



A new model for predicting the total tree height for stems cut-to-length by harvesters in *Pinus radiata* plantations

Chenxi Shan¹ · Huiquan Bi^{2,3} · Duncan Watt⁴ · Yun Li¹ · Martin Strandgard⁵ · Mohammad Reza Ghaffariyan⁵

Received: 6 May 2019 / Accepted: 27 August 2019 / Published online: 16 December 2019
© The Author(s) 2019

Abstract A new model for predicting the total tree height for harvested stems from cut-to-length (CTL) harvester data was constructed for *Pinus radiata* (D.Don) following a conceptual analysis of relative stem profiles, comparisons of candidate models forms and extensive selections of predictor variables. Stem profiles of more than 3000 trees in a taper data set were each processed 6 times through simulated log cutting to generate the data required for this purpose. The CTL simulations not only mimicked but also covered the full range of cutting patterns of nearly 0.45×10^6 stems harvested during both thinning and harvesting operations. The single-equation model was estimated through the multiple-equation generalized method of moments estimator to obtain efficient and consistent parameter estimates in the presence of error correlation and heteroscedasticity that were inherent to the systematic structure of the data. The predictive

performances of our new model in its linear and nonlinear form were evaluated through a leave-one-tree-out cross validation process and compared against that of the only such existing model. The evaluations and comparisons were made through benchmarking statistics both globally over the entire data space and locally within specific subdivisions of the data space. These statistics indicated that the nonlinear form of our model was the best and its linear form ranked second. The prediction accuracy of our nonlinear model improved when the total log length represented more than 20% of the total tree height. The poorer performance of the existing model was partly attributed to the high degree of multicollinearity among its predictor variables, which led to highly variable and unstable parameter estimates. Our new model will facilitate and widen the utilization of harvester data far beyond the current limited use for monitoring and reporting log productions in *P. radiata* plantations. It will also facilitate the estimation of bark thickness and help make harvester data a potential source of taper data to reduce the intensity and cost of the conventional destructive taper sampling in the field. Although developed for *P. radiata*, the mathematical form of our new model will be applicable to other tree species for which CTL harvester data are routinely captured during thinning and harvesting operations.

Project funding: This study was partly funded by Forest and Wood Products Australia Limited (FWPA) through project PNC465-1718: Advanced real-time measurements at harvest to increase value recovery and also supported by Beijing Forestry University through the special fund for characteristic development under the program of Building World-class University and Disciplines.

The online version is available at <http://www.springerlink.com>

Corresponding editor: Yu Lei.

✉ Yun Li
liyun@bjfu.edu.cn

¹ School of Information Science and Technology, Beijing Forestry University, Beijing 100083, People's Republic of China

² Forest Science, NSW Department of Primary Industries – Forestry, Level 12, 10 Valentine Ave, Parramatta, NSW 2150, Australia

³ School of Ecosystem and Forest Sciences, The University of Melbourne, Parkville, VIC 3010, Australia

⁴ Snowy Region, Forestry Corporation of NSW, 76 Capper Street, Tumut, NSW 2720, Australia

⁵ Australian Forest Operations Research Alliance (AFORA), University of the Sunshine Coast, Sippy Downs, QLD 4556, Australia

Keywords Stem profiles · Cut-to-length simulations · Harvester data · Model construction · Nonlinear multiple-equation GMM estimation · Benchmarking prediction accuracy

Introduction

Over the last 40 years, cut-to-length (CTL) harvesters have been increasingly adopted and widely utilised to improve log-harvesting productivity in natural and plantation forests worldwide (e.g., Huyler and LeDoux 1999; Murphy 2003; Gerasimov et al. 2012, 2013; Strandgard et al. 2013; Olivera et al. 2016; Williams and Ackerman 2016; Lu et al. 2018). The widespread utilization has been largely driven by great technological advances in the mechanical design of harvesters and in the harvester head measurement and optimization systems over the same period (Heinimann 2007; Nordfjell et al. 2010; Uusitalo 2010; Malinen et al. 2016). Now modern harvesters equipped with a GPS receiver and a computerized harvester head have become a major source of “big data” for forest management as they constantly capture, accrue and provide a daily flow of spatially explicit and time-stamped data on log production and assortment as well as detailed diameter measurements of harvested stems of individual trees over large operational areas (Sellén 2016; Uusitalo 2017; Lu et al. 2018; Müller et al. 2019; Rossit et al. 2019). Harvester data can provide the total log length but not the total height of individual trees because the top crown section beyond the last cut does not pass through the harvester head and therefore its length cannot be measured. The lack of tree height data represents a stumbling block in the full integration of spatially explicit harvester data with conventional inventory data, remote sensing imagery and LiDAR data for the development of harvester-based inventory systems (e.g., Stendahl and Dahlin 2002; Murphy et al. 2006; Holopainen et al. 2010), for predicting attributes of individual trees, stands and forests (e.g., Rasinmäki and Melkas 2005; Holmgren et al. 2012; Söderberg 2015; Saukkola et al. 2019) and for estimating product recovery (e.g., Peuhkurinen et al. 2008; Barth and Holmgren 2013; Barth et al. 2015; Caccamo et al. 2018; Hauglin et al. 2018). Full integration will provide forest managers with a more detailed view of standing trees over an entire forest area and also allow them to predict spatially the volume and value of a certain product or a product mix that the forest can yield for optimising value recovery in management planning (Hauglin et al. 2018; Lu et al. 2018). Without full data integration, maximum value extraction from harvester data cannot be attained, preventing the

transformation of big data into valuable data for forest management (see Müller et al. 2019).

For the most effective use of harvester data, a necessary first step is to estimate the total height of each harvested tree that was bucked, measured and recorded by the harvester head. This necessity has been well recognised, and several attempts to do so have been made in the estimation of logging residue biomass and in the integration of harvester data with remote sensing data in forest inventory for predicting tree and stand attributes and estimating product recoveries (Varjo 1995; Kiljunen 2002; Maltamo et al. 2010; Möller et al. 2011; Söderberg 2015; Siipilehto et al. 2016; Hauglin et al. 2018; Lu et al. 2018). So far, total tree height has been calculated by estimating the length of the unprocessed top section of individual trees using the only such existing model, that of Varjo (1995), or through an iterative search algorithm using a taper equation as demonstrated well by Lu et al. (2018) and also briefly alluded to by Hauglin et al. (2018). Another ad hoc method was described by Kiljunen (2002), but it cannot be readily applied because of its requirement for diameter measurements at specific heights that are not routinely captured and stored in a harvester database.

Although proved useful in reconstructing individual trees from harvester data (e.g., Palander et al. 2009; Vesa and Palander 2010; Siipilehto et al. 2016), the only such existing model has a rather restricted range of application, only to stems with the small end diameter underbark of the top log between 5 and 10 cm and over a minimum log length of 1.5 m as outlined by Varjo (1995) for the 3 coniferous and broadleaf species in his study. The iterative search algorithm demonstrated by Lu et al. (2018) requires a taper equation and 2 extreme quantile curves that define the search range of total tree height at any given diameter at breast height overbark (DBH) for the species at interest. In comparison to a single predictive equation, it requires a large amount of auxiliary height and diameter data, is somewhat cumbersome to implement in a computer program and also takes longer to reach the optimal estimate. Even after the optimum is reached, it may still be less accurate than the prediction from a single equation as shown by Lu et al. (2018). This approach also relies on extensive bucking simulation through a large taper data set for accuracy testing. Without doing so, the accuracy of tree height estimation cannot be ascertained as in the case of Hauglin et al. (2018).

The weaknesses of these 2 methods highlight the need for an improved and more efficient approach to predict total tree height of individual stems cut-to-length by harvesters. The recent work of Lu et al. (2018) with *Pinus radiata* represented an essential first step towards satisfying this need.

Building upon this work, this paper presents a new model that overcomes the weaknesses of the 2 methods that are currently in use, also using plantation trees of *P. radiata* as an example.

Materials and methods

Notation

DBH	Diameter at breast height overbark (cm)
DOB	Diameter overbark (cm)
H	Total tree height from ground level to tip of the tree (m)
H_s	Average stump height of 0.15 m
H_b	Defined breast height of 1.3 m above ground level
L	Total log length, i.e., sum of lengths of logs and waste sections of a stem (m)
L_{top}	Length of the unprocessed tree top section from the last cutting point to the tip (m)
LED	Large end diameter overbark of a log (cm)
SED	Small end diameter overbark of a log (cm)
SEDTL	SED of the top log (cm), the smallest SED of a cut stem
$d = (SEDTL/DBH)$	Relative diameter that takes any value between 0 and 1
$h = (L + H_s - H_b)/(H - H_b)$	Relative height above breast height
$B = (1 - d)$	The base function
K	Variable exponent that is a function of L , DBH and SEDTL
$R_1 = (L/H)$	Total log length ratio
$T = (DBH - SEDTL)/L$	Average taper over total log length

A schematic diagram of the notation was drawn to aid understanding by readers (Fig. 1).

Data

For the purpose of this study, it would be ideal if the total tree height and stump height of a large enough sample of

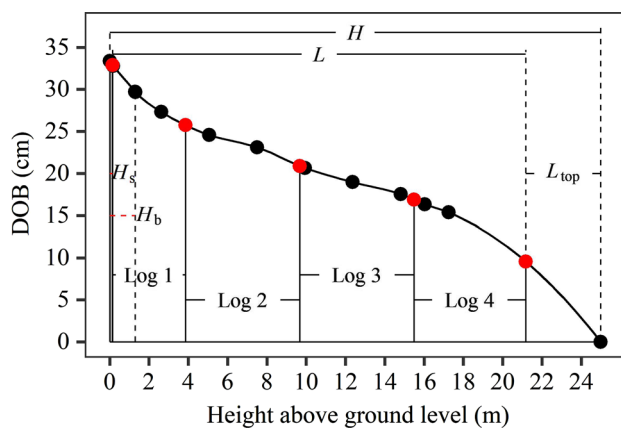


Fig. 1 Diagram illustrating a stem profile cut into 4 logs. Log end diameters (red dots) were derived from the profile that was interpolated from overbark taper measurements (black dots) using piecewise cubic Hermite interpolating polynomials (PCHIP). See the “Notation” section for symbols in the diagram

harvested trees were measured and recorded in addition to log lengths and diameters in the field during harvesting as Lang et al. (2010) did for a small number of sample trees. However, measuring a large number of sample trees across different stand types, site, and age classes is costly and time-consuming, particularly when working with harvesters in field operations amid safety concerns. To overcome the lack of such ideal data, this study followed the intuitive approach of Lu et al. (2018) in combining taper and harvester data through simulated log-cutting to generate a data set that not only included stump height and total tree height, but also mimicked the cutting patterns of harvesters in operational thinning and harvesting.

Taper data

The taper data set contained 3251 trees sampled from *P. radiata* stands across the State of New South Wales (NSW) in Australia over 30 years. These taper trees were sampled from both thinned and unthinned stands over wide ranges of age, site quality, and stand conditions. The lowest measurements of both overbark and underbark diameter were taken at between 0.1 and 0.3 m above ground, then another measurement was usually taken at 0.7 m before reaching the defined breast height of 1.3 m. Measurement intervals above breast height varied between 1.5 and 3 m depending on the height of the sample trees. This data set was previously used by Bi and Long (2001), Bi et al. (2012) and Zhang et al. (2015) for developing taper equations, height–diameter functions and conversion factors for DBH measured at different breast height. It was also used by Lu et al. (2018) in combination with harvester data in log-cutting simulations. These publications provided a

detailed textual and graphical description of this data set, so it was not repeated here. For the log-cutting simulation in this study, 152 trees were excluded, including 66 small trees with DBH less than 10 cm that fell below the minimum size for log-cutting and 86 trees that had underbark but not overbark taper measurements. The remaining 3099 trees in the taper data set had a range of 10–79.1 cm for DBH and 6.4–44.5 m for total tree height. For each tree, a complete stem profile from ground to the tip was constructed numerically by interpolating through its observed values of DOB using piecewise cubic Hermite interpolating polynomials (PCHIP) with an even interval of 0.1 m.

Harvester data

The harvester data came from the Tumut Management Area of the Snowy Region, Forestry Corporation of NSW (FCNSW). Covering the foothills of the Snowy Mountains, this management area has about 90,000 ha of *P. radiata* plantations, representing approximately 46% of FCNSW's total *P. radiata* estate, and produces over 1×10^6 cubic meters of sawlogs and more than 0.6×10^6 cubic meters of pulp wood annually. These plantations were established with an initial stocking of 1000–1100 trees ha^{-1} across all site classes under the stand density management regime adopted for *P. radiata* in NSW since the 1980s (Horne and Robinson 1988). At the present regime, either one or two thinnings are prescribed for more productive sites. The first thinning aims to bring the stocking down to 450–550 trees ha^{-1} at around 14 years of age, and the second thinning further reduces the stocking down to 200–300 trees ha^{-1} at around 23 years of age. For poorer sites, no thinning is carried out before the final harvest generally at the age of 30–35 years.

The harvester data contained a total of 0.502×10^6 stems/trees that were harvested and processed by 20 CTL harvesters in routine thinning and clear-felling operations at 54 sites over a period of 3 years from 2012 to 2014. Each stem was cut into one or more logs that were numbered sequentially from the butt log to the top log, making up a total of 1.808×10^6 logs including approximately 0.240×10^6 waste sections in the data set. For each stem, stump height and stump diameter, i.e. LED of the butt log, were not available. For each log, the length, overbark volume and SED were recorded together with a product description that contained customer identification and product type. For logs without a specific customer and product type, production description was recorded as “random”. The cutting patterns were driven by the sawlog and pulpwood specifications of the local softwood industry consisting of one large 0.7 million tonne pulp and paper mill producing high quality kraft paper for both domestic and international markets and six large and small timber and wood products companies near the town of Tumut in southern NSW.

Data screening and interactive exploratory data analysis

With the aid of the statistical software R 3.5.1 (R Foundation for Statistical Computing, Vienna, Austria) and GGobi 2.1.10, an open source visualization program for exploring high-dimensional data (see Swayne et al. 2006; Cook et al. 2007), multiple rounds of data screening and interactive exploratory data analysis were carried out through a systematic set of procedures to detect inconsistencies in data recording, erroneous or illogical records in the harvester data. Some obvious errors were corrected wherever possible, others errors or anomalies that could not be corrected were removed from the data set. To facilitate this process, DBH was calculated for each stem from the length and SED of its butt log using a system of conversion equations together with linear interpolation as in Lu et al. (2018). In addition, the merchantable height up to 5 cm DOB of trees in the taper data set was analysed in relation to their DBH to obtain an extreme nonlinear conditional quantile curve in the form of the 3-parameter Chapman–Richards function, one of the most commonly used functions for tree height and diameter models (see Huang et al. 1992; Huang 1999; Bi et al. 2000, 2012):

$$\text{MH}_{0.995} = 1.3 + 44.35(1 - e^{-0.0432\text{DBH}})^{1.1395}, \quad (1)$$

where $\text{MH}_{0.995}$ represents the 0.995th nonlinear conditional quantile of merchantable height in meters for a given DBH. The 3 parameters in Eq. 1 were estimated using the R package for quantile regression, quantreg (Koenker 2017, 2018). This extreme quantile curve depicted the relationship between the maximum attainable merchantable height and DBH for *P. radiata* in NSW (Fig. 2). Through this equation, a maximum attainable total log length (L_{\max}) for any given DBH was defined as $\text{MH}_{q=0.995}$ minus 0.15 m, the average stump height. This relationship between L_{\max} and DBH was then used as an upper frontier in the detection of anomalous stems.

After the completion of data screening and filtering, there were 0.448×10^6 stems and 1.581×10^6 logs remaining in the final data set. The DBH of these stems ranged from 10 to 82 cm with an average of 33 cm and their total log length L varied within the range of 2.2–40.8 m with an average of 16.6 m (Fig. 2). The minimum L of 2.2 m represented cases where only one short log was cut from a stem. The SED of the 1.581×10^6 logs varied from 3 to 80 cm, where values larger than 60 cm were found only in 0.3% of the logs. The individual log length ranged from 2.2 to 26.4 m, but lengths longer than 6.4 and 12 m represented less than 0.50% and less than 0.01% of the logs, respectively (Fig. 3). Because the logs were numbered sequentially from the butt log up, the number of logs decreased as log number (i.e., the relative

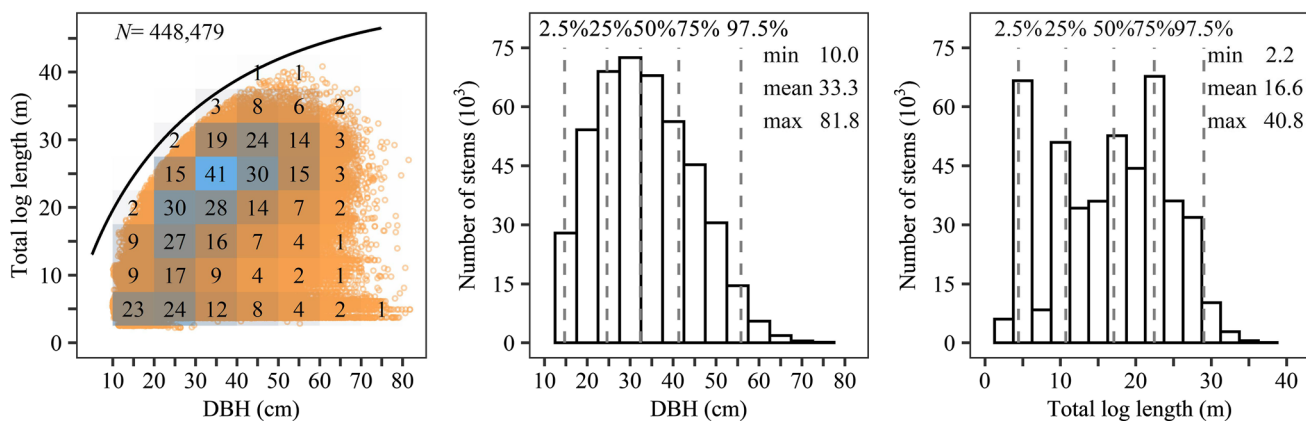


Fig. 2 Total log length (L) in relation to DBH under the curve of the maximum attainable merchantable height defined by Eq. 1 for the 0.448×10^6 stems drawn on the left as clustered heatmaps using the R package Pheatmap (pretty heatmaps). The numbers in the grid cells indicate the number of stems in thousands. The corresponding frequency distributions of DBH (center) and L (right) are shown with percentiles and descriptive statistics

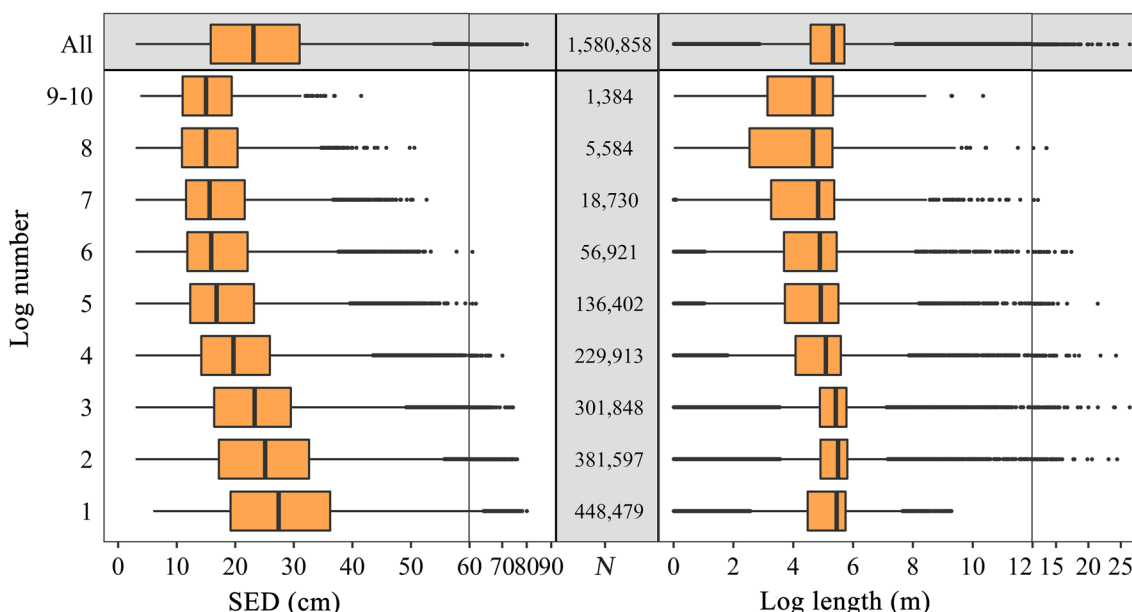


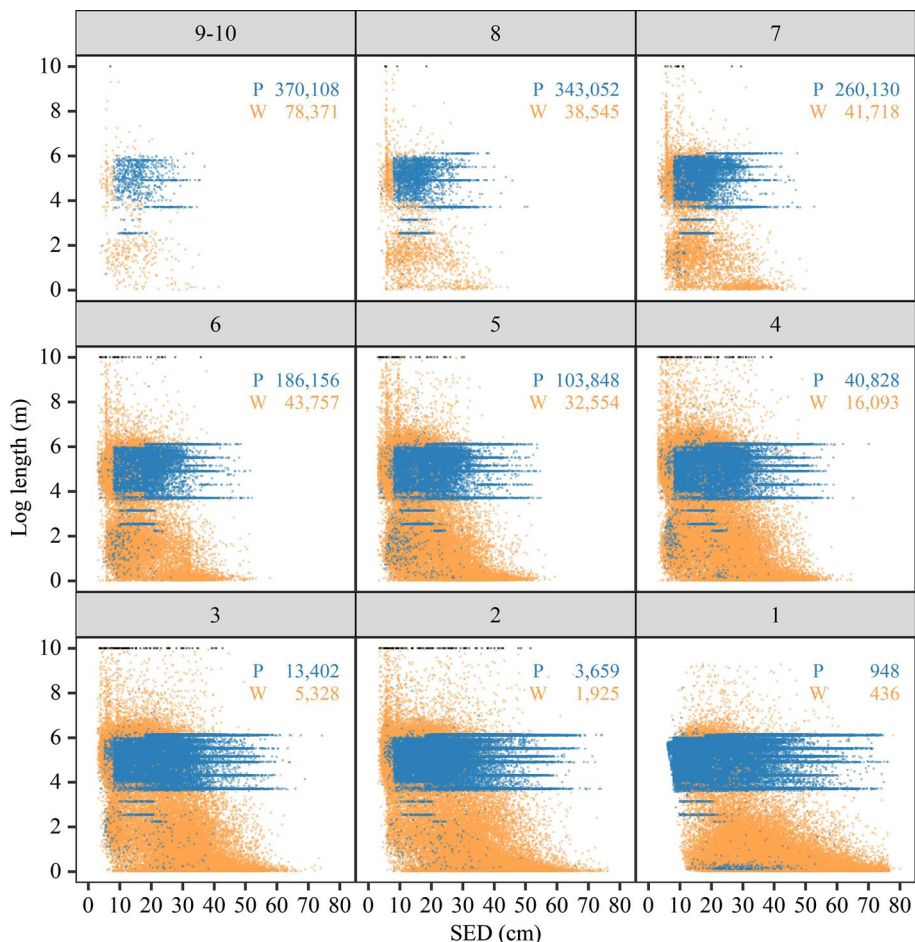
Fig. 3 Boxplots of SED (left) and log length (right) across the sequential log numbers for the 1.58×10^6 logs (including waste sections) cut-to-length from the 0.448×10^6 stems contained in the screened harvester data set. The numbers in the middle vertical stripe indicate the number of logs across the sequence. The boxplots in the top horizontal stripe are for all the logs combined

position of logs) increased. Concomitantly, the SED and length of logs varied both within and across log numbers, with the distribution of log length peaking around a number of specified log lengths (Fig. 4). This final data set was used for deriving cutting patterns for the simulated bucking of the stem profiles of taper trees.

Cut-to-length (CTL) simulations and data generation

Because of its larger volume, the harvester data showed an extremely wide spread in the value of SEDTL at any given DBH, which meant that there could be multiple alternative cutting patterns for a single stem, resulting in many SEDTL values and corresponding total log lengths. To have a more comprehensive representation of such variation in the cutting simulations, SEDTL was plotted against DBH for the 0.448×10^6 stems, and the data cloud was visually inspected

Fig. 4 Scatter plots of log length and SED as an extension of Fig. 3. The sequential log number is shown in the top stripe of each panel. Blue points represent logs tagged with a product code; brown points are either waste sections or random logs that were not tagged with any product type. The number of logs in each of the 2 log categories is indicated in each panel by the string colored in the same scheme starting with P for product or W for waste



through the interactive and dynamic graphical display of GGobi. Then 6 nonparametric conditional quantile curves at $\tau=0.01, 0.25, 0.50, 0.75, 0.90$ and 0.99 were drawn through the data cloud using the QUANTREG procedure in SAS 9.4 (SAS Institute, Cary, NC, USA) to help discern the patterns of conditional variation in SEDTL as DBH increased. Based on the patterns of the 6 nonparametric quantile curves, a parametric model was specified to derive the corresponding parametric conditional quantile curves:

$$SEDTL_{\tau} = a_{\tau} + b_{\tau}DBH^{c_{\tau}}, \tag{2}$$

where $SEDTL_{\tau}$ is the τ th conditional quantile for a given DBH, a_{τ}, b_{τ} and c_{τ} are the corresponding quantile-dependent parameters. Necessary constraints were placed upon parameter a_{τ} during quantile regression to prevent the fitted quantile curves from crossing each other in the close neighbourhood of the minimum DBH of 10 cm. In addition, parameter c_{τ} was set to 1 for the 90th and 99th quantiles as the relationship between $SEDTL_{\tau}$ and DBH became linear nearing the boundary of the data cloud (Fig. 5). For a taper tree with a given DBH in the log cutting simulations, these 6 parametric quantile curves provided 6 consecutive and increasingly

large initial SEDTL values that not only well covered the range of SEDTL variations contained in the harvester data, but also led to 6 systematic and interrelated cutting patterns with 6 different total log lengths ranging from the longest to the shortest that the tree could possibly yield.

The log cutting simulation for each of the 3099 taper trees looped through 6 values of $SEDTL_{\tau}$ starting from $\tau = 0.01$ and ending at $\tau = 0.99$. Within each loop, the simulation went through the following steps: (1) obtaining a value of $SEDTL_{\tau}$ from its DBH and the τ th conditional quantile curve as shown in Fig. 5; (2) deriving the corresponding total log length L by searching through its stem profile from an average stump height of 0.15 m upwards to the height where the DOB was equal to $SEDTL_{\tau}$; (3) from among the 0.448×10^6 stems contained in the harvester data set selecting the stem that was most similar to the taper tree by using a similarity index SI as described below; (4) using the cutting pattern of the most similar stem to locate the sequential cross-cut points on the taper tree’s stem profile with some adjustment if necessary; (5) taking the DOB and height at each cutting point on the numerically interpolated stem profile; (6) exporting the values of stump height, log number,

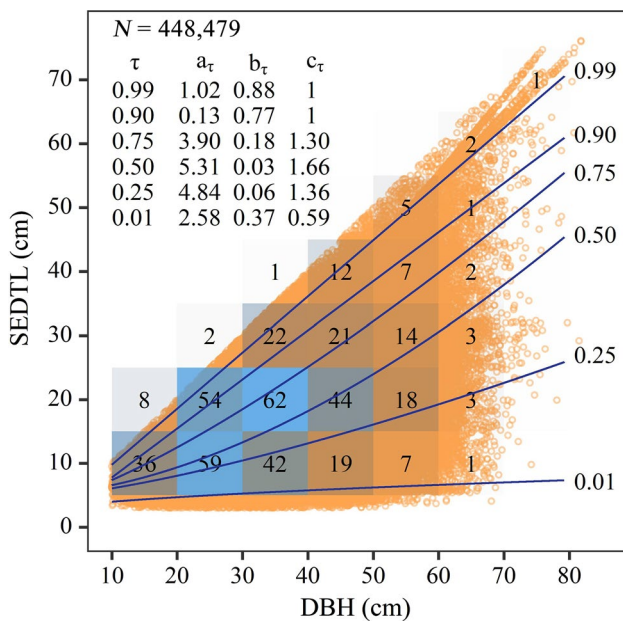


Fig. 5 Small end diameter of the top log (SEDTL) plotted against DBH for the 0.448×10^6 stems in clustered heatmaps using the R package *Pheatmap*. The numbers in the grid cells indicate the number of stems in thousands. Overlaid on top of the heatmaps are 6 conditional quantile curves, each with a number on its right end to indicate the value of the τ th quantile. The curves are defined by Eq. 2 and by the parameters for the 6 quantiles shown in the top left corner of the graph

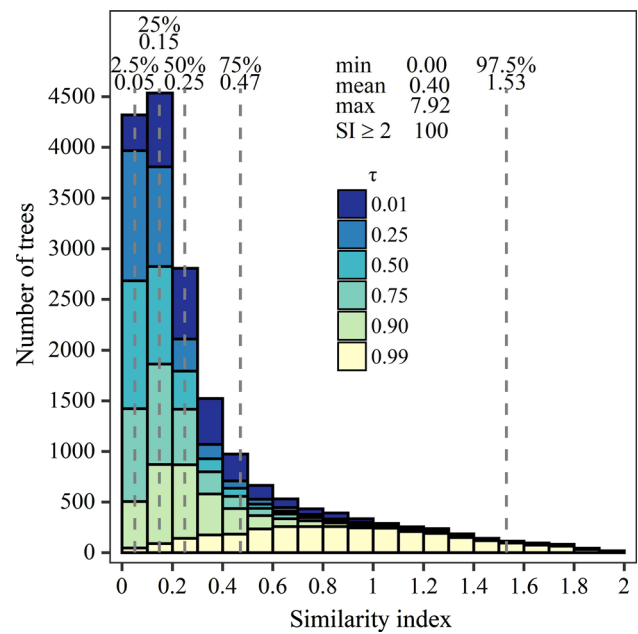


Fig. 6 Frequency distribution of similarity index (SI) for the most similar trees selected from among the 0.448×10^6 trees for the 3099×6 CTL simulations. The colored bar segments for the τ th conditional quantiles represented the cutting patterns derived from the 6 corresponding curves in Fig. 5. Characteristic percentiles are shown together with summary statistics of the distribution where the number of trees with SI greater or equal to 2 was 100 only

log length, LED and SED values, and L_{top} , the length of the tree top section from the last cutting point to the tip to a storage data set. The computer program previously written by Lu et al. (2018) in C# was modified to carry out all steps of the CTL simulations. The final data set contained 6 different but correlated values of SEDTL, L and L_{top} for each taper tree with the same DBH, total tree height H and stump height H_s .

The similarity index (SI) in step 3 was calculated as the distance between the candidate stem and the subject taper tree in a 3-dimensional space as follows:

$$SI = \sqrt{(DBH_H - DBH_T)^2 + (L_H - L_T)^2 + (SEDTL_H - SEDTL_T)^2}, \quad (3)$$

where variables with the subscript H and T denote the attributes of candidate stem from the harvester data and that of the subject taper tree, respectively. The value of SI is zero when the candidate stem and the subject taper tree are identical in the 3 dimensions. The SI values for the 3099×6 most similar stems selected in the log cutting simulations had an L-shaped distribution over a range from 0 to 7.9, with a median of 0.25, a mean of 0.40 and an upper quartile of 0.47. About 91% of the SI values were between 0 and 1 (Fig. 6).

Model derivation

In deriving a model specification for predicting total tree height H , L was considered as a fraction of H as shown in the “Notation” section. In doing so, any given height along a tree stem above breast height was converted to relative height h and thus h varied from 0 at 1.3 m to 1 at the tip of the tree. At the same time, the DOB at a relative height h was converted to relative diameter d , which varied from 1 at 1.3 m to 0 at the tip of the tree. Considering the relative stem profile of the part of the stem above breast height and taking reference of the idea and approach of Bi (2000) in the construction of the trigonometric variable-form taper model, h was expressed as a power function of d and other variables:

$$h = B^K, \quad (4)$$

where the base B is a monotonic function of relative diameter d , varying from 0 to 1 as d changes from 1 to 0, and K is a variable exponent. Nine candidate base functions were constructed, including both linear and trigonometric functions, to approximate the geometrical shape of the relative stem profile through Eq. 4. Because there was no clear geometrical form, biological reason or statistical theory to specify a particular equation form for the variable exponent,

K , an empirical function for K had to be derived through exploratory model building. Variables including DBH, L , d , average stem taper T , and their various transformations and combinations were taken as candidate predictors for K in a linear function. For each of the 9 candidate base functions, Eq. 4 was linearized by taking logarithmic transformation of both sides and then $\ln h$ was regressed against all candidate predictors, each multiplied by $\ln B$, but without the intercept term. The best model with either 4 or 5 predictors for K was determined through stepwise regression in the PROC REG procedure of SAS 9.4 with MAXR (maximum R^2 improvement) as the method of variable selection.

The selected models for all 9 candidate base functions were then further compared with each other in terms of the goodness-of-fit statistics and also subjectively evaluated in terms of their simplicity, elegance, interpretability and applicability. Following this exploratory model building process, the linear model for $\ln h$ was specified as follows:

$$\ln h = (a_1 + a_2L + a_3\sqrt{T} + a_4\sqrt{d^3} + a_5\sqrt{\text{DBH}^3})\ln(1 - d), \tag{5}$$

where a_1 to a_5 are parameters. This linear model was back-transformed from natural logarithm to derive the nonlinear model form for h :

$$h = (1 - d)^{a_1+a_2L+a_3\sqrt{T}+a_4\sqrt{d^3}+a_5\sqrt{\text{DBH}^3}} \tag{6}$$

Because $h = (L + H_s - H_b) / (H - H_b)$ as shown in the ‘‘Notation’’ section, representing the height relative to the total tree height H minus breast height H_b , a model form for predicting H directly was obtained from Eq. 6 as follows:

$$H = \frac{L + H_s - H_b}{(1 - d)^{a_1+a_2L+a_3\sqrt{T}+a_4\sqrt{d^3}+a_5\sqrt{\text{DBH}^3}}} + H_b \tag{7}$$

Although the deterministic structures of Eqs. 5–7 are mathematically equivalent, as statistical models for predicting total tree height, they differ. The differences in their dependent variables and error structures (either multiplicative or additive) could potentially have some impact on the prediction accuracy for H .

Parameter estimation of linear and nonlinear models

Equations 5 and 7 could be estimated through linear and nonlinear least squares regression with the implicit assumption that the error term was an independent and identically distributed random variable. However, this assumption was unrealistic as revealed by the diagnostic analyses of residuals from the regressions. First, the CTL simulations processed each taper tree 6 times using 6 values of $SEDTL_\tau$ obtained from the 6 conditional quantile curves at $\tau = 0.01, 0.25, 0.50, 0.75, 0.90$ and 0.99 (Eq. 2, Fig. 5) and generated

6 sets of values of all variables appearing in Eqs. 5 and 7. Consequently, there was an inherent correlation among the residuals from the same tree. Second, the 6 values of L for each taper tree represented different proportions of its total tree height H . As L decreased and $SEDTL_\tau$ increased with τ , the magnitude of residual variation became increasingly larger, providing a clear indication of the presence of residual heteroskedasticity, particularly for the nonlinear model.

Although the least squares estimators remain unbiased in the presence of residual correlation and heteroskedasticity, they are no longer efficient and their estimates of parameter variances are biased (Gujarati and Porter 2010; Greene 2012). To take both residual correlation and heteroskedasticity into consideration, Eqs. 5 and 7 were not estimated as single equations, but each as a system of 6 equations. For the sake of parsimony, the model specification for estimating Eq. 7 only was given below:

$$\begin{aligned} H_1 &= f(X, \theta) + \varepsilon_1 \\ H_2 &= f(X, \theta) + \varepsilon_2 \\ &\vdots \\ H_6 &= f(X, \theta) + \varepsilon_6, \end{aligned} \tag{8}$$

where H_1 to H_6 represent the same H as in Eq. 7, but corresponding to data generated from the 6 cutting patterns that were based on the 6 conditional quantile curves at $\tau = 0.01, 0.25, 0.50, 0.75, 0.90$ and 0.99 (Fig. 5), $f(X, \theta)$ stands for the nonlinear function in Eq. 7 with X representing a vector of its predictor variables and θ a vector of its 5 parameters, ε_1 to ε_6 are the corresponding error terms. The 6 error terms can be expressed in the matrix algebra notation as follows:

$$\varepsilon = [\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5, \varepsilon_6]' \tag{9}$$

The properties of ε are

$$E(\varepsilon) = 0. \tag{10}$$

and

$$\text{Cov}(\varepsilon) = E(\varepsilon\varepsilon') = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} & \sigma_{16} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} & \sigma_{25} & \sigma_{26} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} & \sigma_{35} & \sigma_{36} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} & \sigma_{45} & \sigma_{46} \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_{55} & \sigma_{56} \\ \sigma_{61} & \sigma_{62} & \sigma_{63} & \sigma_{64} & \sigma_{65} & \sigma_{66} \end{bmatrix} \otimes I_N = \sum \otimes I_N, \tag{11}$$

where $E(\varepsilon)$ and $\text{Cov}(\varepsilon)$ denote the expectation and covariance of ε , σ_{ii} represents the variance of ε_i ($i = 1, \dots, 6$), σ_{ij} represents the covariance between the error term of the i th and the j th equation, ($j = 1, \dots, 6$), \otimes denotes the Kronecker product, $N = 3099$, denotes the number of observations and I_N is an identity matrix of order N . The system of Eq. 8 was fitted to the data generated from the CTL simulations using the generalised method of moments (GMM) through the

MODEL procedure of SAS/ETS with the Marquart method of minimization. The multiple-equation GMM estimator produces efficient parameter estimates under heteroscedastic conditions without any specification of the nature of the heteroscedasticity (Hayashi 2000; Greene 2012). This approach turned the residual correlation in the single-equation regression (7) into the cross-equation error correlation in the system of Eq. 8, which was then characterized by the variance and covariance matrix Σ in Eq. 11 and taken into account in parameter estimation. At the same time, it also avoided the difficulty and complexity in deriving an accurate residual skedastic function conditional upon DBH, L , SEDTL across the 6 log cutting patterns to serve as a weighting function within the frame work of generalized least squares estimation of a single equation. As the same parameter vector θ was shared across the 6 system equations, this approach effectively estimated a single equation model in Eq. 7.

Evaluating and comparing prediction accuracy

To evaluate and compare their predictive performances of Eqs. 5 and 7, a leave-one-tree-out cross-validation approach was adopted to obtain prediction errors from and for trees that were independent of the model building process. In doing so, the system of 6 equations specified for each single equation model was fitted 3099 times. Each time, 6 observations from a single tree were left out from the fitting process. Then the parameter estimates based on data from the remaining 3098 trees were used to compute the corresponding 6 prediction errors for the left-out tree as follows:

$$\hat{\varepsilon}_i = H - \hat{H}_i, \quad i = 1, \dots, 6, \quad (12)$$

where \hat{H}_i represents the predicted value of H for the i th system equation (i.e., i th cutting pattern) and $\hat{\varepsilon}_i$ is its prediction error (see Eq. 8). Upon completion, the repeated fitting and testing process generated 3099×6 prediction errors for each model for the evaluation and comparison of prediction accuracy. This leave-one-tree-out approach was based on the leave-one-out cross-validation method for model selection that was originally introduced by Stone (1974), Allen (1974) and Geisser (1975), but modified for the correlation structure of the data similarly to the cross-validation strategies reviewed by Roberts et al. (2017) for structured ecological data.

The linear and nonlinear models in Eqs. 5 and 7 were further compared with the model of Varjo (1995) in its original linear form:

$$\ln L_{\text{top}} = a_1 + a_2 L + a_3 \ln L + a_4 \ln \text{DBH} + a_5 \text{SEDTL} + a_6 \ln \text{SEDTL} \quad (13)$$

and also in its nonlinear form after back-transformation from log scale:

$$H = H_S + L + L^{a_3} \text{DBH}^{a_4} \text{SEDTL}^{a_6} e^{a_1 + a_2 L + a_5 \text{SEDTL}} \quad (14)$$

Parameters of these 2 single-equation models were also estimated through the system of equations approach as described above. Then prediction errors for each model were generated through the repeated model fitting and testing process following the leave-one-tree-out cross validation approach as previously described.

The prediction accuracy of our model was evaluated and compared with that of Varjo's in both their linear and nonlinear forms graphically and through benchmarking statistics. Scatter plots of the observed values of H against their predicted values with a line of unity slope passing through the origin were evaluated together with prediction error distributions. The benchmarking statistics included the mean error of prediction (MEP), the relative mean error of prediction (RMEP), the mean absolute error of prediction (MAEP), the mean squared error of prediction (MSEP), and the prediction coefficient of determination (R_p^2). These statistics have been commonly used in evaluating the predictive performance of forest models as they assess the size, direction and dispersion of the prediction error as reviewed by Huang et al. (2003). In particular, the MSEP is the measure of prediction accuracy commonly used in the statistical literature since it incorporates both the variance of prediction error and the bias of prediction (Wackerly et al. 1996). In addition to these statistics, skewness and kurtosis (calculated as the excess kurtosis, which is 3 less than the standardized fourth central moment) were also obtained for prediction error distributions.

These benchmarking statistics were calculated both globally over the entire data space and locally over specific subspaces of data. A natural example of such subspace division lay in the data generated by the 6 cutting patterns and represented by the 6 system equations. For the i th system equation, the 4 benchmarking statistics were calculated as follows:

$$\text{MEP}_i = \sum_j^n \hat{\varepsilon}_i / n \quad (15)$$

$$\text{RMEP}_i = 100 \left[\sum_j^n (\hat{\varepsilon}_i / H) \right] / n \quad (16)$$

$$\text{MAEP}_i = \sum_j^n |\hat{\varepsilon}_i| / n \quad (17)$$

$$\text{MSEP}_i = \sum_j^n \hat{\varepsilon}_i^2 / n \quad (18)$$

$$R_{pi}^2 = 1 - \left[\frac{\sum_j^n \hat{\varepsilon}_i^2}{\sum_j^n (H_j - \bar{H})^2} \right], \quad (19)$$

where $i = 1, \dots, 6$, $n = 3099$, i.e., the total number of taper trees, j indicates the j th tree, H_j is its total tree height and \bar{H} is the average height of all n trees. As the 6 cutting patterns represented by the 6 system equations did not have an equal frequency of occurrence, an overall weighted average mean squared error of prediction (WMSEP) was calculated as follows:

$$\text{WMSEP} = \sum_{i=1}^6 w_i \text{MSEP}_i, \quad (20)$$

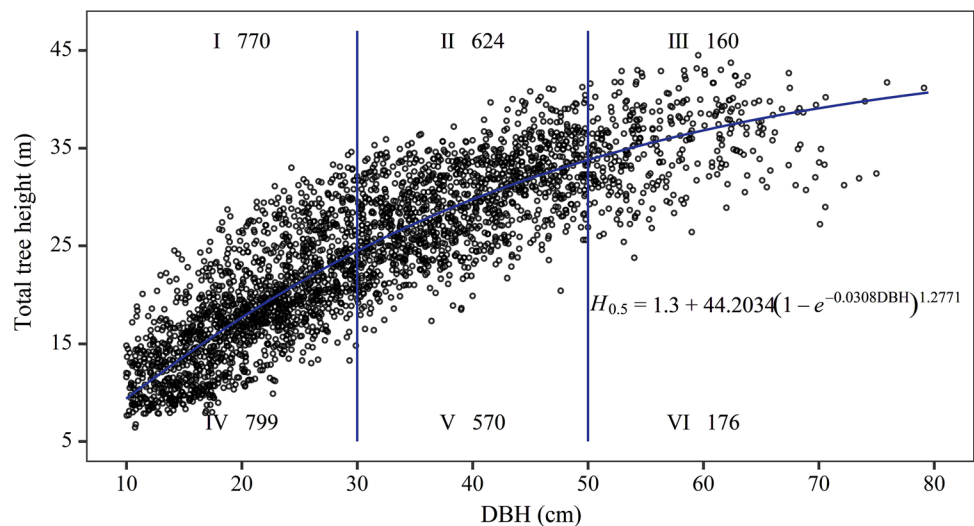
where $w_i = 0.130, 0.245, 0.250, 0.200, 0.120, 0.055$ for $i = 1, \dots, 6$. These weights were the approximate proportions that the 6 quantiles, $\tau = 0.01, 0.25, 0.50, 0.75, 0.90$ and 0.99 , represented over the range from 0 to 1, which were determined by evenly partitioning the distance between any 2 adjacent quantiles and allocating one-half to each quantile. Although analogous to the coefficient of determination R^2 , which ranges from 0 to 1, R_p^2 can range from $-\infty$ to 1 (Nash and Sutcliffe 1970). When $R_p^2 < 0$, the observed mean is a better predictor than the model as the variance of prediction error is larger than the variance of the observed data.

The 6 cutting patterns in the log cutting simulations also generated, as intended, a wide range of d values from 0.07 to 0.97, covering 90% of its theoretically defined interval between 0 and 1. The range of d was therefore divided into 10 even intervals with a width of 0.1, except for the first and last interval, for a further local evaluation and comparison of prediction accuracy. In addition, 3 other subdivisions of data were assessed for the same purpose. First, the height and diameter data of 3099 trees were divided into 6 size classes according to their DBH and H . The 0.50th nonlinear height–diameter quantile curve in the form of the Chapman–Richards equation was obtained by using the R package, `quantreg`. This median curve divided the 3099 trees into 2 halves, i.e., relative taller and shorter trees for a given DBH (Fig. 7). Then the trees in each half were further divided into 3 DBH classes using 30 and 50 cm as points of division, resulting in 6 size classes in total. Second, a ratio of total log length ratio (R_l) was calculated by dividing L with H for each tree through the 3099×6 cutting simulations. Then the entire range of R_l was divided into 9 even intervals with a width of 0.1. Third, the data were divided into 9 groups according to the total number of logs cut from each stem which ranged from 1 to 9 or more.

Results

The variance and covariance matrix Σ as exemplified in Eq. 11 and used in the multiple-equation GMM parameter estimation was

Fig. 7 Total tree height (H) plotted against DBH for the 3099 trees in the taper data set. For local evaluations of prediction accuracy, the entire data space was divided into 6 subspaces marked with Roman numerals. The division was achieved by the median height and diameter curve, for which the equation was plotted on the graph with the 2 vertical lines at DBH equal to 30 and 50 cm. The number next to each Roman numeral indicates the number of trees within that subspace



$$\begin{bmatrix} 0.0012 & 0.0014 & 0.0014 & 0.0014 & 0.0015 & 0.0019 \\ 0.0014 & 0.0033 & 0.0037 & 0.0042 & 0.0047 & 0.0051 \\ 0.0014 & 0.0037 & 0.0051 & 0.0064 & 0.0075 & 0.0077 \\ 0.0014 & 0.0042 & 0.0064 & 0.0125 & 0.0161 & 0.0157 \\ 0.0015 & 0.0047 & 0.0075 & 0.0161 & 0.0267 & 0.0265 \\ 0.0019 & 0.0051 & 0.0077 & 0.0157 & 0.0265 & 0.0367 \end{bmatrix} \text{ and } \begin{bmatrix} 0.41 & 0.54 & 0.57 & 0.61 & 0.61 & 0.68 \\ 0.54 & 1.44 & 1.65 & 1.93 & 2.13 & 2.07 \\ 0.57 & 1.65 & 2.52 & 3.17 & 3.69 & 3.50 \\ 0.61 & 1.93 & 3.17 & 6.10 & 7.66 & 6.97 \\ 0.61 & 2.13 & 3.69 & 7.66 & 12.32 & 11.62 \\ 0.68 & 2.07 & 3.50 & 6.97 & 11.62 & 15.98 \end{bmatrix}$$

for our model in its linear and nonlinear form respectively, and

$$\begin{bmatrix} 0.0748 & 0.0429 & 0.0340 & 0.0300 & 0.0295 & 0.0249 \\ 0.0429 & 0.0581 & 0.0519 & 0.0414 & 0.0333 & 0.0299 \\ 0.0340 & 0.0519 & 0.0534 & 0.0450 & 0.0376 & 0.0341 \\ 0.0300 & 0.0414 & 0.0450 & 0.0507 & 0.0488 & 0.0441 \\ 0.0295 & 0.0333 & 0.0376 & 0.0488 & 0.0567 & 0.0511 \\ 0.0249 & 0.0299 & 0.0341 & 0.0441 & 0.0511 & 0.0493 \end{bmatrix} \text{ and } \begin{bmatrix} 0.92 & 0.85 & 0.61 & 0.75 & 1.34 & 1.41 \\ 0.85 & 1.85 & 2.16 & 2.77 & 3.47 & 3.55 \\ 0.61 & 2.16 & 3.41 & 4.48 & 5.21 & 5.31 \\ 0.75 & 2.77 & 4.48 & 7.26 & 8.80 & 9.21 \\ 1.34 & 3.47 & 5.21 & 8.80 & 12.32 & 13.11 \\ 1.41 & 3.55 & 5.31 & 9.21 & 13.11 & 15.05 \end{bmatrix}$$

for Varjo’s model in its linear and nonlinear form. The diagonal and off-diagonal elements of these matrices showed that error heteroskedasticity and correlation across the 6 cutting patterns were stronger for the nonlinear than the linear models. The values of all elements of Σ for our nonlinear model in Eq. 7 were smaller than that for Varjo’s nonlinear model in Eq. 14, except for σ_{66} . As expected, all elements of the matrices for the 2 linear models were much smaller than the corresponding elements of the matrices for the 2 nonlinear models due to logarithmic transformation. Although

the dependent variable of our new model in its linear form in Eq. 5 was $\ln h$ (log-transformed relative height), some degree of heteroskedasticity was still evident as shown by the trend along the diagonal elements of the matrix. The parameters of our new model estimated in its nonlinear form in Eq. 7 differed slightly from that estimated in its linear form in Eq. 5. In comparison, such comparative differences were much larger for Varjo’s model (Table 1). The linear and nonlinear estimates of 3 parameters, a_2 , a_3 and a_5 in Eqs. 13 and 14 even had opposite signs. For both models, the parameters estimated through their nonlinear forms had appreciably smaller standard errors. The R^2 values, calculated according to Eq. 19 but using the observed and estimated values of H of all trees involved in model fitting, were slightly higher for our model in both linear and nonlinear forms than for Varjo’s.

The scatter plots of observed and predicted values of H and the corresponding distributions of prediction error generated from the leave-one-tree-out cross validation process showed little overall bias in the prediction of H for the 2 models in their linear as well as nonlinear forms (Fig. 8). The smallest MEP of 0.06 m was observed for our model in its nonlinear form, as compared with the MEP of 0.08 m for Varjo’s model also in its nonlinear form. The MEP of our model in its linear form was also smaller than that of Varjo’s model in its linear form. Our model in its nonlinear form also had the smallest value of MSEP of 6.47, as compared with its linear form and also with Varjo’s model in both forms. For both models, the nonlinear form had smaller values of MEP and MSEP than the linear form. The prediction error distributions had little skewness but pronounced kurtosis for the 2 models in both forms, but they were more leptokurtic for our model than Varjo’s (Fig. 8). These results were based on prediction errors for all 6 cutting patterns derived from using the 6 quantile curves that related SEDTL to DBH (Fig. 5). It was apparent from Fig. 8 that prediction errors for the cutting pattern derived from the 6th quantile curve at $\tau = 0.99$ had a much wider spread than all other cutting patterns. When the

Table 1 Parameter estimates and their standard errors in parentheses for the linear and nonlinear forms of our new model (Eqs. 5, 7) and Varjo’s model (Eqs. 13, 14) that are coded as NML, NMN, VML and VMN respectively in the first column of the table

Model	a_1	a_2	a_3	a_4	a_5	a_6	R^2
NML	-0.8771 (0.0101)	0.0132 (0.0003)	1.1457 (0.0052)	1.1575 (0.0076)	-0.0011 (0.0000)		0.91
NMN	-0.8191 (0.0074)	0.0112 (0.0002)	1.1540 (0.0048)	1.0784 (0.0054)	-0.0010 (0.0000)		0.91
VML	0.5063 (0.0268)	0.0139 (0.0005)	-0.0961 (0.0021)	-0.6395 (0.0085)	0.0013 (0.0003)	1.4275 (0.0046)	0.89
VMN	0.0002 (0.0117)	-0.0041 (0.0001)	0.0068 (0.0004)	-0.6713 (0.0004)	-0.0127 (0.0001)	1.7363 (0.0027)	0.90

The R^2 values for the multiple-equation GMM estimation were calculated according to Eq. 19 but using the observed and estimated values of total tree height of all trees involved in model fitting

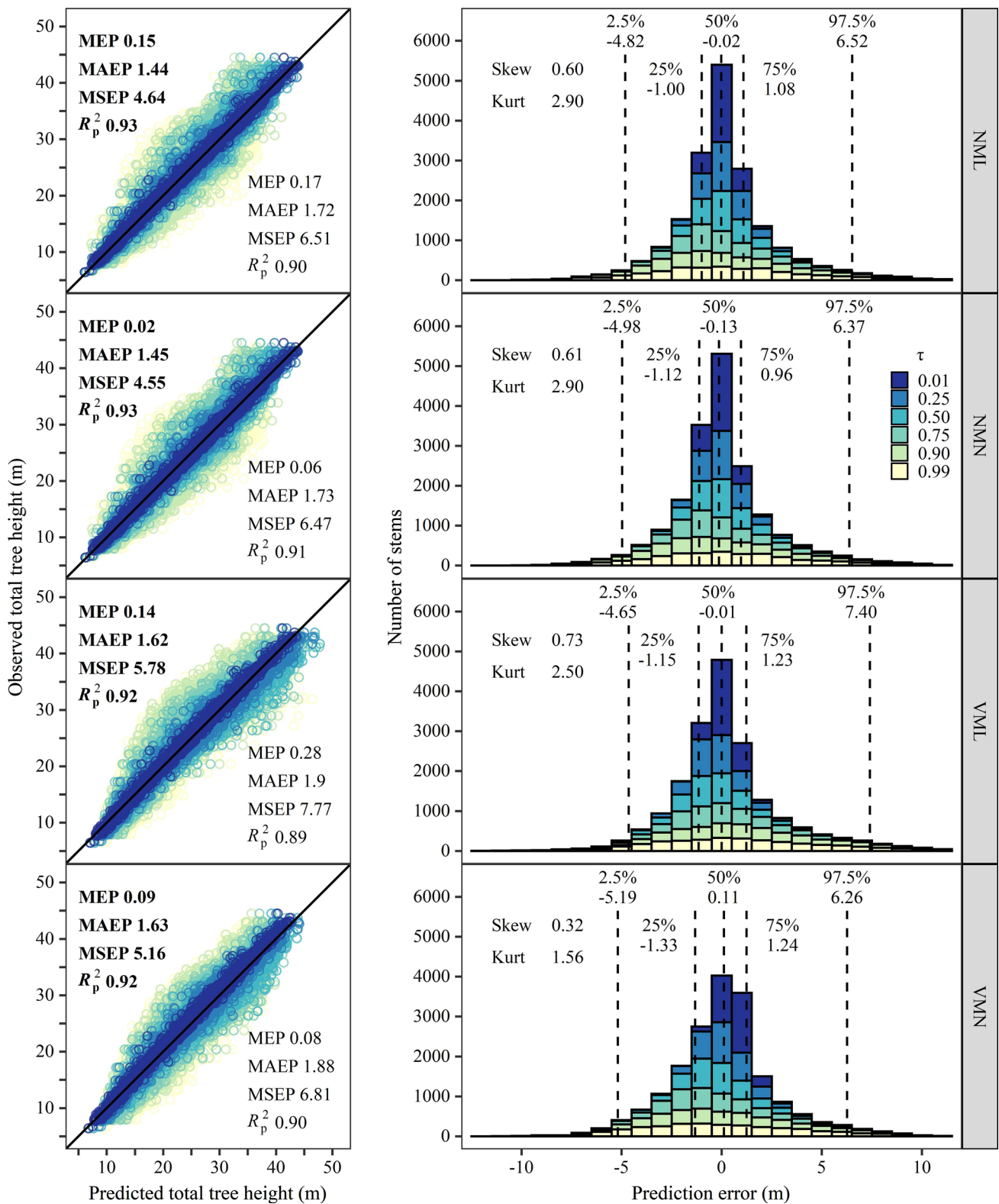


Fig. 8 Observed total tree height plotted against predicted values with a line of unity on the left for the linear and nonlinear forms of our new model (Eqs. 5, 7) and Varjo’s model (Eqs. 13, 14) that are labelled as NML, NMN, VML and VMN respectively in the shaded area on the far right. The corresponding distributions of prediction

error are displayed on the right with characteristic percentiles and values of skewness and kurtosis. The color key is the same as in Fig. 6. The benchmarking statistics in the bottom right corner of the scatter plots were based on all 6 cutting patterns; those in the top left corner were for the first 5 cutting patterns only

Table 2 MSEP for each of the 6 cutting patterns represented by the system of 6 equations (8) and the weighted average (WMSEP) calculated according to Eq. 20 for the linear and nonlinear forms of our new model (Eqs. 5, 7) and Varjo’s model (Eqs. 13, 14) that are coded as NML, NMN, VML and VMN, respectively, in the first column of the table

Model	MSEP ₁	MSEP ₂	MSEP ₃	MSEP ₄	MSEP ₅	MSEP ₆	WMSEP
NML	0.42	1.48	2.55	6.11	12.63	15.90	4.67
NMN	0.41	1.44	2.52	6.09	12.32	16.04	4.61
NMN versus NML	2.3%	2.6%	1.1%	0.4%	2.4%	−0.9%	1.2%
VML	0.46	1.90	3.88	7.53	15.14	17.70	5.79
VMN	0.91	1.85	3.41	7.28	12.34	15.07	5.19
VMN versus VML	−97.7%	2.4%	12.2%	3.3%	18.5%	14.8%	10.4%
NML versus VML	8.8%	22.0%	34.4%	18.8%	16.6%	10.1%	19.4%
NML versus VMN	53.9%	20.1%	25.3%	16.0%	−2.3%	−5.5%	10.1%
NMN versus VML	10.9%	24.0%	35.1%	19.1%	18.6%	9.4%	20.4%
NMN versus VMN	54.9%	22.2%	26.1%	16.3%	0.2%	−6.4%	11.2%

Percentage differences in the 2 benchmarking statistics were calculated for the pairwise comparison of the 4 equations

practically rare cases represented by this cutting pattern were excluded, MSEP became much smaller and R_p^2 higher for the 2 models in both forms (Fig. 8). Our model in its nonlinear form still had the smallest MEP and MSEP among the 2 models and their different forms.

When prediction accuracy was evaluated across the 6 cutting patterns individually, our model in its nonlinear form had smaller values of MSEP than its linear form with the exception of the 6th cutting pattern. However, the differences were small, representing a reduction mostly less than 2.6%. The weighted average of the 6 MSEP values showed even a smaller reduction of 1.2% (Table 2). Varjo’s model in its nonlinear form also had smaller values of MSEP than its linear form except for the first cutting pattern as its MSEP became twice as larger, pointing to a marked deterioration in predictive performance. No matter in what form our model was estimated, it had smaller MSEP values than Varjo’s model in both forms across the first 4 cutting patterns, with the first cutting pattern having the largest reduction of more than 50%. For the last 2 cutting patterns, the MSEP values of our model were either similar to or slightly larger than that of Varjo’s. The weighted average MSEP of our model represented about 20% reduction of that of Varjo’s model in its linear form

and about 10–11% reduction in its nonlinear form (Table 2). For each model, the values of $MSEP_1 - MSEP_6$ in Table 2 corresponded almost exactly to the error variances $\sigma_{11} - \sigma_{66}$, i.e., the diagonal elements of its variance and covariance matrix. In addition to MSEP, the other 4 benchmarking statistics also indicated that our model outperformed Varjo’s in both linear and nonlinear forms. In the nonlinear form in particular, our model had smaller values of MEP, RMEP and MAEP higher values of R_p^2 across the 6 cutting patterns except for the last (Table 3).

As the nonlinear form outperformed the linear form for both models, the comparative performances of the 2 models in further local evaluations of prediction accuracy were only reported for their nonlinear forms for the sake of parsimony. Across the 10 intervals of d , our model had smaller values of MEP, MAEP and MSEP and higher R_p^2 values than Varjo’s over the first 7 intervals where $d \leq 0.80$ (Table 4). In the interval $0.80 < d \leq 0.90$, the predictive performances of the 2 models were about the same. When $0.90 < d \leq 0.95$, our model did not perform as well as Varjo’s. For d greater than 0.95, both models did not perform well as indicated by their large values of MEP, MAEP and MSEP as well as the small values of R_p^2 (Table 4). Over the 6 tree size classes delineated in Fig. 7, our model had smaller values of MEP, MAEP and

Table 3 Values of MEP, RMEP, MAEP and R_p^2 for the nonlinear form of our new model (NMN) in Eq. 7 and that of Varjo’s model (VMN) in Eq. 14 across the 6 cutting patterns derived from the τ th conditional quantile curves in Fig. 5

τ	MEP		RMEP		MAEP		R_p^2	
	NMN	VMN	NMN (%)	VMN (%)	NMN	VMN	NMN	VMN
0.01	−0.03	0.64	−0.13	2.30	0.48	0.76	0.99	0.99
0.25	0.05	0.24	−0.09	0.12	0.88	1.03	0.98	0.97
0.50	0.02	−0.19	−0.50	−1.78	1.17	1.40	0.96	0.95
0.75	−0.03	−0.36	−1.73	−3.49	1.91	2.14	0.91	0.89
0.90	0.06	0.12	−2.44	−2.66	2.79	2.81	0.82	0.82
0.99	0.25	0.04	−0.74	−3.83	3.13	3.16	0.77	0.78
All	0.06	0.08	−0.94	−1.55	1.73	1.88	0.91	0.90

Table 4 Benchmarking statistics of prediction accuracy based on prediction errors for all 6 cutting patterns: MEP, MAEP, MSEP and R_p^2 across the 10 relative diameter classes for the nonlinear form of our new model (NMN) in Eq. 7 and that of Varjo's model (VMN) in Eq. 14 presented in the first and second row in each row group, respectively

Relative diameter d	N	Median Skewness Kurtosis	MEP (m)	MAEP (m)	MSEP	R_p^2
$d \leq 0.20$	1817	-0.15	-0.10	0.48	0.39	0.99
		0.88	0.81	0.85	1.06	0.97
		2.78	(-0.12)	(0.57)	(0.37)	
$0.20 < d \leq 0.30$	954	0.02	0.05	0.46	0.40	0.98
		0.97	0.51	0.64	0.72	0.97
		6.44	(0.10)	(0.71)	(0.55)	
$0.30 < d \leq 0.40$	2599	-0.03	0.04	0.88	1.42	0.98
		0.70	0.30	0.99	1.75	0.97
		2.29	(0.14)	(0.89)	(0.81)	
$0.40 < d \leq 0.50$	3081	-0.12	0.00	1.08	2.14	0.96
		0.60	-0.03	1.27	2.84	0.95
		1.91	(-0.09)	(0.85)	(0.75)	
$0.50 < d \leq 0.60$	741	-0.07	0.22	1.10	2.68	0.98
		1.33	-0.66	1.48	4.00	0.96
		3.55	(-0.33)	(0.74)	(0.67)	
$0.60 < d \leq 0.70$	3125	-0.43	-0.03	1.89	6.02	0.91
		0.76	-0.37	2.13	7.22	0.90
		0.93	(0.08)	(0.89)	(0.83)	
$0.70 < d \leq 0.80$	3197	-0.41	0.12	2.77	12.29	0.83
		0.51	0.14	2.79	12.30	0.83
		0.11	(0.85)	(0.99)	(1.00)	
$0.80 < d \leq 0.90$	2194	-0.22	-0.02	3.13	15.80	0.79
		0.26	0.17	3.17	15.18	0.80
		0.26	(-0.14)	(0.99)	(1.04)	
$0.90 < d \leq 0.95$	873	0.40	0.69	3.06	15.86	0.61
		0.18	-0.37	3.07	14.20	0.65
		0.14	(-1.85)	(0.99)	(1.12)	
$0.95 < d < 0.98$	13	3.14	3.43	3.79	26.42	0.04
		0.83	2.40	2.89	14.05	0.49
		0.35	(1.43)	(1.31)	(1.88)	

The ratios displayed in parentheses in the third row were that of NMN over VMN in the corresponding statistics. N is the number of observations in each relative diameter class. The median, skewness and kurtosis of prediction error distribution presented in a single column were for our new model (NMN) only

MSEP and higher R_p^2 values except for size class III where the comparative performances were reversed (Fig. 9). For both models, the values of MEP were positive over the first 3 size classes and negative over the last 3 size classes. This systematic pattern of bias represented a slight underestimation of total tree height for relatively slender trees and a slight overestimation for relatively squatter trees. However, the bias was relatively small, mostly within ± 1.5 m and within $\pm 5\%$ of observed tree height on average. As with results in Fig. 8, values of MEP, MAEP and MSEP became much smaller and R_p^2 was higher when these benchmarking statistics were based on prediction errors for the first 5 cutting patterns excluding the practically rare cases represented by the 6th cutting pattern (Fig. 9).

Across the 9 intervals of total log length ratio R_l , our model generally outperformed Varjo's as indicated by the benchmarking statistics also based on prediction errors for the first 5 cutting patterns (Table 5). When $R_l \leq 0.2$, i.e., the total log length was less than or equal to 20% of total tree height, the values of MEP, MAEP and MSEP were much larger than that for all other intervals for both models and their R_p^2 was negative. In this first interval, our model did not perform as well as Varjo's. When $0.20 < R_l \leq 0.30$ and $0.30 < R_l \leq 0.40$, our model had a similar or slightly better performance. For the remaining 6 R_l intervals, our model clearly outperformed Varjo's based on the values of MAEP and MSEP. For both models, the bias was small and practically negligible as their MEP values were well within ± 0.8 m. Among the 9 groups of the number of logs cut from

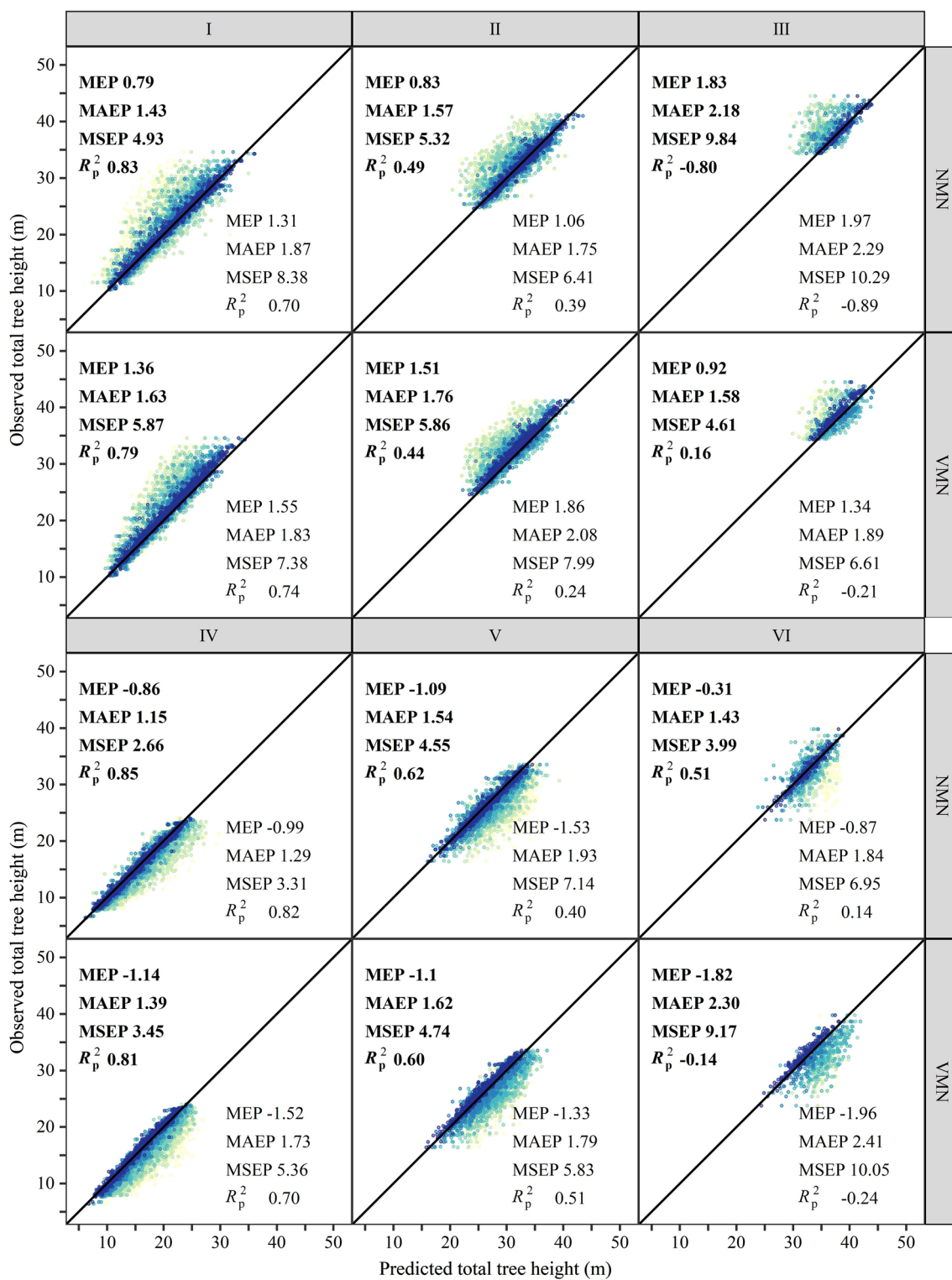


Fig. 9 Observed total tree height plotted against predicted values with a line of unity over the 6 subspaces of data as divided in Fig. 7 for the linear and nonlinear forms of our new model (Eqs. 5, 7) and Varjo's model (Eqs. 13, 14) respectively labelled as in Fig. 8. The

color scheme was the same as in Figs. 6 and 7. The benchmarking statistics in the bottom right corner of the scatter plots were based on all 6 cutting patterns; those in the top left corner were for the first 5 cutting patterns only

Table 5 Benchmarking statistics of prediction accuracy based on prediction errors for the first 5 cutting patterns: MEP, MAEP, MSEP and R_p^2 across the 10 total log length ratio (R_l) classes for the nonlinear form of our new model (NMN) in Eq. 7 and that of Varjo's model (VMN) in Eq. 14 presented in the first and second row in each row group, respectively

Total log length ratio (R_l)	N	Median Skewness Kurtosis	MEP (m)	MAEP (m)	MSEP	R_p^2
0.10 < R_l ≤ 0.20	316	5.46	5.15	5.29	36.26	-0.36
		-0.31	4.59	4.86	31.63	-0.19
0.20 < R_l ≤ 0.30	1647	-0.59	(1.12)	(1.09)	(1.15)	
		0.71	0.83	2.44	9.44	0.82
0.30 < R_l ≤ 0.40	1306	0.19	0.66	2.60	10.61	0.80
		-0.16	(1.26)	(0.94)	(0.89)	
0.40 < R_l ≤ 0.50	1854	-1.53	-0.91	2.84	11.97	0.84
		0.69	-0.90	2.67	10.80	0.86
0.50 < R_l ≤ 0.60	1600	0.33	(1.01)	(1.06)	(1.11)	
		0.12	0.28	1.67	4.89	0.94
0.60 < R_l ≤ 0.70	2436	-0.04	-0.20	2.09	6.91	0.92
		1.06	(-1.40)	(0.80)	(0.71)	
0.70 < R_l ≤ 0.80	2833	-0.99	-0.76	1.79	4.90	0.93
		0.62	-0.85	1.88	5.53	0.92
0.80 < R_l ≤ 0.90	2119	0.67	(0.89)	(0.95)	(0.89)	
		0.09	0.21	1.00	1.92	0.97
0.90 < R_l < 0.99	1384	0.10	0.00	1.33	3.14	0.96
		2.51	(-51.80)	(0.75)	(0.61)	
0.10 < R_l < 0.99	15,495	-0.07	-0.12	0.87	1.37	0.98
		-0.24	0.14	1.04	1.90	0.97
0.10 < R_l < 0.99	15,495	1.26	(-0.84)	(0.84)	(0.72)	
		-0.13	-0.25	0.69	0.90	0.98
0.10 < R_l < 0.99	15,495	-0.59	0.28	0.90	1.31	0.98
		1.73	(-0.89)	(0.76)	(0.69)	
0.10 < R_l < 0.99	15,495	-0.37	-0.37	0.49	0.39	0.99
		-0.43	0.58	0.68	0.66	0.98
0.10 < R_l < 0.99	15,495	3.01	(-0.63)	(0.73)	(0.59)	
		-0.14	0.02	1.45	4.55	0.93
0.10 < R_l < 0.99	15,495	0.79	0.09	1.63	5.16	0.92
		3.28	(0.18)	(0.89)	(0.88)	

The ratios displayed in parentheses in the third row were that of NMN over VMN in the corresponding statistics. N is the number of observations in each relative diameter class. The median, skewness and kurtosis of prediction error distribution presented in a single column were for our new model (NMN) only

a stem, MEP varied between -0.12 and 0.13 m for our model and between -0.51 and 0.85 for Varjo's model (Table 6). The values of MAEP and MSEP of our model were much smaller, and the values of R_p^2 were higher for our model than Varjo's. The comparative advantage of our model increased as the number of logs cut from a stem increased.

Discussion

To our knowledge, this study represents the first rigorous attempt at developing a model for predicting total tree height from CTL harvester data. Based on the global

benchmarking statistics in Table 2, the nonlinear form of our model in Eq. 7 ranked first and its linear form in Eq. 5 second, while the nonlinear form of Varjo's model in Eq. 14 ranked third and its linear form in Eq. 13 last. The superior performance of our nonlinear model was also reflected by the values of the diagonal elements, $\sigma_{11} - \sigma_{66}$, of the variance and covariance matrices in the results section. Our nonlinear model also proved to be far superior than the iterative search algorithm using a taper equation, an existing ad hoc approach demonstrated by Lu et al. (2018) and used by Hauglin et al. (2018). The detailed comparative results formed part of the first author's postgraduate work but were not presented in the present paper for the sake of parsimony and because the iterative search algorithm was already found by Lu et al.

Table 6 Benchmarking statistics of prediction accuracy based on prediction errors for the first 5 cutting patterns: MEP, MAEP, MSEP and R_p^2 across the 9 sequential log number categories for the nonlinear form of our new model (NMN) in Eq. 7 and that of Varjo’s model (VMN) in Eq. 14 presented in the first and second row in each row group, respectively

Number of logs	N	R_l	MEP (m)	MAEP (m)	MSEP	R_p^2
1	1286	0.13	-0.04	2.30	9.68	0.80
		0.34	-0.51	2.55	10.40	0.79
		0.76	(0.08)	(0.90)	(0.93)	
2	4970	0.13	-0.06	1.77	6.39	0.90
		0.48	-0.17	1.91	6.93	0.89
		0.94	(0.38)	(0.93)	(0.92)	
3	3204	0.21	0.11	1.38	4.06	0.89
		0.63	0.17	1.48	4.54	0.88
		0.96	(0.63)	(0.93)	(0.89)	
4	2509	0.22	0.13	1.24	3.12	0.90
		0.70	0.24	1.43	3.83	0.87
		0.95	(0.54)	(0.87)	(0.81)	
5	1578	0.36	0.00	0.99	1.85	0.91
		0.77	0.38	1.23	2.56	0.88
		0.96	(0.01)	(0.81)	(0.72)	
6	1035	0.35	0.02	0.89	1.68	0.91
		0.82	0.53	1.19	2.41	0.87
		0.97	(0.04)	(0.75)	(0.70)	
7	456	0.47	-0.12	0.74	1.06	0.94
		0.86	0.51	0.99	1.55	0.92
		0.97	(-0.24)	(0.75)	(0.69)	
8	253	0.53	-0.05	0.66	1.09	0.93
		0.87	0.67	1.04	1.71	0.89
		0.96	(-0.07)	(0.64)	(0.64)	
9–11	204	0.65	-0.06	0.56	0.64	0.93
		0.91	0.85	1.02	1.53	0.84
		0.96	(-0.07)	(0.54)	(0.42)	

The ratios displayed in parentheses in the third row were that of NMN over VMN in the corresponding statistics. N is the number of observations in each category. The minimum, mean and maximum total log length ratio (R_l) represented by logs in each category were presented in a single column from top down

(2018) to be inferior to Varjo’s model in its original linear form in Eq. 14, the worst performing among the 4 equations reported in this study.

Although the best performer globally, our nonlinear model in Eq. 7 did not perform as well as Varjo’s nonlinear model in Eq. 14 in certain local areas of the data space. When $d > 0.90$, our model had larger MSEP and lower R_p^2 than Varjo’s (Table 4). Where total log length represented less than or equal to 20% of the total tree height, Eq. 7 did not perform as well as the nonlinear form of Varjo’s model. However, the cases with total log length ratio $R_l \leq 0.20$ only accounted for about 2% of the total number of simulated log cuttings (Table 5). Among the 0.448×10^6 stems contained in the screened and filtered harvester data set, such cases accounted for a small proportion of 4.9% approximately. This approximation was based on the values of R_l calculated for individual stems using their predicted total tree height from our nonlinear model in Eq. 7.

A posterior exploratory analysis showed that stems with $R_l \leq 0.20$ were trees with DBH from 10.2 to 81.8 cm, with a mean of 41.0 cm, an upper and lower quartile of 33.8 and 47.5 cm, and a 2.5th and 97.5th percentile of 23.1 and 63.6 cm, respectively. Besides stems with $R_l \leq 0.20$, another subspace of data where Eq. 7 did not perform as well as the nonlinear form of Varjo’s model was the larger slenderer trees with DBH greater than 50 cm and H above the median height–diameter curve as delineated in Fig. 7. These trees had a height and diameter ratio (HDR, i.e., H/DBH) ranging from 0.52 to 0.83 with an average of 0.68. Among the 0.448 million stems, such slender trees represented a small proportion of 3.5% and had HDR (i.e., calculated as predicted H over DBH) from 0.56 to 1.13, with an average of 0.70.

The poorer performance of our model for $R_l \leq 0.20$ and $d \geq 0.90$ reflected a structural weakness of our model. As defined in the “Notation” section, d takes any value between 0 and 1. When the value of SEDTL approaches DBH, $d \rightarrow 1$

and the base function of $(1 - d)$ tends to 0. A very small positive number with a positive exponent as formulated in Eq. 7 would result in a relatively large predicted value of H . Conversely, a positive exponent would lead to a relatively small predicted value of H . As a result, the variance of prediction error became much larger as shown by the large values of MSE in Tables 4 and 5. Based on the benchmarking statistics for the local evaluations of prediction accuracy, the applicable range of our nonlinear model for *P. radiata* stems processed and recorded by CTL harvesters should be where values of $R_1 > 0.20$ and $d < 0.95$. For stems outside of this applicable range, total tree height could still be estimated but indirectly based on model estimates for trees within the applicable range and also within a user-defined neighborhood or local area. Using the estimated H of these trees and their spatial co-ordinates recorded in the harvester data, stem-specific and spatially varying geographically weighted linear or nonlinear height–diameter equations could be derived for more accurate total tree height estimation for such stems following the approach of Zhang et al. (2003) and Caccamo et al. (2018).

The approach of estimating the parameters of a single-equation model through the multiple-equation GMM estimator was adopted in this study specifically to obtain efficient and consistent parameter estimates in the presence of error correlation and heteroscedasticity that were inherent to the systematic structure of data generated by the CTL simulations. This approach proved to be better than all single-equation methods that were evaluated during model derivation and parameter estimation. Although the comparative results were not reported here, these single-equation methods for the log-transformed linear models included (1) least squares regression (LSQ), (2) repeated sampling and fitting through LSQ, each time using data from only 1 of the 6 cutting patterns randomly selected from each tree to avoid error autocorrelation, and (3) LSQ with discrete as well as continuous AR1 (first-order autoregressive) errors. For the nonlinear models, these single-equation methods included (1) nonlinear least squares regression (NLSQ), (2) weighted nonlinear least squares regression (WNLSQ) to overcome heteroscedasticity, (3) repeated sampling and fitting through WNLSQ to reduce heteroscedasticity and at the same time to avoid error autocorrelation, and (4) WNLSQ with both discrete and continuous AR1 errors. In addition, the generalized estimating equations (GEE) implemented in the SAS macro %NLMIX as described by Vonesh (2012) was also attempted, but the estimation was not successful because of difficulties in achieving convergence. All these single-equation methods were compared with the multiple-equation GMM estimator and evaluated through the leave one-tree-out cross-validation process and the benchmarking statistics as described previously. The multiple-equation GMM estimator provided not only efficient and consistent

parameter estimates in the presence of error correlation and heteroscedasticity in the structured data, but also the best predictive performance.

The repeated sampling and fitting also exposed a structural weakness of Varjo's model due to the high degree of multicollinearity among its predictor variables as indicated by the collinearity diagnostics during parameter estimation of its linear as well as nonlinear form. The variance inflation factor was either close to the commonly used benchmark of 10 or much greater than 10 for the predictor variables in the LSQ estimation of a single equation, while the condition index was greater than 110 and 220 in the linear and nonlinear multiple-equation GMM estimations, well above the commonly recognized threshold of 30 (Belsley 1991; Galimacci 1996; Alin 2010; Friendly and Kwan 2009). The high degree of multicollinearity led to highly variable and unstable parameter estimates. As a result, there tended to be either 1 or 2 estimated parameters that were not significantly different from zero as found during repeated sampling and fitting. A small change in the data could result in relatively large changes in parameter estimates, even to extent of switching signs. Therefore, it was not surprising to see (1) the linear and nonlinear estimates of some parameters having opposite signs as shown in Table 1 and (2) the comparatively large difference in the predictive performance between the linear and nonlinear forms of Varjo's model (Table 2). Even for the same linear form, the model fitted by Lu et al. (2018) to data generated from a single cutting pattern based on a much smaller harvester data set had parameters of different signs to that estimated in this study. Although detailed comparative results were not reported here for the sake of parsimony, its predictive performance became much poorer when tested over the much larger data space in this study, reflecting the potential impact of multicollinearity on the predictive ability of the model beyond its original training data space where the nature and degree of multicollinearity differed. Such effects of multicollinearity in linear and nonlinear regression models have long been recognised (Belsley 1984, 1991; Galimacci 1996; Alin 2010; Erkoç et al. 2010). In comparison, our new model in both linear and nonlinear forms did not suffer from the same problem, the condition index was 28 and 23 in the linear and nonlinear multiple-equation GMM estimations, respectively, all below the benchmark value of 30. This relatively weak multicollinearity among predictor variables of our model would certainly contribute to its superior predictive performance.

Interest has been growing in making a greater use of harvester data in forest management and planning among both researchers and managers over the last 10 years (Möller et al. 2011; Olivera and Visser 2016; Roth 2016). Most recently, harvester data analytics has been identified to be essential for the successful transformation of big forestry data into valuable data for management and envisaged to be an integral

and indispensable part in the generation of a virtual forest for subsequent digitization of the wood supply chain within the internet of trees and services in the future (Müller et al. 2019). As shown by Möller et al. (2011), accurate estimation of total tree height for harvested stems represents a necessary basic step in harvester data analytics. Our new model for predicting total tree height will facilitate and widen the utilization of harvester data far beyond the current limited use in the management of radiata pine plantations, i.e., mostly for log production monitoring and reporting. It will enable the full integration of harvester data with conventional inventory data, remote sensing imagery and LiDAR data for the development of harvester-based inventory systems, for the prediction of attributes of individual trees, stands and forests, and for estimating product recovery and residue biomass in radiata pine plantations. It will also facilitate (1) the screening and exploratory analysis of harvester data, (2) calibration and estimation of bark thickness, (3) mapping of site index, (4) development site-specific height–diameter curves, and (5) post-thinning assessment of diameter and height distributions of retained versus removed stems. In addition, accurately estimated total tree height will make harvester data a potential source of taper data to supplement and possibly reduce the intensity and cost of the conventional destructive taper sampling in the field. This list is far from exhaustive. Many other applications of our model are expected to emerge as harvester data analytics become increasingly refined and sophisticated with further development. Although our new model was developed for radiata pine, its mathematical form will be applicable to other tree species in plantations as well as natural forests where CTL harvester data are routinely captured during thinning and harvesting operations.

Acknowledgements We thank Mr. Mike Sutton of the Forestry Corporation of NSW for providing the data license to Beijing Forestry University for this collaborative work. We are indebted to the past and present forestry staff of the Forestry Commission of NSW, State Forests of NSW, Forests NSW, and the Forestry Corporation of NSW who collected the data in the field. Dr. Glen Murphy provided helpful comments on the manuscript.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Alin A (2010) Multicollinearity. Wiley Interdiscip Rev Comput Stat 2(3):370–374
- Allen DM (1974) The relationship between variable selection and data augmentation and a method for prediction. Technometrics 16(1):125–127
- Barth A, Holmgren J (2013) Stem taper estimates based on airborne laser scanning and cut-to-length harvester measurements for pre-harvest planning. Int J For Eng 24(3):161–169
- Barth A, Möller JJ, Wilhelmsson L, Arlinger J, Hedberg R, Söderman U (2015) A Swedish case study on the prediction of detailed product recovery from individual stem profiles based on airborne laser scanning. Ann Sci 72(1):47–56
- Belsley DA (1984) Collinearity and forecasting. J Forecast 3(2):183–196
- Belsley DA (1991) Conditioning diagnostics: collinearity and weak data in regression. Wiley Series in Probability, New York, p 396
- Bi H (2000) Trigonometric variable-form taper equation for Australian eucalypts. For Sci 46(3):397–409
- Bi H, Long Y (2001) Flexible taper equation for site-specific management of *Pinus radiata* in New South Wales, Australia. For Ecol Manag 148(1):79–91
- Bi H, Jurskis V, O'Gara J (2000) Improving height prediction of regrowth eucalypts by incorporating the mean size of site trees in a modified Chapman–Richards equation. Aust For 63(4):257–266
- Bi H, Fox JC, Li Y, Lei Y, Pang Y (2012) Evaluation of nonlinear equations for predicting diameter from tree height. Can J For Res 42(4):789–806
- Caccamo G, Iqbal IA, Osborn J, Bi H, Arkley K, Melville G, Aurik D, Stone C (2018) Comparing yield estimates derived from LiDAR and aerial photogrammetric point-cloud data with cut-to-length harvester data in a *Pinus radiata* plantation in Tasmania. Aust For 81(3):131–141
- Cook D, Swayne DF, Buja A (2007) Interactive and dynamic graphics for data analysis: with R and GGobi. Springer, New York, p 188
- Erkoç A, Tez M, Akay KU (2010) On multicollinearity in nonlinear regression models. Selçuk J Appl Math, Special Issue: 65–72
- Friendly M, Kwan E (2009) Where's Waldo? Visualizing collinearity diagnostics. Am Stat 63(1):56–65
- Galmacci G (1996) Collinearity detection in linear regression models. Comput Econ 9(3):215–227
- Geisser S (1975) The predictive sample reuse method with applications. J Am Stat Assoc 70(350):320–328
- Gerasimov Y, Seliverstov A, Syuney V (2012) Industrial round-wood damage and operational efficiency losses associated with the maintenance of a single-grip harvester head model: a case study in Russia. Forests 3(4):864–880
- Gerasimov Y, Sokolov A, Syuney V (2013) Development trends and future prospects of cut-to-length machinery. Adv Mater Res 705:468–473
- Greene WH (2012) Econometric analysis, 7th edn. Prentice Hall, Boston, p 1232
- Gujarati DN, Porter DC (2010) Essentials of econometrics, 4th edn. Irwin/McGraw-Hill, Boston, p 554
- Hauglin M, Hansen E, Sørngård E, Næsset E, Gobakken T (2018) Utilizing accurately positioned harvester data: modelling forest volume with airborne laser scanning. Can J For Res 48(999):1–10
- Hayashi F (2000) Econometrics. Princeton University Press, Princeton NJ, p 712
- Heinimann HR (2007) Forest operations engineering and management—the ways behind and ahead of a scientific discipline. Croat J For Eng 28(1):107–121

- Holmgren J, Barth A, Larsson H, Olsson H (2012) Prediction of stem attributes by combining airborne laser scanning and measurements from harvesters. *Silva Fenn* 46(2):227–239
- Holopainen M, Vastaranta M, Rasinmäki J, Kalliovirta J, Mäkinen A, Haapanen R, Melkas T, Yu X, Hyypä J (2010) Uncertainty in timber assortment estimates predicted from forest inventory data. *Eur J For Res* 129(6):1131–1142
- Horne R, Robinson GL (1988) Development of basal area thinning prescriptions and predictive yield models for *Pinus Radiata* plantations in New South Wales, 1962–1988. Forestry Commission of New South Wales, Sydney, p 37
- Huang SM (1999) Ecoregion-based individual tree height-diameter models for lodgepole pine in Alberta. *West J Appl For* 14(4):186–193
- Huang SM, Titus SJ, Wiens DP (1992) Comparison of non-linear height-diameter functions for major Alberta tree species. *Can J For Res* 22(9):1297–1304
- Huang SM, Yang Y, Wang Y (2003) A critical look at procedures for validating growth and yield models. In: Amaro A, Reed D, Soares P (eds) *Modelling forest systems*. CABI Publishing, Oxford, pp 271–293
- Huyler NK, LeDoux CB (1999) Performance of a cut-to-length harvester in a single-tree and group selection cut. USDA Forestry Service, Northeastern Research Station, Research Paper NE-711, p 6
- Kiljunen N (2002) Estimating dry mass of logging residues from final cuttings using a harvester data management system. *Int J For Eng* 13(1):17–25
- Koenker R (2017) Quantile regression: 40 years on. *Annu Rev Econ* 9(1):155–176. <https://doi.org/10.1146/annurev-economics-063016103651>
- Koenker R (2018) quantreg: quantile regression. R package version 5.38. <https://cran.r-project.org/package=quantreg>. Accessed 15 Nov 2018
- Lang AH, Baker SA, Greene WD, Murphy GE (2010) Individual stem value recovery of modified and conventional tree-length systems in the southeastern United States. *Int J For Eng* 21(1):7–11
- Lu K, Bi H, Watt D, Strandgard M, Li Y (2018) Reconstructing the size of individual trees using log data from cut-to-length harvesters in *Pinus radiata* plantations: a case study in NSW, Australia. *J For Res* 29(1):13–33
- Malinen J, Laitila J, Väättäin K, Viitamäki K (2016) Variation in age, annual usage and resale price of cut-to-length machinery in different regions of Europe. *Int J For Eng* 27(2):95–102
- Maltamo M, Bollandsås OM, Vauhkonen J, Breidenbach J, Gobakken T, Næsset E (2010) Comparing different methods for prediction of mean crown height in Norway spruce stands using airborne laser scanner data. *Forestry* 83(3):257–268
- Möller JJ, Arlinger J, Hannrup B, Larsson W, Barth A (2011) Harvester data as a base for management of forest operations and feedback to forest owners. In: Ackerman P, Ham H and Gleasure E (eds) *Proceedings of 4th forest engineering conference: innovation in forest engineering—adapting to structural change*. Stellenbosch University, White River, South Africa, 5–7 April 2011, pp 31–35
- Müller F, Jaeger D, Hanewinkel M (2019) Digitization in wood supply—a review on how Industry 4.0 will change the forest value chain. *Comput Electron Agric* 162:206–218
- Murphy G (2003) Procedures for scanning radiata pine stem dimensions and quality on mechanised processors. *Int J For Eng* 14(2):11–21
- Murphy G, Wilson I, Barr B (2006) Developing methods for pre-harvest inventories which use a harvester as the sampling tool. *Aust For* 69(1):9–15
- Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I—A discussion of principles. *J Hydrol* 10(3):282–290
- Nordfjell T, Björheden R, Thor M, Wåsterlund I (2010) Changes in technical performance, mechanical availability and prices of machines used in forest operations in Sweden from 1985 to 2010. *Scand J For Res* 25(4):382–389
- Olivera A, Visser R (2016) Development of forest-yield maps generated from Global Navigation Satellite System (GNSS)-enabled harvester StanForD files: preliminary concepts. *N Z J For Sci* 46(1):3
- Olivera A, Visser R, Acuna M, Morgenroth J (2016) Automatic GNSS-enabled harvester data collection as a tool to evaluate factors affecting harvester productivity in a Eucalyptus spp. harvesting operation in Uruguay. *Int J For Eng* 27(1):15–28
- Palander T, Vesa L, Tokola T, Pihlaja P, Ovaskainen H (2009) Modelling the stump biomass of stands for energy production using a harvester data management system. *Biosyst Eng* 102(1):69–74
- Peuhkurinen J, Maltamo M, Malinen J (2008) Estimating species-specific diameter distributions and saw log recoveries of boreal forests from airborne laser scanning data and aerial photographs: a distribution-based approach. *Silva Fenn* 42(4):625–641
- Rasinmäki J, Melkas T (2005) A method for estimating tree composition and volume using harvester data. *Scand J For Res* 20(1):85–95
- Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, Hauenstein S, Lahoz-Monfort JJ, Schröder B, Thuiller W, Warton DI (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40(8):913–929
- Rossit DA, Olivera A, Céspedes VV, Broz D (2019) A Big Data approach to forestry harvesting productivity. *Comput Electron Agric* 161:29–52
- Roth G (2016) StanForD as a data source for forest management: a forest stand reconciliation implementation case study. M.Sc. thesis, University of Canterbury, New Zealand, p 55
- Saukkola A, Melkas T, Riekkö K, Sirparanta S, Peuhkurinen J, Holopainen M, Hyypä J, Vastaranta M (2019) Predicting forest inventory attributes using airborne laser scanning, aerial imagery, and harvester data. *Remote Sens* 11(7):797
- Sellén D (2016) Big Data analytics for the forest industry: a proof-of-concept built on cloud technologies. M.Sc. thesis, Mid Sweden University, Östersund, Sweden, p 80
- Siipilehto J, Lindeman H, Vastaranta M, Yu X, Uusitalo J (2016) Reliability of the predicted stand structure for clear-cut stands using optional methods: airborne laser scanning-based methods, smartphone-based forest inventory application Trestima and pre-harvest measurement tool EMO. *Silva Fenn* 50(3), 1568. <https://doi.org/10.14214/sf.1568>
- Söderberg J (2015) A method for using harvester data in airborne laser prediction of forest variables in mature coniferous stands. M.Sc. thesis, Swedish University of Agricultural Science, Uppsala, Sweden, p 31
- Stendahl J, Dahlin B (2002) Possibilities for harvester-based forest inventory in thinnings. *Scand J For Res* 17(6):548–555
- Stone M (1974) Cross-validatory choice and assessment of statistical predictions. *J R Stat Soc B* 36(2):111–133
- Strandgard M, Walsh D, Acuna M (2013) Estimating harvester productivity in *Pinus radiata* plantations using StanForD stem files. *Scand J For Res* 28(1):73–80
- Swayne D, Cook D, Bujar A, Lang D, Wickham H, Lawrence M (2006) *GGobi Manual*. <http://www.ggobi.org/docs/manual.pdf>. Accessed 5 Jan 2018
- Uusitalo J (2010) Introduction to forest operations and technology. JVP Forest Systems Oy, Hämeenlinna, p 287
- Uusitalo J (2017) Big data is transforming forestry. www.luke.fi/en/big-data-transforming-forestry. Accessed 1 Mar 2018
- Varjo J (1995) Latvan hukkaosan pituusmallit männylle, kuuselle ja koivulle metsurimittausta varten. In: Verkasalo E (ed) *Puutavaran*

- mittauksen kehittämistutkimuksia 1989–93, Finnish Forest Research Institute, Research Papers 558, pp 21–23 (in Finnish). <https://jukuri.luke.fi/handle/10024/521187>
- Vesa L, Palander T (2010) Modeling stump biomass of stands using harvester measurements for adaptive energy wood procurement systems. *Energy* 35(9):3717–3721
- Vonesh EF (2012) Generalized linear and nonlinear models for correlated data: theory and applications using SAS. SAS Institute, Cary
- Wackerly DD, Mendenhall W, Scheaffer RL (1996) Mathematical statistics with applications. Duxbury Press, Belmont, p 798
- Williams C, Ackerman P (2016) Cost-productivity analysis of South African pine sawtimber mechanised cut-to-length harvesting. *South For J For Sci* 78(4):267–274
- Zhang L, Bi H, Cheng P, Davie CJ (2003) Modelling spatial variations in tree diameter—height relationships. *For Ecol Manag* 189:317–329
- Zhang YH, Li Y, Bi H (2015) Converting diameter measurements of *Pinus radiata* taken at different breast heights. *Aust For* 78(1):1–5

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.