

# A simple area-based model for predicting airborne LiDAR first returns from stem diameter distributions: an example study in an uneven-aged, mixed temperate forest

Rebecca A. Spriggs, Mark C. Vanderwel, Trevor A. Jones, John P. Caspersen, and David A. Coomes

**Version** Post-Print/Accepted Manuscript

**Citation (published version)** Rebecca A. Spriggs, Mark C. Vanderwel, Trevor A. Jones, John P. Caspersen, and David A. Coomes. A simple area-based model for predicting airborne LiDAR first returns from stem diameter distributions: an example study in an uneven-aged, mixed temperate forest. *Canadian Journal of Forest Research*. 45(10): 1338-1350. <https://doi.org/10.1139/cjfr-2015-0018>

**Copyright / License** 'Reprinted with permission. © 2015 Canadian Science Publishing or its licensors'.

## How to cite TSpace items

Always cite the **published version**, so the author(s) will receive recognition through services that track citation counts, e.g. Scopus. If you need to cite the page number of the **author manuscript from TSpace** because you cannot access the published version, then cite the TSpace version **in addition to** the published version using the permanent URI (handle) found on the record page.

**This article was made openly accessible by U of T Faculty.**  
Please [tell us](#) how this access benefits you. Your story matters.



1 **A simple area-based model for predicting airborne LiDAR first**  
2 **returns from stem diameter distributions: an example study in an**  
3 **uneven aged, mixed temperate forest**

4 Rebecca A. Spriggs, Mark C. Vanderwel, Trevor A. Jones, John P. Caspersen, and David A.  
5 Coomes

6

7 **R. A. Spriggs and D. A. Coomes.** Department of Plant Sciences, University of Cambridge,  
8 Cambridge, UK

9 (ras212@cam.ac.uk; dac18@cam.ac.uk)

10 **M. C. Vanderwel.** Department of Biology, University of Regina, 3737 Wascana Parkway,  
11 Regina, Regina, Saskatchewan, Canada.

12 (mark.vanderwel@uregina.ca)

13 **T. A. Jones.** Ontario Ministry of Natural Resources and Forestry, Forest Research and  
14 Monitoring Section, 1235 Queen Street East, Sault Ste. Marie, Ontario, Canada

15 (trevor.jones2@ontario.ca)

16 **J. P. Caspersen.** Faculty of Forestry, University of Toronto, 33 Willcocks Street, Toronto,  
17 Ontario, Canada

18 (john.caspersen@utoronto.ca)

19 \* **Corresponding authors:** R. A. Spriggs and D. A. Coomes (e-mail: ras212@cam.ac.uk and  
20 dac18@cam.ac.uk; phone: 01223333900).

21

22

23 **Abstract:** Tree size distributions are of fundamental importance in ecology, forestry and global  
24 change research. Airborne laser scanning (LiDAR) provides high-resolution information on  
25 canopy structure and has great potential as a tool for mapping and monitoring tree stem diameter  
26 distributions (SDDs) across forest landscapes. We present an area-based allometric model, with  
27 three levels of species specificity, that links ground-based plot data to the height distribution of  
28 LiDAR first returns ( $HD_{LiDAR}$ ), demonstrating the approach with surveys of a mixed, uneven-  
29 aged forest in central Ontario, Canada. Our model translates stem diameters into estimates of  
30 exposed crown area within 1 m height intervals; we then compared those estimates with  $HD_{LiDAR}$   
31 values. This basic approach gave reasonable goodness-of-fits (RMSE = 32%), but accuracy was  
32 improved by adding mechanistic features to adjust crown shapes and crown permeability, and  
33 allow for crown overlap and gaps (RMSE = 17%). The model showed no bias in predicting  
34 LiDAR returns in the mid-to-upper canopy (18-30 m), but tended to underestimate those from  
35 the understory-level (2-8 m) and overestimate returns from the ground and lower canopy (8-18  
36 m). Our model represents an important step towards remote mapping of tree size distributions by  
37 showing that LiDAR first returns can be accurately predicted from standard plot data by  
38 considering a few fundamental canopy properties.

39

40

41

42 *Key words:* tree-size distribution, discrete-return LiDAR, mixed forest, allometry, canopy  
43 model.

44

45

46 **Introduction**

47 Tree stem diameter distributions (SDDs) are a fundamental property of forest structure and are  
48 therefore integral to the productivity, water and nutrient cycles, carbon storage, wildlife habitat  
49 suitability and disturbance risk of the forest (e.g. Parker et al., 1985; Kirschbaum, 1999; Sullivan  
50 et al., 2013; Coomes and Allen, 2007; Lines et al., 2010). Forest managers rely on SDDs to  
51 assess timber volumes and harvest yields (Scheller and Mladenoff 2004). In dynamic models of  
52 global change responses, detailed forest structural information is valuable for setting initial  
53 conditions, particularly when using individual-tree-based approaches (e.g. Köhler and Huth  
54 1998; Scheller and Mladenoff 2004; Purves et al. 2008; Caspersen et al. 2011). Currently, SDDs  
55 are routinely estimated by ground-based surveys, but these are expensive to conduct and based  
56 on sampling rather than complete mapping of survey areas.

57 Airborne LiDAR (light detection and ranging) is a laser-based remote sensing technology  
58 that produces high resolution point clouds of the locations where laser pulses have been reflected  
59 from leaves, branches, stems and the ground (Lim et al. 2003). Airborne LiDAR is increasingly  
60 used by foresters to measure biophysical properties of forests, including stand height, volume  
61 and density, basal area, and leaf area index (e.g. Næsset 2002; Lim et al. 2003; Popescu, Wynne,  
62 and Nelson 2003; Leeuwen and Nieuwenhuis 2010). Being able to map these properties over  
63 entire landscapes is a significant advantage of using airborne LiDAR data with the majority of  
64 work focusing on mapping aboveground carbon (e.g. Asner et al., 2012), biomass (e.g. Zhao et  
65 al., 2009) or timber volume (e.g. Hill et al., 2014). Such properties are important to timber and  
66 carbon accounting and can all be directly estimated from SDDs with some associated species  
67 information, such as the proportion of conifers vs. broadleaves. Possessing information on the  
68 underlying SDDs therefore offers another layer of detail and is particularly important for

69 assessing current and future timber stocks. A number of different approaches have been proposed  
70 for mapping SDDs, which we will go on to discuss. Our ultimate aim is to develop a novel  
71 approach to map SDDs using airborne LiDAR, where stems are binned into user-defined  
72 diameter classes at the scale of 2500 m<sup>2</sup>. In this paper, we explore the link between a forest  
73 stand's stem size distributions and its airborne LiDAR returns; our emphasis is on establishing  
74 the theoretical form of the relationship from basic principles.

75         Predictions of SDDs from airborne LiDAR can be organised into two distinct approaches:  
76 area-based approaches (ABA; e.g. Maltamo et al. 2006; Breidenbach, Gläser, and Schmidt 2008;  
77 Thomas et al. 2008; Jones, Woods, and Lim 2009; Valbuena et al. 2013) and individual tree  
78 detection (ITD; e.g. Dalponte, Bruzzone, and Gianelle, 2011; Yao, Krzystek, and Heurich, 2012;  
79 Vauhkonen and Mehtätalo, 2015). ABAs predominantly use traditional statistical relationships  
80 with the best-fitting LiDAR metrics to estimate the parameters of a distribution describing the  
81 stem diameters, such as the shape and scale parameters of a Weibull distribution (Thomas et al.,  
82 2008; Jones et al., 2009). Breidenbach et al. (2008) predicted the shape and scale of a left and  
83 right truncated Weibull from the first and third LiDAR height quartiles using a log link function.  
84 The main benefits of the statistical methods are that they rely on standard forest inventory data  
85 without any spatial information and are therefore relatively easily parameterised and applied in  
86 new areas at the plot-level scale. The downside of using traditional statistical relationships is that  
87 they are not necessarily intuitive; they are often based on whichever combination of metrics  
88 offered the best fit to the data potentially resulting in over-fitted models. Many of these models  
89 are not robust to extension to other forests where the metrics will be dependent on different  
90 LiDAR acquisition properties (Næsset 2004), as well as offering questionable accuracy in un-  
91 sampled regions of the forest (Vincent et al. 2012, Mitchard et al. 2014).

92 ITD methods can make more informed predictions of SDDs based on scaling relationships  
93 between a crown segmented from the point cloud and the expected associated stem diameter.  
94 Dalponte et al. (2011) presented a method for inferring individual stem diameters and volumes  
95 from crowns segmented using an algorithm developed by Hyyppä et al. (2001). The most  
96 effective LiDAR metrics associated with the segmented crowns were used to predict the stem  
97 diameters. Vauhkonen and Mehtätalo (2015) demonstrated how a transformation function can be  
98 developed to map the cumulative distribution of segmented crown radii onto the cumulative  
99 distribution of stem diameters; they emphasise that their method does not require mapped stems  
100 and makes some compensation for inaccuracies in the delineation process. This method also  
101 draws from the plot-level statistical methods as they selected the best-performing plot-level  
102 LiDAR metric to predict the random plot-level effects to fine tune the transformation function.  
103 Whilst ITD offers a promising alternative, ABA remains the approach favoured by foresters,  
104 particularly in Scandinavia (Xu et al. 2014). ITD requires a minimum point density with most  
105 studies reporting between 8-12 points/m<sup>2</sup>, whereas ABA has been shown to be unaffected by  
106 decreasing the point density to as low as 1 point/m<sup>2</sup> (Jakubowski et al. 2013), and, even at high  
107 point densities, ITD struggles with crown overlap and cannot detect understory trees reliably.  
108 Some progress has been made in combining the two approaches to take advantage of the  
109 strengths of both methods; Xu et al. (2014) demonstrated how the SDD prediction made using  
110 ITD and high density LiDAR can be combined with that made using an ABA and low density  
111 LiDAR to improve the accuracy of the ABA estimates and reduce the bias of the ITD estimates.

112 Our ultimate aim is to offer an alternative approach to those listed above, which balances  
113 the inclusion of realistic properties of the forest canopy with the simplicity to apply the model  
114 across landscapes with low LiDAR point densities. First, we must develop the intermediate link

115 model which bridges the gap between the SDD and the height distribution of LiDAR first returns  
116 ( $HD_{LIDAR}$ ). If all tree crowns in a stand were identically shaped and there was no allometric noise  
117 associated with the crown-stem diameter allometries, and if laser pulses were fired directly  
118 downward and reflected off the upper surface of the forest canopy, then the link between size  
119 distributions and LiDAR signals would be straightforward to describe. From this basic model,  
120 we explore how adding additional complexity improves the predictions of the LiDAR return  
121 distribution ( $HD_{PREDICT}$ ).

122 SDD -  $HD_{PREDICT}$  models can be extremely detailed, as demonstrated in the ray-tracing  
123 literature (e.g. Disney et al., 2010; Goodwin et al., 2007; Sun and Ranson, 2000) where the  
124 interaction of each simulated laser pulse is predicted given a detailed representation of the forest  
125 canopy. These models, such as the LiDAR Interception and Tree Environment (LITE) model  
126 (Goodwin et al. 2007), allow us to explore the effect sensor properties have on the resultant  
127 LiDAR data. For our purposes, we require a model that falls between these two extremes, but  
128 which additional features should be included in our basic model?

129 Within a given plot, our basic model predicts the crowns from non-spatial diameter at  
130 breast height (dbh) measurements based on a single fixed allometry. The predicted crown areas  
131 within 1 m height intervals are summed and the exposed canopy area within each interval is used  
132 to generate the  $HD_{PREDICT}$ , which we compare with the  $HD_{LIDAR}$ . In this way, we are considering  
133 the individual trees, as in ITD methods, but not explicitly. As previously discussed, one of the  
134 limitations of this method is that crown overlap and overtopping reduces the exposed crown area  
135 (Vauhkonen and Mehtätalo 2015), so this is the first feature that should be incorporated.  
136 Drawing from the ray-tracing methods, the fundamental canopy property determining the  
137 behaviour of the LiDAR pulse is the distribution of foliage (Goodwin et al. 2007). Without

138 spatial information and maintaining an ABA, we approximate the distribution of foliage  
139 important to predicting first returns by three additional features: gap fraction, more detailed  
140 crown shapes and crown permeability (Jochem et al. 2010). The combination of these features  
141 should help to refine the predicted area of foliage at given heights.

142 Here, we develop the link between the SDD and the HD<sub>PREDICT</sub> which operates at the plot-  
143 level, yet includes key properties of the forest canopy; the proposed benefits of this method are  
144 that it is simple, requiring only standard forest inventory data, low density LiDAR and low  
145 computational power, it is intuitive, therefore can plausibly be recalibrated to other similar  
146 forests, and has been developed on a relatively complex forest type. A large proportion of the  
147 research in this area is based in alpine, mountain or boreal forests which are therefore largely  
148 comprised of coniferous species. We test the model on an uneven aged, mixed conifer-broadleaf  
149 forest of central Ontario, Canada, which therefore has a relatively complex structure as compared  
150 with many that are used in the literature.

151 We will address two research questions: 1) is a single allometry sufficient or is more  
152 species information required? 2) How does each feature independently improve the basic model  
153 (gap fraction, crown shape refinement, canopy permeability and crown overlap)? In the  
154 discussion, we will discuss the feasibility of using this approach to map SDDs.

155

## 156 **Materials and methods**

### 157 **Study area**

158 The Haliburton Forest and Wildlife Reserve is a 32,000 ha privately owned property located in  
159 the Great Lakes - St. Lawrence Forest Region of central Ontario, Canada (45°13' N, 78°35' W).  
160 The forest grows on undulating terrain (ranging from approximately 400 – 500 m above sea

161 level). Monthly temperatures range from an average of  $-7.5 \pm 3.2$  °C in the winter months to  $17.8$   
162  $\pm 1.5$  °C in the summer months. Average monthly precipitation is  $84 \pm 9.5$  mm (Environment  
163 Canada 2012). The forest is a mixture of broadleaf and conifers species; it is dominated by sugar  
164 maple (*Acer saccharum* Marsh), but several other species are common, including eastern  
165 hemlock (*Tsuga canadensis* (L.) Carrière), American beech (*Fagus grandifolia* Ehrh.), red maple  
166 (*Acer rubrum* L.), balsam fir (*Abies balsamea* (L.) Mill.) and yellow birch (*Betula alleghaniensis*  
167 Britt.). Across the region, forests are extensively managed for timber using selection silviculture,  
168 but there are small patches of unmanaged old-growth stands where the terrain is unsuitable for  
169 harvesting machinery. Natural disturbances include both small-scale gap-disturbances and rare  
170 large-scale wind events (Vanderwel et al. 2008).

171

## 172 **Inventory plot and LiDAR datasets**

173 The ground plot data were comprised of 154 circular plots each with a radius of 28.2 m  
174 corresponding to an area of 2500 m<sup>2</sup> (Fig. 1). We selected the plot locations using a stratified  
175 sampling design: the forest was classified into seven different ecotypes using overstory species  
176 composition and structure, soil type and the moisture regime and then these were further  
177 subdivided into six levels of canopy openness determined by eye from aerial imagery. We  
178 randomly selected plots so that at least two plots represented each of the 42 possible  
179 combinations. Within each plot, all trees with a dbh greater than 8 cm were identified to species  
180 and the dbh measured. The centre points of each plot were georeferenced to within 1 m using a  
181 differential GPS. See Table S.1 for additional plot details.

182 Discrete airborne LiDAR data were collected in August 2009 using a Cessna Turbo 206  
183 Stationair aircraft outfitted with an Optech ALTM 3100 LiDAR four-pass system which is

184 capable of recording up to 100,000 measurements per second with potentially four returns per  
185 pulse, each with an associated intensity. The system was flown at an altitude of 1500 m,  
186 therefore achieving a pulse repetition frequency of 70 kHz, with a pass overlap of 30% (see  
187 Table S.1 for flight specifications and definitions of terms). The average first return point density  
188 was 2 points per m<sup>2</sup>. The precisely geolocated points making up the point cloud were run through  
189 a classification routine to provide an initial surface model (i.e. isolating the points associated  
190 with the ground or close to the ground). The initial surface model was then smoothed using  
191 Optimal Geomatics' proprietary methods to determine a digital elevation model (DEM), which  
192 was checked using control and validation points across the area (horizontal tolerance of less than  
193 0.75 m and consolidated vertical accuracy of 0.20 m). A triangular irregular network (TIN) was  
194 constructed from the points classified as ground returns which then provided a z-coordinate  
195 associated with the ground for every return; all returns were then converted from the height  
196 above sea level to the height above the ground by subtracting these ground z-coordinates. We  
197 used only the first return data in our analyses to increase the generality and interpretability of our  
198 model (Næsset 2004). We clipped the LiDAR data corresponding to each inventoried plot from  
199 the wall-to-wall dataset using ArcGIS 10 and then the first returns were sorted into 1m vertical  
200 height intervals above the ground.

201

## 202 **Description of the basic allometry-based model**

203 Using the forest inventory data, we constructed a model of the size and shape of all of the crowns  
204 (Fig. 2a) in the plot using pre-existing allometric equations (Purves, Lichstein, and Pacala 2007;  
205 Caspersen et al. 2011 – hereby referred to as the published allometric functions). This model was  
206 used to estimate the proportion of exposed canopy area within 1 m height intervals (ECA, the

207 canopy surface which is exposed to laser pulses; Fig. 2b), thereby transforming dbh  
 208 measurements into a prediction of the height distribution of the ECA. The premise of this basic  
 209 model is that LiDAR first returns backscatter off exposed canopy, so the predicted proportion of  
 210 ECA within each 1 m height interval is assumed to be directly proportional to the observed  
 211 distribution of LiDAR first returns ( $HD_{LIDAR}$ ). Note that this model makes use of published  
 212 allometric functions, but does not contain any tuneable parameters.

213 The model requires stem diameter information and so we then explored how adding  
 214 species information improves the predictions. We present three versions of the model: 1) using a  
 215 single allometry for all species, 2) using two allometries (one for conifers and one for  
 216 broadleaves) and 3) using species-specific allometries for the eight most prevalent species plus a  
 217 generic conifer and broadleaf allometry for all other species.

218 The crown components predicted from dbh were tree height, crown depth, maximum  
 219 crown radius and crown shape (Table S.1-Table S.3), from which the exposed crown area within  
 220 height tiers could be calculated for every tree within the plot. Vertical heights,  $h$ , are sorted into  
 221 1 m height tiers denoted as  $[h, h + 1)$ ; this is equivalent to  $h \leq x < h + 1$  where  $x$  is some  
 222 height falling within the interval. Within each height tier, the exposed area of all crowns was  
 223 summed ( $ECA_{[h, h+1)}$ ), and then transformed to a proportion of the total canopy area  
 224 ( $\sum_{i=0}^{\infty} ECA_{[i, i+1)}$ ; where  $i$  denotes the 1 m heights and  $h_{max}$  is the maximum height predicted in  
 225 the plot rounded up to the nearest meter) as follows:

226 (1) 
$$HD_{PREDICT}^{BASIC}[h, h+1) = \frac{ECA_{[h, h+1)}}{\sum_{i=0}^{h_{max}} ECA_{[i, i+1)}}.$$

227 The predicted height distribution of first returns ( $HD_{PREDICT}^{BASIC}$ ) is equivalent to the  
 228 proportional exposed canopy area within each height tier. A sugar maple allometry (Caspersen et

229 al. 2011) was used in the single allometry model and also for the generic broadleaf allometry  
230 used in the two allometry and species-specific model versions, since sugar maples are the  
231 dominant species within the study area. The generic conifer allometry was obtained by re-fitting  
232 the allometries to simulated data aggregated from species-specific predictions of each remaining  
233 coniferous species weighted according to species abundance (Purves et al. 2007, Caspersen et al.  
234 2011). In the species-specific version of the model, allometries for the eight most abundant  
235 species (Purves et al. 2007, Caspersen et al. 2011) were used in the calculations of  $ECA_{[h, h+1)}$   
236 and  $\sum_{i=0}^{h_{max}} ECA_{[i, i+1)}$ ; these species accounted for over 80% of the basal area represented in the  
237 stratified plot inventory. For the remaining species, the generic broadleaf and conifer allometries  
238 were used. See Table S.1 for allometric functions and Table S.2 and Table S.3 for associated  
239 parameter values.

240 Our aim was, ultimately, to develop a simple model which offers the best possible  
241 prediction ( $HD_{PREDICT}$ ) of the observed distribution of LiDAR first returns. The resultant model  
242 provides the intermediate link between SDDs and the  $HD_{LIDAR}$  which is necessary to generate a  
243 suite of height distribution predictions quickly from theoretical SDDs. The model thus offers a  
244 tool for matching any given  $HD_{LIDAR}$  for a similar forest to a most likely SDD; we expand on this  
245 further in the discussion (Fig. 3). We consequently needed to assess the performance of the basic  
246 model, where performance was measured by a reduction in the difference between the observed  
247 and predicted distributions ( $HD_{LIDAR} - HD_{PREDICT}$ ), and then we considered how the additional  
248 features improved model performance.

249

## 250 **Refining the basic model**

251 The basic model assumed that published crown allometries were accurate, that LiDAR first-  
252 returns were backscattered off the outer shell of the forest canopy (i.e. that crowns are  
253 impermeable), that there were no gaps in the canopy, and that individual trees had non-  
254 overlapping crowns. Since we made no spatial considerations, there is an underlying assumption  
255 that tree crowns organise themselves in space to minimise overlap (the perfect plasticity  
256 approximation of Purves et al. 2008) whereas in reality there may be considerable overlap in  
257 some places and gaps in others. In the following section, we describe how the basic model was  
258 refined in order to relax these assumptions. The parameters of the functions used to refine the  
259 basic model were estimated from a 114 plot training dataset; any improvement in fit was then  
260 assessed using the remaining 40 plot test dataset.

261 *Allowing for gaps in the forest canopies:* The proportion of pulses reflected off the ground  
262 (ground returns,  $p_0$ ) was assumed to be exponentially related to the total canopy area of a stand:

263 (2) 
$$p_0 = \exp\left(-\alpha \cdot \frac{\sum_{i=0}^{h_{max}} ECA_{[i, i+1)}}{PA}\right)$$

264 Where  $\alpha$  is a parameter and the second component in the exponential is the total canopy area (i.e.  
265 the sum of the ECA within all height tiers) divided by the plot area (PA; see Table S.4 for all  
266 parameter definitions). Given the estimated  $p_0$ , the predicted proportions of first returns made by  
267 the basic model were scaled to give an  $HD_{PREDICT}$  which sums to 1.

268 *Adjusting crown shapes:* To account for the often sparse datasets of crown shape measurements  
269 and that the crowns represented by the LiDAR first returns may differ from the allometric  
270 predictions (Piboule et al. 2005), we fitted a new crown shape parameter ( $\beta$ ) for all allometries  
271 used in each version of the model (see supplementary material; eqn. (S. 1)). All other parameters  
272 in the allometry were unaltered. The crown shape parameter controls the curvature of the crown

273 ranging from convex ( $\beta < 1$ ) to linear ( $\beta = 1$ ) to concave ( $\beta > 1$ ); this determines how the  
 274 crown radius decreases from the base of the crown to the peak.

275 *Allowing tree crowns to be semi-permeable:* LiDAR pulses may penetrate into a tree crown  
 276 before being backscattered at sufficient intensity to register a return in the laser scanner. Hence,  
 277 first returns are not necessarily from the outer shell of the crown (Gaveau and Hill 2003,  
 278 Chasmer et al. 2006). The following crown permeability (or crown transparency) parameter ( $\varphi$ )  
 279 relaxes the assumption that first returns scatter off the outer shell by allowing a proportion, given  
 280 by  $\varphi$ , to pass through the crown from the height interval above:

$$281 \quad (3) \quad HD_{PREDICT[h, h+1]}^{PERM} = HD_{PREDICT[h, h+1]}^{BASIC} \cdot (1 - \varphi) + HD_{PREDICT[h+1, h+2]}^{BASIC} \cdot \varphi$$

282 **For example, if the crown were a cylinder, then  $(1 - \varphi)$  gives the proportion of the crown**  
 283 **area that would be recorded from the height interval in which the top of the tree falls, and  $\varphi$**   
 284 **gives the proportion that would be recorded from the interval below.**

285 *Accommodating overlap between tree crowns:* There is a strong likelihood that small trees  
 286 within a stand are overtopped or overlapped by neighbours and so are not apparent in the LiDAR  
 287 first-return signal. We assumed that the probability of crowns in a given height tier being  
 288 exposed to LiDAR depends on the canopy area within higher tiers. We incorporated this effect as  
 289 a crown overlap correction factor ( $\theta_{[h, h+1]}$ ) determined by a negative exponential of the  
 290 proportional cumulative canopy area at a particular height where  $\gamma$  controls the shape of the  
 291 exponential; cumulative canopy area refers to the total ECA above this given height.

$$292 \quad (4) \quad \theta_{[h, h+1]} = \exp\left(-\gamma \cdot \frac{\sum_{i=h+1}^{h_{max}} ECA_{[i, i+1]}}{PA}\right)$$

293 *Combining all refinements into the full model:* The crown permeability and crown overlap  
 294 features can be combined into the following:

295 (5)

$$296 \quad HD_{PREDICT[h, h+1]}^{PERM+OVERLAP} = \theta_{[h, h+1]} \cdot HD_{PREDICT[h, h+1]}^{BASIC} \cdot (1 - \varphi) + \theta_{[h+1, h+2]} \cdot HD_{PREDICT[h+1, h+2]}^{BASIC} \cdot \varphi$$

297 The full model incorporating all of the above features: gap fraction, refitted crown shape,  
298 crown permeability and crown overlap is therefore given by:

$$299 \quad (6) \quad HD_{PREDICT[h, h+1]}^{FULL} = \begin{cases} \frac{1}{N} \cdot (p_0 + (1 - p_0) \cdot HD_{PREDICT[0, 1]}^{PERM+OVERLAP}) & \text{for } h = 0 \\ \frac{1}{N} \cdot ((1 - p_0) \cdot HD_{PREDICT[h, h+1]}^{PERM+OVERLAP}) & \text{for } h = 1, 2, 3, \dots \end{cases}$$

300 where N is a normalisation constant calculated such that  $\sum_{h=0}^{h_{max}} HD_{PREDICT[h, h+1]}^{FULL} = 1$

301

302 The  $HD_{PREDICT}$  of the basic model ( $HD_{PREDICT[h, h+1]}^{BASIC}$ ) is defined in eq. 1, the proportion of  
303 ground returns ( $p_0$ ) is defined in eq. 2, crown permeability ( $\varphi$ ) in eq. 3, crown overlap ( $\theta_{[h, h+1]}$ )  
304 in eq. 4 and the combination of the two ( $HD_{PREDICT[h, h+1]}^{PERM+OVERLAP}$ ) in eq. 5.

305 In the discussion, we will address the potential and limitations of using this model to  
306 predict SDDs.

307

### 308 **Model fitting**

309 Building on the basic model, which has no tuneable parameters, we added the four features  
310 detailed above to show how each influenced model performance in each of the three versions of  
311 the basic model. All four features were then combined in the full model. We estimated the  
312 parameter values that minimised the difference between  $HD_{PREDICT}$  and  $HD_{LIDAR}$  in the training  
313 plots, and measured model performance using the test plots.

314 We used a Markov Chain Monte Carlo (MCMC) algorithm, implemented in the Filzbach  
315 C# library (Purves and Lyutsarev 2011), to estimate the parameters in the model features  
316 described above. In each case, the log-likelihood was calculated by assuming that the probability  
317 density function of  $(HD_{LIDAR} - HD_{PREDICT})$  was normally distributed with a mean zero and a  
318 standard deviation  $\sigma$ . We used uninformative priors, and sampled parameters uniformly between  
319 upper and lower bounds that were chosen to be algebraically sensible or biologically reasonable.  
320 For each version of the model, we ran 3 replicate chains of 9,000 iterations each and a burn-in of  
321 1,000 iterations, except for the full model where each chain comprised of 35,000 iterations, with  
322 a burn-in of 5,000 iterations, to allow for the greater number of parameters. All models  
323 comfortably converged within the allotted number of iterations. We retrieved estimates for the  
324 posterior means and 95% credible intervals of all parameters, and used a deviance information  
325 criterion (DIC) to compare support for the models. Model performance was compared by  
326 predicting  $HD_{PREDICT}$  for the 40 plots set aside for testing purposes, and then calculating the  
327 RMSE for each of the test plots by plotting the  $HD_{PREDICT}$  against the corresponding  $HD_{LIDAR}$ .

328

## 329 **Results**

### 330 **Model performance**

331 The full model with all of the features included was the best supported statistically (lowest DIC  
332 and mean RMSE) in each of the three versions of the model (Table 1). Overall, the full model  
333 with the species-specific allometries offered the lowest DIC, but the two allometry version of the  
334 full model predicted the distribution of LiDAR first returns for the test plots (mean RMSE =  
335  $0.0196 \pm 0.0075$ ) equivalently well as the species-specific full model (mean RMSE =  $0.0196 \pm$   
336  $0.0074$ ). When the predictions for each version of the full model were averaged over all of the

337 test plots (Fig. 4), the one-allometry model offered a poorer fit than the two more complex  
338 versions. When focusing on the two- and ten-allometry predictions, the model is seen to capture  
339 the distribution of returns from the top of the canopy and ground, although it slightly over-  
340 predicts frequencies in the mid-height range (~ 15 m) and under-predicts at lower heights. The  
341 performance of the full species-specific (ten-allometry) model is interpreted at the plot level in  
342 Fig. 5 to show examples from the full range of predictions. The RMSE statistics, across all test  
343 plots, for the full model indicated a greatly improved fit compared with the basic model in all  
344 versions of the model (Table 1; Fig. 6). The RMSEs in the full model were clustered more  
345 closely to zero than all of the other models with relatively little spread; this was also supported  
346 by the DIC. The disparity in model performance of the different levels of allometric complexity  
347 is greatly decreased in the full model compared with the basic model with a much smaller  
348 difference in mean RMSE between the one and ten allometry models (basic: 0.0082; full:  
349 0.0012).

### 350 **Individual performance of model features**

351 Each feature in our model led to improvements in model fit, based on comparisons with the  
352 independent test dataset, compared to the basic model (Table 1; Fig. 6), with the exception of the  
353 crown overlap feature. The crown shape adjustment term led to the greatest improvement in fit  
354 almost halving the mean RMSE in the single allometry model, whilst gap fraction only decreased  
355 the mean RMSE slightly and crown overlap had no effect at all when included independently in  
356 the basic model (Table 1 and inset of Fig. 6).

357 *Gap fraction:* The proportion of ground returns was observed to decrease exponentially in  
358 relation to the total canopy area of the forest stands (Fig. 7a). Returns are still reaching the  
359 ground when the crown area exceeds the plot area (about 10% when the total crown area is equal

360 to the plot area), but only about 2% of returns come off the ground when the total crown area  
361 exceeds 1.5 times the plot area.

362 *Crown taper:* Crown shapes generated to maximize similarity between  $HD_{LIDAR}$  and  $HD_{PREDICT}$   
363 were different from the shapes expected from published allometries for six of the ten species  
364 groups (Fig. 8; parameters in Table S.5 ). Balsam fir, ironwood and red maple were predicted to  
365 be more concave (conical) than the original crown shape estimates ( $\beta$  increased), whilst beech,  
366 hemlock and yellow birch were predicted to be more convex (less conical;  $\beta$  decreased). The  
367 crown shapes for sugar maple, white spruce, broadleaf and conifer changed very little from the  
368 original estimates (Fig. 8 compares the original crown shapes to the new crown shapes); these  
369 were all originally parameterised from large sample sizes. With the exception of balsam fir, the  
370 remaining species were originally parameterised using relatively small sample sizes given the  
371 study area was dominated by sugar maples at a prevalence ranging from 63-82% across different  
372 size classes (Caspersen et al. 2011).

373 *Crown permeability:* The tree crowns were estimated to be 65% permeable within a 1 m interval  
374 to LiDAR first returns just entering the crown in the species-specific full model (Fig. 7b); this  
375 value was approximately the same as the estimated permeability when no other features were  
376 incorporated into the model. Crown permeability was the second most influential feature, after  
377 crown taper, when added independently to the basic model (inset of Fig. 6).

378 *Crown overlap:* The crown overlap feature was not useful in improving the similarity between  
379  $HD_{LIDAR}$  and  $HD_{PREDICT}$  when considered alone (inset of Fig. 6), with the crown overlap function  
380 being fitted as a horizontal line at one. When considered in the full model, the crown overlap  
381 function proved to be important in increasing the performance of the model (DIC of the full  
382 single allometry model with crown overlap: -19 057; DIC of the full single allometry model

383 without crown overlap: -18 759). In the full model, crown overlap will tend to increase from the  
384 top of the canopy to the base, which is represented in Fig. 7c. Deeper into the canopy, the  
385 cumulative canopy area will become increasingly large and therefore more of the ECA will be  
386 shaded or overlapped by a neighbouring crown; the overlap correction factor serves to decrease  
387 the ECA as the cumulative canopy area increases. When the cumulative canopy area is equal to  
388 the plot area, approximately 55% of the ECA is recorded by the LiDAR first returns and  
389 therefore 45% of the ECA is predicted to be overlapped or overtopped. In a dense plot, we  
390 expect there to be a greater degree of crown overlap and therefore the more severe penalties in  
391 Fig. 7c would be reached, whereas the ECA in sparser plots would be less severely penalised  
392 overall due to a lower expected degree of overlap amongst crowns.

393

## 394 **Discussion**

395 Airborne LiDAR provides an efficient means of obtaining high-resolution information on forest  
396 structure over forested landscapes, but it is not at present being used to predict tree size  
397 distributions by foresters or ecologists. This paper takes an important step towards developing an  
398 approach for mapping size-distributions remotely, by demonstrating the close links between the  
399 height distribution of LiDAR first returns and SDDs measured in inventory plots. A model based  
400 solely on the basic geometry of tree crowns was not very accurate, but adding a few tuneable  
401 parameters to accommodate for gap fraction, variable crown form, crown permeability and  
402 crown overlap provided enough flexibility to achieve a great improvement in goodness-of-fit.  
403 Here we discuss these tuneable features in more detail, before discussing how the approach could  
404 be used to map size distributions from aircraft.

405

406 **Modelling crown form is essential for accurate HD<sub>LiDAR</sub> prediction**

407 We found predictions based on published crown-shape functions were not sufficient for  
408 replicating the crowns as captured by LiDAR, and that including a tuneable parameter was  
409 essential in order for accuracy. Refining the overall crown shape was more important than  
410 adding detailed information about crown allometries of individual species: a single crown shape  
411 with a tuneable parameter was better supported statistically than having different allometries not  
412 fitted to the LiDAR data for each of the ten species (Table 1). Having an accurate knowledge of  
413 crown shape has been shown to be important for predicting LiDAR pulse height distributions  
414 (Nelson 1997, Sun and Ranson 2000). Van Leeuwen, Coops, and Wulder (2010) developed an  
415 approach for fitting together simple geometric shapes to provide the best approximation of a  
416 given canopy height model and van Leeuwen et al. (2013) used a flexible cone shape to fit to  
417 ground LiDAR data of a coniferous forest. Our findings support the use of simple shapes, since  
418 we found no discernible difference between the predictive accuracy of a model based on only  
419 two allometries compared with models using ten.

420

421 **Crown permeability, overlap and gap fraction**

422 Crown permeability to LiDAR pulses and crown overlap influenced how canopies were  
423 manifested in the point cloud (Li et al. 2012) and thus proved an important property to capture in  
424 the model. The basic model assumes that LiDAR pulses are reflected off the outer shell of the  
425 canopy whereas in reality some LiDAR pulses penetrate into the crown before being first  
426 reflected (Gaveau and Hill 2003, Maltamo et al. 2004). For example, Disney et al. (2010)  
427 estimated that true canopy height was significantly underestimated (~ 4% for broadleaves and ~

428 16% for conifers) as a result of pulse penetration. Our model predicts 65% of first returns will be  
429 reflected within 1 m from where they entered the crown suggesting there is a high degree of  
430 permeation through the crown. Penetration of light is likely to vary with tree size (because leaf  
431 area index is size dependent; Nock, Caspersen, and Thomas 2008) and with the degree of foliage  
432 clumping (Goodwin et al. 2007). Since our model considers the stand as a whole, as opposed to  
433 considering individual trees, these effects might be averaged out, particularly in the full model  
434 where the permeability feature will have some trade-off with the crown overlap correction factor.

435         The crown overlap function was found to be useful when used in combination with the  
436 other features, but redundant when used alone in the basic model. The interplay with the refined  
437 crowns, predicted gaps and permeable foliage meant crown overlap was then an important  
438 feature to include. The crowns overlap more the deeper into the canopy they are with the degree  
439 of overlap determined by the canopy area overhead; a sparsely populated plot will be minimally  
440 penalised, even close to the ground, since the cumulative crown area will be low relative to the  
441 plot area. Competition among crowns in plots with high stem densities has a major influence on  
442 crown form (Canham et al. 2004, Lines et al. 2010, Coomes et al. 2011). The crowns in the basic  
443 model were fixed, based on dbh and species, but the crown overlap function serves to adapt these  
444 fixed predictions to lower the probability of a return being made when the environment  
445 effectively becomes “more competitive”. Whether it is the case that two crowns are simply  
446 overlapping or that one of the crowns has dominated and the other has dropped its foliage in  
447 response (Garber and Maguire 2005; Purves, Lichstein, and Pacala 2007), the ultimate effect on  
448 LiDAR returns is the same. Goodwin, Coops, and Culvenor (2007) noted that the differing  
449 crown shapes in the field will also cause returns to be made from regions other than the fixed  
450 predicted exposed area. Our model does not explicitly measure at which point a crown will

451 become overtopped, nor which tree is overtopped by the other, but the overlap correction factor  
452 assigns a probability that is shared amongst all trees at a given height that they have been  
453 overtopped. The overall degree of overlap in a plot is determined by the canopy area and is  
454 therefore interlinked with the gap fraction.

455         Gap fraction has important influences on LiDAR return distributions, but proved difficult  
456 to capture without any explicit consideration of the location of trees within plots. Our attempts to  
457 model gap fraction using the total crown area was not universally successful, because some plots  
458 contained waterlogged patches within which tree cover was discontinuous (RS, personal  
459 observation) and these patches produced many more ground returns than predicted by our  
460 canopy-cover function (see Bugmann 2001 for a discussion of the link between gap fraction and  
461 canopy cover). An approach, not explored here, is to use the LiDAR point cloud to identify gaps  
462 and use this information in the predictive models (Koukoulas and Blackburn 2004, Gaulton and  
463 Malthus 2010).

464         It is clear that predicting LiDAR height-return distributions from inventory data requires  
465 an understanding of canopy structure and the interaction of LiDAR pulses with the canopy. How  
466 can we use this link to make informative predictions of SDDs across large scales?

467

#### 468 **Mapping SDDs from LiDAR data**

469 Using a mechanistic approach, we have shown close links between SDDs and  $HD_{LIDAR}$ ,  
470 suggesting that it should be possible to map diameter distributions, using airborne LiDAR. In  
471 principle, the LiDAR dataset could be carved up into plot-sized tiles, and the most likely tree size  
472 structure of each tile inferred using the link function developed here (Fig. 3): the  $HD_{LIDAR}$  of a  
473 given tile would be calculated, and the most likely size structure would be obtained by predicting

474 the  $HD_{LIDAR}$  for a particular theoretical SDD, comparing that prediction with the  $HD_{LIDAR}$ , and  
475 then continuing to search through many theoretical SDDs until a close match is found between  
476 observed and predicted  $HD_{LIDAR}$  (see Fig. S.1). Given the accuracy of our model and the  
477 likelihood that multiple SDDs will produce the same  $HD_{LIDAR}$ , our model will be able to produce  
478 a subset of possible SDDs.

479 Mapping of size distributions from LiDAR data will only work if additional layers of  
480 information are provided to the algorithm to help refine its predictions. A single SDD in different  
481 locations will not always produce the same  $HD_{LIDAR}$  as a result of allometric noise; although we  
482 are not proposing to invert the model, but to use it directly to predict an array of SDD-to-  
483  $HD_{PREDICT}$  pairings, noise is still likely to pose an issue (Comerón et al. 2004, Clewley et al.  
484 2012). While our model alone is therefore not adequate to predict the precise structure and  
485 composition of a given stand, it does provide efficient and reliable predictions within a Bayesian  
486 framework that incorporates other sources of information. For example, our study area had  
487 previously been classified into forest types using imagery (Franklin et al. 2000) and including  
488 this information as a prior in the Bayesian analysis can substantially improve its predictive  
489 accuracy. Previous studies have also used statistical relationships drawn from LiDAR metrics to  
490 achieve predictions of plot level statistics which can also be used to better constrain our model  
491 (e.g. Popescu, Wynne, and Nelson 2003; Asner et al. 2010). In this way, we propose drawing on  
492 multiple different sources of information within a flexible framework to make the most confident  
493 predictions when applying our model to estimate SDDs.

494 Recognising individual tree crowns in the LiDAR imagery (“crown segmentation”) is an  
495 alternative approach to the one we have adopted, and has the advantage of providing information  
496 on the location of trees within each plot (Koch et al. 2006). However, this approach has its

497 limitations: it is hard to detect understory trees (Maltamo et al. 2004) and segmentation is often  
498 difficult in dense, deciduous stands (Koch et al. 2006) , although predictions can be improved by  
499 fusing LiDAR with multispectral data (Popescu et al. 2003). Incorporating spatial information  
500 into the model would also improve the predictions made by the gap-fraction feature, (Gaulton  
501 and Malthus 2010), although the extra complexity might compromise the applicability of the  
502 model to large scale mapping.

503

#### 504 **Generalising the approach to other forest types**

505 A significant benefit of developing a model from first principles is that it can be tuned for  
506 applications in different forest types. None of the features included in the model are specific to  
507 the study area, so the model can be re-calibrated for a new study area in which a set of  
508 inventoried plots and the associated LiDAR first return distributions are available. The only  
509 requirements to use the model are reliable measurements, or an existing allometry, predicting the  
510 height and crown size, but not crown taper, for the predominant species; our work suggests that  
511 using published allometries from one or two common species will be sufficient for this as  
512 including ten allometries did not substantially improve model performance. All of the parameters  
513 in our model can be estimated using the non-spatial ground data and low density LiDAR  
514 distributions.

515 We used first returns only in our model to minimise the effects of instrument type (Næsset  
516 2004), recognising that power, pulse frequency, footprint and scan angle all affect signal (Disney  
517 et al. 2010) . The generality of LiDAR derived methods for estimating canopy attributes can be  
518 affected by differences in LiDAR acquisition specifications (Hopkinson 2007). First returns are  
519 most clearly associated with the outer canopy shell, and thus can be related directly to crown

520 allometries. Subsequent returns are reliant on interior canopy properties, such as foliage  
521 clumping (Ni-Meister et al. 2001), and are consequently much more complicated to model  
522 (North et al. 2010). Under different specifications, the internal structure of the canopy will be  
523 captured very differently, and therefore would be hard to represent with a generic model. By  
524 working only with the first returns and summarising these returns into a height distribution, we  
525 reduce the problems associated with different LiDAR devices and increase the applicability of  
526 our model to other temperate or coniferous forest types.

527 LiDAR offers wall-to-wall data for describing heterogeneous forest landscapes that are  
528 difficult to sample with ground plots (Asner et al. 2010, Mitchard et al. 2014). Our study  
529 represents an important step in the development of approaches for mapping SDDs.

530

## 531 **Acknowledgements**

532 We would like to thank the following people for their fieldwork efforts: Elaine Mallory, Jason  
533 Kerr, Michelle Bowman, Gareth Cockwell, Cheryl Widdifield, Assunta Saliola, Alecia  
534 Korkowski, Adam Gorgolewski, Mike Gillespie and Matt Thiel. We thank Drew Purves for his  
535 helpful comments on an earlier version of this manuscript. We thank the anonymous reviewers  
536 for their valuable contributions to improving the manuscript. RAS was funded by a Microsoft  
537 Research scholarship. NSERC, Ontario Power Generation and Haliburton Forest funded other  
538 aspects of this work.

539

## 540 **References**

541 Asner, G. P., J. Mascaro, H. C. Muller-Landau, G. Vieilledent, R. Vaudry, M. Rasamoelina, J. S.  
542 Hall, and M. van Breugel. 2012. A universal airborne LiDAR approach for tropical forest  
543 carbon mapping. *Oecologia* 168:1147–1160.

544 Asner, G. P., G. V. N. Powell, J. Mascaro, D. E. Knapp, J. K. Clark, J. Jacobson, T. Kennedy-  
545 Bowdoin, A. Balaji, G. Paez-Acosta, E. Victoria, L. Secada, M. Valqui, and R. F. Hughes.  
546 2010. High-resolution forest carbon stocks and emissions in the Amazon. *Proceedings of*  
547 *the National Academy of Sciences of the United States of America* 107:16738–42.

548 Breidenbach, J., C. Gläser, and M. Schmidt. 2008. Estimation of diameter distributions by means  
549 of airborne laser scanner data. *Canadian Journal of Forest Research* 38:1611–1620.

550 Bugmann, H. 2001. A review of forest gap models. *Climatic Change* 51:259–305.

551 Canadian climate normals 1971-2000: Haliburton, Ontario. 2012. National Climate Data and  
552 Information Archive.

553 Canham, C. D., P. T. Lepage, and K. D. Coates. 2004. A neighborhood analysis of canopy tree  
554 competition : effects of shading versus crowding *787:778–787*.

555 Caspersen, J. P., M. C. Vanderwel, W. G. Cole, and D. W. Purves. 2011. How stand productivity  
556 results from size- and competition-dependent growth and mortality. *PLoS ONE* 6.

557 Chasmer, L., C. Hopkinson, and P. Treitz. 2006. Investigating laser pulse penetration through a  
558 conifer canopy by integrating airborne and terrestrial lidar. *Canadian Journal of Remote*  
559 *Sensing* 32:116–125.

560 Clewley, D., R. Lucas, M. Moghaddam, and P. Bunting. 2012. The effects of noise on model  
561 inversion for the retrieval of forest structure from SAR data. Pages 7173–7176 IEEE  
562 International Geoscience and Remote Sensing Symposium (IGARSS). IEEE Press, New  
563 York.

564 Comerón, A., F. Rocadenbosch, M. A. López, A. Rodríguez, C. Muñoz, D. García-Vizcaíno, and  
565 M. Sicard. 2004. Effects of noise on lidar data inversion with the backward algorithm.  
566 *Applied Optics* 43:2572–2577.

567 Coomes, D. A., and R. B. Allen. 2007. Mortality and tree-size distributions in natural mixed-age  
568 forests *95:27–40*.

569 Coomes, D. A., E. R. Lines, and R. B. Allen. 2011. Moving on from Metabolic Scaling Theory:  
570 hierarchical models of tree growth and asymmetric competition for light. *Journal of*  
571 *Ecology* 99:748–756.

572 Dalponte, M., L. Bruzzone, and D. Gianelle. 2011. A system for the estimation of single-tree  
573 stem diameter and volume using multireturn LiDAR data. *IEEE Transactions on*  
574 *Geoscience and Remote Sensing* 49:2479–2490.

575 Disney, M. I., V. Kalogerou, P. Lewis, a. Prieto-Blanco, S. Hancock, and M. Pfeifer. 2010.  
576 Simulating the impact of discrete-return lidar system and survey characteristics over young  
577 conifer and broadleaf forests. *Remote Sensing of Environment* 114:1546–1560.

578 Franklin, S. E., R. J. Hall, L. M. Moskal, a. J. Maudie, and M. B. Lavigne. 2000. Incorporating  
579 texture into classification of forest species composition from airborne multispectral images.  
580 International Journal of Remote Sensing 21:61–79.

581 Garber, S. M., and D. A. Maguire. 2005. The response of vertical foliage distribution to spacing  
582 and species composition in mixed conifer stands in central Oregon. Forest Ecology and  
583 Management 211:341–355.

584 Gaulton, R., and T. J. Malthus. 2010. LiDAR mapping of canopy gaps in continuous cover  
585 forests: A comparison of canopy height model and point cloud based techniques.  
586 International Journal of Remote Sensing 31:1193–1211.

587 Gaveau, D. L. A., and R. A. Hill. 2003. Quantifying canopy height underestimation by laser  
588 pulse penetration in small-footprint airborne laser scanning data. Canadian Journal of  
589 Remote Sensing 29:650–657.

590 Goodwin, N. R., N. C. Coops, and D. S. Culvenor. 2007. Development of a simulation model to  
591 predict LiDAR interception in forested environments. Remote Sensing of Environment  
592 111:481–492.

593 Hill, A., J. Breschan, and D. Mandallaz. 2014. Accuracy Assessment of Timber Volume Maps  
594 Using Forest Inventory Data and LiDAR Canopy Height Models. Forests:2253–2275.

595 Hopkinson, C. 2007. The influence of flying altitude, beam divergence, and pulse repetition  
596 frequency on laser pulse return intensity and canopy frequency distribution. Canadian  
597 Journal of Remote Sensing 33:312–324.

598 Hyypä, J., O. Kelle, M. Lehtikoinen, and M. Inkinen. 2001. A segmentation-based method to  
599 retrieve stem volume estimates from 3-D tree height models produced by laser scanners.  
600 IEEE Transactions on Geoscience and Remote Sensing 39:969–975.

601 Jakubowski, M. K., Q. Guo, and M. Kelly. 2013. Tradeoffs between lidar pulse density and  
602 forest measurement accuracy. Remote Sensing of Environment 130:245–253.

603 Jochem, A., M. Hollaus, M. Rutzinger, and B. Höfle. 2010. Estimation of Aboveground Biomass  
604 in Alpine Forests: A Semi-Empirical Approach Considering Canopy Transparency Derived  
605 from Airborne LiDAR Data. Sensors 11:278–295.

606 Jones, T., M. Woods, and K. Lim. 2009. Quantifying diameter and basal area distributions of  
607 uneven-aged tolerant hardwood stands using low density LiDAR. Silvilaser 2009, Oct 14–  
608 16.

609 Kirschbaum, M. U. F. 1999. CenW, a forest growth model with linked carbon, energy, nutrient  
610 and water cycles. Ecological Modelling 118:17–59.

611 Koch, B., U. Heyder, and H. Weinacker. 2006. Detection of Individual Tree Crowns in Airborne  
612 Lidar Data 72:357–363.

613 Köhler, P., and A. Huth. 1998. The effects of tree species grouping in tropical rain forest  
614 modelling Simulations with the individual based model Formind. Ecological Modelling  
615 109:301–321.

- 616 Koukoulas, S., and G. A. Blackburn. 2004. Quantifying the spatial properties of forest canopy  
617 gaps using LiDAR imagery and GIS. *International Journal of Remote Sensing* 25:3049–  
618 3072.
- 619 Van Leeuwen, M., N. C. Coops, T. Hilker, M. a. Wulder, G. J. Newnham, and D. S. Culvenor.  
620 2013. Automated reconstruction of tree and canopy structure for modeling the internal  
621 canopy radiation regime. *Remote Sensing of Environment* 136:286–300.
- 622 Van Leeuwen, M., N. C. Coops, and M. a. Wulder. 2010. Canopy surface reconstruction from a  
623 LiDAR point cloud using Hough transform. *Remote Sensing Letters* 1:125–132.
- 624 Van Leeuwen, M., and M. Nieuwenhuis. 2010. Retrieval of forest structural parameters using  
625 LiDAR remote sensing. *European Journal of Forest Research* 129:749–770.
- 626 Li, W., Q. Guo, M. K. Jakubowski, and M. Kelly. 2012. A New Method for Segmenting  
627 Individual Trees from the Lidar Point Cloud 78:75–84.
- 628 Lim, K., P. Treitz, M. Wulder, B. St-Onge, and M. Flood. 2003. LiDAR remote sensing of forest  
629 structure. *Progress in Physical Geography* 27:88–106.
- 630 Lines, E. R., D. A. Coomes, and D. W. Purves. 2010. Influences of forest structure, climate and  
631 species composition on tree mortality across the eastern US. *PloS one* 5:e13212.
- 632 Maltamo, M., K. Eerikäinen, P. Packalén, and J. Hyyppä. 2006. Estimation of stem volume using  
633 laser scanning-based canopy height metrics. *Forestry* 79:217–229.

634 Maltamo, M., K. Mustonen, J. Hyypä, J. Pitkänen, and X. Yu. 2004. The accuracy of estimating  
635 individual tree variables with airborne laser scanning in a boreal nature reserve. *Canadian*  
636 *Journal of Forest Research* 34:1791–1801.

637 Mitchard, E. T. A., T. R. Feldpausch, R. J. W. Brienen, G. Lopez-Gonzalez, A. Monteagudo, T.  
638 R. Baker, S. L. Lewis, J. Lloyd, C. A. Quesada, M. Gloor, H. ter Steege, P. Meir, E.  
639 Alvarez, A. Araujo-Murakami, L. E. O. C. Aragão, L. Arroyo, G. Aymard, O. Banki, D.  
640 Bonal, S. Brown, F. I. Brown, C. E. Cerón, V. Chama Moscoso, J. Chave, J. a. Comiskey,  
641 F. Cornejo, M. Corrales Medina, L. Da Costa, F. R. C. Costa, A. Di Fiore, T. F. Domingues,  
642 T. L. Erwin, T. Frederickson, N. Higuchi, E. N. Honorio Coronado, T. J. Killeen, W. F.  
643 Laurance, C. Levis, W. E. Magnusson, B. S. Marimon, B. H. Marimon Junior, I. Mendoza  
644 Polo, P. Mishra, M. T. Nascimento, D. Neill, M. P. Núñez Vargas, W. A. Palacios, A.  
645 Parada, G. Pardo Molina, M. Peña-Claros, N. Pitman, C. A. Peres, L. Poorter, A. Prieto, H.  
646 Ramirez-Angulo, Z. Restrepo Correa, A. Roopsind, K. H. Roucoux, A. Rudas, R. P.  
647 Salomão, J. Schiatti, M. Silveira, P. F. de Souza, M. K. Steininger, J. Stropp, J. Terborgh, R.  
648 Thomas, M. Toledo, A. Torres-Lezama, T. R. van Andel, G. M. F. van der Heijden, I. C. G.  
649 Vieira, S. Vieira, E. Vilanova-Torre, V. A. Vos, O. Wang, C. E. Zartman, Y. Malhi, and O.  
650 L. Phillips. 2014. Markedly divergent estimates of Amazon forest carbon density from  
651 ground plots and satellites. *Global Ecology and Biogeography* 23:935–946.

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

654 Næsset, E. 2004. Effects of different flying altitudes on biophysical stand properties estimated  
655 from canopy height and density measured with a small-footprint airborne scanning laser.  
656 Remote Sensing of Environment 91:243–255.

657 Nelson, R. 1997. Modeling forest canopy heights: The effects of canopy shape. Remote Sensing  
658 of Environment 60:327–334.

659 Ni-Meister, W., D. L. B. Jupp, and R. Dubayah. 2001. Modeling Lidar Waveforms in  
660 Heterogeneous and Discrete Canopies. IEEE Transactions on Geoscience and Remote  
661 Sensing 39:1943–1958.

662 Nock, C. a, J. P. Caspersen, and S. C. Thomas. 2008. Large ontogenetic declines in intra-crown  
663 leaf area index in two temperate deciduous tree species. Ecology 89:744–53.

664 North, P., J. Rosette, J. Suárez, and S. Los. 2010. A Monte Carlo radiative transfer model of  
665 satellite waveform LiDAR. International Journal of Remote Sensing 31:1343–1358.

666 Parker, G. R., D. J. Leopold, and J. K. Eichenberger. 1985. Tree dynamics in an old-growth,  
667 deciduous forest. Forest Ecology and Management 11:31–57.

668 Piboule, A., C. Collet, H. Frochot, and J. Dhôte. 2005. Reconstructing crown shape from stem  
669 diameter and tree position to supply light models . I. Algorithms and comparison of light  
670 simulations. Annals of Forest Science 62:645–657.

671 Popescu, S. C., R. H. Wynne, and R. F. Nelson. 2003. Measuring individual tree crown diameter  
672 with lidar and assessing its influence on estimating forest volume and biomass. Canadian  
673 Journal of Remote Sensing 29:564–577.

674 Purves, D., and V. Lyutsarev. 2011. Filzbach user guide version 1.1.

675 Purves, D. W., J. W. Lichstein, and S. W. Pacala. 2007. Crown plasticity and competition for  
676 canopy space: a new spatially implicit model parameterized for 250 North American tree  
677 species. *PloS one* 2:e870.

678 Purves, D. W., J. W. Lichstein, N. Strigul, and S. W. Pacala. 2008. Predicting and understanding  
679 forest dynamics using a simple tractable model. *Proceedings of the National Academy of  
680 Sciences of the United States of America* 105:17018–22.

681 Scheller, R. M., and D. J. Mladenoff. 2004. A forest growth and biomass module for a landscape  
682 simulation model, LANDIS: design, validation, and application. *Ecological Modelling*  
683 180:211–229.

684 Sullivan, T. P., D. S. Sullivan, and P. M. F. Lindgren. 2001. Stand Structure and Small Mammals  
685 in Young Lodgepole Pine Forest : 10-Year Results after Thinning 11:1151–1173.

686 Sun, G., and K. J. Ranson. 2000. Modeling Lidar Returns from Forest Canopies. *IEEE  
687 Transactions on Geoscience and Remote Sensing* 38:2617–2626.

688 Thomas, V., R. D. Oliver, K. Lim, and M. Woods. 2008. LiDAR and Weibull modeling of  
689 diameter and basal area. *Forestry* 84:866–875.

690 Valbuena, R., P. Packalen, L. Mehtätalo, A. García-Abril, and M. Maltamo. 2013. Characterizing  
691 forest structural types and shelterwood dynamics from Lorenz-based indicators predicted by  
692 airborne laser scanning. *Canadian Journal of Forest Research* 43:1063–1074.

693 Vanderwel, M. C., H. C. Thorpe, J. L. Shuter, J. P. Caspersen, and S. C. Thomas. 2008.  
694 Contrasting downed woody debris dynamics in managed and unmanaged northern  
695 hardwood stands. *Canadian Journal of Forest Research* 38:2850–2861.

696 Vauhkonen, J., and L. Mehtätalo. 2015. Matching remotely sensed and field-measured tree size  
697 distributions. *Canadian Journal of Forest Research* 45:353–363.

698 Vincent, G., D. Sabatier, L. Blanc, J. Chave, E. Weissenbacher, R. Péliissier, E. Fonty, J. F.  
699 Molino, and P. Coueron. 2012. Accuracy of small footprint airborne LiDAR in its  
700 predictions of tropical moist forest stand structure. *Remote Sensing of Environment*  
701 125:23–33.

702 Xu, Q., Z. Hou, M. Maltamo, and T. Tokola. 2014. Calibration of area based diameter  
703 distribution with individual tree based diameter estimates using airborne laser scanning.  
704 *ISPRS Journal of Photogrammetry and Remote Sensing* 93:65–75.

705 Yao, W., P. Krzystek, and M. Heurich. 2012. Tree species classification and estimation of stem  
706 volume and DBH based on single tree extraction by exploiting airborne full-waveform  
707 LiDAR data. *Remote Sensing of Environment* 123:368–380.

708 Zhao, K., S. Popescu, and R. Nelson. 2009. Lidar remote sensing of forest biomass: A scale-  
709 invariant estimation approach using airborne lasers. *Remote Sensing of Environment*  
710 113:182–196.

711

712

**Table 1** A performance summary of the different versions of the model. The deviance information criterion (DIC) estimated from the model-fitting plots and summary statistics for the RMSE measured for each test plot is presented for each model (the percentage RMSE is the mean RMSE as a proportion of the range of true values; see equations (S. 2) and (S. 3)). A decrease in both DIC and RMSE corresponds to an improvement in the model.

Model version	No. crown allometries	No. parameters fitted using LiDAR*	DIC	Mean RMSE $\pm$ standard deviation (% RMSE)
Basic allometric	1	0		0.0433 $\pm$ 0.0091 (39.7%)
	2	0		0.0388 $\pm$ 0.0085 (35.3%)
	10	0	-	0.0351 $\pm$ 0.0092 (32.1%)
Gap fraction	1	1	-13 744	0.0417 $\pm$ 0.0090 (38.3%)
	2	1	-14 884	0.0372 $\pm$ 0.0086 (34.0%)
	10	1	-15 596	0.0335 $\pm$ 0.0093 (30.7%)
Crown taper	1	1	-18 253	0.0231 $\pm$ 0.0094 (20.2%)
	2	2	-18 684	0.0221 $\pm$ 0.0077 (19.5%)
	10	10	-18 954	0.0217 $\pm$ 0.0078 (19.1%)
Crown permeability	1	1	-14 345	0.0374 $\pm$ 0.0010 (34.4%)
	2	1	-15 611	0.0331 $\pm$ 0.0085 (30.2%)
	10	1	-16 388	0.0296 $\pm$ 0.0094 (27.2%)
Crown overlap	1	1	-13 374	0.0433 $\pm$ 0.0091 (39.7%)
	2	1	-14 483	0.0388 $\pm$ 0.0085 (35.3%)
	10	1	-15 121	0.0351 $\pm$ 0.0092 (32.1%)
Full	1	4	-19 057	0.0208 $\pm$ 0.0091 (18.2%)
	2	5	-19 634	0.0196 $\pm$ 0.0075 (17.3%)
	10	13	-19 962	0.0196 $\pm$ 0.0074 (17.5%)

713 \*The number of fitted parameters is one greater than the numbers provided here when the error  
714 term used to calculate the likelihood is included.

715

716 **Figure captions**

717 **Fig. 1** The extent of the LiDAR data across Haliburton Forest (grey boxed area; each complete  
718 box is 1 km<sup>2</sup>) and the locations of all of the inventoried plots (black dots; note that the dots are  
719 not to scale).

720  
721 **Fig. 2** (a) Species-specific crown properties predicted from stem diameter at breast height (dbh)  
722 using allometric functions (Table S.1): these include tree height (H), crown depth (V), maximum  
723 crown radius ( $R_{\max}$ ) as well as the crown radius ( $R_h$ ) and crown area ( $CA_h$ ) at a given  
724 measurement height (h); (b)  $ECA_{[h, h+1]}$  (exposed canopy area between height h and h+1) is  
725 calculated by subtracting the crown area at h+1 from the crown area at height h.

726  
727 **Fig. 3** The link model developed in this study can take SDDs classified as conifers or  
728 broadleaves and predicts the height distribution of LiDAR first returns ( $HD_{\text{PREDICT}}$ ) from a few  
729 simple assumptions about canopy structure. We propose that this model can then be used in the  
730 inverse prediction to obtain predicted stand information from an observed LiDAR distribution  
731 ( $HD_{\text{LIDAR}}$ ).

732  
733 **Fig. 4** Histogram of the LiDAR first return distribution summarised across all of the test plots.  
734 The line overlaying the histogram represents the prediction made by the full model in each of the  
735 three allometric versions. The inset presents the deviance from the 1:1 line of the predictions  
736 made by the species-specific full model with the associated RMSE.

737

738 **Fig. 5** Individual LiDAR return distributions for twelve test plots where the histograms represent  
739 the data and the lines show the prediction made by the full species-specific model. The top row  
740 are from the top 25% best fitting plots ( $RMSE < 0.0148$ ), the middle row are from the middle  
741 50% ( $0.0148 \leq RMSE < 0.0253$ ) and the bottom row are from the worst 25% ( $RMSE \geq 0.0253$ ).  
742 The RMSE values for each plot are displayed in the top right corner.

743

744 **Fig. 6** The main figure presents the RMSE of all test plots to compare the performances of the  
745 basic model and full model (containing all features) for each of the model versions using: a  
746 single crown allometry (1), a generic allometry each for broadleaf and conifer (2) and the  
747 species-specific crown allometries (10). A RMSE value of 0 would correspond to a perfect fit.  
748 The inset figure shows the effect of incorporating the features into the basic model individually  
749 in each version of the model. The dotted horizontal lines denote the median RMSE for the single  
750 allometry basic model (upper) and the ten allometry full model (lower) for comparison; gap  
751 fraction, crown taper and crown permeability are all shown to correspond with a significant  
752 improvement in model performance, whilst crown overlap does not improve on the basic model  
753 when considered individually.

754

755 **Fig. 7** Gap fraction in a) is predicted as the proportion of first returns being recorded from the  
756 ground as a function of the total canopy area as a proportion of the plot area. The points are the  
757 actual recorded ground returns. The permeability factor is interpreted in b) where the solid  
758 circles outline the ECA in an example 1 m height interval with the grey shaded area denoting the  
759 area where first returns are recorded in that interval. The lighter shaded area gives the region  
760 where the first returns have permeated from the interval above (fitted proportion of 0.65) and the

761 hashed area gives the region where the first returns will permeate to the interval below. The  
762 crown overlap correction factor in c) is predicted from the cumulative canopy area as a  
763 proportion of the plot area.

764

765 **Fig. 8** Comparison of allometry-derived crown forms (dotted line) and fitted crown forms using  
766 the LiDAR data (solid line) for trees with a dbh of 30cm; the grey shaded area denotes the 95%  
767 confidence interval associated with the fitted crown shape parameter ( $\beta$ ).

768

769

770

771

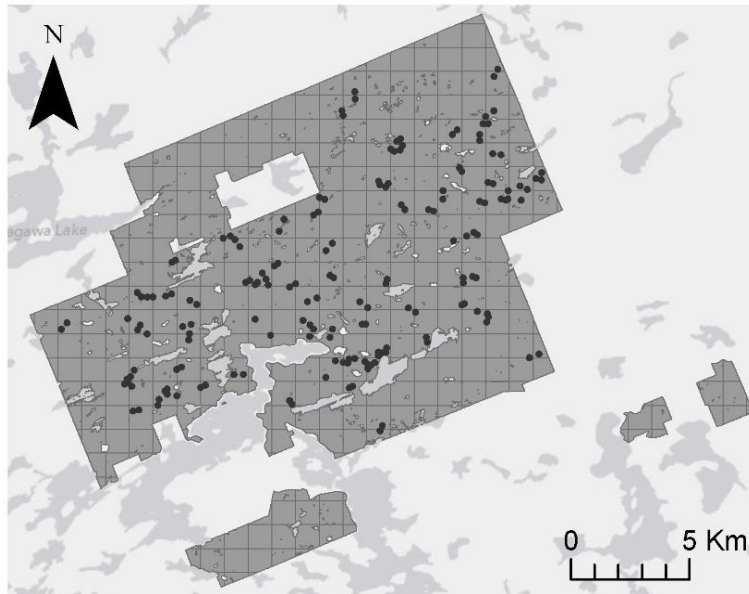
772

773

774

775

776



**Fig. 1** The extent of the LiDAR data across Haliburton Forest (grey boxed area; each complete box is 1 km<sup>2</sup>) and the locations of all of the inventoried plots (black dots; note that the dots are not to scale).

777

778

779

780

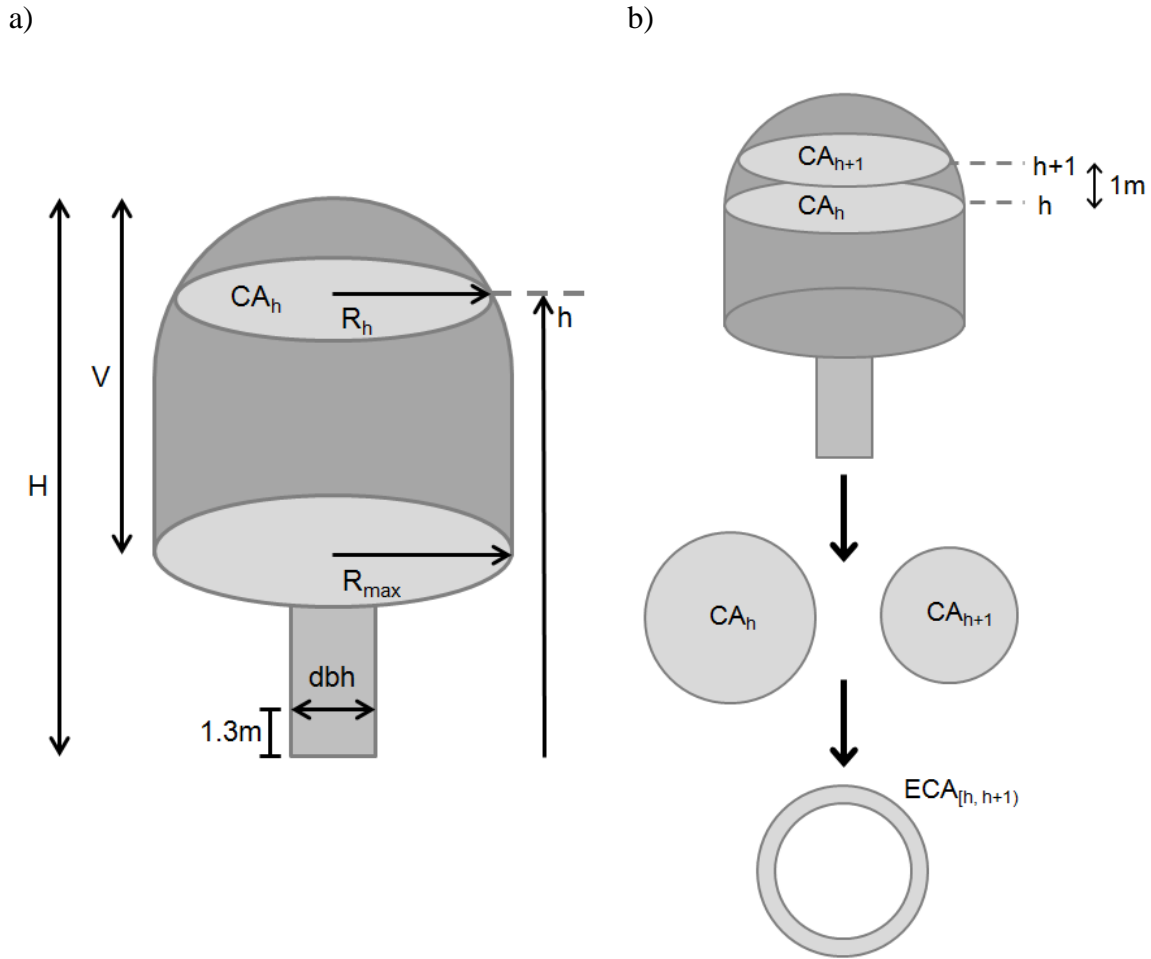
781

782

783

784

785

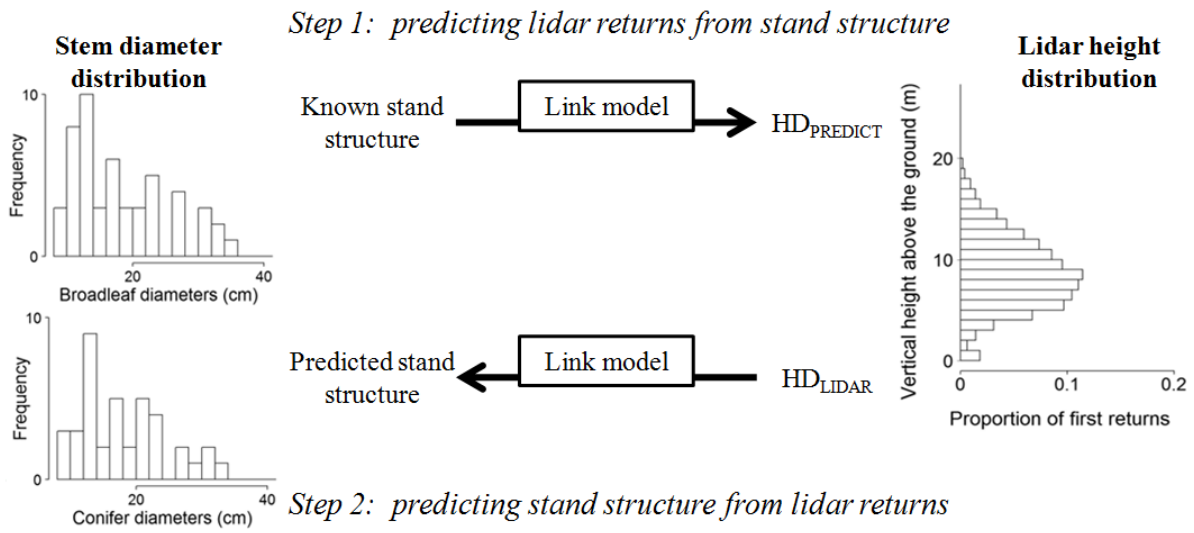


**Fig. 2** (a) Species-specific crown properties predicted from stem diameter at breast height (dbh) using allometric functions (Table S.1): these include tree height (H), crown depth (V), maximum crown radius ( $R_{max}$ ) as well as the crown radius ( $R_h$ ) and crown area ( $CA_h$ ) at a given measurement height (h); (b)  $ECA_{[h, h+1]}$  (exposed canopy area between height h and h+1) is calculated by subtracting the crown area at h+1 from the crown area at height h.

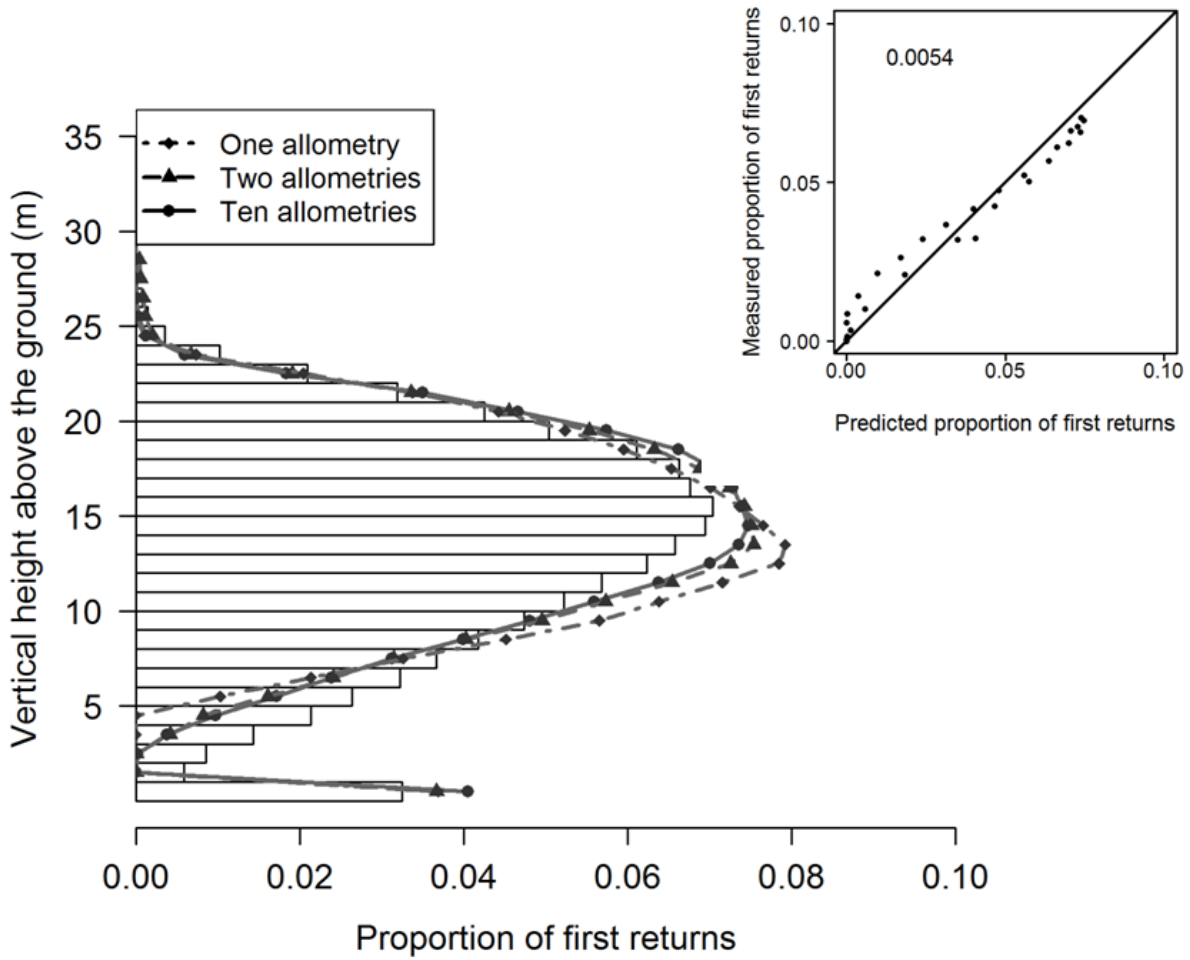
786

787

788



**Fig. 3** The link model developed in this study can take SDDs classified as conifers or broadleaves and predicts the height distribution of LiDAR first returns ( $HD_{PREDICT}$ ) from a few simple assumptions about canopy structure. We propose that this model can then be used in the inverse prediction to obtain predicted stand information from an observed LiDAR distribution ( $HD_{LIDAR}$ ).



**Fig. 4** Histogram of the LiDAR first return distribution summarised across all of the test plots. The line overlaying the histogram represents the prediction made by the full model in each of the three allometric versions. The inset presents the deviance from the 1:1 line of the predictions made by the species-specific full model with the associated RMSE.

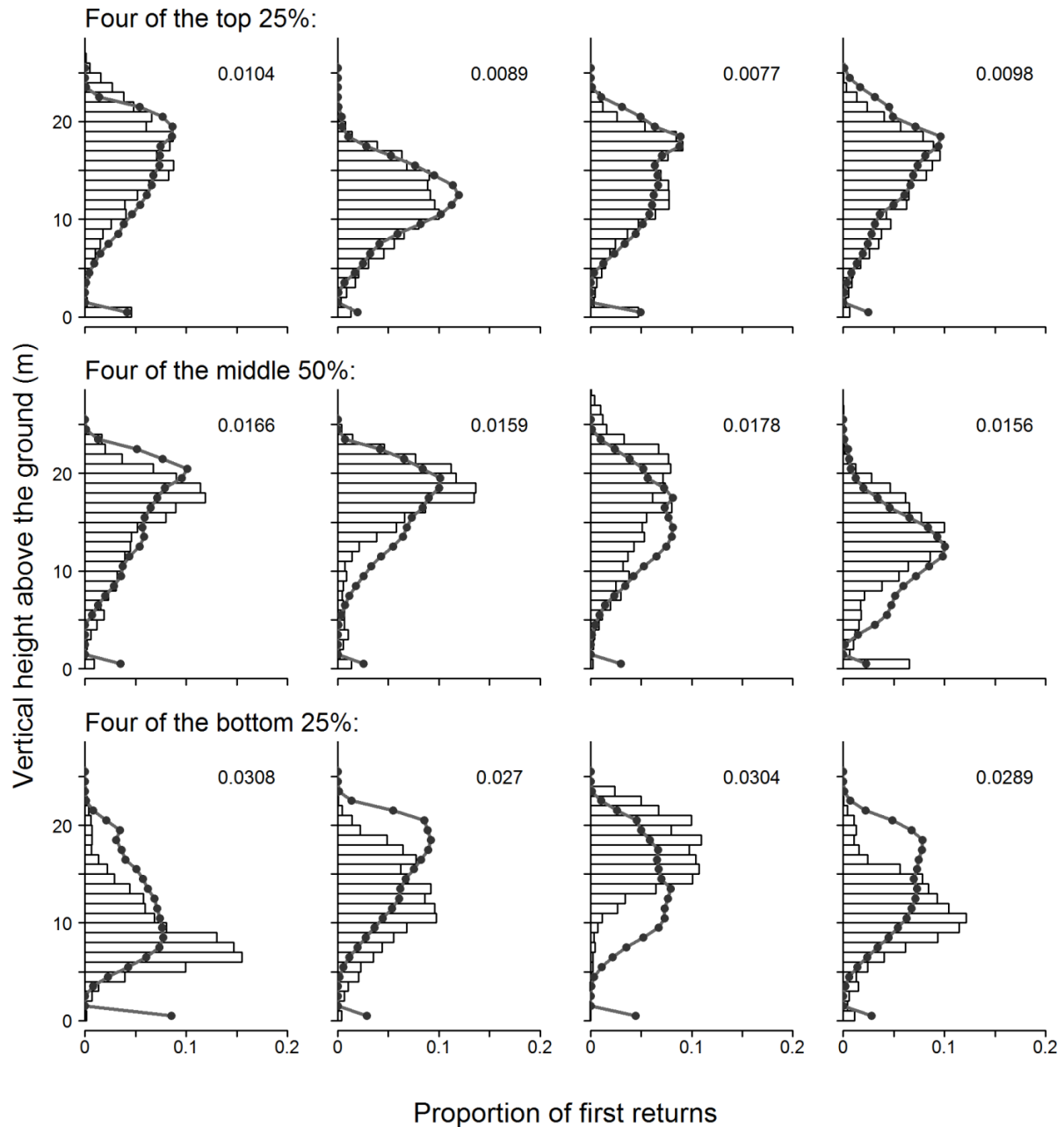
792

793

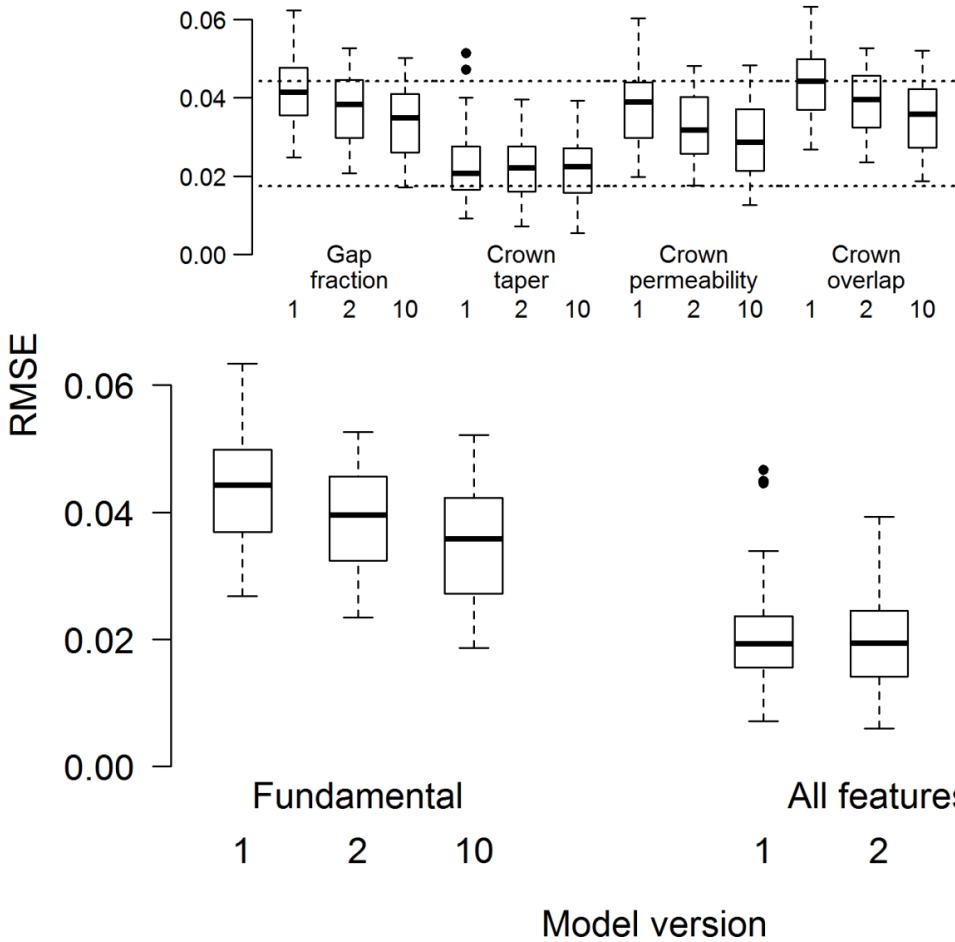
794

795

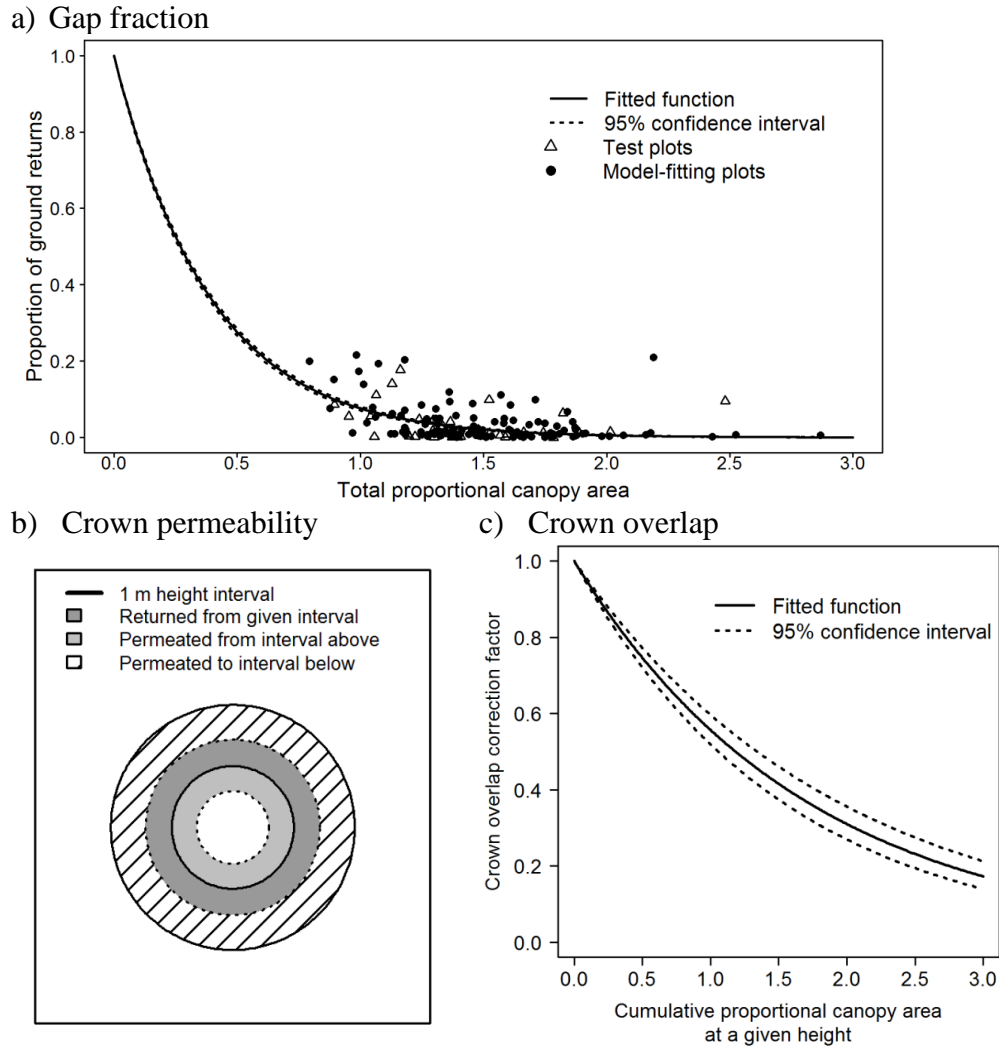
796



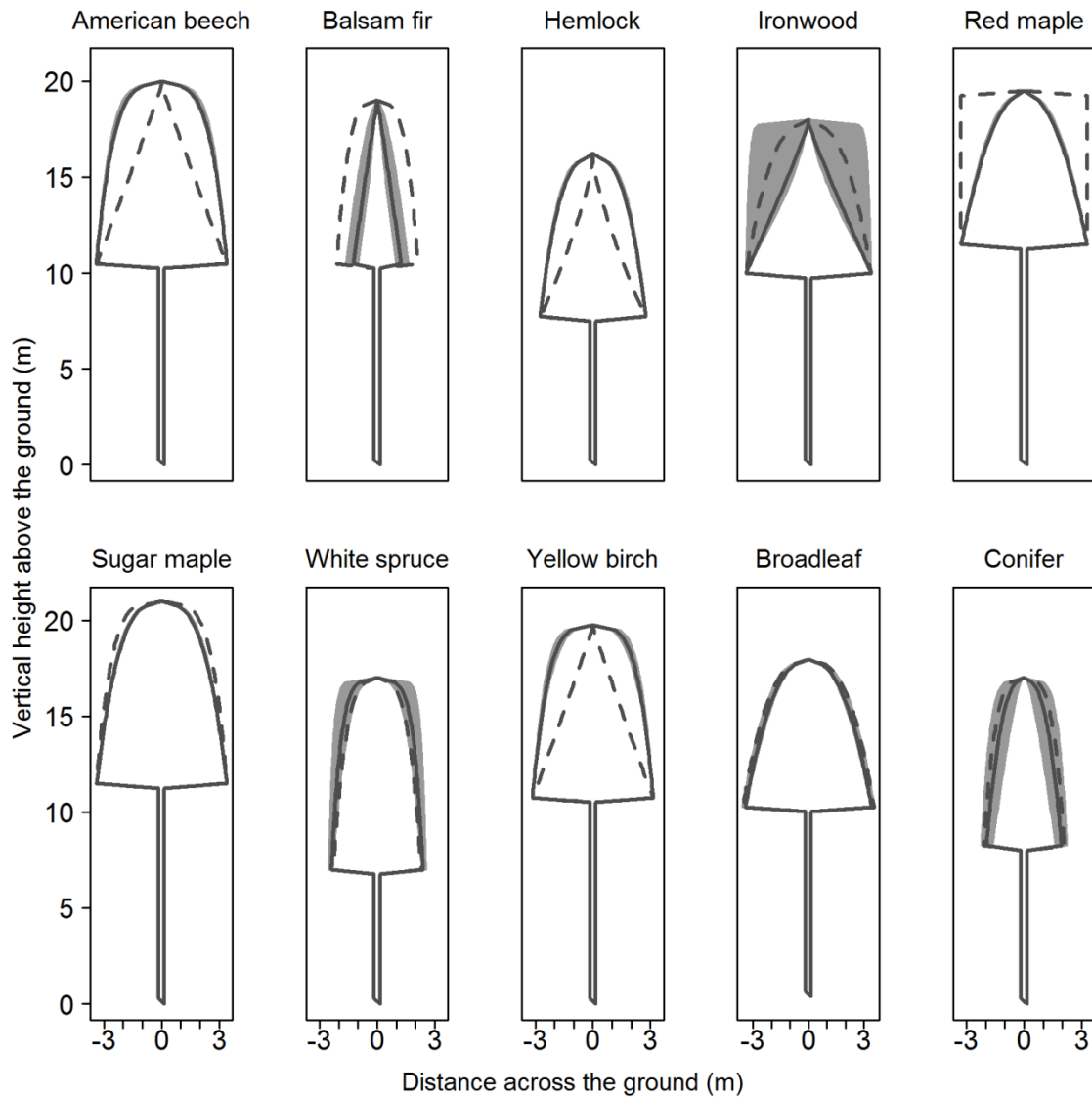
**Fig. 5** Individual LiDAR return distributions for twelve test plots where the histograms represent the data and the lines show the prediction made by the full species-specific model. The top row are from the top 25% best fitting plots ( $RMSE < 0.0148$ ), the middle row are from the middle 50% ( $0.0148 \leq RMSE < 0.0253$ ) and the bottom row are from the worst 25% ( $RMSE \geq 0.0253$ ). The RMSE values for each plot are displayed in the top right corner.



**Fig. 6** The main figure presents the RMSE of all test plots to compare the performances of the basic model and full model (containing all features) for each of the model versions using: a single crown allometry (1), a generic allometry each for broadleaf and conifer (2) and the species-specific crown allometries (10). A RMSE value of 0 would correspond to a perfect fit. The inset figure shows the effect of incorporating the features into the basic model individually in each version of the model. The dotted horizontal lines denote the median RMSE for the single allometry basic model (upper) and the ten allometry full model (lower) for comparison; gap fraction, crown taper and crown permeability are all shown to correspond with a significant improvement in model performance, whilst crown overlap does not improve on the basic model when considered individually.



**Fig. 7** Gap fraction in a) is the proportion of first returns being recorded from the ground as a function of the total canopy area divided by the plot area. The points are the recorded ground returns linked with the total canopy area predicted from the allometries. The permeability factor is interpreted in b) where the solid circles outline the ECA looking from above in an example 1 m height interval with the grey shaded area denoting the area where first returns are recorded in that interval. The lighter shaded area gives the region where the first returns have permeated from the interval above (fitted proportion of 0.65) and the hashed area gives the region where the first returns permeate to the interval below. The crown overlap correction factor in c) is predicted from the cumulative canopy area as a proportion of the plot area.



**Fig. 8** Comparison of allometry-derived crown forms (dotted line) and fitted crown forms using the LiDAR data (solid line) for trees with a dbh of 30cm; the grey shaded area denotes the 95% confidence interval associated with the fitted crown shape parameter ( $\beta$ ).

800 **Supplementary material**

801 Forest details

802 The details of the plots are given in Table S.1. The plots were selected to represent the full

803 variation present in the forest.

**Table S.1** A summary of the forest plot properties where the plots are grouped by percentage ocular coverage. The stems are grouped into diameter classes with the stem diameters given in brackets: small poles ( $8 \leq x \leq 17$ ), large poles ( $17 < x \leq 25$ ), small logs ( $25 < x \leq 37$ ), medium logs ( $37 < x \leq 49$ ) and large logs ( $49 < x \leq 100$ ). The mean basal area (BA; m<sup>2</sup>/ha) and the corresponding standard deviation (s.d.) is given for each of these groups as well as the mean total BA and mean total stem density (no. stems/ha)

Ocular coverage (%)	No. plots	Mean small pole BA (s.d.)	Mean large pole BA (s.d.)	Mean small log BA (s.d.)	Mean medium log BA (s.d.)	Mean large log BA (s.d.)	Mean total BA (s.d.)	Mean total stem density (s.d.)
51-60	32	5.60 (2.82)	4.43 (1.83)	6.14 (2.17)	3.34 (2.16)	2.74 (5.84)	22.25 (6.61)	770 (290)
61-70	22	4.65 (2.27)	4.87 (2.76)	6.94 (2.35)	5.20 (2.80)	2.76 (2.65)	24.41 (5.78)	704 (236)
71-80	44	3.38 (2.28)	4.14 (1.88)	7.16 (2.15)	6.22 (3.39)	5.37 (4.22)	26.70 (6.72)	632 (250)
81-90	41	4.21 (2.53)	3.90 (1.67)	6.33 (2.94)	5.89 (3.25)	4.01 (4.22)	24.35 (6.90)	644 (240)
91-100	15	5.17 (3.29)	5.61 (2.31)	6.55 (2.78)	5.04 (3.93)	3.41 (2.98)	25.78 (3.47)	756 (322)

804

805 **LiDAR specifications**

806 The details of the LiDAR flight are provided in Table S.1.

**Table S.1** The LiDAR flight details for Haliburton Forest, Ontario collected in August 2009.

Settings	Definition	Flight plan specifications
No. lines flown	Number of passes that the aircraft flew	39 passes (3 control)
Altitude	Height at which the aircraft was flown	1500 m
Overlap	The degree to which the passes overlap	30% (15% on either side of the pass)
Speed	Speed at which the plane was flown	120 kts
System PRF	Pulse repetition frequency	70 kHz
Scan frequency	Number of pulses emitted per second	36 Hz
Scan half angle	Angle at which the beam axis was directed away from azimuth	16°
Cross track resolution	Space between pulses perpendicular to the line of flight	0.89 m
Down track resolution	Space between pulses along the line of flight	0.86 m
Point density	Number of first returns recorded per m <sup>2</sup>	2 points m <sup>-2</sup>
Footprint size	Area covered by a single beam on the ground	0.14 m <sup>2</sup>

807

808

809

810

811 **Published allometries**

812 The published allometries used to make predictions for the tree crowns, and thus the canopy  
 813 structure, are detailed in Table S.1.

**Table S.1** The allometric equations used to derive canopy area at a given height from the measurements of diameter at breast height (dbh) for an individual tree of species  $j$ . The equation and coefficients detailed in (Caspersen et al. 2011) were used in preference to (Purves et al. 2007) where available.

Scaling relationship	(a) Caspersen <i>et al.</i> (2011)	(b) Purves <i>et al.</i> (2007)
Height (H)	$H = 1.3 + (\eta_j - 1.3) \cdot (1 - e^{\left(\frac{-\phi_j \cdot dbh}{\eta_j^{-1.3}}\right)})$	$\log(H) = a_j + b_j \cdot \log(dbh)$
Crown depth (V)	$V = \omega_j \cdot H$	As left, where: $\omega_j = (1 - T_j) \cdot C_0(\omega_j) + T_j \cdot C_1(\omega_j)$
Maximum radius ( $R_{max}$ )	$R_{max} = r_{0,j} + \frac{dbh}{40} \cdot (r_{40,j} - r_{0,j})$	As left, where: $r_{0,j} = (1 - T_j) \cdot C_0(r_0) + T_j \cdot C_1(r_0)$ $r_{40,j} = (1 - T_j) \cdot C_0(r_{40}) + T_j \cdot C_1(r_{40})$
Radius at height h ( $R_h$ )	$R_h = R_{max} \cdot \left(\frac{H-h}{H}\right)^{\beta_j}$	As left, where: $\beta_j = (1 - T_j) \cdot C_0(\beta) + T_j \cdot C_1(\beta)$
Crown area at height h ( $CA_h$ )	$CA_h = \pi \cdot R_h^2$	As left.

814 The parameter values for the original allometries are given in Table S.2 for the species which  
815 were included in the analysis. The broadleaf and conifer parameters were derived by re-fitting  
816 the allometric relationships to simulated data which were weighted according to the relative  
817 abundance of all of the species represented by these two allometries. The simulated data were  
818 retrieved by using the height and maximum radius allometries from Table S.1 for each of the  
819 included species to predict data points for a range of dbh measures where the number for each  
820 species is proportional to the measured abundance. The allometries were then re-fit to the  
821 simulated data to retrieve the parameter values for the conifer and broadleaf allometries.

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837

**Table S.2** The species-specific parameter values for the allometric equations used to predict canopy structure from dbh measures with the lower then upper confidence intervals listed below in italics (where available). The \* entries denote that the parameter values have not been directly extracted from the paper, but are derived from re-fitting the allometric relationships to simulated data weighted according to abundance of all included species. The #  $\varpi_j$  values for balsam fir and white spruce were calculated from unpublished data.

Species (j)	Caspersen <i>et al.</i> (2011)					Purves <i>et al.</i> (2007)			
	$\eta_j$	$\phi_j$	$\varpi_j$	$r_{0,j}$	$r_{40,j}$	$\beta_j$	$a_j$	$b_j$	$T_j$
American beech	23.8095	1.29675	0.47493	2.41848	9.0831	1.04203			
	<i>23.6162</i>	<i>1.28351</i>	<i>0.47211</i>	<i>2.31153</i>	<i>8.87821</i>	<i>1.01849</i>	-	-	-
	<i>24.0130</i>	<i>1.31437</i>	<i>0.47726</i>	<i>2.54481</i>	<i>9.18623</i>	<i>1.06318</i>			
Balsam fir			0.4512 <sup>#</sup>				0.1203	0.783	0.278
	-	-	-	-	-	-	-	-	<i>0.276</i>
			-				-	-	<i>0.286</i>
Hemlock	26.9651	0.72939	0.52407	1.55475	8.24872	1.33787			
	<i>25.8911</i>	<i>0.71046</i>	<i>0.51861</i>	<i>1.32560</i>	<i>8.01792</i>	<i>1.28702</i>	-	-	-
	<i>27.8909</i>	<i>0.75505</i>	<i>0.53189</i>	<i>1.82560</i>	<i>8.49993</i>	<i>1.37687</i>			
Ironwood	19.0506	1.5723	0.43857	1.44703	5.48292	0.3837			
	<i>18.6261</i>	<i>1.53716</i>	<i>0.43351</i>	<i>1.36581</i>	<i>5.30517</i>	<i>0.36194</i>	-	-	-
	<i>19.4916</i>	<i>1.60207</i>	<i>0.44383</i>	<i>1.53057</i>	<i>5.67222</i>	<i>0.41123</i>			
Red maple	20.9017	1.65906	0.40956	1.04395	4.03818	0.0001			
	<i>20.2315</i>	<i>1.56883</i>	<i>0.40339</i>	<i>0.88105</i>	<i>3.92248</i>	<i>0.0001</i>	-	-	-
	<i>21.4944</i>	<i>1.77272</i>	<i>0.41678</i>	<i>1.15198</i>	<i>4.15271</i>	<i>0.0001</i>			
Sugar maple	25.3588	1.35017	0.46099	1.52529	4.94851	0.22977			
	<i>25.2651</i>	<i>1.34467</i>	<i>0.45969</i>	<i>1.50483</i>	<i>4.91789</i>	<i>0.22255</i>	-	-	-
	<i>25.4421</i>	<i>1.35479</i>	<i>0.4619</i>	<i>1.55158</i>	<i>4.97733</i>	<i>0.23762</i>			
White spruce			0.5889 <sup>#</sup>				0.1404	0.7354	0.278
	-	-	-	-	-	-	-	-	<i>0.272</i>
			-				-	-	<i>0.296</i>
Yellow birch	24.7702	1.19439	0.45541	4.1811	8.09932	1.03778			
	<i>24.3362</i>	<i>1.14311</i>	<i>0.45016</i>	<i>3.92218</i>	<i>7.79143</i>	<i>0.98285</i>	-	-	-

25.3302 1.23393 0.46208 4.41847 8.33067 1.07811

			0.4479				0.4383	0.5564	0.5585
Broadleaf*	-	-	-	-	-	-	-	-	-
			-				-	-	-
			0.5213				0.1939	0.7005	0.2618
Conifer*	-	-	-	-	-	-	-	-	-
			-				-	-	-

838 The extra parameters required for the Purves *et al.* (2007) allometries, which are not species  
839 specific, are given in Table S.3 and were not altered. These parameters are used to transform the  
840 Purves *et al.* (2007) species-specific parameters into the same form as those used directly in the  
841 Caspersen *et al.* (2011) allometries.

**Table S.3** The parameter values used to convert the species-specific trait score  $T_j$  value into parameters describing the properties of the crown.

Definition	Parameter (P)	$C_0(P)$	$C_1(P)$
Crown depth	$\varpi$	0.95	0.95
Crown radius at height 0 m	$r_0$	0.503	3.126
Crown radius at height 40 m	$r_{40}$	0.5	10.0
Crown shape	$\beta$	0.196	0.511

842 Definitions for all of the terms used in the model equations (eqs. 1-6) are detailed in Table S.4.

843

844

845

846

847

848

849

850

851

852

853

**Table S.4** Definitions of all symbols and terms used throughout the manuscript.

<b>Term</b>	<b>Interpretation</b>
SDD	Stem diameter distribution
RMSE	Root mean square error
DIC	Deviance information criterion
ABA	Area-based approach
ITD	Individual tree detection
Dbh	Diameter at breast height

<b>Symbol</b>	<b>Interpretation</b>
$HD_{LIDAR}/ HD_{PREDICT}$	Height distribution of LiDAR first returns in 1 m intervals (recorded/predicted).
$HD_{PREDICT}^{BASIC}$	Basic model.
$HD_{PREDICT}^{PERM}$	Basic model incorporating crown permeability.
$HD_{PREDICT}^{PERM+OVERLAP}$	Basic model incorporating crown permeability and overlap.
$HD_{PREDICT}^{FULL}$	Basic model incorporating all of the features – full model.
$h/i$	Vertical distance in metres above the ground.
$h_{max}$	Maximum predicted height of the canopy rounded up to the nearest meter.
$ECA_{[h,h+1)}$	Exposed canopy area in the height interval from, and including, height $h$ and up to, but not including, height $h + 1$ .

$\frac{\sum_{i=0}^{h_{max}} ECA_i}{\sum_{i=h+1}^{h_{max}} ECA_i}$	Total exposed canopy area/cumulated exposed canopy area above given interval measured from the upper bound of the interval to the top of the canopy.
$p_0$	Gap fraction
$\alpha$	Parameter in the gap fraction exponential function
$\varphi$	Crown permeability parameter
$\beta$	Parameter controlling crown taper
$\theta_{[h,h+1]}$	Crown overlap correction factor in a given height interval
$\gamma$	Parameter in the crown overlap correction factor exponential function
	Normalisation constant calculated such that
$N$	$\sum_{h=0}^{h_{max}} HD_{PREDICT[h, h+1]}^{FULL} = 1$
PA	Plot area (2500 m <sup>2</sup> )

854 *Model fitting results*

855 The full model includes all features and is represented by eq. 6, but all of the features were tested  
856 independently as well. The resultant fitted parameters used in the different model versions are  
857 presented in Table S.5 along with the 95% confidence intervals for each value.

858 To fit the new crown shape parameters ( $\beta_{j,new}$ ), we recalculated the maximum crown  
859 radius ( $R_{max,new}$ ) as follows:

860 (S. 1) 
$$R_{max,new} = R_{max} \left( \frac{H-V}{H} \right)^{\beta_j - \beta_{j,new}}$$

861 The recalculated maximum crown radius and the new crown shape parameter can be used  
862 in place of the original maximum crown radius and crown shape parameter.

863 The root mean square error (RMSE) was calculated as follows:

864 (S. 2) 
$$RMSE = \sqrt{\frac{\sum_{h=0}^{h_{max}} (HD_{PREDICT}[h, h+1] - HD_{LIDAR}[h, h+1])^2}{h_{max}}}$$

865 Where  $h_{max}$  is the maximum predicted height of the canopy rounded up to the nearest  
 866 meter. The relative RMSE is calculated as:

867 (S. 3) 
$$\%RMSE = \frac{RMSE}{h_{max}}$$

**Table S.5** Parameter values fitted using the LiDAR data from the model-fitting subset of plots.

The values in normal text are from the full model whilst those in italics are from the model independently representing that specific feature. The fitted values are presented with the associated 95% confidence intervals.

Model version	Species (j)	No. crown allometries	Fitted value	Lower 95% CI	Upper 95% CI
Gap fraction ( $\alpha$ )	-	1	2.3056 ( <i>1.8560</i> )	2.2250 ( <i>1.7935</i> )	2.3855 ( <i>1.9183</i> )
		2	2.7709 ( <i>2.3458</i> )	2.6938 ( <i>2.2787</i> )	2.8538 ( <i>2.4109</i> )
		10	2.5618 ( <i>2.2073</i> )	2.4917 ( <i>2.1435</i> )	2.6362 ( <i>2.2828</i> )
	Sugar maple	1	0.4385 ( <i>0.3650</i> )	0.4187 ( <i>0.3541</i> )	0.4600 ( <i>0.3783</i> )
		2	0.4320 ( <i>0.3008</i> )	0.4126 ( <i>0.2842</i> )	0.4542 ( <i>0.3154</i> )
		10	0.3312 ( <i>0.2323</i> )	0.3025 ( <i>0.2166</i> )	0.3521 ( <i>0.2500</i> )
	Conifer	2	0.4910 ( <i>0.3588</i> )	0.4145 ( <i>0.2952</i> )	0.5798 ( <i>0.4046</i> )
		10	0.3850 ( <i>0.1943</i> )	0.1919 ( <i>0.0373</i> )	0.7585 ( <i>0.3487</i> )
	American beech	10	0.2565 ( <i>0.1513</i> )	0.2116 ( <i>0.1245</i> )	0.3055 ( <i>0.1817</i> )
Crown taper ( $\beta_j$ )	Balsam fir	10	0.9390 ( <i>0.5083</i> )	0.6147 ( <i>0.3700</i> )	0.3055 ( <i>0.6444</i> )
	Hemlock	10	0.3613 ( <i>0.2518</i> )	0.3155 ( <i>0.2229</i> )	0.4017 ( <i>0.2751</i> )
	Ironwood	10	1.0711 ( <i>0.0087</i> )	0.0045 ( <i>0.0012</i> )	1.4855 ( <i>0.0324</i> )
	Red maple	10	0.4823 ( <i>0.3257</i> )	0.4233 ( <i>0.2962</i> )	0.5429 ( <i>0.3621</i> )
	White spruce	10	0.2045 ( <i>0.1064</i> )	0.0900 ( <i>0.0429</i> )	0.3366 ( <i>0.1813</i> )
	Yellow birch	10	0.2544 ( <i>0.1638</i> )	0.2007 ( <i>0.1234</i> )	0.3394 ( <i>0.2207</i> )
	Broadleaf	10	0.4213 ( <i>0.3354</i> )	0.3656 ( <i>0.2754</i> )	0.4753 ( <i>0.3687</i> )

Crown permeability ( $\varphi$ )	-	1	0.9334 (0.6307)	0.8548 (0.6010)	0.9924 (0.6704)
		2	0.8097 (0.6557)	0.6835 (0.6049)	0.9055 (0.6930)
		10	0.6507 (0.6635)	0.5575 (0.6313)	0.7351 (0.6886)
Crown overlap ( $\gamma$ )	-	1	0.3349 (0.0004)	0.3013 (0.0001)	0.3707 (0.0012)
		2	0.6092 (0.0006)	0.5606 (0.0001)	0.6693 (0.0022)
		10	0.5854 (0.0004)	0.5168 (0.0001)	0.6556 (0.0012)

868 **What would the accuracy of the approach be in matching LiDAR distributions?**

869 Our planned approach to mapping SDDs is to use the model to generate an  $HD_{\text{PREDICT}}$  for a huge  
870 number of theoretical plots so that for each  $HD_{\text{LIDAR}}$  we can identify a set of SDDs which have  
871 similar  $HD_{\text{PREDICT}}$  values to the observed  $HD_{\text{LIDAR}}$ . If this modelling approach narrows down on  
872 just a few similar-looking SDDs then it has worked well, whereas if a large number of dissimilar  
873 SDDs are identified then the procedure does not have the ability to discriminate alternative size  
874 structures with confidence without the use of additional information. This issue will be  
875 thoroughly analysed in another paper, but as a preliminary test we matched the  $HD_{\text{LIDAR}}$  of each  
876 model plot with that of all the other model plots and measured the RMSE of each pairing (Fig.  
877 S.1); this quantifies how many plots might be matched to a given  $HD_{\text{LIDAR}}$ .

878

879

880

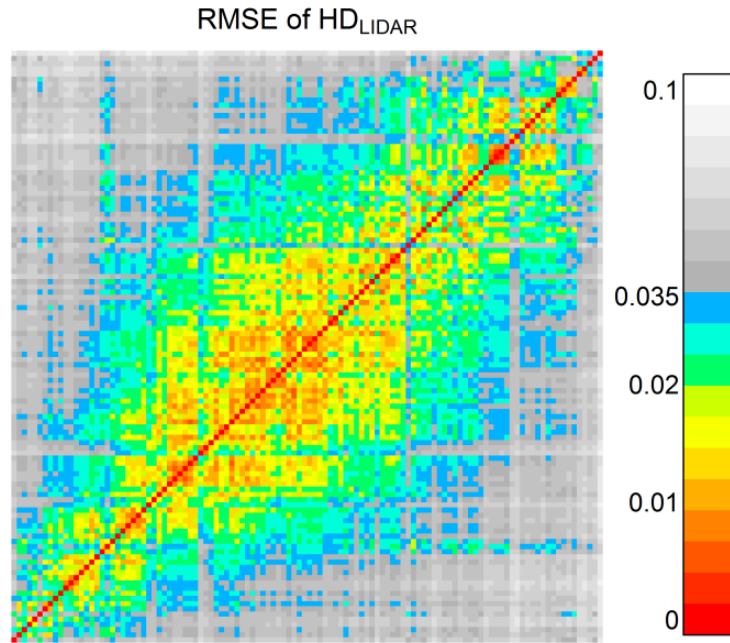
881

882

883

884

885



**Fig. S.1** The matrix give the RMSE of each model plot  $HD_{LIDAR}$  matched with that of each of the other model plots where the diagonal line is where the plot has been matched with itself. The plots have been ordered so that those with similar  $HD_{LIDAR}$  distributions have been grouped together. The grey region are the matched plots with a RMSE less than the lowest RMSE achieved by the ten allometry full model, the blue-green region is the lowest RMSE to mean RMSE and the yellow-red region denotes all of the plots that match with a RMSE equal to or greater than the mean RMSE.

886 Given the accuracy of the ten allometry full model, 20% of the plots, on average, could be  
 887 matched as having the same  $HD_{LIDAR}$  (Fig. S.1); these are denoted by the yellow-red region of  
 888 the matrix.