Confronting terrestrial biosphere models with forest inventory data

Jeremy. W. Lichstein, Ni-Zhang Golaz, Sergey Malyshev, Elena Shevliakova, Tao Zhang, Justin Sheffield, Richard A. Birdsey, Jorge L. Sarmiento, Stephen W. Pacala

APPENDIX A. SUPPLEMENTARY METHODS FOR FIA DATA FILTERING AND ANALYSIS, MEASUREMENT ERROR MODELS, AND LM3V OPTIMIZATION.

Estimating aboveground biomass of FIA plots

We used published allometries (Jenkins et al. 2003) to estimate the aboveground biomass (AGB) and wood growth rate of individual trees from diameter at breast height (dbh) values reported by FIA. Growth was calculated across two plot measurements, with an average remeasurement interval of about five years. To convert the AGB estimates into per-unit-area estimates (which are then summed to yield plot-level AGB values), we multiplied by tree-level expansion factors reported by FIA (equivalent to dividing by the sample area for each tree, which depends on its dbh). A "growth trees per acre" expansion factor was only available for trees with a current (second plot measurement) dbh ≥ 12.7 cm. For smaller trees, we used the previous (first plot measurement) "trees per acre" expansion factor if available (i.e., if previous dbh ≥ 2.5 cm), and we used the current "trees per acre" otherwise (i.e., ingrowth trees with a previous dbh < 2.5 cm, the smallest dbh sampled by FIA). Any errors introduced into our analysis by inaccurate expansion factors for trees with dbh < 12.7 cm should be small, since these trees comprise only a small fraction of total AGB for the stand ages considered in our analysis (> 40 years). Furthermore, such errors should be rare, because growth and non-growth expansion factors are typically the same under FIA's National Sampling Design (Bechtold and Patterson 2005).

Selection criteria for FIA plots

In addition to the criteria specified in the main-text Methods, only FIA plots that met the criteria below were included in our analysis. Note that FIA defines a Condition as a unique combination of land-use/disturbance history and edaphic conditions within an inventory plot.
(1) Plot has only one reported stand age (i.e., if multiple Conditions are present, they must all have same stand age).
(2) The sum of all measured Condition proportions (fraction of plot area in different Conditions) must be > 0.95 and < 1.05 for both current and previous plot measurements. This criterion screens out plots with > 5% inaccessible (un-measurable) area, as well as plot records with database errors.
(3) No individual trees with an unrealistic growth rate between the two plot measurements. We considered growth rate to be realistic if the dbh growth rate was between −1 and 5 cm yr^{-1}, and if the relative growth rate was between −0.5 and 1.5 yr^{-1}. These criteria allow for stem shrinkage (e.g., due to damage or water status) and measurement error, and are intended only to filter out large database errors.
(4) Similar plot-level growth estimates obtained from the two methods illustrated in figure 2 of Clark et al. (2001). Large differences in the two estimates may indicate database errors. Note that we applied the two methods to wood production, not total NPP as in Clark et al. (2001). The first method sums the wood production of all trees on the plot that survived the remeasurement interval, along with the ingrowth of new trees. The second method is given by Equation 1 in our main text. Our specific criterion was that the two estimates agreed to within one standard

deviation; i.e., we calculated the standard deviation of differences between the two methods (each plot yielded one difference), and we excluded plots where the magnitude of difference exceeded one standard deviation. This filtering approach excluded < 1% of the FIA plots, because the distribution of differences was highly non-normal (most differences were close to zero, and a few differences were large). Many of these plot records may in fact be free of errors, because although the two Clark et al. (2001) methods have the same expectation, they are not guaranteed to yield the same value for a given plot. Nevertheless, by excluding these plots, we reduce noise and the number of large database errors in our analysis.

Measurement Error Model methods

The Measurement Error Model (MEM) we fit to FIA data is summarized here and described in detail in Fuller (1987). All equation (Eq.) and page numbers below refer to Fuller (1987). The aim of the MEM is to estimate parameters $\boldsymbol{\beta}$ relating a dependent variable y to a vector of explanatory variables $\mathbf{x}$:

$$E[y] = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-1} x_{k-1},$$

where $E[y]$ is the expected value of y, and k is the dimension of $\boldsymbol{\beta}$ (number of parameters). For sample t in 1, 2, …, n, we assume that y_t and $\mathbf{x}_t$ are observed with error:

$$Y_t = y_t + w_t$$
$$\mathbf{X}_t = \mathbf{x}_t + \mathbf{u}_t$$

where w_t and $\mathbf{u}_t$ are measurement errors that are assumed to be mean-zero, normally distributed random variables. We first present estimators, obtained using linear algebra, for the parameters $\boldsymbol{\beta}$ and their variance-covariance matrix quantifying parameter uncertainty. Terms in the estimators are described in the table below. The maximum likelihood estimator for $\boldsymbol{\beta}$ is (Eq. 2.2.20):

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{M}_{XX} - \mathbf{S}_{uu})^{-1}(\mathbf{M}_{XY} - \mathbf{S}_{uw}),$$

where $\mathbf{M}_{XX}$ is the matrix of mean squared X values; $\mathbf{M}_{XY}$ is the vector of mean XY products; and $\mathbf{S}_{uu}$ and $\mathbf{S}_{uw}$ quantify the measurement error variances and covariances. Note that if $\mathbf{S}_{uu}$ and $\mathbf{S}_{uw}$ contain all zeros, then the expression above reduces to the classical ordinary least squares (OLS) formula, $\widetilde{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\,\mathbf{X'Y}$.

The estimator for the variance-covariance matrix for $\widetilde{\boldsymbol{\beta}}$ is (Eq. 2.2.25):

$$\widehat{\mathbf{V}}\{\widetilde{\boldsymbol{\beta}}\} = n^{-1}\left[\widetilde{\mathbf{M}}_{xx}^{-1}s_{vv} + \widetilde{\mathbf{M}}_{xx}^{-1}\left(\mathbf{S}_{uu}s_{vv} + \tilde{\mathbf{S}}_{uv}\tilde{\mathbf{S}}_{vu}\right)\widetilde{\mathbf{M}}_{xx}^{-1}\right] + d_f^{-1}\widetilde{\mathbf{M}}_{xx}^{-1}[\mathbf{S}_{uu}s_{rr} + \tilde{\mathbf{S}}_{uv}\tilde{\mathbf{S}}_{vu}]\widetilde{\mathbf{M}}_{xx}^{-1}.$$

Terms are defined in alphabetical order (Greek symbols first) in the table below using notation from Fuller (1987).

Term	Definition and notes
$\boldsymbol{\beta}$	True (unknown) parameters values (including an intercept, β_0) relating the true values of y to the true values of $\boldsymbol{x}$.
$\widetilde{\boldsymbol{\beta}}$	Maximum likelihood estimator for $\boldsymbol{\beta}$ (Eq. 2.2.20).
d_f	Degrees of freedom for estimating $\mathbf{S}_{aa}$.
k	Dimension of $\boldsymbol{\beta}$ (i.e., number of parameters).
$\mathbf{M}_{XX} = n^{-1}\sum_{t=1}^{n}\mathbf{X}_t'\mathbf{X}_t$	Matrix of mean squared observed $\mathbf{X}$.
$\mathbf{M}_{XY} = n^{-1}\sum_{t=1}^{n}\mathbf{X}_t'\mathbf{Y}_t$	Matrix of mean cross-products of observed $\mathbf{X}$ and observed $\mathbf{Y}$.
$\widetilde{\mathbf{M}}_{xx} = \mathbf{M}_{XX} - \mathbf{S}_{uu}$	Term in expression for $\widehat{\mathbf{V}}\{\widetilde{\boldsymbol{\beta}}\}$.
N	Sample size.
q_t	Random error in the true value of y that is unrelated to $\mathbf{x}$; q_t are independent, normally distributed random variables with mean 0 and variance σ_{qq} (Eq. 2.2.17). Estimator for σ_{qq} is given by Eq. 2.2.21.
$s_{rr} = (1, -\widetilde{\boldsymbol{\beta}}')\mathbf{S}_{aa}(1, -\widetilde{\boldsymbol{\beta}}')'$	Term in expression for $\widehat{\mathbf{V}}\{\widetilde{\boldsymbol{\beta}}\}$.
$s_{vv} = (n-k)^{-1}\sum_{t=1}^{n}\left(Y_t - \mathbf{X}_t\widetilde{\boldsymbol{\beta}}\right)^2$	Term in expression for $\widehat{\mathbf{V}}\{\widetilde{\boldsymbol{\beta}}\}$. Page 107. The square root of s_{vv} is equivalent to the OLS residual standard error.
$\mathbf{S}_{aa}$	Estimated variance-covariance matrix for $\mathbf{a}_t = (w_t, \mathbf{u}_t)$. $\mathbf{S}_{aa}$ includes $\mathbf{S}_{uu}$ and $\mathbf{S}_{uw}$.
$\mathbf{S}_{uu}$	Estimated variance-covariance matrix for $\mathbf{u}_t$, with zeros in the row and column corresponding to the intercept column (a vector whose elements are all one) in $\mathbf{X}$. $\mathbf{S}_{uu}$ is the lower right corner of $\mathbf{S}_{aa}$.
$\mathbf{S}_{uw}$	Estimated vector of covariances between w_t and of $\mathbf{u}_t$. $\mathbf{S}_{uw}$ is the top row and left column of $\mathbf{S}_{aa}$, excluding the upper left cell, which is the variance of w_t.
$\tilde{\mathbf{S}}_{uv} = \mathbf{S}_{uw} - \mathbf{S}_{uu}\widetilde{\boldsymbol{\beta}}$	Term in expression for $\widehat{\mathbf{V}}\{\widetilde{\boldsymbol{\beta}}\}$.
$\tilde{\mathbf{S}}_{vu} = \tilde{\mathbf{S}}_{uv}'$	Term in expression for $\widehat{\mathbf{V}}\{\widetilde{\boldsymbol{\beta}}\}$.
$t = 1, 2, \ldots, n$	Sample number.
$\mathbf{u}_t$	Vector of measurement errors in explanatory variables $\mathbf{X}_t$.
$\widehat{\mathbf{V}}\{\widetilde{\boldsymbol{\beta}}\}$	Estimator for variance-covariance matrix of $\widetilde{\boldsymbol{\beta}}$ (Eq. 2.2.25).
w_t	Measurement error in dependent variable Y_t.
$\mathbf{x}_t$	True (unknown) vector of explanatory variables (including the value one, for the intercept)
$\mathbf{X}_t = \mathbf{x}_t + \mathbf{u}_t$	Observed k dimensional vector of explanatory variables, including the value 1 for the intercept (Eq. 2.2.1).
$y_t = \mathbf{x}_t\boldsymbol{\beta} + q_t$	True (unknown) value of dependent variable, including a random error component (q_t) that is unrelated to $\mathbf{x}$ and is not due to measurement error.
$Y_t = y_t + w_t$	Observed value of dependent variable, equal to y_t (true value of dependent variable) plus measurement error w_t (Eq. 2.2.18).

Implementing the MEM requires an estimated error covariance matrix $\mathbf{S}_{aa}$, which includes the variance of w_t (measurement error in plot-level FIA growth), the covariance matrix for errors in the explanatory variables $\mathbf{u}_t$ ($\mathbf{S}_{uu}$), and the covariances between w_t and $\mathbf{u}_t$ ($\mathbf{S}_{uw}$). We considered three different estimates for the variance of w_t: we assumed that w_t was 0%, 10%, or 30% of the residual variance from the OLS regression with the lowest AIC. All three assumptions yielded similar qualitative results, and we only report results where the variance of w_t was assumed to be 10% of the residual variance. We assumed that all covariances between w_t and $\mathbf{u}_t$ were zero, because the FIA dataset was derived independently of the driver datasets. As explained in main-text Methods, we lacked quantitative estimates for errors in soil available water capacity (AWC), so we performed our analyses under two different assumptions (AWC error = 10% or 50% of the AWC variance across FIA plots) that should bracket the true AWC error. We assumed that covariances between AWC errors and all other error terms were zero, because the AWC data were derived independently of the other datasets. We used two different sources to estimate the temperature and precipitation terms in $\mathbf{S}_{uu}$: AmeriFlux sites (see main-text Methods for a summary of site selection) and Daly et al. (2008). For the AmeriFlux-based estimate, we performed OLS regression of mean annual temperature and precipitation and their squares (i.e., quadratic terms) from PRISM on the corresponding values from AmeriFlux sites. We then estimated the error variances and covariances from the regression residuals. We set covariances between temperature and precipitation terms to zero, because this was approximately true in the AmeriFlux analysis and because these covariances were not available from Daly et al. (2008). Relaxing this assumption had little impact on the AmeriFlux-based results. To estimate $\mathbf{S}_{uu}$ from Daly et al. (2008), we converted their reported percent error of 4% for both mean annual temperature and precipitation (see page 2048 of Daly et al. 2008) into error variances as follows. First, we multiplied 4% times the mean temperature and precipitation values across FIA plots to obtain mean absolute errors, and then we converted these to variances using the properties of the normal distribution: standard deviation = (mean absolute error)/$\sqrt{(\pi/2)}$. To estimate Daly-based mean absolute errors for the quadratic temperature and precipitation terms, we assumed these terms were double their linear counterparts, as was approximately true in the AmeriFlux analysis. Again, we then converted these to variances according to: standard deviation = (mean absolute error)/$\sqrt{(\pi/2)}$. We estimated Daly-based covariances between linear and quadratic terms (e.g., between temperature and its square) as the product of the two error standard deviations, which was approximately true in the AmeriFlux analysis. The two estimated $\mathbf{S}_{aa}$ matrices are below. The matrices are symmetric, so only the diagonal (error variances) and upper triangle (error covariances) are presented.

Error covariance matrix estimated from AmeriFlux sites.

	FIA growth	Temp	$Temp^2$	Precip	$Precip^2$	Soil AWC
FIA growth	0.000529¶	0	0	0	0	0
Temp		0.260	3.72	0	0	0
$Temp^2$			67.5	0	0	0
Precip				0.00423	0.00750	0
$Precip^2$					0.0147	0
Soil AWC						179.7§

Error covariance matrix estimated from Daly et al. (2008).

	FIA growth	Temp	$Temp^2$	Precip	$Precip^2$	Soil AWC
FIA growth	0.000529¶	0	0	0	0	0
Temp		0.0944	0.189	0	0	0
$Temp^2$			0.378	0	0	0
Precip				0.00188	0.00375	0
$Precip^2$					0.00751	0
Soil AWC						179.7§

¶ Error variance is for 60-80 year-old FIA plots. Error variances for the 40-60 and 60-80 year-old age class are 0.000624 and 0.000618, respectively.

§ Error variance is 10% of soil AWC variance across 60-80 year-old FIA plots. The alternative AWC error (50% of variance across FIA plots) is 5 times this value. Error variances (10% of soil AWC) for the 40-60 and 60-80 year-old age class are 185.8 and 175.5, respectively.

The MEM also requires an estimate of the degrees of freedom (d_f) for estimating the above covariance matrix. For the AmeriFlux-based analysis, we set d_f equal to the number of AmeriFlux sites (15) minus one. For the Daly-based analysis, we set d_f equal to 1000 (any value of this order or larger yields similar results), because a large number of meteorological towers were used by Daly et al. (2008) to estimate PRISM errors.

To calculate the proportion of variance explained (R^2) for MEMs (reported in table S1), we used formulae in Fuller (1987: pp. 96 and 113-114) to estimate the true values of the explanatory variables, and we then calculated R^2 from OLS regressions of FIA growth on these estimated values.

Optimization details

To select grid cells for optimization, we first divided the eastern U.S. grid cells into 25 climate strata defined by five mean-annual-temperature and five mean-annual-precipitation percentile classes (0-20, 20-40, etc.). For example, the coldest and driest stratum included all grid cells whose mean annual temperature and precipitation were both in the 0-20th percentiles. More FIA plots per grid cell were available in the northeastern compared to the southeastern U.S. To reduce the north-south data imbalance, we included the two grid cells with the largest number of FIA plots in each of the 10 warmest strata (warmest two temperature classes; all five precipitation classes); whereas in each of the 15 coolest strata (coolest three temperature classes; all five precipitation classes), we included only the single grid cell with the most FIA plots. We excluded from the optimization grid cells with < 20 FIA plots, as these cells provide relatively little information for optimization but carry the same computational costs as more data-rich grid cells. The optimization included 2313 FIA plots, 23 grid cells, and 20 out of 25 possible climate strata (Fig. 2a,b).

For each optimization step, we executed 23 single-grid-cell LM3V runs configured to match the mean schedule (stand age and time of sampling) of FIA plots in each of the 23 optimization grid cells. For FIA plots in a given grid cell, let $\bar{A}_0$ be the mean stand age at the time of the first measurement, and let $\bar{Y}_0$ and $\bar{Y}_t$ be the mean first and second measurement years, respectively. LM3V was initialized with 1 kg C m^{-2} of vegetation biomass in year $\bar{Y}_0 - \bar{A}_0$ and run to year $\bar{Y}_t$. If $\bar{Y}_t$ was greater than 2006, we set $\bar{Y}_t = 2006$, because the Sheffield et al. (2006) meteorology was only available through 2006. Growth was calculated in LM3V between $\bar{Y}_0$ and $\bar{Y}_t$ according to main-text Equation 1. No spin-up was needed for soil C or nutrient pools because the optimization analysis was restricted to wood growth of live vegetation and because we studied a C-only version of LM3V in which soil biogeochemistry has no effect on vegetation. Because LM3V was configured to match the mean schedule of FIA plots in each grid cell, there was an inexact match for any given plot. The mismatch in stand age should have little impact on our results, because growth is nearly independent of age in both LM3V and in the 40-100 year-old FIA plots used in the optimization.

We implemented a modified Gauss-Newton algorithm according to Fletcher (1987) to minimize the sum of squared differences between predicted growth (from the 23 selected LM3V grid cells) and observed growth (from the 2313 FIA plots in the 23 grid cells). For each of the 23 selected grid cells, each Gauss-Newton iteration required three LM3V runs to estimate derivatives (by forward differencing) of the residuals with respect to each of the three optimized parameters (A_l, A_r, and V_{Cmax}). Both the current search direction and step length are derived from these derivatives in the basic Gauss-Newton algorithm. To improve the robustness of the basic algorithm, we replaced the default step length with an "acceptable point" identified from a line search algorithm (Fletcher 1987: pp. 26-40), which required an additional ~5-10 runs per iteration. We terminated the optimization when the cost function (sum of squares) decreased by < 0.01% between successive iterations. We calculated approximate 95% confidence intervals for each optimized parameter from an approximate variance-covariance matrix available in the Gauss-Newton context (Fletcher 1987: p. 112). Tests with artificial data showed that our parameter estimates and confidence intervals were nearly identical to those obtained from the 'nls' (nonlinear least squares) function in the R software package (R Core Team 2012).

Literature Cited

Clark, D. A., S. Brown, D. W. Kicklighter, J. Q. Chambers, J. R. Thomlinson, and J. Ni. 2001. Measuring net primary production in forests: Concepts and field methods. Ecological Applications 11:356–370.

Daly, C., M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. P. Pasteris. 2008. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. International Journal of Climatology 28:2031–2064.

Fletcher, R. 1987. Practical methods of optimization. John Wiley & Sons, Chichester.

Fuller, W. A. 1987. Measurement error models. John Wiley & Sons, New York.

Jenkins, J. C., D. C. Chojnacky, L. S. Heath, and R. A. Birdsey. 2003. National-scale biomass estimators for United States tree species. Forest Science 49:12–35.

R Core Team. 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Sheffield, J., G. Goteti, and E. F. Wood. 2006. Development of a 50-year high-resolution global dataset of meteorological forcings for land surface modeling. Journal of Climate 19:3088–3111.