forests

MDPI

Article

Data Assimilation in Forest Inventory: First Empirical Results

Mattias Nyström [1,*], Nils Lindgren [1], Jörgen Wallerman [1], Anton Grafström [1], Anders Muszta [1], Kenneth Nyström [1], Jonas Bohlin [1], Erik Willén [2], Johan E. S. Fransson [1], Sarah Ehlers [1], Håkan Olsson [1] and Göran Ståhl [1]

Received: 15 September 2015; Accepted: 4 December 2015; Published: 11 December 2015
Academic Editor: Joanne C. White

[1] Department of Forest Resource Management, Swedish University of Agricultural Sciences, 90183 Umeå, Sweden; nils.lindgren@slu.se (N.L.); jorgen.wallerman@slu.se (J.W.); anton.grafstrom@slu.se (A.G.); anders.muszta@slu.se (A.M.); kenneth.nystrom@slu.se (K.N.); jonas.bohlin@slu.se (J.B.); johan.fransson@slu.se (J.E.S.F.); sarah.ehlers@slu.se (S.E.); hakan.olsson@slu.se (H.O.); goran.stahl@slu.se (G.S.)
[2] Skogforsk, 75183 Uppsala, Sweden; erik.willen@skogforsk.se
* Correspondence: mattias.nystrom@slu.se; Tel.: +46-90-786-83-16

Abstract: Data assimilation techniques were used to estimate forest stand data in 2011 by sequentially combining remote sensing based estimates of forest variables with predictions from growth models. Estimates of stand data, based on canopy height models obtained from image matching of digital aerial images at six different time-points between 2003 and 2011, served as input to the data assimilation. The assimilation routines were built on the extended Kalman filter. The study was conducted in hemi-boreal forest at the Remningstorp test site in southern Sweden (lat. 13°37′ N; long. 58°28′ E). The assimilation results were compared with two other methods used in practice for estimation of forest variables: the first was to use only the most recent estimate obtained from remotely sensed data (2011) and the second was to forecast the first estimate (2003) to the endpoint (2011). All three approaches were validated using nine 40 m radius validation plots, which were carefully measured in the field. The results showed that the data assimilation approach provided better results than the two alternative methods. Data assimilation of remote sensing time series has been used previously for calibrating forest ecosystem models, but, to our knowledge, this is the first study with real data where data assimilation has been used for estimating forest inventory data. The study constitutes a starting point for the development of a framework useful for sequentially utilizing all types of remote sensing data in order to provide precise and up-to-date estimates of forest stand parameters.

Keywords: data assimilation; extended Kalman filter; forestry; image matching; photogrammetric point clouds; digital aerial images; forest inventory

1. Introduction

Accurate information about forest stands is one of the keys to successful forest management and for efficient wood supply to the forest industry. Modern planning tools, such as the Heureka system [1], can be applied to support decision making in forestry. These tools rely on accurate information about the stands in the target forest area and have the potential to provide solutions to the spatiotemporal planning problem that go beyond what can be achieved by human intuition [2].

Traditionally in many countries, stand-level forest information has been collected through recurrent campaigns. That is, every 10–20 years data have been collected in the field from all stands in a forest holding in order to update the information. However, the transfer to digital

databases in combination with the increasing flow of low-cost data that are becoming regularly provided through remote sensing motivates a shift towards continuous updating. Several new remote sensing techniques are emerging, and we are entering an unprecedented era of information richness. Modern 3D remote sensing techniques like point clouds from laser scanning [3] and stereo-view image matching of digital aerial images (denoted "image matching") [4] are being rapidly developed and applied. Digital aerial images are regularly acquired in many countries, and it has recently been shown that point clouds from aerial images can be used to estimate canopy heights and correlated variables with good accuracy [5–8]. In a similar way, accurate forest height estimates can be obtained frequently from interferometric SAR data [9,10] as well as from radargrammetry [11]. This remote sensing development might imply a paradigm shift regarding how information is collected and compiled for purposes of forest management planning.

Data assimilation can be used to continuously combine models and new sensor data in an optimal way, offering great potential for making use of new sources of forest information. Existing information about a forest area is forecasted using a model that provides an estimate for the date of the next data acquisition and an estimate of the precision of the forecasted information. Thus, the precision of the forecasted information can be compared with the precision of the new information. In the assimilation step, the two sources of information are combined through weights that are inversely proportional to their uncertainties (Figure 1). The combined estimate is then forecasted to the time-point of the next data acquisition, and repeated when new data become available.

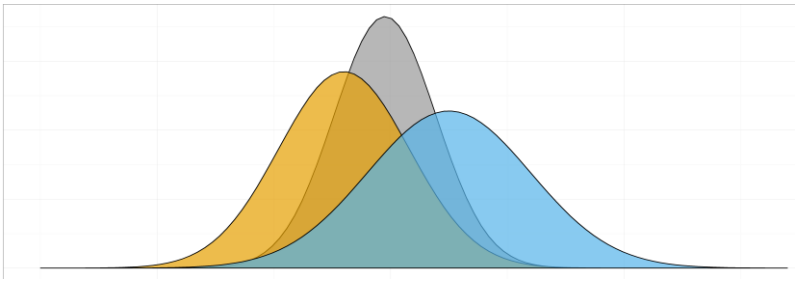

Figure 1. An illustration of the basic principle of data assimilation applied to a Gaussian model and estimate. The figure shows how the prior forecast (the center of the orange distribution) is updated to the posterior forecast (the center of the grey distribution) when a new estimate (the center of the blue distribution) is taken into account. Notice that the grey distribution is narrower than the orange distribution, indicating that the posterior forecast is more precise (*i.e.*, the estimate has lower variance) than the prior forecast.

The success of data assimilation in areas such as meteorology is well documented [12]. Data assimilation of time series of satellite data has also been used in research studies for calibration of physical models of ecosystem functions [13]. Such models typically produce estimates of ecosystem variables such as gross primary production, net primary production and ecosystem respiration [14]. Data assimilation techniques have also been proposed for operational forest inventory [15–18], but we are not aware of any previous data assimilation studies where estimates like those usually produced and employed for forest management planning have been made using real data.

However, in order to realize the potential in the context of forest inventory and management, the data assimilation techniques need to be adapted to this field of application. This involves development of new growth models from which not only growth predictions can be made but also the uncertainty of those models can be ascertained. All new estimates from remote sensing or field surveys should be used to adjust the forecasted forest information to the extent motivated by the uncertainties of the new estimates as compared to the uncertainties of the forecasts. Thus, through data assimilation, forest planning would benefit from having up-to-date and accurate information

available as input to its different planning processes, from long-term strategic planning to short-term optimization of the value chains linking forests with forest industry.

Two common approaches to data assimilation are Kalman filtering [19] and Bayesian methods [20]. Kalman filters assume that the errors of the forecasts and the new measurements based on sensor data are normally distributed whereas Bayesian methods can handle any distribution of errors. In the basic setting, the Kalman filter assumes linear forecasting models, so that the variance of forecasted variables can be calculated without approximation. However, several further developments of the Kalman filter are available as well, such as the extended Kalman filter which uses a Taylor approximation to linearize non-linear prediction models. Bayesian methods assume the true state to be a random variable and predict entire joint probability distributions of the variables of interest.

In a previous simulation study made by our research group [15], several challenges for applying data assimilation to forest information were identified. These included non-linear growth models, temporal correlation of errors from growth models, poorly known uncertainty of estimates, spatiotemporally correlated inventory errors, and the need in some cases to handle discrete data, such as individual trees within description units. Thus, any future system for data assimilation in forestry would need to be developed through a series of research studies where the different challenges are addressed.

The objective of the present study was to present our first empirical results of the application of data assimilation to forest stand data. In our previous study [15], the results were based on theoretical assumptions and the article provided case examples of the potential benefits of applying data assimilation. In the present study, we applied the data assimilation technique to empirical estimates based on point clouds from image matching from six time-points obtained over an eight year period (2003–2011) calibrated with forest estimates from circular field plots. Estimates were compiled for three variables: stem volume (V), basal area (BA) and Lorey's mean height (H_L). Data assimilation using the extended Kalman filter was compared to two methods used in practice for estimation of forest variables: (i) estimates of the target variables using point clouds from image matching from the most recent time-point (denoted "most recent estimate") and (ii) forecasting the stand development with growth models from the initial state estimate (denoted "forecast").

2. Materials and Methods

2.1. Study Area

The study was carried out at the forest estate Remningstorp in southwestern Sweden (lat. 13°37′ N; long. 58°28′ E). This 1500 ha estate is covered primarily by well-managed, productive forest. The forest is dominated by Norway spruce (*Picea abies*) and Scots pine (*Pinus sylvestris*), with some deciduous forest of mainly birch (*Betula pendula* and *Betula pubescens*). Generally, the area is rather flat, with ground elevations ranging from 120 to 145 m above sea level.

2.2. Field Data

In this study, field data were applied for three different purposes: (i) for developing models linking the remotely sensed data with the ground conditions for the target variables (V, BA, H_L), (ii) for developing growth models, and (iii) for validating the results at the endpoint of the data assimilation period.

For the development of models estimating the state of the target variables from the remote sensing data, we used field data from sample plots (denoted "training plots") distributed across the Remningstorp study area. Training plots were available from two different field campaigns during the study period (data set 1 and 2 in Figure 2). All trees greater than 4 cm diameter at breast height on the plots were calipered, and a sub-sample of the trees was selected for measurements of height and age. Site index (SI) [21] was also assessed on the plots, using the Swedish system for site index

based on ground- and field-layer vegetation as indicators of site productivity. Lorey's mean height (H_L) [22], basal area (BA) and stem volume (V) [23] were calculated for the plots using all trees greater than 4 cm in diameter at breast height. The stem volume was defined as the total volume of the entire stem above the stump, including bark but excluding branches. In Table 1, a summary of the training plot data available for the development of remote sensing based models is given.

Table 1. Field inventories of training plots used for the development of predictive models estimating the state of the target variables from the remote sensing data.

No	Inv. Year	No of Plots	Radius (m)	V (m^3/ha) min/mean/max	BA (m^2/ha) min/mean/max	H_L (m) min/mean/max
1	2004	849	10	0/277/1050	0/28/80	0/19/34
2	2010	247	10	0/202/697	0/22/60	0/16/33

V = Stem volume, BA = Basal area, H_L = Lorey's mean height, based on all calipered trees greater than 4 cm diameter at breast height.

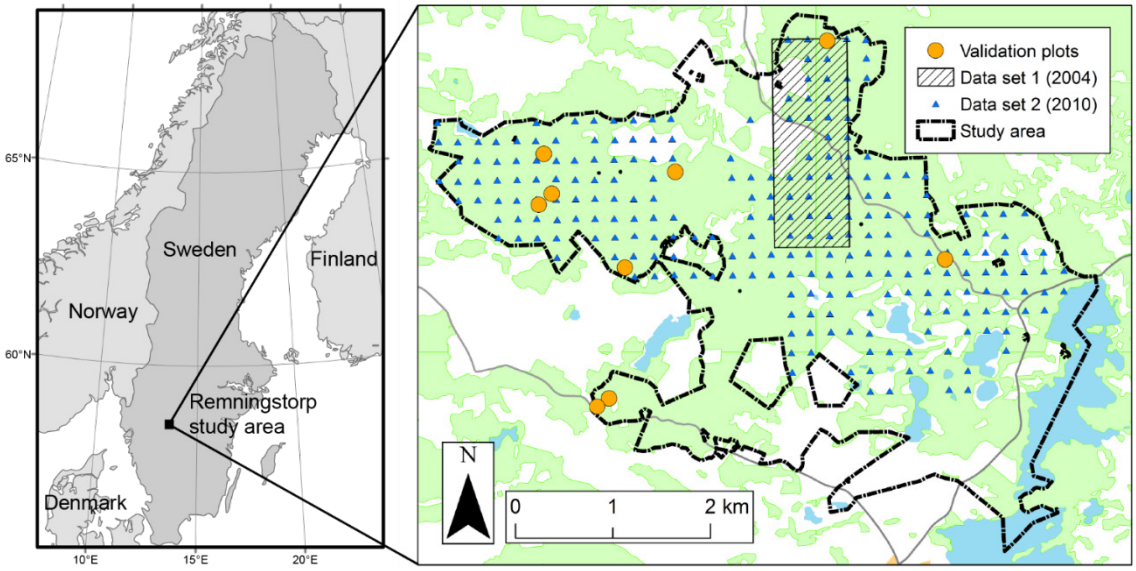

Figure 2. Overview of the study area, the location of the field inventoried training plots, and the validation plots used for assimilation. Data set 1 consists of 849 field plots located in a regular grid of 40 m within the cross-hatched area. ©Lantmäteriet I2014/00764.

The field data used for developing the growth models were derived from the Swedish National Forest Inventory (NFI) [24]. The NFI each year lays out approximately 10,000 field plots across the land area of Sweden. A majority of the plots are permanent and are revisited every fifth year. The permanent plots from the NFI were used for developing the growth models and the corresponding models for estimating the variance of the growth predictions (see Appendix A).

The third set of field data was the large plots (denoted "validation plots") used for validating the results of the data assimilation and the two methods used in practice for estimating forest variables (most recent estimate and forecast). The validation plots were selected from a larger dataset of 52 plots with 40 m radius whose positions had initially been located in the interior of fairly homogeneous stands. After having removed all plots that had been subject to cutting or major damages during the study period, as well having omitted the only multi-storied stand, nine plots remained. The nine validation plots (Table 2, Figure 2) were classified into four different forest type classes for which different growth models based on the Swedish NFI data were used. The forest type class was based on the conditions at the time-point of validation (2011). The composition was recorded as spruce stand if more than 65% of the total stem volume was Norway spruce, pine stand if more than 65% of

the total stem volume was Scots pine, mixed conifer stand if more than 65% of the total stem volume was conifer, and mixed stand if the total stem volume of broadleaved trees was between 35 and 65%. None of the validation plots had, according to the field inventories, been subject to management (such as thinning or clear-felling) during the period from the first acquisition of images used to the last. All the validation plots (except two) were inventoried in the same growth season as the last aerial images were acquired. The two other plots inventoried during the growth season of 2012 and 2013 were back-casted to the growth season of the field inventory (2011), using the growth models developed in this study.

Table 2. Summary of the validation plots, *i.e.*, the nine large plots with 40 m radius that were used to validate the results of the data assimilation, at growth season 2011.

ID	Inv. Growth Season [1]	SI (m/100 years)	Age (years)	H_L (m)	V (m^3/ha)	BA (m^2/ha)	Forest Type Class
10	2011	25	39	21.5	389	39.3	Spruce stand
116	2011	30	55	25.3	501	43.4	Spruce stand
151	2011	31	44	19.8	283	29.3	Spruce stand
211	2011	32	41	21.6	308	30.5	Spruce stand
212	2011	32	40	20.2	241	26.4	Mixed stand
325	2011	31	43	22.2	386	37.0	Spruce stand
351	2011	31	46	20.5	265	27.1	Spruce stand
515	2012	25	26	12.1	115	20.6	Mixed stand
517	2013	30	30	13.1	185	30.8	Mixed stand

[1] Growth season when the plot was measured in field.

2.3. Remote Sensing Data

Aerial images were acquired with the Intergraph Z/I Imaging Digital Mapping Camera (DMC) system [25] operated by Lantmäteriet (Swedish National Land Survey) for six different survey campaigns with partial or complete coverage of the study site (Table 3). All datasets were acquired at 4800 m above ground level (a.g.l.) except for the 2003 dataset which was acquired at 3000 m a.g.l., resulting in a ground sampling distance of approximately 0.48 m and 0.30 m, respectively. The images were acquired with 60% along-track overlap and 30% across-track overlap, resulting in at least one stereo model for each position on the ground. The images were block triangulated using bundle adjustment and radiometrically corrected by Lantmäteriet, as part of their operational aerial image production, creating a pan-sharpened false colour composite image with 8-bit radiometric resolution.

Table 3. Digital aerial images and training plots used for modeling the target variables.

Acquisition Date	Growth Season [1]	Leaf-on/leaf-off	Altitude (m a.g.l.)	Field Ref. Data Set [2]	No. of Training Plots Used
13 October 2003	2003	leaf-on	3000	1	361
28 June 2005	2005	leaf-on	4800	1	416
26 May 2007, 3 June 2007	2006	leaf-on	4800	1	258
1 September 2009	2009	leaf-on	4800	2	214
2 May 2010	2009	leaf-off	4800	2	214
23 May 2012	2011	leaf-on	4800	2	166

[1] Refers to whether the acquisition was performed before or after the 15th of June. See section *Estimation of forest state from aerial image matching data* for details. [2] Refers to Table 1.

Image matching was performed using the SURE software [26,27] to produce point cloud data for each dataset. SURE generates a height value for each point using cost-based matching, similar to the semi-global matching method [28]. All possible stereo pairs were used, so that both along-track and cross-track stereo images were incorporated. The point clouds from all stereo-matched image

pairs were merged into one point cloud with a point density ranging from 2 to 46 points/m^2 (mean 8 points/m^2) depending on the amount of overlap, the height of acquisition, and object occlusion within the 10 m radius training plots. Finally, the height values of the point cloud were transformed from height above sea level to height above ground level by subtracting a digital terrain model (DTM). The DTM used to normalize the photogrammetric point cloud was created from airborne laser scanning (ALS) data acquired using a Leica ALS 60/23 on April 21, 2011 under leaf-off conditions. The DTM used was the national DTM from Lantmäteriet; its grid size is 2 m and its vertical accuracy (RMS) is 0.2 m [29] and the density of the ALS point cloud used to create the DTM was 0.7 last returns/m^2.

In this study, an area based approach [30] was used. Metrics describing the point cloud data, such as height distribution and spatial density characteristics, were calculated for every training plot using Fusion software [31]. For the validation plots, raster of each metric was calculated using a grid cell size of 18 m × 18 m, which corresponds to the size of training plot *i.e.*, 314 m^2.

2.4. Estimation of Forest State from Aerial Image Matching Data

The state of each validation plot was estimated for each image acquisition using the corresponding raster with metrics. Estimation models were trained with training plots, namely data sets 1 and 2 described in Table 1. In order to obtain as good temporal matches as possible, sample plot data were either fore- or back-casted for short time periods to correspond to the time-points of the image acquisitions. This was done using growth models in the forestry planning and analysis system Heureka [1,32]. In southern Sweden, the growth of tree-shoots occurs mainly in June. Therefore, if the images were acquired after the 15th of June, the growth for that particular year was included when the field reference data were computed. For example, the last images (acquired the 23rd of May 2012) will be defined as belonging to growth season 2011, as they were acquired before the breaking point (15th of June). Therefore, we consistently denote the last estimate to be from year 2011 in this study to avoid confusion. Plots that had been subject to logging operations in the time between the field inventory and the remote sensing acquisition were identified through management operations registers provided by the land manager. Affected plots were excluded from the training data, thus resulting in a different number of training plots used for the different acquisitions. Linear regression was applied for modeling Lorey's mean height (H_L) and non-linear regression for modeling stem volume (V) and basal area (A), using the aerial imagery metrics as predictors. First, model expressions and metrics were chosen using studies of correlation and regression residuals of the reference data surveyed in 2010 and metrics from the aerial images acquired in 2010 (Equations (1)–(3)).

$$H_L = \beta_0 + \beta_1 P95 + \beta_2 D + \varepsilon \quad (1)$$

$$V = \exp\left(\beta_3 + \beta_4 P95 + \beta_5 \ln\left(P95\right)\right) + \varepsilon \quad (2)$$

$$BA = \exp\left(\beta_6 + \beta_7 P30 + \beta_8 \ln\left(P95\right)\right) + \varepsilon \quad (3)$$

where P30 and P95 are the 30th and 95th percentiles of the height distribution of image matching points at each plot and D is the ratio of matched points with heights larger than 2 m above ground to all matched points at each plot. Second, the parameters β were then re-estimated for each remaining time-point of aerial images and corresponding reference data. Third, the state of H_L, V and BA for each of the raster elements in the validation plots was estimated at each time-point using the developed models (Equations (1)–(3)) and, for each time-point, the corresponding data and parameter estimates. The residual variances of V and BA were not constant, rather there were increasing trends with respect to the predicted values. As accurate error variance estimates are crucial in the assimilation model, predictive models of the error standard deviations for V and BA and each time-point were developed using linear regression. In other words, models predicting the residual standard deviation from the predicted value of each respective variable were developed.

These models were fitted using data of standard deviations calculated for residuals of ten equal intervals of predicted values.

2.5. Data Assimilation

In this study, we applied the extended Kalman filter [19] for the data assimilation. Only univariate models were used. For simplicity, influential factors such as SI, tree species classification and age were assumed to be measured without error to simplify the modeling. We limited the study in this way in order to more easily interpret causes and effects as well as to compare the usefulness of applying data assimilation for estimation of the three different target variables.

Data assimilation was used for modeling the target variables over time using remote sensing data for adjusting the development given by the growth model. Figure 3 shows a flow chart describing the forecasting and assimilation steps in this study. The notations in this article are the same as in [15].

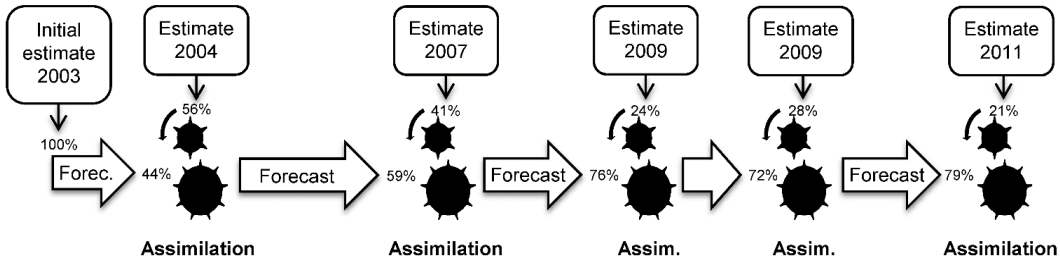

Figure 3. Illustration of the data assimilation procedure used in this study. The years state the growth season for each remote sensing data acquisition. The percentage is an example of the Kalman Gain, *i.e.*, the amount of information that is included from each source when assimilating. The Kalman Gain value is different for each 18 m × 18 m pixel within the validation plot. There were two acquisitions with growth season 2009 and, therefore, no forecast of the growth was needed between these two.

The development over time of the target variable can thus be formulated as:

$$X_t = f(x_{t-1}, W_t, t-1) = x_{t-1} + g(x_{t-1}, t-1) + W_t \quad (4)$$

where the random variable X_t denotes the state of the target variable at time t. The model describes the conditional distribution of X_t, given that the observed value of the variable at time $t-1$ was x_{t-1}. Thus, the forecasted value at time t was given by the previous value, to which the expected growth $g(x_{t-1}, t-1)$ and a random term W_t) were added. The growth models were developed through regression analysis. Details about the methods used for developing the growth models, and the corresponding model outputs, are given in Appendix A.

The growth was estimated for a five-year period; predictions for any shorter periods were obtained through computing a share corresponding to the share of the length of the prediction period. Model precision was assessed through analyzing the variance of the residual errors. It was found that the residuals were heteroscedastic and thus a separate model was applied for estimating the variance of the predictions. This was made by dividing the dataset into groups based on the predicted values and deriving a simple linear regression model for the variance prediction (see Table A3). We assumed the random term W_t to be normally distributed with zero mean and a variance q_t^2 that is dependent on the previous state as well as on time; $q_t = \alpha + \beta g(x_{t-1}, t-1)$, where $t \geqslant 1$ and α and β are parameters used for estimating the standard deviation of the residual errors. The parameters were different for the three target variables and for different forest type classes.

The deviations between the state of the target variable and estimates of it—based on remotely sensed data—were assumed to be independent normally distributed random variables. These normal distributions were assumed to have zero mean and time-dependent variances according to:

$$Z_t = x_t + V_t \tag{5}$$

where Z_t is the (random) estimator, x_t is the true state at time-point t, and V_t is a random deviation with zero mean and variance, r_t^2, which is estimated from the residual errors obtained in connection with the development of estimators for the target variables based on the remotely sensed data, and $\widetilde{z}_t$ is used as notation for the actual value observed of Z_t. Similarly, $\widetilde{x}_t$ is used as notation for the actual value obtained through the growth predictions, *i.e.*, $\widetilde{x}_t = f(\hat{x}_{t-1}, 0, t-1)$, where $\hat{x}_{t-1}$ is the assimilated variable at time-point $t-1$; its variance is denoted p_{t-1}^2. In case no estimates are made at time-point t, the assimilated variable $\hat{x}_t$ will obtain the value $\widetilde{x}_t$ and the variance $\widetilde{p}_t^2 = a_t^2 p_{t-1}^2 + q_t^2$. This is a consequence of the use of first order Taylor linearization, in connection with the extended Kalman filter, to linearize the growth model in order to compute the variance. Thus, $a_t = \dfrac{d}{dx} f(\hat{x}_{t-1}, 0, t-1)$, or in other words, a_t is the partial derivative of the growth model with respect to the target variable.

In case an estimate is made at time-point t, the forecast and the estimate are weighted inversely proportional to the variances to obtain the assimilated variable, *i.e.*,

$$\hat{x}_t = (1 - K_t)\,\widetilde{x}_t + K_t \widetilde{z}_t \tag{6}$$

with the Kalman gain $K_t = \dfrac{\widetilde{p}_t^2}{\widetilde{p}_t^2 + r_t^2}$. If the estimator (Equation (5)) has considerable variance then the Kalman gain becomes almost zero, implying that the estimate does not contribute to the assimilation. On the other hand, if the estimator is very precise, then the Kalman gain becomes almost 1, implying that the forecast does not contribute much to the assimilation. Equation (6) is based on the assumption that $\widetilde{x}_t$ and $\widetilde{z}_t$ are independent. Further, the variance of the assimilated variable is:

$$p_t^2 = (1 - K_t)\,\widetilde{p}_t^2. \tag{7}$$

The assimilation steps were conducted for each raster element (18 m × 18 m). For purposes of comparison, we also estimated the state of the validation plots based on two methods used in practice for estimation of forest variables: use of only the most recent remotely sensed data trained with simultaneously surveyed training plots (Table 1, data set 2), and forecasting of the first estimate using growth models. Both these methods and the assimilation were performed on single raster cells and the mean value of the raster cells with its center within the validation plots was calculated for the year of validation (2011). Figure 4 shows a flowchart of the implementation in the code of the data assimilation framework.

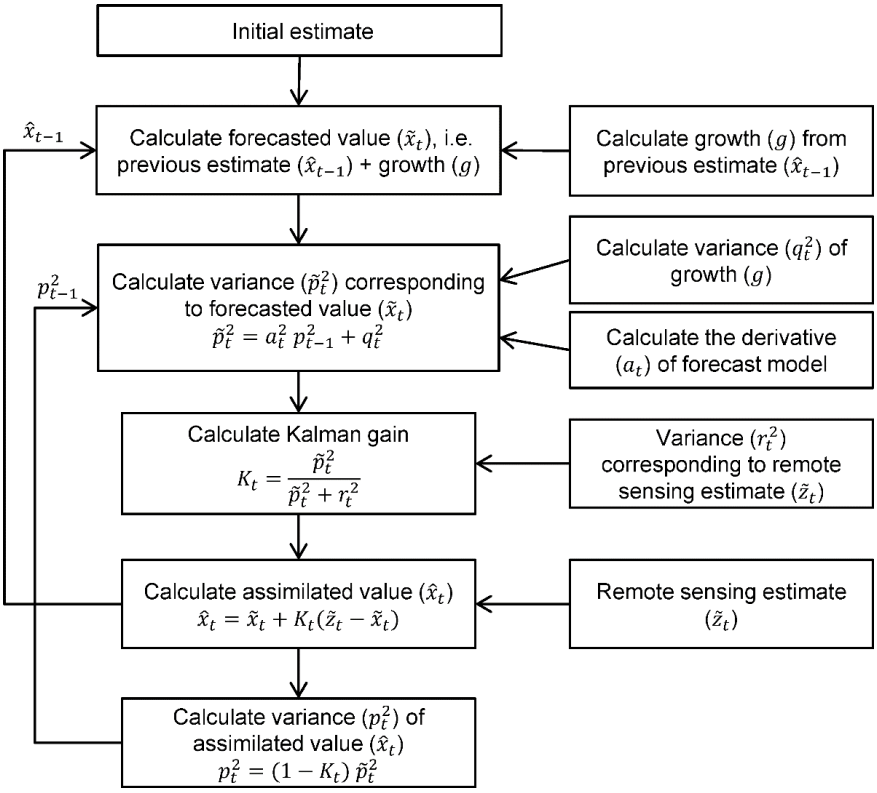

Figure 4. Flowchart of the implementation of the data assimilation framework.

Validation of the three methods (assimilation, most recent estimate, and forecast) was made by calculating the deviation (e_i) from the field measured value for the nine validation plots at the last time-point. For each method, the deviations were calculated by subtracting the field measured value from the plot mean value of the intersecting raster cells. In addition, the root mean squared error ($RMSE$) and mean deviation (MD) was calculated for the three target variables and methods as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \tag{8}$$

$$MD = \frac{1}{n}\sum_{i=1}^{n} e_i \tag{9}$$

3. Results and Discussion

The initial state (2003) was estimated from point clouds obtained from image matching. The validation was made with field data for the 40 m radius validation plots reflecting the state in 2011. Figure 5 shows the deviation from the field measured value for each of the three methods (*i.e.*, assimilation, most recent estimate, forecasting) for the nine validation plots. Table 4 shows the $RMSE$ of the deviation from the field measurements and Table 5 shows the corresponding mean deviation (MD). A positive MD means that the plot value is on average overestimated. It can be seen that the $RMSEs$ are smaller using data assimilation compared to forecasting the value from the first (initial state) remote sensing prediction. The full strength of data assimilation will probably first be seen when data from multiple sources are combined. For example, if data are first acquired using airborne laser scanning (ALS) and later acquired with a technique that has lower accuracy, we will be able to update the high accuracy acquisition from the ALS with new data. In a similar way,

the data assimilation framework can be used for maintaining information from earlier high quality measurements, for example from field visits, and combining it, rather than replacing it, with new information from remote sensing.

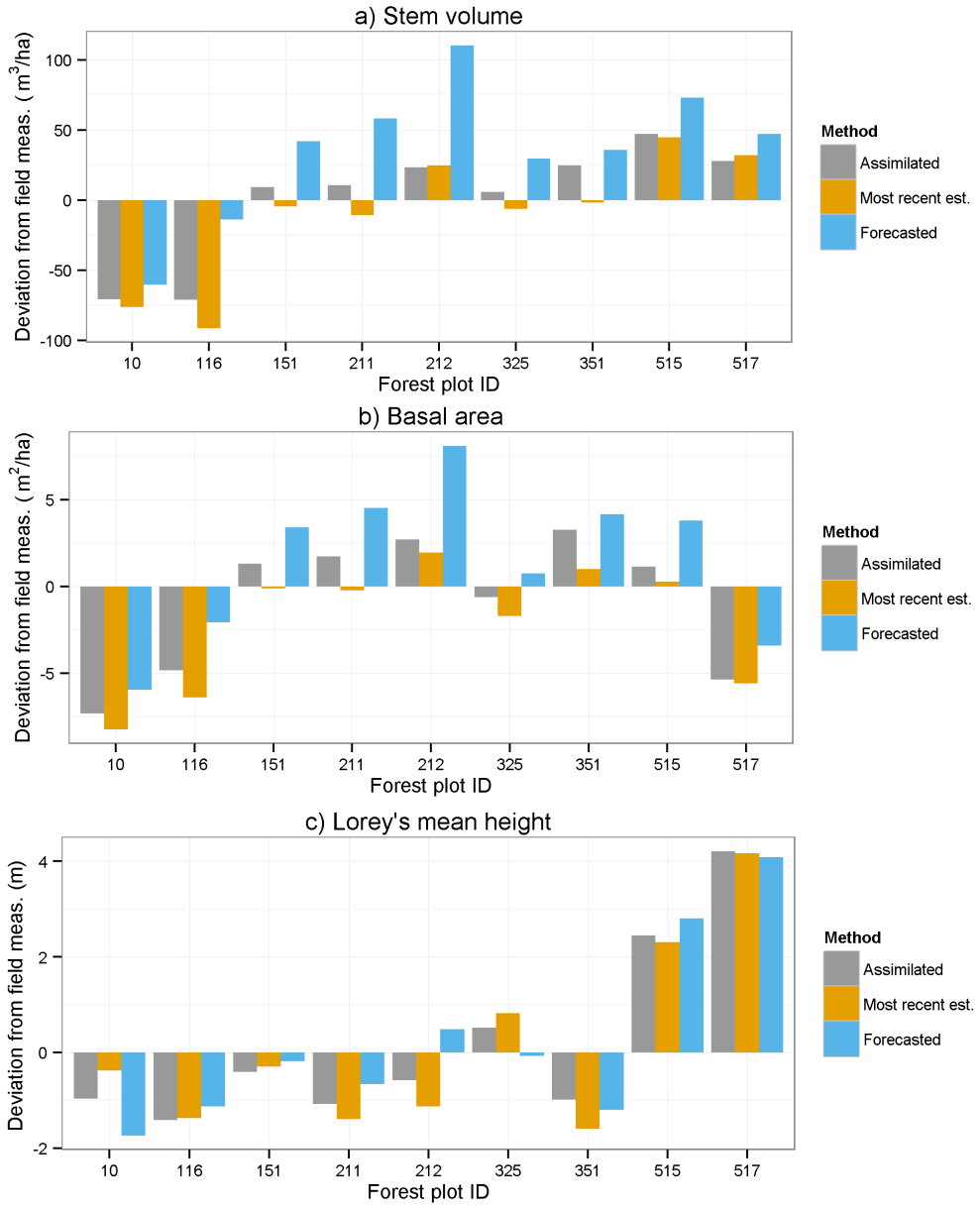

Figure 5. Deviation at year 2011 between field measurement and estimates of (**a**) stem volume; (**b**) basal area, and (**c**) Lorey's mean height using different methods. A positive value means that the plot has an overestimated value.

Table 4. Root mean squared error (*RMSE*) of the deviation from the field measurement 2011 for the nine assimilated plots. In parentheses is relative *RMSE*.

Target Variable	Assimilated	Most Recent Estimate	Forecasted
V	40.1 (13.5%)	44.7 (15.0%)	58.5 (19.7%)
BA	3.80 (12.0%)	4.05 (12.8%)	4.49 (14.2%)
H_L	1.81 (9.3%)	1.86 (9.6%)	1.86 (9.5%)

Table 5. Mean deviation (*MD*) from the field measurement 2011 for the nine assimilated plots. A positive *MD* means that the value is on average overestimated compared to the field measurements. In parentheses is relative *MD*.

Target Variable	Assimilated	Most Recent Estimate	Forecasted
V	0.73 (0.3%)	−10.0 (−3.4%)	35.7 (12.0%)
BA	−0.89 (−2.8%)	−2.12 (−6.7%)	1.47 (4.7%)
H_L	0.19 (1.0%)	0.12 (0.6%)	0.27 (1.4%)

The results of the validation shows that the accuracy of estimates are similar to other studies of image matching based forest inventory [5,6]. Tables 4 and 5 show that the assimilated values on average for all stands and variables are better than the most recent estimates, as well as the forecasted estimates. The only exception is that the mean deviation of Lorey's mean height is lower when using only the most recent estimate. This might be explained by the fact that the photogrammetric point cloud contains particularly accurate information about tree height, and then the gain by data assimilation to reduce noise is less [15]. It is also evident from Figure 5 that two of the mixed stands (515 and 517) had comparatively large over estimates of the canopy height. This might be because broadleaved trees have more leaves and branches in the upper part of the canopy than cone-shaped spruces. Thus, developing separate regression models for different tree species groups might improve the results in future studies.

Figure 6 shows the growth of validation plot 212. Each line represents one raster cell within the 40 m radius validation plot. The black dots represent remote sensing based estimates. In the fifth acquisition (2 May 2010), there is one clear outlier at 14.5 m, but the assimilation remained stable despite this. The result of the last assimilation (the rightmost triangle) should be compared with the field measured value (the red cross). It should also be noticed that the acquisition dates are shown on the x-axis, but it is the growth season that determines the growth prediction, and, therefore, the slope is different between the acquisitions.

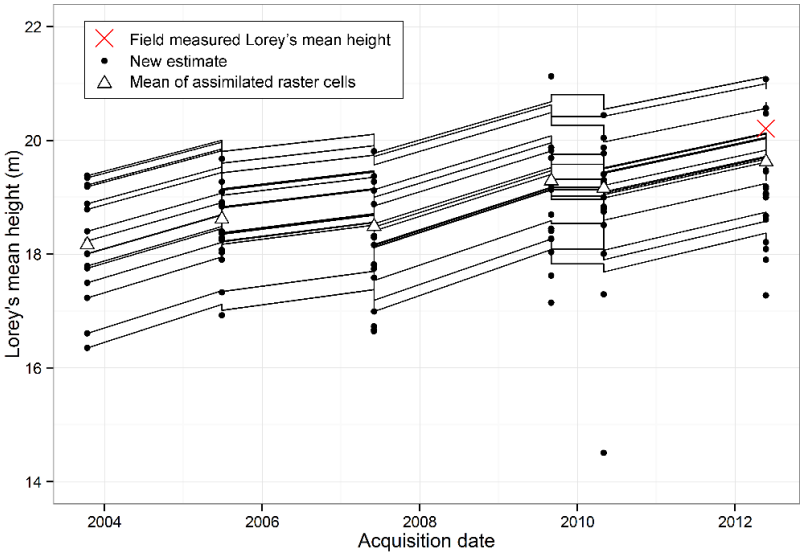

Figure 6. Data assimilation of Lorey's mean height on validation plot 212. The black dots are the estimates from remote sensing where each dot represents one 18 m × 18 m raster cell within the validation plot. The triangles are the mean value of the assimilation for each time-point. The red cross is the field measured Lorey's mean height for the same growth season (2011) as the last acquisition. The black lines are the current estimate for each raster cell. Note that the lines jump either up or down depending on the outcome of the assimilation at each acquisition time-point.

In this study, the data are all from the same remote sensing method, *i.e.*, point clouds from image matching of aerial images. Therefore, it is likely that the estimation errors are correlated, which would imply that too much weight is being put on the forecasted information as compared to the new information. This could explain why data assimilation did not perform much better than using information from the most recent estimate (Table 4). Methods to compensate for this will be developed in future studies but requires non-standard filters for the data assimilation. During the course of the study, a test was made where smaller weights were consistently given to the forecasted values in the assimilation step as compared to the weights assigned by the Kalman filter. This led to improved assimilation results, which indicates that the issue of correlated errors needs to be more carefully addressed in future studies.

In a future version of the data assimilation application, the models for prediction of the stand attributes ought to be evaluated in a simultaneous setting to handle cross-correlated errors across models. However, in the present study, we analyzed the different growth characteristics separately over a short time period using simple growth models where the non-static predictors (V, BA, H_L) were not used as predictor variables across the growth models to reduce cross-correlation effects in the forecasts.

The first images were from growth season 2003 and the last from 2011. Thus, the assimilation period is rather short (eight years) considering the time perspective in Nordic forestry where the rotation period is typically 80–100 years. In an operational case, we would have a model that is continuously updated when new data become available and probably spans over much longer time periods than eight years. The starting point for the assimilation is an estimate based on the first aerial images. The short time period might give fairly good growth predictions; however, since the initial state itself is an estimate, it is difficult to evaluate how the forecasts perform.

In this study, the assimilation was conducted on 18 m $\times$ 18 m raster cells. Further research is needed to investigate at what level the assimilation should be performed. An alternative could be to assimilate directly at stand level. However, the modeling units for the growth forecasts and the estimates based on remotely sensed data need to correspond, and it must be possible to acquire high quality field reference data for purposes of modeling. These issues point towards raster based approaches to data assimilation being more straightforward than stand based approaches. However, procedures need to be developed where the precision of aggregated raster elements can be estimated.

We are entering an era with a frequent flow of data from different sources and of different types. An example of this is the potential availability of several optical and radar satellite images per year, point clouds from digital photogrammetry with a few years' time interval, and laser scanning data with five to ten years' time interval. In addition, there will also be different types of field reference data, for example, from field plots, harvesters and ground based laser scanners. The data assimilation technique offers a potential method for utilizing all of these data sources, even if some of these sources would not be sufficient for use in operational forest planning when used alone. The combination of these different data sources will be the subject of further studies.

4. Conclusions

This study presents the first empirical results of data assimilation applied to the estimation of forest variables. A system for data assimilation has been developed, implemented and validated. The input to the data assimilation was canopy height models obtained from image matching of digital aerial images at six different time points during the growth season between 2003 and 2011. The study showed that data assimilation of stem volume, basal area and Lorey's mean height resulted in marginally better accuracy than estimates made from the last available aerial images, but substantially better accuracy than estimates from the first available aerial images forecasted to the endpoint. There is a strong potential to further develop the data assimilation concept. Among the benefits are that data from all relevant remote sensors can be utilized in proportion to their information content and complement each other in the assimilation. Furthermore, there is no

need for each new remote sensing data set to cover the whole area of interest. When remote sensing data are missing for an area of interest at any time point, predictions for that area will be obtained by the growth models. In a similar way, a database with field reference plots could be continuously updated with new ground truth data, for example, from harvesters, even if the whole area of interest is not covered each time. The field reference plots will be forecasted, and old plots with cutting or suspected damages according to records and/or remote sensing change detection will be omitted when the next set of remotely sensed data becomes available.

Acknowledgments: This study was financed by the Swedish Forest Society Foundation, the Swedish National Space Board and the Kempe foundation. We are thankful to Heather Reese for her comments on the manuscript. Finally, we thank the anonymous reviewers for their constructive comments.

Author Contributions: Mattias Nyström made the main part of the analysis and the writing of the article; Nils Lindgren contributed second most to the analysis and made several of the figures; Jörgen Wallerman calibrated the field data and made the remote sensing analysis; Anton Grafström and Anders Muszta contributed with the statistical framework; Kenneth Nyström made the growth functions; Jonas Bohlin created the photogrammetry point clouds; Erik Willén contributed with end-user views; Johan Fransson managed the field data collection; Sara Ehlers was the main author of the first article based on simulations, which this article with empirical data builds on; Håkan Olsson and Göran Ståhl initiated and managed the project. All authors checked and contributed to the final text.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Stand Level Growth Models for Stem Volume, Basal Area, and Mean Height Increment

The stand development (forecast) regarding stem volume (V), basal area (BA), and basal area weighted mean height (Lorey's mean height, H_L) were simulated with simple and robust growth models in the present study. The growth models were based on data from permanent sample plots (plot size 314 m^2) of the Swedish National Forest Inventory established during 1983–1987 and re-inventoried three to four times between 1988 and 2010. Stand aggregates were computed at each plot and measurement. In the present study, we only used plots with an increment period of five growth seasons and where no harvest had been performed (registered) during the increment period. The net increments were derived as the difference between estimated yield attributes (V, BA, H_L) from two consecutive inventories. The growth data were divided into four forest types according to species composition based on standing volume at initial stage (see *Field Data* for a definition of the forest types). Summary statistics of some stand variables by forest type are presented in Table A1.

Table A1. Mean and standard deviation (SD) of some stand characteristics for the different forest types used in the modeling (*n* = number of plots).

	Stand Characteristics							
	V_i	V	BA_i	BA	H_i	H_L	Age	SI
	Spruce stands (*n* = 4467)							
Mean	28.4	169	2.7	21.7	1.2	15.7	82	23
SD	23.8	118	2.8	10.6	1.6	6.1	46	6
	Pine stands (*n* = 6235)							
Mean	21.1	115	2.5	17.0	1.2	12.8	74	20
SD	18.0	89	2.6	9.7	1.4	5.5	43	4
	Mixed conifers stands (*n* = 2751)							
Mean	25.4	162	2.6	22.0	1.2	15.0	77	22
SD	20.9	111	2.8	10.9	1.5	5.6	40	5
	Mixed stands (*n* = 1678)							
Mean	22.4	105	2.8	17.2	1.4	11.6	58	21
SD	20.8	89	3.2	10.5	1.7	5.6	37	6

Note: V_i = 5 years net volume increment (m^3/ha), V = stem volume (m^3/ha), BA_i = 5 years basal area increment (m^2/ha), BA = basal area (m^2/ha), H_i = 5 years basal area weighted mean height increment (m), H_L = basal area weighted mean height (m), Age = basal area weighted stand age (years), SI = site index [21] according to site properties expressed as expected height at 100 years total age for dominant height (m).

The modeling approach used in the present study is rather straightforward and the different growth models were evaluated and estimated independent of each other. The following general exponential model was used to describe the net increment for all models (Y):

$$Y = \exp\left(b_0 + \sum b_i X_i\right) + \epsilon \quad \text{(A1)}$$

where b_0 is a constant, b_i a vector of coefficients for the independent variables (X_i) and ϵ is a random error component. The independent variables in the model were restricted to size of the current yield variable (Y), stand age (Age), site index (SI) [21], as well as their transformations and interactions among these variables.

The traditional logarithmic transformation of the dependent variable and use of linear regression is not an option in our case since we sometimes have a negative increment in our data due to mortality, unregistered cutting, measurement errors, *etc.* To remove severe outliers a preliminary model was fitted to the data and, in a second fit, we used only observations having residuals within three standard deviations, *i.e.*, approximately 1% of the data were removed from the current data set. The final parameter estimates were evaluated with ordinary nonlinear regression. The parameter estimates are presented in Table A2a–c.

Table A2a. Coefficients for the net volume growth (V_i, 5 years) functions.

	Spruce Stands		**Pine Stands**		**Mixed Conifer Stands**		**Mixed Stands**	
Variable	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**
Intercept	0.2207	0.2404	−0.6707	0.2118	0.8171	0.3111	0.3930	0.3754
V	−0.0003	0.0001	-	-	-	-	-	-
ln(V)	0.3907	0.02286	0.3706	0.0147	0.3867	0.0254	0.3867	0.0307
ln(Age)	−0.2712	0.0605	−0.2897	0.0474	−0.3721	0.0775	−0.1555	0.1095
Age	−0.0048	0.0010	−0.0046	0.0008	−0.0038	0.0012	−0.0091	0.0022
ln(SI)	0.8671	0.0579	1.1857	0.0536	0.7302	0.0784	0.6961	0.0822
ln(1 + Pns)					0.2805	0.1313		
MSE	335		200		317		302	
N (obs.)	4467		6235		2751		1678	

Table A2b. Coefficients for the net basal area growth (BA_i, 5 years) functions.

	Spruce Stands		**Pine Stands**		**Mixed Conifer Stands**		**Mixed Stands**	
Variable	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**
Intercept	0.6755	0.2641	0.3408	0.2620	0.2890	0.3559	0.4635	0.4020
BA	−0.0222	0.0024	−0.0071	0.0025	−0.0130	0.0035	−0.0140	0.0047
ln(BA)	0.3172	0.0268	0.3599	0.0342	0.1860	0.0423	0.1718	0.0411
ln(Age)	−0.5512	0.0271	−0.7863	0.0471	−0.5307	0.0379	−0.4725	0.0509
ln(SI)	0.6712	0.0667	1.0175	0.0647	0.8280	0.0963	0.7032	0.1047
BA/Age	-	-	−0.5603	0.1130				
MSE	5.8		4.8		6.4		8.6	
N (obs.)	4467		6235		2751		1678	

Table A2c. Coefficients for the basal area weighted mean height growth (H_i, 5 years) functions.

	Spruce Stands		Pine Stands		Mixed Conifer Stands		Mixed Stands	
Variable	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**
Intercept	−0.4226	0.4161	0.2966	0.3105	0.0679	0.4963	−1.6143	0.8103
H_L	−0.2156	0.0149	−0.2598	0.0145	−0.2301	0.0222	−0.1472	0.0335
ln(H_L)	0.4275	0.0551	0.4880	0.1488	0.4585	0.0876	0.2891	0.1214
Age	-		-		-		−0.0144	0.0039
ln(Age)	-		-		-		0.3900	0.1686
ln(SI)	0.4140	0.1268	0.1488	0.1035	0.2800	0.1620	0.5533	0.1883
$H_L \times SI$	0.0039	0.0004	0.0058	0.0005	0.0040	0.0007	0.0020	0.0009
MSE	1.92		1.46		1.84		2.24	
N (obs.)	4467		6235		2751		1678	

Note: V = stem volume (m^3/ha), BA = basal area (m^3/ha), H_L = basal area weighted mean height (m), Age = basal area weighted stand age (years), SI = site index [21] according to site properties expressed as expected height at 100 years total age for dominant height (m), Pns = proportion volume of Norway spruce.

Residual analysis indicates that the present functions fit the data well (e.g., Figures A1 and A2) and might be useful to update inventory registers at least for shorter periods, as done in the present study. However, the residual variances increased slightly with the size of the prediction for the volume and basal area increment model, and decreased for the mean height increment model. To be able to plug in accurate estimates of the prediction errors into the data assimilation application, the residual standard deviation (SD) was regressed as a linear function of the predicted value (see Table A3).

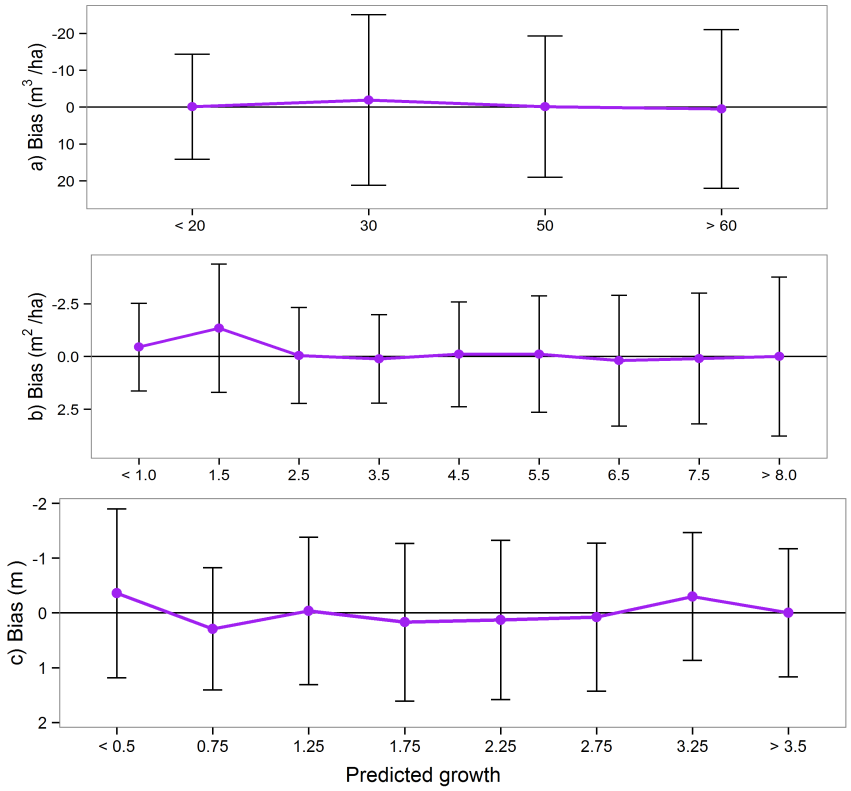

Figure A1. Mean residuals (Bias) and standard deviation (SD) in different classes of predicted growth for the different growth models, (**a**) V_i (stem volume growth); (**b**) BA_i (basal area growth), and (**c**) H_i (Lorey's mean height growth), for spruce stands.

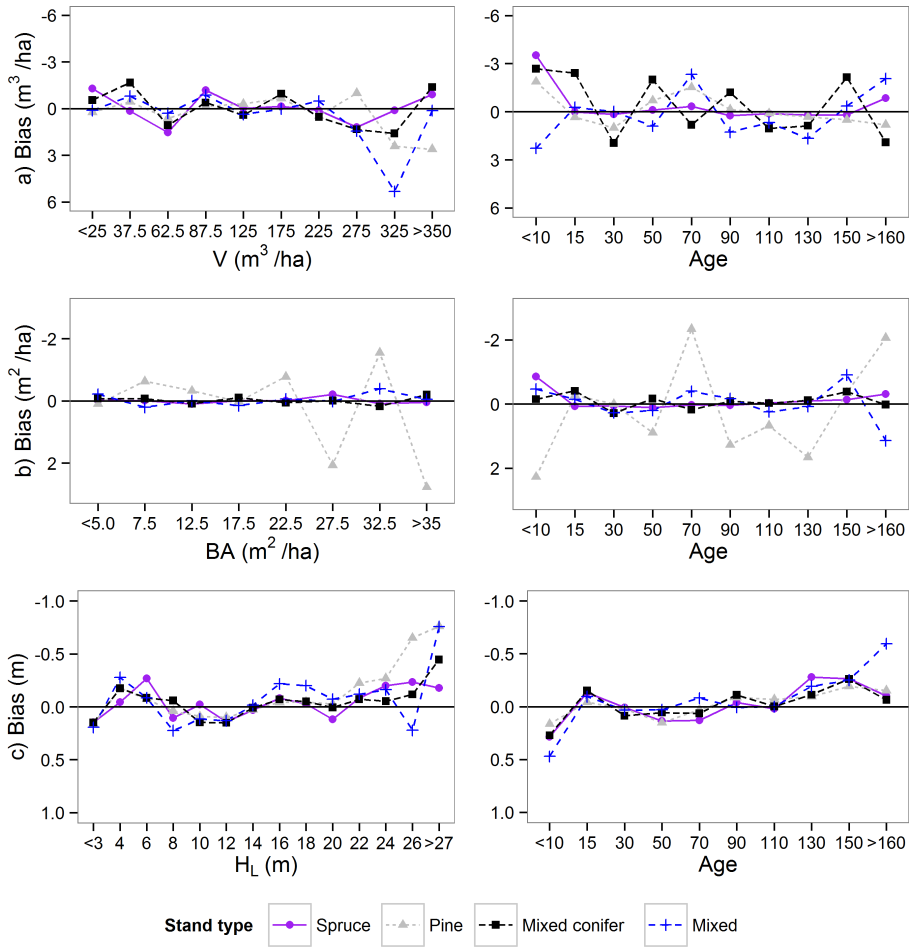

Figure A2. Mean residual (Bias) in different classes of the yield variable and stand age for the different growth models, (**a**) V_i (stem volume growth); (**b**) BA_i (basal area growth), and (**c**) H_i (Lorey's mean height growth).

Table A3. Relationships for estimations of the residual standard deviation (SD) as a function of predicted increment. $SD = \alpha + \beta\hat{Y}$.

	Spruce Stands		Pine Stands		Mixed Conifer Stands		Mixed Stands	
Function	**α**	**β**	**α**	**β**	**α**	**β**	**α**	**β**
Volume increment, V_i	13.67	0.147	8.68	0.248	14.58	0.113	11.20	0.260
Basal area increment, BA_i	1.92	0.183	1.43	0.263	1.83	0.260	2.65	0.096
Mean height increment, H_i	1.55	−0.113	1.34	−0.131	1.46	−0.103	1.71	−0.133

References

1. Wikström, P.; Edenius, L.; Elfving, B.; Eriksson, L.O.; Lämås, T.; Sonesson, J.; Öhman, K.; Wallerman, J.; Waller, C.; Klintebäck, F. The Heureka forestry decision support system: An overview. *Math. Comput. For. Nat. Resour. Sci.* **2011**, *3*, 87–94.
2. Baskent, E.Z.; Keles, S. Spatial forest planning: A review. *Ecol. Model.* **2005**, *188*, 145–173. [CrossRef]
3. Næsset, E.; Gobakken, T.; Holmgren, J.; Hyyppä, H.; Hyyppä, J.; Maltamo, M.; Nilsson, M.; Olsson, H.; Persson, A.; Söderman, U. Laser scanning of forest resources: The Nordic experience. *Scand. J. For. Res.* **2004**, *19*, 482–499. [CrossRef]

4. Leberl, F.; Irschara, A.; Pock, T.; Meixner, P.; Gruber, M.; Scholz, S.; Wiechert, A. Point Clouds: Lidar *versus* 3D Vision. *Photogramm. Eng. Remote Sens.* **2010**, *76*, 1123–1134. [CrossRef]
5. Bohlin, J.; Wallerman, J.; Fransson, J.E.S. Forest variable estimation using photogrammetric matching of digital aerial images in combination with a high-resolution DEM. *Scand. J. For. Res.* **2012**, *27*, 692–699. [CrossRef]
6. Gobakken, T.; Bollandsås, O.M.; Næsset, E. Comparing biophysical forest characteristics estimated from photogrammetric matching of aerial images and airborne laser scanning data. *Scand. J. For. Res.* **2015**, *30*, 73–86. [CrossRef]
7. Pitt, D.G.; Woods, M.; Penner, M. A comparison of point clouds derived from stereo imagery and airborne laser scanning for the area-based estimation of forest inventory attributes in boreal Ontario. *Can. J. Remote Sens.* **2014**, *40*, 214–232. [CrossRef]
8. Stepper, C.; Straub, C.; Pretzsch, H. Using semi-global matching point clouds to estimate growing stock at the plot and stand levels: Application for a broadleaf-dominated forest in central Europe. *Can. J. For. Res.* **2015**, *45*, 111–123. [CrossRef]
9. Soja, M.J.; Persson, H.; Ulander, L.M.H. Estimation of forest height and canopy density from a single InSAR correlation coefficient. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 646–650. [CrossRef]
10. Kugler, F.; Schulze, D.; Hajnsek, I.; Pretzsch, H.; Papathanassiou, K.P. TanDEM-X Pol-InSAR performance for forest height estimation. *IEEE Geosci. Remote Sens.* **2014**, *52*, 6404–6422. [CrossRef]
11. Karjalainen, M.; Kankare, V.; Vastaranta, M.; Holopainen, M.; Hyyppä, J. Prediction of plot-level forest variables using TerraSAR-X stereo SAR data. *Remote Sens. Environ.* **2012**, *117*, 338–347. [CrossRef]
12. Rabier, F. Overview of global data assimilation developments in numerical weather-prediction centres. *Q. J. R. Meteorol. Soc.* **2005**, *131*, 3215–3233. [CrossRef]
13. Loew, A. Preface: Remote sensing data assimilation special issue. *Remote Sens. Environ.* **2008**, *112*, 1257. [CrossRef]
14. Quaife, T.; Lewis, P.; de Kauwe, M.; Williams, M.; Law, B.E.; Disney, M.; Bowyer, P. Assimilating canopy reflectance data into an ecosystem model with an Ensemble Kalman Filter. *Remote Sens. Environ.* **2008**, *112*, 1347–1364. [CrossRef]
15. Ehlers, S.; Grafström, A.; Nyström, K.; Olsson, H.; Ståhl, G. Data assimilation in stand-level forest inventories. *Can. J. For. Res.* **2013**, *43*, 1104–1113. [CrossRef]
16. Dixon, B.L.; Howitt, R.E. Continuous forest inventory using a linear filter. *For. Sci.* **1979**, *25*, 675–689.
17. Czaplewski, R.L.; Alig, R.J.; Cost, N.D. Monitoring land/forest cover using the Kalman filter: A proposal. In Proceedings of the IUFRO Forest Growth Modeling and Prediction Conference, Minneapolis, MN, USA, 23–27 August 1987; pp. 1089–1096.
18. Czaplewski, R.L.; Thompson, M.T. Opportunities to improve monitoring of temporal trends with FIA panel data. In Proceedings of the Forest Inventory and Analysis (FIA) Symposium 2008, Park City, UT, USA, 21–23 October 2008; pp. 1–55.
19. Welch, G.; Bishop, G. *An Introduction to the Kalman Filter*; University of North Carolina: Chapel Hill, NC, USA, 2006.
20. Dowd, M. Bayesian statistical data assimilation for ecosystem models using Markov Chain Monte Carlo. *J. Mar. Syst.* **2007**, *68*, 439–456. [CrossRef]
21. Hägglund, B.; Lundmark, J.E. Site index estimation by means of site properties. Scots pine and Norway spruce in Sweden. *Stud. For. Suec.* **1977**, *138*, 38.
22. Söderberg, U. *Funktioner för Skogsindelning: Höjd, Formhöjd Och Barktjocklek för Enskilda Träd (Functions for Forest Management: Height, form Height and Bark Thickness of Individual Trees)*; Institutionen for skogstaxering: Umeå, Sweden, 1992.
23. Brandel, G. *Volymfunktioner för Enskilda Träd: Tall, Gran Och Björk*; Sveriges lantbruksuniversitet, Institutionen för skogsproduktion: Garpenberg, Sweden, 1990. (In Swedish)
24. Fridman, J.; Holm, S.; Nilsson, M.; Nilsson, P.; Ringvall, A.H.; Ståhl, G. Adapting National Forest Inventories to changing requirements—The case of the Swedish National Forest Inventory at the turn of the 20th century. *Silva Fenn.* **2014**, *48*, 1–29. [CrossRef]
25. Hinz, A.; Dörstel, C.; Heier, H. DMC—The digital sensor technology of Z/I-Imaging. In *Photogrammetric Week 01*; Wichmann Verlag: Heidelberg, Germany, 2001; pp. 93–103.

26. Rothermel, M.; Wenzel, K.; Fritsch, D.; Haala, N. SURE—Photogrammetric surface reconstruction from imagery. In Proceedings of the LC3D Workshop, Berlin, Germany, 4–5 December 2012; pp. 1–9.
27. Wenzel, K.; Rothermel, M.; Haala, N.; Fritsch, D. SURE—The ifp software for dense image matching. In Proceedings of the Photogrammetric Week 2013, Heidelberg, Germany, 4–6 September 2013; pp. 59–70.
28. Hirschmüller, H. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 328–341. [CrossRef] [PubMed]
29. Lundgren, J.; Owemyr, P. *Noggrannhetskontroll av Laserdata för ny Nationell Höjdmodell*; Högskolan i Gävle: Gävle, Sweden, 2010.
30. Næsset, E. Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. *Remote Sens. Environ.* **2002**, *80*, 88–99. [CrossRef]
31. McGaughey, R.J. *FUSION/LDV: Software for LiDAR Data Analysis and Visualization*; FUSION Version 3.42; USDA Forest Service, Pacific Northwest Research Station, University of Washington: Seattle, WA, USA, 2014; p. 175.
32. Fahlvik, N.; Elfving, B.; Wikström, P. Evaluation of growth models used in the Swedish forest planning system Heureka. *Silva Fenn.* **2014**, *48*. [CrossRef]