Accurate Simulation of Both Sensitivity and Variability for Amazonian Photosynthesis: Too Much to Ask

Sarah Gallup[1], Ian T. Baker[1], John L Gallup[2], Natalia Restrepo-Coupe[3], Katherine Haynes[1], Nicholas Geyer[1], and A Scott Dennning[1]

[1]Colorado State University
[2]Portland State University
[3]UTS

November 22, 2022

Abstract

Causes of climate predictions' uncertainty include wide spread in modeled gross primary productivity (GPP) for evergreen broadleaf forests. Deterministic predictions inherently lack the portion of variability that a regression's error term summarizes. Omitted predictors' contribution to error represent simulations' necessary underestimation of real variability. Earth system model outputs with high variability relative to reference data warrant skeptical examination. We compare three statistical and 15 process models to site-level means, seasonal amplitude and driver responsiveness of GPP as calculated at six Amazon eddy covariance (EC) towers. Current month's weather determines only 12% of the variability in EC GPP, implying that models whose predicted GPP's variability approaches that of EC GPP probably are substantially hypersensitive to weather drivers. Roughly half the models have stronger seasonal GPP variability than ECs show, and inaccurately identify the timing of annual minimum GPP. Responses to temperature and light for some highly seasonal models are of the opposite sign as EC GPP's. Strongly seasonal models' deepest dip in photosynthesis both occurs later in the dry season and is more severe than EC estimates. Excessive reactivity to drivers appears to cause the high simulated variability of the strongly seasonal models.

Accurate Simulation of Both Sensitivity and Variability for Amazonian Photosynthesis: Too Much to Ask

Sarah M. Gallup[1], Ian T. Baker[2], John L. Gallup[3], Natalia Restrepo-Coupe[4], Katherine D. Haynes[2], Nicholas M. Geyer[2]and A. Scott Denning[1,2]

[1]Graduate Degree Program in Ecology, Colorado State University, Fort Collins, Colorado, USA

[2]Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado, USA

[3]Department of Economics, Portland State University, Portland, Oregon, USA

[4]

Key Points:

- Regression logic is cause to doubt predictions whose variability is unrealistically high.
- A suite of models poorly reproduce tower estimates of Amazon rainforest gross primary productivity.
- Highly seasonal models predict stronger GPP reactivity to meteorology than is likely to be true.

Corresponding author: Sarah Gallup, `sgallup@colostate.edu`

Abstract

Causes of climate predictions' uncertainty include wide spread in modeled gross primary productivity (GPP) for evergreen broadleaf forests. Deterministic predictions inherently lack the portion of variability that a regression's error term summarizes. Omitted predictors' contribution to error represent simulations' necessary underestimation of real variability. Earth system model outputs with high variability relative to reference data warrant skeptical examination. We compare three statistical and 15 process models to site-level means, seasonal amplitude and driver responsiveness of GPP as calculated at six Amazon eddy covariance (EC) towers. Current month's weather determines only 12% of the variability in EC GPP, implying that models whose predicted GPP's variability approaches that of EC GPP probably are substantially hypersensitive to weather drivers. Roughly half the models have stronger seasonal GPP variability than ECs show, and inaccurately identify the timing of annual minimum GPP. Responses to temperature and light for some highly seasonal models are of the opposite sign as EC GPP's. Strongly seasonal models' deepest dip in photosynthesis both occurs later in the dry season and is more severe than EC estimates. Excessive reactivity to drivers appears to cause the high simulated variability of the strongly seasonal models.

1 Key Words

Amazon

model benchmarking

gross primary productivity (GPP)

Multi-Scale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP)

outliers

tropical rainforest

regression

seasonality

Simple Biosphere Model (SiB)

variability

2 Plain Language Summary

Global climate models must accurately represent many processes, including the pace at which plants convert sunlight, water and CO_2 into sugar. The Amazon rainforest is enormous and extremely biologically productive, so the region strongly influences the world's cycling of CO_2. Measurements from instruments on towers in the Amazon, despite imperfections, seem to be the most accurate estimates of rainforest plant productivity rates that exist. We compare "tower" estimates to 18 global models, focusing on rainforests' subtle dry v. wet seasons.

Modeled monthly plant productivity poorly matches tower estimates. About half the models have more seasonal variation than the towers and half have less. Simple equations that use current month's temperature, light, and rainfall describe model output quite closely, especially for the weakly seasonal models.

Reality is more variable than are mathematical model predictions that accurately describe the results of a particular change in inputs. Why, then, do some models have stronger seasonal swings than tower estimates? One cause is using a descriptive equation that overlooks models' non-linear responses to weather. But the main reason is that when weather changes, tower estimates of plant productivity change less than it do predictions from models with strong seasonal swings.

3 Introduction

Modeling tropical plant productivity accurately is important to the accuracy of global climate predictions because rainforests are so large and productive that their gross primary productivity (GPP) represents about 34% of the terrestrial total [*Beer et al.*, 2010]. Rainforest productivity largely drives interannual variability in global CO_2 concentrations [*Bousquet et al.*, 2000; *Rödenbeck et al.*, 2003; *Wenzel et al.*, 2014]. Positive feedbacks to change in rainforest productivity amplify change in CO_2 [*Christoffersen et al.*, 2014; *Harper et al.*, 2014; *Zemp et al.*, 2017].

The spread in simulations of current rainforest productivity [*Ardö*, 2015; *Malhi et al.*, 2009] is wide, and typically greater than for other biomes [*Anav et al.*, 2015; *Beer et al.*, 2010; *Cavaleri et al.*, 2015; *Friedlingstein et al.*, 2006; *Jung et al.*, 2020; *Mystakidis et al.*, 2016], shows of the need for better modeling for tropical rain forests. Cross-model differences represent material uncertainty about future rainforest productivity. Large differences in tropical GPP persist even after removing model differences in simulated precipitation [*Malhi et al.*, 2009; *Poulter et al.*, 2010a].

3.1 Predictions tend to have lower variance than source data.

For assessing a non-stochastic model's responses to climate change, it is helpful to consider a trade-off between accurate responsiveness to drivers and accurate variance of predicted outcomes. A model's sensitivity, or responsiveness, can be characterized as marginal change in outcome per unit change in a predictor [*Friedlingstein et al.*, 2006; *Hamby*, 1994]. In a regression, responsiveness corresponds to the slope coefficients. For models of GPP, large slopes imply stronger responses to changing climate. Tropical GPP sensitivity is critical because it describes the extent to which climate will continue to alter rainforest activity and even its viability.

In a model, a simplification by definition, omitted drivers cause some of any outcome's real variability. Predictions cannot include the variability that this portion of random error contributes because the model has no information about the missing drivers. For a regression, if modeled responses to included drivers, or sensitivities, are accurate and other sources of model error modest, omitted variables will still make variability of the predictions unrealistically low. We label the inherent tendency for predicted outcomes to have lower variability than true outcomes as "flattening".

An idealized illustration explains why predictions have low variability. Posit a regression that predicts its outcome y_i as a response to one predictor, x_i. The model is perfectly accurate, meaning that its responsiveness, β, exactly equals the true responsiveness of y_i to x_i. The model's only source of error is omitted predictors, which account for the random error term ϵ_i. There is no uncertainty due to imperfections of measurement, sampling, or specification. Variance of the predictions from this nearly perfect regression is necessarily smaller than the source data's variance.

$$\begin{aligned} sampled\ real\ world\ observation:\ y_i &= \beta x_i + \epsilon \\ regression\ prediction:\ \hat{y}_i &= \beta x_i \end{aligned} \tag{1}$$

The error term adds variance only to the source data but not to predicted values (Equation 1, [*Greene*, 2012, Chapter 3]). Text S1 expresses the argument formally.

The regression could describe for a particular model the responsiveness of rainforest GPP to temperature.

$$\widehat{GPP}_i = \beta * temperature_i + intercept \tag{2}$$

If the GPP model is based on enzyme kinetics, there is no internal β for temperature, but instead a variety of other calculations involving temperature. For example, one parameter could be Q10, the exponent for a rate multiplier to rubisco carboxylation per increase of 10^{o}C. The value of Q10 may be derived from bench or field research. The parameter is not estimated directly for rainforest due to lack of source data, and because in theory Q10 is constant for all chlorophyll [but see *Alster et al.*, 2020]. Other steps within the model may further affect GPP's temperature responsiveness. The temperature to which Q10 is applied may be modified from ambient to account for degree of shading. There may be adjustments for each plant functional type's (PFT's) optimum temperature. Temperature may affect GPP indirectly through vapor pressure deficit. The descriptive regression summarizes as a linear approximation the effects of the process model's more complex underlying calculations.

As a thought experiment, assume that the net result of imperfections in the GPP model is that effective rainforest temperature sensitivity, or the descriptive regression's only β, is twice the true sensitivity. A consequence is higher variance of the $\hat{y}_i$'s. As shown in Equation 3, exaggerating the regression coefficient by a factor of two quadruples the predictions' variance.

$$\begin{aligned} \hat{y} &= 2\beta x + intercept \\ \sigma^2_{\hat{y}} &= \frac{1}{n-1}\sum_{n=1}^{i}(2\beta x_i - \overline{2\beta x})^2 = \frac{4}{n-1}\sum_{n=1}^{i}(\beta x_i)^2 \end{aligned} \tag{3}$$

More generally, when the slope of a regression with one predictor changes by a multiplied constant, the variance of predictions made from the same x-values used to fit the model changes by the squared amount of the multiplier. A slope multiplier larger than one increases predictions' variance, and a multiplier smaller than one decreases variance.

If only because models by definition simplify reality, every regression model has a non-zero error term. While error due to omitted variables embodies real-world variability, measurement error, sampling error, and misspecification of mathematical form represent the data and model's uncertainty about included aspects of the real world [*Vicari et al.*, 2007]. More statistical noise from any source increases ϵ and causes more flattening, or less variability of predicted outputs.

There is a trade off between accurate responsiveness of $\hat{y}_i$ to x_i and accurate variance of predicted outcomes. Either deliberately or inadvertently, changing the regression slope can adjust variance to any level including to equal observed variance. But deviations from the optimal regression fit to the data carry a cost of less accurate modeled responsiveness to change in the observed driver(s). The right answer as measured by accurate variability of outcomes may result from the wrong reason of excessive model sensitivity.

A model can predict only what it "knows" about. Numeric calculations simulate the processes for which equations are included, and the consequences of the influences for which driver data are provided. If the model's sensitivities are accurate, outputs have only as much variability as the included processes and drivers create. The maximum portion of true outcome variance that a model whose responsiveness to drivers is perfectly accurate can simulate is the portion that the included processes and drivers in fact determine. The more completely a model with accurate driver responsiveness includes all true determinants of its outcome, the larger and closer to correct will be its predictions' variance.

The variability that model errors contribute and that otherwise is missing from deterministic predictions can be added back in directly. If random statistical noise is added, then predictions can have both realistic variance and accurate reactivity to predictors. If instead the introduced noise is correlated with drivers, such as by drawing predictions from a probability distribution at calculated percentiles, then effective driver slope(s) will be altered as described in Equation 3. Stochastic modeling has computational and other complications, and is rare in full earth system models (ESMs).

3.2 Flattening applies to Earth System Models.

Flattening occurs within parameterized ESM calculations. Many hard-coded model parameters are "essentially a smaller model within the larger model" [*Dahan*, 2010]. For example, in the Community Land Model (CLM5.0) each PFT's stomatal resistance parameter originated in a regression fitted to a global database of conductances [https://escomp.github.io/ctsm-docs/doc/build/html/tech_note/index.html section 2.9.3,

Table 2.9.1 of values from *De Kauwe et al.*, 2015]. Real variability in resistances caused by variables omitted from the source equation and subsequently from the ESM remains unexplained and unmodeled.

Climate models' inner workings are decisively more intricate than a single linear regression. But flattening is a tendency of any deterministic numeric prediction, including non-linear equations, transformations of variables (Text S1) and, like entire ESMs, complex combinations of equations with feedbacks. Inclusion of processes may be indirect, such as a model forced with satellite data that is a proxy for deciduous leaves' annual cycle. The driver data itself may be simulated, as weather is in fully-coupled ESM runs. In all of these situations, predictions from the model will lack the portion of real variance determined by omitted drivers and processes.

ESMs simulate systems so complex that omitted processes and drivers loom large. Because significant determinants of real variability are missing, the connection between the accuracy of predictions' variance and the accuracy of driver sensitivity provides a diagnostic tool. If the variance of a predicted outcome is higher than or even close to a benchmark's variance, offsetting excessive sensitivity to drivers could be a cause.

3.3 The Amazon is likely to become warmer, with more variable rainfall.

Change in tropical GPP as represented in ESMs depends largely on four environmental drivers: ambient CO_2, precipitation, temperature, and top of canopy insolation. Prediction accuracy depends on correctly simulating rainforest GPP responses to changes in the forcings. Like the rest of the world, rainforests are experiencing consistently increasing ambient CO_2. Models concur that the region's precipitation will become more variable [*Bathiany et al.*, 2018; *Chadwick et al.*, 2015; *Feng et al.*, 2013], a trend for which there already are observational indications at least on the drying side [*Fu et al.*, 2013; *Gloor et al.*, 2013; *Li et al.*, 2008; *Lopes et al.*, 2016]. The direction of change in rainforest mean rainfall rather than in its increasing variability is uncertain, however [*Gloor et al.*, 2012; *Li et al.*, 2006; *Poulter et al.*, 2010a].

Rainforest temperature is projected to rise and may already have changed measurably [*Corlett*, 2011; *Jiménez-Muñoz et al.*, 2013]. Temperature increases are likely to vary regionally within the basin [*Gloor et al.*, 2012]. Deforestation and temperature may have an amplifying feedback on GPP at landscape scales, with lower plant productivity causing more sunlight to become sensible rather than latent heat, and higher temperatures further reducing photosynthesis rates.

In the sometimes-cloudy tropics light can limit photosynthesis, especially for lower leaves in thick canopies. Amazonian surface insolation has increased slightly in recent decades due to reduced cloudiness [*Barkhordarian et al.*, 2017], though the trend's robustness is unclear [*Wielicki et al.*, 2002]. One likely cause of increased surface light, seasonal changes in the sizes of both Pacific and Atlantic Ocean warm pools, has an uncertain anthropogenic signal [*Arias et al.*, 2011]. Pollution and tropical fires on the other hand, reduce top of canopy insolation. Both reflect human behavior, an influence whose uncertainty increases the spread in trend predictions for most weather parameters.

This paper explores the fidelity of modeled Amazonian GPP to eddy covariance (EC) flux tower data, with an emphasis on the accuracy trade-offs that flattening presents. Credible representations of responsiveness to change in weather are especially important for climate models because they describe the direction and strength of trends in GPP in response to increasing concentrations of greenhouse gasses. If responsiveness is too weak, forecasts will be unreasonably reassuring. Excessively strong weather responsiveness will predict faster and more dramatic dieback of the rainforest [*Cox et al.*, 2013].

4 Methods

We compare EC GPP to 15 process models and three statistical models. Methods summarized in this section are described further in Text S2. The statistical models, Fluxcom, Wecann and VPM, each have fared well in global accuracy intercomparisons. SG3 runs for Multi-scale synthesis and Terrestrial Model Intercomparison Project's [MsTMIP; *Huntzinger et al.*, 2014; *Wei et al.*, 2014] 14 process models have common initial land cover maps, land use and land cover change, spin-up procedures, and atmospheric CO_2 and weather inputs. To MsTMIP we added SiB4 [*Haynes et al.*, 2019a,b], a recent major revision to the participating SiB3 model and which now has prognostic phenology.

To compare process models to statistical models that rely on satellite data, the evaluation period is limited to 2000 - 2010. Wecann starts in 2007 with the earliest satellite dataset for solar-induced fluorescence. Wecann is included only in basinwide comparisons because its shorter period for approximating seasonal cycles at tower sites would increase Wecann's apparent variability compared to all other models.

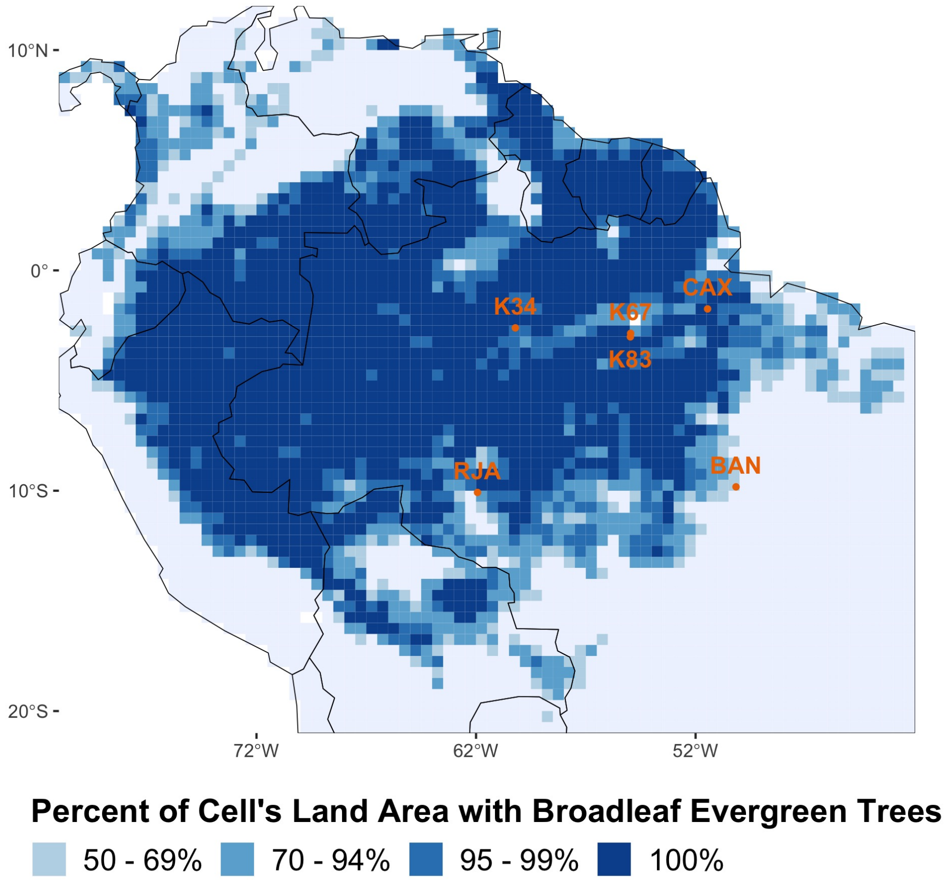


Figure 1: Study area, which includes all shaded cells. Orange dots are locations of eddy covariance towers. Based on MsTMIP's PFT classifications, a high proportion of the study area is almost pure rainforest. The land PFT distribution in 86% of study cells is at least 90% evergreen broadleaf forest.

For basinwide comparisons, the study area is northern South America, the world's largest rainforest that we refer as the Amazon although we do not use a strict watershed boundary. The study cells are limited to 42° - 81°W and 12°N - 21°S, excluding Central America both north of 7°N and west of 77.5°W. Selecting grid cells whose MsTMIP tiled PFTs are at least 50% evergreen broadleaf forest (EBF) limits the study to rainforest vegetation. Fig. 1 shows the portion of each study cell that MsTMIP codes as EBF.

We compare the process and statistical models to six EBF eddy covariance sites from the Large-scale Biosphere-Atmosphere Experiment in Amazonia: Rio Javaés-Bananal (BAN), Caxiuanã (CAX), Manaus Kilometer 34 (K34), Tapajos Kilometer 67 (K67), Tapajos

Kilometer 83 (K83), and Reserva Jaru (RJA) [*Restrepo-Coupe et al.*, 2013]. Fig. 1 maps the towers' locations. Each process or statistical model's GPP estimates for the grid cell containing a site are matched to the months for which data exists at each tower.

Eddy covariance towers are flawed benchmarks for GPP. Measured net ecosystem exchange is a small residual whose much larger offsetting components of GPP and ecosystem respiration must be modeled. A particularly thorny issue is lack of closure in energy budgets. Calculated energy fluxes leaving a site do not equal measured energy entering [*da Rocha et al.*, 2009; *Jung et al.*, 2019; *von Randow et al.*, 2004]. Where GPP seasonal variation is smaller than average, as in the tropics, closure corrections introduce more noise [*Clark et al.*, 2017; *Tramontana et al.*, 2016]. These weaknesses are serious. Nevertheless, and partly on faith, we take tower estimates to be the best reference data available, and their GPP responsiveness to individual drivers as true to the extent of being qualitatively strongly positive, strongly negative, or weak.

We define 'site' as an eddy covariance location. 'EC' refers more specifically to measurements made at a tower site and their derivatives. Unless noted, the weather driver data used to assess modeled GPP's responsiveness is MsTMIP's. Precipitation is monthly total in mm. Light is monthly mean top of canopy short-wave radiation under all sky conditions, in Wm^{-2}. Mean monthly temperature is measured in ^{o}C, and GPP in $gCm^{-2}d^{-1}$.

5 Results

5.1 Modeled GPP mean and variance grossly differ from EC estimates.

An optimistic hypothesis that each model's simulated GPP mean and variance match EC estimates is easily rejected. The lines in Fig. 2 enclosing the EC mean and variance mark wide 99^{th} percentile bootstrapped confidence intervals. Averaged across the six sites and all months of each tower's operation, nearly all model estimates are outside the ECs' confidence intervals. The EC variance reflects only calculated mean monthly GPP, however, and does not include the considerable additional uncertainty from EC modeling and measurements.

For individual sites, at least one model severely underestimates mean GPP and at least one model's mean is more than twice EC GPP (Fig. S7). On average one or two models' means are credible matches to EC data. The variance of two process models' simulated GPP is higher than EC variance for all sites, while one is too low for every site. No model's variance is within target range for every site. One statistical model is credible for five of six sites.

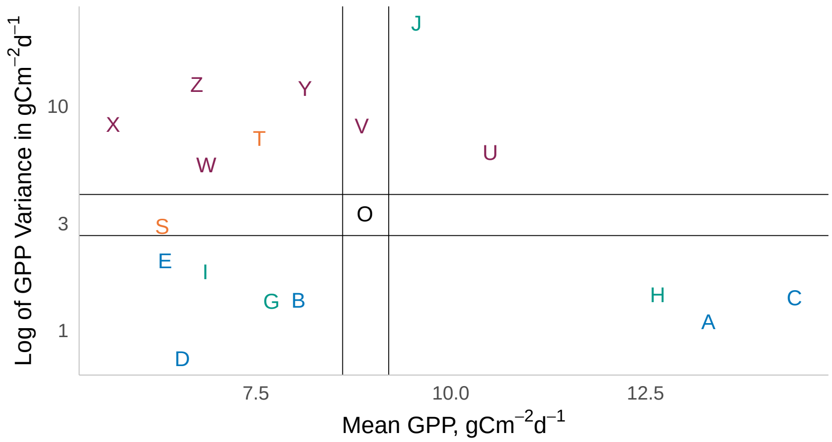


Figure 2: Comparison of EC GPP to process and statistical model means and variances across all site-months. Variance, on the y-axis, has a $\log_{10}$ scale. Model "O" is EC GPP. Names of other lettered models are shown in Fig. 3. Lines bracketing EC estimates are 99th percentile confidence bounds. For all but two models both means and variances fall outside the confidence bounds.

Which models are outliers differs across sites, and some models have very different relative performance among sites. For example, model J's variance is an outlier in Fig. 2 due to its puzzling GPP close to zero for several grid cells near K67, while its GPP is well above average at most other sites.

Correlations with EC GPP are remarkably low. Individual models' average across all sites ranges from -0.16 to 0.45, with a grand mean of only 0.12 (range across sites: -0.32 - 0.66, Fig. S8). Overall correlations with EC GPP for four models are statistically indistinguishable from zero. CAX and K67 have especially weak matches, with negative correlations for 12 and 14 of the 17 models respectively. The most closely simulated site is RJA, where EC GPP is especially variable and average correlation across all models is 0.66.

Data with larger magnitudes tend to have larger variance than data with smaller magnitudes, which tends to make GPP variance of mildly responsive models lower for strongly responsive models. But most models in Fig. 2 show the opposite pattern, with higher variance and lower mean GPP than the EC towers or the opposite. Means with large absolute values do not correspond to large variances. The opposing tendencies mean that for these GPP models, whatever causes differences in means does not explain variability. The causes of differences in variance need to be considered directly. Based on the logic of flattening, we hypothesize that high variance models may be overly sensitive to drivers. Before an exploration of model responsiveness, in the next section a more robust descriptor of model variability is assigned.

5.2 Seasonal cycle amplitude characterizes a model's GPP variability.

To explore the connection between flattening and sensitivity, and to compare models to EC GPP based on the relative variance of their predictions, one possibility is to rank the models based on variances in EC estimates. An alternative metric of variability that is more closely related to the biology being simulated is the amplitude of a model's seasonal cycle. Unlike a variance calculation, it takes into account the sequencing of observations. A Fourier series approximation smooths a site's GPP across outliers, uneven numbers of observation years per month and missing data. Earth's annual insolation cycle is sinusoidal, giving Fourier transformations inherent good fit for some ecological cycles. Characterizing seasonal cycles of GPP with four pairs of Fourier terms is a compromise between overfitting versus forcing unrealistic simplification. The first pair can be thought of as creating an annual cycle, the second allows for asymmetric shoulders, and the third and fourth provide for limited shaping of the annual peak and trough.

We label the difference between maximum and minimum months of a site's mean annual Fourier cycle as seasonal amplitude. In Fig. 3 and elsewhere, models are listed in increasing order of their seasonal amplitude averaged across the EC sites and indicated with black dots. Colors indicate how a model's amplitude compares to the EC amplitude, both averaged across all sites. The nine 'mild' models with weaker mean seasonal cycles than ECs are shown in blue or green. The eight 'lively' models whose cycles are stronger are colored red or orange. The intensity distinctions for dark blue or red break at one standard deviation from the EC mean. Most notable is how widely the seasonal swings differ across models, by a factor of 8.2. The difference means roughly that model Z's simulated trees vary in productivity eight times as much during a year as do model A's. The only model that would switch between the categories of mild versus lively if rankings were determined by variance instead is Model J, whose very high variance (Fig. 2) is due to anomalous GPP at one site.

ECs are a benchmark for model seasonality, albeit one with arguable accuracy. The mildest model varies during the year a third (0.35) as much as does EC GPP. The most strongly responsive model's mean site amplitude is almost triple (2.9 times) that of the ECs.

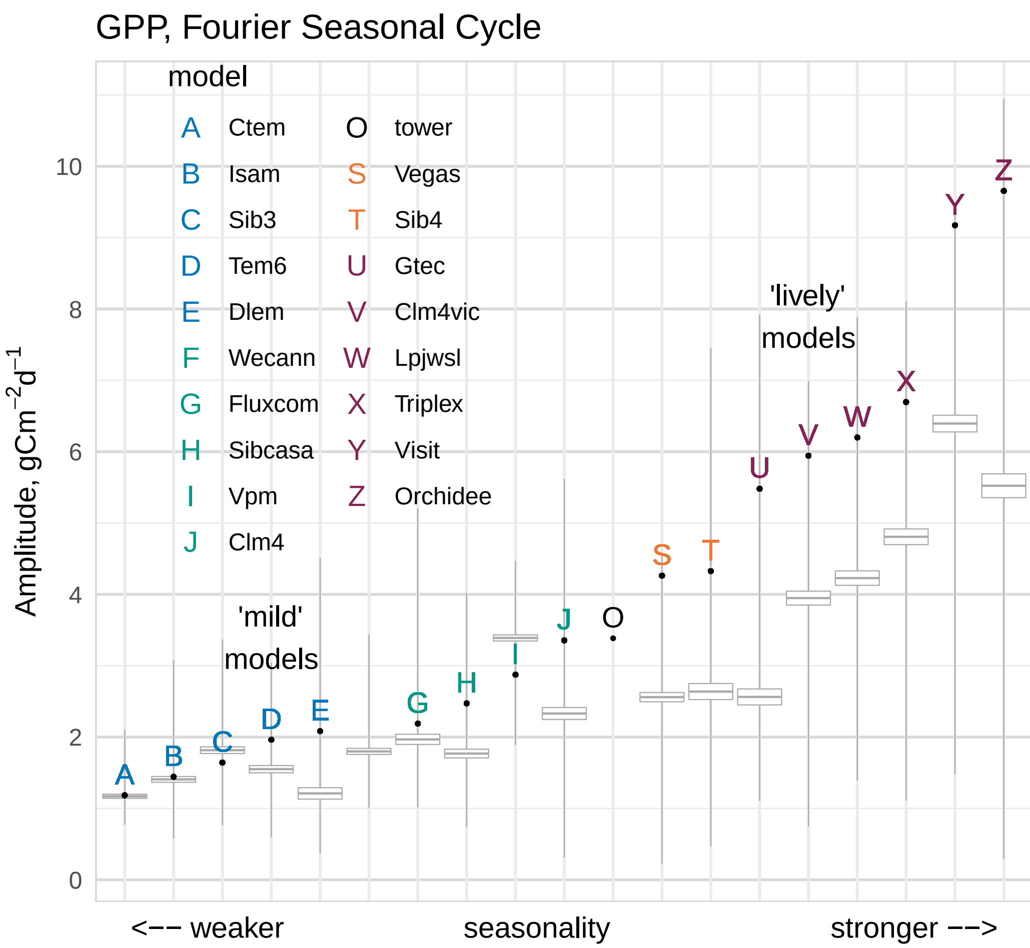


Figure 3: GPP's seasonal amplitude, with models' means across six sites ranked on the x-axis and marked by black dots labeled with colored letters. Grey boxes show a model's amplitude tendency across all rainforest cells in the Amazon. For both sites and basin-wide, the range in seasonal amplitude across models approaches an order of magnitude.

The models' degree of seasonality differs strongly also at individual sites. At no site is a mild model's mean amplitude larger than four $gCm^{-2}d^{-1}$ (Fig. S9), while few of the most responsive models, shown in red, have a seasonal amplitude below four $gCm^{-2}d^{-1}$ at any site. As climate parameters that affect productivity shift over time, a lively model is likely to predict greater change in rainforest carbon fixation per unit area than a mild model.

Each mean amplitude summarizes only six data points, so the mean EC GPP, 3.4 $gCm^{-2}d^{-1}$ has large uncertainty. Based on a t-test, only the liveliest model's amplitude is outside a 95% confidence interval around the EC mean. K34 and K67 are located close enough to each other to have somewhat similar climate. With four degrees of freedom rather than five to reflect possible pseudoreplication, no model is outside the credible interval. There are too few data points to tighten the confidence interval by bootstrapping.

ECs' differing periods of operation preclude a temporally exact comparison between a model's basinwide and site tendencies. Fig. 3 shows in grey a box plot of a 95% confidence interval around the mean seasonal amplitude for all Amazon rainforest cells in all months. Whiskers on the grey boxes represent the tenth and ninetieth percentiles of cell amplitudes. The 20% of each model's cells that are outliers are not shown. As explained in the methods section, there is no EC mean for model F, whose ranking is an approximation. For the most mild models, shown in blue, EC amplitudes are roughly representative of the entire basin. For all the strongly lively red models but one, EC sites have moderately stronger seasonality than do basin-wide means. The extent to which the ECs are typical of the Amazon as a whole decreases with model seasonal amplitude.

Comparing seasonal amplitudes to interannual variability (Fig. S10) reinforces how strong the consequences of seasonal cycles are for simulated GPP. The difference between highest and lowest year's mean GPP from 2000 to 2010 for individual models ranges from 0.1 to 1.2 $gCm^{-2}d^{-1}$. The models' mean basin-level seasonal amplitude ranges are several times larger, from 1.3 to 6.0 $gCm^{-2}d^{-1}$. The range in seasonal cycle amplitudes across models, 1.2 to 9.7 $gCm^{-2}d^{-1}$, approximately equals the grand mean of monthly GPP across all models, 8.9. For GPP in the Amazon, understanding what drives variation within a year explains much more about a model's tendencies than do determinants of its interannual variability.

5.3 EC GPP barely responds to current weather.

Compared to many temperate locations, rainforests are always moist, always warm and always green. But even the wettest tropical forests have subtle annual cycles in rain, temperature and light. Our null hypothesis is that a simple linear combination of these three drivers largely describes monthly mean GPP. If so, weather's cycles might logically also set the timing of the simulated annual GPP cycle. Importantly, the drivers' individual influences on GPP could be parsed and evaluated [*Hamby*, 1994]. With differing trends expected for each driver, a model with retrospectively credible responsiveness to each is more likely to predict reliably.

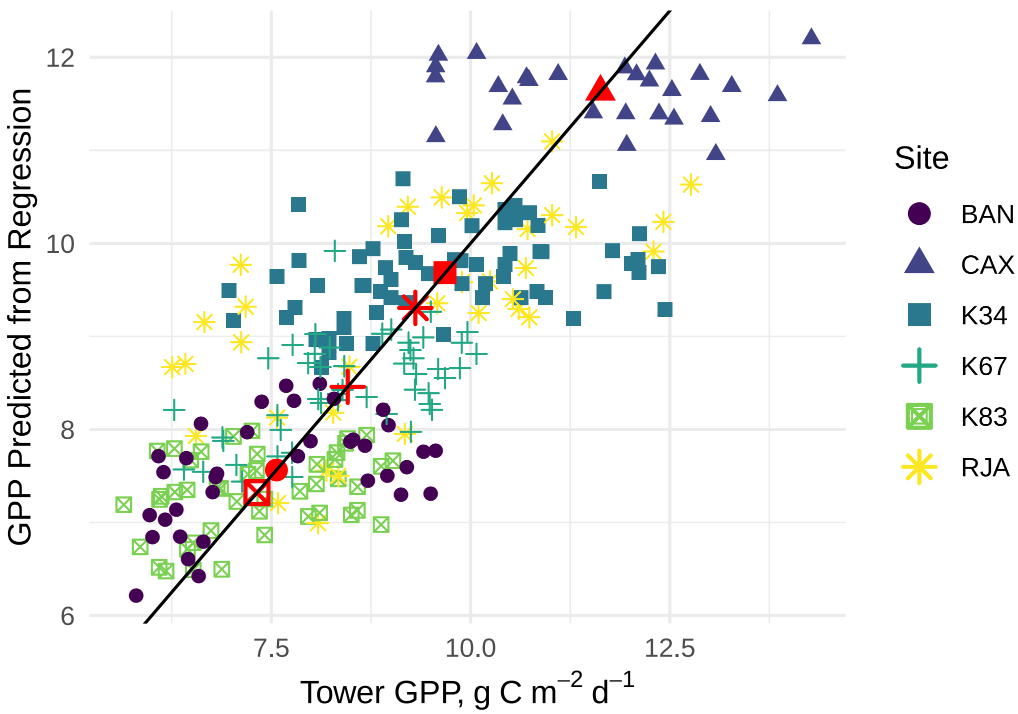


Figure 4: Monthly GPP for each eddy covariance tower compared to paired values predicted from a regression on MsTMIP rain, light, and temperature with site intercepts. Individual intercepts force each site's mean predicted value to equal the mean EC value, indicated with red symbols. Differences between sites are the main source of the prediction's power, with minimal EC GPP responsiveness to weather.

To judge models' responsiveness requires EC benchmarks. A linear regression with only current month's rain, temperature and light explains a small 12% of the variability in EC GPP (not shown). Rain's coefficient but no other is statistically significant.

GPP varies substantially between sites (Fig. 4). Individual cell intercepts, or fixed effects, segregate out the undetermined sources of location-specific differences that cause a particular site's outcome to differ by a consistent increment over time. For process models of GPP, potential underlying causes of site differences include soil depth and fertility, species assemblage in the spectacularly diverse tropics, herbivory, tree age distribution that reflects disturbance history, local geology that affects flooding and subsurface hydrology, and others. Most of these causes for site differences are impossible to parameterize globally, and therefore from an ESM perspective constitute impenetrable statistical noise. Site intercepts represent differences between locations that may or may not be modeled accurately, and if not, are apt to involve omitted variables. But separate intercepts allow a focus on how consistently the three weather drivers cause GPP to change at all locations even if site characteristics significantly influence long-term baseline productivity.

Equation 4 predicts monthly mean EC GPP from MsTMIP weather, with site intercepts added.

$$GPP = Intercepts \ + \ 0.0054 * Rain \ + \ 0.019 * Light \ + \ 0.52 * Temperature \tag{4}$$
$$P\text{-}values : Rain = 0.00; \ Light = 0.01; \ Temperature = 0.00$$
$$Adjusted \ R^2 = 0.59; \quad Residual \ standard \ error = 1.2; \quad n = 260$$

Fig. 4 compares EC GPP on the x-axis to paired predictions from Equation 4 on the y-axis. Site intercepts account for most of the spread in the EC dataset. Due to the weak predictive power of the weather variables, paired values for individual sites in Fig. 4 do not otherwise cluster near the 1:1 black line of perfect prediction. As measured by the adjusted r^2, the regression terms determine 59% of the variability. Site-level differences account for 81% of the regression's predictive power [*Chevan and Sutherland*, 1991], leaving 19% of explained variability, or only 12% of total variability, predicted by the current month's environmental attributes. Contrary to the initial hypothesis, current weather has little influence on EC GPP.

Each site's predicted values are flattened, with less variability on the y-axis than the source values on the x-axis. While the EC GPP data points have a variance of 3.3, the variance of matched but flattened predictions is 2.0, or 61% as large.

If EC meteorology for the comparable cell is used rather than MsTMIP weather, the regression fit with site effects degrades slightly ($r^2 = 0.54$). The only terms whose p-value is $\leq .10$ are four site intercepts and the slope for rain. GPP is statistically unrelated to either light or temperature. Why site-specific weather should be less predictive than regional weather is unclear. Perhaps soil moisture is influential, reflects regional recharge, and overwhelms highly localized rainfall differences. Also, conceivably a geographically broader weather summary more accurately represents conditions across ECs' full footprints than does weather at point locations chosen to represent the upwind area. The unexpectedly weaker fit with site weather is convenient, however. Errors in representing the true values of the drivers cause attenuation bias in regression coefficients, or weaker sensitivity. If MsTMIP weather were a worse fit than site weather, comparisons of EC driver sensitivities to model sensitivities would be less straightforward.

5.4 Lively models respond more strongly to current weather than mild models.

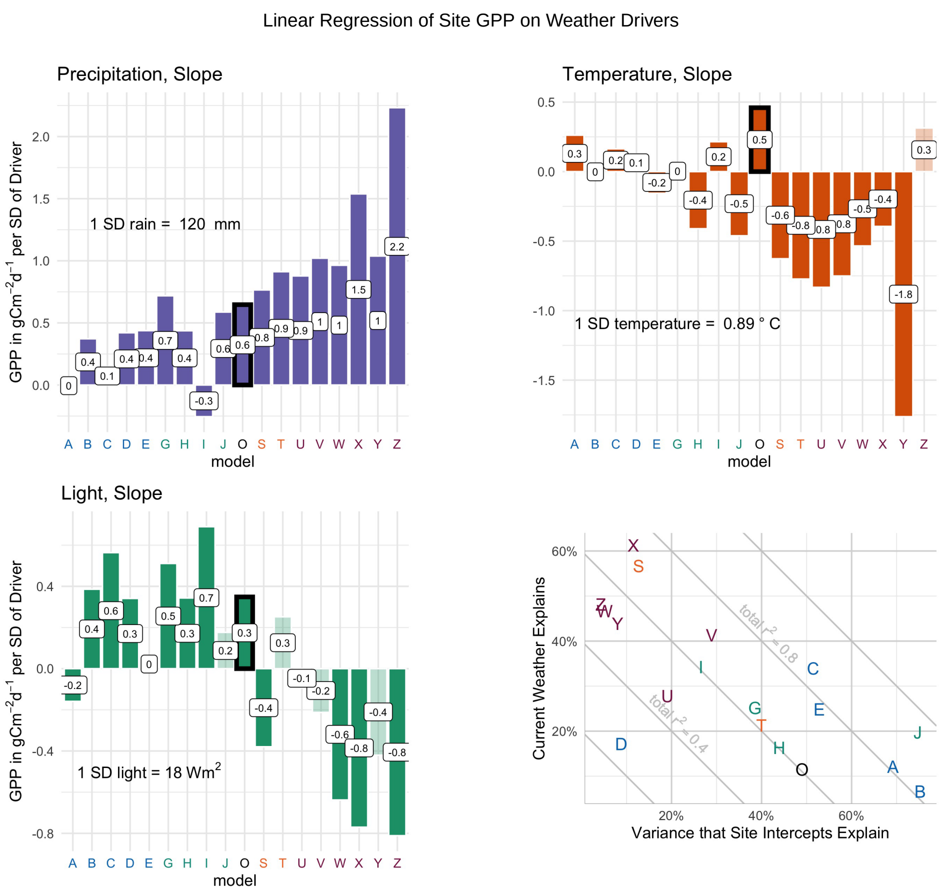


Figure 5: Linear regression slopes of GPP on current month's temperature, radiation, and precipitation for each model across six sites. Drivers are in units of standard deviation across all sites and months. Pale bars are coefficients whose p-value exceeds .05. EC slopes are outlined in black. The lower right panel shows the residual standard error and adjusted r^2 for each regression. Slopes, or driver responsiveness, range among models across an order of magnitude.

Whether flattening is as strong an influence on a model's simulations as on EC predictions depends on (a) whether current weather describes similarly little of the model's GPP variability, and (b) how reasonably a linear sum describes the mathematical form of current weather's relationship to modeled GPP. The three driver variables included in Eq. 4 describe EC GPP's weak but statistically significant responses to each. Based on the Akaike Information Criterion fit, all three drivers should be retained. All would be kept even if the full combination were less than ideal, however, in order to explore next each model's potentially differing emphases among the weather elements.

The vigor with which some models respond to rain, light and temperature contrasts sharply with EC GPP's weak responses. The regressions whose responsiveness slopes appear in Fig. 5 are simple additive models with site-specific intercepts, parallel to the characterization above of EC GPP. To facilitate comparison across drivers, environmental variables are shown in units of standard deviations. A regression slope of +0.5 in Fig. 5 indicates a tendency for a half $gCm^{-2}d^{-1}$ increase in GPP to result from an increase of one standard deviation in the driver. Among the models, statistically significant coefficients for rain range from -0.26 to 2.2, for temperature from -1.8 to 0.26, and for light from -0.81 to 0.69. Ambient CO_2 is an insignificant predictor of historical site GPP in nearly all models (Fig. S11) and is not included in Fig. 5's regressions.

Among the GPP drivers, rain is the most consistent predictor across models. Its sign is positive for all but one model, and its influence statistically significant ($p \leq 0.05$) for all but one. The magnitude of rain responsiveness varies substantially, from 0.1 to 2.2 $gCm^{-2}d^{-1}$ of GPP per 120 mm increase in a month's precipitation. GPP increases with rain in all but 2 models. Models' rank for rain slopes almost matches that of seasonal amplitudes. The ECs' responsiveness to rain, outlined in black, sits solidly in the middle. For rain, the main difference among models is the response strength.

No environmental variable has a monopoly on GPP. Mean absolute slopes differ by less than a factor of two: 0.75 for rain, 0.46 for temperature, and 0.40 for light. Responses to temperature differ more than they do to rain. Temperature is statistically insignificant for four models. For the mildest models, temperature's influence is positive, as it is for EC GPP, and very weak. For the liveliest, GPP strongly declines. For light, statistically unreliable slopes exist across the range of slope magnitudes. On average mild models have nearly as strong a GPP response to light as do lively models but light's effect is in the opposite direction. For temperature and light, there is as much disagreement between models in what direction GPP responds as how strongly.

The descriptive regressions largely characterize GPP for all models. For the mild models as a group, residual standard error (RSE) averages 8% of site GPP and mean r^2 is 0.69 (Fig. S12). For the lively models, average RSE is 24% and mean r^2 is 0.58. GPP and current weather have an even closer linear connection for seasonally mild models than for livelier models.

Differences in mean cell GPP, or among intercepts, are substantial and influential. Except for one model, the range in intercepts as a percent of mean site GPP is 12 - 55%. The exception, model J, has intercepts that vary by more than 100% of mean GPP due to one outlier site. Given that site intercepts are largely a proxy for omitted variables, it is not surprising that they are relatively less influential on process models than on EC GPP. The wide range in site GPP compared to relatively modest slopes for driver values means that site means constitute most of the descriptive regression's explanatory ability, as they do for the ECs.

As shown in Fig. 5's lower right panel, for mild models site intercepts explain more of GPP's variance (mean = 49%) than does weather (mean = 21%). For lively models, site intercepts explain a smaller share (mean = 16%) than does weather (mean = 43%)

A model's responses to drivers in the six cells with eddy covariance towers are generally similar to its responses across the Amazon (Figs. S11 and S12), suggesting that assessing model responses for the Amazon by comparing them with EC estimates of GPP is a reasonable application of scarce benchmarking data. But the similar mean tendencies smooth across considerable spatial differences (Fig. S13). For percent of variance that a simple regression explains, the most striking spatial pattern is that for almost every model there are areas where a linear relationship of current environmental conditions plus site intercepts describes change in GPP almost completely, and other places where it explains little.

Weather's stronger influence on responsive model variability compared to the mild models is consistent with lively models' generally steeper slopes for each weather driver. The principle of flattening suggested, and the data summarized in Fig. 5 confirm, that models with high variance, the strongly seasonal models, are on average overly responsive to drivers.

5.5 Models' differing rainforest non-linearities are not benchmarked.

Compared to regressions that characterize reality, those that describe model output are unusually clean. All of the modeled values for the study period and area typically are accessible, yielding a census with no sampling errors nor errors in measuring outputs. Some or all of the inputs to the model's calculations may also be known exactly, as is MsTMIP weather. Only two sources of stochastic noise remain in the regressions that describe modeled GPP: omitted drivers and misspecification. This section considers alternative specifications, first interactions between weather drivers then non-linear responses.

Adding interaction terms to the regression that describes EC GPP minimally increases its total explanatory power, while diluting evidence of individual environmental drivers' influence. The same predictors as Equation 4 were used plus all possible crosses for the three weather drivers: rain times temperature, etc. The resulting regression has 5% more explanatory power than the model without interactions ($r^2 = 0.62$). But no single or combined weather driver has a significant slope.

Unfortunately, there are too few months of noisy EC data to resolve a non-linear regression form or assess the accuracy of flex points. It is possible that daily GPP estimates from the towers would better resolve non-linear responsiveness, or might share with monthly means having too much random uncertainty. Fig. 6 highlights differences in model responses to extreme temperatures. On the x-axes monthly mean temperatures are grouped by deciles basinwide. GPP on the y-axis is displayed as z-scores to remove model differences in mean and variability for each cell. What remains is the degree to which a model's response to extremes of temperature are anomalous compared to its responses to currently more typical temperatures.

The descriptive regressions with simple linear forms of the drivers do reliably indicate mean tendencies across the range of driver values for the EC months sampled. But as climate shifts, the models' responsiveness to more extreme values will be most relevant. For most models, responses have particularly strong deviation at high temperature from their mean responsiveness. GPP rises continuously with temperature in about a third of models. The rest eventually flex into declining plant productivity. Most of the mild models simulate that rainforest is more productive in the warmest of months, while lively models reach the opposite conclusion. Mild models H and J resemble the lively models in this respect, simulating their very lowest GPP at peak temperatures.

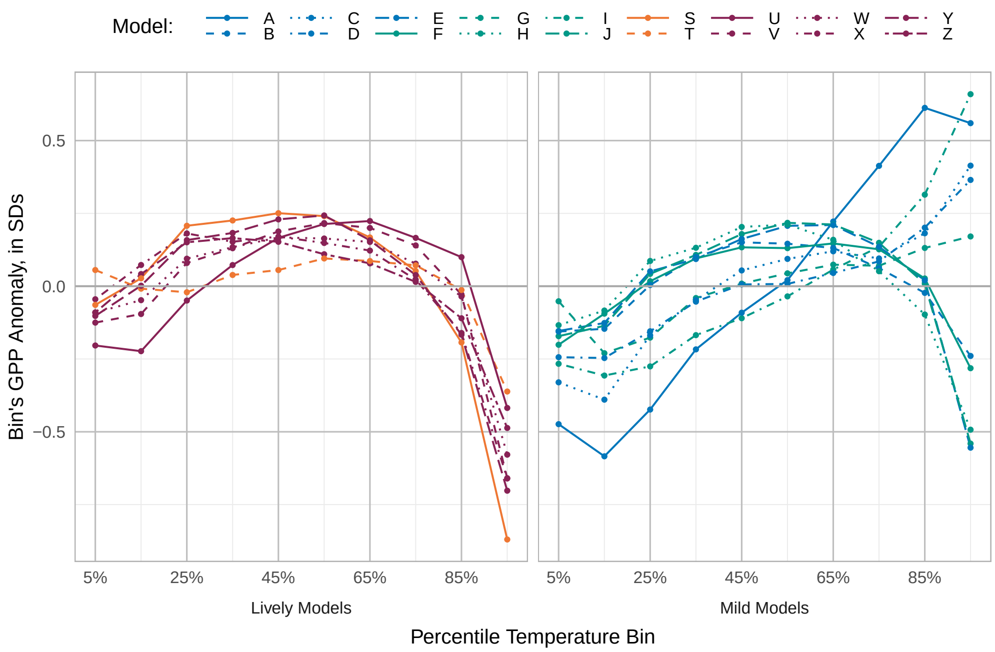


Figure 6: Non-linearity in modeled cell-level GPP responses to temperature across the Amazon basin. Monthly cell average temperature, on each panel's x-axis, is binned by deciles for all cells and months. GPP is scaled as the number of standard deviations from a particular model and cell's mean. Mild models are in the left panel and lively models in the right. At high temperature, GPP falls markedly in lively models, while the response of mild models varies widely.

Models' responses to rain and light also are non-linear (Text S4). Models with the strongest, steepest GPP response to increasing rain tend to have only modest response to increasing light and vice versa. Most lively models respond strongly to rain but have below-average GPP in the brightest months. In contrast, most of the mild models simulate their highest rainforest GPP with typical, middle decile rain amounts, and below-average GPP in the wettest months.

5.6 Lively models simulate strong, rapid drops in dry season GPP.

Responsiveness is a critical characteristic of GPP models because it describes how the model represents the consequences of climate change. An overly responsive model will predict more change than is realistic. The phase, or timing, of modeled GPP's seasonality is a corroborative assessment of responsiveness' accuracy that can be benchmarked. Seasonal timing may also suggest which model processes cause any mismatches. A slightly more re-

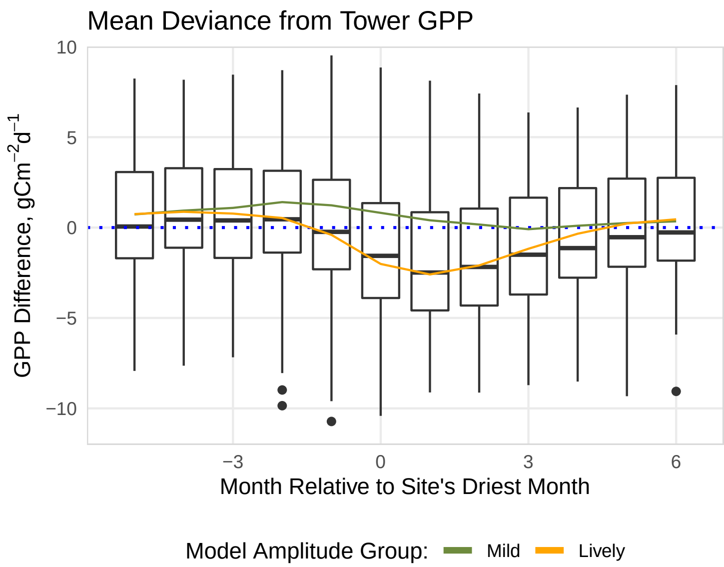


Figure 7: Seasonal deviations of GPP from EC estimates for models grouped by seasonal amplitude. On the y-axis, zero represents an exact match to EC GPP. Green and orange lines are mean deviances for mild and lively models respectively. Zero on the x-axis is the month at each EC tower with lowest average rainfall. Boxes show tendencies of all models as a group compared to ECs, with crossbars marking overall median deviations. Divergences are especially large for lively models in the latter half of the dry season.

cent version of Model J illustrates, for example, that an ESM's seasonal GPP timing can be correct in most of the world but have major inaccuracies in the world's wettest regions [*Collier et al.*, 2018, Fig. 5d].

One way to characterize seasonal timing focuses on the month of lowest GPP, as showing when the modeled forest experiences its greatest stress. Across all sites, the month with lowest modeled GPP is on average 2.6 months different from EC estimates (Fig. S3). Every model matches at least one site's time of minimum EC GPP to within one month. But with one exception, every model also is at least 5 months different from the EC estimate at one or more sites, or essentially has an opposite seasonal cycle.

Fig. 7 summarizes seasonal timing tendencies for models generalized by responsiveness group. Zero on the figure's x-axis is each site's long-term average driest month, with one month before the driest month shown as -1, two months after as +2, etc. Boxes for each month enclose the 25^{th} and 75^{th} percentiles of all models' GPP deviances from the EC estimates. Dots indicate outlier models. In months whose median value of the box plot is above zero, most models at most sites simulate higher GPP than EC estimates. Taller boxes late in the dry season show when the widest spread among models occurs.

Mild models on average, shown with the green line in Fig. 7, have relatively little systematic difference from EC estimates over the course of a year. Mildness is defined as a dampened seasonal cycle, not by similarity to EC GPP, so this result is not inevitable. Mild models tend to exceed EC estimates slightly in the two months before the driest month, or early in the dry season. A possible mechanism is insufficient modeled water stress. In contrast, lively models, whose average the orange line tracks, simulate lower GPP for 5 months starting with the driest month. For lively models, lack of plant available water during the dry season may more strongly curtail GPP than EC data suggest.

At individual sites, the mean differences between mild and lively models are sharper, with more variation in timing relative to the dry season (Fig. S14). But the overall patterning at individual sites is similar to the means (Fig. S4), with lively models on average differing more from EC estimates than mild models except at the K83 site. During transition months on either side of the dry season, boxes overlap the zero line, showing that on average the models simulate GPP close to EC estimates during the shoulder seasons. At the RJA site there is little seasonal pattern in differences between models and EC, and mild models match EC GPP more closely throughout the year. At other sites, both groups of models estimate lower GPP during the dry season than do ECs, and higher during the wettest months. The tendency for all models to simulate GPP that on average is lowest relative to EC estimates during the late dry season suggests challenges in modeling soil moisture.

6 Analysis

Flattening describes the tendency for the variance of predictions made from otherwise accurate equations with significant omitted variables or other statistical noise to be reduced. In light of flattening, this paper addresses the extent of modeled GPP's seasonal variability in Amazonian rainforest, how strongly current weather variables determine GPP at six eddy covariance sites, and the fidelity of seasonal timing. lively models are defined as those with higher seasonal amplitude than EC GPP, while mild models are less seasonal (Fig. 3).

6.1 Summary of Findings

Both process and statistical models struggle to reproduce EC estimates of Amazonian rainforest gross primary productivity. Mean and/or variance of all models' GPP falls outside of 99th percentile confidence intervals (Fig. 2). Flattening helps interpret benchmark comparisons of the variances. Flattening's degree of influence is a function of how completely a model's drivers determine predicted outcomes, leaving only a small error term in the descriptive regression. Model outcome variances that are similar to benchmark variance indicate model skill only if included drivers explain most of the variability in the reference outcomes. Otherwise, or worse if modeled variance exceeds its reference equivalent, excessive model sensitivity to drivers has overwhelmed flattening. Multiple metrics in this study suggest that lively models are overly responsive, while the mild models appear most likely to represent accurately the mean GPP consequences of climate shifts for rainforests.

The regression that predicts EC GPP might appear strong enough that process and statistical models' predicted GPP variance should be nearly as high. The regression for EC GPP that includes both weather and site-specific intercepts explains a total of 59% of variability (Equation 4). However, the descriptive regression treats intercepts as the fixed outcome of unspecified variables. With reference data for only six intercepts, this paper does not explore the important component of GPP accuracy that resides in site means and their drivers.

Separated from site effects, a linear combination of current month's rain, temperature and light explains only about an eighth (12%) of EC GPP's total variance, equal to 0.38 $gCm^{-2}d^{-1}$. Although there are severe difficulties in "measuring" GPP at a flux tower, the portion of variability explained is so low that qualitative conclusions seem warranted. Current month's weather is a weak linear determinant of rainforest GPP, and the amount of GPP variability due to weather is small. In terms of comparing modeled GPP variance to the eddy covariance estimates, flattening is a strong influence because included drivers do not largely explain EC GPP.

For the lively models, the weight of evidence favors excessive sensitivity to weather drivers. The models pass flattening's indirect test of hypersensitivity; GPP variability of all the lively models as measured by both simple variance (Fig. 2) and seasonal amplitude (Fig. 3) exceeds EC GPP variability. Direct comparisons of descriptive regression slopes are even stronger evidence of excessive sensitivity. Responsiveness to rain is stronger in every lively model than for EC GPP (1.17 average v. 0.48, Fig. 5). Perhaps in counterbalance, all statistically significant slopes for temperature and light for each lively model are of the opposite sign from EC GPP's. Finally, mismatched seasonal cycles also imply that lively models have excessive responsiveness to at least current rain. For highly seasonal models, the annual minimum in photosynthesis tends to be both later in the dry season than EC estimates, and more severe (Fig. 7).

Whether mild models are overly sensitive is less clear. Net flattening does occur, which makes excessive driver sensitivity less likely. Mild models' seasonal GPP amplitudes and in most cases variances are below EC GPP's (Figs. 2, 3). However, it is possible that other flattening influences are sufficiently strong to counteract excessive driver responsiveness. Two main contributors to the flattening are likely. One is model misspecification noise due to non-linearity in responses to weather (Figs. 6 and S2). Non-linearities were assessed only qualitatively due to EC data limitations. The second likely cause of low variance in mild models' GPP predictions is low spread in site intercepts. The range in site means for EC GPP is 4.0 $gCm^{-2}d^{-1}$. Except for one outlier, the ranges of GPP site means for mild models all are smaller, 0.8 to 3.5. Both model misspecification and low sensitivity to site mean differences could flatten the GPP predictions.

Since excessive driver responsiveness in mild models could coexist with net flattening, the sensitivity needs to be assessed more directly. One test is whether weather predictors as a group explain an appropriate amount of mild model responsiveness. If model sensitivity to weather perfectly matched the ECs', weather would explain the same absolute amount of variability as it does for the towers, 0.38 $gCm^{-2}d^{-1}$. For mild models, the average variance that weather explains is 0.79 $gCm^{-2}d^{-1}$ (range across models = 0.09 - 4.60). Given the degree of uncertainty in EC GPP, this check seems at most suggestive that some mild models respond too weakly. Direct comparison of driver slopes indicates that the mild model group's sensitivities to weather is reasonable overall, although there is quite a bit of spread among models (Fig. 5). Average mild model responsiveness to rain and light is similar to that of EC GPP, while temperature responsiveness is lower but at least of the same sign.

6.2 Assessment of Findings

Our results are specific to the scales of time and space for which they are calculated: monthly means for 6 tower sites between 2000 and 2010. Driver strengths can vary with time integration. At the K67 flux tower, for example, vapor pressure deficit and total and diffuse light largely determined hourly averaged GPP, while a derived index that also included leaf area index was better at explaining monthly averages [*Wu et al.*, 2017]. A model that reproduces hourly photosynthetic fluxes well may still have substantial biases in annual totals [*Keenan et al.*, 2012]. Spatial amalgamation even more strongly affects variability [*Rödig et al.*, 2018]. Given the grossly finite mean annual amount of atmospheric water, even though precipitation may drive local variability of GPP, temperature largely determines global variability in net land:atmosphere carbon exchange [*Jung et al.*, 2017].

Our analysis agrees with prior studies that have found rainforest GPP in global vegetation models reacts excessively to weather [*Ahlström et al.*, 2017; *Baker et al.*, 2008; *Cleveland et al.*, 2015; *Huang et al.*, 2016; *Li et al.*, 2017; *Parazoo et al.*, 2014; *Piao et al.*, 2013; *Poulter et al.*, 2009; *Restrepo-Coupe et al.*, 2016; *von Randow et al.*, 2013; *Zhu et al.*, 2016]. Excessive rainforest GPP seasonality was reported for an earlier version of Model J at K67 [*Sakaguchi et al.*, 2011], and for Model I at a flux tower in Guyana [*Zhu et al.*, 2018]. However, we found also that a few mild models have weak responses to weather.

While future trends in Amazon light and rain are uncertain, temperature are expected to rise [*Jiménez-Muñoz et al.*, 2013]. This makes model responses to temperatures that currently are outliers particularly important [*Cavaleri et al.*, 2015]. Ten of the study models have a statistically significant negative response, consistent with *Huntingford et al.* [2013] and *Poulter et al.* [2010a]. The mismatch between EC and model responses to temperature is striking (Fig. 5). Not one is as welcoming of warmer temperature as EC GPP. Disproportionately strong responses to high temperature (Fig. 6) mean that differences between model predictions will increase as the climate warms.

Independent indicators of tropical plant response to higher temperatures are mixed. At La Selva, Costa Rica, the net of GPP and respiration fell strongly with increases over 12 years in daily minimum temperatures [*Clark et al.*, 2013]. Temperature appears to be a positive and stronger driver of net ecosystem exchange globally, and precipitation a weaker driver, than is represented in most dynamic global vegetation models [*Wang et al.*, 2013]. For some models, EBF's optimal temperature parameter is influential. *MacBean et al.* [2018] found that one MsTMIP model's temperature flex point (Fig. 6) seems too low. Optimal temperature for EBF is difficult to specify well including because the real value may vary among the Amazon's thousands of tree species.

Mismatches in GPP seasonal timing (Fig. 7, Text S5), consistent with *Poulter et al.* [2010b], suggest that during the dry season plants experience less water stress than modeled. An exploratory assessment of cumulative rain as a proxy for soil moisture (Text S6) shows that while at individual sites cumulative rain is an important predictor, no lag duration works well everywhere. GPP's site-specific dependence peaks at three months at CAX and eight at BAN. When the only lag durations whose cumulative rain predictor is statistically significant at all sites, six or seven months, is used at all sites, the regression has a worse fit than if only current month's rain is included. Sites' differing optimal lag periods cancel each other when generalized and blur the importance for GPP of seasonal drought.

6.3 A water- versus light-limitation dichotomy poorly characterizes the models.

GPP increases with light in mild models, and falls with light in lively models (Fig. 5). The lively models also respond more to light strongly. It is tempting to associate each group with either strong response to light at the expense of temperature sensitivity or vice versa. If more rain causes GPP to rise enough, and tropical rain clouds reduce radiation, then GPP necessarily would fall with increasing light.

Clouds' opposing consequences for photosynthesis of more rain but less light [*Huete et al.*, 2006; *Nemani et al.*, 2003] has been labeled as light-limited versus water-limited [*Arias et al.*, 2011; *Baker et al.*, 2013, 2019; *Myneni et al.*, 2007]. If a site is water-limited, GPP falls during sunny periods due to soil dryness. At light-limited sites water is more plentiful and GPP rises during sunny periods [*Graham et al.*, 2003]. Some observational evidence contradicts the hypothesis that light limitations and water limitations represent a trade-off in the tropics, however, and instead that GPP may have little response to variations in light [*Restrepo-Coupe et al.*, 2013]. For four of the nine mild models, GPP is lowest during a dark month for at least one site, but for at least one other site GPP is lowest during a dry month when light is likely to be stronger (Fig. S3).

The dichotomy requires that light and rain be anti-correlated, with a tendency for increases in tropical clouds both to reduce light and to increase rain. In the MsTMIP driver data rain explains only 3% of the variation in light, and the correlation of EC GPP and light is negative (-0.14). Neither in the MsTMIP driver data does low rainfall bring more sensible heat in presumed Bowen ratio response to drought stress; Rain explains similarly little (4%) of the variation in temperature. While the lively models do respond strongly to precipitation, as would be expected for modeling approaches that focus mostly on water limitations, there are few indications the mild models are strongly light-limited.

6.4 Hindcast GPP's weak responses to CO_2 do not reveal predicted responses to CO_2.

CO_2 was a significant predictor of GPP only for two especially mild process models (Fig. S11). The CO_2 slope for five models was statistically indistinguishable from zero Much more than for the other weather drivers, however, the effect of CO_2 is likely to differ in coming decades from either real or modeled responses for 2000-2010.

Over time CO_2 can overwhelm trends in other environmental drivers due to its persistence and much larger relative change [*Fisher et al.*, 2013]. And GPP responses to CO_2 are unlikely to be linear, mainly because multiple and sometimes conflicting components shape a net trade-off between CO_2 fertilization and increased water use efficiency [*Swann et al.*, 2016]. Land surface models differ substantially in how strongly they amplify atmospheric CO_2 increases [*Piao et al.*, 2013]. As an example, this study's model D, Tem6, simulates logarithmic increase [*Jain et al.*, 1994]. Other methods than comparison with this EC dataset will be needed to assess the accuracy of model sensitivity to CO_2 in rainforest vegetation.

6.5 Fluxcom GPP's low variance logically reflects flattening.

Statistical models of GPP derive from a limited set of core time series data: satellites, flux towers, and ground-based weather observations. The statistical models included in this study use all three. There is no fourth independent source data with which the statistical models can be benchmarked. Some tentative conclusions about each statistical model's accuracy for the Amazon still can be drawn from the comparisons made above to flux towers, however.

There is conspicuous absence of a Fluxcom (Model G) assessment that features rainforests even though rainforest is the most productive PFT on the planet and for climate perhaps the most important. One comparison of GPP from 53 eddy covariance towers to an early version of Fluxcom included EBF sites only in Australia and Italy [*Joiner et al.*, 2014].

The defining source for Fluxcom is EC data. Reassuringly, Fluxcom's gross fit with EC GPP is among the closest for this study's models. Only model B exceeds Fluxcom's overall correlation with EC GPP (r = 0.37, Fig. S8). While Fluxcom's mean GPP in site cells is, like all but one model, outside EC credible bounds, it is among the half-dozen models closest to the EC mean. Fluxcom's rain responsiveness slope is one of the closest to EC estimates (Fig. 5). Its scaled temperature slope, 0.006, is much shallower than the ECs' slope of 0.460. But so to some degree is every other model's. The sign of Fluxcom's slope for light matches that of the EC towers although its response is stronger by almost half. Fluxcom's month of peak GPP is within two months of EC estimates for all sites except RJA (Fig. S4), again among the best matching of models.

Fluxcom's complex algorithms resemble linear regression in ways that make flattening applicable. For example, in model tree ensembles, one of Fluxcom's options, machine learning stratifies spatially and temporally defined outcomes into bins but each simulated value is the prediction of a particular bin's regression. Predictions from regressions are systematically flattened, with lower variance than the underlying data.

Flattening could help explain *Piao et al.* [2013]'s finding that Fluxcom's GPP is less variable than any of ten DGVMs'. In our study, Fluxcom's GPP variance (1.4 v. ECs' 3.3 $gCm^{-2}d^{-1}$) and seasonal amplitude for site cells (2.2 v. 3.4 $gCm^{-2}d^{-1}$) are about half as variable. Fluxcom and Wecann are less accurate for evergreen broadleaf forests (EBF) than their global average [*Alemohammad et al.*, 2017; *Badgley et al.*, 2019]. Reasons for the poor performance include the dearth of both eddy covariance towers and clear satellite retrievals in the tropics [*Tramontana et al.*, 2016; *Jung et al.*, 2020]. Fluxcom would thus likely have especially well-smoothed tropical predictions.

Fluxcom's globally low GPP variability has been called "undersampled" [*Piao et al.*, 2013], "poorly captured" [*Tramontana et al.*, 2016], underestimated for reasons that are not fully clear [*Jung et al.*, 2020], and, on the product's website, "too small" (https://www.bgc-jena.mpg.de/geodb/projects/Data.php). The tendency bears consideration when using Fluxcom. But in light of flattening, we disagree that Fluxcom's mildness is necessarily a weakness. The low variance appears instead to be an inherent consequence of omitted drivers and flattening, and suggests theoretically good potential for relatively accurate driver responsiveness globally.

Wecann's (Model F's) temporal span prevents reasonable direct comparisons to the EC towers used in this study. In each comparison across the Amazon basin, Wecann resembles Fluxcom (Text S3 and Figs. 6, S2, S10, S11, and S12). GPP differs at the warmest decile

(Fig. 6). Wecann is slightly less sensitive to light than is Fluxcom and more so to CO_2 (Fig. S11). Current weather explains less of the variance. But the models' residual standard errors (Fig. S12), which tend to indicate the extent of mismatch in a cycle's phase [*Taylor*, 2001], are similar. The detailed nature of these differences reinforce Wecann and Fluxcom's overall similarity.

While source data for Fluxcom and Wecann feature eddy covariance towers, VPM (Model) emphasizes satellite sources. It uses especially tight coupling and straightforward equations to merge remotely sensed products. VPM's unusual spatial pattern of rain responsiveness in the Amazon (Fig. 1, Model I) corresponds to likely seasonal trends in cloud contamination of satellite data retrievals. Cloud cover and therefore scarcity of acceptable satellite retrievals globally tends to peak in the tropics and during the wettest months. Gap-filled or missing data therefore overlap heavily with periods of high greenness and potentially of peak rainforest GPP. Photosynthetically active radiation from the NCEP II weather reanalysis is the model's multiplicative component that cloudiness is most likely to skew. Radiation from weather reanalyses was specifically omitted as an input to another statistical model due to the product's high uncertainty [*Gentine et al.*, 2019; *Jung et al.*, 2011].

VPM is more strongly anticorrelated with EC GPP than any other model (r = -0.19). It is the only model for which rain's linear regression slope is negative. GPP falls with increasing precipitation for all but the driest decile (Fig. S2). VPM's response to light also is an outlier, increasing at every decile with no inflection point (Fig. 6). In terms of the month with lowest GPP at each site, while other models' average timing difference from ECs ranges from 1.5 to 3.3 months, VPM averages 4.5 months (Fig. S3). The phase of VPM's seasonal cycle is nearly opposite that of EC GPP at most sites. At no site does VPM simulate minimum GPP during a dry season month. While VPM's GPP estimates are unrepresentative of the best reference data available for the Amazon, a logically underlying reason of cloud contamination applies most strongly to the tropics and could have little effect on VPM's accuracy elsewhere.

In summary, Fluxcom's match with EC GPP is weak. Their correlation is 0.37. But the fit is better than for almost all process models. Wecann compares similarly. The fidelity does not establish Fluxcom's or Wecann's veracity, merely their anticipated conformity with EC GPP in all its imperfections. Of this study's models, Fluxcom and Wecann appear to be the best currently available wall-to-wall estimates of mean Amazon GPP.

6.6 Flattening has practical implications for ESMs.

Flattening certainly does not mean that ESM outputs necessarily have low variability. GPP in the Amazon is a case in point. For many ESM subprocesses, other sources of model uncertainty and imperfection remain large enough to obscure and/or overwhelm flattening due to omitted random variables.

Flattening is likely to be a dominant problem for models privileged to have high accuracy (modest RSE) combined with low precision (low r^2). These models suffer little from uncertainty about included predictors' responsiveness but lack some key drivers. Currently it is flattening's side effects that are most pertinent to ESMs. As described below, they affect studies of climate outliers; intermediate model calculations; trade-offs in the development, assessment, and use of models; and consequences of increasing model complexity.

There is enormous practical value in understanding change in the frequencies of droughts, wildland fires, heat waves, floods, and tropical cyclones [*Katz and Brown*, 1992]. Studies of climate outliers' consequences often base predictions on driver change or z-scores [*Abatzoglou and Williams*, 2016, for example] rather than on the variability of predictions directly, sidestepping any bias in variance of predictions that results from flattening. A conceptually similar option is to build model predictor metrics from simulated history [*Camargo*, 2013] rather than from observed driver history [*Westerling et al.*, 2011].

Feedbacks mean that variability internal to a model can affect mean outcomes. Precipitation intensity in the Amazon rainforest is an illustration. In typical ESM runs, rainfall depth is spread uniformly across a grid cell and can simulate overabundant frequency and duration of light mist [*Baker et al.*, 2019]. Rain reaching the soil is a step function that breaks when rain exceeds the depth that leaf surfaces can store. Modeled rainforest foliage remains wet far longer than it actually does after the region's legend cloudbursts. So much water evaporates from leaves that soil recharge is weak, in turn reducing GPP. Cloud superparameterization distributes rain among and within virtual sub-cells rather than spreading it uniformly, improving representation of soil moisture. Greater variability in rainfall intensity within each cell was required for simulating accurate mean rates of plant processes. But the superparameterization adjustment creates new inaccuracies in modeled global precipitation patterns [*Phillips*, 2019].

When variance of intermediate outputs needs to be similar to actual variability, feasible workarounds may be scarce. Theoretically they include explicit addition of statistical noise in the form of deterministic or random draws from exogenous distributions [*Pelletier*, 1997; *Khodaparast et al.*, 2008], finer temporal or spatial scaling, and ensemble modeling of probabilistic outcomes. An example of the first is LPJmL's semi-stochastic distribution of monthly rainfall to individual days [*Poulter et al.*, 2010a]. Fuzzy parameters can be used to address not uncertainty in parameter estimates [*Hoffman and Miller*, 1983; *Ersoy and Yünsel*, 2006] but as a proxy for statistical noise due to omitted variables.

Driver sensitivity affects both trends and variability of outcomes [*Nijsse et al.*, 2019]. As long as models lack some influential real drivers, optimizing the predicted variance of predictions can compromise the accuracy of responsiveness to drivers and vice versa. Ideally the trade-offs are deliberate and publicly documented. The short list of diagnostics to which the Max Planck Institute's ESM was finely-tuned for CMIP5 included both means and variabilities [*Mauritsen et al.*, 2012]. Especially when omitted variables are known to be highly influential, conversations about the relative importance of sensitivity versus variance for specific equations, models or applications may be fruitful. For example, flattening may affect benchmark selection; Fluxcom appears likely to be reasonable if noisy reference data for GPP's driver sensitivity.

Weather sensitivity describes, by definition, the consequences of a changing climate. It is important to focus at least as much on the accuracy of sensitivity as on predictions' variance despite possibly less certain benchmark data. Accuracy of outcome variances already is integral to ILAMB *Collier et al.* [2018] and to many model intercomparison projects [*Houghton et al.*, 2001; *Jupp et al.*, 2010; *Li et al.*, 2019]. Responsiveness to drivers tends to be more difficult to benchmark than the variance of outputs. But inadvertently prioritizing spread over responsiveness in accuracy assessments can be counterproductive.

As ESMs become increasingly accurate for the processes and drivers they include, and incorporate more of the drivers that cause real variability, simulated variance will tend to rise. Processes and drivers added over time to the ESMs in IPCC's Comprehensive Assessment Reports have done little to reduce uncertainty around mean temperature trends but higher complexity has increased the accuracy of simulated variability [*Dahan*, 2010]. There are practical limits to adding global drivers to climate models, however.

7 Conclusions

We compared 15 process models and three statistical models to GPP estimates from six eddy covariance towers in the Amazon rainforest. Models split almost equally between weaker and stronger responsiveness than EC data to the environmental drivers of current month's rain, temperature, and light. Most striking about the wide spreads across modeled GPP's means, responsiveness to drivers, and amplitude of the annual cycle is how little virtually every model resembles EC estimates. Similarity to Amazon flux towers is one of many important ESM accuracy metrics. Models that poorly replicate driver responsiveness in tropical rainforest may do very well by other criteria. The implication, provided EC GPP is somewhat realistic, is that lively models overreact to rain and have opposite signs of response for light and temperature. Since temperature is likely to rise and rain to become more variable, the liveliest models may substantially exaggerate the Amazon's future change and peril.

As this article's title asserts, accurate deterministic simulation of both sensitivity to drivers and variability for Amazonian GPP is unattainable. The reason is that weather explains so little of EC GPP variability that flattening is a strong influence. Of wider relevance is the role of omitted processes and uncertainty due to other sources of modeling error in reducing the variability of model predictions. It is generically appropriate to be more skeptical of too much variability than of too little. In the interest of accurate sensitivity, low variance of predictions relative to a benchmark may sometimes deserve acclaim.

Acknowledgments

We are grateful for the many people who managed and operated flux towers in the Amazon, developed the models assessed, and participated in MsTMIP. Ideas shared freely at Cross Validated (Stackexchange) improved myriad details of this study. GPP for MsTMIP [*Huntzinger et al.*, 2014] was downloaded from https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1225, and driver data [*Wei et al.*, 2014] from https://daac.ornl.gov/NACP/guides/NACP_MsTMIP_Model_Driver.html. For statistical model GPP, Fluxcom [*Jung et al.*, 2019; *Tramontana et al.*, 2016] was downloaded from https://www.bgc-jena.mpg.de/geodb/projects/Data.php, Wecann [*Alemohammad et al.*, 2017] from https://avdc.gsfc.nasa.gov/pub/data/project/WECANN/, and VPM [*Xiao et al.*, 2005; *Zhang et al.*, 2017] from https://figshare.com/articles/Monthly_GPP_at_0_5_degree/5048011.

EC data collection and analysis was funded by the National Aeronautics and Space Administration (NASA) LBA investigation CD-32, and NASA LBA-DMIP project (NNX09AL52G). EC GPP used in this study [*Restrepo-Coupe et al.*, 2013] is archived at [reviewers - archiving is underway. Url will go here.] SiB4 GPP derives from coauthors' prior work. The authors declare no conflict of interest. We thank the Department of Energy (DE-SC0014438) and the National Aeronautics and Space Administration (NNX14AI52G) for funding and support.

References

Abatzoglou, J. T., and A. P. Williams (2016), Impact of anthropogenic climate change on wildfire across western US forests, *Proceedings of the National Academy of Sciences*, *113*(42), 11,770–11,775, doi:10.1073/pnas.1607171113.

Ahlström, A., J. G. Canadell, G. Schurgers, M. Wu, J. A. Berry, K. Guan, and R. B. Jackson (2017), Hydrologic resilience and Amazon productivity, *Nature Communications*, *8*(1), doi:10.1038/s41467-017-00306-z.

Alemohammad, S. H., B. Fang, A. G. Konings, F. Aires, J. K. Green, J. Kolassa, D. Miralles, C. Prigent, and P. Gentine (2017), Water, Energy, and Carbon with Artificial Neural Networks (WECANN): A statistically based estimate of global surface turbulent fluxes and gross primary productivity using solar-induced fluorescence, *Biogeosciences*, *14*(18), 4101–4124, doi:10.5194/bg-14-4101-2017.

Alster, C. J., J. C. von Fischer, S. D. Allison, and K. K. Treseder (2020), Embracing a new paradigm for temperature sensitivity of soil microbes, *Global Change Biology*, *26*(6), 3221–3229, doi:10.1111/gcb.15053.

Anav, A., P. Friedlingstein, C. Beer, P. Ciais, A. Harper, C. Jones, G. Murray-Tortarolo, D. Papale, N. C. Parazoo, P. Peylin, S. Piao, S. Sitch, N. Viovy, A. Wiltshire, and M. Zhao (2015), Spatiotemporal patterns of terrestrial gross primary production: A review, *Reviews of Geophysics*, *53*(3), 785–818, doi:10.1002/2015RG000483.

Ardö, J. (2015), Comparison between remote sensing and a dynamic vegetation model for estimating terrestrial primary production of Africa, *Carbon Balance and Management*, *10*(1), doi:10.1186/s13021-015-0018-5.

Arias, P. A., R. Fu, C. D. Hoyos, W. Li, and L. Zhou (2011), Changes in cloudiness over the Amazon rainforests during the last two decades: Diagnostic and potential causes, *Climate Dynamics*, *37*(5-6), 1151–1164, doi:10.1007/s00382-010-0903-2.

Badgley, G., L. D. L. Anderegg, J. A. Berry, and C. B. Field (2019), Terrestrial gross primary production: Using NIRV to scale from site to globe, *Global Change Biology*, *25*(11), 3731–3740, doi:10.1111/gcb.14729.

Baker, I., A. Harper, H. da Rocha, A. Denning, A. Araújo, L. Borma, H. Freitas, M. Goulden, A. Manzi, S. Miller, A. Nobre, N. Restrepo-Coupe, S. Saleska, R. Stöckli, C. von Randow, and S. Wofsy (2013), Surface ecophysiological behavior across vegetation and moisture gradients in tropical South America, *Agricultural and Forest Meteorology*, *182-183*, 177–188, doi:10.1016/j.agrformet.2012.11.015.

Baker, I. T., L. Prihodko, A. S. Denning, M. Goulden, S. Miller, and H. R. da Rocha (2008), Seasonal drought stress in the Amazon: Reconciling models and observations, *Journal of Geophysical Research: Biogeosciences*, *113*, 1–10, doi:10.1029/2007JG000644.

Baker, I. T., A. Denning, D. A. Dazlich, A. B. Harper, M. D. Branson, D. A. Randall, M. C. Phillips, K. D. Haynes, and S. M. Gallup (2019), Surface-Atmosphere Coupling Scale, the Fate of Water, and Ecophysiological Function in a Brazilian Forest, *Journal of Advances in Modeling Earth Systems*, *11*(8), 2523–2546, doi:10.1029/2019MS001650.

Barkhordarian, A., H. von Storch, E. Zorita, P. C. Loikith, and C. R. Mechoso (2017), Observed warming over northern South America has an anthropogenic origin, *Climate Dynamics*, doi:10.1007/s00382-017-3988-z.

Bathiany, S., V. Dakos, M. Scheffer, and T. M. Lenton (2018), Climate models predict increasing temperature variability in poor countries, *Science Advances*, *4*(5), eaar5809, doi:10.1126/sciadv.aar5809.

Beer, C., M. Reichstein, E. Tomelleri, P. Ciais, M. Jung, N. Carvalhais, C. Rodenbeck, M. A. Arain, D. Baldocchi, G. B. Bonan, A. Bondeau, A. Cescatti, G. Lasslop, A. Lindroth, M. Lomas, S. Luyssaert, H. Margolis, K. W. Oleson, O. Roupsard, E. Veenendaal, N. Viovy, C. Williams, F. I. Woodward, and D. Papale (2010), Terrestrial Gross Carbon Dioxide Uptake: Global Distribution and Covariation with Climate, *Science*, *329*(5993), 834–838, doi:10.1126/science.1184984.

Bousquet, P., P. Peylin, P. Ciais, C. Le Quéré, P. Friedlingstein, and P. P. Tans (2000), Regional changes in carbon dioxide fluxes of land and oceans since 1980, *Science*, *290*(5495), 1342–1346, doi:10.1126/science.290.5495.1342.

Camargo, S. J. (2013), Global and Regional Aspects of Tropical Cyclone Activity in the CMIP5 Models, *Journal of Climate*, *26*(24), 9880–9902, doi:10.1175/JCLI-D-12-00549.1.

Cavaleri, M. A., S. C. Reed, W. K. Smith, and T. E. Wood (2015), Urgent need for warming experiments in tropical forests, *Global Change Biology*, *21*(6), 2111–2121, doi:10.1111/gcb.12860.

Chadwick, R., P. Good, G. Martin, and D. P. Rowell (2015), Large rainfall changes consistently projected over substantial areas of tropical land, *Nature Climate Change*, *6*, 177–181, doi:10.1038/nclimate2805.

Chevan, A., and M. Sutherland (1991), Hierarchical Partitioning, *The American Statistician*, *45*(2), 90, doi:10.2307/2684366.

Christoffersen, B. O., N. Restrepo-Coupe, M. A. Arain, I. T. Baker, B. P. Cestaro, P. Ciais, J. B. Fisher, D. Galbraith, X. Guan, L. Gulden, B. van den Hurk, K. Ichii, H. Imbuzeiro, A. Jain, N. Levine, G. Miguez-Macho, B. Poulter, D. R. Roberti, K. Sakaguchi, A. Sahoo, K. Schaefer, M. Shi, H. Verbeeck, Z.-L. Yang, A. C. Araújo, B. Kruijt, A. O. Manzi, H. R. da Rocha, C. von Randow, M. N. Muza, J. Borak, M. H. Costa, L. G. Gonçalves de Gonçalves, X. Zeng, and S. R. Saleska (2014), Mechanisms of water supply and vegetation demand govern the seasonality and magnitude of evapotranspiration in Amazonia and Cerrado, *Agricultural and Forest Meteorology*, *191*, 33–50, doi:10.1016/j.agrformet.2014.02.008.

Clark, D. A., D. B. Clark, and S. F. Oberbauer (2013), Field-quantified responses of tropical rainforest aboveground productivity to increasing CO_2 and climatic stress, 1997-2009, *Journal of Geophysical Research: Biogeosciences*, *118*(2), 783–794, doi:10.1002/jgrg.20067.

Clark, D. A., S. Asao, R. Fisher, S. Reed, P. B. Reich, M. G. Ryan, T. E. Wood, and X. Yang (2017), Reviews and syntheses: Field data to benchmark the carbon cycle models for tropical forests, *Biogeosciences*, *14*(20), 4663–4690, doi:10.5194/bg-14-4663-2017.

Cleveland, C. C., P. Taylor, K. D. Chadwick, K. Dahlin, C. E. Doughty, Y. Malhi, W. K. Smith, B. W. Sullivan, W. R. Wieder, and A. R. Townsend (2015), A comparison of plot-based satellite and Earth system model estimates of tropical forest net primary production, *Global Biogeochemical Cycles*, *29*(5), 626–644, doi:10.1002/2014GB005022.

Collier, N., F. M. Hoffman, D. M. Lawrence, G. Keppel-Aleks, C. D. Koven, W. J. Riley, M. Mu, and J. T. Randerson (2018), The International Land Model Benchmarking (ILAMB) System: Design, Theory, and Implementation, *Journal of Advances in Modeling Earth Systems*, pp. 1–24, doi:10.1029/2018MS001354.

Corlett, R. T. (2011), Impacts of warming on tropical lowland rainforests, *Trends in Ecology & Evolution*, *26*(11), 606–613, doi:10.1016/j.tree.2011.06.015.

Cox, P. M., D. Pearson, B. B. Booth, P. Friedlingstein, C. Huntingford, C. D. Jones, and C. M. Luke (2013), Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability, *Nature*, *494*(7437), 341–344, doi:10.1038/nature11882.

da Rocha, H. R., A. O. Manzi, O. M. Cabral, S. D. Miller, M. L. Goulden, S. R. Saleska, N. Restrepo-Coupe, S. C. Wofsy, L. S. Borma, P. Artaxo, G. Vourlitis, J. S. Nogueira, F. L. Cardoso, A. D. Nobre, B. Kruijt, H. C. Freitas, C. von Randow, R. G. Aguiar, and J. F. Maia (2009), Patterns of water and heat flux across a biome gradient from tropical forest to savanna in Brazil, *Journal of Geophysical Research*, *114*, 1–8, doi: 10.1029/2007JG000640.

Dahan, A. (2010), Putting the Earth System in a numerical box? The evolution from climate modeling toward global change, *Studies in History and Philosophy of Modern Physics*, *41*(3), 282–292, doi:10.1016/j.shpsb.2010.08.002.

De Kauwe, M. G., J. Kala, Y.-S. Lin, A. J. Pitman, B. E. Medlyn, R. A. Duursma, G. Abramowitz, Y.-P. Wang, and D. G. Miralles (2015), A test of an optimal stomatal conductance scheme within the CABLE land surface model, *Geoscientific Model Development*, *8*(2), 431–452, doi:10.5194/gmd-8-431-2015.

Ersoy, A., and T. Y. Yünsel (2006), Geostatistical Conditional Simulation for the Assessment of the Quality Characteristics of Çayırhan Lignite Deposits, *Energy Exploration and Exploitation*, *24*(6), 391–416.

Feng, X., A. Porporato, and I. Rodriguez-Iturbe (2013), Changes in rainfall seasonality in the tropics, *Nature Climate Change*, *3*(9), 811–815, doi:10.1038/nclimate1907.

Fisher, J. B., M. Sikka, S. Sitch, P. Ciais, B. Poulter, D. Galbraith, J.-E. Lee, C. Huntingford, N. Viovy, N. Zeng, A. Ahlstrom, M. R. Lomas, P. E. Levy, C. Frankenberg, S. Saatchi, and Y. Malhi (2013), African tropical rainforest net carbon dioxide fluxes in the twentieth century, *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1625), 20120,376–20120,376, doi:10.1098/rstb.2012.0376.

Friedlingstein, P., P. Cox, R. Betts, L. Bopp, W. von Bloh, V. Brovkin, P. Cadule, S. Doney, M. Eby, I. Fung, G. Bala, J. John, C. Jones, F. Joos, T. Kato, M. Kawamiya, W. Knorr, K. Lindsay, H. D. Matthews, T. Raddatz, P. Rayner, C. Reick, E. Roeckner, K.-G. Schnitzler, R. Schnur, K. Strassmann, A. J. Weaver, C. Yoshikawa, and N. Zeng (2006), Climate–Carbon Cycle Feedback Analysis: Results from the C^{4} MIP Model Intercomparison, *Journal of Climate*, *19*(14), 3337–3353, doi:10.1175/JCLI3800.1.

Fu, R., L. Yin, W. Li, P. A. Arias, R. E. Dickinson, L. Huang, S. Chakraborty, K. Fernandes, B. Liebmann, R. Fisher, and R. B. Myneni (2013), Increased dry-season length over southern Amazonia in recent decades and its implication for future climate projection, *Proceedings of the National Academy of Sciences*, *110*(45), 18,110–18,115, doi: 10.1073/pnas.1302584110.

Gentine, P., A. Massmann, B. R. Lintner, S. Hamed Alemohammad, R. Fu, J. K. Green, D. Kennedy, and J. Vilà-Guerau de Arellano (2019), Land–atmosphere interactions in the tropics – a review, *Hydrology and Earth System Sciences*, *23*(10), 4171–4197, doi:10.5194/hess-23-4171-2019.

Gloor, M., L. Gatti, R. Brienen, T. R. Feldpausch, O. L. Phillips, J. Miller, J. P. Ometto, H. Rocha, T. Baker, B. de Jong, R. A. Houghton, Y. Malhi, L. E. O. C. Aragão, J.-L. Guyot, K. Zhao, R. Jackson, P. Peylin, S. Sitch, B. Poulter, M. Lomas, S. Zaehle, C. Huntingford, P. Levy, and J. Lloyd (2012), The carbon balance of South America: A review of the status, decadal trends and main determinants, *Biogeosciences*, *9*(12), 5407–5430, doi:10.5194/bg-9-5407-2012.

Gloor, M., R. J. W. Brienen, D. Galbraith, T. R. Feldpausch, J. Schöngart, J.-L. Guyot, J. C. Espinoza, J. Lloyd, and O. L. Phillips (2013), Intensification of the Amazon hydrological cycle over the last two decades, *Geophysical Research Letters*, *40*(9), 1729–1733, doi: 10.1002/grl.50377.

Graham, E. A., S. S. Mulkey, K. Kitajima, N. G. Phillips, and S. J. Wright (2003), Cloud cover limits net CO_2 uptake and growth of a rainforest tree during tropical rainy seasons, *Proceedings of the National Academy of Sciences*, *100*(2), 572–576.

Greene, W. H. (2012), *Econometric Analysis*, 7th ed ed., Prentice Hall, Boston.

Hamby, D. M. (1994), A Review of Techniques for Parameter Sensitivity Analysis of Environmental Models, *Environmental Monitoring and Assessment*, *32*, 135–154, doi: 10.1007/BF00547132.

Harper, A., I. T. Baker, A. S. Denning, D. A. Randall, D. Dazlich, and M. Branson (2014), Impact of Evapotranspiration on Dry Season Climate in the Amazon Forest, *Journal of Climate*, *27*(2), 574–591, doi:10.1175/JCLI-D-13-00074.1.

Haynes, K. D., I. T. Baker, A. S. Denning, R. Stöckli, K. Schaefer, E. Y. Lokupitiya, and J. M. Haynes (2019a), Representing Grasslands Using Dynamic Prognostic Phenology Based on Biological Growth Stages: 1. Implementation in the Simple Biosphere

Model (SiB4), *Journal of Advances in Modeling Earth Systems*, pp. 1–26, doi:10.1029/2018MS001540.

Haynes, K. D., I. T. Baker, A. S. Denning, S. Wolf, G. Wohlfahrt, G. Kiely, R. C. Minaya, and J. M. Haynes (2019b), Representing Grasslands Using Dynamic Prognostic Phenology Based on Biological Growth Stages: Part 2. Carbon Cycling, *Journal of Advances in Modeling Earth Systems*, pp. 1–17, doi:10.1029/2018MS001541.

Hoffman, F. O., and C. W. Miller (1983), Uncertainties in environmental radiological assessment models and their implications, in *Annual Meeting of the National Council of Radiation Protection and Measurements*, pp. 1–57, Washington D.C., doi:[noDOI].

Houghton, R. A., K. T. Lawrence, J. L. Hackler, and S. Brown (2001), The spatial distribution of forest biomass in the Brazilian Amazon: A comparison of estimates, *Global Change Biology*, *7*(7), 731–746, doi:10.1111/j.1365-2486.2001.00426.x.

Huang, Y., S. Gerber, T. Huang, and J. W. Lichstein (2016), Evaluating the drought response of CMIP5 models using global gross primary productivity, leaf area, precipitation, and soil moisture data, *Global Biogeochemical Cycles*, *30*(12), 1827–1846, doi:10.1002/2016GB005480.

Huete, A. R., K. Didan, Y. E. Shimabukuro, P. Ratana, S. R. Saleska, L. R. Hutyra, W. Yang, R. R. Nemani, and R. Myneni (2006), Amazon rainforests green-up with sunlight in dry season, *Geophysical Research Letters*, *33*(6), doi:10.1029/2005GL025583.

Huntingford, C., P. Zelazowski, D. Galbraith, L. M. Mercado, S. Sitch, R. Fisher, M. Lomas, A. P. Walker, C. D. Jones, B. B. B. Booth, Y. Malhi, D. Hemming, G. Kay, P. Good, S. L. Lewis, O. L. Phillips, O. K. Atkin, J. Lloyd, E. Gloor, J. Zaragoza-Castells, P. Meir, R. Betts, P. P. Harris, C. Nobre, J. Marengo, and P. M. Cox (2013), Simulated resilience of tropical rainforests to CO2-induced climate change, *Nature Geoscience*, *6*(4), 268–273, doi:10.1038/ngeo1741.

Huntzinger, D., C. Schwalm, Y. Wei, R. Cook, A. Michalak, K. Schaefer, A. Jacobson, M. Arain, J. Fisher, D. Hayes, M. Huang, S. Huang, A. Ito, H. Lei, C. Lu, F. Maignan, J. Mao, N. Parazoo, C. Peng, S. Peng, B. Poulter, D. Ricciuto, H. Tian, X. Shi, W. Wang, N. Zeng, F. Zhao, and Q. Zhu (2014), NACP MsTMIP: Global 0.5-deg Terrestrial Biosphere Model Outputs in Standard Format, doi:10.3334/ORNLDAAC/1225.

Jain, A. K., D. J. Wuebbles, and H. S. Kheshgi (1994), Integrated Science Model for Assessment of Climate Change, in *Annual Meeting and Exhibition of the Air and Waste Management Association*, pp. 1–19, Cincinnati, OH.

Jiménez-Muñoz, J. C., J. A. Sobrino, C. Mattar, and Y. Malhi (2013), Spatial and temporal patterns of the recent warming of the Amazon forest, *Journal of Geophysical Research: Atmospheres*, *118*(11), 5204–5215, doi:10.1002/jgrd.50456.

Joiner, J., Y. Yoshida, A. Vasilkov, K. Schaefer, M. Jung, L. Guanter, Y. Zhang, S. Garrity, E. Middleton, K. Huemmrich, L. Gu, and L. Belelli Marchesini (2014), The seasonal cycle of satellite chlorophyll fluorescence observations and its relationship to vegetation phenology and ecosystem atmosphere carbon exchange, *Remote Sensing of Environment*, *152*, 375–391, doi:10.1016/j.rse.2014.06.022.

Jung, M., M. Reichstein, H. A. Margolis, A. Cescatti, A. D. Richardson, M. A. Arain, A. Arneth, C. Bernhofer, D. Bonal, J. Chen, D. Gianelle, N. Gobron, G. Kiely, W. Kutsch, G. Lasslop, B. E. Law, A. Lindroth, L. Merbold, L. Montagnani, E. J. Moors, D. Papale, M. Sottocornola, F. Vaccari, and C. Williams (2011), Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, *Journal of Geophysical Research: Biogeosciences*, *116*(G3), doi:10.1029/2010JG001566.

Jung, M., M. Reichstein, C. R. Schwalm, C. Huntingford, S. Sitch, A. Ahlström, A. Arneth, G. Camps-Valls, P. Ciais, P. Friedlingstein, F. Gans, K. Ichii, A. K. Jain, E. Kato, D. Papale, B. Poulter, B. Raduly, C. Rödenbeck, G. Tramontana, N. Viovy, Y.-P. Wang, U. Weber, S. Zaehle, and N. Zeng (2017), Compensatory water effects link yearly global land CO2 sink changes to temperature, *Nature*, *541*(7638), 516–520, doi:10.1038/nature20780.

Jung, M., S. Koirala, U. Weber, K. Ichii, F. Gans, Gustau-Camps-Valls, D. Papale, C. Schwalm, G. Tramontana, and M. Reichstein (2019), The FLUXCOM ensemble of global land-atmosphere energy fluxes, *Scientific Data*, *6*(1), 74.

Jung, M., C. Schwalm, M. Migliavacca, S. Walther, G. Camps-Valls, S. Koirala, P. Anthoni, S. Besnard, P. Bodesheim, N. Carvalhais, F. Chevallier, F. Gans, D. S. Goll, V. Haverd, P. Köhler, K. Ichii, A. K. Jain, J. Liu, D. Lombardozzi, J. E. M. S. Nabel, J. A. Nelson, M. O'Sullivan, M. Pallandt, D. Papale, W. Peters, J. Pongratz, C. Rödenbeck, S. Sitch, G. Tramontana, A. Walker, U. Weber, and M. Reichstein (2020), Scaling carbon fluxes from eddy covariance sites to globe: Synthesis and evaluation of the FLUXCOM approach, *Biogeosciences*, *17*(5), 1343–1365, doi:10.5194/bg-17-1343-2020.

Jupp, T. E., P. M. Cox, A. Rammig, K. Thonicke, W. Lucht, and W. Cramer (2010), Development of probability density functions for future South American rainfall, *New Phytologist*, *187*(3), 682–693, doi:10.1111/j.1469-8137.2010.03368.x.

Katz, R. W., and B. G. Brown (1992), Extreme Events in a Changing Climate: Variability is More Important than Averages, *Climatic Change*, *21*, 289–302.

Keenan, T. F., E. Davidson, A. M. Moffat, W. Munger, and A. D. Richardson (2012), Using model-data fusion to interpret past trends, and quantify uncertainties in future projections, of terrestrial ecosystem carbon cycling, *Global Change Biology*, *18*(8), 2555–2569, doi:10.1111/j.1365-2486.2012.02684.x.

Khodaparast, H. H., J. E. Mottershead, and M. I. Friswell (2008), Perturbation methods for the estimation of parameter variability in stochastic model updating, *Mechanical Systems and Signal Processing*, *22*(8), 1751–1773, doi:10.1016/j.ymssp.2008.03.001.

Li, F., M. V. Martin, M. O. Andreae, A. Arneth, S. Hantson, J. W. Kaiser, G. Lasslop, C. Yue, D. Bachelet, M. Forrest, E. Kluzek, X. Liu, S. Mangeon, J. R. Melton, D. S. Ward, A. Darmenov, T. Hickler, C. Ichoku, B. I. Magi, S. Sitch, G. R. V. D. Werf, C. Wiedinmyer, and S. S. Rabin (2019), Historical (1700-2012) global multi-model estimates of the fire emissions from the Fire Modeling Intercomparison Project (FireMIP), *Atmospheric Chemistry and Physics*, *19*(19), 12,545–12,567, doi:10.5194/acp-19-12545-2019.

Li, L., Y.-P. Wang, J. Beringer, H. Shi, J. Cleverly, L. Cheng, D. Eamus, A. Huete, L. Hutley, X. Lu, S. Piao, L. Zhang, Y. Zhang, and Q. Yu (2017), Responses of LAI to rainfall explain contrasting sensitivities to carbon uptake between forest and non-forest ecosystems in Australia, *Scientific Reports*, *7*(1), doi:10.1038/s41598-017-11063-w.

Li, W., R. Fu, and R. E. Dickinson (2006), Rainfall and its seasonality over the Amazon in the 21st century as assessed by the coupled models for the IPCC AR4, *Journal of Geophysical Research: Atmospheres*, *111*(D2), doi:10.1029/2005JD006355.

Li, W., R. Fu, R. I. N. Juárez, and K. Fernandes (2008), Observed change of the standardized precipitation index, its potential cause and implications to future climate change in the Amazon region, *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1498), 1767–1772, doi:10.1098/rstb.2007.0022.

Lopes, A. V., J. C. H. Chiang, S. A. Thompson, and J. A. Dracup (2016), Trend and uncertainty in spatial-temporal patterns of hydrological droughts in the Amazon basin: Hydrological Droughts in the Amazon, *Geophysical Research Letters*, *43*(7), 3307–3316, doi:10.1002/2016GL067738.

MacBean, N., F. Maignan, C. Bacour, P. Lewis, P. Peylin, L. Guanter, P. Köhler, J. Gómez-Dans, and M. Disney (2018), Strong constraint on modelled global carbon uptake using solar-induced chlorophyll fluorescence data, *Scientific Reports*, *8*(1), doi:10.1038/

s41598-018-20024-w.

Malhi, Y., L. E. O. C. Aragão, D. Galbraith, C. Huntingford, R. Fisher, P. Zelazowski, S. Sitch, C. McSweeney, and P. Meir (2009), Exploring the likelihood and mechanism of a climate-change-induced dieback of the Amazon rainforest, *Proceedings of the National Academy of Sciences*, *106*(49), 20,610–20,615, doi:10.1073/pnas.0804619106.

Mauritsen, T., B. Stevens, E. Roeckner, T. Crueger, M. Esch, M. Giorgetta, H. Haak, J. Jungclaus, D. Klocke, D. Matei, U. Mikolajewicz, D. Notz, R. Pincus, H. Schmidt, and L. Tomassini (2012), Tuning the climate of a global model, *Journal of Advances in Modeling Earth Systems*, *4*(3), doi:10.1029/2012MS000154.

Myneni, R. B., W. Yang, R. R. Nemani, A. R. Huete, R. E. Dickinson, Y. Knyazikhin, K. Didan, R. Fu, R. I. Negron Juarez, S. S. Saatchi, H. Hashimoto, K. Ichii, N. V. Shabanov, B. Tan, P. Ratana, J. L. Privette, J. T. Morisette, E. F. Vermote, D. P. Roy, R. E. Wolfe, M. A. Friedl, S. W. Running, P. Votava, N. El-Saleous, S. Devadiga, Y. Su, and V. V. Salomonson (2007), Large seasonal swings in leaf area of Amazon rainforests, *Proceedings of the National Academy of Sciences*, *104*(12), 4820–4823, doi:10.1073/pnas.0611338104.

Mystakidis, S., E. L. Davin, N. Gruber, and S. I. Seneviratne (2016), Constraining future terrestrial carbon cycle projections using observation-based water and carbon flux estimates, *Global Change Biology*, *22*(6), 2198–2215, doi:10.1111/gcb.13217.

Nemani, R. R., C. D. Keeling, H. Hashimoto, W. M. Jolly, S. C. Piper, C. J. Tucker, R. B. Myneni, and S. W. Running (2003), Climate-Driven Increases in Global Terrestrial Net Primary Production from 1982 to 1999, *Science*, *300*(5625), 1560–1563.

Nijsse, F. J. M. M., P. M. Cox, C. Huntingford, and M. S. Williamson (2019), Decadal global temperature variability increases strongly with climate sensitivity, *Nature Climate Change*, *9*(8), 598–601, doi:10.1038/s41558-019-0527-4.

Parazoo, N. C., K. Bowman, J. B. Fisher, C. Frankenberg, D. B. A. Jones, A. Cescatti, O. Pérez-Priego, G. Wohlfahrt, and L. Montagnani (2014), Terrestrial gross primary production inferred from satellite fluorescence and vegetation models, *Global Change Biology*, *20*(10), 3103–3121, doi:10.1111/gcb.12652.

Pelletier, J. D. (1997), Analysis and Modeling of the Natural Variability of Climate, *Journal of Climate*, *10*, 1331–1342.

Phillips, M. C. (2019), Multiple Scales of Surface-Atmosphere Coupling in an Earth System Model, Doctoral, Colorado State University, Fort Collins CO.

Piao, S., S. Sitch, P. Ciais, P. Friedlingstein, P. Peylin, X. Wang, A. Ahlström, A. Anav, J. G. Canadell, N. Cong, C. Huntingford, M. Jung, S. Levis, P. E. Levy, J. Li, X. Lin, M. R. Lomas, M. Lu, Y. Luo, Y. Ma, R. B. Myneni, B. Poulter, Z. Sun, T. Wang, N. Viovy, S. Zaehle, and N. Zeng (2013), Evaluation of terrestrial carbon cycle models for their response to climate variability and to CO_2 trends, *Global Change Biology*, *19*(7), 2117–2132, doi: 10.1111/gcb.12187.

Poulter, B., U. Heyder, and W. Cramer (2009), Modeling the Sensitivity of the Seasonal Cycle of GPP to Dynamic LAI and Soil Depths in Tropical Rainforests, *Ecosystems*, *12*(4), 517–533, doi:10.1007/s10021-009-9238-4.

Poulter, B., L. Aragão, U. Heyder, M. Gumpenberger, J. Heinke, F. Langerwisch, A. Rammig, K. Thonicke, and W. Cramer (2010a), Net biome production of the Amazon Basin in the 21st century, *Global Change Biology*, *16*(7), 2062–2075, doi:10.1111/j.1365-2486.2009.02064.x.

Poulter, B., F. Hattermann, E. Hawkins, S. Zaehle, S. Sitch, N. Restrepo-Coupe, U. Heyder, and W. Cramer (2010b), Robust dynamics of Amazon dieback to climate change with perturbed ecosystem model parameters, *Global Change Biology*, *16*(9), 2476–2495, doi: 10.1111/j.1365-2486.2009.02157.x.

Restrepo-Coupe, N., L. R. Hutyra, A. C. da Araujo, L. S. Borma, B. Christoffersen, O. M. Cabral, P. B. de Camargo, F. L. Cardoso, A. C. L. da Costa, D. R. Fitzjarrald, M. L. Goulden, B. Kruijt, J. M. Maia, Y. S. Malhi, A. O. Manzi, S. D. Miller, A. D. Nobre, C. von Randow, L. D. A. Sá, R. K. Sakai, J. Tota, S. C. Wofsy, F. B. Zanchi, and S. R. Saleska (2013), What drives the seasonality of photosynthesis across the Amazon basin? A cross-site analysis of eddy flux tower measurements from the Brasil flux network, *Agricultural and Forest Meteorology*, *182-183*, 128–144, doi:10.1016/j.agrformet.2013.04.031.

Restrepo-Coupe, N., N. M. Levine, B. O. Christoffersen, L. P. Albert, J. Wu, M. H. Costa, D. Galbraith, H. Imbuzeiro, G. Martins, A. C. da Araujo, Y. S. Malhi, X. Zeng, P. Moorcroft, and S. R. Saleska (2016), Do dynamic global vegetation models capture the seasonality of carbon fluxes in the Amazon basin? A data-model intercomparison, *Global Change Biology*, *23*(1), 191–208, doi:10.1111/gcb.13442.

Rödenbeck, C., S. Houweling, M. Gloor, and M. Heimann (2003), CO_2 flux history 1982–2001 inferred from atmospheric data using a global inversion of atmospheric transport, *Atmospheric Chemistry and Physics*, *3*(6), 1919–1964, doi:10.5194/acp-3-1919-2003.

Rödig, E., M. Cuntz, A. Rammig, R. Fischer, F. Taubert, and A. Huth (2018), The importance of forest structure for carbon fluxes of the Amazon rainforest, *Environmental Research Letters*, *13*(5), 054,013, doi:10.1088/1748-9326/aabc61.

Sakaguchi, K., X. Zeng, B. J. Christoffersen, N. Restrepo-Coupe, S. R. Saleska, and P. M. Brando (2011), Natural and drought scenarios in an east central Amazon forest: Fidelity of the Community Land Model 3.5 with three biogeochemical models, *Journal of Geophysical Research*, *116*(G1), doi:10.1029/2010JG001477.

Swann, A. L. S., F. M. Hoffman, C. D. Koven, and J. T. Randerson (2016), Plant responses to increasing CO_2 reduce estimates of climate impacts on drought severity, *Proceedings of the National Academy of Sciences*, *113*(36), 10,019–10,024, doi:10.1073/pnas.1604581113.

Taylor, K. E. (2001), Summarizing multiple aspects of model performance in a single diagram, *Journal of Geophysical Research: Atmospheres*, *106*(D7), 7183–7192, doi:10.1029/2000JD900719.

Tramontana, G., M. Jung, C. R. Schwalm, K. Ichii, G. Camps-Valls, B. Ráduly, M. Reichstein, M. A. Arain, A. Cescatti, G. Kiely, L. Merbold, P. Serrano-Ortiz, S. Sickert, S. Wolf, and D. Papale (2016), Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms, *Biogeosciences*, *13*(14), 4291–4313, doi:10.5194/bg-13-4291-2016.

Vicari, A. S., A. Mokhtari, R. A. Morales, L.-A. Jaykus, H. C. Frey, B. D. Slenning, and P. Cowen (2007), Second-Order Modeling of Variability and Uncertainty in Microbial Hazard Characterization, *Journal of Food Protection*, *70*(2), 363–372, doi:10.4315/0362-028X-70.2.363.

von Randow, C., A. O. Manzi, B. Kruijt, P. J. de Oliveira, F. B. Zanchi, R. L. Silva, M. G. Hodnett, J. H. C. Gash, J. A. Elbers, M. J. Waterloo, F. L. Cardoso, and P. Kabat (2004), Comparative measurements and seasonal variations in energy and carbon exchange over forest and pasture in South West Amazonia, *Theoretical and Applied Climatology*, *78*(1-3), doi:10.1007/s00704-004-0041-z.

von Randow, C., M. Zeri, N. Restrepo-Coupe, M. N. Muza, L. G. G. de Gonçalves, M. H. Costa, A. C. Araujo, A. O. Manzi, H. R. da Rocha, S. R. Saleska, M. A. Arain, I. T. Baker, B. P. Cestaro, B. Christoffersen, P. Ciais, J. B. Fisher, D. Galbraith, X. Guan, B. van den Hurk, K. Ichii, H. Imbuzeiro, A. Jain, N. Levine, G. Miguez-Macho, B. Poulter, D. R. Roberti, A. Sahoo, K. Schaefer, M. Shi, H. Tian, H. Verbeeck, and Z.-L. Yang (2013), Inter-annual variability of carbon and water fluxes in Amazonian forest, Cerrado and pasture sites, as simulated by terrestrial biosphere models, *Agricultural and Forest Meteorology*, *182-183*, 145–155, doi:10.1016/j.agrformet.2013.05.015.

Wang, W., P. Ciais, R. R. Nemani, J. G. Canadell, S. Piao, S. Sitch, M. A. White, H. Hashimoto, C. Milesi, and R. B. Myneni (2013), Variations in atmospheric CO2 growth rates coupled with tropical temperature, *Proceedings of the National Academy of Sciences*, *110*(32), 13,061–13,066, doi:10.1073/pnas.1219683110.

Wei, Y., S. Liu, D. Huntzinger, A. Michalak, N. Viovy, W. Post, C. Schwalm, K. Schaefer, A. Jacobson, C. Lu, H. Tian, D. Ricciuto, R. Cook, J. Mao, and X. Shi (2014), NACP MsTMIP: Global and North American Driver Data for Multi-Model Intercomparison, doi: 10.3334/ORNLDAAC/1220.

Wenzel, S., P. M. Cox, V. Eyring, and P. Friedlingstein (2014), Emergent constraints on climate-carbon cycle feedbacks in the CMIP5 Earth system models, *Journal of Geophysical Research: Biogeosciences*, *119*(5), 794–807, doi:10.1002/2013JG002591.

Westerling, A. L., M. G. Turner, E. A. H. Smithwick, W. H. Romme, and M. G. Ryan (2011), Continued warming could transform Greater Yellowstone fire regimes by mid-21st century, *Proceedings of the National Academy of Sciences*, *108*(32), 13,165–13,170, doi: 10.1073/pnas.1110199108.

Wielicki, B. A., T. Wong, R. P. Allan, A. Slingo, J. T. Kiehl, B. J. Soden, C. T. Gordon, A. J. Miller, S.-K. Yang, D. A. Randall, F. Robertson, J. Susskind, and H. Jacobowitz (2002), Evidence for Large Decadal Variability in the Tropical Mean Radiative Energy Budget, *Science*, *295*(5556), 841–844.

Wu, J., K. Guan, M. Hayek, N. Restrepo-Coupe, K. T. Wiedemann, X. Xu, R. Wehr, B. O. Christoffersen, G. Miao, R. da Silva, A. C. de Araujo, R. C. Oliviera, P. B. Camargo, R. K. Monson, A. R. Huete, and S. R. Saleska (2017), Partitioning controls on Amazon forest photosynthesis between environmental and biotic factors at hourly to interannual timescales, *Global Change Biology*, *23*(3), 1240–1257, doi:10.1111/gcb.13509.

Xiao, X., Q. Zhang, S. Saleska, L. Hutyra, P. De Camargo, S. Wofsy, S. Frolking, S. Boles, M. Keller, and B. Moore (2005), Satellite-based modeling of gross primary production in a seasonally moist tropical evergreen forest, *Remote Sensing of Environment*, *94*(1), 105–122, doi:10.1016/j.rse.2004.08.015.

Zemp, D. C., C.-F. Schleussner, H. M. J. Barbosa, M. Hirota, V. Montade, G. Sampaio, A. Staal, L. Wang-Erlandsson, and A. Rammig (2017), Self-amplified Amazon forest loss due to vegetation-atmosphere feedbacks, *Nature Communications*, *8*, 14,681, doi: 10.1038/ncomms14681.

Zhang, Y., X. Xiao, X. Wu, S. Zhou, G. Zhang, Y. Qin, and J. Dong (2017), A global moderate resolution dataset of gross primary production of vegetation for 2000–2016, *Scientific Data*, *4*, 170,165, doi:10.1038/sdata.2017.165.

Zhu, J., M. Zhang, Y. Zhang, X. Zeng, and X. Xiao (2018), Response of Tropical Terrestrial Gross Primary Production to the Super El Niño Event in 2015, *Journal of Geophysical Research: Biogeosciences*, *123*(10), 3193–3203, doi:10.1029/2018JG004571.

Zhu, Z., S. Piao, R. B. Myneni, M. Huang, Z. Zeng, J. G. Canadell, P. Ciais, S. Sitch, P. Friedlingstein, A. Arneth, C. Cao, L. Cheng, E. Kato, C. Koven, Y. Li, X. Lian, Y. Liu, R. Liu, J. Mao, Y. Pan, S. Peng, J. Peñuelas, B. Poulter, T. A. M. Pugh, B. D. Stocker, N. Viovy, X. Wang, Y. Wang, Z. Xiao, H. Yang, S. Zaehle, and N. Zeng (2016), Greening of the Earth and its drivers, *Nature Climate Change*, *6*(8), 791–795, doi:10.1038/nclimate3004.

Supporting Information for "Accurate Simulation of Both Sensitivity and Variability for Amazonian Photosynthesis: Too Much to Ask"

Sarah M. Gallup[1], Ian T. Baker[2], John L. Gallup[3], Natalia Restrepo-Coupe[4], Katherine D. Haynes[2], Nicholas M. Geyer[2] and A. Scott Denning[1,2]

[1]Graduate Degree Program in Ecology, Colorado State University, Fort Collins, Colorado, USA
[2]Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado, USA
[3]Department of Economics, Portland State University, Portland, Oregon, USA
[4]

Contents

Text S1: Difference between the variances of outcomes and of predictions

Let $y_i = h(\mathbf{x}_i, \beta) + \varepsilon_i$ be the equation of an outcome y_i to be predicted, where $\mathbf{x}_i$ is a $K \times 1$ vector of explanatory variables, β is a $K \times 1$ vector of parameters, ε_i is a random error

Corresponding author: Sarah Gallup, `sgallup@colostate.edu`

term with mean 0 and variance σ^2 and $i = 1, \ldots, n$ is an index of observations. Let $h(\cdot, \cdot)$ be a known function, which if it were linear would be $h(\mathbf{x}_i, \beta) = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{Ki}$.

Consider the prediction of a specific outcome, y_i corresponding to the explanatory variables $\mathbf{x}_i$. If we have a consistent estimator $\hat{\beta}$ of the parameters β, then the predicted value of y_i is $\hat{y}_i = h(\mathbf{x}_i, \hat{\beta})$. Since $\hat{\beta}$ is consistent, as the sample size n becomes large its limit in probability is precisely β, so in the limit $h(\mathbf{x}_i, \hat{\beta})$ becomes indistinguishable from $h(\mathbf{x}_i, \beta)$. However, even in the probability limit, $\hat{y}_i \xrightarrow{p} E[y_i|\mathbf{x}] \neq y_i$ because $y_i = h(\mathbf{x}_i, \beta) + \varepsilon_i$ also has the random error term ε_i.

The variance of y_i has two parts, $V[y_i] = V[h(\mathbf{x}_i, \beta)] + V[\varepsilon_i]$, assuming $\mathbf{x}_i$ is uncorrelated with ε_i, as typically is necessary for $\hat{\beta}$ to be consistent. In large samples, $V[\hat{y}_i] \xrightarrow{p} V[h(\mathbf{x}_i, \beta)]$, but the variance of y_i is larger:

$$V[y_i] = V[h(\mathbf{x}_i, \beta)] + \sigma^2$$

due to the presence of the random error term ε_i in y_i.

Text S2: Methods Details

2.1 Process Models

The Multi-scale synthesis and Terrestrial Model Intercomparison Project [MsTMIP; *Huntzinger et al.*, 2014; *Wei et al.*, 2014] isolates land model GPP structural responsiveness from output differences due to varying inputs. Variants of four of the models participate in IPCC's forecasts. Comparing runs based on standardized drivers is important for the Amazon because its rainfall differs strongly across ESMs [*Ahlström et al.*, 2017; *Huntingford et al.*, 2013; *Jupp et al.*, 2010; *Li et al.*, 2006; *Poulter et al.*, 2010]. MsTMIP did not prescribe how modelers should distribute monthly meteorology into the shorter time steps at which many models run. Forcing all models with the same weather does omit feedbacks between weather and GPP [*Gloor et al.*, 2013; *Harper et al.*, 2014] at times scales longer than a single month.

All participating MsTMIP models provided outputs of GPP, respiration, and closely-derived net ecosystem productivity. While MsTMIP invited additional variables, their absence for at least a varying third of models hampers comparative analysis. Runs represent each model's configuration in about 2014. A subsequent update of CLM, for example, specifically addressed previously excessive modeled tropical GPP [*Oleson and Lawrence*, 2013, p. 9].

Weather reanalyses are less certain for the tropics than for midlatitudes [*Clark and Clark*, 2011; *Li et al.*, 2006; *Malhi and Wright*, 2004]. The only striking outliers in the MsTMIP meteorology were retained. From 4000 to 8,597 mm of rain in January, 2000 is assigned to 56 half-degree grid cells. Otherwise the highest monthly rainfall anywhere in the study area in any month is 2,431 mm. Following convention for the wet tropics [*Saleska et al.*, 2003], dry season is defined as months when long-term mean precipitation is below 100 mm, or less than the approximate maximum plants can metabolize in real time [*Aragão et al.*, 2007].

SiB4 meteorology and land cover drivers were developed in conjunction with the new model version and differ slightly from MsTMIP's.

2.2 Statistical Models

MsTMIP models plus SiB4 are referred to as process models, since each simulates the biological determination of GPP. In contrast, data assimilation estimates of global GPP, labeled as statistical models, simulate retrospectively from remotely sensed inputs. Ideally they are sufficiently accurate to benchmark process models [*Jung et al.*, 2011; *Zhang et al.*, 2017]. They are driven not by MsTMIP weather but by closely-related weather reanalyses.

Global satellite inputs temper Fluxcom's and extrapolate cleaned eddy covariance flux estimates. Fluxcom is widely used as reference GPP globally [for example, *Anav et al.*, 2015; *Bonan et al.*, 2011; *Collier et al.*, 2018; *Jung et al.*, 2019; *Malavelle et al.*, 2019; *Mystakidis et al.*, 2016; *Piao et al.*, 2013; *Tramontana et al.*, 2016; *Xu et al.*, 2015]. We use the half-degree resolution product from the multivariate adaptive regression splines algorithm.

Wecann is similar in both concept and results to Fluxcom. With additional input from GOME-2's solar-induced fluorescence, Wecann fits tower sensible heat, latent heat, and GPP slightly better than does Fluxcom [*Alemohammad et al.*, 2017]. Wecann's one degree resolution is coarser than MsTMIP's. We attribute each value to four half-degree cells, and note below adjustments made to avoid artificially narrowed confidence intervals.

The third statistical model, vegetation photosynthesis model (VPM) is a light-use efficiency model. VPM applies deliberately few and non-fitted numeric constants to temperature reanalysis and multiple MODIS and SPOT satellite products [*Xiao et al.*, 2005; *Zhang et al.*, 2017]. For a test year in North America, VPM provided the median estimate compared to six other global GPP models [*Zhang et al.*, 2016]. Being even more heavily dependent on satellite data than is Fluxcom, VPM is likely to be less accurate in the cloudy tropics than elsewhere, and less accurate for the tropical wet season than for the dry season.

2.3 Study Boundaries

Selecting EBF tiles within cells is not workable because MsTMIP models' GPP estimates are not available for individual PFTs. To assess GPP that is representative for each cell's vegetation (Fig. 1) requires that cell values should be an average only across land area. The MsTMIP models [*Chapin et al.*, 2006], SiB4, and WeCann [*personal communication, Alemohammad, 2020*] give GPP as a mean value across both land and water. VPM [*Zhang et al.*, 2017] and Fluxcom [*Jung et al.*, 2019] provide GPP as means for a cell's land area only. All GPP datasets except VPM and Fluxcom were adjusted by the cell's water fraction in the MsTMIP PFT map. In summary metrics, months are treated as if they are equally long.

2.4 Eddy Covariance Fluxes

While there is some utility in simply comparing models, knowing their true accuracy is far more useful. The statistical models are candidate reference data sets, but despite circularity issues addressed below, we wish to assess their accuracy as well. For the Amazon, GPP from individual eddy covariance towers is the only remaining option. ECs measure exchange between the land surface and the atmosphere of CO_2 and other gasses that vegetation affects. From measurements related to net ecosystem exchange, the large and opposing contributions of GPP and ecosystem respiration are modeled.

EC GPP was further limited to the study period starting in 2000, cutting off a few months each at sites CAX, K34, and RJA. The merged EC dataset offers eleven GPP algorithm options. Consistent with *Baker et al.* [2013], we use "GEP_model." The two ECs in the Tapajos National Forest are about a dozen kilometers apart in stands with different logging histories. Due to their related synoptic weather and seasonality, for this study the K67 and K83 sites are best considered as pseudo-replicates.

Six sites in the South American rainforest cannot fully represent the region's range in either plant productivity or other criteria. For most study models the six cells that contain a flux tower site encompass less than a third of the model's central 98% of range in mean annual GPP across the Amazon. However, EC site cells typically fall on both sides of a model's median GPP. EC cells also tend to have high GPP (Text S3), which is useful because similarly annual high productivity is uncommon at other flux towers globally whose tendencies might otherwise help constrain modeling of the tropics.

Text S3: Representativeness of EC Sites

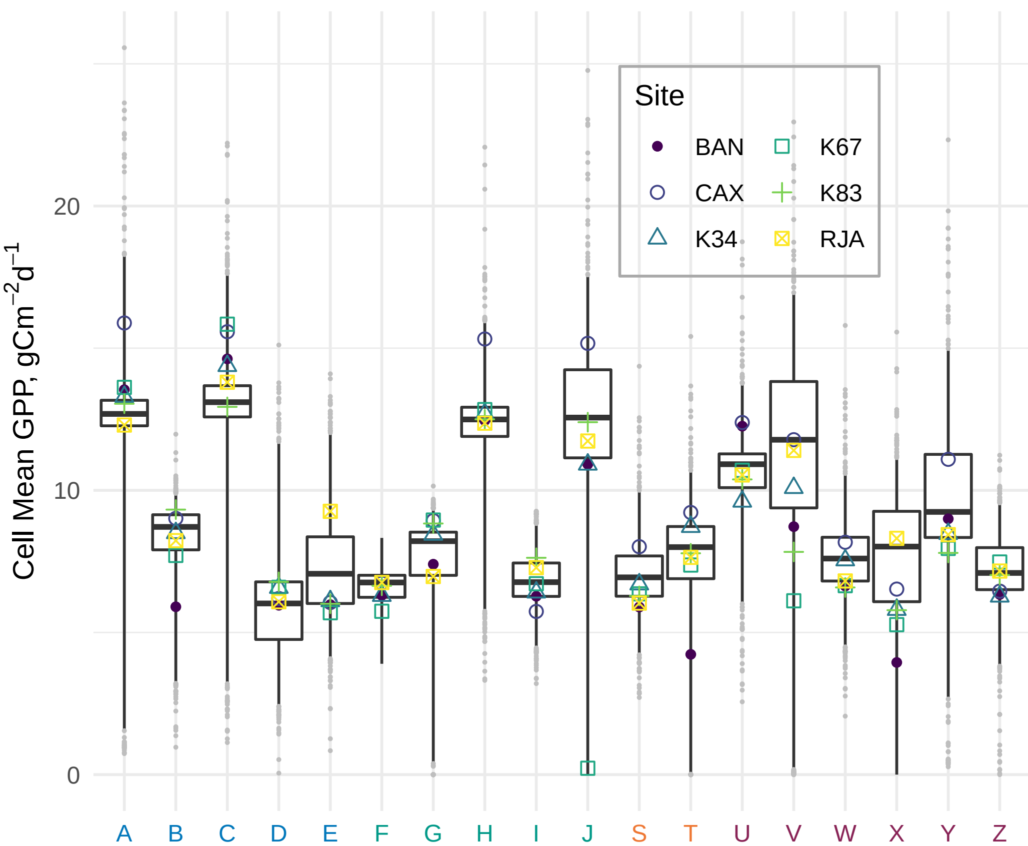


Figure S1: Distribution of cell-level GPP averaged across all study months and the entire study area. A horizontal bar marks a model's median GPP. Boxes encompass the central 50% of a model's cells as ranked by mean GPP, outside of which all grey points indicate all but the most extreme 2% of cells. Colored symbols mark EC cells. Especially for the some of the more responsive models, toward the left, the EC sites are in cells with above-average productivity.

From the perspective of the models being contrasted, how representative are the EC sites of the entire Amazon? One basis of comparison is how much of the range in mean GPP for the 11-year study period the six cells that contain a flux site capture compared to the range across all EBF cells. EC sites are compared to the basin's range as defined by its 1^{st} and 99^{th} percentile values. The outlier is model J, for which EC sites span 85% of the watershed's range in GPP due largely to its near-zero GPP for K67. Among the remaining models, mean GPP for cells with ECs cover from 9% and 51% of the range of the central 98% of Amazon rainforest cells. The median among the models in span that ECs represent is 29%. For most of the study models, the six ECs represent less than a third of the range in mean grid-cell GPP.

On the other hand, the EC sites represent a portion of the rainforest GPP range that is especially valuable to match. There are disproportionately few flux towers in the exceptionally productive tropics than there are in some of the world's less productive biomes. If plant responsiveness across the global range of environmental driver values is mostly continuous although non-linear, then ECs elsewhere may help constrain modeling of the low end of the rainforest productivity spectrum. By this criterion, the most useful EBF flux towers are at the most productive sites.

The six EC cells do have a higher mean GPP than is typical of the Amazon basin. The median cell-level annual GPP across all the models that the most productive of the sites represents as a percentile of each model's central 98% of study cells is 85%, with a range of 50% to 99% . In most but not all models, the most productive flux site cell is CAX. For the least rather than most productive of the EC cells, usually the cell that contains BAN, model percentiles range from 4% and 49% with a median of 17%. While the cells with eddy covariance data cover only a limited portion of the Amazon basin's range in mean annual GPP, they are in relatively productive sites for which closely-related alternative data are least plentiful.

Text S4: Non-Linearities in GPP Responses to Rain and Light

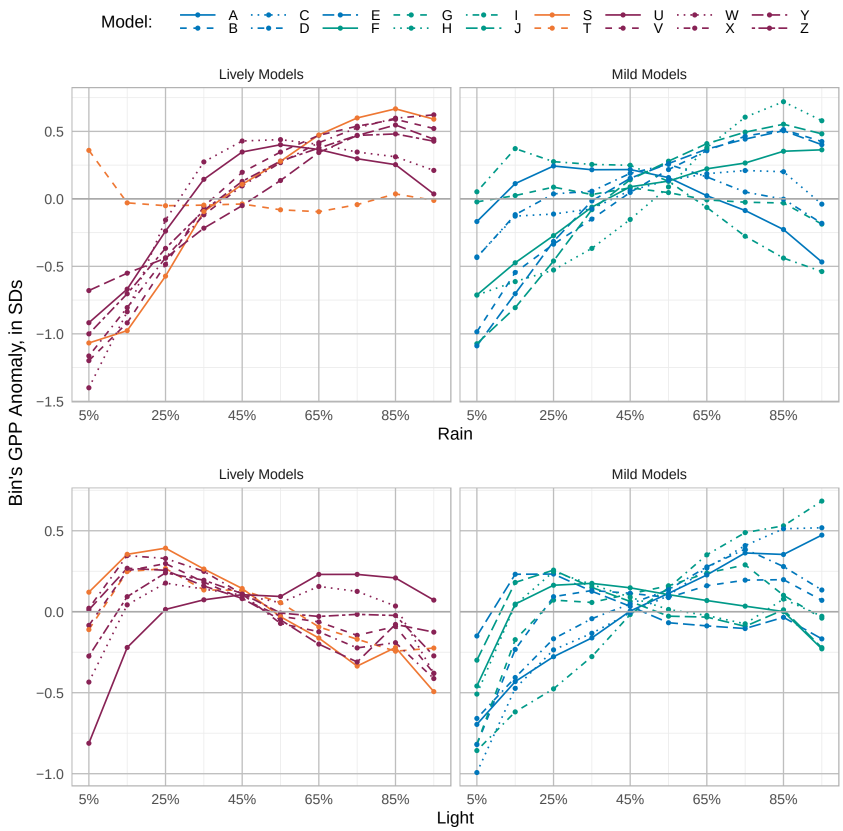


Figure S2: Non-linearity in modeled GPP's responses to rain and light, parallel to the main paper's Fig. 6 for temperature.

In Fig. S2, GPP is shown as z-score relative to a particular model and cell's mean across the study period. Shared MsTMIP driver data is binned by deciles. Mild models are in the left panels, and lively models on the right. Nearly all models simulate GPP as falling in the very wettest months although in only a few mild models is it below average. In the driest 10% of months GPP is below average in all but two models.

GPP increases with rain at the driest deciles and falls in the very wettest months in all but one model. One difference from temperature is that responses to rain for individual mild models are more nearly linear, and models diverge from each other for the driest months in approximately reverse rank as they do at the wettest. Model T, a statistical model, is an exception, with no discernable trend in response to varying rain. The inflection point at which GPP switches from increasing to falling with more rain ranges among models from the second to ninth deciles of current month's rain. The drop begins at drier levels in most of the mild models. For some of the lively models, more rain corresponds to higher GPP until about the top two deciles. The disparities represent disagreement about what moisture is optimal for EBF, although do not reveal how modeled soil moisture mediates these responses within many of the models.

Radiation's responsiveness too is almost uniformly curved, consistent with a classic light response curve [*Baker et al.*, 2019]. In the darkest months nearly all models simulate low GPP. Consistent with differences that *Rogers et al.* [2017] noted, the flex points of models' light saturation divide into two groups, one slightly below 200 Wm^{-2} and the other near 220 Wm^{-2}. Some of the lively models show strong drops in plant productivity in especially bright months.

Text S5: Months of Modeled GPP's Seasonal Peaks at EC Sites

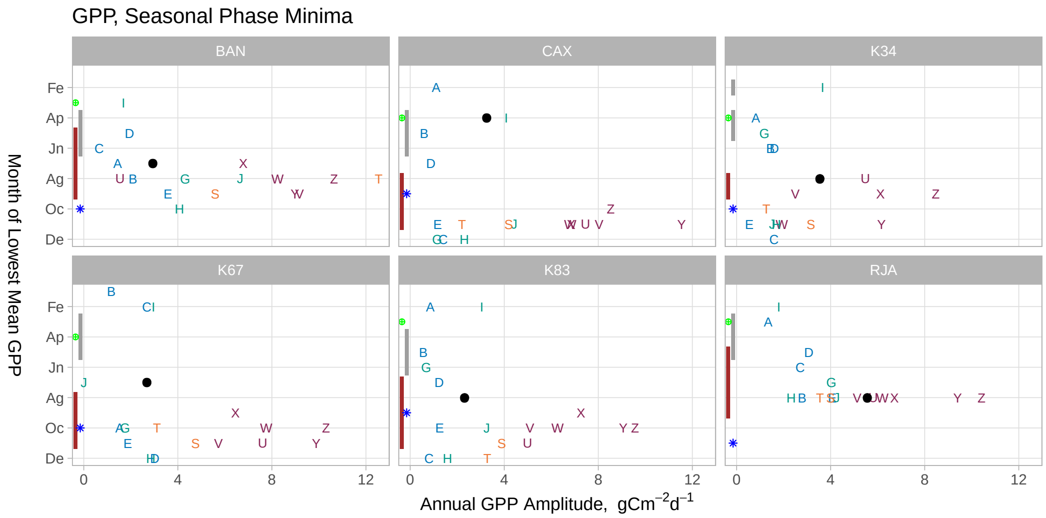


Figure S3: For each site, the month of lowest mean monthly GPP for each mode. The x-axis and symbol colors rank models by extent of seasonality. Grey bars on each site's y-axis indicate the three months with the least light, while a blue star marks the brightest month. Brown bars show the dry season. The wettest month is indicated with a green dot. Mild models are more likely to simulate minimum GPP during a dark month, while lively models' lowest GPP typically occurs during the dry season.

Fig. S3 shows that models tend to fall into the same groups for seasonal phase as they do for seasonal amplitude, both reflecting their relative responsiveness to drivers. Mild and lively models have nearly identical mean timing differences between EC and modeled GPP, of 2.9 and 2.3 months respectively. The model groups differ in the direction of differences. Models with little seasonality tend to simulate the year's lowest GPP before or early in the dry season. For every lively model except Model I, GPP is lowest at every site either during the dry season or in the first month afterward (Fig. S3). Most of the lively models' minima occur late in the dry season when modeled soil moisture presumably is lowest. Again excepting Model I, no lively model simulates minimum GPP during the three darkest months for any site.

Patterns are similar for month of highest rather than of lowest mean GPP (Fig. S4). EC GPP at all sites but CAX peak 2-5 months after the last dry season month. Half of the models, a mix of mild and lively, match CAX' timing to the extent of peaking during its four-

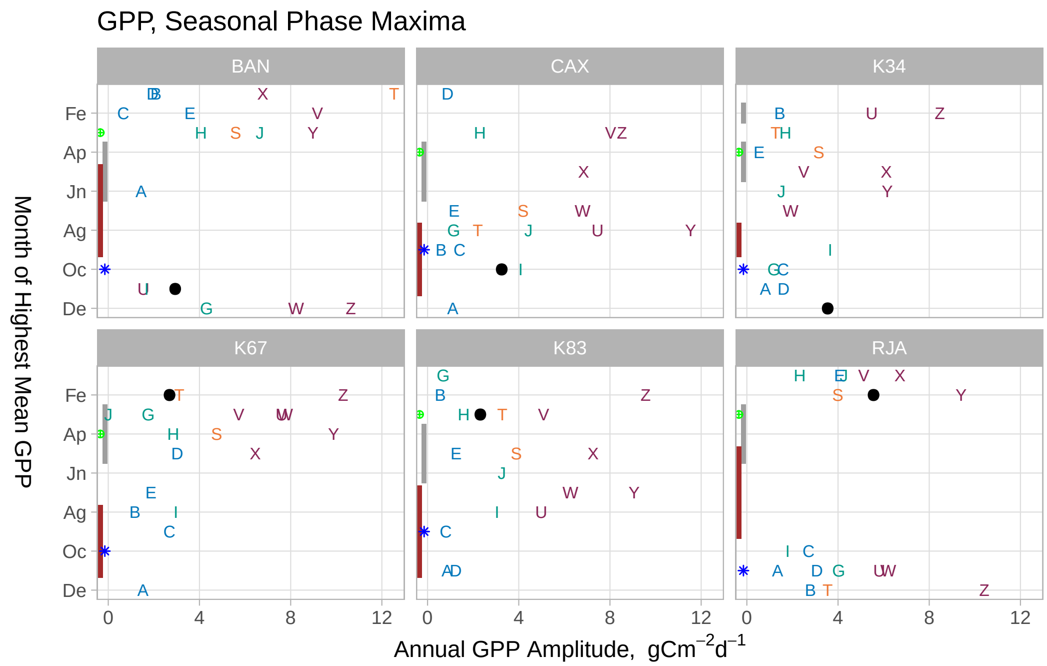


Figure S4: Parallel to Fig. S3 of month of lowest GPP at each site, but showing month of highest GPP. While the peak and trough of the seasonal phase is not consistently offset by six months, overall patterns of model responses are similar between lowest and highest GPP timing.

month dry season. Peak month is most accurately modeled for RJA, where neither the EC tower nor any model reaches maximum GPP during the dry season.

The wet tropics have a modest annual cycle in both leaf area index and the photosynthetic capacity of average leaves [*Albert et al.*, 2018; *Borchert et al.*, 2002; *Dahlin et al.*, 2017; *Doughty and Goulden*, 2008; *Goulden et al.*, 2004; *Samanta et al.*, 2012; *Wilson et al.*, 2001; *Wu et al.*, 2016] but see [*Morton et al.*, 2016]. Seasonal rainforest leaf phenology is thus far absent from most process models of GPP *Albert et al.* [2019]. While there is at least speculative logic for the timing of each site's maximum plant stress, the dominant mechanism appears to vary across sites. At Tapajos, sites K34 and K83, the lowest EC GPP occurs early in the dry season (Fig. S4) and corresponds to the annual peak of leaf exchange. The timing at K67 and BAN also is reasonably consistent with a leaf demography hypothesis. RJA reaches its lowest EC GPP late in its stark dry season. CAX's minimum is during a dark month in the middle of its mildly wetter season. The timing of rainforest leaf exchange may respond to a continuum of water v. light limitation even if instantaneous GPP does not *Albert* *et al.* [2019]. The mismatches suggest that adding tropical leaf seasonality could improve the accuracy of modeled GPP.

Text S6: Cumulative rainfall's influence on EC GPP

Although site intercepts and current weather in simple regressions explain on average 64% of modeled GPP's variance, on average about a third remains unexplained. For the ECs, the correlation of soil moisture with GPP is -0.31. Lively models' lower GPP than ECs during the dry season suggests that modeled rainforest plants experience more severe water stress than real plants. Process models simulate and track soil moisture, often at multiple depths. Cumulative water deficit may be a key variable omitted from the descriptive regressions. Soil moisture output is not available for enough models, but an indirect indicator of its effects is the strength of connection between modeled GPP and cumulative rain over recent months. The added explanatory power of cumulative rainfall is one way to characterize the strength of a site's hydrologic memory.

Testing soil moisture modeling directly requires reasonable reference GPP across the basin's spectrum of annual precipitation. Local plants logically adapt to the degree of drought they experience episodically [*Corlett*, 2016], and satellite data imply that sensitivity to tropical drought is spatially heterogeneous [*Bonal et al.*, 2016; *Feldpausch et al.*, 2016]. Soil moisture varies markedly also at fine scales, making it difficult to measure [*Broedel et al.*, 2017; *Huang et al.*, 2016] or model [*Parazoo et al.*, 2014] for even an EC footprint. Benchmarking modeled soil moisture across the Amazon is therefore particularly challenging.

It would be helpful if instead accumulated rain were a rough proxy for soil moisture. Rain summed over periods ranging from only the most recent month to the entire last year explain greatly varying portions of individual sites' GPP variability. Each point in Fig. S5, represents a regression of GPP on MsTMIP temperature, light and accumulated rain, optimized for a single site except the summary line for all sites. Dots and solid connecting lines mark regressions whose rain coefficients are statistically significant at $p \leq 0.05$. Dashed lines pass through r^2 of regressions whose rain coefficients fail the significance test. The maximum predictive power of a full year for all sites is of little practical consequence. A full year's cumulative rain is significant at only one site.

For each site, the point for one month of lag in Fig. S5 shows how much of the variation in GPP current weather alone explains. The difference from each site's peak value indicates how much more information rain history can add to current month's weather. The legend lists each site's maximum fit improvement due to cumulative rain. As with current

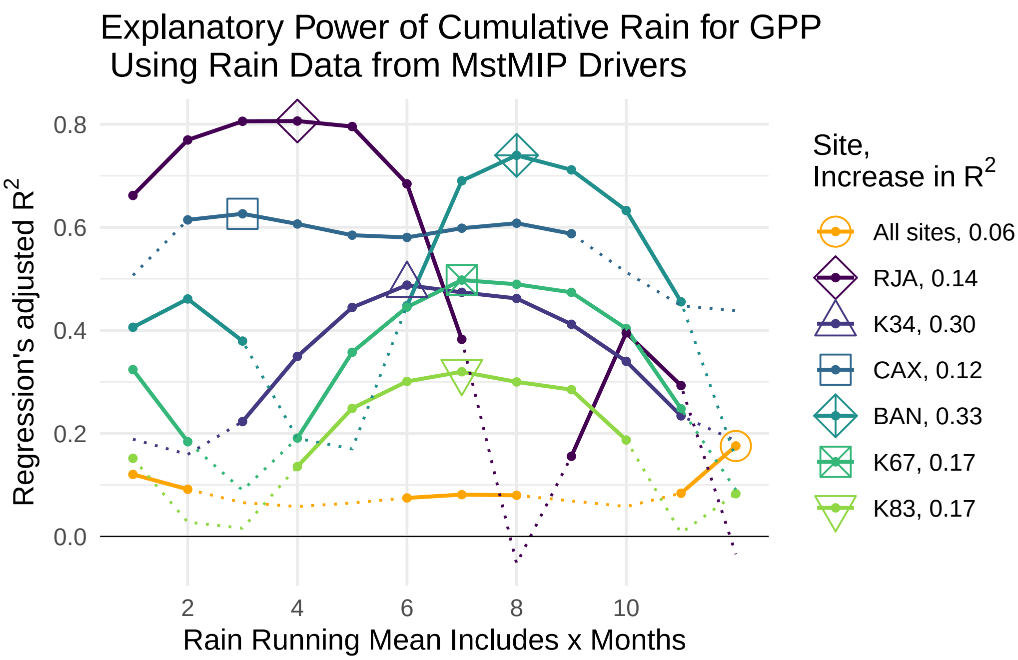


Figure S5: For each site, the rain accumulation duration best predicts EC GPP. The x-axis shows the number of months before and including the current month that are totaled. The y-axis indicates the portion of variability explained by regressing each rain accumulation period plus current month's light and temperature on GPP. Large open symbols indicate each site's accumulation duration with the most explanatory power. The increase in r^2 listed in the legend equals the difference between the optimal formula and one that uses only current month's rain. The month with the most explanatory power varies so much that no single duration explains more of the variability across all sites than can current month's rain alone.

month predictors of GPP, weather measured near each EC has slightly less explanatory power (Fig. S6).

We attempted to predict each site's best rain lag period. The negative coefficient on annual average rain, -0.010 months of optimal lag per mm increase, implies that a longer lag and possibly greater soil moisture retention capacity or deep rootedness exist at relatively dry sites. Correlations are very low for latitude, dry season length, and mean rain during the three driest months. Site mean annual precipitation explained 73% of the variation in best lag length.

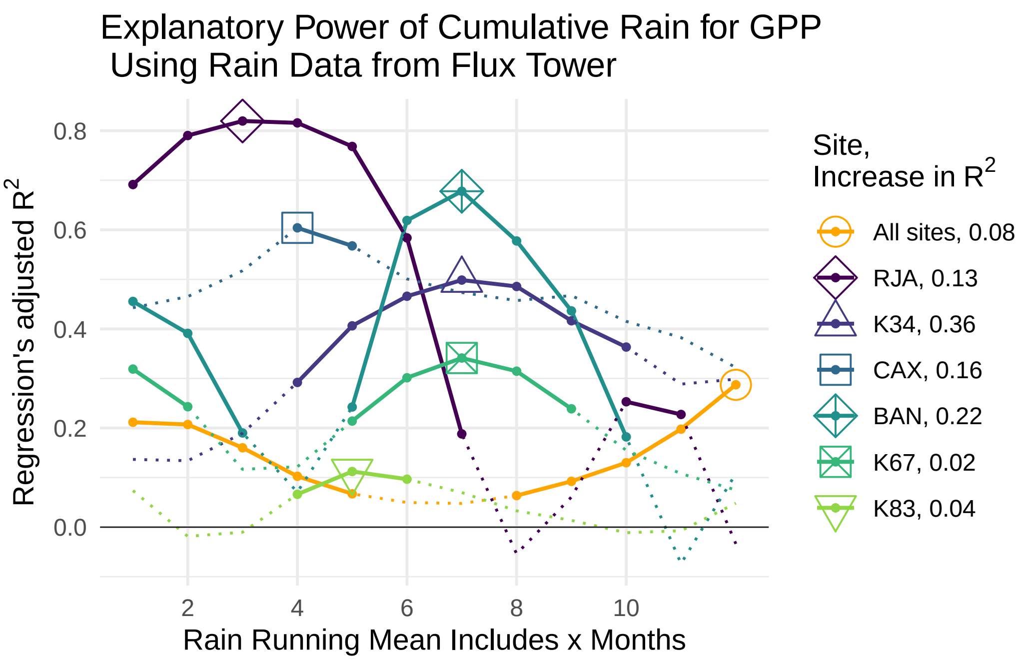


Figure S6: Parallel to Fig. S5 of rain lag duration versus explanatory power, using EC meteorology rather than MsTMIP weather reanalysis rain. The sharpest differences in results for the two meteorology sources are the worse fits of EC rain for K67 and K83. EC rain does slightly better than MsTMIP rain for K34 and CAX.

Highlighting the difficulties in developing modeling equations that are reasonable for all sites are differences in Fig. S5's site-specific light and temperature coefficients. Particularly for light, responsiveness is typically two to eight times stronger at individual ECs than when calculated across all sites simultaneously. Light at one or more rain lags is significant at only three sites, one only at cumulations longer than 9 months. That light nonetheless is significant for the regression across all sites for every lag period option suggests that light may be partly a surrogate for omitted drivers correlated with latitude. Temperature differs more across sites, significant in regressions for 2 to 7 inconsistent groupings of the 12 possible lag periods.

Figure S7: Mean and Variance of GPP for Each Site

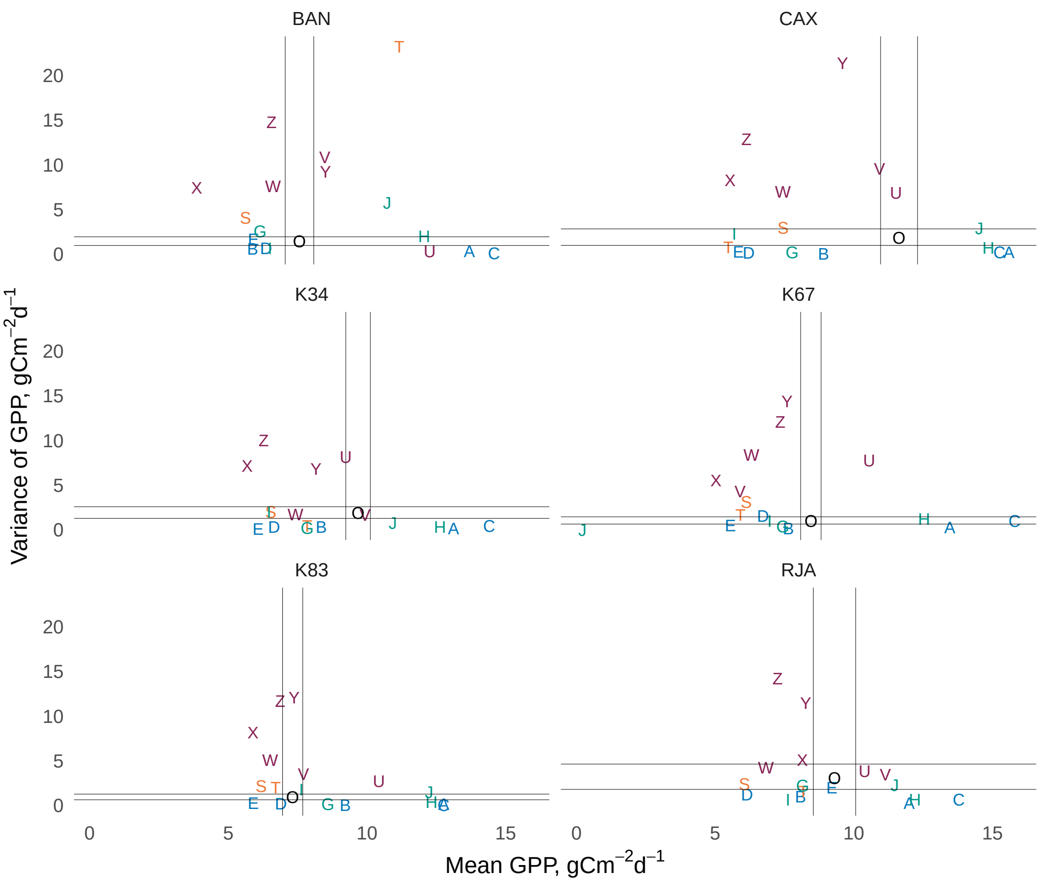


Figure S7: For each eddy covariance site, comparison of EC estimates to the means and variances in GPP estimates from each statistical or process model. Vertical and horizontal lines bracketing Model "O", EC estimates, are 99th percentile confidence bounds. For most models, both mean and variance fall outside the confidence bounds, with some models higher and some lower. For some models, their GPP variability exceeds that of EC estimates even more markedly than does mean GPP.

Comparing GPP's mean and variance for individual EC sites to each model relative to the mean rather than by the absolute variance further contradicts the possibility that high variance is simply due to high absolute GPP. Six models' variance at one or more sites exceeds 100% of the site's mean (Fig. 2). These six models are among the eight whose overall variance is larger than overall EC variance. Site-level variance for the model with the highest overall variance ranges from 160 to 226% of the site's mean modeled GPP. In contrast, the model with the lowest variance as a percent of mean modeled GPP ranges from 1% to 8% at individual sites.

Figure S8: Correlations of EC GPP with Process and Statistical Models

model	BAN	CAX	K34	K67	K83	RJA	All
A	-0.07	0.17	0.45	-0.07	0.18	0.07	0.36
B	0.69	0.03	0.74	-0.42	0.50	0.77	0.39
C	0.21	-0.32	-0.22	0.09	-0.05	0.37	0.25
D	0.64	0.41	0.67	-0.36	0.42	0.12	0.03
E	0.40	-0.60	-0.16	-0.52	-0.03	0.87	0.19
G	0.82	0.12	0.43	-0.08	0.65	0.85	0.37
H	0.02	-0.54	-0.26	-0.48	-0.10	0.89	0.34
I	0.33	0.68	-0.04	-0.45	-0.19	0.11	-0.19
J	0.34	-0.61	-0.27	0.30	-0.18	0.87	0.20
S	0.25	-0.70	-0.19	-0.38	0.01	0.80	0.17
T	0.49	-0.59	0.12	-0.28	0.12	0.85	-0.05
U	0.89	-0.45	0.28	-0.38	-0.01	0.75	0.02
V	0.41	-0.55	-0.14	-0.34	0.01	0.87	0.34
W	0.61	-0.53	0.02	-0.34	0.06	0.76	0.18
X	0.58	-0.74	-0.09	-0.19	0.02	0.78	0.13
Y	0.05	-0.68	-0.59	-0.50	-0.16	0.75	-0.03
Z	0.81	-0.50	0.32	0.19	0.31	0.77	0.21
Mean	0.44	-0.32	0.06	-0.25	0.09	0.66	0.17

Figure S8: Simple correlations between a process or statistical model's GPP at a particular site to its EC estimate. Letter colors correspond to models' seasonal amplitude relative to that of the ECs, with green and blue models milder and orange and red models livelier. The mean months that a EC operated, or number of paired values per site correlation, is 43. The column on the far right, for all sites, is the variance calculated on all pairs of EC GPP with the other model, across all sites and months, not the mean of the six site-level variances. A small squared correlation suggests *prima facie* that there is only random connection between a model and EC estimate.

259 **Figure S9: Seasonal Cycle Amplitudes for Each EC Site**

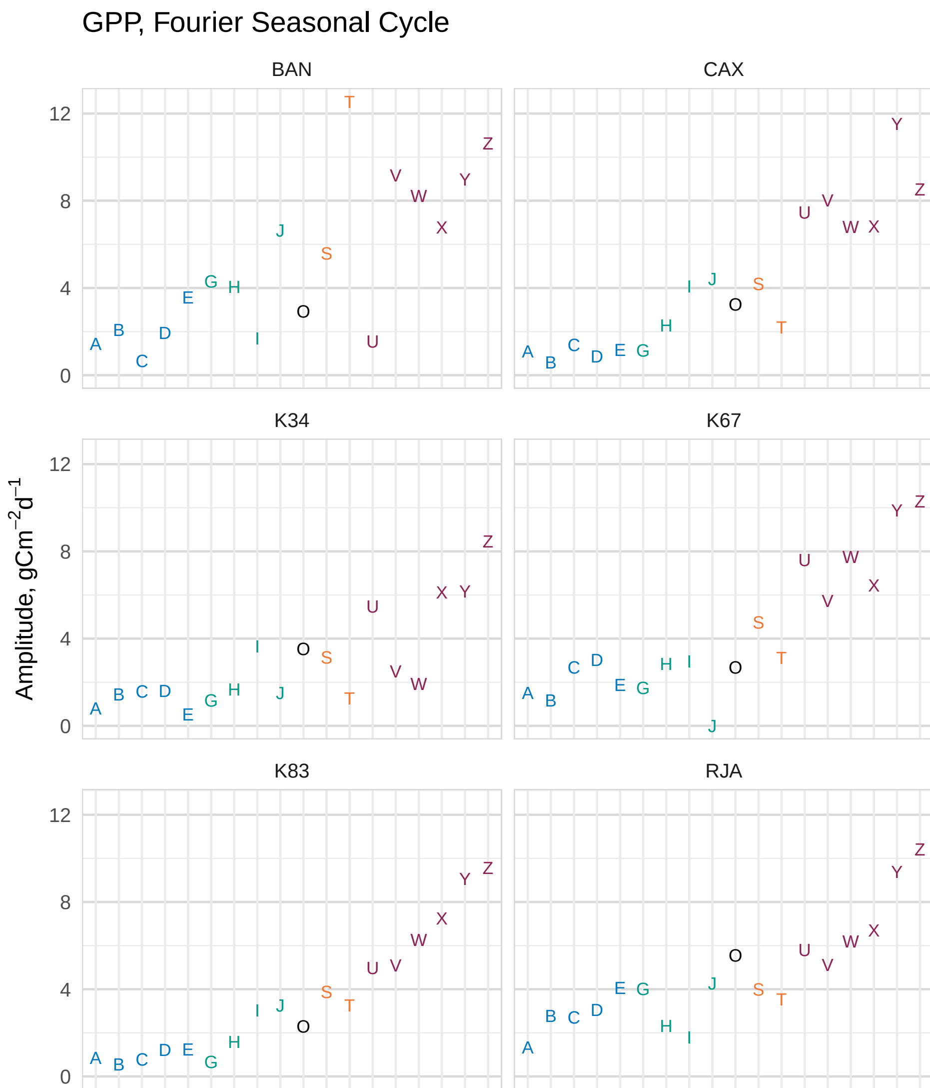


Figure S9: Parallel to the main paper's Fig. 3 of seasonal amplitudes, showing each site separately. Steadily increasing values from left to right indicate moderate similarity in site-level amplitude ranking and the ranking of mean amplitudes.

260 **Figure S10: Yearly Mean GPP by Model**

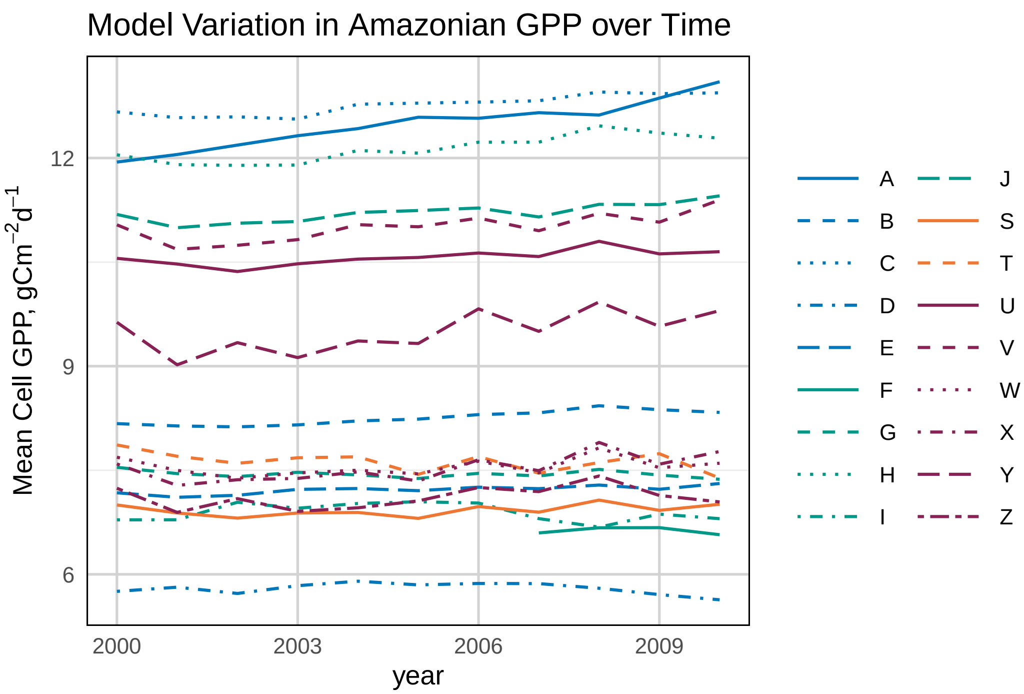


Figure S10: Yearly mean GPP for Amazon rainforest model cells. Interannual variability of individual models is much smaller than the differences among models in a particular year. Even driven by identical climate inputs, different models simulate consequentially different GPP. In every year the mean reconstructed GPP for the Amazon is over twice as high for the highest three models as for the lowest two. The differences mean that ESMs' predictions of tropical GPP several reflect not only substantial differences in meteorological predictions but also in tropical GPP's model structure and parameterization.

Figures S11 and S12: GPP Responses to Environmental Drivers Across the Entire Amazon Basin

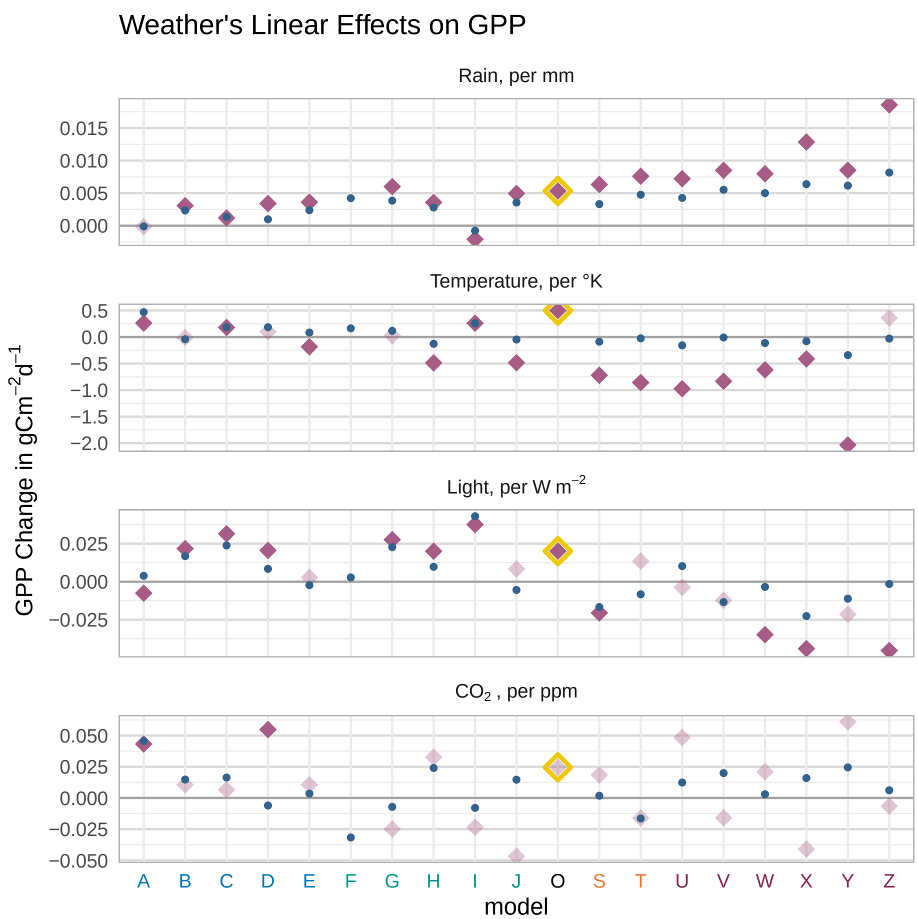


Figure S11: Fig. 5 shows each model's slope of GPP for EC cells only with driver units in z-scores. This figure summarizes tendencies also across the entire basin, displays slopes per unit value of each driver, and includes slopes for CO_2. The large purple diamonds are for the EC cells only while the blue dots are for the entire Amazon. EC estimates are highlighted with a yellow background. Site slopes with probability $\leq$0.05 are semi-transparent. With hundreds of cells, p$\leq$0.05 for all basin-level predictors. Model responsiveness at EC sites generally mimics their basinwide responsiveness.

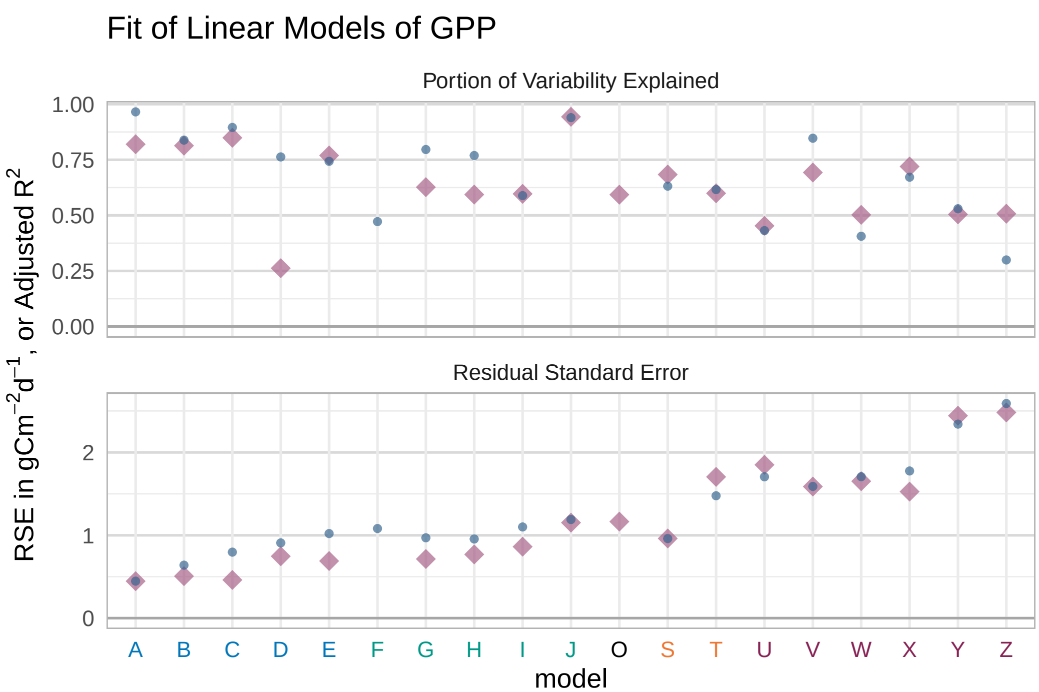


Figure S12: Please see caption for previous figure. Fits for Model F are not directly comparable because they summarize a 1° spatial grid while all other models have ½°, or approximately four times as many cells in the Amazon.

Figure S13: Maps of Driver Responsiveness by Model

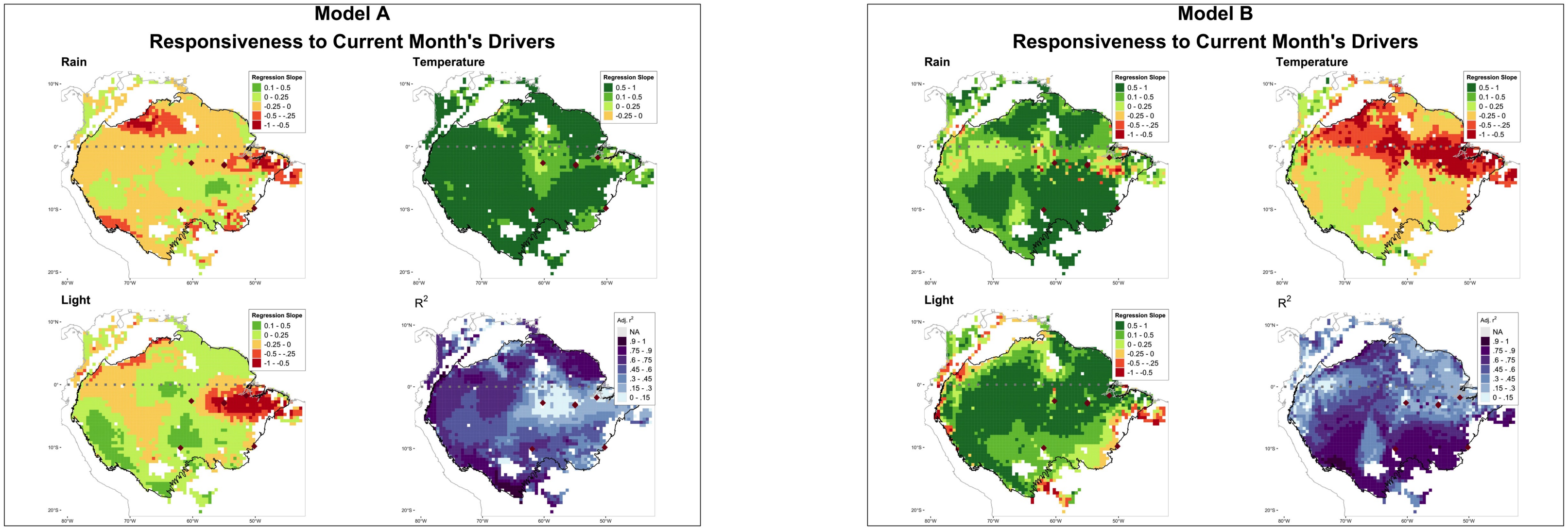
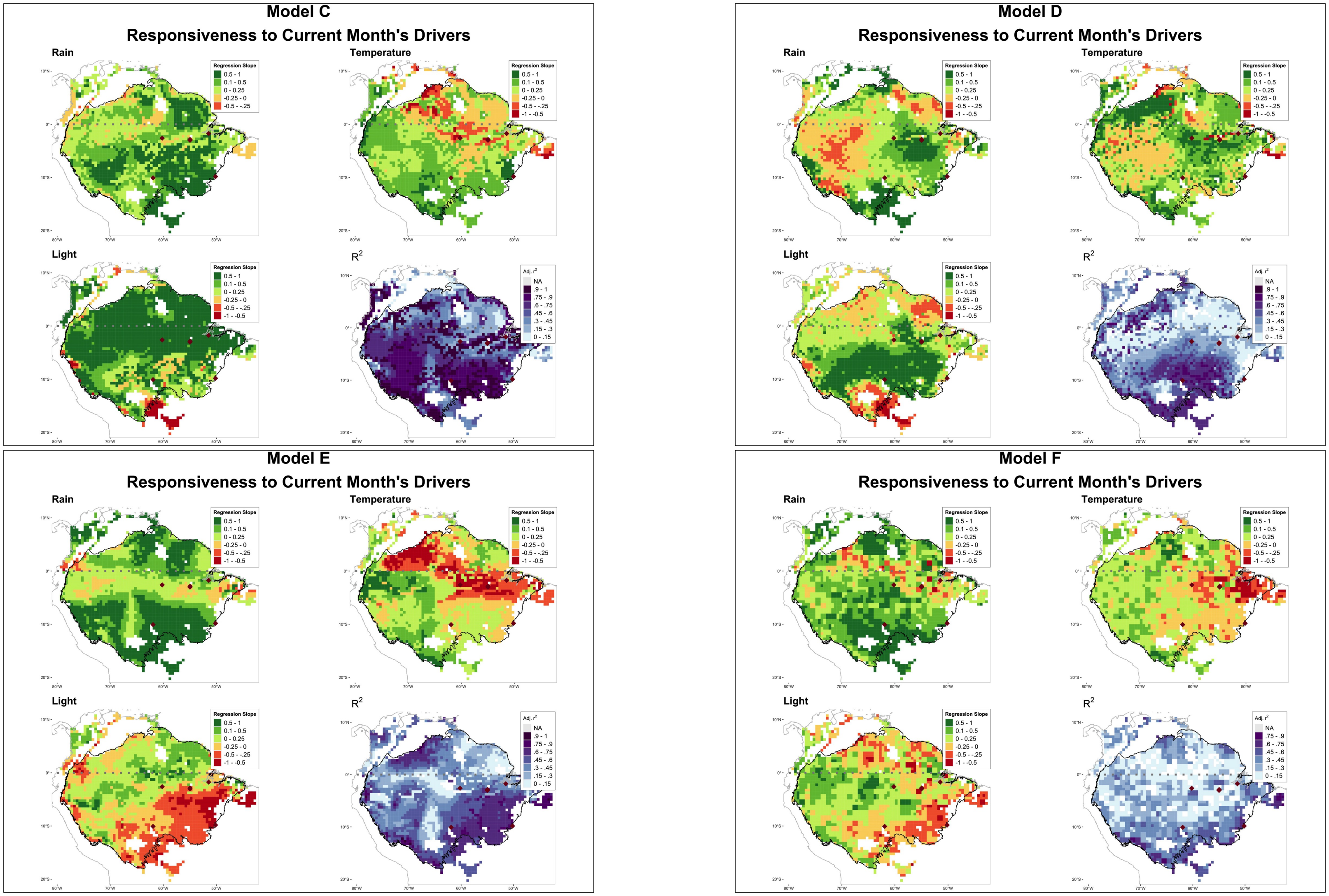
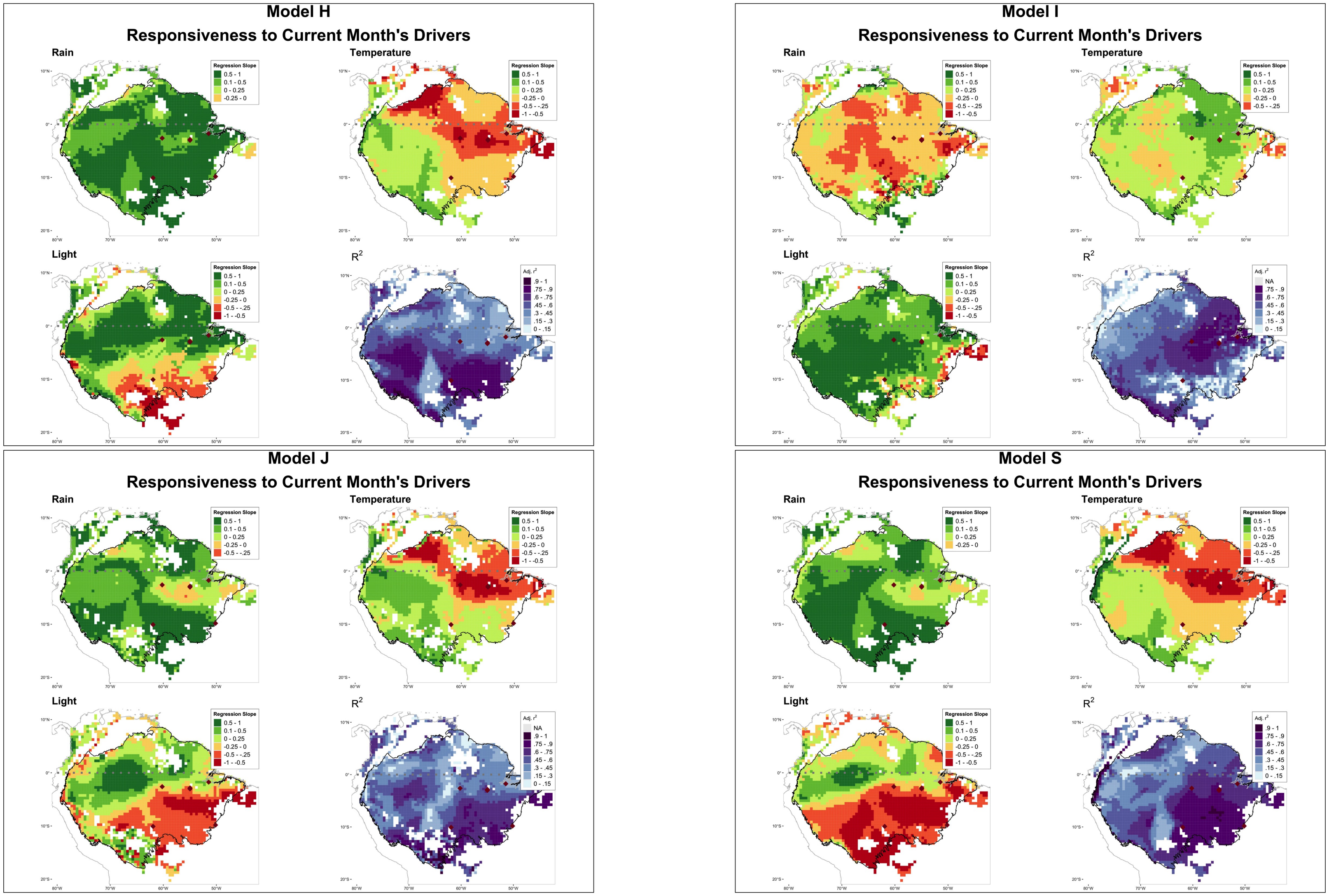


Figure S13: Spatial patterns in driver slopes and regression's portion of GPP variability explained for each model. Each model's four panels are arranged in the same locations as the panels of Fig. 5. For the EC sites in Fig. 5, the regression form allows each site to have a different intercept while driver slopes are forced to be identical at every site. For the model maps, in contrast, every cell's slopes as well as intercept are calculated individually.

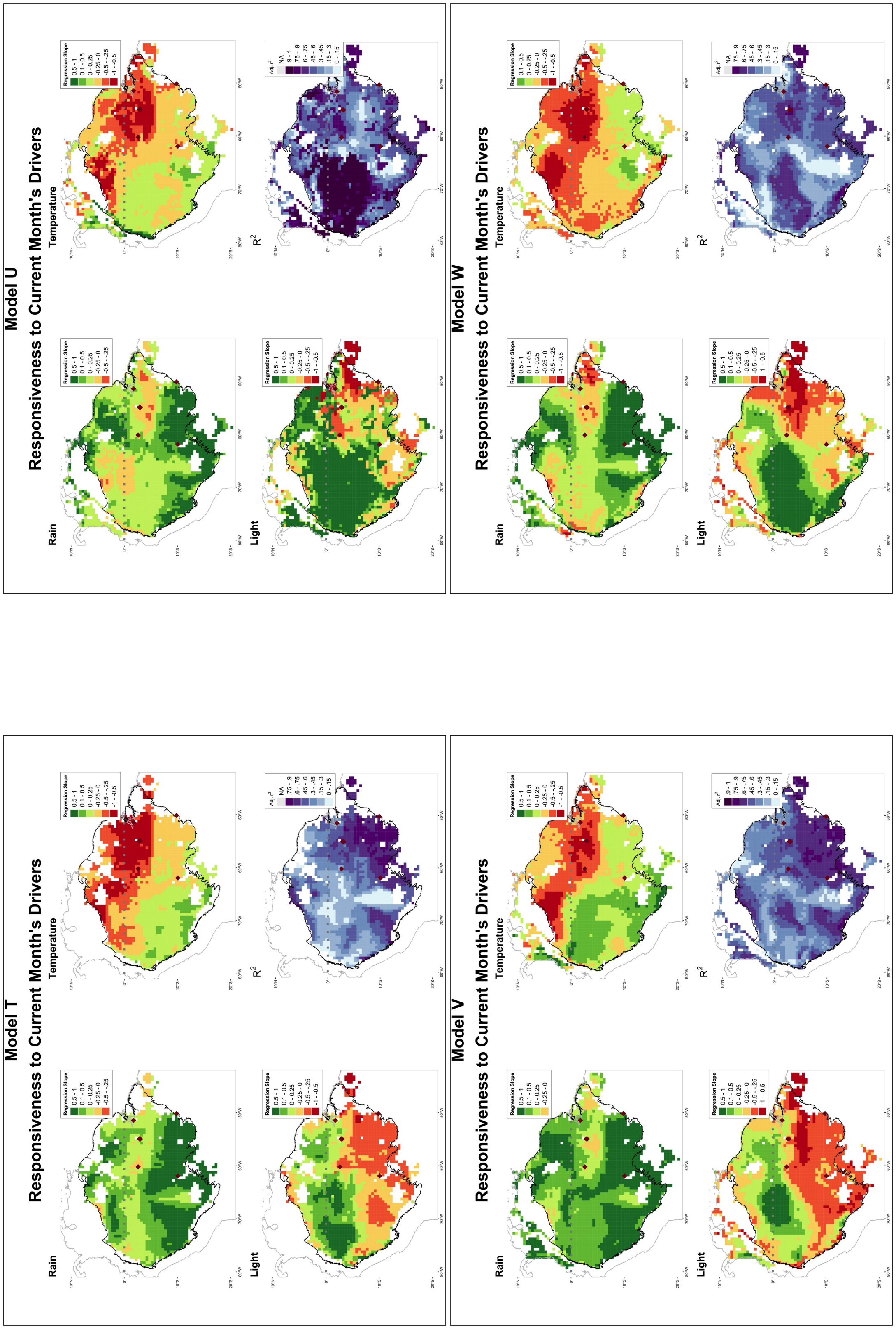
All but Models D and I respond most weakly to rain in an east-west swath south of the equator. There, twice-yearly overcrossing of the Intertropical Convergence Zone may reduce the distinctiveness of monthly mean rain. For the especially wet northwest, models respond similarly to rain, differing mostly in the extent but not general location of places where GPP increases most in the rainiest months. Temperature's spatial pattern is similar, but the central band has stronger rather than weaker response, and extends farther to the north on the west end of the basin. GPP is least responsive near the basin's periphery for models with weak seasonal cycles. For strongly seasonal models, instead there is a sharp tendency to switch from a positive response to more light north of 5°S to negative, or lower GPP with more light, farther south.



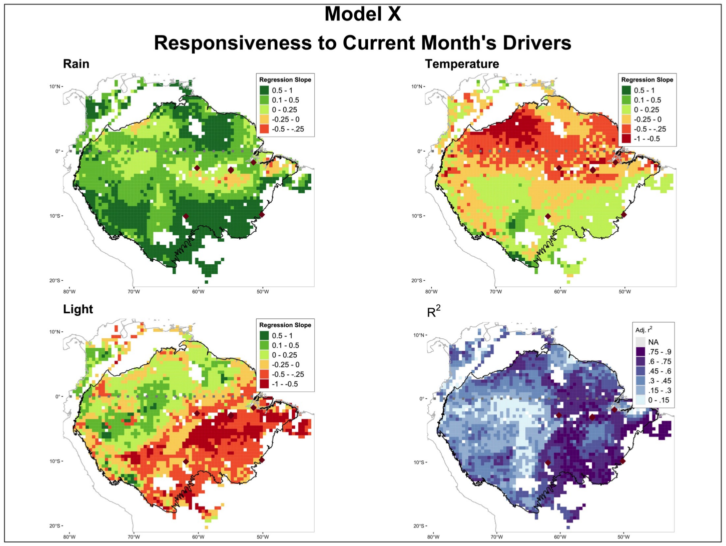
Please see caption below first figure in this section.

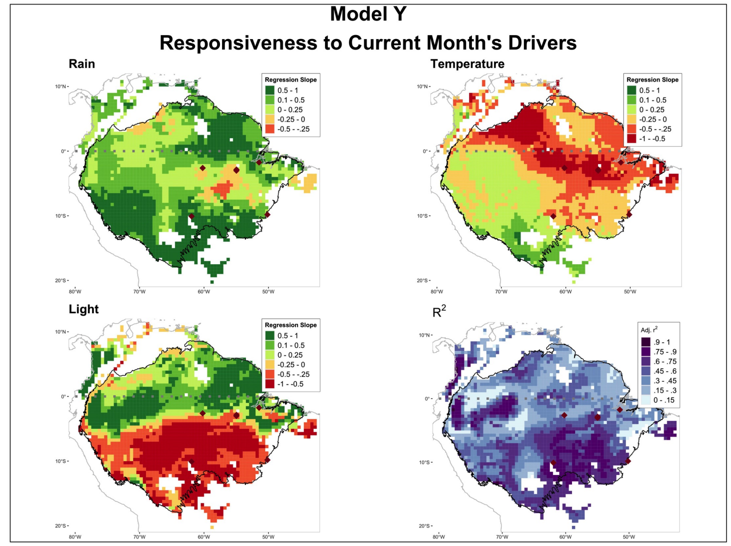


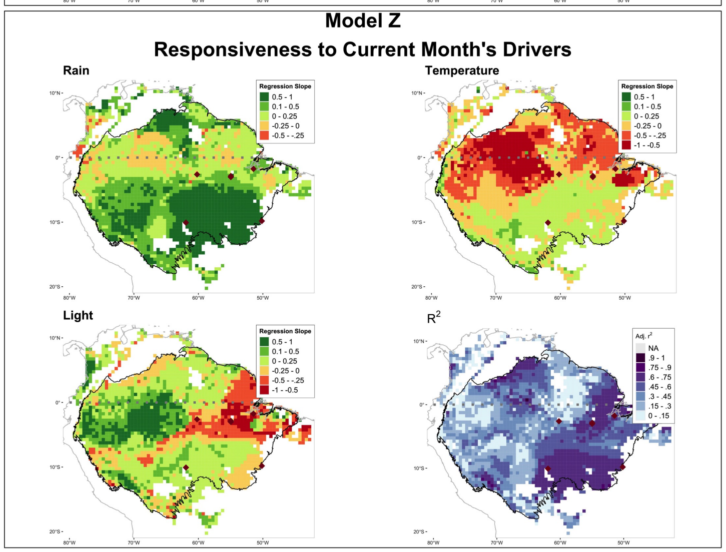
Please see caption below first figure in this section.



Please see caption below first figure in this section.







Please see caption below first figure in this section.

263 **Figure S14: Phase of Site-Level Seasonality**

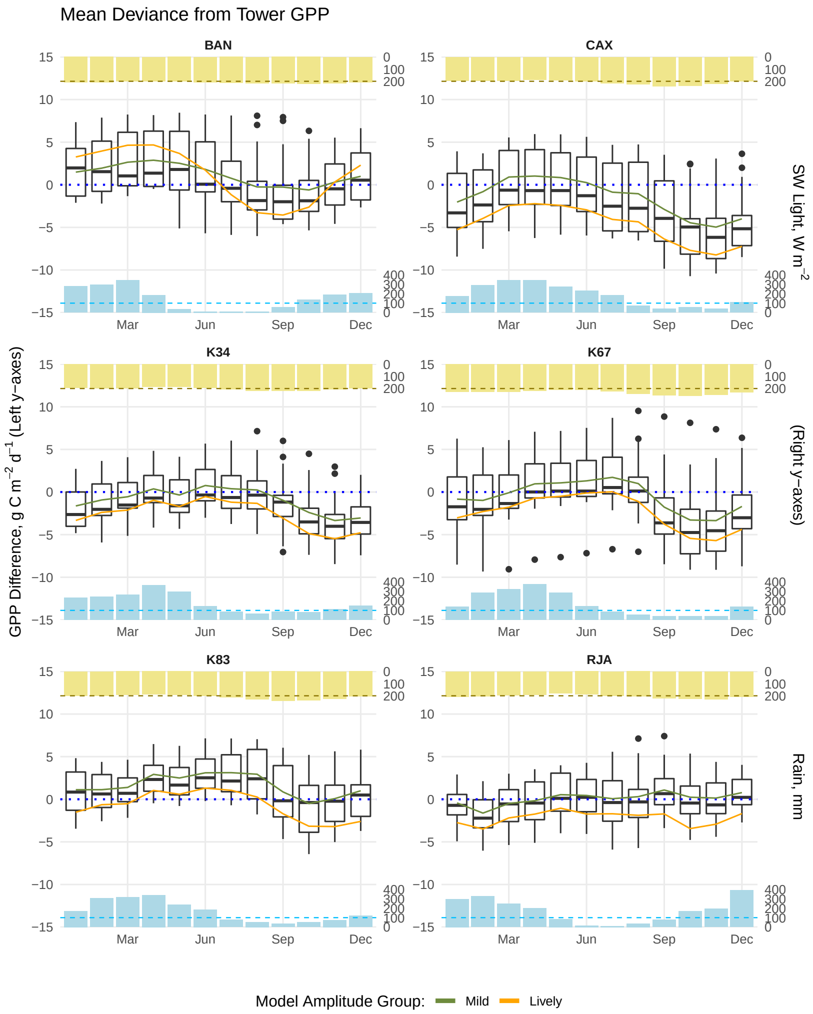


Figure S14: Parallel to the main paper's Fig. 7 for each site separately, showing deviance in modeled GPP from EC estimates. Blue and yellow bars at the top of each panel show mean monthly insolation and rainfall. Relative variations in light are small, and the line at 200 W m^{-2} is an arbitrary visual reference. The dotted line for rain defines dry season months. At most sites, divergences are largest for lively models late in the dry season.

References

Ahlström, A., J. G. Canadell, G. Schurgers, M. Wu, J. A. Berry, K. Guan, and R. B. Jackson (2017), Hydrologic resilience and Amazon productivity, *Nature Communications*, *8*(1), doi:10.1038/s41467-017-00306-z.

Albert, L. P., J. Wu, N. Prohaska, P. B. de Camargo, T. E. Huxman, E. S. Tribuzy, V. Y. Ivanov, R. S. Oliveira, S. Garcia, M. N. Smith, R. C. O. Junior, N. Restrepo-Coupe, R. da Silva, S. C. Stark, G. A. Martins, D. V. Penha, and S. R. Saleska (2018), Age-dependent leaf physiology and consequences for crown-scale carbon uptake during the dry season in an Amazon evergreen forest, *New Phytologist*, *219*(3), 870–884, doi:10.1111/nph.15056.

Albert, L. P., N. Restrepo-Coupe, M. N. Smith, J. Wu, C. Chavana-Bryant, N. Prohaska, T. C. Taylor, G. A. Martins, P. Ciais, J. Mao, M. A. Arain, W. Li, X. Shi, D. M. Ricciuto, T. E. Huxman, S. M. McMahon, and S. R. Saleska (2019), Cryptic phenology in plants: Case studies, implications, and recommendations, *Global Change Biology*, *25*(11), 3591–3608, doi:10.1111/gcb.14759.

Alemohammad, S. H., B. Fang, A. G. Konings, F. Aires, J. K. Green, J. Kolassa, D. Miralles, C. Prigent, and P. Gentine (2017), Water, Energy, and Carbon with Artificial Neural Networks (WECANN): A statistically based estimate of global surface turbulent fluxes and gross primary productivity using solar-induced fluorescence, *Biogeosciences*, *14*(18), 4101–4124, doi:10.5194/bg-14-4101-2017.

Anav, A., P. Friedlingstein, C. Beer, P. Ciais, A. Harper, C. Jones, G. Murray-Tortarolo, D. Papale, N. C. Parazoo, P. Peylin, S. Piao, S. Sitch, N. Viovy, A. Wiltshire, and M. Zhao (2015), Spatiotemporal patterns of terrestrial gross primary production: A review, *Reviews of Geophysics*, *53*(3), 785–818, doi:10.1002/2015RG000483.

Aragão, L. E. O. C., Y. Malhi, R. M. Roman-Cuesta, S. Saatchi, L. O. Anderson, and Y. E. Shimabukuro (2007), Spatial patterns and fire response of recent Amazonian droughts, *Geophysical Research Letters*, *34*(7), doi:10.1029/2006GL028946.

Baker, I., A. Harper, H. da Rocha, A. Denning, A. Araújo, L. Borma, H. Freitas, M. Goulden, A. Manzi, S. Miller, A. Nobre, N. Restrepo-Coupe, S. Saleska, R. Stöckli, C. von Randow, and S. Wofsy (2013), Surface ecophysiological behavior across vegetation and moisture gradients in tropical South America, *Agricultural and Forest Meteorology*, *182-183*, 177–188, doi:10.1016/j.agrformet.2012.11.015.

Baker, I. T., A. Denning, D. A. Dazlich, A. B. Harper, M. D. Branson, D. A. Randall, M. C. Phillips, K. D. Haynes, and S. M. Gallup (2019), Surface-Atmosphere Coupling Scale, the Fate of Water, and Ecophysiological Function in a Brazilian Forest, *Journal of Advances in Modeling Earth Systems*, *11*(8), 2523–2546, doi:10.1029/2019MS001650.

Bonal, D., B. Burban, C. Stahl, F. Wagner, and B. Hérault (2016), The response of tropical rainforests to drought—lessons from recent research and future prospects, *Annals of Forest Science*, *73*(1), 27–44, doi:10.1007/s13595-015-0522-5.

Bonan, G. B., P. J. Lawrence, K. W. Oleson, S. Levis, M. Jung, M. Reichstein, D. M. Lawrence, and S. C. Swenson (2011), Improving canopy processes in the Community Land Model version 4 (CLM4) using global flux fields empirically inferred from FLUXNET data, *Journal of Geophysical Research*, *116*(G2), G02,014, doi:10.1029/2010JG001593.

Borchert, R., G. Rivera, and W. Hagnauer (2002), Modification of Vegetative Phenology in a Tropical Semi-deciduous Forest by Abnormal Drought and Rain, *Biotropica*, *34*(1), 27–39, doi:10.1111/j.1744-7429.2002.tb00239.x.

Broedel, E., J. Tomasella, L. A. Cândido, and C. von Randow (2017), Deep soil water dynamics in an undisturbed primary forest in central Amazonia: Differences between normal years and the 2005 drought, *Hydrological Processes*, *31*(9), 1749–1759, doi:10.1002/hyp.11143.

Chapin, F. S., G. M. Woodwell, J. T. Randerson, E. B. Rastetter, G. M. Lovett, D. D. Baldocchi, D. A. Clark, M. E. Harmon, D. S. Schimel, R. Valentini, C. Wirth, J. D. Aber, J. J. Cole, M. L. Goulden, J. W. Harden, M. Heimann, R. W. Howarth, P. A. Matson, A. D. McGuire, J. M. Melillo, H. A. Mooney, J. C. Neff, R. A. Houghton, M. L. Pace, M. G. Ryan, S. W. Running, O. E. Sala, W. H. Schlesinger, and E.-D. Schulze (2006), Reconciling Carbon-cycle Concepts, Terminology, and Methods, *Ecosystems*, *9*(7), 1041–1050, doi:10.1007/s10021-005-0105-7.

Clark, D. A., and D. B. Clark (2011), Assessing Tropical Forests' Climatic Sensitivities with Long-term Data, *Biotropica*, *43*(1), 31–40, doi:10.1111/j.1744-7429.2010.00654.x.

Collier, N., F. M. Hoffman, D. M. Lawrence, G. Keppel-Aleks, C. D. Koven, W. J. Riley, M. Mu, and J. T. Randerson (2018), The International Land Model Benchmarking (ILAMB) System: Design, Theory, and Implementation, *Journal of Advances in Modeling Earth Systems*, pp. 1–24, doi:10.1029/2018MS001354.

Corlett, R. T. (2016), The Impacts of Droughts in Tropical Forests, *Trends in Plant Science*, *21*(7), 584–593, doi:10.1016/j.tplants.2016.02.003.

Dahlin, K. M., D. D. Ponte, E. Setlock, and R. Nagelkirk (2017), Global patterns of drought deciduous phenology in semi-arid and savanna-type ecosystems, *Ecography*, *40*(2), 314–323, doi:10.1111/ecog.02443.

Doughty, C. E., and M. L. Goulden (2008), Seasonal patterns of tropical forest leaf area index and CO_2 exchange, *Journal of Geophysical Research*, *113*, doi:10.1029/2007JG000590.

Feldpausch, T. R., O. L. Phillips, R. J. W. Brienen, E. Gloor, J. Lloyd, G. Lopez-Gonzalez, A. Monteagudo-Mendoza, Y. Malhi, A. Alarcón, E. Álvarez Dávila, P. Alvarez-Loayza, A. Andrade, L. E. O. C. Aragao, L. Arroyo, G. A. Aymard C., T. R. Baker, C. Baraloto, J. Barroso, D. Bonal, W. Castro, V. Chama, J. Chave, T. F. Domingues, S. Fauset, N. Groot, E. Honorio Coronado, S. Laurance, W. F. Laurance, S. L. Lewis, J. C. Licona, B. S. Marimon, B. H. Marimon-Junior, C. Mendoza Bautista, D. A. Neill, E. A. Oliveira, C. Oliveira dos Santos, N. C. Pallqui Camacho, G. Pardo-Molina, A. Prieto, C. A. Quesada, F. Ramírez, H. Ramírez-Angulo, M. Réjou-Méchain, A. Rudas, G. Saiz, R. P. Salomão, J. E. Silva-Espejo, M. Silveira, H. ter Steege, J. Stropp, J. Terborgh, R. Thomas-Caesar, G. M. F. van der Heijden, R. Vásquez Martinez, E. Vilanova, and V. A. Vos (2016), Amazon forest response to repeated droughts, *Global Biogeochemical Cycles*, *30*(7), 964–982, doi:10.1002/2015GB005133.

Gloor, M., R. J. W. Brienen, D. Galbraith, T. R. Feldpausch, J. Schöngart, J.-L. Guyot, J. C. Espinoza, J. Lloyd, and O. L. Phillips (2013), Intensification of the Amazon hydrological cycle over the last two decades, *Geophysical Research Letters*, *40*(9), 1729–1733, doi:10.1002/grl.50377.

Goulden, M. L., S. D. Miller, H. R. da Rocha, M. C. Menton, H. C. de Freitas, A. M. e Silva Figueira, and C. A. D. de Sousa (2004), Diel and Seasonal Patterns of Tropical Forest CO_2 Exchange, *Ecological Applications*, *14*(sp4), 42–54, doi:10.1890/02-6008.

Harper, A., I. T. Baker, A. S. Denning, D. A. Randall, D. Dazlich, and M. Branson (2014), Impact of Evapotranspiration on Dry Season Climate in the Amazon Forest, *Journal of Climate*, *27*(2), 574–591, doi:10.1175/JCLI-D-13-00074.1.

Huang, Y., S. Gerber, T. Huang, and J. W. Lichstein (2016), Evaluating the drought response of CMIP5 models using global gross primary productivity, leaf area, precipitation, and soil moisture data, *Global Biogeochemical Cycles*, *30*(12), 1827–1846, doi:

10.1002/2016GB005480.

Huntingford, C., P. Zelazowski, D. Galbraith, L. M. Mercado, S. Sitch, R. Fisher, M. Lomas, A. P. Walker, C. D. Jones, B. B. B. Booth, Y. Malhi, D. Hemming, G. Kay, P. Good, S. L. Lewis, O. L. Phillips, O. K. Atkin, J. Lloyd, E. Gloor, J. Zaragoza-Castells, P. Meir, R. Betts, P. P. Harris, C. Nobre, J. Marengo, and P. M. Cox (2013), Simulated resilience of tropical rainforests to CO2-induced climate change, *Nature Geoscience*, *6*(4), 268–273, doi:10.1038/ngeo1741.

Huntzinger, D., C. Schwalm, Y. Wei, R. Cook, A. Michalak, K. Schaefer, A. Jacobson, M. Arain, J. Fisher, D. Hayes, M. Huang, S. Huang, A. Ito, H. Lei, C. Lu, F. Maignan, J. Mao, N. Parazoo, C. Peng, S. Peng, B. Poulter, D. Ricciuto, H. Tian, X. Shi, W. Wang, N. Zeng, F. Zhao, and Q. Zhu (2014), NACP MsTMIP: Global 0.5-deg Terrestrial Biosphere Model Outputs in Standard Format, doi:10.3334/ORNLDAAC/1225.

Jung, M., M. Reichstein, H. A. Margolis, A. Cescatti, A. D. Richardson, M. A. Arain, A. Arneth, C. Bernhofer, D. Bonal, J. Chen, D. Gianelle, N. Gobron, G. Kiely, W. Kutsch, G. Lasslop, B. E. Law, A. Lindroth, L. Merbold, L. Montagnani, E. J. Moors, D. Papale, M. Sottocornola, F. Vaccari, and C. Williams (2011), Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, *Journal of Geophysical Research: Biogeosciences*, *116*(G3), doi:10.1029/2010JG001566.

Jung, M., S. Koirala, U. Weber, K. Ichii, F. Gans, Gustau-Camps-Valls, D. Papale, C. Schwalm, G. Tramontana, and M. Reichstein (2019), The FLUXCOM ensemble of global land-atmosphere energy fluxes, *Scientific Data*, *6*(1), 74.

Jupp, T. E., P. M. Cox, A. Rammig, K. Thonicke, W. Lucht, and W. Cramer (2010), Development of probability density functions for future South American rainfall, *New Phytologist*, *187*(3), 682–693, doi:10.1111/j.1469-8137.2010.03368.x.

Li, W., R. Fu, and R. E. Dickinson (2006), Rainfall and its seasonality over the Amazon in the 21st century as assessed by the coupled models for the IPCC AR4, *Journal of Geophysical Research: Atmospheres*, *111*(D2), doi:10.1029/2005JD006355.

Malavelle, F. F., J. M. Haywood, L. M. Mercado, G. A. Folberth, N. Bellouin, S. Sitch, and P. Artaxo (2019), Studying the impact of biomass burning aerosol radiative and climate effects on the Amazon rainforest productivity with an Earth system model, *Atmospheric Chemistry and Physics*, *19*(2), 1301–1326, doi:10.5194/acp-19-1301-2019.

Malhi, Y., and J. Wright (2004), Spatial patterns and recent trends in the climate of tropical rainforest regions, *Philosophical Transactions of the Royal Society B: Biological Sciences*, *359*(1443), 311–329, doi:10.1098/rstb.2003.1433.

Morton, D. C., J. Rubio, B. D. Cook, J.-P. Gastellu-Etchegorry, M. Longo, H. Choi, M. Hunter, and M. Keller (2016), Amazon forest structure generates diurnal and seasonal variability in light utilization, *Biogeosciences*, *13*(7), 2195–2206, doi:10.5194/bg-13-2195-2016.

Mystakidis, S., E. L. Davin, N. Gruber, and S. I. Seneviratne (2016), Constraining future terrestrial carbon cycle projections using observation-based water and carbon flux estimates, *Global Change Biology*, *22*(6), 2198–2215, doi:10.1111/gcb.13217.

Oleson, K. W., and D. M. Lawrence (2013), Technical description of version 4.5 of the Community Land Model (CLM).

Parazoo, N. C., K. Bowman, J. B. Fisher, C. Frankenberg, D. B. A. Jones, A. Cescatti, O. Pérez-Priego, G. Wohlfahrt, and L. Montagnani (2014), Terrestrial gross primary production inferred from satellite fluorescence and vegetation models, *Global Change Biology*, *20*(10), 3103–3121, doi:10.1111/gcb.12652.

Piao, S., S. Sitch, P. Ciais, P. Friedlingstein, P. Peylin, X. Wang, A. Ahlström, A. Anav, J. G. Canadell, N. Cong, C. Huntingford, M. Jung, S. Levis, P. E. Levy, J. Li, X. Lin, M. R. Lomas, M. Lu, Y. Luo, Y. Ma, R. B. Myneni, B. Poulter, Z. Sun, T. Wang, N. Viovy, S. Zaehle, and N. Zeng (2013), Evaluation of terrestrial carbon cycle models for their response to climate variability and to CO_2 trends, *Global Change Biology*, *19*(7), 2117–2132, doi:10.1111/gcb.12187.

Poulter, B., L. Aragão, U. Heyder, M. Gumpenberger, J. Heinke, F. Langerwisch, A. Rammig, K. Thonicke, and W. Cramer (2010), Net biome production of the Amazon Basin in the 21st century, *Global Change Biology*, *16*(7), 2062–2075, doi:10.1111/j.1365-2486.2009.02064.x.

Rogers, A., B. E. Medlyn, J. S. Dukes, G. Bonan, S. von Caemmerer, M. C. Dietze, J. Kattge, A. D. B. Leakey, L. M. Mercado, Ü. Niinemets, I. C. Prentice, S. P. Serbin, S. Sitch, D. A. Way, and S. Zaehle (2017), A roadmap for improving the representation of photosynthesis in Earth system models, *New Phytologist*, *213*(1), 22–42, doi:10.1111/nph.14283.

Saleska, S. R., S. D. Miller, D. M. Matross, M. L. Goulden, S. C. Wofsy, H. R. da Rocha, P. B. de Camargo, P. Crill, B. C. Daube, H. C. de Frietas, L. Hutyra, M. Keller, V. Kirchhoff, M. Menton, J. W. Munger, E. Hammond Pyle, A. H. Rice, and H. Silva (2003), Carbon in Amazon Forests: Unexpected Seasonal Fluxes and Disturbance-Induced Losses, *Science*, *302*(5650), 1554–1557, doi:10.1126/science.1091165.

Samanta, A., Y. Knyazikhin, L. Xu, R. E. Dickinson, R. Fu, M. H. Costa, S. S. Saatchi, R. R. Nemani, and R. B. Myneni (2012), Seasonal changes in leaf area of Amazon forests from leaf flushing and abscission, *Journal of Geophysical Research: Biogeosciences*, *117*(G1), doi:10.1029/2011JG001818.

Tramontana, G., M. Jung, C. R. Schwalm, K. Ichii, G. Camps-Valls, B. Ráduly, M. Reichstein, M. A. Arain, A. Cescatti, G. Kiely, L. Merbold, P. Serrano-Ortiz, S. Sickert, S. Wolf, and D. Papale (2016), Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms, *Biogeosciences*, *13*(14), 4291–4313, doi: 10.5194/bg-13-4291-2016.

Wei, Y., S. Liu, D. Huntzinger, A. Michalak, N. Viovy, W. Post, C. Schwalm, K. Schaefer, A. Jacobson, C. Lu, H. Tian, D. Ricciuto, R. Cook, J. Mao, and X. Shi (2014), NACP MsTMIP: Global and North American Driver Data for Multi-Model Intercomparison, doi: 10.3334/ORNLDAAC/1220.

Wilson, K. B., D. D. Baldocchi, and P. J. Hanson (2001), Leaf age affects the seasonal pattern of photosynthetic capacity and net ecosystem exchange of carbon in a deciduous forest, *Plant, Cell & Environment*, *24*(6), 571–583, doi:10.1046/j.0016-8025.2001.00706.x.

Wu, J., L. P. Albert, A. P. Lopes, N. Restrepo-Coupe, M. Hayek, K. T. Wiedemann, K. Guan, S. C. Stark, B. Christoffersen, N. Prohaska, J. V. Tavares, S. Marostica, H. Kobayashi, M. L. Ferreira, K. S. Campos, R. da Silva, P. M. Brando, D. G. Dye, T. E. Huxman, A. R. Huete, B. W. Nelson, and S. R. Saleska (2016), Leaf development and demography explain photosynthetic seasonality in Amazon evergreen forests, *Science*, *351*(6276), 972–976, doi:10.1126/science.aad5068.

Xiao, X., Q. Zhang, S. Saleska, L. Hutyra, P. De Camargo, S. Wofsy, S. Frolking, S. Boles, M. Keller, and B. Moore (2005), Satellite-based modeling of gross primary production in a seasonally moist tropical evergreen forest, *Remote Sensing of Environment*, *94*(1), 105–122, doi:10.1016/j.rse.2004.08.015.

Xu, L., S. S. Saatchi, Y. Yang, R. B. Myneni, C. Frankenberg, D. Chowdhury, and J. Bi (2015), Satellite observation of tropical forest seasonality: Spatial patterns of carbon exchange in Amazonia, *Environmental Research Letters*, *10*(8), 084,005, doi:10.1088/1748-9326/10/8/084005.

Zhang, Y., X. Xiao, C. Jin, J. Dong, S. Zhou, P. Wagle, J. Joiner, L. Guanter, Y. Zhang, G. Zhang, Y. Qin, J. Wang, and B. Moore (2016), Consistency between sun-induced chlorophyll fluorescence and gross primary production of vegetation in North America, *Remote Sensing of Environment*, *183*, 154–169, doi:10.1016/j.rse.2016.05.015.

Zhang, Y., X. Xiao, X. Wu, S. Zhou, G. Zhang, Y. Qin, and J. Dong (2017), A global moderate resolution dataset of gross primary production of vegetation for 2000–2016, *Scientific Data*, *4*, 170,165, doi:10.1038/sdata.2017.165.