Available online at www.sciencedirect.com

ScienceDirect

Forest Ecology and Management 255 (2008) 1759–1771

Forest Ecology and Management

www.elsevier.com/locate/foreco

Evaluating alternative implementations of the Lake States FVS diameter increment model

Bharat Pokharel, Robert E. Froese *

School of Forest Resources and Environmental Science, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931, USA

Received 14 August 2007; received in revised form 28 November 2007; accepted 29 November 2007

Abstract

We evaluated the STEMS family of diameter increment models that have been incorporated in the Lake States variant of the Forest Vegetation Simulator. We paired validation using regression-based equivalence tests with evaluation of trends in errors across species and predictor variables, using independent data from the Michigan Forest Inventory and Analysis program. Our evaluation shows that 10-year increment bias is substantial, almost 17% on average, and our tests failed to validate the model for every one of the 30 most common tree species in the region. A comparative analysis among all alternative implementations demonstrated that error arose from structural weaknesses in the underlying model. Furthermore, the way the model is currently implemented in the Forest Vegetation Simulator partly masks poor performance at the tree level, but likely amplifies error at the stand level, a particularly troubling result in many conceivable applications. Our results also affirm that a simple adjustment factor as a function of dbh provides an inadequate correction of prediction bias. We argue that the diameter increment model needs to be re-engineered.

Keywords: Diameter increment model; Equivalence test; FVS; Model validation; Model evaluation; STEMS

1. Introduction

Vegetation simulation models are valuable tools for sustainable management and scientific investigations in forested ecosystems (Vanclay, 1994; Frelich, 2002; Kimmins et al., 2004; Davis et al., 2005). In the Great Lakes region of the Upper Midwest, the Lake States Variant of the Forest Vegetation Simulator (LS-FVS – Dixon, 2007) has been important and extensively used in these regards. Versions of LS-FVS have been key components of US National Forest Plans and Revisions (e.g., the Hiawatha National Forest in Michigan – USDA Forest Service, 1986, 2006) and model-enhanced regional forest inventory (Hansen, 1990; Schmidt et al., 1997). Henderson (2007) used LS-FVS to enhance ecological succession models. An emerging application of FVS is in modelling carbon pools (e.g., Hoover, 2007) and carbon in sequestration projects (e.g., Smith, 2007).

The accuracy of model projections is of critical importance in many foreseeable applications. For example, the Chicago Climate Exchange has recently approved the use of the TWIGS model (The Woodman’s Ideal Growth Projection System – Miner et al., 1988) for estimating forest carbon sequestration that may be saleable as offsets on the exchange. Smith’s (2007) case study relied on TWIGS for this purpose. Following Smith’s analysis, and assuming tree volume increment corresponds with carbon offset revenue, a hypothetical 10% over-prediction by the model could imply a 31% over-estimation of net offset dollar value, an amount that is clearly substantial. Notably, LS-FVS is based upon the same fundamental growth equations found in LS-TWIGS (Bush and Brand, 1995).

In this study, we critically evaluate the key driving-component in LS-FVS: the large-tree diameter increment (DI) sub-model. We use data from the Forest Inventory and Analysis (FIA) program from Michigan and focus on the disparate ways the DI model has been implemented within the larger model framework since first introduced in the Lake States. We expect the model to be imperfect, so we use equivalence tests (*sensu* Froese and Robinson, 2007) to identify imperfection that is informative or meaningful.

* Corresponding author. Tel.: +1 906 487 2723; fax: +1 906 487 2915.
E-mail addresses: bpokhare@mtu.edu (B. Pokharel), froese@mtu.edu (R.E. Froese).

0378-1127/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.foreco.2007.11.035

1.1. Lake States FVS

LS-FVS is a single-tree, aspatial forest growth modelling framework (*sensu* Robinson and Ek, 2000) and comprises a collection of component models that predict tree and stand conditions based on initial state and subsequent disturbances. The model is ‘‘diameter driven’’; a projection cycle begins by first predicting diameter increment for each modelled tree, which is subsequently used directly or indirectly as a predictor in other component functions (Dixon, 2007). The current LS-FVS release relies upon a diameter increment function adapted from a series of earlier models (Bush and Brand, 1995): TWIGS (Miner et al., 1988), its precursor STEMS (the Stand and Tree Evaluation and Modelling System – Belcher et al., 1982) and the original implementation known as ‘‘A Generalized Forest Growth Projection System’’ (USDA Forest Service, 1979).

These models share a common approach. The annual increment in diameter is posited as a function of two basic components: the potential (POT), which is reduced by a competition modifier (MOD) (Hahn and Leary, 1979). The MOD is a multiplier, taking a value between 0 and 1, that accounts for deviations in the subject tree from the potential increment attainable in the absence of competition. The model structure in the current LS-FVS release includes a diameter adjustment factor (DAF) to account for bias revealed in a validation exercise (Holdaway, 1985), and is structured as follows (Bush and Brand, 1995):

$$\Delta \mathrm{dbh}_{\mathrm{annual}} = \mathrm{POT} \cdot \mathrm{MOD} + \mathrm{DAF}$$

$$\mathrm{POT} = \alpha_1 - \alpha_2 \cdot \mathrm{dbh}^{\alpha_3} + \alpha_4 \cdot \mathrm{SI} \cdot \mathrm{CR} \cdot \mathrm{dbh}^{\alpha_5}$$

$$\mathrm{MOD} = 1 - \exp\left(- f(R) \cdot g(\mathrm{AD}) \cdot \left[\frac{(\mathrm{BAMAX} - \mathrm{BA})}{\mathrm{BA}} \right]^{1/2} \right)$$

$$f(R) = \beta_1 \cdot \left[1 - \exp\left(\beta_2 \cdot \frac{\mathrm{dbh}}{\mathrm{AD}} \right) \right]^{\beta_3} + \beta_4$$

$$g(\mathrm{AD}) = \theta_1 \cdot (\mathrm{AD} + 1)^{\theta_2}$$

$$\mathrm{DAF} = \eta_1 \cdot \mathrm{dbh} + \eta_2 \cdot \mathrm{dbh}^2 + \eta_3$$

The POT is modelled as a function of diameter at breast height (dbh), site index (SI) and crown ratio (CR). The MOD includes three groups of factors (Holdaway, 1984): the approach of current stand basal area (BA) to the maximum for the species in question (BAMAX); in $f(R)$ the tree dbh relative to stand average dbh (AD) as an index of social position; and, in $g(AD)$ a function characterizing the average stand diameter effect. Then, α_1–α_5, β_1–β_4, θ_1–θ_2 and η_1–η_3 are species-specific coefficients (Bush and Brand, 1995).

Though the DI model in LS-FVS is mathematically the same as in TWIGS and STEMS (the latter developed before the DAF), it was implemented in a different framework. TWIGS and STEMS are annual models, while FVS defaults to a decadal interval. For simplicity, the first version of LS-FVS generated a single annual increment prediction and multiplied it by 10 (Bush and Brand, 1995). In 2005 the LS-FVS program logic was updated so the DI model iterated within the 10-year LS-FVS cycle (USDA Forest Service, 2005b), bringing it in line with earlier implementations.

1.2. Past validation studies

Evaluations of STEMS, TWIGS and LS-FVS, using independent data, have generally shown the models to over-predict diameter increment. Smith (1983) used data from the Upper Peninsula of Michigan and found STEMS over-predicted increment for 18–22 species. Holdaway and Brand (1983) tested STEMS with data collected across the three Lake States. They also reported that the model over-predicted diameter increment for most tree species, and that over-prediction was significantly correlated with dbh. Holdaway (1985) developed the DAF to correct for this bias; however, this adjustment was developed using another independent data set, collected exclusively from northern Wisconsin. Holdaway’s (1985) adjustment was implemented region wide, creating STEMS85 (Holdaway and Brand, 1986), and carried over into TWIGS and LS-FVS despite this geographic restriction. Results from a follow-up study (Holdaway and Brand, 1986) showed over-prediction was reduced in Michigan by this change, though bias increased in Minnesota. Guertin and Ramm (1996) tested TWIGS and Canavan and Ramm (2000) tested LS-FVS using a small set of plots on the Manistee National Forest and found that the models consistently over-predict diameter increment for the selected northern hardwood tree species. Smith-Mateja and Ramm (2002) reported that LS-FVS over-predicts diameter increment for *RP* (*Pinus resinosa* Ait.– see Table 1 for species codes, common and Latin names) plantations under different management regimes.

Common to all past validation studies is that models were run as complete units while evaluation metrics included both tree- and stand-level variables. In other words, both dbh increment and mortality sub-models were used to project test data for multiple cycles with periodic updating of stand statistics. This is true also for Holdaway’s (1985) DAF. She projected entire stands 10–15 years with STEMS and then derived tree-level annualized prediction errors to develop the DAF. With this approach accuracy for a given species and prediction interval can be expected to become progressively more confounded across species, plots and model sub-components; errors for a given tree affect future predictions because they affect the calculation of stand-level competition variables such as BA and AD during updating. While this may be a fair test of the model in application at the stand level, it compromises conclusions about inherent weaknesses in model components, such as the diameter increment model for a given species, because the relative effects of the confounding are not estimated.

1.3. Objectives

Our overall goal was to reconcile disparate past evaluations, adjustments, and implementations of the LS-FVS diameter increment model using geographically extensive test data. We

Table 1
Common and scientific names of species examined in this study and their associated species codes

Species code	Common name	Scientific name
AB	American beech	*Fagus grandifolia* Ehrh.
AE	American elm	*Ulmus americana* L.
BA	Black ash	*Fraxinus nigra* Marsh.
BC	Black cherry	*Prunus serotina* Ehrh.
BF	Balsam fir	*Abies balsamea* (L.) Mill.
BO	Black oak	*Quercus velutina* Lam.
BP	Balsam poplar	*Populus balsamifera* L.
BS	Black spruce	*Picea mariana* (Mill.) B.S.P.
BT	Bigtooth aspen	*Populus grandidentata* Michx.
BW	American basswood	*Tilia americana* L.
EH	Eastern hemlock	*Tsuga canadensis* (L.) Carr.
GA	Green ash	*Fraxinus pennsylvanica* Marsh.
JP	Jack pine	*Pinus banksiana* Lamb.
NP	Northern pin oak	*Quercus ellipsoidalis* E.J. Hill
PB	Paper birch	*Betula papyrifera* Marsh.
QA	Quaking aspen	*Populus tremuloides* Michx.
RM	Red maple	*Acer rubrum* L.
RN	Red pine (natural)	*Pinus resinosa* Ait.
RO	Northern red oak	*Quercus rubra* L.
RP	Red pine (plantation)	*Pinus resinosa* Ait.
SC	Scotch pine	*Pinus sylvestris* L.
SM	Sugar maple	*Acer saccharum* Marsh.
SV	Silver maple	*Acer saccharinum* L.
TA	Tamarack	*Larix laricina* (Du Roi) K. Koch
WA	White ash	*Fraxinus americana* L.
WC	Northern white-cedar	*Thuja occidentalis* L.
WO	White oak	*Quercus alba* L.
WP	White pine	*Pinus strobus* L.
WS	White spruce	*Picea glauca* (Moench) Voss
YB	Yellow birch	*Betula alleghaniensis* Britton

divided this goal into four objectives: (1) to conduct a formal validation of the model as currently implemented using equivalence tests (Berger and Hsu, 1996; Wellek, 2003); (2) to evaluate the impact of alternative implementations of the same underlying model formulation; (3) to estimate the influence of confounding through annual updating on the validation test; and (4) to determine if a new DAF is warranted and if so to calibrate it using the test data.

2. Materials and methods

We examined four implementations of the DI model: (1) the base POTMOD function as implemented in STEMS in 1982 (''STEMS''), (2) the base function plus DAF as implemented in STEMS85, TWIGS and the current version of the Lake States FVS variant (''LS-FVS''), (3) the STEMS85/TWIGS/LS-FVS implementation with competition variables interpolated to reduce confounding (''INTERP''), and (4) the initial implementation in LS-FVS (version 1), which extrapolated a 1-year prediction by multiplying it by 10 (''EXTRAP''). As much as possible we structured our evaluation to isolate the DI model from other FVS model components (e.g., simulation of mortality or crown ratio change) or case-specific user adjustments (e.g., DI multipliers or BAMAX adjustments).

We used tests of equivalence for validation instead of the traditional approach to statistical testing (Robinson and Froese, 2004). In the equivalence approach, the test hypotheses are reversed from their usual arrangement. The null hypothesis is of *dissimilarity*; e.g., that the mean prediction is dissimilar to the mean measured value and the model is biased. Then a model is validated only if it is truly unbiased and the test is sufficiently powerful to detect that. For a discussion see, e.g., Froese and Robinson (2007), Robinson et al. (2005), Robinson and Froese (2004), Wellek (2003) and Berger and Hsu (1996).

2.1. Testing and evaluation data

We used data from two successive FIA inventories of Michigan in this study: cycle 4 data collected between 1980 and 1984 and cycle 5 data collected between 1990 and 1993. Data from the new, annualized FIA design (FIA cycle 6, Bechtold and Patterson, 2005) were also available, but were not used because FIA privacy policy prohibits explicit linking across cycles 5 and 6. We limited the analysis to Michigan to constrain the analytical effort and to permit assessment of the impact of extrapolation of earlier bias adjustments outside of the range of the calibration data.

In Michigan, the FIA cycles 4 and 5 sampling design used a 10-point cluster with probability proportional to size, a basal area factor of 8.61 $\mathrm{m}^2\,\mathrm{ha}^{-1}$ and a 12.5 cm breakpoint dbh (Doman et al., 1981; USDA Forest Service, 1991). The cluster was augmented by three 0.0014 ha fixed-area subplots to sample saplings 2.5–12.5 cm dbh. If any point fell outside of forested conditions, it was moved according to pre-determined criteria. We limited the data set to species with at least 100 observations (Tables 1 and 2); the final testing data included 38,047 trees across 3369 field locations with measured increment between cycles 4 and 5, representing a wide range of stand age, species composition, stand structure, growing conditions and stocking in Michigan.

We derived a standardized 10-year observed diameter increment for each tree alive at both cycles based on a linear approximation:

$$\Delta\mathrm{dbh}_{10} = \frac{\mathrm{dbh}_2 - \mathrm{dbh}_1}{T_2 - T_1} \times 10$$

where dbh is the diameter at breast height, $\Delta\mathrm{dbh}_{10}$ is the diameter increment for a period of 10 years and T is the fractional year (year plus month/12) of measurement. We assigned the species-specific SI measured at cycle 4 to each tree when possible. For trees without measured SI we followed the protocol in LS-FVS to estimate SI. Where possible, we used conversion equations developed by Carmean and Vasilevsky (1971) and Carmean (1979), with conversion through *QA* if there was no direct conversion available. For the very few trees (177 trees across 5 plots) that had no measured SI values, we followed the protocol in LS-FVS and assumed an *RP* site index of 18 m.

2.2. Generating model predictions

We derived predicted 10-year diameter increment differently depending on the scenario being simulated. In each case, we derived stand-level predictors (BA, AD) using beginning of

Table 2
Summary of increment measured and predicted using different implementations of the LS-FVS large-tree diameter increment model

Species code	No. plots	No. trees	Measured increment (cm dec^{-1})		Mean predicted increment (cm dec^{-1})				Mean prediction error (%)				σ_{pe} LS-FVS (cm dec^{-1})
			$\bar{y}$	σ	STEMS	LS-FVS	INTERP	EXTRAP	STEMS	LS-FVS	INTERP	EXTRAP	
AB	199	557	2.72	1.87	3.27	2.99	3.05	2.98	−20.2	−9.9	−12.1	−9.6	1.63
AE	77	107	5.16	4.21	5.40	4.32	4.29	4.60	−4.7	16.3	16.9	10.9	3.81
BA	253	648	1.81	1.38	2.51	2.14	2.18	2.30	−38.7	−18.2	−20.4	−27.1	1.33
BC	324	731	2.55	2.02	3.70	3.56	3.61	3.79	−45.1	−39.6	−41.6	−48.6	1.94
BF	571	1225	2.08	1.43	2.98	2.59	2.65	2.83	−43.3	−24.5	−27.4	−36.1	1.25
BO	167	692	2.96	2.10	4.38	2.79	2.76	2.95	−48.0	5.7	6.8	0.3	1.94
BP	169	365	3.52	2.17	4.37	3.98	4.09	4.10	−24.1	−13.1	−16.2	−16.5	2.10
BS	377	1232	1.37	1.08	2.54	1.45	1.47	1.66	−85.4	−5.8	−7.3	−21.2	1.02
BT	338	1095	3.45	1.92	3.99	4.32	4.42	4.38	−15.7	−25.2	−28.1	−27.0	1.79
BW	323	1141	2.45	1.83	3.76	3.00	3.06	3.05	−53.5	−22.4	−24.9	−24.5	1.86
EH	310	1151	2.49	1.52	1.57	2.12	2.17	2.07	36.9	14.9	12.9	16.9	1.43
GA	99	197	3.60	2.23	3.95	3.41	3.41	3.68	−9.7	5.3	5.3	−2.2	1.92
JP	308	1335	2.26	1.59	3.32	3.07	3.17	3.40	−46.9	−35.8	−40.3	−50.4	1.39
NP	49	144	3.72	1.96	5.14	3.61	3.62	3.72	−38.2	3.0	2.7	0.0	1.92
PB	563	1433	1.60	1.31	2.54	2.23	2.31	2.37	−58.8	−39.4	−44.4	−48.1	1.31
QA	815	2142	3.81	2.01	4.60	4.15	4.28	4.25	−20.7	−8.9	−12.3	−11.5	1.97
RM	1259	4734	2.38	1.82	2.80	2.57	2.64	2.71	−17.6	−8.0	−10.9	−13.9	1.65
RN	117	170	3.76	2.10	4.70	4.63	4.72	5.10	−25.0	−23.1	−25.5	−35.6	1.72
RO	493	2023	3.00	2.03	3.60	3.22	3.29	3.31	−20.0	−7.3	−9.7	−10.3	1.90
RP	297	1463	3.63	1.87	5.47	5.53	5.72	6.18	−50.7	−52.3	−57.6	−70.2	1.66
SC	44	283	3.47	2.01	3.35	2.84	2.77	3.20	3.5	18.2	20.2	7.8	2.04
SM	875	5346	2.09	1.70	2.77	2.43	2.51	2.44	−32.5	−16.3	−20.1	−16.7	1.51
SV	62	332	5.22	3.60	3.56	3.27	3.15	3.36	31.8	37.4	39.7	35.6	3.94
TA	161	377	2.20	1.41	1.87	1.92	1.96	1.94	15.0	12.7	10.9	11.8	1.47
WA	216	535	3.88	2.61	4.45	3.37	3.30	3.44	−14.7	13.1	14.9	11.3	2.71
WC	666	4317	1.64	1.14	2.83	1.91	1.94	2.02	−72.6	−16.5	−18.3	−23.2	1.10
WO	289	971	2.10	1.45	1.92	2.29	2.35	2.30	8.6	−9.0	−11.9	−9.5	1.23
WP	451	1532	4.48	3.47	4.30	4.31	4.37	4.37	4.0	3.8	2.5	2.5	2.90
WS	341	734	2.75	2.09	6.44	6.48	6.68	7.19	−134.2	−135.6	−142.9	−161.5	2.56
YB	410	1035	1.82	1.30	1.99	1.82	1.88	1.85	−9.3	0.0	−3.3	−1.6	1.25
ALL	3369	38047	2.31	1.86	3.12	2.70	2.76	2.84	−35.1	−16.9	−19.5	−22.9	1.70

Diameter increment has been standardized to a 10-year interval. Prediction error follows the convention where error is calculated as measured minus predicted increment; i.e., negative errors represent over-prediction by the model.

cycle conditions and sampling fractions associated with each tree. Sampling fractions were held constant for live trees throughout 10 years of simulation (McRoberts, 2001). We held SI constant and used default values for BAMAX in all simulations as well.

To generate a STEMS prediction, we generated annual increment predictions for each tree, updating stand statistics (AD, BA) each year, for 10 consecutive iterations. We updated CR as well by linearly interpolating measured values at the two FIA inventories. To generate an LS-FVS prediction, we used the same procedure but with the DAF applied to each prediction. In both cases, we accounted for the impact of tree mortality on BA by linearly depleting sampling fraction for trees that died sometime between the two FIA inventories. We took a different approach for INTERP, with the goal of approximately isolating the confounding of prediction errors on consecutive updates of stand statistics. Instead of annual updating and depleting sampling fraction for dead trees, we linearly interpolated known live-tree BA and AD directly between measured values at each FIA inventory. Finally, we generated an EXTRAP prediction by simply multiplying the first annual LS-FVS prediction by 10. In this case, no accounting for mortality was necessary.

2.3. Evaluation and validation tests

Model evaluations were conducted in a regression framework, with measured increment as the response and modelled increment as the predictor (Robinson et al., 2005). In this approach, tests can be interpreted to provide two related conceptual validations of the underlying model (Froese and Robinson, 2007). For example, bias is the difference between the intercept, shifted to the mean, and the mean modelled value:

$$\text{bias} = (b_0 + b_1 \cdot \bar{\hat{y}}) - \bar{\hat{y}}.$$

A test of the intercept compares mean increment predicted by the model to that observed in the data, providing a population-level validation of the model as a prediction tool. A test of the slope provides an assessment of the model's ability to match individual tree or plot-level differences by comparing the slope to the slope of the 1:1 line. Because this compares predicted deviation to observed deviation, it may be interpreted as one possible validation of the underlying model structure, examining to a degree whether the result ''is correct for the right reasons'' (Robinson et al., 2005).

We used the two-one-sided test (Wellek, 2003) for all tests of equivalence. The details of this test are described elsewhere (e.g., Berger and Hsu, 1996; Wellek, 2003; Robinson and Froese, 2004) and are only briefly summarized here. In each test, a performance metric is compared to an ideal reference value; e.g., in a regression context, mean measured increment is compared to mean predicted increment, and the slope is compared to 1. The decision is made by first nominating a region within which differences between test and reference data are considered negligible. Then, two one-sided confidence intervals of size α are constructed around the selected metric of model performance. The interval is compared to the nominated indifference region and, if completely contained by it, the null hypothesis of dissimilarity is rejected for that test. In other words, if the metric and reference value are truly similar and a test is powerful enough both the metric and the entire confidence interval will be contained within the indifference region.

We nominated our indifference intervals to match the context for each equivalence test. For tests of equivalence between predicted against measured increment, we selected an indifference interval similar to that implied by measurement error tolerances in U.S. forest inventory programs. For dbh or dbh increment, this is commonly 0.25 cm dec^{-1} (e.g., USDA Forest Service, 1991, 2005a), which for our test data was approximately 10% of mean 10-year predicted increment. We adopted a less stringent criterion of ±25% of the ideal value of 1 for the slope, based on the argument that concordance between individual predictions is less important than population-level accuracy.

For tests of equivalence between values predicted using different model implementations, we used Wellek's (2003) suggested ''liberal'' choice: ±74% of the standard deviation, in this case the residual standard deviation from the regression model. Wellek (2003, pp. 11–13) shows this value for the two-one-sided test and Gaussian parameters to be approximately equivalent to well-established criteria used by the Food and Drug Administration in bioequivalence studies. Because indifference regions are inherently subjective, and to aid interpretation, we also calculated the minimum percentage that would have resulted in a successful validation.

The FIA cycles 4 and 5 inventories used variable-radius plots where the probability of tree selection was proportional to dbh. Under this unequal probability sampling, estimates of tree attributes, such as prediction error, may be biased if they are correlated with diameter (Lappi and Bailey, 1987; Overton and Stehman, 1995). To eliminate design bias in general we used analytical methods that weight observations by the inverse of their sampling probability. Statistical analyses were all performed in R (R Development Core Team, 2007) utilizing functions for estimating means and parameters of linear models with sampling weights from the ''survey'' package (Lumely, 2004). Variance estimates from these functions are produced via Taylor series approximation.

2.4. Development of a new DAF

We calibrated a new DAF as a quadratic function of dbh for each species following Holdaway's (1985) methodology. This DAF was intended to replace the current DAF; thus, the response variable in this case was tree-level prediction error from the STEMS implementation of the DI model. To evaluate whether the quadratic function of dbh was alone sufficient we also generated scatterplots of STEMS prediction error against each of the prediction variables used in the DI model (i.e., dbh, SI, CR, BA and AD). These plots were augmented with design-unbiased linear regression lines to evaluate the importance of observed trends.

3. Results

3.1. Observed increment and model performance

Observed mean diameter increment across the entire testing data was modest at 2.31 cm dec^{-1}, corresponding to just over 1 mm radially per year (Table 2). Increment varied widely among species, ranging from just 1.37 cm dec^{-1} for *BS* up to 5.22 cm dec^{-1} for *SV*. Variance of actual increment was roughly related to the mean; for example, increment for *WP*, *SV* and *AE* had both the largest means and variances, and *BS*, *PB* and *WC* had the smallest. Just five species comprised nearly half of all observations: *SM*, *RM*, *WC*, *QA* and *RO*, listed in decreasing order, each had more than 2000 individual samples. Mean increment ranged widely in these species as well, from 1.64 cm dec^{-1} for *WC* up to 3.81 cm dec^{-1} for *QA*. The least-abundant species, in decreasing order, were *GA*, *RN*, *NP* and *AE*, each represented by less than 200 sample trees.

Considering predictions for the entire data set, the largest mean predicted decadal increment was made by the original STEMS model, at 3.12 cm dec^{-1}, and the smallest by the LS-FVS release, at 2.70 cm dec^{-1} (Table 2). The INTERP and EXTRAP implementations were intermediate, both less than 2.85 cm dec^{-1}, and were closer to LS-FVS than to STEMS. All four implementations over-predict increment substantially, at worst with more than 35% over-prediction by STEMS and at best still a 17% over-prediction by the LS-FVS release.

Across species, STEMS bias ranged from 134% over-prediction for *WS* to nearly 37% under-prediction for *EH* (Table 2). The range was similar for other implementations of the DI model. The species with the largest absolute bias was *WS* for each implementation, reaching nearly 162% over-prediction for EXTRAP. The species with the lowest absolute bias differed across implementations; bias was nearly zero for *SC* under STEMS, *YB* under LS-FVS, *WP* under INTERP, and *NP* under EXTRAP. For nearly every species, increment predicted by either INTERP or EXTRAP was greater than that predicted by LS-FVS. Correspondingly, for species where LS-FVS resulted in over-prediction, that over-prediction always increased under INTERP but not always under EXTRAP.

In terms of absolute bias, among species LS-FVS was most often the best performing implementation and STEMS most often the worst. LS-FVS produced the smallest bias among alternatives for 16 of 30 species and never produced the largest bias (Table 2). In contrast, STEMS produced the largest bias among alternatives for 20 of 30 species. Despite poor average performance, however, STEMS was still the best implementation for 7 species, notably including *WS*, the species for which all implementations performed poorly, and *RP*, a relatively fast-growing species well-represented in the test data.

LS-FVS was usually superior when all implementations produced an over-prediction. In other words, when all implementations over-predicted increment, LS-FVS over-predicted increment the least. Results for the remaining species were less consistent. Where STEMS was superior, it was sometimes because STEMS produced a smaller over-prediction (e.g., *RP*) and sometimes a smaller under-prediction (e.g., *SV*) than other implementations. The same is true for species where INTERP or EXTRAP were superior.

3.2. Tests of equivalence

Equivalence tests of the intercept for LS-FVS predictions against measured values failed to validate the model for every species (Fig. 1). The mean predicted increment fell within the indifference interval for ten species (*AB*, *BO*, *BS*, *GA*, *NP*, *QA*, *RM*, *RO*, *WO*, *WP* and *YB*). For some of these species, mean increment was very close to the centre of the indifference region, and failure to reject was due to large confidence intervals (e.g., *YB*). However, for most of these species mean increment was near the extreme of the indifference interval, and rejection would have occurred only if the confidence interval had been extremely small. For some species (e.g., *AE*, *SV* and *WA*) confidence intervals were very large, reflecting small sample sizes and high native variability in observed increment (Fig. 1; Table 2). The smallest indifference interval that would result in approval of the test for the model was just over 16%, for *RM*. For all but one species the width of the confidence interval exceeded the width of the indifference interval, so even if the model could be perfectly unbiased it would still have not been validated.

Equivalence tests of the slope for LS-FVS predictions against measured values failed to validate the model for all species but *WO* and *SM* (Fig. 1). For six other species (*AB*, *BO*, *EH*, *GA*, *JP* and *RN*) the estimated slope was within the indifference interval but the null hypothesis of dissimilarity was not rejected because of wide confidence intervals. Three of these species, *AB*, *BO*, and *GA*, in addition to *WO*, suffered the same result in tests of the intercept. Yet, unlike results for the intercept, for nearly 2/3 of the species examined the width of the confidence interval was less than the indifference interval and failure to validate the model was because the estimated slope was substantially less than 1. Slopes were often very small, estimated at less than 0.75 for 20 species, and less than 0.50 for 7 of those 20.

Equivalence tests between predictions from EXTRAP against LS-FVS generally failed to validate comparisons of the intercept, but succeeded in validating comparisons of the slope (Fig. 2). Results from equivalence tests between INTERP and LS-FVS were very similar (results not shown), though indifference and confidence interval widths were about half as large. This reflects greater uniformity between predictions from these two implementations as contrasted to comparisons between LS-FVS and EXTRAP. For tests of the slope, comparisons of either INTERP or EXTERP against LS-FVS resulted in rejection of dissimilarity in nearly every case (Fig. 2, results for INTERP not shown). Estimated slopes were, however, much more uniform across species and closer to 1 for comparisons between INTERP and LS-FVS; for all but three species these slopes were greater than 0.95. When EXTRAP is compared to LS-FVS, estimated slopes were much less than 0.95 for more than half of the species examined. For tests of the intercept, comparisons of either INTERP or EXTRAP against LS-FVS resulted in rejection of dissimilarity for only a few species (Fig. 2, results for INTERP not shown). Differences were, on average, much larger when EXTRAP is compared to LS-FVS, and more variable across species. The results for *JP*, *RN*, *RP*, *SC* and *WS* were particularly pronounced when EXTRAP is compared but otherwise not unusual when INTERP is compared.

3.3. Constructing a new DAF

Scatterplots of model prediction errors against predictor variables, augmented with design-unbiased regression lines, revealed a complex set of relationships (Fig. 3). Three basic patterns were revealed. First, for some species (e.g., *QA*), prediction errors could be apparently unrelated to dbh and most other predictors, except one (e.g., basal area). Second, prediction errors could be strongly related to many predictors, including but not limited to dbh (e.g., *RP*). Third, prediction errors could be at best very weakly related to any model predictor (e.g., *SM*).

In preliminary analyses, recalibrations of Holdaway's (1985) additive DAF, using STEMS prediction errors as the response variable, were statistically significant ($p < 0.05$) for 22 of 30 species (data not shown). However, R^2 values never exceeded 11%, and for the five most abundant species never exceeded 3%. Because preliminary results using the quadratic function of dbh were not promising, and scatterplots suggested other predictors might be more powerful descriptors of the trend, further efforts to recalibrate the DAF were not made.

4. Discussion

Tests of models against independent data are bound to find failures, if not because models are imperfect representations of systems, but because often the data sets used to develop and test models simply come from different samples. Thus, the relevant question in testing is not ‘‘does a model fail’’ but rather ‘‘what failures are important’’. Equivalence tests are a useful decision-making tool in this regard. Furthermore, every failure is inseparably linked to the test context. We believe that criteria for an important failure need to be determined with respect to the context, which includes the intended application or objective. Pairing equivalence tests with analyses of error patterns can help understand where and why the model fails.

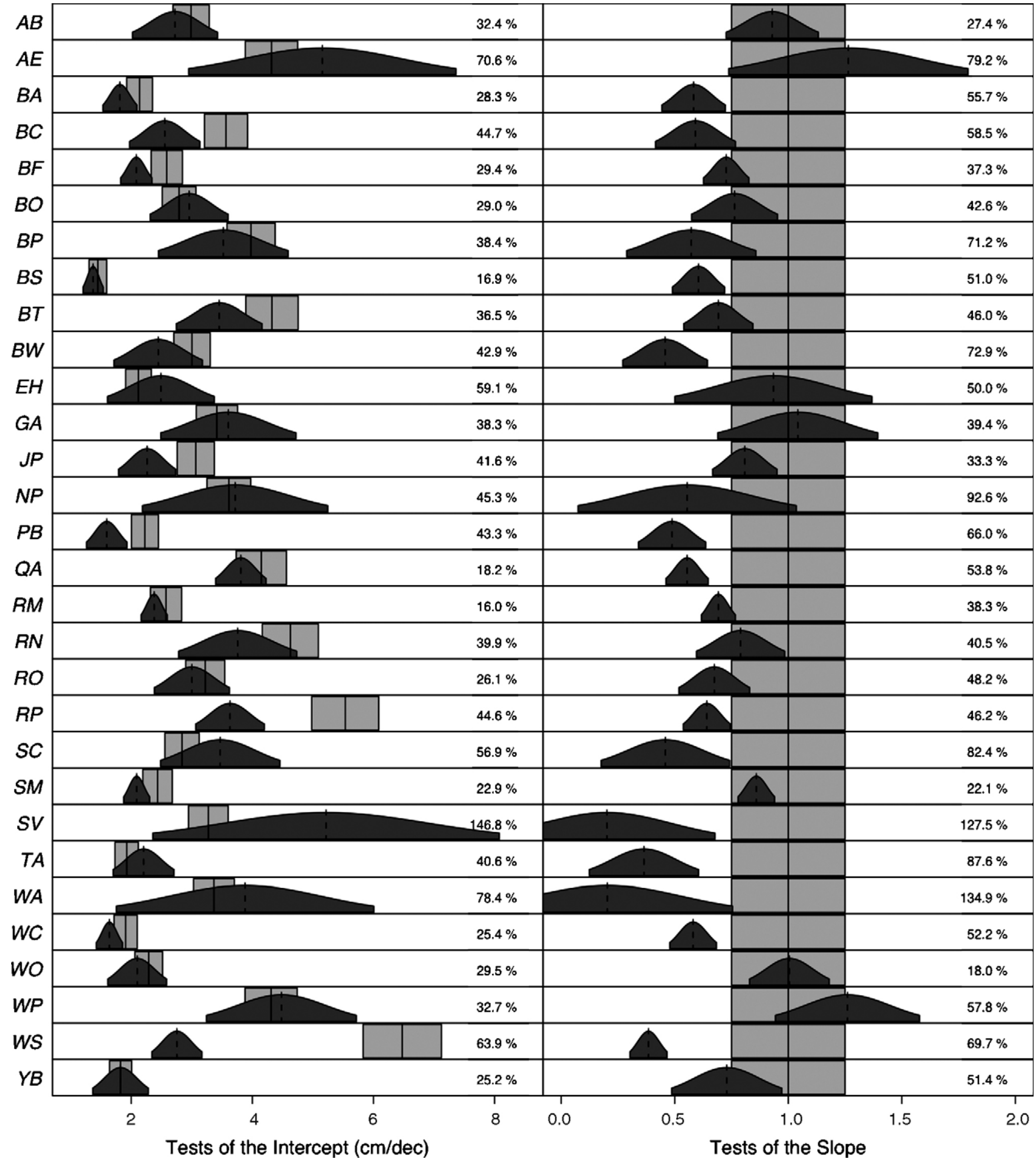

Fig. 1. Results of regression-based tests of equivalence between measured increment and increment predicted by LS-FVS. Light grey bars represent the indifference interval and dark grey truncated Gaussian densities represent the confidence interval. The former are centred on the mean predicted increment for tests of the intercept and a value of 1 for tests of the slope. The latter are centred on the mean measured increment for tests of the intercept or the estimated coefficient for tests of the slope. Where the confidence interval is fully contained by the indifference interval the null hypothesis of dissimilarity is rejected. Stated percentages are the minimum indifference interval that would be required to reject dissimilarity.

In our study, we found performance of the STEMS DI model family, including the current implementation in LS-FVS, to be surprisingly poor. Our results showed that the DI model in application as a predictive tool fails validation using equivalence tests, and by a wide margin. More troubling are results that suggest fundamental problems with the underlying model structure. Together these imply that predictions from the model are not only too large on average, but that they are poorly linked to the underlying processes that explain variability in observed increment. Simulated stand structures will diverge

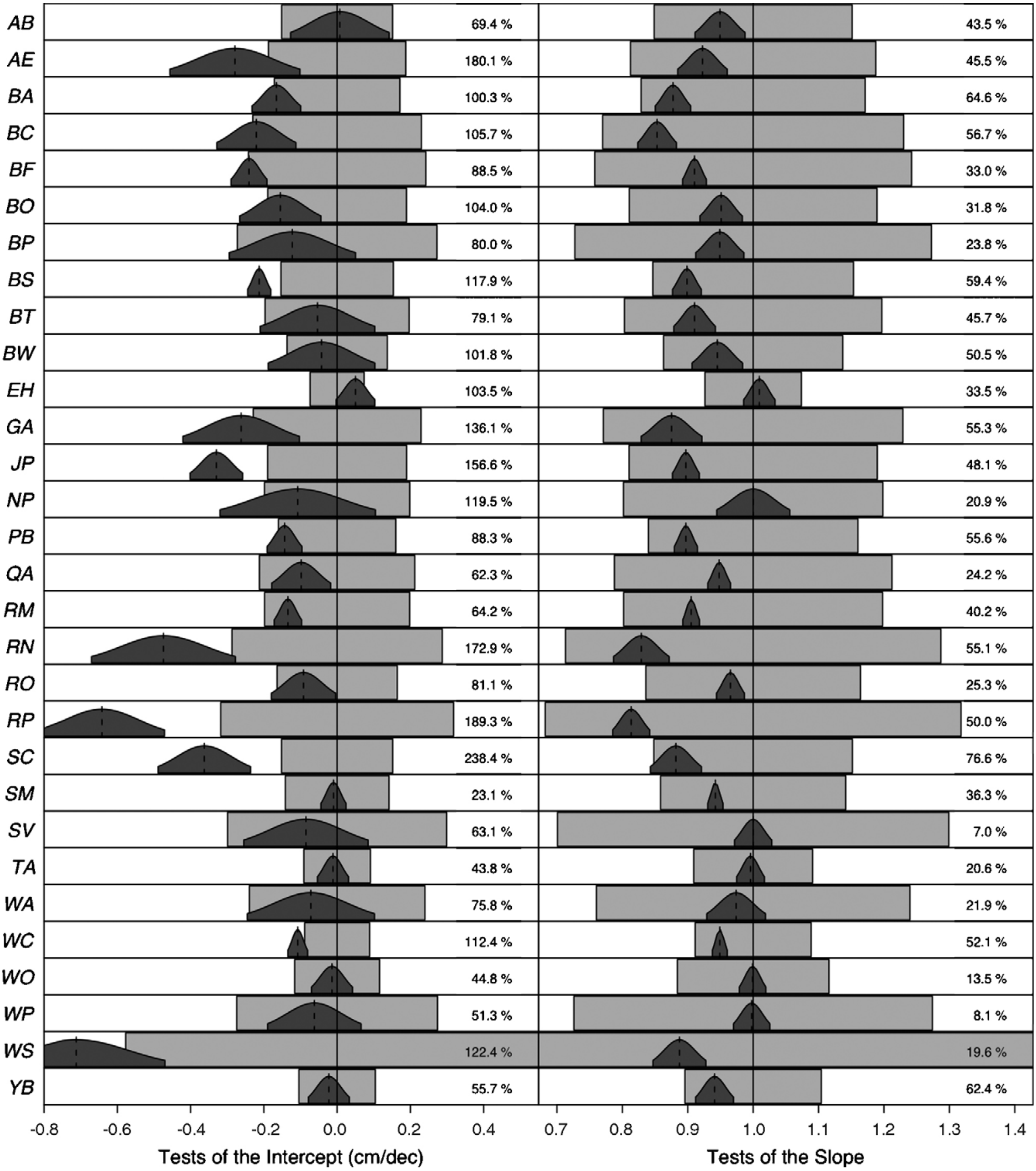

Fig. 2. Results of regression-based tests of equivalence between increment predicted by LS-FVS and EXTRAP. Light grey bars represent the indifference interval and dark grey truncated Gaussian densities represent the confidence interval. The former are centred on the mean prediction from EXTRAP for tests of the intercept and a value of 1 for tests of the slope. The latter are centred on the mean prediction from EXTRAP for tests of the intercept or estimated coefficient for tests of the slope. Where the confidence interval is fully contained by the indifference interval the null hypothesis of dissimilarity is rejected. Stated percentages are the minimum indifference interval that would be required to reject dissimilarity. To simplify the illustration indifference and confidence intervals for tests of the intercept were centred by subtracting the mean prediction from EXTRAP.

from reality, and this inconspicuously feeds back into tree-level evaluations, masking part of the underlying problem. We believe the DI model needs to be re-engineered rather than just re-calibrated or adjusted.

4.1. Comparisons with related studies

We found STEMS performance to be worse than that reported in earlier studies. For example, our results showed

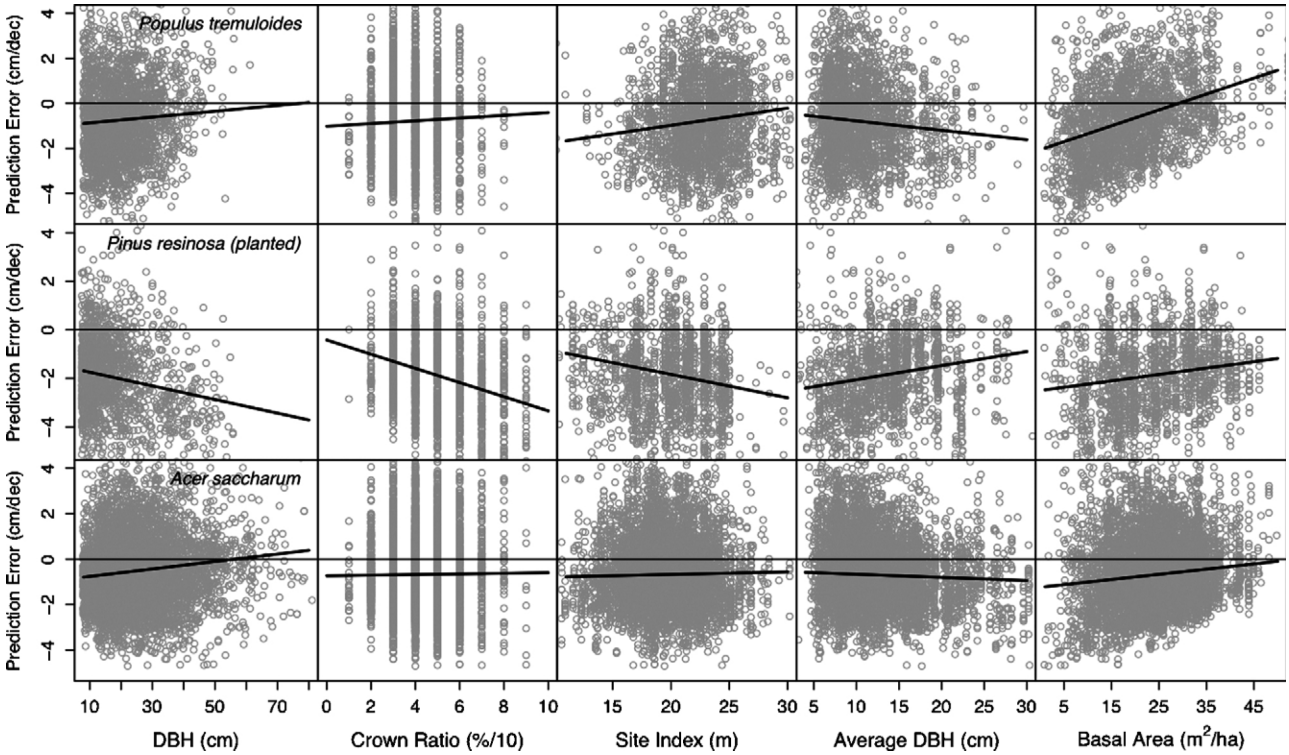

Fig. 3. Scatterplots of 10-year prediction errors against predictor variables for selected tree species, augmented by design-unbiased regression lines. These results are for the underlying STEMS model, without addition of the DAF. Negative errors represent over-prediction by the model.

greater over-prediction than Smith (1983) found for five of the seven most abundant species in his Michigan data. Similarly, when the estimated errors for Michigan National Forests reported by Holdaway and Brand (1983) are combined, they suggest an overall over-prediction of about 0.46 cm dec^{-1}, which is only half as much as shown in our results. When Holdaway and Brand (1986) repeated this analysis using STEMS85, which included the DAF, over-prediction for the same Michigan data declined to about 0.11 cm dec^{-1}, a 76% reduction. In our study, the application of the DAF reduced the over-prediction by just 52%.

Notably, Holdaway and Brand (1986) recommended against using the DAF for *WS* because of concerns about performance, and the DAF was never implemented for this species in STEMS85 or successors. We re-calculated bias under LS-FVS as if the DAF had been implemented for *WS* and found that over-prediction declined from 134 to 68%. While this is still a substantial error, it is a marked improvement and should be implemented in the public LS-FVS release. This has the additional advantage of bringing the model for *WS* in line with all other changes made in 1985.

4.2. LS-FVS performance and validation using equivalence tests

We focused our formal evaluation and validation on LS-FVS because it is the implementation currently available in the software and represents the state-of-the-art for the STEMS family in Michigan. And our results showed LS-FVS model performance on FIA test data to be very poor. Though only half as much as STEMS, the model still over-predicted 10-year dbh increment on average by 17%. Given that this bias approaches twice the chosen indifference criteria, it is not surprising that the model failed validation using equivalence tests for every species in the test data.

To further illustrate this over-prediction, we arbitrarily selected two FIA plots and calculated the associated merchantable volume, using Hahn's (1984) equations (Fig. 4). Both plots were well stocked and of about mean SI; one was a northern hardwood stand dominated by *SM* and the other a nearly pure stand of *RP*. Over-prediction occurs across the typical range of diameter in both stands, and the corresponding total 10-year merchantable volume over-prediction amounted to 6% of standing volume for the *SM* plot, and over 14% in the *RP* plot. Over-prediction as a proportion of observed increment was obviously far greater.

More problematic than overall bias, however, is the apparently poor correspondence between increment measured and predicted by the DI model. This is revealed in evaluation of slope estimates in regression-based comparisons of LS-FVS to measured increment, for which the model failed validation using equivalence tests for all but one species. The magnitude of the estimated slope in most cases was very small, as low as 0.25 for *WA*, but often little more than 0.50 for the most abundant species in the test data. This result is of particular concern. If the model, through the predictors and their coefficients, suggests increment should be relatively larger or smaller than the norm, then the slope indicates the degree to

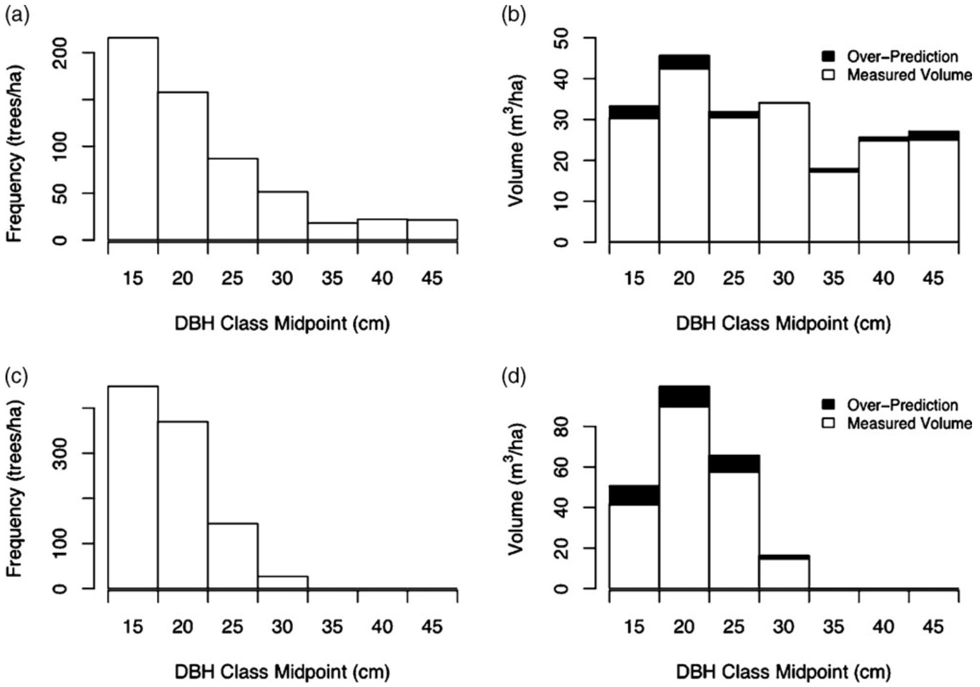

Fig. 4. Histograms illustrating initial density and merchantable volume and volume over-prediction implied by LS-FVS after 10 years for arbitrarily selected examples in the FIA data. Plots (a) and (b) are for a sugar maple dominated northern hardwood stand and (c) and (d) for a red pine plantation.

which measurements are, on average, correspondingly larger or smaller than the norm. The very small slope estimates show that the way increment in the test data is functionally related to the predictors is not very well represented by the model. In other words, the model is not very generalizable beyond the calibration data set and thus there are some very basic weaknesses in the DI model structure.

The considerably variable performance of the DI model across species is also troubling. Many forest types in the test data are comprised of mixtures of five or more species; consider a mixed-northern hardwood forest composed of *SM*, *RM*, *YB*, *EH* and *BW*. Bias under LS-FVS for these species was estimated at −16.3, −8.0, 0.0, 14.9 and −22.4%, respectively. In the absence of some other compensating factor, not only will overall increment be biased, but also differences in relative species performance suggest modelled dynamics will be far different from the actual dynamics. The implications wherever LS-FVS is used for sustainable forest management planning, allowable annual cut calculations or calculation of forest carbon stocks, among many foreseeable applications, may be dramatic.

4.3. Impact of alternative implementations (INTERP and EXTRAP)

Differences between INTERP and LS-FVS suggest that successive over-prediction by the LS-FVS DI model provides a negative feedback when errors are permitted to compound, dampening the 10-year trend. For example, after 10 compounded simulations with LS-FVS we found the accumulated stand basal area to be, on average, more than 2.0 m^2 ha^{-1} greater than that actually observed. In parallel, when INTERP was constrained so that stand-level BA and AD matched observed values, we found mean increment and increment prediction errors after 10 years to be generally larger than for LS-FVS. In these simulations, all else was equal, including the treatment of mortality.

It appears, however, that a 1-year prediction multiplied by 10 is on average a reasonable approximation of 10-year compounded increment, but only for some species. This approach essentially ignores the effects of growth and mortality on subsequent cycles within the 10-year interval. In reality, in some stands, increasing BA might produce more intense competition and diminish observed increment. In others, mortality might result in some release. Diameter increment is usually greatest for medium-sized trees, with all else equal. With a one-times-10 prediction, accumulated growth is ignored; thus, progression towards or away from this peak is not considered either. Since these effects could suggest over or under-prediction, depending on each tree and plot, it is not implausible that, depending on stand conditions, they might cancel each other out.

These results are echoed in the equivalence tests of the slope between INTERP and EXTRAP against LS-FVS. Dissimilarity was rejected in nearly every instance, for both sets of tests, and the estimated slopes were nearly always less than one. However, the pattern of estimated slopes was notably different between the two comparisons. Estimates were uniform across species in comparisons of INTERP to LS-FVS, and relatively close to 1, but they were much more variable when EXTRAP was compared to LS-FVS. The former illustrates the region-wide effect of the general tendency to over-prediction, while the latter illustrates the species and stand-specific appropriateness of approximating 10 cycles with a single prediction.

4.4. A new DAF

We argue that a new calibration of the DAF as a quadratic function of dbh alone is not an effective revision to the original STEMS model. The original DAF (Holdaway, 1985) clearly improves model performance, but our results show that this is true largely because the DAF scales predictions on average, not because prediction errors are particularly correlated with dbh, the sole explanatory variable used in the DAF function. The small slope estimates in equivalence tests of LS-FVS against observed increment, as well as the inconsistent and sometimes strong correlation between error and model predictors other than dbh, suggests there is a larger problem with the model structure. While recalibration of the DAF would improve model performance on average, here we agree with Holdaway and Brand's (1983) argument that performance across domains (say, a range of SI) is often more important than performance at the population level. Improving the STEMS family performance across domains will require re-engineering of the underlying model to resolve the structural issues behind poor performance in the first instance.

4.5. Why is performance of the STEMS family so poor?

Factors not represented in the DI model likely explain some of the differences observed in this study. For example, climate and silvicultural practices have cumulative effects that change with time and are not explicitly considered in the model structure (Froese and Robinson, 2007). Even differences in sampling design between fitting and test data can contribute to bias, though the magnitude can be difficult to estimate (Stage and Wykoff, 1998). We did not explore case-specific adjustments, such as the self-calibration routines, DI multipliers or BAMAX adjustments that can be used to ‘‘tune’’ the model (Dixon, 2007). Though the implications of these would be interesting to examine, our goal was to isolate the underlying model, upon which long-term projections ultimately depend.

A number of earlier studies have suggested that the underlying STEMS structure needs to be re-examined, rather than just recalibrated. The most straightforward evidence for a basic structural weakness in STEMS is provided by Lessard (2000) and McRoberts et al. (2000). They re-calibrated STEMS using FIA data from Minnesota, following the same procedures as used for the original model. However, when they evaluated performance on the 25% of the data they reserved for validation, the new version performed no better than the original STEMS model. This is despite the new version having the advantage of training on essentially the same population as the test data, where differences in climate, silvicultural practices and plot design should not have been an issue.

The calibration procedure may condition the model to be biased when applied to new data. Development and calibration of STEMS followed a two-step procedure: first the POT function is fit, and then the POT predicted for each tree and used in the composite equation to fit the MOD component. Holdaway (1984) suggested that, as a result, the MOD could explain underlying processes or could merely compensate for weaknesses in the POT. Adding the DAF is an explicit attempt to compensate for underlying weaknesses, and essentially makes this a three-step procedure. Holdaway (2000) and Lessard et al. (2001) note that the process of fitting in steps makes analytical derivation of parameter covariances and estimates of prediction uncertainty impossible, and may introduce bias because estimation in later stages does not allow parameters to vary in response to the totality of the data.

Some authors have questioned how POT is defined and measured. Reed et al. (2003) argued that the potential is difficult to observe because true potential growth is rarely, if ever, achieved in reality. Even if potential is achieved, Reed et al. (2001) suggested that potential increment may not be achieved every year; thus, using a single observation of potential over-estimates the longer-term trend, or might select for a measurement error (Vanclay, 1994). Lessard et al. (2001) address these and two-step concerns by simply replacing POT with average growth, viewing MOD as departure from this average, and estimating the model parameters in one step. This approach appears to out-perform STEMS (Holdaway, 2000; Lessard et al., 2001) and should be explored further in the context of LS-FVS.

5. Conclusions

Our validation and evaluation of the LS-FVS DI model family suggests that a re-examination of the underlying POT times MOD model structure and choice of predictor variables is warranted. While the way the DI model is implemented in the current release (LS-FVS) appears to perform best among alternatives, over-prediction of diameter increment is substantial and the model failed validation both as a predictive tool and as an abstraction of the underlying biological system. Moreover, bias varies by species and is confounded through successive projections that are sensitive to stand structure; this suggests forecasts of stand dynamics may be particularly poor. Many different results point to structural weaknesses in the model itself; hence, a recalibration of the DAF would provide little benefit.

Acknowledgements

We thank Christopher Webster for helpful comments on an initial draft of this manuscript, and an anonymous reviewer for suggestions that greatly enhanced clarity and accuracy in our analysis. Figures depicting equivalence test results incorporate some ideas suggested by Andrew Robinson. Funding for this study was provided by Michigan Technological University and the McIntire-Stennis Cooperative Forestry Research Program.

References

Bechtold, W.A., Patterson, P.L. (Eds.), 2005. The enhanced forest inventory and analysis program—national sampling design and estimation procedures. General Technical Report SRS-80. USDA Forest Service, Southern Research Station, Asheville, North Carolina.

Belcher, D.W., Holdaway, M.R., Brand, G.J., 1982. A description of STEMS: the stand and tree evaluation and modeling system. General Technical

Report NC-079. USDA Forest Service, North Central Forest Experiment Station, St. Paul, Minnesota.

Berger, R.L., Hsu, J.C., 1996. Bioequivalence trials, intersection tests and equivalence confidence sets. Stat. Sci. 11, 283–319.

Bush, R.R., Brand, G.J., 1995. Lake States TWIGS geographic variant of the Forest Vegetation Simulator. Unpublished Report. USDA Forest Service, Forest Management Service Center, Fort Collins, Colorado.

Canavan, S.J., Ramm, C.W., 2000. Accuracy and precision of 10 year predictions for Forest Vegetation Simulator—Lake States. North. J. Appl. For. 17, 62–70.

Carmean, W.H., 1979. Site index comparisons among northern hardwoods in northern Wisconsin and upper Michigan. Research Paper NC-169. USDA Forest Service, North Central Forest Experiment Station, St. Paul, Minnesota.

Carmean, W.H., Vasilevsky, A., 1971. Site-index comparisons for tree species in Northern Minnesota. Research Paper NC-065. USDA Forest Service, North Central Forest Experiment Station, St. Paul, Minnesota.

Davis, L.S., Johnson, K.N., Bettinger, P.S., Howard, T.E., 2005. Forest Management: To Sustain Ecological, Economic, and Social Values. Waveland Press, Inc., Long Grove, Illinois.

Dixon, G., 2007. Essential FVS: A user's guide to the Forest Vegetation Simulator (Revised). USDA Forest Service, Forest Management Service Center, Fort Collins, Colorado.

Doman, A.P., Ennis, R., Weigel, D., 1981. North Central Resources Evaluation Field Instructions. USDA Forest Service, North Central Forest Experiment Station, St. Paul, Minnesota.

Frelich, L.E., 2002. Forest Dynamics and Disturbance Regimes: Studies from Temperate Evergreen-Deciduous Forests. Cambridge University Press, New York.

Froese, R.E., Robinson, A.P., 2007. A validation and evaluation of the Prognosis individual-tree basal area increment model. Can. J. For. Res. 37, 1438–1449.

Guertin, P.J., Ramm, C.W., 1996. Testing Lake States TWIGS: five-year growth projections for upland hardwoods in Northern Lower Michigan. North. J. Appl. For. 13, 182–188.

Hahn, J.T., 1984. Tree volume and biomass equations for the Lake States. Research Paper NC-250. USDA Forest Service, North Central Forest Experiment Station, St. Paul, Minnesota.

Hahn, J.T., Leary, R.A., 1979. Potential diameter growth functions. General Technical Report NC-049. USDA Forest Service, North Central Forest Experiment Station, St. Paul, Minnesota, pp. 22–26.

Hansen, M.H., 1990. A comprehensive sampling system for forest Inventory based on an individual tree growth model. Ph.D. Dissertation. University of Minnesota, St. Paul, Minnesota.

Henderson, E., 2007. Development of state and transition model assumptions used in national forest plan revision. Presented at the 3rd Forest Vegetation Simulator Conference, February 13–15, 2007, Ft. Collins, Colorado. USDA Forest Service, Forest Management Service Center (available from http://www.fs.fed.us/fmsc/fvs/fvs_conf_presentations.shtml [accessed 13 August 2007]).

Holdaway, M.R., 1984. Modeling the effect of competition on tree diameter growth as applied in STEMS. General Technical Report NC-094. USDA Forest Service, North Central Forest Experiment Station, St. Paul, Minnesota.

Holdaway, M.R., 1985. Adjusting STEMS growth model for Wisconsin forests. Research Paper NC-267. USDA Forest Service, North Central Forest Experiment Station, St. Paul, Minnesota.

Holdaway, M.R., 2000. The AFIS tree growth model for updating annual forest inventories in Minnesota. In: Hansen, M.H., Burk, T.E. (Eds.), Integrated tools for natural resources inventories in the 21st century, Proceedings of the IUFRO Conference, August 16–20, 1998, Boise, Idaho, USA. General Technical Report NC-212. USDA Forest Service, North Central Research Station, St. Paul, Minnesota, pp. 507–514.

Holdaway, M.R., Brand, G.J., 1983. An evaluation of STEMS tree growth projection system. Research Paper NC-234. USDA Forest Service, North Central Forest Experiment Station, St. Paul, Minnesota.

Holdaway, M.R., Brand, G.J. 1986. An evaluation of Lake States STEMS85. Research Paper NC-269. USDA Forest Service, St. Paul, Minnesota.

Hoover, C.M., 2007. Kane Experimental Forest carbon inventory: a case study introducing FVS carbon. Presented at the 3rd Forest Vegetation Simulator Conference, February 13–15, 2007, Ft. Collins, Colorado. USDA Forest Service, Forest Management Service Center (available from http://www.fs.fed.us/fmsc/fvs/fvs_conf_presentations.shtml [accessed 13 August 2007]).

Kimmins, J.P., Scoullar, K.A., Mailly, D., 2004. Models and their role in ecology and resource management. In: Kimmins, J.P. (Ed.), Forest Ecology: A Foundation for Sustainable Forest Management and Environmental Ethics in Forestry. Prentice Hall, New Jersey.

Lappi, J., Bailey, R.L., 1987. Estimation of the diameter increment function or other tree relations using angle-count samples. For. Sci. 33, 725–739.

Lessard, V.C., 2000. Calibration of the STEMS diameter growth model using FIA data. In: Hansen, M.H., Burk, T.E. (Eds.), Integrated tools for natural resources inventories in the 21st century, Proceedings of the IUFRO Conference, August 16–20, 1998, Boise, Idaho, USA. General Technical Report NC-212. USDA Forest Service, North Central Research Station, St. Paul, Minnesota, pp. 525–532.

Lumely, T., McRoberts, R.E., Holdaway, M.R., 2001. Diameter growth models using Minnesota Forest Inventory and Analysis data. For. Sci. 47, 301–310.

Lumely, T., 2004. Analysis of complex survey samples. J. Stat. Soft. 9 (8), 1–19 (available at http://www.jstatsoft.org/v09/i08/paper.pdf [accessed 13 August 2007]).

McRoberts, R.E., 2001. Imputation and model-based updating techniques for annual forest inventories. For. Sci. 47, 322–330.

McRoberts, R.E., Holdaway, M.R., Lessard, V.C., 2000. Comparing the STEMS and AFIS growth models with respect to the uncertainty of predictions. In: Hansen, M.H., Burk, T.E. (Eds.), Integrated tools for natural resources inventories in the 21st century, Proceedings of the IUFRO Conference, August 16–20, 1998, Boise, Idaho, USA. General Technical Report NC-212. USDA Forest Service, North Central Research Station, St. Paul, Minnesota, pp. 539–548.

Miner, C.L., Walters, N.R., Belli, M.L., 1988. A guide to the TWIGS program for the North Central United States. General Technical Report NC-125. USDA Forest Service, North Central Forest Experiment Station, St. Paul, Minnesota.

Overton, W.S., Stehman, S.V., 1995. The Horvitz-Thompson theorem as a unifying perspective for probability sampling: with examples from natural resource sampling. Am. Stat. 49, 261–268.

R Development Core Team, 2007. R: a language and environment for statistical computing. R Foundation for Statistical Computing (available from http://www.r-project.org), Vienna, Austria.

Reed, D.D., Jones, E.A., Tomé, M., Araújo, M.C., 2003. Models of potential height and diameter for *Eucalyptus globulus* in Portugal. For. Ecol. Manage. 172, 191–198.

Reed, D.D., Tomé, M., Araújo, M.C., Jones, E., 2001. A re-examination of potential-modifier dimensional growth models. In: Rennolls, K. (Ed.), Proceedings of IUFRO 4.11 Conference on Forest Biometry, Modelling and Information Science. University of Greenwich, London (available at http://cms1.gre.ac.uk/conferences/iufro/proceedings/Reed2potmod.pdf [accessed 13 August 2007]).

Robinson, A.P., Duursma, R.A., Marshall, J.D., 2005. A regression-based equivalence test for model validation: shifting the burden of proof. Tree Physiol. 25, 903–913.

Robinson, A.P., Ek, A.R., 2000. The consequences of hierarchy for modeling in forest ecosystems. Can. J. For. Res. 30, 1837–1846.

Robinson, A.P., Froese, R.E., 2004. Model validation using equivalence tests. Ecol. Model. 176, 349–358.

Schmidt, T.L., Spencer, J.S.J., Bertsch, R. 1997. Michigan's forests, 1993: an analysis. Resource Bulletin NC-179. USDA Forest Service, North Central Forest Experiment Station, St. Paul, Minnesota.

Smith-Mateja, E.E., Ramm, C.W., 2002. Validation of the forest vegetation simulator growth and mortality predictions on red pine in Michigan. In: Crookston, N.L., Havis, R.N. (Eds.), Proceedings of Second Forest Vegetation Simulation Conference, 12–14 February, 2002, Fort Collins, Colorado, USA. USDA Forest Service, Proceedings RMRS-25, pp. 38–44.

Smith, M., 2007. Carbon market may offer opportunities for forest landowners. The Forestry Source (available at http://www.mfpp.org/Data/carbonmarket.forestry_source_2007.pdf [accessed 13 August 2007]).

Smith, W.B., 1983. Adjusting the STEMS regional forest growth model to improve local predictions. Research Note NC-297. USDA Forest Service, North Central Experiment Station, St. Paul, Minnesota.

Stage, A.R., Wykoff, W.R., 1998. Adapting distance independent forest growth models to represent spatial variability: effects of sampling design on model coefficients. For. Sci. 44, 224–238.

USDA Forest Service, 1979. A generalized forest growth projection system: applied to the Lake States region. General Technical Report NC-049. North Central Forest Experiment Station, St. Paul, Minnesota.

USDA Forest Service, 1986. Land and Resource Management Plan: Hiawatha National Forest. USDA Forest Service, Escanaba, Michigan (available from http://www.fs.fed.us/r9/hiawatha/revision/current_plan/welcome_plan.html [accessed 13 August 2007]).

USDA Forest Service, 1991. North Central Region Forest Inventory and Analysis: Field Instructions. North Central Research Experiment Station, St. Paul, Minnesota.

USDA Forest Service, 2005a. Forest Inventory and Analysis North Central National Core Field Guide, Volume 1: Field Data Collection Procedures for Phase 2 plots, Version 2. USDA Forest Service, North Central Research Station, St. Paul, Minnesota.

USDA Forest Service, 2005b. Forest Vegetation Simulator Bulletin. Forest Management Service Center, USDA Forest Service, Fort Collins, Colorado (available at http://www.fs.fed.us/fmsc/ftp/fvs/docs/bulletins/614-100405.txt [accessed 13 August 2007]).

USDA Forest Service, 2006. Hiawatha National Forest 2006 Forest Plan. USDA Forest Service, Milwaukee, Wisconsin (available at http://www.fs.fed.us/r9/hiawatha/revision/2006/ForPlan.pdf [accessed 13 August 2007]).

Vanclay, J.K., 1994. Modelling Forest Growth and Yield: Application to Mixed Tropical Forest. CAB International, Wallingford, UK.

Wellek, S., 2003. Testing Statistical Hypotheses of Equivalence. Chapman and Hall, London.