Remote Sensing of Environment 115 (2011) 1978–1996

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Characterizing forest canopy structure with lidar composite metrics and machine learning

Kaiguang Zhao [a,*], Sorin Popescu [b], Xuelian Meng [c], Yong Pang [d], Muge Agca [b]

[a] Center on Global Change & Dept. of Biology, Duke University, Durham, NC 27708, United States
[b] Spatial Sciences Lab., Dept. of Ecosystem Science and Management, Texas A&M University, College Station, TX 77450, United States
[c] Department of Civil and Environmental Engineering and Geodetic Science, The Ohio State University, Columbus, OH 43210, United States
[d] Institute of Forest Resource and Information Technology, Chinese Academy of Forestry, Beijing 10091, China

ARTICLE INFO

Article history:
Received 6 January 2011
Received in revised form 22 March 2011
Accepted 2 April 2011
Available online 12 May 2011

Keywords:
Lidar
Laser scanner
Canopy
Biomass
Carbon
Machine learning
Gaussian process
Support vector machine
Forest fuel

ABSTRACT

A lack of reliable observations for canopy science research is being partly overcome by the gradual use of lidar remote sensing. This study aims to improve lidar-based canopy characterization with airborne laser scanners through the combined use of lidar composite metrics and machine learning models. Our so-called composite metrics comprise a relatively large number of lidar predictors that tend to retain as much information as possible when reducing raw lidar point clouds into a format suitable as inputs to predictive models of canopy structural variables. The information-rich property of such composite metrics is further complemented by machine learning, which offers an array of supervised learning models capable of relating canopy characteristics to high-dimensional lidar metrics via complex, potentially nonlinear functional relationships. Using coincident lidar and field data over an Eastern Texas forest in USA, we conducted a case study to demonstrate the ubiquitous power of the lidar composite metrics in predicting multiple forest attributes and also illustrated the use of two kernel machines, namely, support vector machine and Gaussian processes (GP). Results show that the two machine learning models in conjunction with the lidar composite metrics outperformed traditional approaches such as the maximum likelihood classifier and linear regression models. For example, the five-fold cross validation for GP regression models (vs. linear/log-linear models) yielded a root mean squared error of 1.06 (2.36) m for Lorey's height, 0.95 (3.43) m for dominant height, 5.34 (8.51) m^2/ha for basal area, 21.4 (40.5) Mg/ha for aboveground biomass, 6.54 (9.88) Mg/ha for belowground biomass, 0.75 (2.76) m for canopy base height, 2.2 (2.76) m for canopy ceiling height, 0.015 (0.02) kg/m^3 for canopy bulk density, 0.068 (0.133) kg/m^2 for available canopy fuel, and 0.33 (0.39) m^2/m^2 for leaf area index. Moreover, uncertainty estimates from the GP regression were more indicative of the true errors in the predicted canopy variables than those from their linear counterparts. With the ever-increasing accessibility of multisource remote sensing data, we envision a concomitant expansion in the use of advanced statistical methods, such as machine learning, to explore the potentially complex relationships between canopy characteristics and remotely-sensed predictors, accompanied by a desideratum for improved error analysis.

1. Introduction

Lidar remote sensing of canopies far goes beyond the proof-of-concept phase and thus far has become a well-established tool for monitoring terrestrial ecosystems, with great potential for continued technological advances (Reutebuch et al., 2005). Commercial airborne lidar systems currently available, for example, can easily reach a pulse repetition frequency of more than 100 kHz and some are even able to simultaneously manipulate multiple pulses in the air. Compared to traditional remote sensing such as multispectral imaging and radar, lidar (especially small-footprint airborne laser scanner) provides superior capabilities for accurately measuring vegetation structure with no apparent sign of saturation in predicting high biomass and leaf area index (LAI) over dense forests (Koch, 2010). Such advantages spur the increasing availability and widespread use of lidar data in canopy studies, which, on the other hand, connotes the growing importance of reliable vegetation information in practical and scientific applications such as natural resource inventory and more interestingly, studies of vegetation responses and functioning in a changing climate (Hurtt et al., 2004). An early review of basic lidar principles in the context of forestry applications is given in Lim et al. (2003b).

* Corresponding author. Tel.: +1 979 739 9981.
E-mail address: lidar.rs@gmail.com (K. Zhao).

0034-4257/$ – see front matter
doi:10.1016/j.rse.2011.04.001

The recent use of lidar for characterizing forest landscapes encompasses a wide array of applications targeted at estimating various vegetation structural parameters across a range of scales (van Leeuwen & Nieuwenhuis, 2010). These biophysical parameters include, but are not limited to, individual tree dimension variables such as height, crown width, crown base height, crown volume, and diameter at breast height (dbh) (Coops et al., 2004; Popescu & Wynne, 2004; Falkowski et al., 2006; Kato et al., 2009; Popescu & Zhao, 2008), and at the canopy level, canopy height, fractional vegetation cover, LAI, total aboveground biomass and component biomass, basal area, timber volume, and canopy fuel parameters such as canopy base height, available canopy fuel, canopy bulk density and coarse wood debris (Andersen et al., 2005; Holmgren et al., 2003; Jensen et al., 2008; Næsset, 2002; Zhao et al., 2009). Meanwhile, some studies used lidar to examine ecological roles of vegetation for various purposes. For example, Quirino et al. (2009) assessed the power of lidar-derived variables for directly predicting soil respiration at several sites in a Virginia forest. Cook et al. (2009) modeled plant productivity through the combined use of vegetation variables derived from lidar and Quickbird. A recent study by Falkowski et al. (2009) successfully classified forest succession stages in a complex, mixed coniferous forest using lidar metrics. There also are several lidar studies that derived useful indices for mapping wildlife habitat and evaluating habitat suitability, especially for many endangered species (Hyde et al., 2006). Overall, past experiences with forestry applications accentuate the great potential of lidar for remotely measuring canopy structures (Koch, 2010; van Leeuwen & Nieuwenhuis, 2010). Moreover, part of these studies clearly articulate the virtues of integrating lidar with other remote sensing data such as multispectral images (Erdogy & Moskal, 2010), high-resolution (Mutlu et al., 2008) or hyperspectral images, and radar backscattering data (Hyde et al., 2006), although a few studies point out that for estimating some vegetation variables like LAI, the gain from the combined use of lidar and multispectral data is only marginal compared to the use of lidar alone (Jensen et al., 2008; Zhao & Popescu, 2009), possibly due to the insufficiency of statistical models used (Zhao & Popescu, 2009).

Regarding the processing and analysis of lidar data for canopy information extraction, lidar users generally keep two aprons, one laden with computational tools borrowed from disciplines such as signal/image processing and pattern recognition (Falkowski et al., 2006; Kato et al., 2009; Meng et al., 2009; Pang et al., 2008; Popescu & Zhao, 2008), and another apron stocked with statistical analysis tools for inferring relationships and making prediction (Andersen et al., 2005; García et al., 2010; Næsset et al., 2005; van Aardt et al., 2006; Zhao et al., 2009). The boundary between the two classes of tools is not dichotomous and often, their use is complementary as has been previously demonstrated (Popescu et al., 2002). A typical example within the first class of tools is segmenting lidar point clouds or more frequently, canopy height models for individual tree crown delineation, of which maximum filtering, template matching, k-means clustering, and the watershed transform are common algorithms (Falkowski et al., 2006; Pang et al., 2008; Popescu et al., 2003). Other studies customized computer algorithms to directly derive crown base height and dbh of individual trees from lidar point clouds (Kato et al., 2009; Popescu & Zhao, 2008). In contrast, the vast majority of lidar research on canopy structural characteristics is built upon the use of statistical tools taken from the second apron, and of particular importance therein is not only the selection of appropriate statistical procedures (e.g., models) but more important, the choice of effective lidar predictors (i.e., metrics) (Lefsky et al., 2005; McRoberts et al., 2010; Næsset, 2002). Typically, lidar metrics from discrete-return data are obtained in a bookkeeping manner by first enumerating all the points within an analysis unit (e.g., grid, plot, or stand) to select part or all of the lidar points and then computing the desired statistics as metrics from the selected lidar points (Næsset et al., 2005; Reutebuch et al., 2005). Common lidar metrics include mean, median, maximum and percentile heights, canopy density metrics, truncated mean height, quadratic mean height, variance in height, coefficient of variation, etc. (Erdogy & Moskal, 2010; Lim & Treitz, 2004; Næsset et al., 2005). These metrics are not equally predictive for estimating a given canopy variable; then, statistical procedures come into play to determine which metrics should enter models. This paradigm is typical of many lidar applications in canopy studies (Næsset, 2002).

The large data volume and high dimensionality of lidar point clouds preclude their direct use as input to statistical models, thereby stressing the necessity to reduce and synthesize raw lidar data into various metrics at a spatial resolution commensurate with the analysis unit (Andersen et al., 2005; García et al., 2010; Reutebuch et al., 2005). In such a sense, the conversion of lidar point clouds into metrics can be considered as some encoders that, on one hand, transform raw lidar data into a consistent format reconciliatory with statistical models and on the other hand, reduce the dimensions of lidar point data to a manageable degree. Apparently, this encoding process involves two conflicting goals — maintaining a parsimony of lidar metrics versus preserving as much information inherent in the raw lidar data as possible. To select metrics for estimating a canopy variable, one usually follows practical guides such as the physical linkage of metrics to the variable in question, the avoidance of multi-collinearity, and the optimization of certain statistical measures like R-squared, the Akaike information criterion and the Bayesian information criterion (García et al., 2010; Lim & Treitz, 2004; Roberts et al., 2005). Recently, Zhao et al. (2009) proposed the use of lidar-derived canopy height profile (CHP) in functional models to estimate forest biomass. In analogy to a hyperspectral spectrum, this metric is a discretized curve representing the vertical distribution of lidar heights over canopies and thus, it tends to retain more information of raw lidar data than many commonly used lidar metrics. However, due to its high-dimensionality and the inherent correlation between its components, CHP needs to be used in dedicated models like the functional model of Zhao et al. (2009). Alternatively, advanced supervised learning techniques such as machine leaning may be referred because of their usefulness in tackling high-dimensional problems, as confirmed by earlier studies pertaining to hyperspectral/multi-angular remote sensing (Durbha et al., 2007; Zhao et al., 2008).

The superior performances of machine learning over classical methods have been demonstrated in a preponderance of comparative research for general remote sensing applications (Durbha et al., 2007; van der Heijden et al., 2004; Zhao et al., 2008). For the past few decades, machine learning has been an active research field yielding a rich set of computer tools aimed to discover patterns and relationships in data (Evgeniou et al., 2000). For example, recent advances in kernel machines promote the novel use of Gaussian processes for Bayesian-based supervised learning (Rasmussen & Williams, 2006; Zhao et al., 2008). All these machine learning tools, which continue to evolve and expand, generally provide more powers in capturing the implicit, potentially nonlinear and complex relationships between dependent and independent variables (Evgeniou et al., 2000), that is, a problem typical of many information extraction tasks in remote sensing (Næsset et al., 2005), but thus far, the use of machine learning for estimating canopy structural variables in lidar remote sensing remains rather limited (Zhao et al., 2009).

The overall purpose of this study is to explore effective methods for characterizing canopy structure using small-footprint discrete-return lidar. To this end, we attempt to implement two specific objectives, i.e., (1) to derive comprehensive lidar metrics that can serve as versatile predictors for various canopy characteristics, and (2) to investigate the usefulness of machine learning techniques for estimating canopy characteristics from the metrics derived in (1). In particular, for the first objective, we introduced a lidar canopy density composite metric as well as its other equivalent forms (e.g., lidar quantile composite metric) in hopes to take better advantage of information inherent in the raw lidar point data. Each of the composite metrics contains a number of components and thus has a

large dimension. For the second objective, we considered two popular machine learning techniques, i.e., support vector machines (SVM) and Gaussian processes (GP). Both techniques are kernel machines that resort to the "kernel" trick for circumventing the curse of dimensionality. To examine the utility of the proposed metrics and kernel learning methods, lidar data acquired over a temperate forest in East Texas, USA, were used in conjunction with coincident field measurements to estimate a range of canopy structural variables.

2. Materials

2.1. Study area

Our study area is a forested region in the eastern half of Texas, the southern United States (30° 42′ N, 95° 23′ W), covering approximately 4800 ha. The region mainly comprises pine plantations of various developmental stages, old growth pine stands in the Sam Houston National Forest, many of them with a natural pine stand structure, and upland and bottomland hardwoods. Much of the southern U.S. has forest types similar to the ones included in our study area, with similar forest types, productivity, and patterns of land use change. A mean elevation of 85 m, with a minimum of 62 m and a maximum of 105 m, and gentle slopes characterize the topography of the study area. In addition, a 2.5-meter-resolution land cover map is available that is used to mainly differentiate pines, hardwoods and mixed forests, and this map was derived from a Quickbird scene acquired in February, 2004, with a maximum likelihood classifier.

2.2. Field inventory data

Field inventory data were collected on 58 circular plots over the study area in May–June, 2004. These include 36 plots of 0.1 ha in size and 22 plots of 0.01 ha. The 0.01-ha plots were located in very dense and uniform young pine stands while the 0.1-ha plots were located in stands of relatively complex structure. A total of 1004 trees were tallied with respect to height, crown width, height to crown base, dbh, species, and crown class (Kraft). The protocol for field mapping of individual trees and the comparison to airborne lidar-measured trees are described in great detail in Popescu and Zhao (2008).

2.3. Airborne lidar data

The lidar data were acquired with a Leica-Geosystems ALS40 during the leaf-off season in early March 2004 from an average altitude of 1000 m above the ground level. In this flight campaign, the lidar system was configured to scan a swath ±10° from nadir, and to record up to two returns per laser pulse, i.e., first and last. The reported horizontal and vertical accuracies are 20–30 cm and 15 cm, respectively. The entire study area was covered by a cross-hatch grid of flight lines, with 19 and 28 flight lines in a north–south direction and east–west direction, respectively. On average, the swath width is 350 m and the laser point density is 2.6 per m^2, translating to an average distance between laser points of about 0.62 m. A digital elevation model (DEM) at a 2-m spatial resolution was derived by our data vendor with its proprietary package. The ground elevation underlying any given lidar point was calculated by bi-linearly interpolating the DEM, and the entire point cloud was registered relative to the ground by subtracting ground elevations from the original lidar above-sea-level elevations such that ground echoes are zero in height.

3. Methodological background

3.1. Canopy structure characterization

Canopy structure represents the spatial and temporal organization of such aboveground vegetative elements as foliage, stems, twigs, and branches. To conceptualize structural complexities for canopies of diverse forms, simplified representations have to be sought (Fig. 1), as exemplified by the routine use of structural variables such as canopy height, LAI, canopy cover, biomass density, stem density, and basal area for quantifying the average canopy conditions at scales of interest (Parker, 1995; Jennings et al., 1999). Field-based measurements of forest canopies can be made at various levels of detail ranging from a single leaf to a landscape, with both direct and indirect methods possible (Parker, 1995). Direct methods are preferred when the variable of interest is easily measurable (e.g., dbh of trees using a diameter-tape) or when a relatively high accuracy is needed (e.g., destructive sampling for weighing biomass). Indirect field methods usually involve the use of statistical equations or physically-based models (Englund et al., 2000). For example, as will be illustrated later in this paper, tree biomass can be estimated from dbh through existing allometric equations (Jenkins et al., 2003), and LAI can be retrieved with hemispherical photography based on some light–canopy interaction principles (Englund et al., 2000). In addition, common practice requires sampling procedures/protocols in order to extend individual measurements at sub-unit levels to an overall estimate at a higher level of analysis unit (e.g., plots, stands or even up to geographic regions) (Schreuder et al., 1993). For instance, tree-level attributes measured on subplots as in the plot design of forest inventory analysis could be converted to a suite of canopy fuel characteristics such as canopy height, canopy base height (CBH) and canopy bulk density (CBD) (Reinhardt et al., 2006), as will be detailed later. Also, observations of sight occlusion along a transect or at random points can be translated to canopy cover estimates with appropriate design-based estimators (Englund et al., 2000). Furthermore, field-based measurements are generally combined with remote sensing-based predictors to generate spatially-explicit maps of canopy characteristics over large regions (McRoberts et al., 2010), as demonstrated in the following by using lidar-based predictors (Fig. 1).

3.2. Lidar metrics

Lidar metrics, as surrogates for canopy structural variables, are simply some descriptive statistics of laser ranging data (Fig. 1). Theoretically, there can be an infinite number of possible metrics derivable from a given subset of lidar points. The forms of possible metrics also depend on lidar data characteristics, such as how many returns are recorded per pulse and whether or not auxiliary information such as intensity and scan angle is available (Zhao et al., 2009). For example, given lidar intensity information, it is possible to compute thresholded height metrics based on only those lidar points with intensity above a prescribed threshold, as used in Lim et al. (2003a). In forest canopy studies, typical lidar metrics that have been previously used include canopy densities, mean and percentile heights, second-order height statistics (Hudak et al., 2009; Lim & Treitz, 2004) and even some composite metrics such as the canopy height distribution and the canopy quantile function of Zhao et al. (2009).

Algorithmically, most lidar metrics can be computed by two equivalent means. The first is to literally implement the relevant computation and the second is to indirectly refer to an underlying empirical lidar height distribution (Zhao et al., 2009). For example, mean lidar height $\bar{h}_{all}$ and variance of all returns s^2_{all} can be computed either explicitly by

$$\bar{h}_{all} = \sum_{i=1}^{N} z_i / N, \tag{1}$$

$$s^2_{all} = \sum_{i=1}^{N} \left(z_i - \bar{h}_{all}\right)^2 / N$$

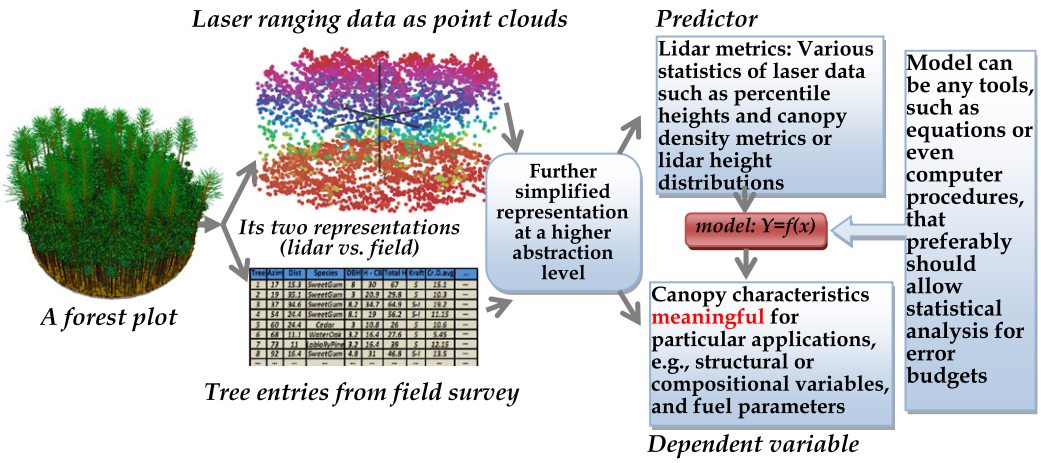

Fig. 1. A typical workflow for canopy characterization with a grid-based lidar approach: Canopy characteristics synthesized from field survey data are linked to lidar-derived metrics through various predictive models (see Section 3).

where N is the total number of lidar returns within the analysis unit (i.e., a pixel or a patch) and z_is are heights of individual returns, or alternatively by

$$\bar{h}_{all} = \int h \cdot p(h) \cdot dh \approx \sum_{i=0}^{N_{bin}-1} h_i \cdot p(h_i) \cdot \Delta h, \quad (2)$$

$$s^2_{all} = \int \left(h - \bar{h}_{all}\right)^2 \cdot p(h) \cdot dh \approx \sum_{i=0}^{N_{bin}-1} \left(h_i - \bar{h}_{all}\right)^2 \cdot p(h_i) \cdot \Delta h$$

where $p(h)$ refers to a lidar height distribution function denoting the probability of finding lidar points around height h. In practice, its numerical value $p(h_i)$ needs to be obtained by counting the fraction of lidar points in the height interval $[h_i - \Delta h/2,\ h_i + \Delta h/2]$ with $h_i = i \cdot \Delta h$. Herein Δh is a user-defined size chosen to divide the vertical height range into N_{bin} bins/intervals, assuming that lidar points have been topographically detrended by some ground-filtering algorithms so that ground echoes are zero-valued in height (Meng et al., 2009). As another example, the popular lidar percentile-height metrics can be computed in the above two fashions, that is, either by directly sorting lidar points with respect to height or by indirectly relying on the empirical height distribution $\{p(h_i)\}|_{i=0}^{N_{bin}-1}$ of Eq. (2). It becomes clear that the indirect option is more appealing when the analysis unit contains a myriad of lidar points that make the direct sorting computationally expensive.

3.3. Statistical models

Lidar metrics need to be linked with field-based canopy variables to establish predictive frameworks (Fig. 1). Despite their popularity in common practice, parametric statistical models only represent a restrictive class of functions and often lack flexibilities in capturing the unknown relationship f underlying a given dataset $[\mathbf{x}_i, y_i]$, $i = 1,\ldots, n$. For general tasks of supervised learning, recent years have witnessed the widespread use of nonparametric machine learning models for implicitly inferring f from $[\mathbf{x}_i, y_i]_{i=1,\ldots,n}$ (Rasmussen & Williams, 2006). The applications of such machine learning models are often found under the guise of pattern recognition, data mining and artificial intelligence (Durbha et al., 2007; Rasmussen & Williams, 2006). As opposed to linear regression, most of these models are versatile enough to discover complicated nonlinear relationships. Besides, some of them are useful in alleviating or even circumventing the difficulties caused by the curse of dimensionality (Durbha et al., 2007; Rasmussen & Williams, 2006; Zhao et al., 2008), which is a common problem in fitting models with a large number of predictors, as is the case of the present study. In particular, this study examined two widely used machine learning models, namely, SVM and GP, which will be detailed later in Section 5.

4. Lidar composite metric

As seen in Eqs. (1) and (2), the use of lidar height distribution to indirectly calculate common metrics unarguably highlights its important roles in lidar-based canopy characterization. To make it more explicit, the lidar height distribution $p(h)$ or its numerical version $\{p(h_i)\}|_{i=0}^{N_{bin}-1}$ retains information essential for calculating most of the commonly used metrics, such as mean height, quantile heights, and coefficient of variation, given that a relatively small height bin Δh is used to ensure numerical accuracy. This information-rich property of $p(h)$ serves as one fundamental argument elicited by Zhao et al. (2009) to substantiate their choice of CHP as predictor for aboveground biomass. However, the CHP metric of Zhao et al. (2009) was computed only using either first returns or lidar-derived canopy height models. More generally, it is expected that multiple lidar-derived canopy height distributions, each associated with a particular type of lidar returns, could provide an improved characterization of the vertical distributions of lidar points. Our lidar dataset, for example, features up to two returns per pulse: First returns most likely reflect from canopy tops and last returns from below the canopy top so that they characterize different layers of the canopy (Fig. 2a). As a result, rather than using only a single height profile of all returns $p(h)$, the combined use of two CHPs (i.e., $p_f(h)$ and $p_l(h)$), one for first returns and another for last returns, has potential to enhance the utility of lidar in depicting canopies (Fig. 2b). However, because the derivation of the two CHPs is done separately, the inherent information concerning the topological correspondence between first and last returns is missed and not captured by the two CHPs. To recover such missing linkage, we next propose several new composite metrics that partially account for the penetration process of laser pulses (Fig. 2c–f).

4.1. Canopy density metric

The thrust for our newly proposed lidar metrics is to encode more information of raw lidar data into them, though at the expense of a higher metric dimension (Fig. 2). This is in the same spirit as the CHP profile metric of Zhao et al. (, 2009). Realizing that earlier research rarely took account of the lidar information at the individual pulse level, we extend the CHP metric of Zhao et al. (2009) to differentiate first and last returns and further consider the topological correspondence between the two types of returns (i.e., the ties between first and

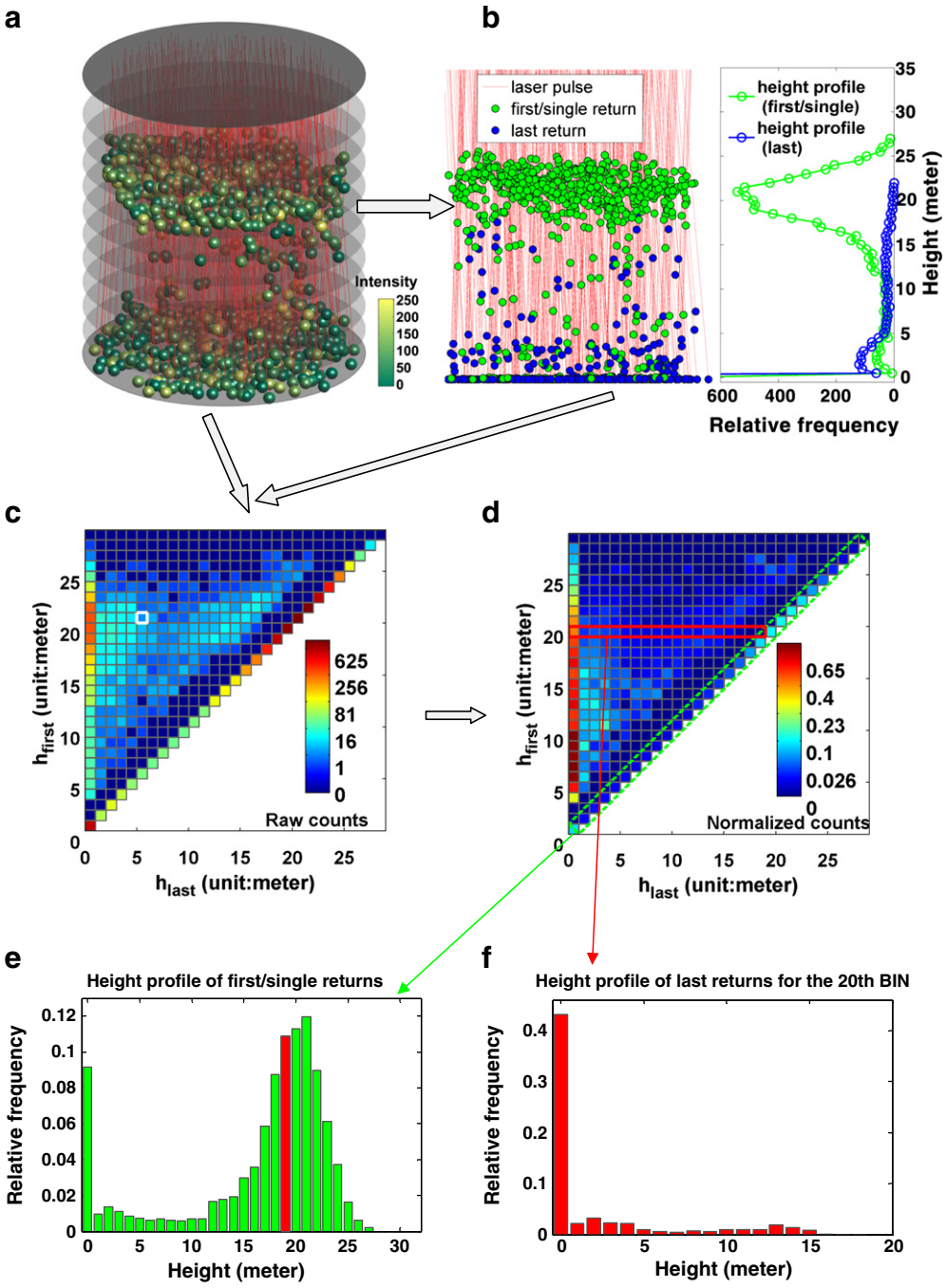

Fig. 2. A schematic explaining the derivation of lidar composite metrics from raw laser point clouds. Notice how the original information associated with raw lidar points is being progressively discarded from top to bottom for the convenience of "encoding": (a) a full 3D rendering of lidar point cloud with red, thin lines depicting the penetration paths of laser pulses, and the color ramp indicating echo intensity. (b) A 2D side-view of the same lidar point cloud, along with two vertical profiles of lidar height distributions for first/single returns (green) and last returns (blue), respectively. (c) A matrix representation of lidar point counts where rows and columns denote the height bins for first/single and last returns, respectively, and the count denotes the number of laser returns that fall into certain height bins (see Section 4.1 for more details). (d) The same as (c) except that the non-diagonal elements of each row are normalized with respect to the diagonal element (i.e., the last element of each row) and that the diagonal elements are normalized with respect to the total count of first/single returns. (e) The profile of the normalized diagonal of (d) as highlighted by the green dash-line rectangle. (f) An example of the normalized row profile as highlighted by the red solid-line rectangle of (d).

last returns of a pulse as implied by the red lines in Fig. 2a and b). Such correspondence information can be inferred implicitly from the order in which lidar returns are written to an ASCII or LAS file (personal communication, Dr. Ross Nelson at NASA). Our lidar data contain three types of returns, i.e., first, last and single, of which single returns correspond to pulses that have only one return. For simplicity, single and first returns are combined together when counting point occurrence to derive the metrics (i.e., first/single in our latter presentation). As in most studies, we consider only the height dimension of a lidar point z_i and abandon other information such as horizontal xy coordinates and intensity value.

We derive the new metric by counting the vertical distribution of first/single returns and additionally counting the occurrence of last returns with respect to both their heights relative to the ground and their penetration depths from the associated first returns (Fig. 2c). Specifically, the occurrence of first/single returns is counted for each

height-bin $[(i\text{-}1/2)\cdot\Delta h,(i+1/2)\cdot\Delta h]$ with $i=0,1,\ldots,N_{max}$ where Δh is again the bin size and N_{max} is the number of bins. For reference, the count of first/single returns for the ith bin is denoted by $n_i^{fs}=n_i^f+n_i^s$ where the superscripts, fs, f and s, are not exponents but just symbols for labeling return types. Further, because a last return is uniquely associated with a first return, all last returns can be correspondingly divided into N_{max} distinct groups according to the bin indices of their associated first returns. It apparently holds that the number of last returns in the ith group (n_i^l) equals that of first returns in the ith bin (n_i^f), viz. $n_i^l=n_i^f, i=0,1,\ldots,N_{max}$. Also, the ith group of last returns is all confined to a truncated height range $[0,h_i+\Delta h/2]$ because in height, no last returns exceed their associated first returns. Subsequently, the vertical distribution of last returns in the ith group is derived by counting their occurrence along $[0,h_i+\Delta h/2]$. To implement this counting, another bin size δh needs to be specified to partition the range $[0,h_i+\Delta h/2]$ into a series of individual bins. Note that δh is not necessarily the same as the bin size Δh that is used to count first/single returns.

Obviously, the resulting metric described above is not single-valued but instead is a composite that includes a list of components, each representing a count of lidar points that meet certain criteria or fall into a specified height bin (Fig. 2c). The total number of components in this metric depends on the height range chosen for deriving lidar height distributions as well as the height-bin sizes (Δh and δh) chosen for discretizing this height range. In addition, we can organize the components of the composite metric neatly into a matrix format, as depicted in Fig. 2c where approximately half of the matrix elements are occupied and the rest are empty and invalid. In this matrix, the ith row (the origin is at the left bottom) is made up of both the count of the first/single returns for the ith bin and a number of counts of last returns for the associated ith group of last returns. To be more specific, in the ith row, the rightmost nonempty element records the count of first/single returns in the ith height bin $[(i\text{-}1/2)\cdot\Delta h,(i+1/2)\cdot\Delta h]$ (e.g., the red-filled bar in Fig. 2e); the remaining nonempty elements characterize the vertical distribution for the ith group of last returns, with the element at the jth column being the count of those last returns that are within the height bin $[(j-1/2)\cdot\delta h,\ (j+1/2)\cdot\delta h]$ (Fig. 2f). Note that $j<ceiling[(i+1/2)\Delta h/\delta h]$ where the *ceiling*[] is the function to fetch the nearest upper integer. It is reminded again that the ith group of last returns are those that have their associated first returns falling into the ith height bin $[(i\text{-}1/2)\cdot\Delta h,(i+1/2)\cdot\Delta h]$.

To better serve as predictors, the counts in the aforementioned composite metric or matrix need to be properly normalized (Fig. 2c and d). The most straightforward way should be to directly divide the counts by the number of all lidar returns. Alternatively, a more interesting way is to normalize the counts of first/single and last returns separately, that is, the counts for first/single returns are normalized with respect to the total number of first/single returns while the counts of last returns in ith group are normalized with respect to the total number of first/single returns of the ith bin. After normalization, the resulting metric values become fractions that gain some probability interpretation. The normalized composite metric can be viewed as a detailed set of canopy density predictors and thus provides a consistent, systematic way to depict the vertical patterns of lidar returns as well as canopy vertical structure (Fig. 2c and d). Symbolically, the normalized metric can be denoted by

$$\mathbf{P}\left\{h_l,h_{fs}\right\}=\left\{p\left(h_{fs}\right),p\left(h_l|h_{fs}\right)\right\} \tag{3}$$

where $p(h_{fs})$ denotes the probability of observing first/single returns around height h_{fs}, and $p(h_l|h_{fs})$ denotes the probability of finding last returns around height h_l, given that the associated first returns occur at height h_{fs}. In the subsequent analysis of this study, we will resort to the use of the second normalization scheme.

4.2. Quantile height metric or hybrid metric

The composite metric of lidar point density described above, $\mathbf{P}\{h_l,\ h_{fs}\}$, can be converted to an equivalent composite metric of quantile heights $\mathbf{Q}\{p_l,\ p_{fs}\}$ without any loss of information. Such a conversion is guaranteed by the mathematical equivalence of a probability density function to its quantile function, and the same logic has also been adopted in Zhao et al. (2009) that demonstrated the equivalence of canopy density and quantile height metrics in predicting aboveground biomass. Due to the separate treatment of first/single and last returns in the density metric $\mathbf{P}\{h_l,h_{fs}\}$, its equivalent quantile height metric $\mathbf{Q}\{p_l,p_{fs}\}$ should accordingly handle first/single and last returns differently, i.e.,

$$\mathbf{Q}\left\{p_l,p_{fs}\right\}=\left\{q\left(p_{fs}\right),q\left(p_l|p_{fs}\right)\right\} \tag{4}$$

where $0\le p_l\le 1$; $0\le p_{fs}\le 1$; $q(p_{fs})$ is converted equivalently from the density $p(h_{fs})$ and it represents the p_{fs} quantile height of first/single returns; $q(p_l|p_{fs})$ is converted from $p(h_l|h_{fs})$ and it is the p_l quantile height of last returns given the associated first returns hitting at the height of $q(p_{fs})$.

Moreover, because the lidar point density metric $\mathbf{P}\{h_l,h_{fs}\}$ contain multiple components (i.e., $p(h_{fs})$ and $p(h_l|h_{fs})$), it is possible to select only a portion of components to be converted to the forms of quantile height function or even cumulative distribution function and keep the remaining components intact as lidar canopy densities. As a result, we arrive at a variety of hybrid composite metrics that mathematically should be all equivalent to $\mathbf{P}\{h_l,\ h_{fs}\}$ or $\mathbf{Q}\{p_l,\ p_{fs}\}$ in terms of the information content. For example, one possible hybrid composite metric can be symbolized as

$$\mathbf{QP}\left\{p_{fs},h_l\right\}=\left\{q\left(p_{fs}\right),p\left(h_l|q\left(p_{fs}\right)\right)\right\} \tag{5}$$

which means that first/single returns are characterized by a quantile function $q(p_{fs})$ and last returns by conditional density distributions $p(h_l|q(p_{fs}))$.

In practice, numerical values of the lidar composite metrics inevitably need to be computed from lidar data at a discrete set of heights or quantiles. This has been demonstrated by the counting of lidar point occurrence within a given height bin when we introduce $\mathbf{P}\{h_l,h_{fs}\}$. Likewise, a numerical representation of $\mathbf{Q}\{p_l,p_{fs}\}=\{q(p_{fs}),\ q(p_l|p_{fs})\}$ could be obtained at a discrete set of quantiles, namely $q(p_{fs,\ i})|_{i=1,\ldots,\ N_{fs}}$ and $q(p_{l,\ j}|p_{fs,\ i})|_{i=1,\ldots,\ N_{fs};\,j=1,\ldots,\ Nl,i}$ where N_{fs} and $N_{l,\ i}$ are the numbers of quantile intervals used to discretize the quantile height functions. Due to such discretization, the equivalence of different composite metrics does not strictly hold, but is subject to some numerical truncation errors. In practice, the discretized versions of $\mathbf{P}\{h_l,\ h_{fs}\}$, $\mathbf{Q}\{p_l,\ p_{fs}\}$ or $\mathbf{QP}\{p_{fs},\ h_l\}$ may contain dozens, hundreds or even thousands of components, depending on the desired level of detail at which lidar height distributions are quantified. For ease of reference, we call components of the composite metrics as predictors in our latter presentation, and hence, our new lidar density or height metric is typically high-dimensional and consists of a number of predictors.

5. Machine learning models

SVM and GP, known as kernel machines, have become standard procedures for supervised learning, especially nonlinear regression and classification problems. The successful applications of SVM in remote sensing have been well documented (Durbha et al., 2007) whereas only several studies have investigated the utility and effectiveness of GPs for remote sensing classification problems such as classifying hyperspectral images (e.g., Zhao et al., 2008). In terms of model formulation, several heuristic rules exist to inspire the

derivation of SVM and GP. For example, in its simplest form, the binary SVM classifier can be intuitively motivated by searching for a separating hyperplane that yields the maximum margin between the two classes in a transformed feature space (Fig. 3a). In contrast, GP tends to model the unknown function f with a posterior mean GP that is obtained by updating prior GPs in light of the observed data (Fig. 3b). It is beyond the scope of this study to delve into their mathematical details, but for completeness, we briefly present a general, unified formulation of SVM and GP from a regularization perspective. More theoretical and technical details as well as the performance comparison of SVM and GPs for classification are referred to Rasmussen and Williams (2006) and Zhao et al. (2008). Due to the "black-box" nature of the two techniques for most practical use, readers, if not interested, may skip this section without too many losses in understanding our results.

5.1. Formulation of SVM and GP

Like most other regression techniques, learning a kernel machine model such as SVM or GP from a training sample $[\mathbf{x}_i, y_i]_{i=1,\ldots,n}$ is to seek an implicit function $f(\mathbf{x})$ that minimizes the difference between the observed and predicted values of the target variable while at the same time requiring that $f(\cdot)$ should generalize well to predictions at new out-of-sample data points. Mathematically, these two goals can be combined and translated into the minimization of the following functional with respect to the function f (Evgeniou et al., 2000):

$$J[f] = L(\mathbf{y}, \mathbf{f}) + \frac{C}{2}||f||^2_{\mathcal{H}} \tag{6}$$

where the first term on the right is a loss function that defines the cost incurred by the deviation of the predicted target values $\mathbf{f} = [f(\mathbf{x}_1),\ldots, f(\mathbf{x}_n)]^T$ from the observed ones $\mathbf{y} = [y_1,\cdots, y_n]^T$; the second term is called a regularizer that quantifies the smoothness of function f, with smoother f yielding lower values of $||f||^2_{\mathcal{H}}$; C is a scalar parameter that controls the tradeoff between the two terms, and its value can be either predetermined *a priori* or tuned from training samples (Rasmussen & Williams, 2006). Note that a functional (e.g., $J[f]$) can be loosely thought as a function of functions. In what follows, we separately explicate the two terms of Eq. (6).

$L(\mathbf{y}, \mathbf{f})$ represents the sum of individual losses caused by the mismatch between observed and predicted values, i.e., $L(\mathbf{y}, \mathbf{f}) = \sum l(y_i, f_i)$. SVM and GPs choose different functional forms for $l(\cdot,\cdot)$,

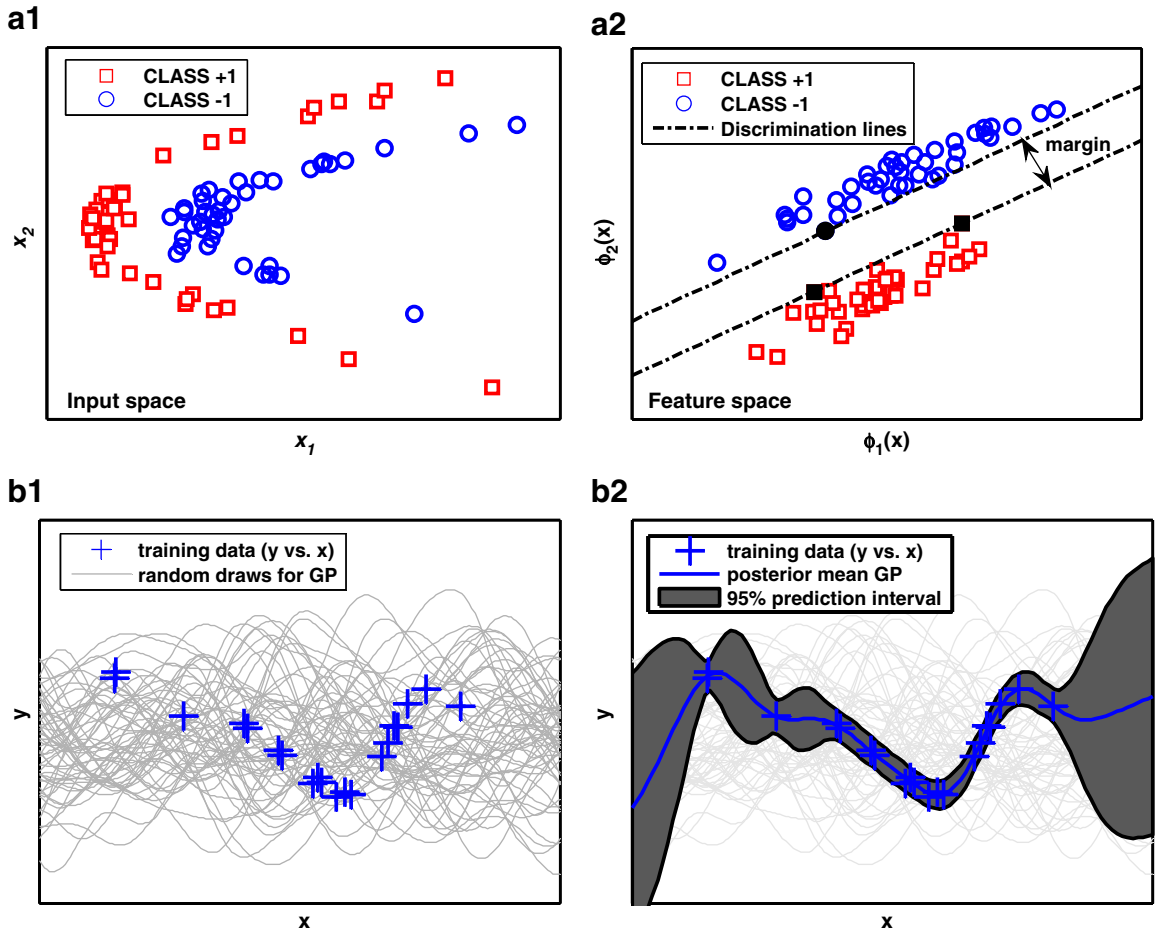

Fig. 3. Geometrical interpretations of (a) support vector machine for binary classification and (b) Gaussian processes for nonlinear regression. (a1) A 2D input space ($x1, x2$) loaded with two linearly non-separable classes (i.e., $+1$ and -1); (a2) to linearly separate the two classes, the input space ($x1, x2$) is transformed into a kernel-induced feature space ($\phi_1(\mathbf{x}), \phi_2(\mathbf{x})$) where in SVM seeks a distinguishing hyper-plane that produces a maximal separation margin. (b1) Random draws from an underlying Gaussian process are equally likely used as candidate functions to approximate the unknown nonlinear relationship *a priori*; and (b2) in light of the training data (plus signs), some draws are more likely to be the true function than others, and the most probable draw is the blue solid curve that is interpreted as a posterior GP mean function (i.e., the best fit), together with some uncertainty estimates as indicated by the shaded area.

which also depends on whether regression or classification is concerned. Its standard form is summarized as follows (Rasmussen & Williams, 2006):

$$l(y,f) = \begin{bmatrix} (y-f)^2 & \text{GP}-\text{regression} \\ -\,logp(y|f) & \text{GP}-\text{classification} \\ (|y-f|-\varepsilon)_+ & \text{SVM}-\text{regression} \\ (1-y\cdot f)_+ & \text{SVM}-\text{classification} \end{bmatrix} \quad (7)$$

where the response variable y is continuous for regression and discrete for classification, e.g., taking value of either $+1$ or -1 for binary classification. With regard to the loss $l(y,f)$, GP regression uses the common squared error; GP classification uses the negative logarithm of likelihood where $p(y|f)$ denotes the probability of the class label being y (i.e., $+1$ or -1) conditioned on the function value $f(\mathbf{x})$ (Rasmussen & Williams, 2006). In contrast, SVM uses the $\varepsilon-$insensitive loss function and the hinge function for regression and classification, respectively, and therein, the notation $(x)_+$ stands for $\max(x,0)$ that equals x for $x>0$ and 0 otherwise. As depicted in Fig. 3a, the use of $(x)_+$ in SVM implies that only those training data points with positive losses will contribute to the model inference and the rest are useless and hence can essentially be discarded. This property known as sparsity is a salient advantage of SVM (Hsu et al., 2003). Those data points contributing to the loss are termed as support vectors, thus hinting the name of SVM. Conversely, GPs in its original formulation do not enjoy the sparsity property because all data points are involved in model inference, although there do exist some fast, sparse GP models such as relevance vector machine and informative vector machine that aim to approximate the corresponding full GP models for practical applications with large-scale datasets.

The regularizer $||f||^2_{\mathcal{H}}$ measures how regular or smooth a function $f(\mathbf{x})$ is. The norm $||\cdot||_{\mathcal{H}}$ involved is defined upon a so-called reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ (Rasmussen & Williams, 2006). Of particular note is that each RKHS $\mathcal{H}$ is uniquely associated with a positive definite kernel function $k(\mathbf{x},\mathbf{x}')$ and accordingly, any $f(\mathbf{x})$ in the function space $\mathcal{H}$ can be expressed as

$$f(\mathbf{x}) = \sum_i \alpha_i \cdot k(\mathbf{x},\mathbf{x}_i) \quad (8)$$

where α_i are scalar coefficients. As a result, $||f||^2_{\mathcal{H}}$ becomes (Rasmussen & Williams, 2006)

$$||f||^2_{\mathcal{H}} = <f,f>_{\mathcal{H}} = \sum_i \sum_j \alpha_i \alpha_j k\left(\mathbf{x}_i,\mathbf{x}_j\right) \quad (9)$$

where the first equality assumes that the norm is induced from $<\cdot,\cdot>_{\mathcal{H}}$ (i.e., the inner product of $\mathcal{H}$) and the second has used the reproducing property of a RKHS $<f(\cdot),k(\cdot,\mathbf{x})>_{\mathcal{H}} = f(\mathbf{x})$. For practical purposes, common kernels of $k(\cdot,\cdot)$ include, but are not limited to, the polynomial, the radial basis function (RBF), the exponential, the neural network, and the automatic relevance determination functions (ARD) (Rasmussen & Williams, 2006). The most widely used one among these is probably the RBF that takes the form of:

$$k(\mathbf{x},\mathbf{x}') = exp\left(-\frac{||\mathbf{x}-\mathbf{x}'||^2}{2\sigma^2}\right) \quad (10)$$

where the scalar parameter σ^2 can be interpreted as a characteristic length beyond which function values at two points will become irrelevant. Another interesting kernel is the ARD that is a natural extension to the RBF and has the following form (Neal, 1996):

$$k(\mathbf{x},\mathbf{x}') = exp\left[-\sum_{k=1}^{m}\frac{(x_i-x_i')^2}{2\sigma_i^2}\right] \quad (11)$$

which uses separate characteristic-length parameters σ_i^2 for different dimensions x_i of $\mathbf{x}$. A large value for σ_i^2 tends to downplay the contribution of x_i to $k(\mathbf{x},\mathbf{x}')$, with an extremely large σ_i^2 essentially annihilating x_i from the model inference. Such use of different σ_i^2 allows us to determine the relative importance of each variable x_i, thus providing an appealing scheme for variable selection especially if the number of candidate variables is large (i.e., more commonly known as feature selection in remote sensing literature) (Neal, 1996).

5.2. Learning of SVM and GP

Substitution of Eqs. (7) and (9) into the objective functional Eq. (6) suggests that the minimization of $J[f]$ with respect to the function f reduces to that with respect to a discrete set of scalar coefficients α_i, provided that all the other necessary parameters are supplied beforehand, viz.

$$J[f] = J\left[\alpha_i\,|\,k(\cdot,\cdot),\sigma^2,\cdots,C\right]. \quad (12)$$

These pre-specified or tuned parameters are often called hyperparameters, which, for example, include kernel type $k(\cdot,\cdot)$, kernel parameter (σ^2 or σ_i^2), the tradeoff weight C, and the tolerance parameter ε for SVM regression. Given a training dataset $[\mathbf{x}_i,y_i]_{i=1,\ldots,n}$, the optimization of $J[f]$ for α_i conditioned on a set of fixed hyperparameter values is here called model learning/fitting whereas the tuning of these hyperparameters with respect to $[\mathbf{x}_i,y_i]_{i=1,\ldots,n}$ is called model training. A brief description of model training is deferred to the next sub-section. Mathematically, the existence of solutions α_i to the minimization of Eq. (12) is guaranteed by the convexity of all the loss functions of Eq. (7) (Rasmussen & Williams, 2006). Of the optimal solution $\hat{\alpha}_i$ (i = 1, …, n), those $\hat{\alpha}_i$ associated with support vectors are nonzero and all the others are zero, embodying the sparsity property of SVM. In contrast, all $\hat{\alpha}_i$ for GP models are generally nonzero, although some of them may be very small in magnitude. Numerically, the minimization of $J[f]$ for SVM is frequently solved via the sequential minimal optimization algorithm (van der Heijden et al., 2004), and that for GPs is solved directly by matrix inversion (Rasmussen & Williams, 2006).

5.3. Model training and practical considerations

Model training herein refers to selecting sensible values for hyperparameters (e.g., kernel σ_i^2 and tradeoff C parameters) based on an observed dataset. This process is critical to ensure fitting a good SVM or GP model because in most cases, predefined values of these hyperparameters are unavailable *a priori* and random guesses are far from sufficient. The typical training procedure for SVM is to employ a nested grid-searching approach to pinpointing sensible parameter values based on cross-validation in terms of minimizing some error criteria (Hsu et al., 2003). In contrast, GP models are rooted inherently in a Bayesian framework, therefore enabling the Bayesian learning of sensible parameters with the well-known maximum-likelihood type II (ML-II) method (Zhao et al., 2008). In this work, the optimization procedure used in the ML-II method is a modified Polak–Ribière conjugate gradient approach (Rasmussen & Williams, 2006).

The optimization for model training used to tune hyperparameters is in most cases non-convex, which differs from the minimization of $J[f]$ to learn model coefficients α_i conditioned on a given set of hyperparameters. Therefore, the optimized hyperparameters are likely local extremes, potentially leading to model overfitting (Seeger, 2004). No universal solutions exist to safeguard against this problem, yet a simple, practical remedy is to run the optimization for multiple times from different initial random values of hyperparameters and then choose the solution with the best

performance, as used in Zhao et al. (2008). For this study, 500 times was used when training GP models.

Exact values of hyperparameters optimized via model training are also contingent on the data ranges of predictors and response variables. For numerical convenience, it is common practice to linearly scale data to a predefined interval (Hsu et al., 2003), e.g., [0, 1] as chosen in our following experiments. However, in GP models, the transformation of response variables (i.e., y) is often treated differently in that y is first centered with respect to its means and then normalized to have a unit variance. This scaling for y is preferred because GP models usually assume a zero-mean function in formulation (Zhao et al., 2008). Correspondingly, when making prediction, results need to be appropriately back-transformed to the original scale.

Another practical trick is pertinent to the use of the ARD kernel that typically has a large number of characteristic-length parameters σ_i to be tuned through model training (e.g., hundreds or thousands of parameters). The computation for high-dimensional optimization is feasible with gradient-based optimizers as in the ML-II method of GP models but is prohibitively daunting with the nested grid-searching approach of SVM (Zhao et al., 2008). To alleviate computation demands as well as take advantage of the ARD kernel for SVM, the following expedient is often found helpful: first train a GP model with the ARD kernel; then apply the learned characteristic-length parameters to scale the predictors accordingly; and use the scaled predictors to train a SVM model with the RBF kernel by the grid-searching approach. Although such a combined use of the ARD and SVM is not theoretically grounded, it often helps to better exploit the information content of predictors and thus improve model performance.

6. Experiments

Using the lidar and field data collected in East Texas, we conducted several experiments to examine the degree to which the combination of our high-dimensional composite lidar metrics and machine learning models could improve estimation of canopy attributes at the plot level, as compared to classical classification and regression models. In these experiments, the hybrid form of lidar composite metric $\mathbf{QP}\{p_{fs}, h_l\}$ was used, which include both canopy density and quantile height predictors (Eq. 5). Specifically, the hybrid composite metric was calculated by employing a varying height-bin and quantile interval to discretize the involved lidar height distributions and quantile function: Quantile heights of first/single returns for $q(p_{fs})$ were obtained at p_{fs} of 1%–5% and 95%–100% with an increment of 1%, and 10% to 90% with an increment of 5%; canopy density predictors of last returns for $p(h_l|q(p_{fs}))$were obtained with a height-bin of 1 m below the height of 5 m, and a height-bin of 5 m for the range of 5 m up to 40 m. As a result, the hybrid composite metric consists of 165 predictors, 8 canopy density predictors of which, however, were removed due to being uniformly zero-valued, thus leaving a total of 157 predictors for SVM and GP models. On the other hand, another independent set of 138 common laser metrics was calculated, including mean heights, variances and coefficients of variation for first/single returns, last returns, and all returns, 50 percentile heights of first/single and last returns, and 80 canopy density metrics of various types of returns. As opposed to the hybrid metric for machine learning models, these 138 common metrics, together with their quadratic and logarithmic transformations, served as candidate variables for conventional linear or log-linear regression models.

6.1. Use of SVM to classify forest types

The first experiment aims to distinguish pine from hardwood/mixed forests. This is a trivial yet potentially difficult binary classification problem because our lidar metrics are based on ranging measurements and thus, unlike multispectral indices, lack spectral information that normally provides critical clues to assess forest compositional status. To train and test the SVM classifier, we randomly select two samples, each containing 5000 pixels at a 20-m resolution, with one being used as training data and another as test data. Class labels of these pixels were indirectly obtained from the Quickbird-derived classification map by aggregating it from a resolution of 2.5 m to 20 m: a 20-m pixel is defined as pine if it has at least 50% of its covered 2.5-m pixels being pine; otherwise, it is labeled as hardwood/mixed. Apparently, this definition is arbitrary but it serves well our purpose of assessing the performance of SVM. These Quickbird-derived class labels were assumed to be ground-truth in the experiment, which is acceptable as far as our purpose of testing the utility of SVM and our lidar metrics is concerned. Furthermore, to account for the effect of training sample size, we selected a series of sub-samples from the full set of the 5000 training pixels, with a gradually increasing size from 100 to 5000 incremented by 100. This resulted in a total of 50 new training samples, each of which was used to train a separate SVM. The trained SVMs were then validated on the other independent 5000 test pixels. As a comparison, we also applied a conventional classifier, i.e., the commonly used maximum likelihood classifier (MLC), to the same training and test samples as for SVM. Since MLC is not robust enough in tackling high-dimension problems, we first performed principal component analysis (PCA) to the whole set of common lidar metrics and then, chose only the top principal components as new input variables for MLC.

6.2. Use of GP to estimate canopy structural variables

As additional experiments, GP regression models were fitted to relate the lidar composite metric to a total of 10 canopy structural variables at the plot level, including Lorey's height (LH), dominant height (DH), basal area (BA), aboveground and belowground biomass (AGB and BGB), canopy base height (CBH), canopy ceiling height (CCH), canopy bulk density (CBD), available canopy fuel (ACF), and LAI. Of these 10 variables, DH is the mean height of trees within the dominant Kraft class; CBH, CCH, CBD and ACF are four basic canopy fuel parameters that find important applications in forest fuels management practice (Reinhardt et al., 2006).

Field-based estimates for the canopy variables were derived from the plot-level survey data, and in a loose sense, all these variables except LAI can be deemed as some linear or nonlinear transformations of individual tree attributes. More specifically, the general dbh-based allometric equations of Jenkins et al., 2003 are used to calculate AGB and BGB; a program called FuelCalc was used to calculate the four canopy fuel characteristics (Reinhardt et al., 2006). Therein, a fundamental feature is the vertical distribution of crown fuel which is calculated by first distributing the crown fuel of each tree between its crown base and its top according to species-specific allometry, then summing fuels in a 1 foot height-BIN for all trees of a plot, and finally smoothing the profile with a running-mean filter (e.g., a window size of 15 ft). From this smoothed profile, CBD is defined as the maximum of the running mean; CBH and CCH are defined as the lowest and highest heights at which the running mean exceeds 0.012 kg/m^3, respectively (Reinhardt et al., 2006). In the calculation of the four fuel parameters, we used the built-in species-specific allometry of FuelCalc. As to LAI, the two inversion algorithms employed in Zhao and Popescu (2009) were applied to the hemiphotos to generate two sets of *in-situ* LAI. We took the average of the two sets of LAI as more reliable field estimates. Furthermore, in the following model fitting and validation, BA, AGB, BGB and ACF were logarithmically transformed to mitigate the heteroskedasticity of error variances. Of particular note is that the anti-log transformation of predictions needs to be adjusted by some bias-correction factor (e.g., based on the estimated error variances as used in this study) (Sprugel, 1983).

Due to the limited number of field plots, cross-validation was considered when evaluating GP models. Rather than using the leave-one-out cross-validation (LOOCV), five-fold cross validation (FFCV) was chosen such that the observed sample is partitioned into five complementary subsets of about equal size: In each fold, four out of the five subsets is combined as the training sample, with the remaining one as out-of-sample validation data; this is repeated five times in total; and validation results from the five folds are combined to compute statistical measures for assessing how the model possibly performs on independent observations. Compared to the LOOCV that uses only a single observation for validation each time, FFCV considers a much larger proportion of out-of-sample observations as well as a smaller training sample per fold, thus possibly allowing for a more stringent and conservative model evaluation than LOOCV. Additionally, randomness is introduced due to the many possible choices for splitting the sample. To quantify such random variations, we partitioned the sample at random for a total of 1000 times and ran FFCV on each of the resulting samples.

For comparison purposes, linear or log-linear regression models were also developed and validated for each canopy variable using the same FFCV procedure as mentioned above. The independent variables entering these models have been determined using the entire set of sample via the stepwise regression procedure before we performed cross-validation.

7. Results

Descriptive statistics for the field-based estimates of the 10 canopy variables are summarized in Table 1. For LH and DH, four 0.01-ha plots have been excluded from consideration because these plots are either treeless or contain only low-stature tree saplings that are close to zero in height and also because they correspond to high leverage points that artificially boost model accuracies. For the fuel parameters, seven plots were eliminated because their canopy fuels are less than 0.012 kg/m^3 along the whole crown fuel profiles and no valid values exist for CBH, CCH and CBD according to the definitions in FuelCalc (Reinhardt et al., 2006). For LAI, five plots are eliminated because the associated hemiphotos are dominated by open skies and do not allow reliable LAI calculation.

Classification results of forest types show that the lidar data, acquired in a leaf-off season, are useful to discriminate between pines and hardwood/mixed through both the SVM and MLC classifiers. The final input variables chosen for MLC are the first 19 PCA components of the 138 common lidar metrics, which explained 97.54% of the original variance and produced the best accuracy for 26 out of 50 runs as compared to the use of other numbers of PCA components. Over these 50 runs that were trained with a gradually increasing sample size ranging from 100 to 5000 incremented by 100, both the overall accuracies of SVM and MLC tended to first increase dramatically with the training sample size and then steadily saturate to a plateau (Fig. 4). Specifically, for the first 5 runs that used a training sample size of less than 500 pixels, SVM and MLC performed similarly, but for the remaining runs, SVM consistently outperformed MLC (p-value $\ll 0$ from a paired-t test), with an averaged overall accuracy of 80.68% for MLC versus 82.27% for SVM. Such a gain of 1.69%, though small in magnitude, is statistically significant at the 0.05 level when assessed using the z-statistic (p-value = 0.0147).

To further illustrate, the SVM and MLC trained with 500 training pixels were applied to a subset of the study area, and the resulting classified maps are shown in Fig. 5. Both the class maps (Fig. 5c and d) exhibit a pattern similar to that of the reference Quickbird-derived forest type map (Fig. 5b). The confusion matrices obtained for the two classified maps indicate that SVM produced slightly better accuracies than MLC (Table 2), which however is only marginally significant (p-value = 0.07 for the z-statistic). It is also revealed in Table 2 that hardwood/mixed pixels are more prone to misclassification than pines, and this behavior was actually observed consistently throughout all the 50 runs.

As shown in Figs. 6 and 7, regression results demonstrate that both GP and common linear/log-linear models are capable of estimating canopy variables from laser-based metrics with reasonable accuracies. Regarding the final model forms, linear/log-linear models selected a varying number of lidar metrics as predictors for different canopy variables; the respective equations that were fitted on the entire set of plot-level data using stepwise regression are listed in Table 3. Coefficients of determination for these models were relatively high for height-related canopy variables but low for mass-related canopy variables, e.g., with the maximum R^2 of 0.89 attained for CCH and the minimum of 0.474 for Log(BGB).

Fig. 6 depicts the scatters of observed vs. predicted canopy variables resulting from one particular FFCV out of 1000 runs. It is observed that GP models significantly improved predictions over the linear/log-linear models in terms of both correlation coefficients (r) and root mean squared errors (RMSEcv). Data points of predicted versus observed values for the GP models followed more tightly around the 1:1 line, and such superior performances of GP models

Table 1
Statistics of field estimates of canopy characteristics.

	Number of plots	Minimum	Maximum	Mean	Standard deviation
Lorey's height (meter)	54	3.05	29.42	17.92	6.07
Dominant height (meter)	54	3.05	32.75	17.20	6.44
Basal area (m^2/ha)	51	10.23	63.28	25.72	10.80
Aboveground biomass (Mg/ha)	51	56.15	257.01	119.00	50.49
Belowground biomass (Mg/ha)	51	11.04	66.45	27.55	11.89
Canopy base height (meter)	51	2.75	21.04	11.44	4.43
Canopy ceiling height (meter)	51	8.54	36.73	20.72	6.86
Canopy bulk density (kg/m^3)	51	0.0124	0.1289	0.054	0.028
Available canopy fuel (kg/m^2)	51	0.14	0.85	0.36	0.16
LAI (m^2/m^2)	53	0.08	3.52	1.99	0.78

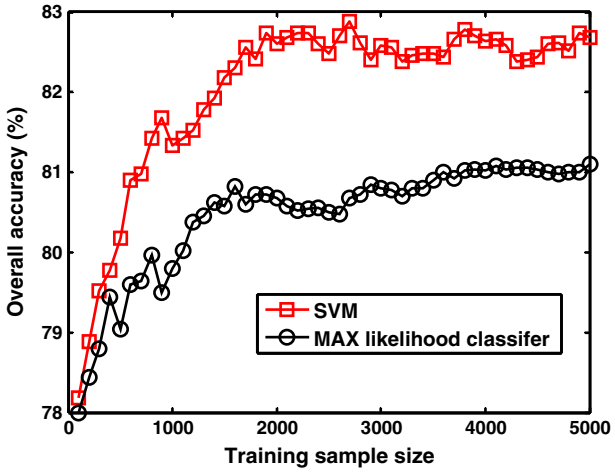

Fig. 4. Comparison of overall accuracies between SVM and maximum likelihood classifier for the forest type classification using a range of training sample sizes.

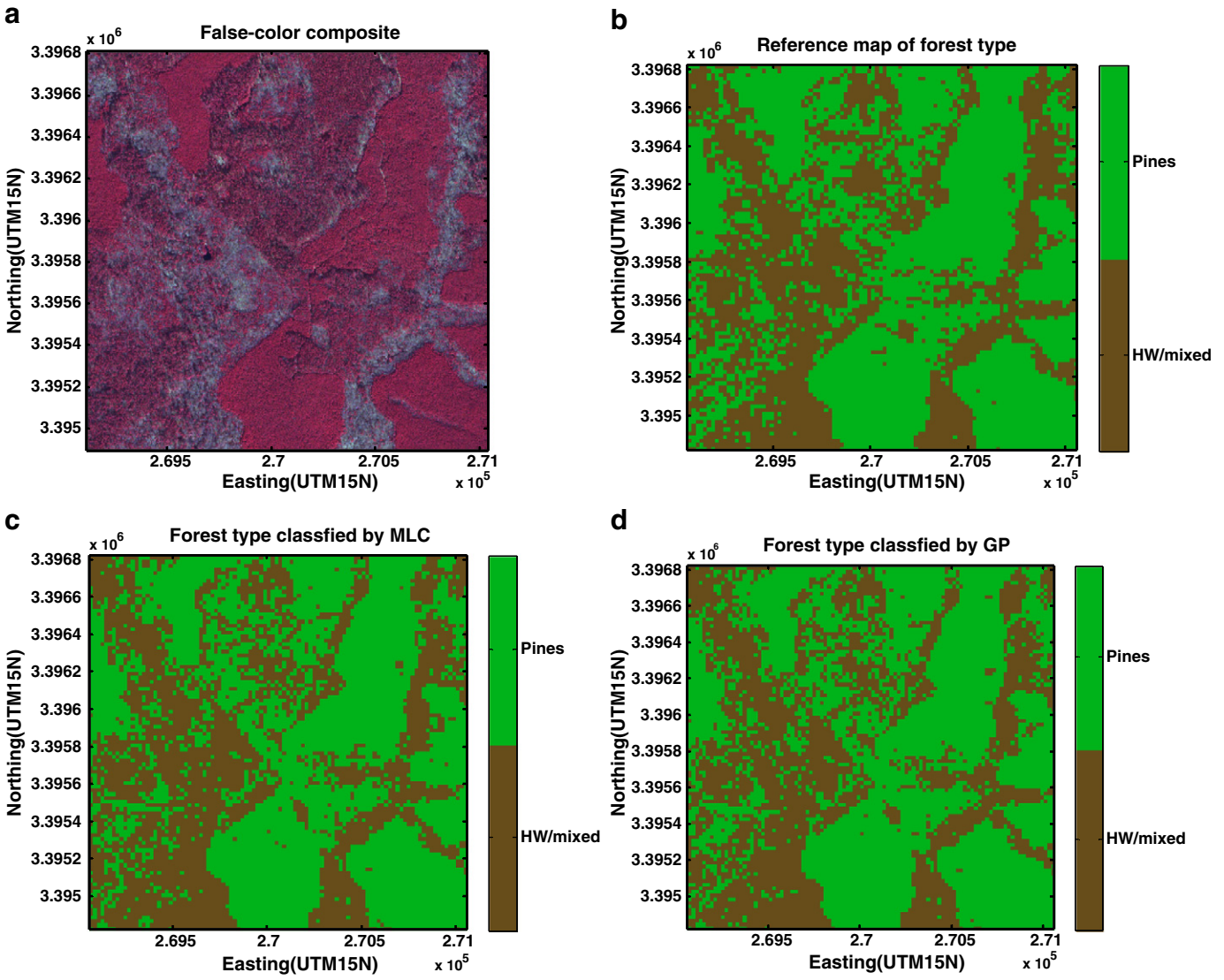

Fig. 5. Classification maps over a subset of the study area: (a) the false-color Quickbird composite, with a band assignment of near-infrared, green and red for RGB channels, respectively; (b) the reference forest type map classified directly from the Quickbird image; (c) the classified map from lidar data using MLC and (d) the classified map using SVM.

were noted consistently in all the 1000 runs of cross-validations. This has been clearly illustrated in the box-and-whisker plot of Fig. 7 comparing correlation coefficients of predicted vs. observed variables for the two types of models. The mean correlation coefficients and RMSEcv averaged over these 1000 runs are detailed in Table 4.

In addition, the scatterplots of Fig. 6 show that the prediction intervals of both GP and linear/log-linear models were realistic uncertainty estimates, as indicated by the number of times that the horizontal error bars intercept the 1:1 lines. Compared to linear/log-linear models, GP models yielded narrower prediction intervals for all the 10 canopy variables (Fig. 6), which accords with the higher correlation coefficients (Fig. 7) and lower RMSEcv in the cross-validation results (Table 4). Moreover, uncertainty intervals from the GP models could be more indicative of the true magnitudes of prediction errors than those of the linear models. Using Log(AGB) as an example, the scatterplots of Fig. 8 depict, for the GP model, a linear increase in estimated errors that is proportional to the absolute residuals but, for the log-linear model, a flat pattern that suggests some relatively constant error estimation regardless of the true residuals. The reduced major axis (RMA) fitting to the data points in the scatterplot of Fig. 8 shows that the best fitted line for the GP model is statistically indistinguishable from the 1:1 line (p value<0.001), which provides further positive evidence that the GP model yielded more informative uncertainty estimation.

The fitted GP and linear/log-linear models were also applied to the entire lidar data to produce spatially-explicit maps (Fig. 9). To avoid proliferation of figures, here we only show the resultant maps of predicted LH over the same sub-area as in Fig. 5. The overall patterns

Table 2
Confusion matrix for the SVM classification of forest types (within the parentheses are the corresponding results for the MLC).

Reference label	Classified label		
	Mixed	Pine	Producer's accuracy
Mixed	3240 (3228)	848 (860)	79.26% (78.96%)
Pine	846 (913)	5066 (4999)	85.69% (84.56%)
User's accuracy	79.29% (77.76%)	85.66% (85.47%)	83.06% (82.27%)

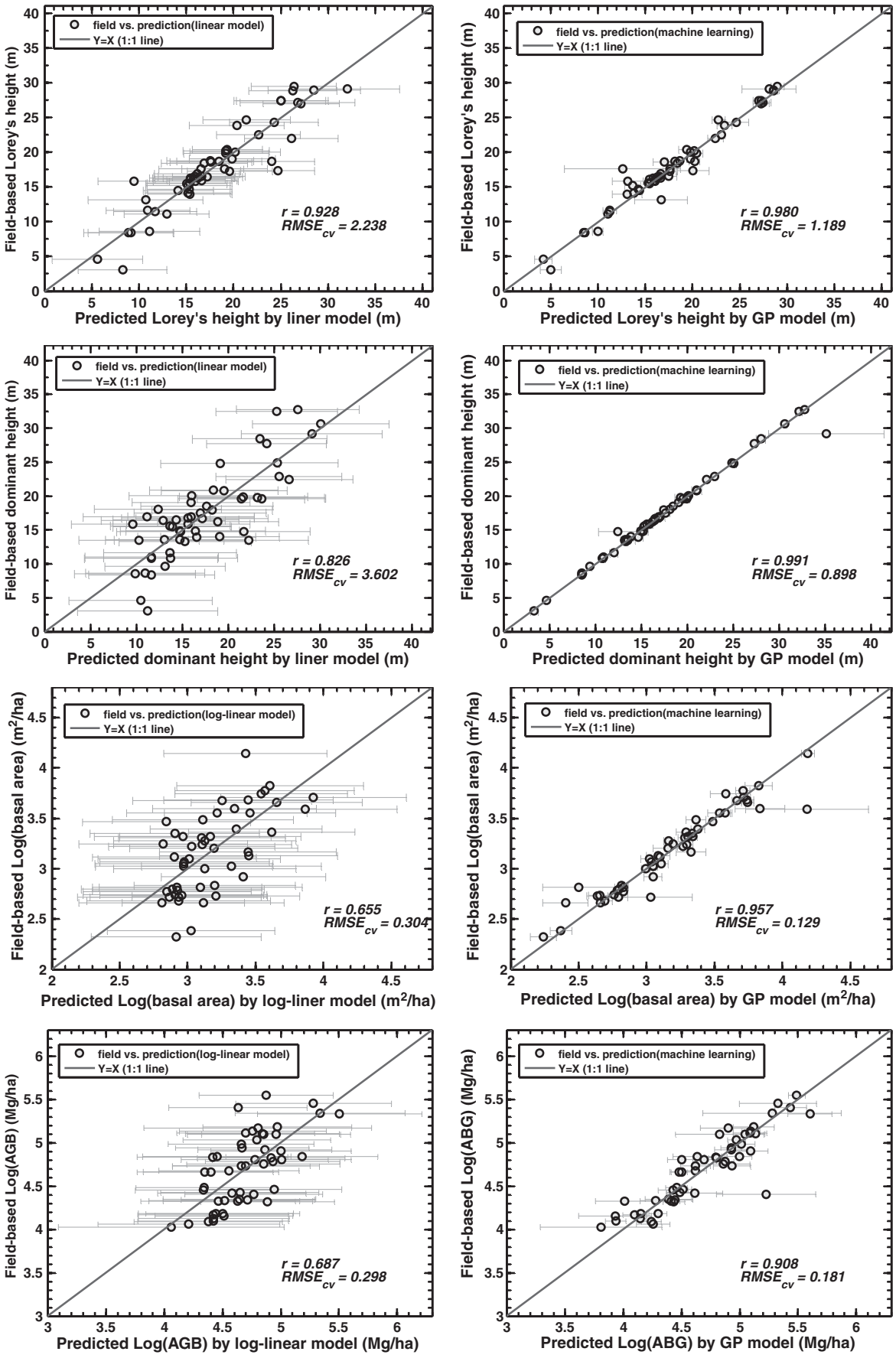

Fig. 6. Scatterplots of field-based vs. lidar-predicted canopy characteristics as obtained from a particular run of the five-fold cross validation analysis: horizontal error bars indicate the 95% prediction intervals. The left panel refers to linear regression models, and the right to GP regression models.

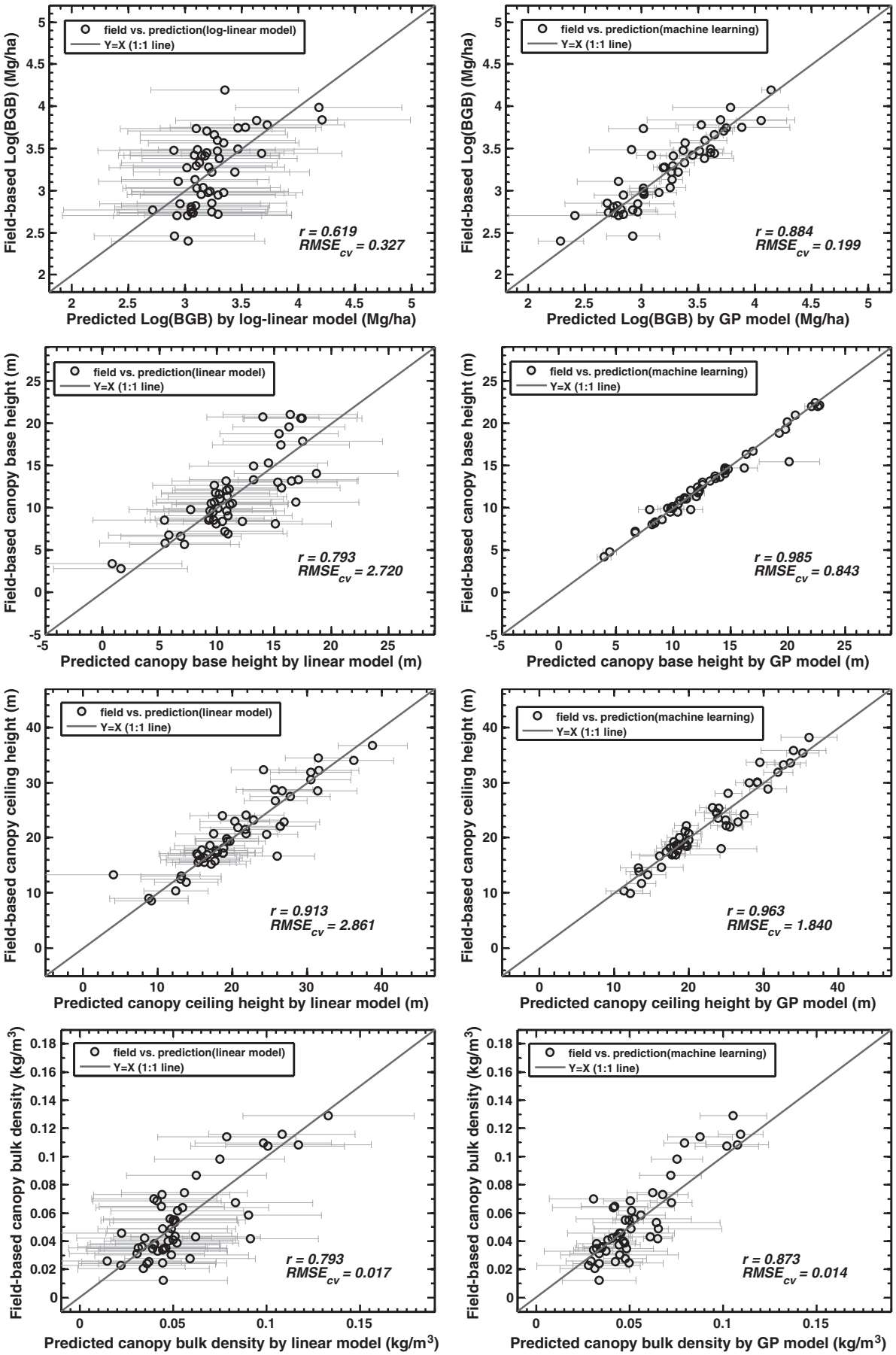

Fig. 6. (*continued*).

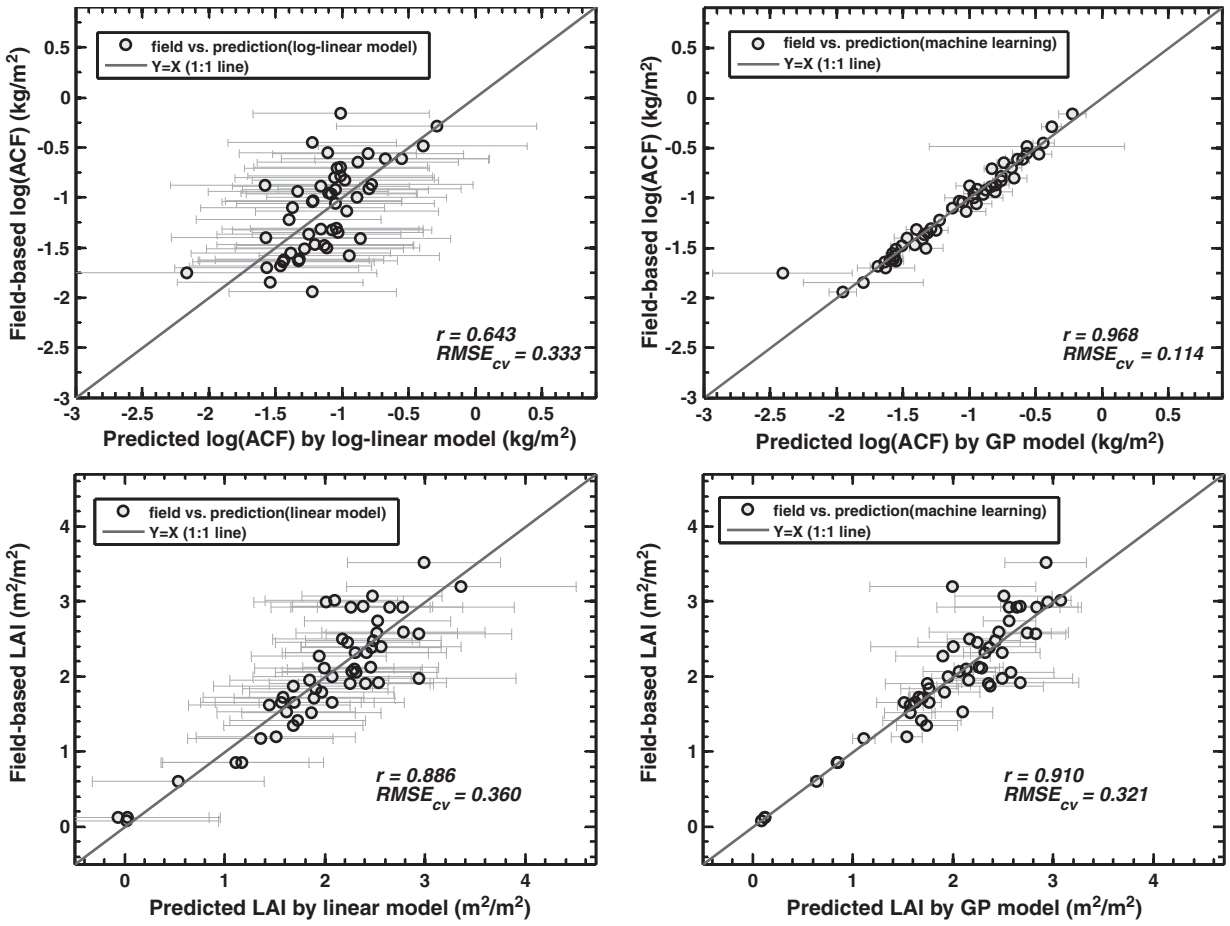

Fig. 6. (*continued*).

of the two maps are similar to each other, but many fine-scale disparities exist, as also revealed in the scatterplot between the two (Fig. 9c). Of particular note is that the GP models produced unrealistically negative LH values at a few pixels. Likewise, unrealistic predictions by GP occasionally occurred to several other canopy variables such as DH, CBH, and CCH.

8. Discussion

This study exemplifies that the use of detailed lidar metrics in conjunction with machine learning models could boost the efficacy of lidar in canopy characterization. Lidar data contain a wealth of information, including not only positional (xyz) and intensity

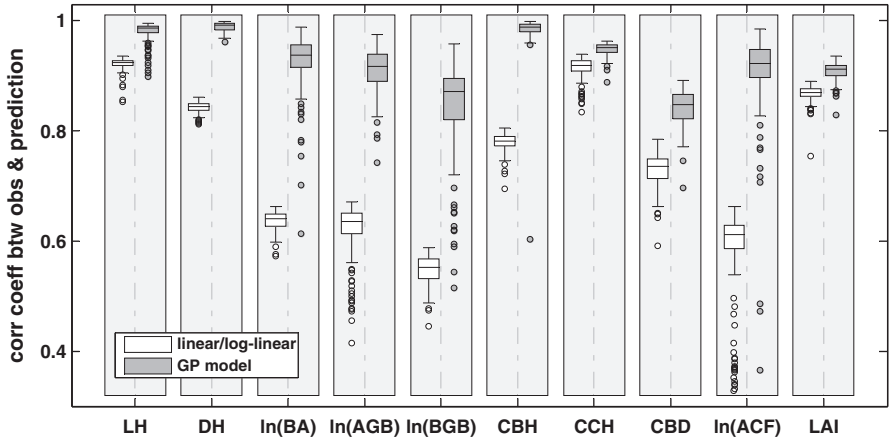

Fig. 7. Box-and-whisker plots of correlation coefficients between observed and predicted canopy characteristics for the 1000 runs of five-fold cross-validation. On each column associated with a canopy variable, the left half refers to the box plot of the linear/log-linear model and the right half refers to that of the GP model. Notice that the upper locations of the boxes for GP models relative to those for linear/log-linear models indicate better performance of GP models in terms of the measure of correlation coefficient.

Table 3
The best linear/log-linear models fitted with step regression using all the field samples. The naming convention for lidar predictors is as follows: "h" or "d" indicates that the predictor is a percentile height or density metric; "s" means the standard deviation of lidar heights; "a" or "f" refers to all or first returns. For example, h_{f02} denotes the 2% percentile of first returns only; $d_{f[0-5]}$ denotes the percentage of first returns falling into the height range of 0–5 m; and $d_{[30-35][25-30]}$ denotes the percentage of last returns that fall into 25–30 m and at the same time have the associated first returns falling to 30–35 m. Units for the canopy variables are the same as in Table 1.

Canopy variable	Equation	R2	RMSE
LH	$3.20 - 1.165 \cdot h_{f02} + 0.851 \cdot h_{f90} + 692.1 \cdot d_{[20-25][20-25]} - 95.71 \cdot d_{[30-35][15-30]}$	0.894	1.45
DH	$6.12 + 10.71 \cdot d_{[20-25][0-5]} + 0.1434 \cdot s_f + 0.0683 \cdot h^2_{f06} - 1924 \cdot d^2_{[30-35][25-30]}$	0.789	3.07
Log(BA)	$3.201 + 0.002895 \cdot h^2_{f25} - 366.26 \cdot d^2_{[10-15][6-9]} - 1.085 \cdot d^2_{f[0-5]}$	0.53	0.287
Log(AGB)	$4.40 - 0.242 \cdot h_{f02} + 0.0411 \cdot h_{f35} + 8969 \cdot d^2_{[20-25][10-15]}$	0.513	0.298
Log(BGB)	$2.765 + 0.00275 \cdot h^2_{f30} - 335.5 \cdot d^2_{[10-15][6-9]} + 0.1327 \cdot \log(h_{f60})$	0.474	0.314
CBH	$-8.43 + 1.528 \cdot h^2_{a01} - 206.38 \cdot d_{[15-20][15-20]} + 506.01 \cdot d_{[20-25][20-25]} - 0.0252 \cdot h^2_{f98}$	0.722	2.43
CCH	$2.71 - 5.64 \cdot h_{f75} + 10.16 \cdot h_{f80} - 3.559 \cdot h_{f85}$	0.894	2.30
CBD	$0.0475 - 0.00481 \cdot h_{f10} + 0.54 \cdot d_{[10-15][3-6]} + 0.000620 \cdot h^2_{f15} - 7.539 \cdot d^2_{[15-20][5-10]} - 9.80 \cdot d^2_{[30-35][5-10]}$	0.70	0.0163
Log(ACF)	$-1.724 + 0.0473 \cdot h_f - 0.0740 \cdot h^2_{f02} + 0.00189 \cdot h^2_{f25}$	0.473	0.326
LAI	$3.69 - 4.187 \cdot d_{a[0-3]} + 0.512 \cdot \log(d_{f[0-13]})$	0.805	0.356

measurements but also auxiliary variables such as scan angle, return types, and even topological linkage between returns (Reutebuch et al., 2005). However, regression models of standard grid-based approaches typically choose only a few metrics (Lim & Treitz, 2004), which potentially lose some important information useful for predicting canopy variables of interest. In contrast, specialized procedures, such as point-clouds segmentation algorithms that could handle individual lidar points, are capable of exploiting much more information inherent in the raw lidar data, but for various practical and technical reasons, most of these procedures at the current stage are infeasible for applications over extensive regions, especially over forested area of complex structures (Reitberger et al., 2009). To improve the use efficiency of lidar data in grid-based statistical modeling approaches, we proposed to simultaneously use a comprehensive set of metrics. One most salient reason for doing so is to preserve a considerable amount of the raw information, therefore holding promise for enhancing predictive powers of lidar data. Another important feature of our composite metrics is to take into account the correspondence of first and last returns belonging to the same pulse, which is expected to encode certain geometrical structure of canopies in a rather subtle and implicit way (Fig. 2). Of particular note, the composite metrics of this study are specific to discrete-return lidar data with up to two returns per pulse, but the same logic can be adopted for data with multiple returns per pulse in an effort to preserve a plethora of original information.

A notable disadvantage of our composite metrics is the potential high dimensionality of predictors that include variables with strong correlations (i.e., multi-collinearity). This essentially precludes their use in classical linear regression models unless the predictors are first pre-screened with respect to some selection criteria (Næsset et al., 2005). Actually, the same dilemma has also been faced by the hyperspectral image users who are interested in deriving biophysical or biochemical variables from a large number of contiguous spectral bands (Landgrebe, 2002). Rather than performing dimensionality reduction or feature selection, a common remedy they found effective is to refer to advanced modeling approaches, such as SVM and GPs, that are capable of tackling high-dimensional regression/classification problems (Durbha et al., 2007; Zhao et al., 2008). Results of this study add new experiential evidence to the large body of literature proving the effectiveness of advanced statistical models for remote sensing applications. Although we only examined the applicability of SVM and GP upon high-dimensional composite metrics, there is little doubt that the two techniques are also useful when only a parsimony of common lidar metrics are available. Therefore, their use as competitive alternatives to conventional methods is encouraged in future studies for establishing predictive frameworks based on conventional laser metrics from either discrete-return or waveform lidar.

The dimension of our composite metrics depends on the bin sizes specified for discretization, with a larger bin giving a coarser representation of lidar point distribution. In practice, it is appealing to vary the bin size in computing the composite metrics, as implemented in this study, so that the portions of height profiles with smooth variations are discretized using a larger bin and those with abrupt variations using a finer bin (Zhao et al., 2009). Such a varying bin size allows us to maintain a tradeoff between reducing the metric dimension and preserving details of the lidar height distributions. Because each group of last returns contains far fewer points compared to the total number of first/single returns, it is advisable to use a coarser bin for characterizing the distribution of each group of last returns than that of first/single returns. Furthermore, a non-overlapping bin has been implicitly assumed in this study; for practical purposes, it is viable to employ an overlapping bin that enables a more continuous numerical representation of the lidar height distribution, especially for low-density lidar data. This study did not go into details to investigate how the bin sizes will affect prediction performances, which could be examined in future work. To provide some preliminary clue on the bin-size effect, we recall the finding of Zhao et al. (2009) that a bin size as large as 5 m could produce aboveground estimates almost as good as those with smaller bins when using the CHP as predictor.

Results of this study also suggest that the kernel machine, GP, is applicable for both point and interval estimation (i.e., mean and standard deviation). Although the interval estimation theory for linear regression has long been well established, our cross-validation results reflect that error estimation from the GP model was more suggestive of the actual magnitudes of unknown uncertainties for our predictions (Fig. 8). The advantages of GP models for error bar estimation were also highlighted in many previous studies (Neal, 1996; Rasmussen & Williams, 2006; Seeger, 2004). For example, Chu

Table 4
Two statistical measures (correlation coefficients r and RMSEcv) calculated using the observed and predicted canopy variables from the five-fold cross validation. Values in parenthesis for Log(BA), Log(ABG) and Log(BGB) are those computed on the anti-logarithm scale. Units for the ten canopy variables are the same as in Table 1.

Statistics	Model	LH	DH	Log(BA)	Log(AGB)	Log(BGB)	CBH	CCH	CBD	Log(ACF)	LAI
r	GP	0.983	0.992	0.923 (0.901)	0.909 (0.902)	0.838 (0.835)	0.984	0.946	0.844	0.91 (0.915)	0.907
	Linear	0.921	0.941	0.634 (0.620)	0.623 (0.598)	0.549 (0.559)	0.778	0.916	0.73	0.572 (0.554)	0.868
RMSEcv	GP	1.06	0.95	0.163 (5.34)	0.170 (21.4)	0.2288 (6.54)	0.75	2.2	0.015	0.195 (0.068)	0.33
	Linear	2.36	3.43	0.589 (8.51)	0.568 (40.5)	0.5698 (9.88)	2.76	2.76	0.02	0.360 (0.133)	0.39

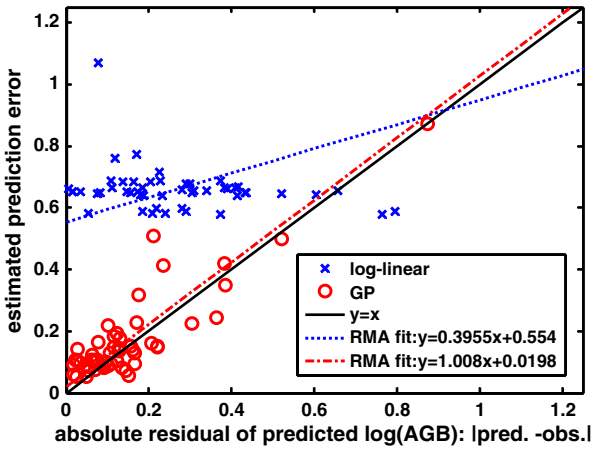

Fig. 8. Estimated prediction errors versus absolute residuals using Log(ABG) as an example: a comparison between the log-linear (blue cross) and GP (red circle) models. Note that the two dashed lines are the reduced major axis fits to the respective scatterplots and that the RMA fit for the GP model is statistically indistinguishable from y = x.

et al. (2004) showed that the prediction errors associated with the GP estimates for a laser fluctuation dataset were indicative of the estimation quality. Meanwhile, it must be realized that multiple linear regression and GP models are based on two contrasting paradigms, that is, frequentist versus Bayesian; hence, in a strict sense, uncertainty estimates for the two types of models should be interpreted differently. GP, as a Bayesian model, is supposed to provide a more natural interpretation conducive to our probabilistic understandings of error intervals (Neal, 1996; Zhao et al., 2008).

Along with the usefulness of GP for error estimation, we call for more attention paid to uncertainty analysis when employing lidar or other remote sensing data to retrieve land surface variables simply because a point estimate without knowledge about its uncertainty is less informative or even useless. This is also because error estimation has not been adequately addressed in many previous studies or sometimes even advertently overlooked (Chave et al., 2005; McRoberts et al., 2010). In fact, no universal methods exist for uncertainty analysis, due primarily to the great diversity of remote sensing-based information retrieval procedures. Generally, statistically-based approaches could allow a rigorous error assessment in a probabilistic sense, and uncertainty estimates are made possible without recourse to another independent ground-reference data. However, the exact error assessment for statistical estimators is somehow complicated by the many possible model choices for the same problem, especially for those involving the use of both design-based and model-based prototypes (Gregoire, 1998; McRoberts et al., 2010). In contrast, for non-statistical approaches such as those that rely on numerically inverting physically-based forward models, a ground-truth dataset independent of designing the inversion procedures is necessitated in the validation phase for error

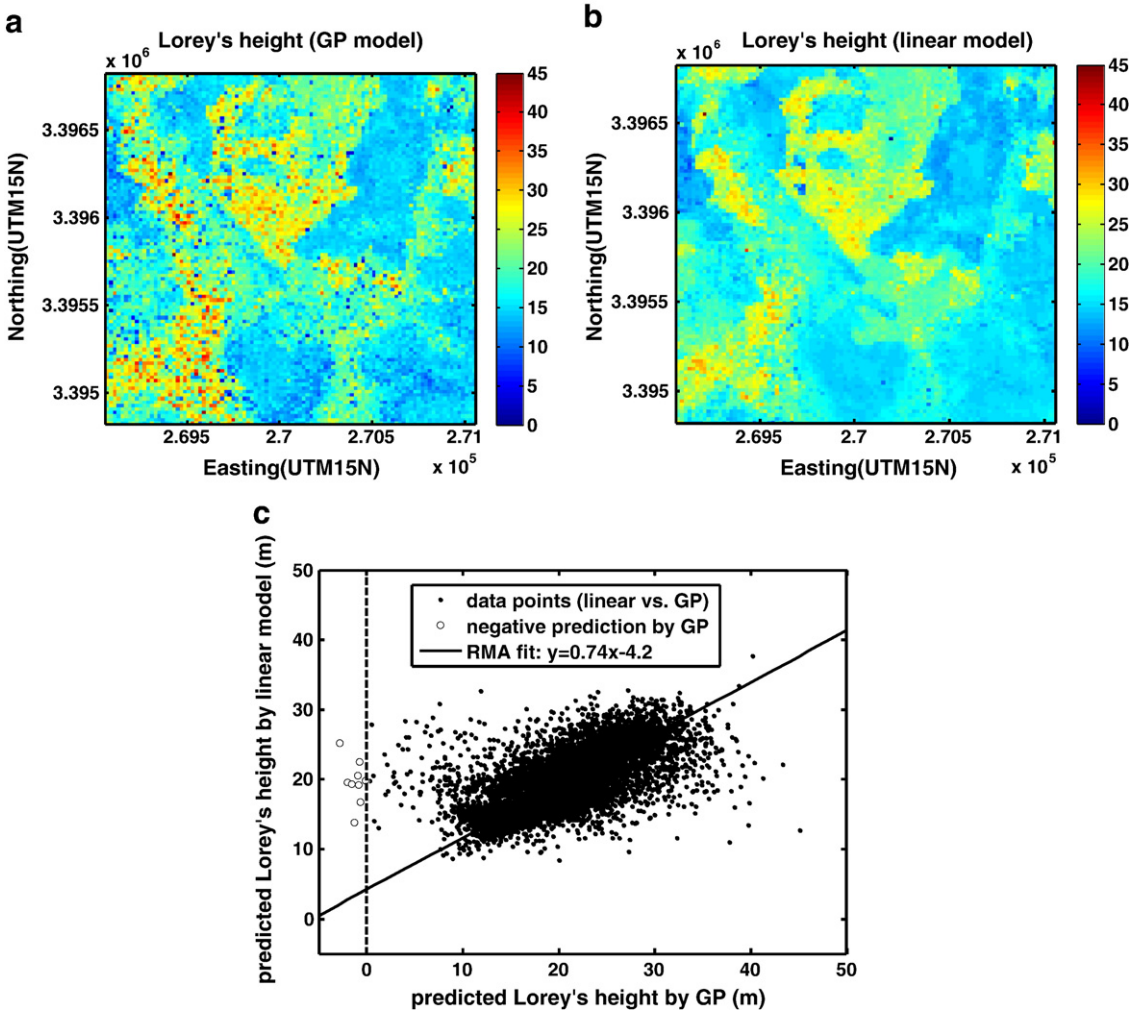

Fig. 9. A subset of Lorey's height map generated by the GP model (a) and linear model (b), together with the scatterplots between the two maps (c).

assessment (Morisette et al., 2002). For example, the uncertainties associated with MODIS LAI and albedo products are often directly evaluated by comparing them to observations of higher accuracies because the inversion algorithms involve some complicated radiation transfer models and cannot be easily incorporated into a rigorous statistical framework for error analysis (Morisette et al., 2002; Zhao & Popescu, 2009). Similarly, accuracies of lidar-derived tree variables by some dedicated, non-statistical individual-tree delineation methods are often assessed directly with respect to field measurements (Kato et al., 2009; Popescu, 2007; Popescu & Zhao, 2008). In practice, to assess the uncertainty for a non-statistical approach, another viable method is to use a simple regression model to relate the reference values to the estimates produced by the non-statistical approach. For example, Popescu and Zhao (2008) regressed the field-measured tree height against the lidar height derived by a varying-window filter method; the fitted equation serves at least two purposes: (1) to correct for the downward bias in the lidar-derived tree height, and (2) to treat the lidar tree height as a new predictor to estimate the true field-measured height, thus offering a statistical means to assess the errors associated with new height predictions.

With the increasing availability of remotely-sensed data from various sources (McRoberts et al., 2010), it is our vision that the future of remote sensing applications will expect an increase in use of machine learning as the relevant techniques become more mature. A more recent example in lidar canopy remote sensing is Yu et al. (2010) that successfully employed a machine learning technique called random forest to estimate individual tree attributes. Overall, the current use of machine learning, especially in lidar research, still remains limited due in part to some practical factors such as the unfamiliarity of these tools to a larger community, the unavailability of user-friendly implementations, the requirement for an in-depth understanding of underlying theories to avoid misuse, and the higher computation demand than their classical counterparts (Durbha et al., 2007; Zhao et al., 2008). In this study, machine learning was primarily employed as a prediction tool. The role of machine learning in practical lidar applications can be further expanded by applying it as an exploratory tool. Specifically, when it remains uncertain whether or what lidar predictors are useful for estimating a particular canopy characteristic (either a structural or compositional variable), machine learning models could be first fitted using a comprehensive set of lidar metrics like the ones of this study for the purpose of testing. The testing results then serve as guidelines to assist searching for appropriate lidar metrics and model forms needed for conventional regression methods. For example, our cross-validation results for GP models gave an r-value of 0.838 for Log (BGB), and chances are little to find a linear regression model that has a better performance, thus setting a yardstick for those who attempt to develop linear models.

Relating lidar metrics to canopy variables via functional relationships, whether implicit or explicit, is characteristic of many supervised learning tasks in remote sensing (Næsset et al., 2005). Such tasks depend strongly on field observations, which is particularly true for the kernel machines of this study because GP and SVM are essentially template-matching models (see Eq. 8). Similar to the common k-nearest neighborhood method, the effective use of kernel machine models is guaranteed only if enough observations (i.e., templates) are available (Evgeniou et al., 2000; Rasmussen & Williams, 2006). For instance, the effects of observation availability on estimation can be partially revealed in Fig. 3b2 where predictions at locations closer to observations are more accurate with smaller 95% error intervals than those at distant locations. Meanwhile, data used for machine learning models are preferably observed across the full range of the variable in question to minimize the risk of extrapolation (Zhao et al., 2009). Unlike a linear model, extrapolation for GP models gives rise to considerable variability in prediction, as indicated in Fig. 3b2 by the much wider 95% confidence intervals near the two ends of the predictor axis. In forest surveys, field samples are often randomly observed according to some sampling protocols that are tailored to some specific design-based estimators (Schreuder et al., 1993); in most probability, such field observations are not optimal for fitting machine learning models. Furthermore, various errors associated with field observations will be propagated along the chain of model inference, therefore contributing to the overall prediction uncertainty (Chave et al., 2005). As stated in earlier research (e.g., Popescu & Zhao, 2008), one important source of error is the misregistration between lidar data and field plots, which on average is about 2–3 m in magnitude for this study. The mis-registration error can be further confounded by the plot boundary effects, e.g., the exclusion of a tree that has its stem outside the plot but with a significant portion of its crown falling within the plot, or vice versa (Andersen et al., 2005). Other possible errors are concerned with calculation of in-situ values for canopy variables (Chave et al., 2005). For example, our *in-situ* LAI does not represent true leaf area but is an effective one that does not account for either clumping or the differentiation between wood and foliage. Also, inaccuracies of tree allometry contribute to errors in field-estimated biomass or canopy fuels (Zhao & Popescu, 2009).

Unrealistic predictions by GP models due to extrapolation place a limit to the range of predicted values that can be deemed as valid and reliable estimates, as suggested by the negative predicted values for LH (Fig. 9). Our results indicate that due to their nonlinearity, GP models are more sensitive to extrapolation than linear regression models. Indeed, cautions should be exercised for any models when interpreting extrapolated values. In the meantime, it must be noted that the negative predictions in canopy heights, though useless and unrealistic, do not invalidate the GP model inference because the estimated prediction intervals associated with these negative estimation are large enough so as to still contain the true values with a very high probability. To correct for these extreme predictions in practical applications, the following two options could be useful. The first one is to simply truncate the unrealistic values into a pre-defined valid range. The second is to employ an ensemble of models, from which a better estimate can be synthesized in terms of certain synthesis rules: the most straightforward rule probably is to take the average or sometimes the weighted average over the ensemble; this logic, known as model averaging, has been extensively used in many disciplines such as Bayesian statistical modeling (Denison et al., 2002). For models considered in this study, an alternative synthesis rule is to just combine the GP and linear/log-linear models and choose the prediction that has a smaller uncertainty interval.

Although fine-scale maps with good accuracies can be generated from lidar with approaches such as the machine learning models used here, further research is in need to develop lidar approaches as regional inventory tools. As discussed in Næsset and Gobakken (2008), one possibility is to apply airborne lidar as a sampling tool by acquiring data only over selected units (e.g., local areas); lidar-based models as developed in this study can be applied to each unit; and the information derived for these units could be further complemented by affordable optical imagery to attain estimates for a large region, e.g., within a multi-phase sampling framework. As an extreme case of sampling units, airborne profiling laser measurements along individual transects, spaced a few kilometers apart, have been used in combination with data such as MODIS and ICESat/GLAS to estimate biomass and carbon stock over extensive geographical regions (Nelson et al., 2009).

In the foreseen future, lidar will continue to be one of the most promising remote sensing tools for characterizing forests, with one current priority being the deployment of lidar for mapping and monitoring biomass and carbon of forest ecosystems in support of global change studies (Hurtt et al., 2004; Koch, 2010; Reutebuch et al., 2005). A factor of practical concern for implementing this is to maintain a reasonable cost. Due to the supervised learning nature,

lidar-based estimation models usually are developed using spatially- and temporally-coincident lidar and field data (Lefsky et al., 2005; Næsset, 2002; van Aardt et al., 2006). Such dependence discourages transplanting models from one region to another or to the same region at a different time unless the relevant discrepancies can be quantified, although there are a few studies indicating the generality of lidar-biomass models for extensive geographic regions (Lefsky et al., 2005). Notably, most lidar metrics depend on lidar systems, system and flight configurations (e.g., flight altitude, maximum scan angle, sampling density, and recorded echoes per pulse), and vegetation phenology (e.g., leaf-off or leaf on), which all restrict the re-usability of lidar-based models (Næsset, 2009a, 2009b). Comparatively, those lidar metrics derived from only first returns are less sensitive to changes in lidar acquisition settings, and they are preferred when building a model intended to be applied multiple times for monitoring biomass over time (Næsset, 2009a). One such example again is the functional biomass model of Zhao et al. (2009) that is dictated only by tree allometry; this model used canopy height profiles that can be derived either from lidar canopy height models or first returns. On the other hand, the requirement for temporal coincidence of lidar and field data could be relaxed if the interest is in predicting forest status at the time of field observation; this relaxation is ensured by the statistical nature of the relevant models, i.e., seeking a statistically-based function relationship between variables. For example, Zhao and Popescu (2009) successfully used laser data acquired on a leaf-off date to estimate LAI of leaf-on trees through regression analysis. Furthermore, given no landscape-changing disturbances, it is even possible to use the same lidar data to map forest statuses at multiple times by relating the lidar metrics to the respective field observations via different regression models.

9. Conclusions

The utility of lidar for characterizing forest ecosystems, e.g., mapping forest biomass, has been unarguably confirmed by a large body of literature, yet the capacity of lidar data for estimating canopy structural characteristics might have not been fully attained with conventional grid-based regression approaches, thus still leaving considerable room for improvement. This study partly fills this room by incorporating some high-dimensional lidar composite metrics into machine learning models. Built upon the canopy height profile metric of Zhao et al. (2009), our proposed composite metrics represent an even more comprehensive set of laser-based predictors, aiming to preserve as much information of raw lidar data as possible. In particular, these metrics indirectly encode the geometrical correspondence between first and last lidar returns, that is, an aspect that has been ignored in previous research. The two kernel machines we employed, i.e. SVM and GPs, are powerful tools to build data-driven nonlinear models for relating canopy variable to lidar metrics. The combined use of the high-dimensional composite metrics and machine learning models partly alleviates the burden of searching for physically, ecologically or statistically meaningful predictors, which expedites testing on the predictive power of lidar for a given canopy structural variable. Results from the case study for the Eastern Texas forest prove that the machine learning models improve upon both the point and interval estimates of a number of canopy structural characteristics as compared to classical regression models. Results also highlight that our composite metrics can serve as universal predictors for estimating many canopy characteristics or even classifying forest types. Further, the use of such advanced machine learning models is expected to increase in future research, e.g., when fusing lidar with other remote sensing data to estimate surface biophysical attributes. In addition, research efforts are still needed regarding the affordable, operational use of lidar for monitoring forest status over time for extensive regions.

Acknowledgment

This research was partly supported by a NASA New Investigator grant (NNX08AR12G). We thank Curt Stripling, all forestry personnel of the Texas Forest Service, and Alicia Griffin for their help with field data collection/compilation. Special thanks are due to Dr. Elizabeth Reinhardt and Mr. Duncan Lutes at the Rocky Mountain Research Station of US Forest service for their help in instructing on the use of FuelCalc. Last but not the least, we greatly appreciate the constructive comments from the three reviewers.

References

Andersen, H. E., McGaughey, R. J., & Reutebuch, S. E. (2005). Estimating forest canopy fuel parameters using lidar data. *Remote Sensing of Environment*, *94*, 441–449.

Chave, J., Chust, G., Condit, R., Perez, R., & Lao, S. (2005). Error propagation and scaling for tropical forest biomass estimates. In O. L. Phillips, & Y. Malhi (Eds.), *Tropical forests and global atmospheric change* (pp. 155–163). : Oxford University Press.

Chu, W., Keerthi, S. S., & Ong, C. J. (2004). Bayesian support vector regression using a unified loss function. *IEEE Transactions on Neural Networks*, *15*(1), 29–44.

Cook, B. D., Bolstad, P. V., Næsset, E., Anderson, R. S., Garrigues, S., Morisette, J. T., et al. (2009). Using LiDAR and quickbird data to model plant production and quantify uncertainties associated with wetland detection and land cover generalizations. *Remote Sensing of Environment*, *113*, 2366–2379.

Coops, N. C., Wulder, M. A., Culvenor, D. S., & St-Onge, B. (2004). Comparison of forest attributes extracted from fine spatial resolution multispectral and lidar data. *Canadian Journal of Remote Sensing*, *30*(6), 855–866.

Denison, D., Holmes, C., Mallick, B., & Smith, A. F. M. (2002). *Bayesian methods for nonlinear classification and regression*. London: Wiley.

Durbha, S. S., King, R. L., & Younan, N. H. (2007). Support vector machines regression for retrieval of leaf area index from multi-angle imaging spectroradiometer. *Remote Sensing of Environment*, *107*(1–2), 348–361.

Englund, S. R., O'Brien, J. J., & Clark, D. B. (2000). Evaluation of digital and film hemispherical photography and spherical densiometry for measuring forest light environments. *Canadian Journal of Forest Research*, *30*, 1999–2005.

Erdogy, T. L., & Moskal, L. M. (2010). Fusion of LiDAR and imagery for estimating forest canopy fuels. *Remote Sensing of Environment*, *114*, 725–737.

Evgeniou, T., Pontil, M., & Poggio, T. (2000). Regularization networks and support vector machines. *Advances in Computational Mathematics*, *13*(1), 1–50.

Falkowski, M. J., Evans, J. S., Martinuzzi, S., Gessler, P. E., & Hudak, A. T. (2009). Characterizing forest succession with lidar data: An evaluation for the Inland Northwest, USA. *Remote Sensing of Environment*, *113*, 946–956.

Falkowski, M. J., Smith, A. M. S., Hudak, A. T., Gessler, P. E., Vierling, L. A., & Crookston, N. L. (2006). Automated estimation of individual conifer tree height and crown diameter via two-dimensional spatial wavelet analysis of lidar data. *Canadian Journal of Remote Sensing*, *32*, 153–161.

García, M., Riaño, D., Chuvieco, E., & Dansond, F. (2010). Estimating biomass carbon stocks for a Mediterranean forest in central Spain using LiDAR height and intensity data. *Remote Sensing of Environment*, *114*, 816–830.

Gregoire, T. G. (1998). Design-based and model-based inference in survey sampling: Appreciating the difference. *Canadian Journal of Forest Research*, *28*, 1429–1447.

Holmgren, J., Nilsson, M., & Olsson, H. (2003). Estimation of tree height and stem volume on plots using airborne laser scanning. *Forest Science*, *49*(3), 419–428.

Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification. URL:. http://www.csie.ntu.edu.tw/cjlin/libsvm/index.html (last date accessed: 16 June 2010).

Hudak, A. T., Evans, J. S., & Stuart, A. M. (2009). LiDAR utility for natural resource managers. *Remote Sensing*, *1*, 934–951.

Hurtt, G. C., Dubayah, R., Drake, J., Moorcroft, P. R., Pacala, S. W., Blair, J. B., et al. (2004). Beyond potential vegetation: Combining lidar data and a height-structured model for carbon studies. *Ecological Applications*, *14*, 873–883.

Hyde, P., Dubayah, R., Walker, W., Blair, J. B., Hofton, M., & Hunsaker, C. (2006). Mapping forest structure for wildlife habitat analysis using multi-sensor (lidar, SAR/InSAR, ETM+, Quickbird) synergy. *Remote Sensing of Environment*, *102*, 63–73.

Jenkins, J. C., Chojnacky, D. C., Heath, L. S., & Birdsey, R. A. (2003). National-scale biomass estimators for United States tree species. *Forest Science*, *49*, 12–35.

Jennings, S. B., Brown, N. D., & Sheil, D. (1999). Assessing forest canopies and understorey illumination: canopy closure, canopy cover, and other measures. *Forestry*, *72*, 59–73.

Jensen, J. L. R., Humes, K. S., Vierling, L. A., & Hudak, A. T. (2008). Discrete return lidar-based prediction of leaf area index in two conifer forests. *Remote Sensing of Environment*, *112*, 3947–3957.

Kato, A., Monika, M., Schiess, P., Swanson, M. E., Calhoun, D., & Stuetzle, W. (2009). Capturing tree crown formation through implicit surface reconstruction using airborne lidar data. *Remote Sensing of Environment*, *113*, 1148–1162.

Koch, B. (2010). Status and future of laser scanning, synthetic aperture radar and hyperspectral remote sensing data for forest biomass assessment. *ISPRS Journal of Photogrammetry and Remote Sensing*, 581–590.

Landgrebe, D. (2002). Hyperspectral image data analysis as a high dimensional signal processing problem. *Special Issue of the IEEE Signal Processing Magazine*, *19*(1), 17–28.

Lefsky, M. A., Hudak, A. T., Cohen, W. B., & Acker, S. A. (2005). Geographic variability in lidar predictions of forest stand structure in the Pacific Northwest. *Remote Sensing of Environment*, *95*, 532–548.

Lim, K. S., & Treitz, P. M. (2004). Estimation of above ground forest biomass from airborne discrete return laser scanner data using canopy-based quantile estimators. *Scandinavian Journal of Forest Research*, *19*, 558–570.

Lim, K., Treitz, P., Baldwin, K., Morrison, I., & Green, J. (2003a). Lidar remote sensing of biophysical properties of tolerant northern hardwood forests. *Canadian Journal of Remote Sensing*, *29*, 648–678.

Lim, K., Treitz, P., Wulder, M., St-Onge, B., & Flood, M. (2003b). LiDAR remote sensing of forest structure. *Progress in Physical Geography*, *27*, 88–106.

McRoberts, R. E., Cohen, W. B., Næsset, E., Stehman, S. V., & Tomppo, E. O. (2010). Using remotely sensed data to construct and assess forest attribute maps and related spatial products. *Scandinavian Journal of Forest Research*, *25*(4), 340–367.

Meng, X., Wang, L., Silván, J. L., & Currit, N. (2009). A multi-directional ground filtering algorithm for airborne LIDAR. *ISPRS Journal of Photogrammetry and Remote Sensing*, *64*(1), 117–124.

Morisette, J. T., Privette, J. L., & Justice, C. O. (2002). A framework for the validation of MODIS land products. *Remote Sensing of Environment*, *83*, 77–96.

Mutlu, M., Popescu, S. C., Stripling, C., & Spencer, T. (2008). Assessing surface fuel models using lidar and multispectral data fusion. *Remote Sensing of Environment*, *112*, 274–285.

Næsset, E. (2002). Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. *Remote Sensing of Environment*, *80*, 88–99.

Næsset, E. (2009a). Effects of different sensors, flying altitudes, and pulse repetition frequencies on forest canopy metrics and biophysical stand properties derived from small-footprint airborne laser data. *Remote Sensing of Environment*, *113*, 148–159.

Næsset, E. (2009b). Influence of terrain model smoothing and flight and sensor configurations on detection of small pioneer trees in the boreal-alpine transition zone utilizing height metrics derived from airborne scanning lasers. *Remote Sensing of Environment*, *113*, 2210–2223.

Næsset, E., Bollandsas, O. M., & Gobakken, T. (2005). Comparing regression methods in estimation of biophysical properties of forest stands from two different inventories using laser scanner data. *Remote Sensing of Environment*, *94*(4), 541–553.

Næsset, E., & Gobakken, T. (2008). Estimation of above- and below-ground biomass across regions of the boreal forest zone using airborne laser. *Remote Sensing of Environment*, *112*, 3079–3090.

Neal, R. M. (1996). *Bayesian learning for neural networks.* New York: Springer 206 pp.

Nelson, R., Ranson, K. J., Sun, G., Kimes, D. S., Kharuk, V., & Montesano, P. (2009). Estimating Siberian timber volume using MODIS and ICEsat/GLAS. *Remote Sensing of Environment*, *113*, 691–701.

Pang, Y., Lefsky, M., Andersen, H., Miller, M. E., & Sherrill, K. (2008). Validation of the ICEsat vegetation product using crown-area-weighted mean height derived using crown delineation with discrete return lidar data. *Canadian Journal of Remote Sensing*, *34*(Suppl. 2), S471–S484.

Parker, G. G. (1995). Structure and microclimate of forest canopies. In M. D. Lowman, & N. M. Nadkarni (Eds.), *Forest canopies* (pp. 73–106). New York: Academic Press.

Popescu, S. C. (2007). Estimating biomass of individual pine trees using airborne Lidar. *Biomass and Bioenergy*, *31*, 646–655.

Popescu, S. C., & Wynne, R. H. (2004). Seeing the trees in the forest: Using lidar and multispectral data fusion with local filtering and variable window size for estimating tree height. *Photogrammetric Engineering & Remote Sensing*, *70*, 589–604.

Popescu, S. C., Wynne, R. H., & Nelson, R. H. (2002). Estimating plot-level tree heights with LIDAR: Local filtering with a canopy-height based variable window size. *Computers and Electronics in Agriculture*, *37*(1–3), 71–95.

Popescu, S. C., Wynne, R. H., & Nelson, R. H. (2003). Measuring individual tree crown diameter with LIDAR and assessing its influence on estimating forest volume and biomass. *Canadian Journal of Remote Sensing*, *29*(5), 564–577.

Popescu, S. C., & Zhao, K. G. (2008). A voxel-based lidar method for estimating crown base height for deciduous and pine trees. *Remote Sensing of Environment*, *112*(3), 767–781.

Quirino, V., Wynne, R., Seiler, J., & Thomas, V. (2009). Assessing the contribution of small-footprint lidar data for estimating soil respiration. In S. Popescu, R. Nelson, K. Zhao, & A. Neuemshwander (Eds.), *Silvilaser 2009 proceedings. College Station, Texas, USA, October 14–16, 2009*978-1-61623-997-8 CD-ROM.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning.* Cambridge, Massachusetts: The MIT Press 248 pp.

Reinhardt, E. D., Lutes, D., & Scott, J. (2006). FuelCalc: A method for estimating fuel characteristics. *Proceedings — Fuels management – How to measure success.* In P. L. Andrews, & B. W. Butler (Eds.), *Proceedings RMRS-P-41, Fort Collins, CO.* (pp. 273–282) Rocky Mountain Research Station: USDA Forest Service.

Reitberger, J., Schnörr, C., Krzystek, P., & Stilla, U. (2009). 3D segmentation of single trees exploiting full waveform LIDAR data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *64*, 561–574.

Reutebuch, S. E., Anderson, H. -E., & McGaughey, R. J. (2005). Light detection and ranging (LIDAR): An emerging tool for multiple resource inventory. *Journal of Forestry*, 286–292.

Roberts, S. D., Dean, T. J., Evans, D. L., McCombs, J. W., Harrington, R. L., & Glass, P. A. (2005). Estimating individual tree leaf area in loblolly pine plantations using lidar-derived measurements of height and crown dimensions. *Forest Ecology and Management*, *213*, 54–70.

Schreuder, H. T., Gregoire, T. G., & Wood, G. B. (1993). *Sampling methods for multiresource forest inventory.* : J. Wiley and Sons, Inc. 446 pp.

Seeger, M. (2004). Gaussian processes for machine learning. *International Journal of Neural Systems*, *14*(2), 69–106.

Sprugel, D. G. (1983). Correcting for bias in log-transformed allometric equations. *Ecology*, *64*, 209–221.

van Aardt, J. A. N., Wynne, R. H., & Oderwald, R. G. (2006). Forest volume and biomass estimation using small-footprint Lidar-distributional parameters on a per-segment basis. *Forest Science*, *52*(6), 636–649.

van der Heijden, F., Duin, R. P. W., de Ridder, D., & Tax, D. M. J. (2004). *Classification, parameter estimation and state estimation: An engineering approach using MATLAB.* New York: Wiley 440 pp.

van Leeuwen, M., & Nieuwenhuis, M. (2010). Retrieval of forest structural parameters using LiDAR remote sensing. *European Journal of Forest Research*, *129*(4), 749–770.

Yu, X., Hyyppä, J., Vastaranta, M., Holopainen, M., & Viitala, R. (2010). Predicting individual tree attributes from airborne laser point clouds based on the random forests technique. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(1), 28–37.

Zhao, K., & Popescu, S. (2009). Lidar-based mapping of leaf area index and its use for validating GLOBCARBON satellite LAI product in temperate forest of the southern USA. *Remote Sensing of Environment*, *113*, 1628–1645.

Zhao, K., Popescu, S., & Nelson, R. (2009). Lidar remote sensing of forest biomass: A scale-invariant estimation approach using airborne lasers. *Remote Sensing of Environment*, *133*, 182–196.

Zhao, K., Popescu, S. C., & Zhang, X. (2008). Bayesian learning with Gaussian processes for supervised classification of hyperspectral data. *Photogrammetric Engineering & Remote Sensing*, *74*(10), 1223–1234.