Are Terrestrial Biosphere Models Fit for Simulating the Global Land Carbon Sink?

Christian Seiler[1], Joe R. Melton, Vivek K. Arora, Stephen Sitch, Pierre Friedlingstein, Peter Anthoni, Daniel Goll, Atul K. Jain, Emilie Joetzjer, Sebastian Lienert, Danica Lombardozzi, Sebastiaan Luyssaert, Julia E. M. S. Nabel, Hanqin Tian, Nicolas Vuichard, Anthony P. Walker, Wenping Yuan, and Sönke Zaehle
[1]Environment and Climate Change Canada

Research Problem

Context: The Global Carbon Project estimates that the terrestrial biosphere has absorbed about one-third of anthropogenic CO_2 emissions during the 1959-2019 period. This sink-estimate is derived from an ensemble of terrestrial biosphere models (**TBMs**) rather than direct observations.

Motivation: Evaluate how well TBMs reproduce the processes that drive the terrestrial carbon sink.

Research Problem: What level of agreement between model output and observation-based reference data is adequate considering that reference data are prone to uncertainties?

Methods

Model Scores: Model scores quantify how similar model outputs are compared to reference data, such as remotely sensed gross primary productivity (GPP).

Benchmark Scores: Benchmark scores quantify how similar independently derived reference data are, e.g. remotely sensed GPP versus spatially up-scaled GPP.

AMBER: Automated Model Benchmarking R-package (AMBER) computes scores based on five statistical metrics (bias, centralized root-mean-square error, timing of seasonal peaks, interannual variability, spatial patterns). Scores range from 0 to 1, where higher values imply stronger agreement.

Benchmarking: Comparing model scores against benchmark scores shows how well a model performs when considering how uncertain reference data are (Figure 1).

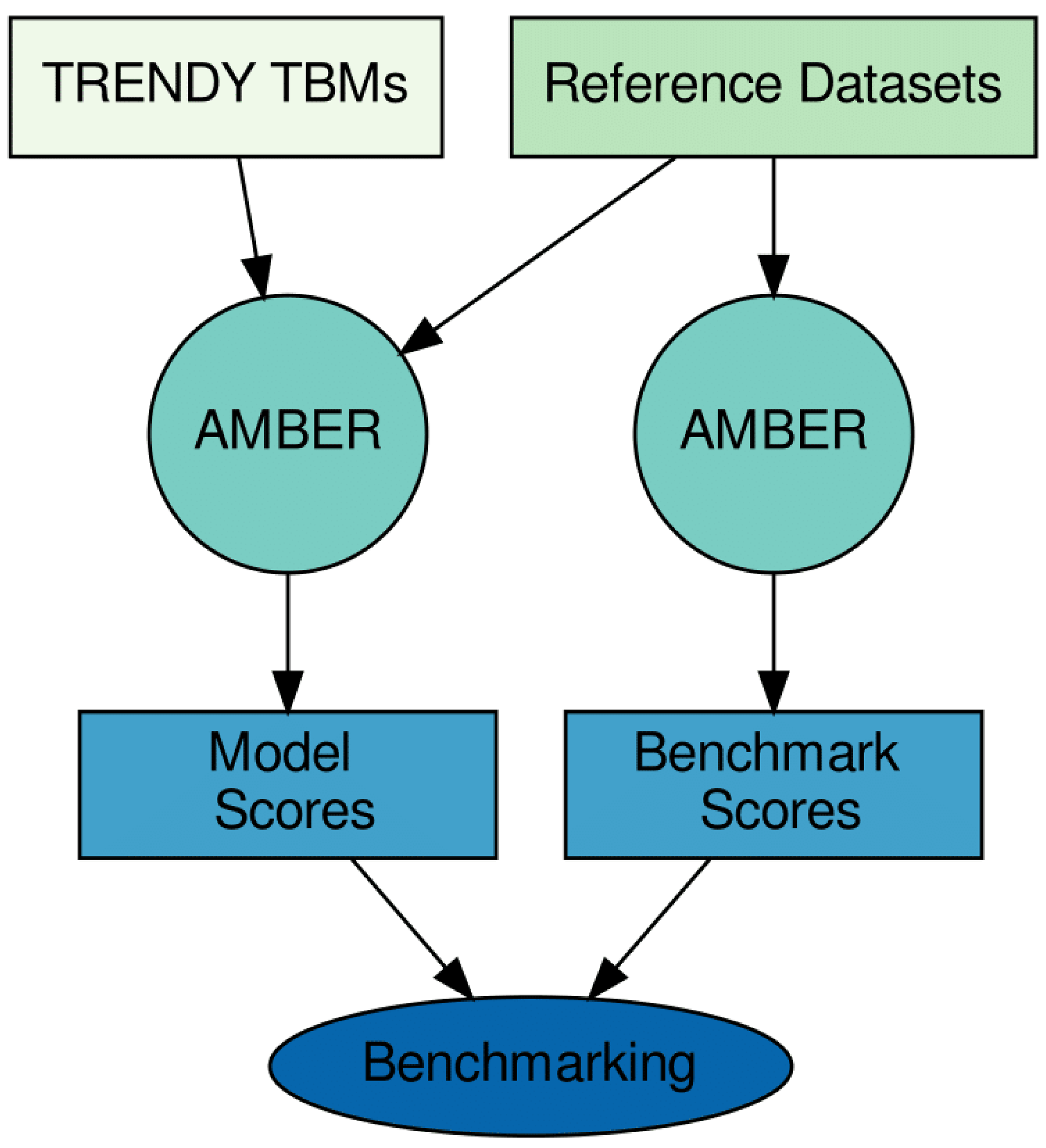

Figure 1: Benchmarking Terrestrial Biosphere Models (TBMs) that form part of the TRENDY ensemble used for estimating the strength of the land carbon sink.

Results

How much do observations differ? Reference data can differ considerably, causing benchmark scores to be low (Figure 2).

How well do models perform? Model scores are often of similar magnitude as benchmark scores, implying that model performance is reasonable given how different reference data are (Figure 2).

What are some of the main model deficiencies? Negative soil organic carbon bias in high latitudes, positive leaf area index bias, large model divergence with respect to GPP in boreal regions and humid tropics (not shown).

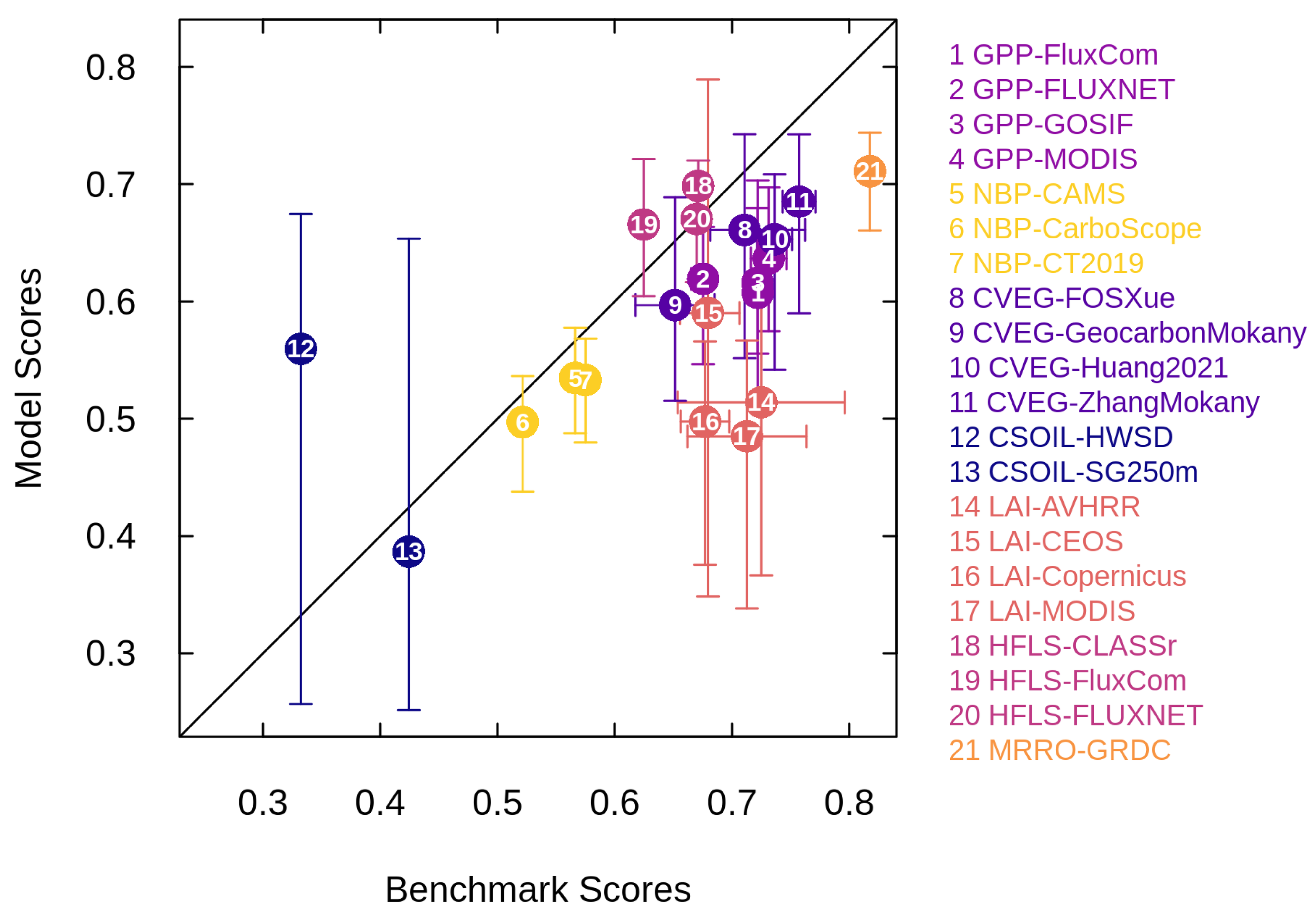

Figure 2: Each dot gives the mean model score for a particular reference data set. The vertical bars cover the total range of model scores, with one value per model. The horizontal bars give the total range of benchmark scores for variables with more than two reference data sets. The variables are gross primary productivity (GPP), net biome productivity (NBP), vegetation carbon (CVEG), soil organic carbon (CSOIL), leaf area index (LAI), latent heat flux (HFLS), and annual streamflow (MRRO).

Key Points

- We quantify differences between independently derived observations to disentangle model deficiencies from observational uncertainties.
- Differences between models and observations are often similar compared to differences between independently derived observations.
- Future work should address biases in soil organic carbon, leaf area index, and the large spread of gross primary productivity among models.

Data Availability

Explorer: https://cseiler.shinyapps.io/AmberTrendy2020/

Data and code: https://doi.org/10.5281/zenodo.5670387

Publication: https://doi.org/10.1029/2021MS002946

Contact: christian.seiler@ec.gc.ca