa Monica Mountains using multiple endmember spectral mixture models," *Remote Sens. Environ.*, vol. 65, no. 3, pp. 267–279, Sep. 1998.

[45] J.-B. Feret *et al.*, "PROSPECT-4 and 5: Advances in the leaf optical properties model separating photosynthetic pigments," *Remote Sens. Environ.*, vol. 112, no. 6, pp. 3030–3043, 2008.

[46] O. M. Kvalheim, R. Arneberg, O. Bleie, T. Rajalahti, A. K. Smilde, and J. A. Westerhuis, "Variable importance in latent variable regression models," *J. Chemom.*, vol. 28, no. 8, pp. 615–622, Aug. 2014.

[47] E. A. Walter-Shea and J. M. Norman, "Leaf optical properties," in *Photon-Vegetation Interactions: Applications in Optical Remote Sensing and Plant Ecology*, R. B. Myneni and J. Ross, Eds. Berlin, Germany: Springer, 1991, pp. 229–250.

[48] P. J. Curran, "Remote sensing of foliar chemistry," *Remote Sens. Environ.*, vol. 30, no. 3, pp. 271–278, 1989.

[49] F. Castaldi, A. Palombo, F. Santini, S. Pascucci, S. Pignatti, and R. Casa, "Evaluation of the potential of the current and forthcoming multispectral and hyperspectral imagers to estimate soil texture and organic carbon," *Remote Sens. Environ.*, vol. 179, pp. 54–65, Jun. 2016.

[50] I. Herrmann, A. Pimstein, A. Karnieli, Y. Cohen, V. Alchanatis, and D. J. Bonfil, "LAI assessment of wheat and potato crops by VENμS and Sentinel-2 bands," *Remote Sens. Environ.*, vol. 115, no. 8, pp. 2141–2151, Aug. 2011.

[51] P. A. Pellissier, S. V. Ollinger, L. C. Lepine, M. W. Palace, and W. H. McDowell, "Remote sensing of foliar nitrogen in cultivated grasslands of human dominated landscapes," *Remote Sens. Environ.*, vol. 167, pp. 88–97, Sep. 2015.

[52] M. Ewald *et al.*, "LiDAR derived forest structure data improves predictions of canopy N and P concentrations from imaging spectroscopy," *Remote Sens. Environ.*, vol. 211, pp. 13–25, Jun. 2018.

[53] T. P. Dawson, P. J. Curran, P. R. J. North, and S. E. Plummer, "The propagation of foliar biochemical absorption features in forest canopy reflectance: A theoretical analysis," *Remote Sens. Environ.*, vol. 67, no. 2, pp. 147–159, Feb. 1999.

[54] S. L. Ustin *et al.*, "Retrieval of foliar information about plant pigment systems from high resolution spectroscopy," *Remote Sens. Environ.*, vol. 113, pp. S67–S77, Sep. 2009.

[55] M. M. Verstraete, B. Pinty, and R. B. Myneni, "Potential and limitations of information extraction on the terrestrial biosphere from satellite remote sensing," *Remote Sens. Environ.*, vol. 58, no. 2, pp. 201–214, 1996.

[56] M. Disney, "Remote sensing of vegetation: Potentials, limitations, developments and applications BT," in *Canopy Photosynthesis: From Basics to Applications*, K. Hikosaka, Ü. Niinemets, and N. P. R. Anten, Eds. Dordrecht, The Netherlands: Springer, 2016, pp. 289–331.

[57] A. Lo, H. Chernoff, T. Zheng, and S.-H. Lo, "Why significant variables aren't automatically good predictors," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 45, pp. 13892–13897, Nov. 2015.

[58] A. Lo, H. Chernoff, T. Zheng, and S.-H. Lo, "Framework for making better predictions by directly estimating variables' predictivity," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 50, pp. 14277–14282, Nov. 2016.

[59] K. Gränsbo *et al.*, "Chromosome 9p21 genetic variation explains 13% of cardiovascular disease incidence but does not improve risk prediction," *J. Intern. Med.*, vol. 274, no. 3, pp. 233–240, Mar. 2013.

[60] M. D. Ward, B. D. Greenhill, and K. M. Bakke, "The perils of policy by p-value: Predicting civil conflicts," *J. Peace Res.*, vol. 47, no. 4, pp. 363–375, Mar. 2010.

[61] R. S. Kenett and G. Shmueli, "Clarifying the terminology that describes scientific reproducibility," *Nature Methods*, vol. 12, p. 699, Jul. 2015.

[62] P. J. Castaldi, I. J. Dahabreh, and J. P. A. Ioannidis, "An empirical assessment of validation practices for molecular classifiers," *Briefings Bioinf.*, vol. 12, no. 3, pp. 189–202, May 2011.

[63] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf, "Correcting sample selection bias by unlabeled data," in *Proc. 19th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2006, pp. 601–608.

[64] C. Bernau *et al.*, "Cross-study validation for the assessment of prediction algorithms," *Bioinformatics*, vol. 30, no. 12, pp. i105–i112, Jun. 2014.

[65] H. H. Zhou, V. Singh, S. C. Johnson, and G. Wahba, "Statistical tests and identifiability conditions for pooling and analyzing multisite datasets," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 7, pp. 1481–1486, Jan. 2018.

[66] E. Bareinboim and J. Pearl, "Causal inference and the data-fusion problem," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 27, pp. 7345–7352, Jul. 2016.

[67] M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann, "Unsupervised domain adaptation by domain invariant projection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 769–776.

[68] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, May 2015.

[69] P. Köhler, L. Guanter, H. Kobayashi, S. Walther, and W. Yang, "Assessing the potential of sun-induced fluorescence and the canopy scattering coefficient to track large-scale vegetation dynamics in Amazon forests," *Remote Sens. Environ.*, vol. 204, pp. 769–785, Jan. 2018.

[70] N. Altman and M. Krzywinski, "Association, correlation and causation," *Nature Methods*, vol. 12, no. 10, pp. 899–900, Oct. 2015.

[71] D. Lewis, "Causation as influence," *J. Philos.*, vol. 97, no. 4, pp. 182–197, 2000.

[72] S. V. Ollinger, P. B. Reich, S. Frolking, L. C. Lepine, D. Y. Hollinger, and A. D. Richardson, "Nitrogen cycling, forest canopy reflectance, and emergent properties of ecosystems," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 27, p. E2437, Jul. 2013.

[73] G. P. Asner, R. E. Martin, C. B. Anderson, and D. E. Knapp, "Quantifying forest canopy traits: Imaging spectroscopy versus field survey," *Remote Sens. Environ.*, vol. 158, pp. 15–27, Mar. 2015.

[74] N. T. Ilangakoon *et al.*, "Constraining plant functional types in a semi-arid ecosystem with waveform lidar," *Remote Sens. Environ.*, vol. 209, pp. 497–509, Mar. 2018.

[75] E. N. Stavros *et al.*, "ISS observations offer insights into plant function," *Nature Ecol. Evol.*, vol. 1, Jun. 2017, Art. no. 0194.

[76] S. L. Ustin, "Remote sensing of canopy chemistry," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 3, pp. 804–805, Jan. 2013.

[77] J. Stuckens, B. Somers, S. Delalieux, W. W. Verstraeten, and P. Coppin, "The impact of common assumptions on canopy radiative transfer simulations: A case study in *Citrus* orchards," *J. Quant. Spectrosc. Radiat. Transf.*, vol. 110, nos. 1–2, pp. 1–21, Jan. 2009.

[78] B. Pinty, T. Lavergne, R. E. Dickinson, J.-L. Widlowski, N. Gobron, and M. M. Verstraete, "Simplifying the interaction of land surfaces with radiation for relating remote sensing products to climate models," *J. Geophys. Res.*, vol. 111, no. D2, p. D02116, 2006, doi: 10.1029/2005JD005952.

[79] P. R. Moorcroft, G. C. Hurtt, and S. W. Pacala, "A method for scaling vegetation dynamics: The ecosystem demography model (ED)," *Ecol. Monogr.*, vol. 71, no. 4, pp. 557–586, Nov. 2001.

[80] D. Medvigy, S. C. Wofsy, J. W. Munger, D. Y. Hollinger, and P. R. Moorcroft, "Mechanistic scaling of ecosystem function and dynamics in space and time: Ecosystem Demography model version 2," *J. Geophys. Res.*, vol. 114, no. G1, p. G01002, Jan. 2009.

[81] D. M. Lawrence *et al.*, "Parameterization improvements and functional and structural advances in Version 4 of the Community Land Model," *J. Adv. Model. Earth Syst.*, vol. 3, no. 1, p. M03001, 2011, doi: 10.1029/2011MS00045.

[82] P. J. Sellers, "Canopy reflectance, photosynthesis, and transpiration, II. The role of biophysics in the linearity of their interdependence," *Remote Sens. Environ.*, vol. 21, no. 2, pp. 143–183, Mar. 1987.

[83] R. A. Fisher *et al.*, "Vegetation demographics in Earth System Models: A review of progress and priorities," *Global Change Biol.*, vol. 24, no. 1, pp. 35–54, 2018.

[84] L. Yáñez-Rausell *et al.*, "Estimation of spruce needle-leaf chlorophyll content based on dart and paras canopy reflectance models," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 4, pp. 1534–1544, Apr. 2015.

Susan Ustin received the B.S. and M.A. degrees from California State University at East Bay, Hayward, CA, USA, and the Ph.D. degree in botany from the University of California at Davis, Davis, CA, USA.

She is currently a Professor of environmental and resource sciences with the Department of Land, Air and Water resources, University of California at Davis. Her research interests include remote sensing of environment and landscape analysis utilizing optical, microwave, and thermal scanners, radiation interactions in plant canopies, and application to hydrological and ecological problems.

Jessica J. Mitchell received the B.S. degree in geography and environmental systems from the University of Maryland at Baltimore County, Baltimore, MD, USA, and the Ph.D. degree in engineering and applied science–geosciences from Idaho State University, Pocatello, ID, USA.

She is currently the Spatial Analysis Lab Director of the Montana Natural Heritage Program, University of Montana, Missoula, MT, USA. Her research interests include applied remote sensing research with environmental assessment experience to advance new tools for land management decision making and quantitative impact analysis.

Yi Qi received the Ph.D. degree in geography from The University of Utah, Salt Lake City, UT, USA, in 2014.

He is currently an Assistant Professor with the University of Nebraska–Lincoln, Lincoln, NE, USA. His research interests include developing remote sensing and geographical information system methodology to understand vegetation dynamics and interaction with climatic change and ecological disturbance, remote sensing of live fuel moisture content, imaging spectroscopy of plant biochemistry and carbon dioxide, wildfire danger, and fire ecology.

Hamid Dashti received the M.S. degree in remote sensing and geographical information system from Shahid Beheshti University, Tehran, Iran. He is currently pursuing the Ph.D. degree with Boise State University, Boise, ID, USA.

His research interests include canopy radiative transfer modeling, ecosystem modeling, and vegetation parameter estimation from hyperspectral remote sensing.

Nayani T. Ilangakoon received the B.S. degree in geology from the University of Peradeniya, Peradeniya, Sri Lanka, and the M.S. degree in geology from Bowling Green State University, Bowling Green, OH, USA. She is currently pursuing the Ph.D. degree with Boise State University, Boise, ID, USA.

Her research interests include developing full-waveform LiDAR processing algorithms and characterizing canopy structure, plant functional types, and functional diversity.

Nancy F. Glenn received the B.S. degree in geological engineering from the University of Nevada at Reno (UNR), Reno, NV, USA, the M.S. degree in civil engineering from the University of California at Berkeley, Berkeley, CA, USA, and the Ph.D. degree in geoengineering from UNR.

She directs the Boise Center Aerospace Laboratory, Boise, ID, USA, where her research focuses on the soil and vegetation characterization with LiDAR and imaging spectroscopy (hyperspectral) for ecological structure and function and modeling. Her research interests include dryland ecosystems and understanding how these ecosystems respond to the changes in climate and disturbance.

Alejandro N. Flores received the B.S. and M.S. degrees in civil engineering from Colorado State University, Fort Collins, CO, USA, in 2001 and 2003, respectively, and the Ph.D. degree in hydrology from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2009.

He is currently an Associate Professor with the Department of Geosciences, Boise State University, Boise, ID, USA, and the Director of the Lab for Eco-hydrological Applications and Forecasting, Boise, ID, USA. His research interests include terrestrial system modeling, remote sensing, data assimilation, and integrated human–hydrologic systems modeling.

José Luis Silván-Cárdenas received the B.S. degree in computational engineering from the National Autonomous University of Mexico, Mexico City, Mexico, and the Ph.D. degree in geographical information system from Texas State University, San Marcos, TX, USA, in 2009.

His research interests include machine learning, image processing, and pattern recognition.

Lucas P. Spaete received the B.A. degree in program in the environment from the University of Michigan, Ann Arbor, MI, USA, and the M.S. degree in forest ecology and management from the Michigan Technological University, Houghton, MI, USA.

He is currently a Senior Spatial Analyst with Superior Geospatial, Superior, WI, USA. His research interests include remote sensing of vegetation using a combination of optical and LiDAR sensors.

Kaiguang Zhao received the B.S. degree in physics from Beijing Normal University, Beijing, China, and the Ph.D. degree in forestry from Texas A&M University, College Station, TX, USA, in 2008.

He is currently an Assistant Professor with Ohio State University, Columbus, OH, USA. His research interests include mapping, monitoring, modeling, and managing terrestrial environments across scales, especially in the context of global environmental changes.

Marie-Anne de Graaff received the B.S. degree in forestry and nature management and the Ph.D. degree in environmental science from Wageningen University, Wageningen, The Netherlands.

She is currently an Associate Professor with Boise State University, Boise, ID, USA. Her lab broadly studies how changes in climate and land use affect ecosystem processes that drive the global carbon cycle.