Kershaw et al. Forest Ecosystems (2017) 4:15
DOI 10.1186/s40663-017-0102-2

Forest Ecosystems

RESEARCH Open Access

An imputation/copula-based stochastic individual tree growth model for mixed species Acadian forests: a case study using the Nova Scotia permanent sample plot network

John A. Kershaw Jr[1*], Aaron R. Weiskittel[2], Michael B. Lavigne[3] and Elizabeth McGarrigle[4]

Abstract

Background: A novel approach to modelling individual tree growth dynamics is proposed. The approach combines multiple imputation and copula sampling to produce a stochastic individual tree growth and yield projection system.

Methods: The Nova Scotia, Canada permanent sample plot network is used as a case study to develop and test the modelling approach. Predictions from this model are compared to predictions from the Acadian variant of the Forest Vegetation Simulator, a widely used statistical individual tree growth and yield model.

Results: Diameter and height growth rates were predicted with error rates consistent with those produced using statistical models. Mortality and ingrowth error rates were higher than those observed for diameter and height, but also were within the bounds produced by traditional approaches for predicting these rates. Ingrowth species composition was very poorly predicted. The model was capable of reproducing a wide range of stand dynamic trajectories and in some cases reproduced trajectories that the statistical model was incapable of reproducing.

Conclusions: The model has potential to be used as a benchmarking tool for evaluating statistical and process models and may provide a mechanism to separate signal from noise and improve our ability to analyze and learn from large regional datasets that often have underlying flaws in sample design.

Keywords: Nearest neighbor imputation, Copula sampling, Individual tree growth model, Mortality, Ingrowth, Mixed species stand development, Acadian forests, Nova Scotia

Background

Forest management planning requires long-term forecasts of resource flows (Baskerville 1986). Whether the forest management plan is focused on timber flows (Clutter et al. 1983; Baskerville 1986), carbon offsets (Birdsey 2006; MacLean et al. 2014), or other ecosystem services (Pretzsch et al. 2008), forest growth and yield models play a key role in the forest management process (Clutter et al. 1983; Weiskittel et al. 2011a). In addition to providing forecasts of long-term resource flows, growth and yield models are used to design silviculture interventions (Maguire et al. 1991; Barrett and Davis 1994), and to assess changes in factors such as fire risk (Keyes and O'Hara 2002) or wildlife habitat (MacLean et al. 2010). Because growth and yield models have such diverse applications, numerous types of models evolved over the years (Leary 1988; Weiskittel et al. 2011a). For much of the last 30 years, growth and yield research has concentrated on developing individual tree growth and yield models (Adlard 1995; Fox et al. 2007; Weiskittel et al. 2011a).

All individual tree growth models function similarly. A list of trees, usually representing some small areal extent (i.e., a "plot") are inputted into the model and the model predicts how those trees change over some time horizon. Changes in tree diameter at breast height (DBH) and total height (HT),

* Correspondence: kershaw@unb.ca
[1]University of New Brunswick, Fredericton, NB, Canada
Full list of author information is available at the end of the article

as well as survival rates, are typically predicted (Weiskittel et al. 2011a). Some models may predict changes in other tree attributes such as height to crown base, crown ratio, crown width, and so on. Some models also predict new trees entering the list (i.e., ingrowth), but these equations are often highly imprecise due to the stochastic nature of regeneration processes (e.g., Li et al. 2011). Models differ primarily in terms of (1) what tree-level and stand-level information is required, and (2) how the underlying growth, survival, and ingrowth functions were derived (Flewelling et al. 1986; Dixon et al. 1991; Adlard 1995; Weiskittel et al. 2011a).

Dixon et al. (1991) suggest three components that are essential for any modelling system: 1) an understanding of the process or relationships being modelled; 2) mathematical, statistical, and computational techniques and equipment capable of handling the problem; and 3) experimental or survey data. Many of the individual tree models used in forest management planning are statistically derived (Weiskittel et al. 2011a), Most of the more widely used statistical models are derived from large regional growth and yield permanent plot networks (e.g., Flewelling et al. 1986; Dixon 2002; Woods and Robinson 2008; Weiskittel et al. 2013). Although these extensive datasets allow for developing robust equations that potentially extrapolate well, the parametric model forms and covariates generally used are relatively simplistic and may not fully leverage the available data. As a result, forest modellers are often accused of having the above list of components in reverse order (Adlard 1995).

More sophisticated statistical methods for deriving growth and yield models (e.g., Bayesian approaches or Big Data Analytic approaches) have seen more limited applications to date, but may result in more robust model behaviour as well as allow better insights into the underlying processes driving tree and stand development. However, no matter how complex the underlying equations, statistical models generally predict the average growth conditions given a realization of a set of independent variables. While their deterministic output has been particularly useful for forest management planners (Flewelling et al. 1986; Weiskittel et al. 2011a), this deterministic nature also has been one of the main criticisms of statistically derived models (Dixon et al. 1991; Vanclay 1991; Johnsen et al. 2001). Assessment of uncertainty and the propagation of prediction errors through the various equations over time is a complicated undertaking (Vanclay 1991; Fox et al. 2007). This complexity of adequately quantifying uncertainty is the result of the number of equations involved, the various sources of uncertainty (e.g. measurement vs. model error), and the questionable assumption of error independence among the equations. However, the uncertainty in growth and yield models can be quite high. For example, Weiskittel et al. (2016) found that standard error for total stand volume ranged from 4 to 6% after forty years of projection for relatively homogeneous Douglas-fir (*Pseudotsuga menziesii* (Mirb.) Franco var. *menziesii*) plantations in the Pacific Northwest using an individual tree growth and yield model. Likely, this projection uncertainty is even higher for more complicated stand structures, forest types, and management regimes.

Vanclay (1991) reviewed approaches used to incorporate stochasticity into growth projection systems, and considered many approaches to be either naïve attempts or ad hoc swindles. Vanclay (1991) proposed that change be expressed using probabilistic functions. Deterministic predictions can be obtained by using these functions to represent proportions of populations, while stochastic predictions can be obtained by using these functions to represent probabilities for individuals. Vanclay's (1991) approach produced a system of compatible deterministic-stochastic predictions. Fox et al. (2007) reviewed the use of hierarchical mixed effects models to incorporate structured spatial and temporal stochasticity directly into growth equations. They argue that the very nature of the data used in growth and yield modelling (i.e., trees measured within plots, repeatedly over time) necessitates the need to incorporate structured stochasticity into models and discuss several ways this may be accomplished.

In an alternative approach to incorporating stochasticity into growth models, McGarrigle et al. (2013) developed a stand-level model based on what they call informed random walks. In their approach, a large regional growth and yield database, assembled for the forests of northeastern North America (Weiskittel et al. 2013), was compiled into a stand-level reference database, and the resulting stand-level model alternately utilizes imputation-based selection of nearest neighbors from the reference database and copula sampling to grow stands through time. The structure of the model was fairly simple, but was able to produce a wide range of stand trajectories and behaviors across a range of stand conditions defined within Reineke's (1933) stand density space (McGarrigle et al. 2013). This model provided some unique insights into the variability in prediction uncertainty across the stand trajectory space. McGarrigle (2013) further demonstrated how this model could be useful for model evaluation and benchmarking.

In this study, we propose an individual tree imputation/copula model similar to the stand-level model developed by McGarrigle et al. (2013). The model structure and development is presented, and the model behavior is compared with the Acadian variant of FVS (Weiskittel et al. 2014) using a set of stand ideotypes and the extensive Nova Scotia, Canada permanent sample plot network. Potential applications of this model and future extensions are discussed.

Methods

Study area

The data used in this study is a subset of the Acadian Forest growth and yield dataset (Weiskittel et al. 2013).

In this study, we only used data from the Nova Scotia Permanent Sample Plot (NSPSP) system (Fig. 1) as it is comprehensive in terms of species and stand structure, has used a consistent measurement protocol, and has a long history of establishment. Nova Scotia forests contain a diverse group of species growing on a wide range sites resulting from the varied soils, climate, and elevation (NSDNR 2008). Nova Scotia is in the Acadian Forest region (Rowe 1972; Loo and Ives 2003), a transitional forest region between the northern hardwoods to the south and boreal forests to the north.

Data compilation

The NSPSP contains 2897 plots with 2 to 8 remeasurement periods spanning a time range of 5 to 40 years. Plots were 0.04 ha circular plots (11.28 m radius). Plot centers are permanently monumented and global position system (GPS) coordinates archived to facilitate relocation and remeasurement. At plot establishment, all trees ≥10.0 cm in diameter at breast height (DBH; BH = 1.3 m) and within the 11.28 m radius where tagged with a unique tree number and measured, as described below. At subsequent measurements, dead or cut trees were noted and any new trees ≥10.0 cm DBH (i.e., ingrowth) were tagged with a unique tree number and measured. Species was identified for all tagged trees, DBH was measured to the nearest 0.1 cm, and total height (HT) was measured to the nearest 0.1 m. The plot data were compiled into two reference databases to be used in the individual tree growth model developed in this study.

The first reference data base was the individual tree growth and survival database (TREE). Each tree on each plot in each measurement interval was a potential record for the TREE database. Survivorship status was noted (1 if survived, 0 if died) and any cut trees were removed from the TREE database. DBH and HT at the beginning and end of each measurement interval were used to calculate individual tree growth. Growth was annualized by dividing by the number of years in the measurement interval. Several one-sided and two-sided competition factors were calculated and the importance of these variables for growth was evaluated using a generalized boosted regression model (Kuhn 2008) which helped identify a subset of competition factors to be used during the imputation step in the growth model. All competition measures were calculated at the beginning of the measurement interval.

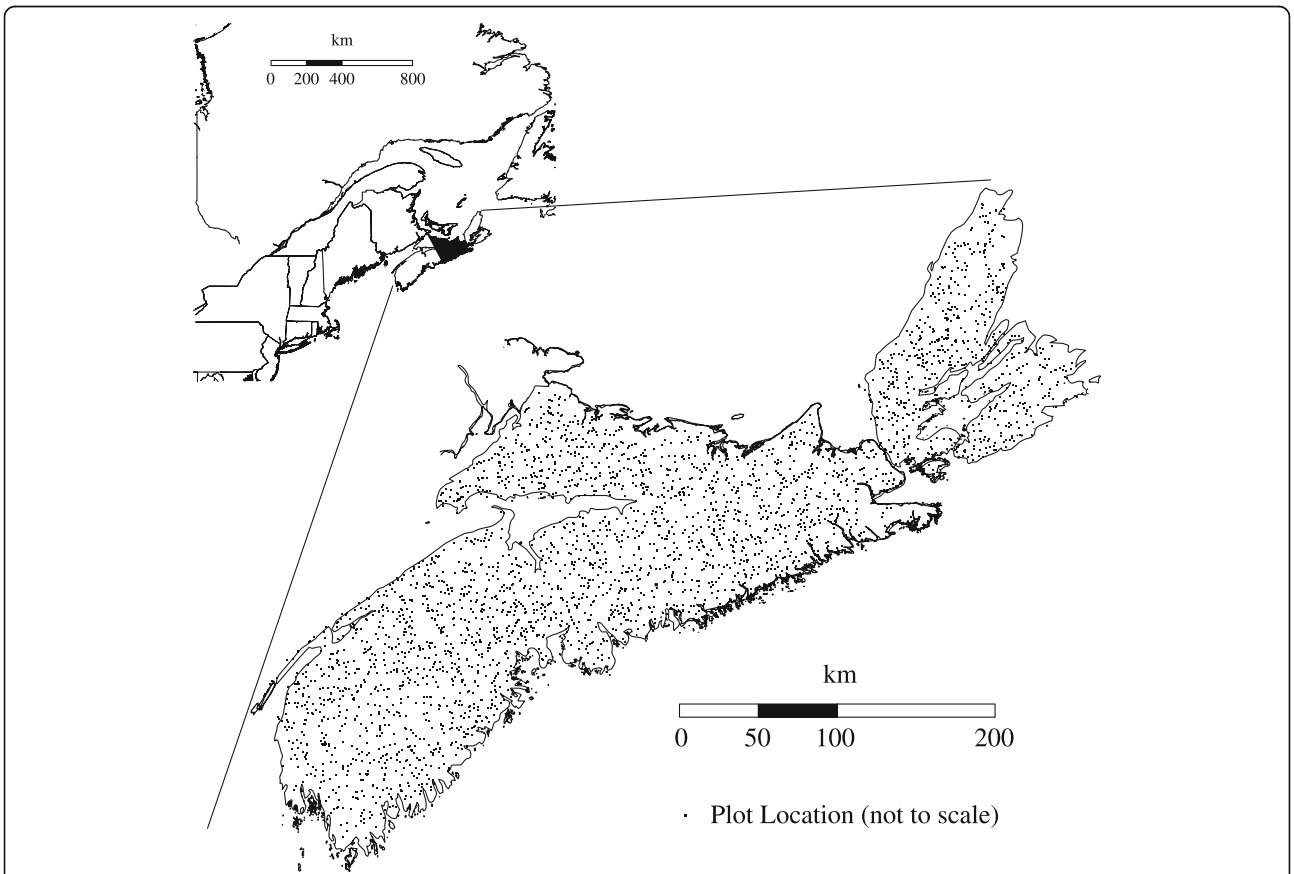

Fig. 1 Map showing the location of Nova Scotia, Canada in northeastern North America and a detailed map of Nova Scotia showing permanent sample plot locations

The final set of competition measures used in this study were: basal area per ha:

$$\mathrm{BAPHA} = 0.00007854 \sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\mathrm{DBH}_i^2 \cdot \mathrm{TF}_i\right) \quad (1)$$

where, BAPHA = basal area per ha (m^2 ha^{-1}), DBH = diameter at breast height (cm), and TF = tree per ha expansion factor (Kershaw et al. 2016, Chap. 9); basal area of trees of larger DBH than subject tree:

$$\mathrm{BALD} = 0.00007854 \sum_{i=1}^{n} \left(\mathrm{DBH}_i^2 \cdot \mathrm{TF}_i \cdot I(\mathrm{DBH}_i > \mathrm{DBH}_s)\right) \quad (2)$$

where, BALD = basal area of trees of larger DBH than subject tree (m^2 ha^{-1}), I is an identity function (I = 1 if $\mathrm{DBH_i} > \mathrm{DBH_s}$, otherwise I = 0), and s denotes the subject tree; stand density index (McCarter and Long 1986) of trees taller than the subject tree:

$$\mathrm{SDITT} = \left(\sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\mathrm{TF_i} \cdot \mathrm{I}[\mathrm{HT_i} > \mathrm{HT_s}]\right)\right) \left(\frac{\overline{\mathrm{DBH}_Q}}{25}\right)^{1.6} \quad (3)$$

where SDITT = stand density index of trees taller than the subject tree (stems/ha), $\overline{DBH_Q}$ = quadratic mean diameter of tree taller than the subject tree:

$$\overline{\mathrm{DBH}_Q} = \sqrt{\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\mathrm{DBH}_i^2 \cdot \mathrm{TF}_i \cdot \mathrm{I}[\mathrm{HT}_i > \mathrm{HT}_s]\right)}{\sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\mathrm{TF}_i \cdot \mathrm{I}[\mathrm{HT}_i > \mathrm{HT}_s]\right)}} \quad (4)$$

1.6 is the power coefficient of Reineke's (1933) stand density index, and 25 is the reference DBH; relative DBH:

$$\mathrm{relDBH} = \frac{\mathrm{DBH}_s}{\max(\mathrm{DBH}_i)} \quad (5)$$

where relDBH = relative diameter; and a climate-based site index value (CSI, Weiskittel et al. 2011b). The final TREE database was composed of Species, DBH, HT, BAPHA, BALD, SDITT, relDBH, CSI, Survivor Status, annualized DBH growth (AGDBH), and annualized HT growth (AGHT). There were 4,235,533 records in TREE (See Additional file 1: Table S1 for example reference and target tree records).

The second reference database was for Ingrowth (INGROW). Ingrowth, defined as trees growing across the minimum DBH measurement threshold of 10.0 cm, was summarized at the plot-level for each measurement interval and the ingrowth trees were extracted. As for the individual tree growth reference database, several stand-level variables were compiled and boosted regression was used to screen variables important for predicting number of ingrowth trees per ha per year (ANINGR). Basal area per ha (BAPHA), net basal area growth per ha per year (AGBA), basal area mortality per ha per year (MortBA), and BA for balsam fir (*Abies balsamea* (L.) Mill.), black spruce (*Picea mariana* (Mill.) B.S.P.), red spruce (*Picea rubens* (Sarg.)), white pine (*Pinus strobus* L.), red maple (*Acer rubrum* L.), sugar maple (*Acer saccharum* Marsh.), aspen (*Populus* spp.), white birch (*Betula papyrifera* Marsh.), and yellow birch (*Betula alleghaniensis* Britt.) were identified as important predictors by the boosted regressions. In addition to these basal area measurements, stand density index (SDI), quadratic mean diameter (Dq), maximum DBH, and maximum HT were selected as important stand-level variables.

The INGROW reference database was composed of two data tables. The first table was INGROW.PROB, and contained the annualized per ha ingrowth rates and the plot-level variables described above. The number of ingrowth trees per ha was calculated by counting the number of ingrowth trees per plot and multiplying by TF, and subsequently was annualized by dividing by the number of years in the measurement interval. INGROW.PROB had a record for each plot over each measurement interval and contained 14,143 data records. The second table, INGROW.LIST, contained the individual ingrowth tree records, coupled with their associated plot-level variables described above. INGROW.LIST contained 60,611 records (See Additional file 1: Tables S2 and S3 for example reference and target ingrowth probability and tree list records).

The two reference databases were used in the imputation steps in the model described below and constitute the data drivers behind the imputation/copula individual tree model developed in this study.

Model development

The model developed here is an individual tree extension of the stand-level informed random walk model developed by McGarrigle et al. (2013). The imputation/copula individual tree model (I/C model) was developed in R (R Development Core Team 2016) using parallel processing extensions on a Linux operating system. Inputs into the model include a compiled tree list, number of replicates to run, number of years to forecast, and number of nearest neighbors to select in the imputation steps. The model uses multiple imputation via the yaiImpute package (Crookston and Finley 2008) to select nearest neighbors to the subject tree or subject plot. Empirical distributions based on the "k" nearest neighbors are built and copula sampling used to estimate average growth and ingrowth rates. The general flow of the algorithm is shown in pseudocode in Fig. 2. Details of each component of the algorithm are described in the following sections.

```
INPUT tree list, TREES, INGROW.PROB, INGROW.LIST
INPUT number of replicates (REP), number of years (YEARS), number of nearest neighbors (Knn)
DO PARALLEL 1:REPS
    FOR 1:YEARS
        IMPUTE Knn nearest neighbors for each tree in tree list from reference database TREES
        FOR 1:number of trees
// Survival Calculations
            CALCULATE annual survival probability from Knn nearest neighbors
            DETERMINE tree survival status
// Tree Growth Calculations
            IF tree survives
                ESTIMATE empirical kernel density of annualized dbh growth
                ESTIMATE empirical kernel density of annualized ht growth
                ESTIMATE correlation between dbh growth and ht growth
                SAMPLE copula(kernel(dbh growth),kernel(ht growth),cor)
                CALCULATE average dbh and ht growth
            END IF
        NEXT tree
// Ingrowth Calculations
        IMPUTE Knn nearest neighbors from INGROW.PROB
        SAMPLE distance weighted neighbors to estimate Poisson λ
        SAMPLE Poisson(λ) to estimate number of ingrowth trees (NING)
        IMPUTE NING nearest neighbors from INGROW.LIST
        ADD ingrowth trees to tree list
        UPDATE tree.list
        SAVE tree.list
    NEXT year
END DO PARALLEL
RETURN tree list
EXIT
```

Fig. 2 Pseudocode for the general individual tree imputation/copula model algorithm

Individual tree survival and growth

Individual tree survival and growth are estimated from the TREE reference database described above. For each year in the simulation cycle, a specified number of nearest neighbors (Knn) are selected from the TREE reference database for each tree in the tree list.

Tree survival probability is estimated from the Knn nearest neighbors. The proportion of trees surviving, P_S, is calculated by counting the number of nearest neighbors that survive the measurement interval divided by Knn. This proportion is converted into and annualized probability using:

$$\mathrm{p(Survival)} = \mathrm{P}_s^{1/Y} \tag{6}$$

where p(Survival) = the annualized tree survival probability; P_S = proportion of Knn neighbors surviving measurement intervals; Y = number of years in measurement intervals. For the NSPSP, Y = 5 years for all measurement intervals. Tree survival is determined by generating a uniform (U[0,1]) random number. If U[0,1] > p(Survival), then the tree dies, otherwise the tree survives and growth is determined.

Tree growth is estimated using random sampling from a Normal copula (Genest and MacKay 1986; Nelsen 2006) as implemented by McGarrigle et al. (2013). Copulas are constructed for each tree at each time step using the annualized DBH and height growth observed on the surviving Knn nearest neighbors. The copula marginals are modelled as empirical kernel density estimates (Scott 1992) and the correlation between annualized DBH and height growth is estimated using Pearson's correlation coefficient (Zar 2009). Following McGarrigle et al.'s (2013) algorithm, estimates of annualized DBH and height growth are obtained as follows:

1) Two sets of n standard Normal variates are randomly generated ($N_d(0,1)$, $N_h(0,1)$) and column bound to form matrix [**N**]
2) A partial correlation matrix is formed using Pearson's correlation coefficient estimated from the annualized DBH and height growth values:

$$[\mathrm{P}] = \begin{bmatrix} 1 & \rho(\Delta\mathrm{dbh}, \Delta\mathrm{ht}) \\ \rho(\Delta\mathrm{dbh}, \Delta\mathrm{ht}) & 1 \end{bmatrix} \tag{7}$$

3) The columns of **[N]** are correlated by matrix multiplying [**N**] by Choleski's (Andersen et al. 1999) upper decomposition of [**P**]:

$$[\mathrm{C}] = [\mathrm{N}]\%*\%chol([\mathrm{P}]) \tag{8}$$

4) The Normal marginal distributions are converted into U[0,1] distributions by applying the inverse Normal distribution to the columns of [**C**]:

$$[\mathrm{U}] = \left[\mathrm{N}^{-1}(\mathrm{C}_1), \mathrm{N}^{-1}(\mathrm{C}_2)\right] \tag{9}$$

[U] is the Normal copula (Genest and MacKay 1986).

5) The annualized growth samples are then obtained by applying a quantile function, derived from the kernel density estimates, based on the surviving nearest neighbors, to the columns of [**U**]:

$$[\Delta] = [\text{qkernel}_{\text{dbh}}(U_1), \text{qkernel}_{\text{ht}}(U_2)] \quad (10)$$

DBH growth is estimated as the average of the first column of [**Δ**] and HT growth is estimated as the average of the second column of [**Δ**]. New estimates of DBH and HT are obtained by adding the growth estimates to the current values of DBH and HT. The average of 10 random samples from each copula was used in this study; however, the model structure allows users to change this value.

Ingrowth

As described above, the ingrowth reference database is composed of two reference tables: INGROW.PROB and INGROW.LIST. The first table, INGROW.PROB, is used to determine if ingrowth occurs and how many ingrowth trees to add. The Knn nearest plots to the plot being grown are selected from INGROW.PROB using yaImpute package in R. One nearest neighbor plot is randomly selected from the Knn plots with probability inversely proportional to neighbor distance. The annualized number of ingrowth trees on the selected plot record is used as the density parameter, λ, and a random Poisson variate (Inn) is generated and used as the number of ingrowth records to add to the plot at the current year in the projection cycle.

Individual ingrowth tree records are then selected from the INGROW.LIST reference table. Similar to the process describe above, 2*Inn nearest tree records are selected from INGROW.LIST using the yaImpute package in R. Inn tree records are randomly selected with probability inversely proportion to nearest neighbor distance from current plot. Selection is carried out with replacement, thus the same ingrowth record may be added to the tree list more than once for any given year in the projection cycle.

Once all ingrowth trees are added, the plot is updated, calculating all of the required one- and two-side competition measures and stand composition measures required for the next year in the projection cycle. The tree list at the end of each cycle is stored separately so that trends over time can be analyzed after all cycles and replicates are completed. Replicates were completed in a parallel processing loop significantly reducing the computation time required to carry out all of the simulations.

Stand Ideotypes

Standard input tree lists, that we term "ideotypes", were created. The idea behind creating ideotypes was to have standardized stand structures that only varied by Dq, mean HT, species composition, and density. We used 100 tree records as the basis of our ideotypes. These 100 trees are replicated the required number of times to create a tree list for any density (trees/ha), and a small amount of random noise is added to each DBH and HT, to make each tree record unique without substantially changing the distribution. DBH and HT of the 100 base records were simulated using a Normal copula with a correlation coefficient of 0.60. The uniform marginals are then removed by applying "standardized" Weibull quantile functions. For the DBH distribution we used a three parameter Weibull distribution with a location parameter of 0.5, scale parameter of 1.0 and a shape parameter of 2.0 (right skewed. Negative exponential-like distribution). For HT, we used a two parameter Weibull distribution with a scale parameter of 1 and a shape parameter of 3.6 (symmetric distribution). The standardized distributions for DBH and HT are shown in Fig. 3. Species are then assigned to the resulting standardized DBH-HT pairs using a finite mixture distribution (Zhang et al. 2001; Liu et al. 2002) based on

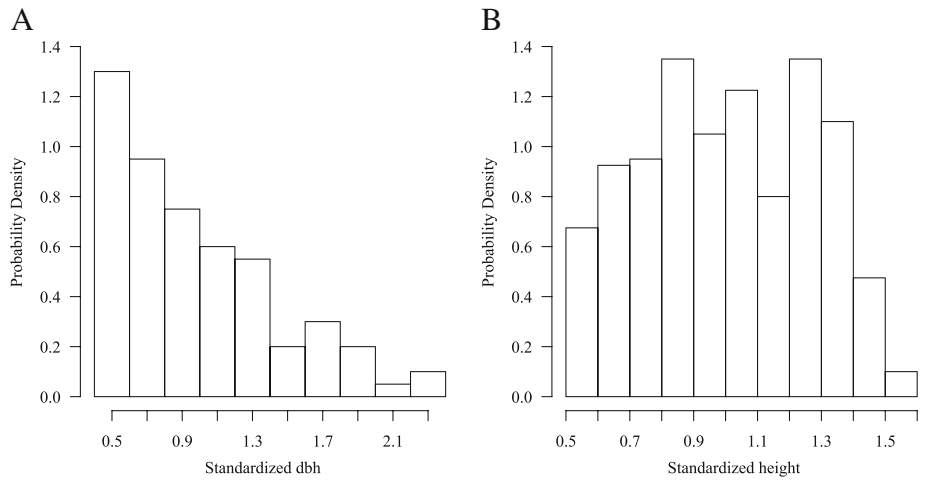

Fig. 3 Standardized DBH (**a**) and height (**b**) distributions for the tree list ideotypes

relative DBHs of the species mixtures. Four different stand types, based on species composition, were created (Table 1). Densities and quadratic mean diameters were applied to the standard 100-tree tree list to produce a set of stand structures that covered the range of conditions observed across Reineke's (1933.) stand density index space (Fig. 4).

Model calibration and comparison

To determine optimum number of nearest neighbors for the imputation step (bandwidth, McGarrigle 2013), 100 of the oldest intact plots were selected. Plot selection criteria included: no cutting during the entire plot measurement period; no catastrophic mortality (> 30% BAPHA in any one measurement cycle); and minimum BAPHA ≥10 m^2 ha^{-1}. The initial plot measurements were used as the input tree lists in the I/C model and each plot was grown for the same period over which observed measurements were available. Five bandwidths (K) were used for nearest neighbor selection in the imputation step: 25, 30, 40, 50, and 60 nearest neighbors. Sample size for the copula step was held at 10 samples (McGarrigle et al. 2013). For each bandwidth, percent error was calculated for mortality rate, DBH growth, HT growth, ingrowth rate, and ingrowth species composition. Species composition errors were measured using:

$$\%\mathrm{Error}_{\mathrm{Species}} = 100\frac{\sqrt{\sum_{i=1}^{s}\left(O_i - P_i\right)^2}}{\sum_{i=1}^{s} O_i} \qquad (11)$$

where, s = number of different species; O_i = number of observed ingrowth trees of species I; P_i = number of predicted ingrowth trees of species i. The number of species, s, is the unique set of species in both the observed and predicted ingrowth trees. Optimal bandwidth was determined to be the bandwidth that minimized these percent errors.

Table 1 Species composition (per cent stems/ha) of the four ideotype stand types

Species Name		Stand Type[a]			
Common	Scientific	INHW	IHMW	BS	TSW
balsam fir	*Abies balsamea*	11	27	1	6
black spruce	*Picea mariana*			91	
red spruce	*Picea rubens*	12	21		71
eastern white pine	*Pinus strobus*		1	1	3
northern white-cedar	*Thuja occidentalis*		1		2
eastern hemlock	*Tsuga canadensis*		1		3
red maple	*Acer rubrum*	41	27	4	8
yellow birch	*Betula alleghaniensis*		1		
white birch	*Betula papyrifera*	30	13	3	5
white ash	*Fraxinus americana*	1	2		
bigtooth aspen	*Populus grandidentata*	2	1		1
quaking aspen	*Populus tremuloides*	3	5		1

[a]*INHW* intolerant hardwood, *IHMW* intolerant hardwood, mixedwood, *BS* black spruce, *TSW* tolerant softwood

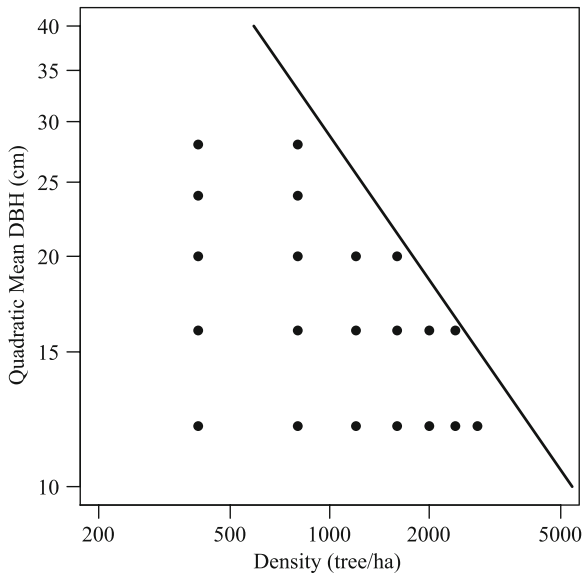

Fig. 4 Location of stand ideotype combinations in Reineke's (1933) stand density space

For model performance comparisons, each stand ideotype combination (stand type × Dq × density) was grown in the I/C model for 25 years with 25 replicates. These tree lists were also projected using the Acadian variant of FVS (Weiskittel et al. 2013). Comparisons of model performances were based on visual assessments of stand-level trajectories. Bandwidths of 25 and 40 (the estimated optimal bandwidths) were used for selection of nearest neighbors in the imputation step.

Results

Bandwidth selection

Average percent mortality error ranged from about 24% to 32% across the bandwidths used in the study (Fig. 5a). Both the range in mortality error and average mortality error were minimized with K = 40 nearest neighbors. Average DBH growth error ranged from about 12% to 18%, and was relatively unaffected by bandwidth until it exceeded 40 (Fig. 5b). DBH growth error increased steadily and sharply for K = 50 and 60. The minimum average DBH growth error was observed at K = 30; however, the minimum range in error was observed at K = 50. Average HT growth errors ranged from 12% to 17% (Fig. 5c), and was unaffected by bandwidth; however, the range in error generally increased with increasing bandwidth.

Similar to the average HT growth errors, ingrowth rate errors were not greatly influenced by bandwidth (Fig. 6a); however, ingrowth errors were almost double the error rates observed for DBH and HT growth, yet were similar to the error rates observed in mortality. Average ingrowth

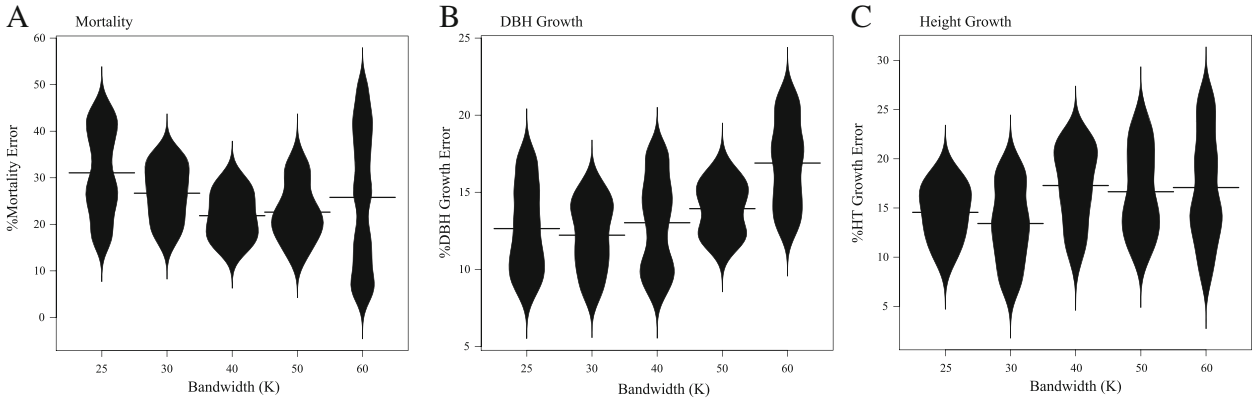

Fig. 5 Bean plots of percent error by bandwidth (k) for: **a**) mortality; **b**) DBH growth; and **c**) height growth

errors ranged from about 28% to 31% and generally increased slightly across the range of bandwidths. Errors in ingrowth species composition ranged from about 20% to almost 50% across the range of bandwidths (Fig. 6b). Both the average error and range in error increased rapidly with increasing bandwidth.

Copula sample intensity

As long as sample size was >1 and not excessive (generally <30), the number of samples selected from the copula, and averaged to produce the DBH and HT growth estimates, had very little influence on model behavior and stand trajectories (not shown). A single sample produced extremely variable individual tree and stand trajectories. As sample size increased, the variability in individual tree and stand trajectories decreased. For all analyses shown in this paper, we used a sample size of 10 as a compromise between large variations in tree and stand development and smoothing out the variability completely. Sample size had much greater effect on ideotypes at the lowest densities and greatest Dq; however, trends were similar across the entire range of density and Dq.

Model comparisons

Figure 7 compares observed plot trajectories for five plots from the NSPSP database with predictions from FVS-Acadian and the I/C Model developed here. Plots were chosen based on observation longevity and to highlight interesting differences in behavior between the FVS model and the I/C model. For Plot 700, both models predict trajectories that are very close to the observed plot trajectory. This stand was similar in structure and composition to the intolerant hardwood-mixedwood ideotype. For Plot 2 the two models have similar predictions; however, neither reflect the observed plot trajectory very well. Plot 2 is a softwood dominated type and during the first two measurement cycles experienced significant balsam fir mortality, most likely due to natural stand breakup or possibly spruce budworm defoliation. In Plots 44 and 479, the I/C model produces predictions that very closely agree with the observed trajectories, while the FVS predictions deviate substantially. Plot 44 is an intolerant hardwood type and the FVS model predicted substantial mortality in this stand while the I/C model did not predict much change in

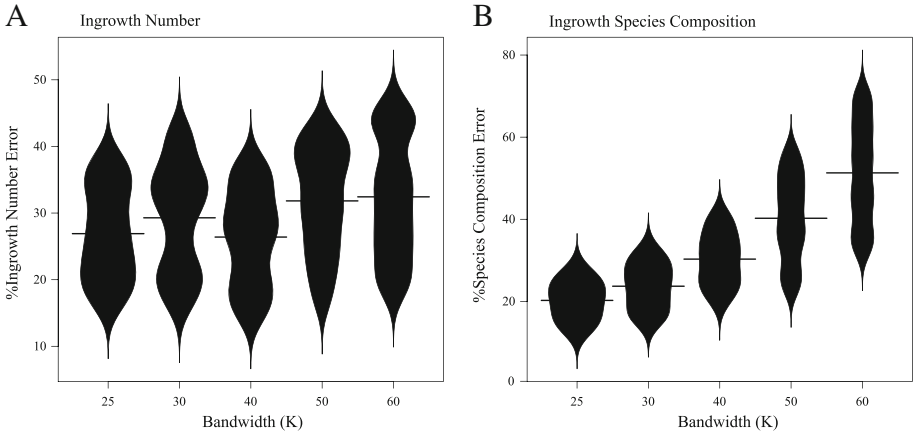

Fig. 6 Bean plots of percent error by bandwidth (k) for ingrowth: **a**) number; and **b**) species composition

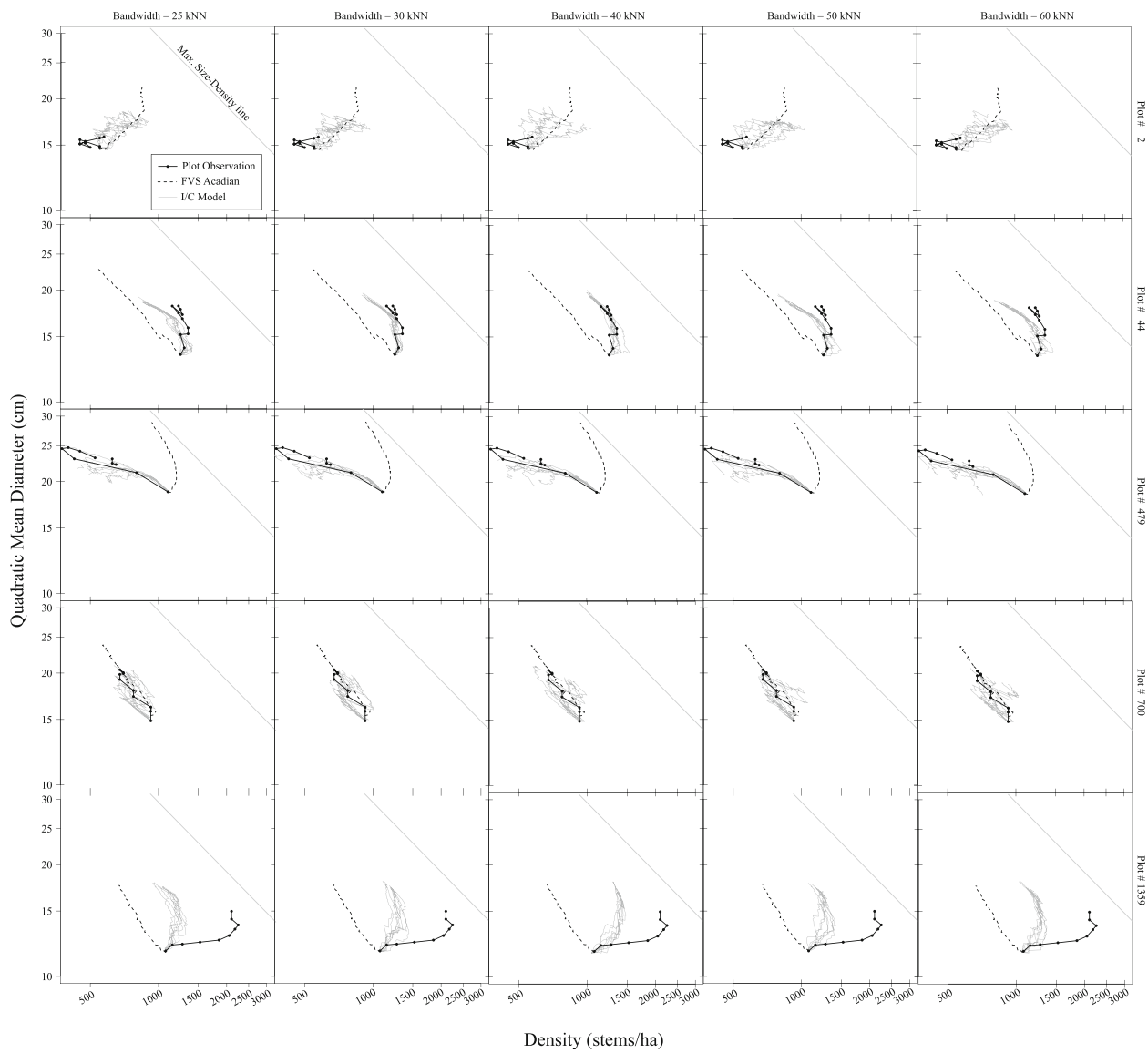

Fig. 7 Comparison of observed plot trajectories versus predictions from FVS-Acadian and the I/C Model for 5 selected NSPSP plots

either mortality or ingrowth, especially at optimal bandwidth (k = 40). Plot 479 is balsam fir dominated and experienced substantial mortality through the measurement period, again most likely due to spruce budworm defoliation. Unlike Plot 2, in this case, the I/C model was able to chose appropriate nearest neighbor trees to capture the mortality dynamics. Plot 479 was much more heavily dominated by balsam fir while red spruce was the dominant species in Plot 2. Plot 1359 represents a case where the observed trajectory, the FVS predicted trajectory, and the I/C model trajectory are all very different. Plot 1359 is another intolerant hardwood-mixedwood stand with a large amount of red maple. It appears that small-diameter red maple and red spruce mortalities are over-predicted in the FVS model. It also appears that the ingrowth of red spruce is generally under-observed (at least relative to the rates observed in Plot 1359) in the NSPSP database (i.e., this may reflect a lack of younger stands in the database).

Figure 8 shows the development trajectories from FVS-Acadian and the I/C model for the IHMW and BS stand ideotypes and k = 25 and 40. For both ideotypes, k = 25 resulted in trajectories from the I/C model that were substantially different from those obtained using FVS-Acadian. For the lower relative density stands, differences were less pronounced; however, as relative density approach the maximum size-density line, the trajectories from the I/C model had increasing deviations from the FVS-Acadian projections. Most of the source of deviation at the higher relative densities was due to errors in mortality predictions (Fig. 5a), resulting in trajectories that were much flatter

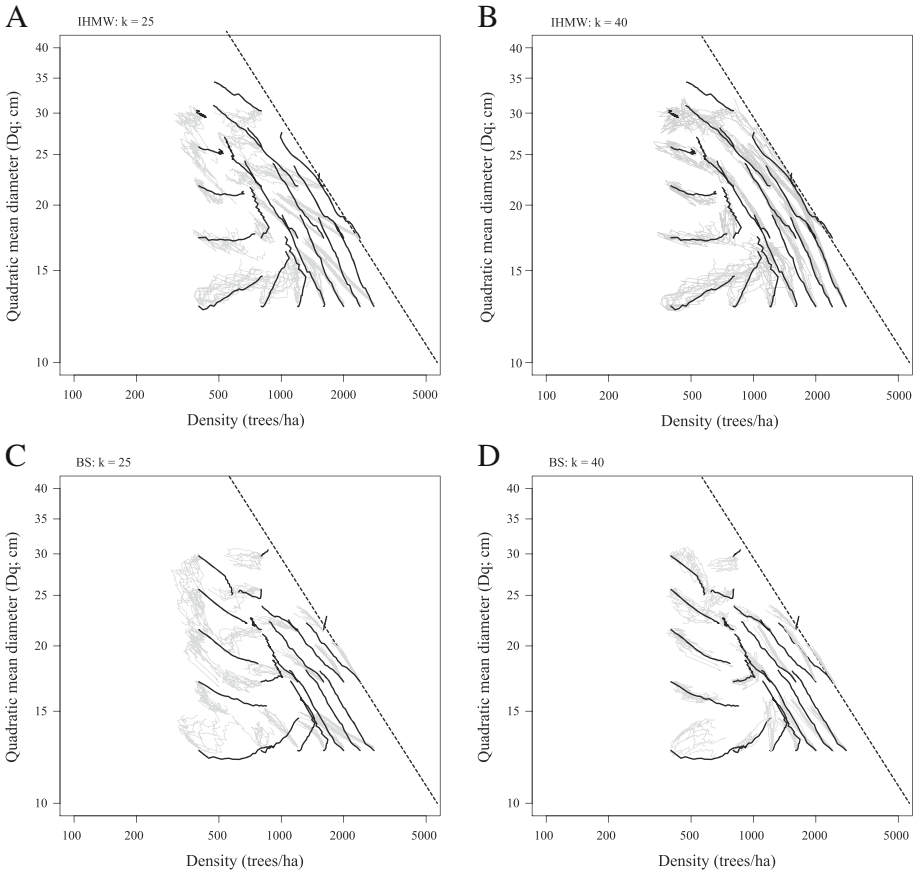

Fig. 8 Comparisons of stand development trajectories of stand ideotypes by bandwidth (k): **a**) intolerant hardwood, mixedwood ideotype with k = 25; **b**) intolerant hardwood, mixedwood ideotype with k = 40; **c**) black spruce ideotype with k = 25; **d**) black spruce ideotype with k = 40. Grey lines are replicate projections from the I/C model and the thick black line is the projection from the Acadian variant of FVS. The clusters of projections represent the ideotype starting points shown in Fig. 4

(i.e., higher mortality rates and lower changes in Dq) than what were associated with the FVS-Acadian projections. When k = 40, these flat trajectories were almost completely eliminated, and the two models more closely agreed with one another. The lower ideotype densities and the upper quadratic mean DBHs were notably exceptions. When Dq = 28 cm, all ideotypes had trajectories from the I/C model that often went in opposite directions from the trajectories obtained from FVS-Acadian.

Trajectories for the intolerant hardwood ideotypes (not shown) were very similar to those for the intolerant hardwood - mixedwood type, while the tolerant softwood ideotypes (not shown) were more similar to those for the BS type. However, the tolerant softwood type did not have as flat trajectories as those observed for the other types when k = 25, suggesting that mortality predictions for tolerant species might be less sensitive to bandwidth.

Discussion

The model developed here is purely data-driven with minimal assumptions that include: 1) that the set of nearest neighbors is adequate to predict the growth and survival of individual trees and the number and species composition of ingrowth trees; and 2) there is an optimal or preferred set of covariates from which to match nearest neighbors. Given these simple assumptions, the model developed here was capable of producing complex stand trajectories, generally consistent with those obtained from FVS-Acadian, a statistically derived model with several complex equations. The results obtained here are very similar to those obtained by McGarrigle et al. (2013) and McGarrigle (2013) for stand-level modelling dynamics.

While model performance was generally good, and in agreement with predictions from FVS, a widely accepted statistical model, there is still room for improvement. Errors associated with DBH and HT growth were similar to error levels obtained by other growth studies in the region using nonlinear mixed effects models (Russell et al. 2011; Russell 2012; Russell et al. 2014) and are consistent with error levels commonly observed in many growth and yield studies (e.g., Ek and Monserud 1979; Bragg 2003; Weiskittel et al. 2011a). DBH growth was not strongly influenced by

bandwidth except when bandwidth was greater than 40 nearest neighbors. HT growth was not influenced by any of the bandwidths tested in the study.

On the other hand, both mortality and ingrowth had substantial error rates (Figs. 5 and 6) with ingrowth species composition being poorly predicted (Fig. 6). Mortality and ingrowth were very sensitive to bandwidth choice (Figs. 5 and 6). Percent mortality error was reduced from about 30% to around 20% as bandwidth increased from 25 to 40 nearest neighbors and remained more-or-less constant for bandwidths greater than 40. The larger errors associated with small bandwidths most likely reflects the influence each tree has on mortality probability. When k = 25, each tree represents 4% of the population, and, since survival probability is calculated as the proportion of trees surviving the measurement interval, each dead tree represents a 4% reduction in survival. At k = 40, this reduction is reduced to 2.5%. Overall mortality in the database averaged about 2.8% annually. Larger bandwidths would reduce the impact individual trees have on mortality rates; however, selecting individuals that are increasingly more different than the target tree might increase error rates in other ways, which might be what is contributing to the small increase in mortality error when k = 60 (Fig. 5a).

Ingrowth rate was not influenced greatly by the bandwidths explored in this study (Fig. 6a); however, ingrowth species composition prediction was very sensitive to bandwidth, with errors increasing steadily with increasing bandwidth. The primary source of species composition errors was due to species not found on the plot being selected during the imputation and sampling process. Even though plot-level species composition was included in the reference variables used in the imputation step, this only represents local seed source availability (Ek et al. 1996; Archambault et al. 2009; Arseneault et al. 2011), and does not include any consideration of nearby seed sources. More explicit georeferencing of the plots and inclusion of a physical distance from target plot may reduce the incidence of stray species appearing as ingrowth. Additionally, the climate-based site index (CSI) was the only site factor included in the I/C model in this study. CSI, like all site index measures, is a composite measure of site productivity based on height modified by climate (Weiskittel et al. 2011b) and there are many other factors that influence regeneration species composition (Bakken and Cook 1998; Bataineh et al. 2013). Incorporation of local climatic variation and edaphic factors might improve ingrowth species composition.

Finally, the data used in this study came only from Nova Scotia, a relatively narrow portion of the range associated with many of the species. This relatively narrow range of data may explain some of the differences in predictions between the Acadian FVS model, developed from a much larger regional database, and the I/C model developed here. Additional testing of the model approach using the full Acadian Forest database is required, but will also require substantial increases in computational resources. The current model formulation was run on a minicomputer with 2 quadcore processors and 128Gb of RAM using the multiprocessing capability in R (R Development Core Team 2016). Twenty-five year projections with 25 replicates required 15–20 h to run depending upon bandwidth. The imputation step was the most time-consuming portion of the algorithm and increased with increasing bandwidth. Obviously, while this complexity and required computing time might limit application of the I/C model, potential gains in efficiencies are likely in the future.

Multiple imputation has found many applications in forestry and ecology over the past several years. Random forest imputation has become the standard analytical tool for LiDAR analyses (e.g., Hudak et al. 2008; Dassot et al. 2011; Gregoire et al. 2011; Hayashi et al. 2015). Imputation techniques are also widely used in forest inventory to estimate missing values (Eskelson et al. 2009b), and to spatially allocate forest inventory data over the landscape (McRoberts 2001; McRoberts and Tomppo 2007; Eskelson et al. 2009a; Falkowski et al. 2010). While imputation techniques have found many applications for static forest inventory estimates, growth and stand dynamics applications have seen much less attention. To our knowledge, this is only the second model to apply an imputation-copula approach to model forest stand dynamics (see McGarrigle et al. 2013 for the other example), and the first to apply it to individual tree dynamics.

Conclusions

Over the past several decades there has been much debate over the appropriateness of statistical versus process models (e.g., Dixon et al. 1991; Adlard 1995; Johnsen et al. 2001). While it is generally scientifically appropriate to engage in debate regarding model approach/philosophy, model structure, and model complexity, the role these large regional datasets play should not be downgraded or ignored. Even given many underlying experimental and sampling design flaws, these datasets contain a wealth of information on forest change that only now are we more fully understanding. Big data analytics, which has been utilized widely in other fields, may enable forest modellers to gain new insights into the underlying processes and better separate signal from noise.

Additional file

Additional file 1: Table S1. Example reference and target data for the survival and DBH and HT growth (TREE) reference database. **Table S2.** Example reference and target data for the Ingrow Probability (INGROW.PROB) reference dataset. **Table S3.** Example reference and target data records for ingrowth tree list (INGROW.LIST) reference database (XLSX 22 kb)

Authors' contributions

JAK developed the program and ran all of the simulations and analyses and coordinated the manscript writing, ARW contributed significantly to the writing of the introduction and discussion and ran all of the FVS projections, MBL provided significant input into the manuscript through many reviews and helped structure the model to be easily expanded to include climatic variables, HEM provided significant input throughout the manuscript preparation and conceived the original stand-level model from which this model was developed. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]University of New Brunswick, Fredericton, NB, Canada. [2]University of Maine, Orono, ME, USA. [3]Atlantic Forestry Centre, Natural Resources Canada, Fredericton, NB, Canada. [4]Nova Scotia Natural Resources, Forestry Division, Truro, NS B2N 0G9, UK.

Received: 25 June 2017 Accepted: 22 August 2017
Published online: 14 September 2017

References

Adlard PG (1995) Myth and reality in growth estimation. For Ecol Manag 71:171–176
Andersen E, Bai Z, Bischof C, et al (1999) LAPACK Users' Guide, Third Edition http://www.netlib.org/lapack/lug/index.html. Accessed 5 Apr 2017
Archambault L, Delisle C, Larocque GR (2009) Forest regeneration 50 years following partial cutting in mixedwood ecosystems of southern Quebec, Canada. For Ecol Manag 257:703–711
Arseneault JE, Saunders MR, Seymour RS, Wagner RG (2011) First decadal response to treatment in a disturbance-based silviculture experiment in Maine. For Ecol Manag 262:404–412
Bakken PN, Cook JE (1998) Regeneration potential of six habitat types common to north-central Wisconsin. North J Appl For 15:116–123
Barrett TM, Davis LS (1994) Using tree growth and yield simulators to create ecological yield tables for silvicultural prescriptions. West J Appl For 9:91–94
Baskerville G (1986) Understanding forest management. For Chron 62:339–347
Bataineh M, Kenefic LS, Weiskittel AR et al (2013) Influence of partial harvesting and site factors on the abundance and composition of natural regeneration in the Acadian Forest of Maine, USA. For Ecol Manag 306:96–106
Birdsey RA (2006) Carbon accounting rules and guidelines for the United States Forest sector. J Environ Qual 35:1518–1524
Bragg DC (2003) Optimal diameter growth equations for major tree species of the midsouth. South J Appl For 27:5–10
Clutter JL, Fortson JC, Pienaar LV, et al (1983) Timber management. A quantitative approach, first edn. John Wiley & Sons, New York
Crookston NL, Finley AO (2008) yalmpute: an R package for kNN imputation. J Stat Softw 23:1–16
Dassot M, Constant T, Fournier M (2011) The use of terrestrial LiDAR technology in forest science: application fields, benefits and challenges. Ann For Sci 68: 959–974
Dixon GE (2002) Essential FVS: a user's guide to the forest vegetation simulator. USDA, Forest Service, Forest Management Service Center, Ft. Collins, CO.
Dixon RK, Meldahl RS, Ruark GA, Warren WG (eds) (1991) Process modeling of forest growth responses to environmental stress. Timber Press, Portland, OR
Ek AR, Monserud RA (1979) Performance and comparison of stand growth models based on individual tree and diameter-class growth. Can J For Res 9:231–244
Ek AR, Robinson AP, Radtke PJ, Walters DK (1996) Regeneration imputation models and analysis for forests in Minnesota. University of Minnesota, Minnesota Agricultural experiment Station
Eskelson BNI, Temesgen H, Barrett TM (2009a) Estimating current forest attributes from paneled inventory data using plot-level imputation: a study from the Pacific northwest. For Sci 55:64–71
Eskelson BNI, Temesgen H, LeMay VM, et al (2009b) The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. Scand J For Res 24:235–246
Falkowski MJ, Hudak AT, Crookston NL et al (2010) Landscape-scale parameterization of a tree-level forest growth model: a k-nearest neighbor imputation approach incorporating LiDAR data. Can J For Res 40:184–189
Flewelling JW, Curtis RO, Hyink DM (1986) Forest growth models in the 1990s: functions, sources, needs. In: Oliver CD, Harley DM, Johnson JA (eds) Douglas-fir stand management for the future. University of Washington, Seattle, WA, College of Forest Resources, pp 364–369
Fox JC, Bi H, Ades PK (2007) Spatial dependence and individual-tree growth models: I. Characterising spatial dependence. For Ecol Manag 245:10–19
Genest C, MacKay J (1986) The joy of copulas: bivariate distributions with uniform marginals. Am Stat 40:280–283
Gregoire TG, Ståhl G, Næsset E et al (2011) Model-assisted estimation of biomass in a LiDAR sample survey in Hedmark County, Norway. Can J For Res 41:83–95
Hayashi R, Kershaw JA Jr, Weiskittel AR (2015) Evaluation of alternative methods for using LiDAR to predict aboveground biomass in mixed species and structurally complex forests in northeastern North America. Math Comput For Nat Resour Sci 7:49–65
Hudak AT, Crookston NL, Evans JS et al (2008) Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. Remote Sens Environ 112:2232–2245
Johnsen KH, Samuelson L, Teskey RO et al (2001) Process models as tools in forestry research and management. For Sci 47:2–8
Kershaw JA Jr, Ducey MJ, Beers TW, Husch B (2016) Forest Mensuration, 5th edn. Wiley/Blackwell, Hobokin, NJ
Keyes CR, O'Hara KL (2002) Quantifying stand targets for silvicultural prevention of crown fires. West J Appl For 17:101–109
Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 28:1–26
Leary RA (1988) Some factors that will affect the next generation of Forest growth models. In: Ek, a.R., S.R. Shifley and T.E. Burk (eds.) forest growth modelling and prediction. USDA, Forest Service, North Central Forest Experiment Station General Techinal Report GTR-NC-120, pp 22–32
Li R, Weiskittel AR, Kershaw JA Jr (2011) Modeling annualized occurrence, frequency, and composition of ingrowth using mixed-effects zero-inflated models and permanent plots in the Acadian Forest region of North America. Can J For Res 41:2077–2089
Liu C, Zhang L, Davis CJ et al (2002) A finite mixture model for characterizing the diameter distributions of mixed-species forest stands. For Sci 48:653–661
Loo J, Ives N (2003) The Acadian Forest: historical condition and human impacts. For Chron 79:462–474
MacLean DA, Adams G, Pelletier G et al (2010) Forest dynamics, succession and habitat relationships under differing levels of silviculture. Sustainable Forest Management Network, Edmonton, AB
MacLean RG, Ducey MJ, Hoover CM (2014) A comparison of carbon stock estimates and projections for the northeastern United States. For Sci 60:206–213
Maguire DA, Kershaw JA Jr, Hann DW (1991) Predicting the effects of silvicultural regime on branch size and crown wood core in Douglas-fir. For Sci 37:1409–1428
McCarter JB, Long JN (1986) A lodgepole pine density management diagram. West J Appl For 1:6–11
McGarrigle E (2013) Stand structure and development models for the Acadian Forest region. PhD Dissertation, University of New Brunswick, Fredericton, NB, Canada
McGarrigle E, Kershaw JA Jr, Ducey MJ, Lavigne MB (2013) A new approach to modeling stand-level dynamics based on informed random walks: influence of bandwidth and sample size. Forestry 86:377–389
McRoberts RE (2001) Imputation and model-based updating techniques for annual forest inventories. For Sci 47:322–330
McRoberts RE, Tomppo EO (2007) Remote sensing support for national forest inventories. Remote Sens Environ 110:412–419
Nelsen RB (2006) An introduction to copulas, 2nd edn. Springer, New York
NSDNR (2008) State of the Forest Report 1995–2005: Nova Scotia Forests in Transition. Nova Scotia Department of Natural Resources
Pretzsch H, Grote R, Reineking B et al (2008) Models for forest ecosystem management: a European perspective. Ann Bot 101:1065–1087
R Development Core Team (2016) R: a language and environment for statistical computing. R Found. Stat. Comput. Vienna Austria, In http://www.R-project.org. Accessed 8 Apr 2017
Reineke LH (1933) Perfecting a stand-density index for even- aged forests. J Agric Res 46:627–638
Rowe JS (1972) Forest regions of Canada. Canadian Forestry Service
Russell MB (2012) Modeling individual tree and snag dynamics in the mixed -species Acadian forest. PhD Dissertation, University of Maine, Orono, ME
Russell MB, Weiskittel AR, Kershaw JA Jr (2011) Assessing model performance in forecasting long-term individual tree diameter versus basal area increment for the primary Acadian tree species. Can J For Res 41:2267–2275

Russell, MB, Weiskittel, AR, Kershaw, JA (2014) Comparing strategies for modeling individual-tree height and height-to-crown base increment in mixed species Acadian forests of northeastern North America. Eur J For Res 133(6):1121–1135.
Scott DW (1992) Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons, New York
Vanclay JK (1991) Compatible deterministic and stochastic predictions by probabilistic modelling of individual trees. For Sci 37:1656–1663
Weiskittel AR, Hann DW, Kershaw JA Jr, Vanclay JK (2011a) Forest growth and yield modeling, 2nd edn. Wiley/Blackwell, New York
Weiskittel AR, Kershaw JA Jr, Hennigar C (2014) Refinement of Forest vegetation simulator individual tree model growth and yield model for the Acadian region. University of Maine, Orono, ME, Cooperative Forest Research Unit. 2013 Annual Report. pp. 36-41. Available online: https://umaine.edu/cfru/files/2015/05/Annual-Report-2013.pdf
Weiskittel AR, Wagner RG, Seymour RS (2011b) Refinement of the Forest vegetation simulator, northeastern variant growth and yield model: phase 2. University of Maine, Orono, ME, Cooperative Forest Research Unit
Weiskittel AR, Wilson DS, Kuehne C (2016) Forecasting Douglas-fir response to forest management: evaluating alternative approaches and growth model projection uncertainty. 2016 Western Mensurationists Annual Meeting. Skamania Lodge, Stevenson, WA. Available online: http://www.westernmensurationists.org/m2016/Weiskittel_Aaron.pdf
Woods M, Robinson DCE (2008) Development of FVSOntario: a Forest vegetation simulator variant and application software for Ontario. USDA, Forest Service, Rocky Mountain Research Station
Zar JH (2009) Biostatistical analysis, 5th edn. Pearson, New York
Zhang L, Gove JH, Liu C, Leak WB (2001) A finite mixture of two Weibull distributions for modeling the diameter distributions of rotated-sigmoid, uneven-aged stands. Can J For Res 31:1654–1659. https://doi.org/10.1139/x01-086