Model-data fusion bias due to driver errors

Jeremy W. Lichstein[1], Ni-Zhang Golaz[2], Sergey Malyshev[3], Elena Shevliakova[3], Tao Zhang[1], Richard A. Birdsey[4], Stephen. W. Pacala[3]

[1]University of Florida, Dept. of Biology; [2]Geophysical Fluid Dynamics Laboratory; [3]Princeton University, Dept. of Ecology and Evolutionary Biology; [4]USDA Forest Service

Model-data fusion bias

- There is a growing effort to test and improve C-cycle models with a variety of data sources.
- The standard approach to parameter optimization is to minimize the discrepancy between model output and observations; e.g., minimize the sum of squared deviations.
- This approach is biased if there are errors in the meteorological, soil, or other variables used to drive the C-cycle model. To see why, consider the well-known "Errors-in-X" problem…

The Errors-in-X problem

Errors in explanatory variables (X) bias statistical analyses that ignore these errors, whereas errors in response variables (Y) cause uncertainty but no bias.

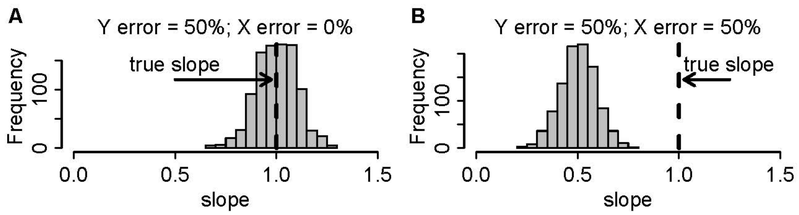


Figure 1. Distribution of estimated slopes from 1000 ordinary least squares (OLS) regressions of randomly generated data with a **true slope of one**. Measurement errors in X account for (A) 0% and (B) 50% of the observed variance in X. Measurement errors in Y account for 50% of the observed variance in Y in both cases.

Demonstration with the GFDL LM3V global land model and forest inventory data (U.S. Forest Service FIA plots)

- We quantified the response of aboveground wood growth in LM3V (Shevliakova et al. 2009) to climate and soil variables.
- We compared these LM3V responses to FIA responses estimated with:
  1) OLS models that ignore Errors-in-X, and
  2) Measurement Error Models (MEMs; Fuller 1987) that account for Errors-in-X.
- We also optimized LM3V vegetation parameter values using a standard optimization approach that ignores Errors-in-X.

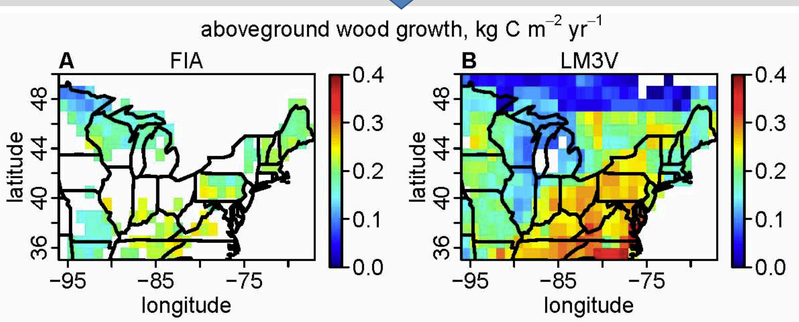


Figure 2. Growth in 60-80 year-old FIA plots (grid-cell mean) and 70 year-old LM3V vegetation (temperate-deciduous-tree functional type).

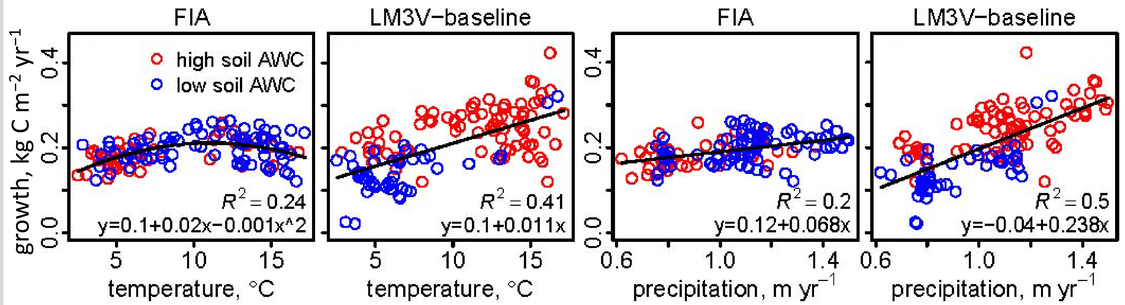
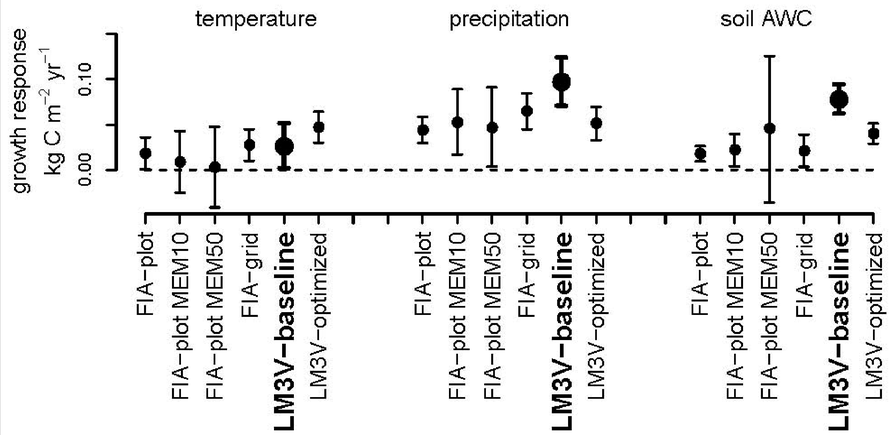


Figure 3. Grid-cell values in Fig. 2 vs. mean annual temperature and precipitation. Blue/red points have low/high values of soil available water capacity (AWC). **LM3V appears too sensitive to temperature, precipitation, and AWC.** The analysis is restricted to grid cells with ≥ 10 remeasured FIA plots in the 60-80 year-old age class.



LM3V appears too sensitive to precipitation and soil when Errors-in-X are ignored, but not when Errors-in-X are accounted for.

Figure 4. Response of aboveground biomass growth to temperature, precipitation, and soil available water capacity (AWC) estimated from FIA data and LM3V (baseline and optimized models). From left to right, the 6 growth responses are from: (1) ordinary least squares (OLS) analysis of plot-scale FIA data; (2-3) Measurement Error Model (MEM) analysis of plot-scale FIA data using temperature and precipitation error estimates derived from AmeriFlux sites and assuming 10% ("MEM10") or 50% ("MEM50") of AWC variance is due to measurement error; (4) OLS analysis of FIA data aggregated to 1° grid cells; (5) baseline LM3V model (shown in bold); and (6) LM3V with optimized parameter values (Table 1).

Optimized fine-root allocation is too high, as expected from the Errors-in-X problem.

Source	Leaf allocation	Fine-root allocation	V_{Cmax}
LM3V-baseline	0.52	0.25	40
LM3V-optimized	0.264 (CI: 0.249-0.279)	0.466 (CI: 0.459-0.472)	17.7 (CI: 16.9-18.4)
Observations	0.28 (range: 0.24-0.35)	0.36 (range: 0.22-0.44)	40.7 (CI: 39.2-42.2)

Table 1. Vegetation parameter values for temperate deciduous trees in baseline and optimized LM3V models vs. observations. Leaf and fine-root allocation are NPP fractions. V_{Cmax} is maximum photosynthetic rate (μmol CO_2 m^{-2} s^{-1}) at 15°C. Observations of leaf and fine-root allocation are biometric estimates from Luyssaert et al. (2007). V_{Cmax} observations are from Kattge et al. (2009).

Can we avoid the Errors-in-X problem by relying on intensive study sites (e.g., AmeriFlux) with small measurement errors?

- We performed a power analysis to determine the number of intensive study sites needed to accurately estimate the relationship between growth and mean annual precipitation within the eastern U.S.
- Different hypothetical values of the "true" growth vs. precipitation slope were derived from LM3V and FIA data.
- To have an 80% chance of estimating a slope within 20% of the true value requires at least 37 intensive study sites within the eastern U.S.!

An unbiased alternative for model-data fusion

1) Use Measurement Error Models to estimate the unbiased functional response of real ecosystems (e.g., response of growth to climate).
2) Fit the functional response of the C-cycle model to the functional response estimated from data (rather than fitting C-cycle predictions to observations). Let $f_D(\theta)$ and $f_M(\theta)$ be the probability distributions of statistical parameters θ that describe the data and model functional responses, respectively. We seek to minimize the Kullback–Leibler divergence: $D_{KL}(f_D||f_M) = \int \ln[f_D(\theta)/f_M(\theta)]f_D(\theta)d\theta$.

Literature cited

- Fuller, W. A. 1987. *Measurement error models*. John Wiley & Sons.
- Kattge et al. 2009. Quantifying photosynthetic capacity and its relationship to leaf nitrogen content for global-scale terrestrial biosphere models. *Global Change Biol.* 15:976–991.
- Luyssaert et al. 2007. CO_2 balance of boreal, temperate, and tropical forests derived from a global database. *Global Change Biol.* 13:2509–2537.
- Shevliakova et al. 2009. Carbon cycling under 300 years of land use change: Importance of the secondary vegetation sink. *Global Biogeochem. Cy.* 23:GB2022.

Funding:

A Core Element of the U.S. Global Change Research Program
North American Carbon Program
Continental Carbon Budgets, Dynamics, Processes, and Management

4th NACP All-Investigators Meeting
February 4 - 7, 2013, Albuquerque, NM