Peer Review File

Manuscript Title: Amazon forest biogeography predicts resilience and vulnerability to drought

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

Comments

In this work, the authors combine valuable complementary sources of data – remotely sensed photosynthetic indices, ground-based demographic data, climate, soil, and topographic information – with the aim to better understand the processes underlying the spatially heterogeneous nature of Amazonian forest drought responses. The overarching goal of the study is to use a mix of community features (e.g. tree height) and abiotic factors (topography, climate, soil fertility) to predict the photosynthetic indices during three major droughts in the past two decades, to then define a biogeography of Amazon drought resilience and vulnerability. A key result of the study is the improved explanation of drought response variation by accounting for (a proxy of) water table depth in southern Amazonia, highlighting regions of deeper water table depths as being the ones corresponding to the increased browning and tree mortality. Similarly, highlighting the interaction between water table depth and the duration of drought events to understand forest vulnerability spatio-temporal variations importantly draws attention to the urgent necessity to measure and better account for hydrological and edaphic/topographic data to improve the critical forecasts of future paths of carbon dynamics in the Amazon under alternative climate change scenarios.
The study is novel, well written, its hypotheses clearly stated, and its objectives are of utmost importance given the urgency of getting a better mechanistic understanding of Amazonian forest responses to extreme weather events such as droughts.

I have a series of minor suggestions and questions, listed below, as well as a small number of bigger concerns – all of which addressable, in my opinion.
These bigger points relate to 1) the predictive statistical approach used in combination with causal inference objectives and language, 2) the definition of resilience and how it is assumed to be related to the remotely-sensed photosynthetic vegetation indices, and 3) the absence of soil texture in the study, in spite of its potential relevance during droughts for soil water retention capacity.

Note that line numbers only appear every 5 lines. I will therefore refer to line numbers as the closest previous line number followed by a dot and the line number from there (e.g. 42.3 for the third line of the 42th line chunk).
Major points

I. Data analyses and causal inferences

I have two points to raise regarding the thinking behind the statistical approach of the study, as well as two suggestions.

1) Predictive versus causal approaches:
While having PAR, VPD, P and MCWD together in the same model is not a problem for the purpose of statistical prediction, it may hinder/obscure interpretations when it comes to drawing causal inferences from the fitted model coefficients (e.g. Pearl et al. 2016, Rohrer 2018, McElreath 2020, Arif and MacNeil 2022). Here the authors include all predictors in a single model (as well as all pairwise interactions between them) irrespectively of the effects these might have the ones on the others (or at least the authors do not address their consideration of this point). For example, I would assume that PAR influences temperature, which influences VPD, hence creating a causal relation between PAR and VPD. Similarly, P partly causes MCWD.
Having a causal model (e.g. a directed acyclic graph, DAG) of the assumed causal relations among explanatory variables and with the response variable is a necessary condition to be able to draw causal inferences from the coefficients of a fitted statistical model (McElreath 2020, Addicott et al. 2022, Arif and McNeill 2022). Without such causal model to accompany a statistical model, interpreting all predictors' coefficients as if their magnitude were the reflect of their ecological importance for the response makes the (hidden) assumption that all predictors are independent of one another – which is rarely the case, and probably not here – and leads to mixing total causal effects, partial causal effects, and potentially confounds (i.e. non-causal paths relating an explanatory variable to the response of interest). Simple rules such as the "back-door criterion" used in light of a causal diagram including all important variables are very helpful to guide the careful selection of the variables to include/exclude to correctly assess the causal effect of a given predictor (Addicott et al. 2022, Arif and MacNeil 2002).

2) Model selection: My second "statistical point" is regarding the use of a "dredging"' model selection approach, consisting of exploring all possible combination of the predictors (and their interactions) to then select the few ones with the lowest AIC and R2. If this approach ensures the best statistical prediction, it does not ensure a correct causal model, and can lead to wrong causal interpretations (Addicott et al. 2022, McElreath 2020), hence potentially confusing the objective to understand the mechanisms underlying the spatial heterogeneity in drought responses of the Amazon.

While using a model selection approach aimed at prediction is not a problem, it is very difficult to do so without confounding "best" resulting models, once selected, with "models providing the best causal understanding of the data". Note also that replacing "causal" by "association" does not really solve things, as two variables can for example be very associated because they have a common cause (but manipulating one would never change the other), or because of a collider bias (see Rohrer 2018 for an example). It quickly becomes difficult to interpret the results of a predictive model while totally avoiding implications of causality, in the absence of a causal model explicitly laying out the causal assumptions underlying the relations among the variables of the system.

I appreciate that the above is a common practice, but it is also a risky – or wrong – one when combined with a causal interpretation language as is often the case, and is the case here. I would appreciate it if the authors could take the above into account, one way or the other, and if they decide to keep their current analytical approach, to then justify how or why they consider they can do so, in light of the above considerations.

Suggestion to address this while keeping the changes to the current analytical framework to a minimum:
a) As explained above, good predictive models are not necessarily good causal models (Addicott et al. 2022, Arif and MacNeill 2022), and while I can imagine that "forest ecotypes" may indeed be important – it would make sense to expect so, to some extent –, drawing causal inferences, both qualitatively and quantitatively, requires making explicit ecological assumptions regarding causal relations among the variables of the system (for example, using directed acyclic graphs, DAGs), to guide a careful choice of variable to include or not to include together in a model, for inference about a covariate of interest (see backdoor criterion; Arif and MacNeill 2022).
If the aim is to draw causal inference, as is suggested by the objectives and the language used in the results regarding ecotypes (e.g. lines 215 and below), I recommend generating a DAG, or a small set of sound alternative DAGS if necessary, and using the backdoor criterion (and if necessary frontdoor criterion) to justify inclusion/exclusion of control variables for the models aiming to assess the causal effect of each covariate (HAND, tree height, etc). This would allow testing and quantifying the effects of each of these in terms of total causal effects (hence also allowing comparing their importance). Different models including different control variables may be necessary, depending on the causal and non-causal paths highlighted by the backdoor criterion, to capture the causal effects of different explanatory variables.
Note also that the functional form of the relations among variables is not included in a DAG, which therefore would not hinder the use of GAMs. In practice, little change to the current analyses; only an explicit causal justification for the inclusion/exclusion of covariates in each model, to ensure causal interpretability of fitted coefficients.

b) The predictive approach – AIC-based model selection following "dredging" – used here could be maintained for the purpose of prediction only, and to compare the predictive power of a model considering these variables and the concept of ecotype (regardless of their exact causal relations / creation of confounds) with a model only considering climate.
In this sense, the predictive model would not especially serve the interpretation of the detail inside the model, but would show the added predictive value of HAND, tree height etc, while the models structured based on a causal model (e.g. a DAG) would allow drawing causal inferences in terms of mechanisms (line 215.5).

While this may look like much, I believe that the above could relatively easily be added to the present analytical framework and would provide a correct and more justifiable/robust basis for the interpretation and implications of the findings – that is, a crucial point given the importance of the objectives of the study.

References
Arif and MacNeil 2022, Ecological Monographs: https://doi.org/10.1002/ecm.1554
Arif and MacNeil 2022, Ecology Letters: https://doi.org/10.1111/ele.14033
Addicott et al. 2022, Frontiers in Ecology and the Environment: https://doi.org/10.1002/fee.2530
McElreath 2020. Statistical Rethinking, Second Edition.

Rohrer 2018, Advances in Methods and Practices in Psychological Sciences: https://doi.org/10.1177/2515245917745629

II. Definition of resilience

1) The concept of resilience has a central role in the study, yet it is only defined in brackets in a sentence centred on something different (line 110.3). Moreover, there are many definitions of resilience out there (e.g. Grafton et al. 2019, Nikiinmaa et al. 2020). Resilience needs to be clearly conceptually defined and explicitly related to the measurements taken as proxies of resilience in the paper (with justifications). I suggest starting the paragraph where the concept appears for the first time (close to line 110) with this definition and links to the data of the study.

2) Following up on the previous point: Why would an increase in photosynthesis/greening correspond to a higher photosynthetic resilience? If resilience defines the capacity of a system to return to its stable or equilibrium state following a disturbance, then how does an increase in photosynthesis *during* a disturbance indicate resilience? It surely is a notable response, but not one that I would think to relate to the concept of resilience at first. In addition, would a forest that would not change its photosynthesis rate during the drought (i.e. no greening, but no browndown either) not be as resilient as a forest increasing photosynthesis (if resilience were related to photosynthesis activity during the stress)? The former forest would simply maintain its functions without showing either greening or excess mortality. Would resilience not be as much about recovery *following* a drought (resilience sensu stricto) as it is about photosynthesis response *during* the drought ("resistance" dimension of resilience sensu latto)? It appears that the authors may be referring mostly to the *resistance* property of the multidimensional notion of resilience (e.g. Grafton et al. 2019, Nikiinmaa et al. 2020). Is that correct? If so, properly defining this upon first use in the main text (and in more details in the Methods) would help conceptual clarity, by defining a more precise framework to formulate and think of the questions and results.

References
Grafton et al. 2019, Nature Sustainability: https://doi.org/10.1038/s41893-019-0376-1
Nikiinmaa et al. 2020, Current Forestry Reports: https://doi.org/10.1007/s40725-020-00110-x.
III. Soil texture

The authors aim to address key drivers of forest sensitivity and resilience to drought, and focus their efforts on a proxy of water table depth, a proxy of peak dry season water deficit, and soil fertility, to go beyond climate data alone.
I would expect a key missing ingredient determining soil hydrological stress to be soil texture (and ideally structure, e.g. approximated by bulk density). Droughts can be attenuated or amplified depending on the clayey/sandy nature of the soil, which relates to its useful water retention capacity; not only the water table depth.
Could this be included in the analyses, even if only for a subset of the whole dataset? This should at least be acknowledged and discussed in relation to the results. How much soil texture variation is there within the three major biogeographic regions defined by the authors, and how is this expected to affect the biogeography of drought as perceived through the current analyses? Would interaction terms among soil texture, HAND, and/or MCWD not be expected to be necessary to obtain a more

complete picture of the spatial distribution of drought responses?

Minor comments

- First sentence of the introduction: I'd remove the parentheses for the dates.

- The beginning of Fig. 1 legend could be shortened quite a bit: instead of repeating three times what maps 1,2,3 represent, only changing the year they correspond to, I would suggest writing this only once (like for the insets). Something like: "Standardized anomalies of drought Enhanced Vegetation Index (EVI, a proxy of photosynthetic capacity) in drought-affected pixels (see Fig. S1) for the droughts of 2005 (A), 2010 (B), and 2015 (C), with ellipses highlighting green-up and browndown patterns that correspond to shallow and deep water tables (see Fig. 2A)." And no need to explain what a standardised anomaly is in the legend I'd say.

- Still in Fig. 1 legend: If I am understanding correctly, the delta EVI are statistically estimated, is that right? If so, it would be useful to provide some measure of uncertainty around those estimates (e.g. confidence intervals). Note also that the closing parenthesis for the delta EVI of 2005 is missing.

- Line 50.4: "a strong decrease" of what? Greenness I suppose, but please make it explicit.

- Line 60.1: "exhibiting green-up, or reduced mortality": Are we to understand that these are one and the same, or two alternatives corresponding to resilient scenarios? I am asking as it is stated further up that the 2005 drought saw simultaneous greening up and increased mortality from ground-based data, which seems at odd with the content of the parenthesis here.

- Line 60.3: I would remove "well-" in "well-validated" (unnecessary), or provide a more specific definition than "well" about what is special about the indices.

- Line 65.2: I suggest clearly defining "forest ecotype" at the end of the sentence line 65.2, as it is very unusual to use the term "ecotype" – typically reserved for *intraspecific* levels of *genetic* variation related to different habitats – to characterise *community*-level differences mostly related to *abiotic* factors.
Even better might be to use a different word altogether (e.g. "habitat", simply), to avoid confusion using "ecotype" for a definition not corresponding to its common one in biology/ecology.

- Line 75.3-4: I am not sure to follow the parenthesis explanation for the hypothesis. I would understand that fast turnover due to a dominance of fast growers would cause a strategic disadvantage to slow growers (with drought resistance traits) on fertile soils (if one forgets about soil texture and potential edaphic droughts, regardless of fertility). The authors could probably rephrase/clarify their justification for this hypothesis.

- Line 95.2-5: This part lacks some details to be clear.
1) I am guessing the authors mean "Drought-affected regions *for each drought year* (…)". Correct? If so, please clarify.
2) "The 20-year mean" is mentioned here for the first time: specify what are the reference years

(e.g. “for a 20-year baseline based on XXXX-XXXX”).
3) Regarding drought duration: lines 95.3-5: Please clarify “as the *number of months* that their *monthly* CWD remained below the *20-year* mean MCWD”, if this is indeed how duration was calculated.

- Lines 100.4-5: Does the “across multiple droughts” mean that the data for the three droughts were used in a single analysis, or were combined a posterior from separate analysis on each drought? Please clarify what the input data was, in this sentence.

- Line 110-115: see resilience point raised earlier, with respect to definition and justification of hypothesis.

- Line 115.1: The GOSIF acronym was not introduced before.

- Line 115.3-4: I would rephrase slightly to replace positive “greening” anomalies and negative “browning” anomalies by positive vegetation index anomalies (i.e. “greening”) and negative anomalies (i.e. “browning”). To avoid confusion with a negative “browning” anomaly being less browning than the baseline, for example.

- Line 120.1-3: The “because” in the sentence suggests a capacity to draw causal conclusions from the observation of higher solar radiation anomalies during periods of greening.
This sentence warrants either more caution in the phrasing, replacing “apparently” by “potentially” and “which were particularly advantageous” by “which may have been …”, or requires a stronger justification for the causal interpretation (probably the former; see also comment with respect to Fig. 2C).

- Fig. 2B: No legend provided for the empty symbols.

- Fig. 2C: Comment directly related to the one of line 120 about solar radiation:
1) The legend mentions a figure of EVI anomaly *sensitivity* to PAR anomalies, but it is unclear whether Fig. 2C presents a model output, or plots the data simply. I am guessing it is the former, as I see error bars, but the error bars are not mentioned in the legend, and it would be helpful to specify the model this result came from (even if briefly).
2) The legend does not say whether the upper panel of Fig. 2C is for the relation from all drought severities or one in particular. Is this a model fit across all drought severities?
3) Solar radiation has a very high value at 1.5 of PAR anomaly for severe droughts, that has nothing even close to it for the modest and medium droughts. How robust is the association between EVI anomalies and PAR without this extreme value, in the model? Does it remain after removing it, or is it entirely driven by this value?
4) Regarding the causal inference implied by “sensitivity of EVI to PAR”: Unless justified by making the underlying causal assumptions explicit somewhere (i.e. having a causal model or diagram underlying the statistical model), implying a cause from an association is very delicate. Solar radiation also causes temperature, which causes VPD, both of which also potential drivers of a greening/browning, for example. See more about this in the main points raised further above.

- Line 124.4-125.5: Lots of information in a 7-line sentence. I suggest putting up front the main message, and dividing this in shorter sentences to help the reader.

- Line 130: “less negative (less browning) anomalies” could simply be written “less browning”, to ease the reading. (The same sort of thinking applies to many long sentences throughout the manuscript). Note also that since the structuration of anomalies with water table depth has already been presented at this stage, the whole “with less negative (…) over deep water tables” ending the paragraph is redundant and could simply be removed.

- Line 180.5: The authors cite 31-33 to support that tree height is a proxy for rooting depth. Chitra-Tarak et al. 2021 (31) does not show this association however (unjustified citation, to be removed). Brum et al. 2019 (32) indeed show a relation between DBH and oxygen stable isotope ratio in the xylem used as proxy of rooting depth, based on 12 species in a seasonal Amazonian forest, while Tumber-Davila et al. 2019 (34) show a relation between shoot height and rooting depth in woody species, but mostly across biomes, with no clear information (that I could find) regarding such a relationship in the moist tropics (most likely a minor proportion of their dataset). On top of removing the citation of 31, I would advise acknowledging how little is still understood (or even available in terms of data) regarding rooting depth and its determinants in the (neo)tropics. A brief discussion regarding this source of uncertainty here would be a valuable addition to a perspectives paragraph. This could go together with the perspectives addressing soil texture (in case it could not be included in the analyses).

- Lines 209.5-210.1: See previous comment regarding the use of “ecotype” in the study. Note also that ecotype has been used in previous sections but is being defined here, in parenthesis. The concept of ecotype or whatever other term is used should be clearly defined upon first use.

- Line 210.2: See conceptual comment regarding actual match between greening up and resilience, as well as regarding the definition of resilience.

- Lines 215.4-5: The attribution language and causal inference drawn in this paragraph relate to my concerns regarding the absence of a causal model to accompany the statistical models used. The authors use a predictive approach, selecting models based on an information criterion mostly, and then interpret the output in causal terms (causal language of mechanisms, sensitivity to X or Y, etc, throughout the results). See related point at the beginning of the review.

- Lines 475-480: Consider having two sentences here.

- Lines 525.3-530: Speaking of results only in terms of them being “statistically significant” or not is insufficient to highlight their importance or the relevance of the related signal. Some of the figures in Fig. S8 to Fig. S10 are apparently driven by a few extreme points, potentially leading to significant p-values given the relatively low number of observations, and there is a substantial and/or asymmetric variation around the regression lines in some of these figures. The R2s and these considerations should be acknowledged in the text to provide a more complete depiction of these relations.

- Line 565.4: Ecotype variable: A key element that should be added at the beginning of this section is the definition of what the authors call “forest ecotypes”, before diving into the different elements they considered to make up the variable. This is particularly necessary as the term “ecotype” is typically defined as concerning geographically distinct groups of individuals within a same species (i.e. populations) that have genotypically-defined adaptations to certain environments, but is not often used to characterise different communities irrespectively of their species composition. This would ensure a clear understanding of what the authors mean.
Similarly, given the structure of Nature papers, I would re-define “forest ecotypes” upon first use in the main text (outside the Methods section).

Line 595: Does “community-averaged” mean a community-weighted mean? If so, please specify.

- Lines 605 to 615.5: How do they authors expect the very different sample size for shallow and deep water table plots (i.e. 25 vs 321) to affect their results and conclusions, or the uncertainty around these?

- Line 615.2: What were the demographic baselines used to define ‘demographic anomalies’?

- Line 635: closing parenthesis to be removed.

- Line 640.4: Please remove “tree height” from the variables contributing the most to the first PCA axis. Fig S3A clearly shows a minimal contribution of this variable to the first axis. Checking the variable contributions would confirm this.

- Line 640.5: You could name this second axis something more specific than “climate”, since the two variables contributing to it are both related to water deficit, or water stress.

- Line 645: Fig. S3A does not show “three nearly uncorrelated axes” (as in PCA axes), since only two axes are displayed in that plane. What appears is correctly interpreted for PCA axes 1 and 2 (or will be, once “tree height” will have been removed from the important variables of axis 1), but soil fertility is a variable contributing mostly to the first PCA axis, and somewhat to the second too. In that sense, it would be okay to maintain that it captures some complementary information to that of [Fabaceae proportion + wood density] and [MCWD var. and minimum precip.], but it is incorrect to speak of uncorrelated axes. Or change terminology to speak of “three complementary dimensions of forests”?
I also suggest acknowledging that the “vegetation axis” is also a “soil fertility axis”, negatively correlated to Fabaceae abundance and wood density.

- Line 750.1: Add reference for the R package mgcv.

- Lines 740.3-760.1: see two statistical points made at the beginning of the review, regarding predictive vs causal approaches, and model selection.

- Line 830.2: specify if “concentration” refers to effective cationic exchange capacity.
Line 970.5: Location of the code to reproduce the analyses and graphs is missing.

This is an impressive and important work, and I hope the authors will be able to address the above comments.

David Bauman

Referee #2 (Remarks to the Author):

A. Summary of the key results

This paper looks at the association between indices of forest vegetation responses and drought in response to landscape predictors, and incorporates demographic data from forest tree plots. The authors reveal a heterogenous response—termed a biogeography of forest drought responses—and suggest a mechanism for its occurrence based on regional variation in average forest canopy height ("tree height"), coarse measures of height above major drainages, taken to be measures of water table, and soil fertility. They use heterogeneity in responses to 2005, 2010, and 2015 droughts, the latter particularly strong in the N. Amazon, as tests of the mechanisms. The authors suggest that a predictive understanding heterogeneous responses to drought are important for multiple factors, from conservation to understanding resilience.

B. Originality and significance: if not novel, please include reference

The ideas and hypotheses generated bring together a large amount of disparate information to answer questions of broad relevance and interest. Each of the pieces has been discussed, but the inclusion in one effort to generate a biogeography of drought with a mechanistic interpretation is new.

C. Data & methodology: validity of approach, quality of data, quality of presentation

I have a number of problems with the approach, even though I am sympathetic to the exigencies of large-scale remotely sensed data, aggregated and interpolated structural and soils data, etc. and detail a number of those below. This ranges from the remotely sensed indices of drought response and their concordance, what the derived variables are really giving information about (which affects attribution of effects observed).

Though the title and the paper refers to Amazonian forests, it appears that the authors did not constrain the analyses to only forests, but rather included large areas of tropical savanna including Brazilian cerrado and edaphic savanna in Boliva, as well as deforested areas/agriculture, and high elevation Andean grasslands. Within forested areas, Andean forests (albeit within the Amazon drainage), higher elevation forests on Guianan teupis, and the large areas of degraded forests in the Amazon, are also included. This could lead to having greater than half of the inference space for the paper not being in what is traditionally thought of as Amazonian forests, with the accompanying influence on the GAM fitted models and predictor variable influence. If the authors did extensive masking, exclusion of non-forest habitat, etc., this should be stated in the paper, and these areas should be excluded from the figures.

The figures are well done, but reach a level of complexity in some cases that makes them opaque. It would be good to think about how to simplify the presentation in the more complex cases.

D. Appropriate use of statistics and treatment of uncertainties

Following on point C, many of the statistics are appropriate, but there are areas for improvement noted in the detailed comments below, particularly in attribution of results and inference. More diagnostics of the GAM fit and partial plots would be important.

E. Conclusions: robustness, validity, reliability

In many cases the authors make very strong conclusions from weak relationships, or data open to multiple interpretations (many cases are noted below). The authors need to be their own harshest critics and present alternative hypotheses, and alternative interpretations. The authors spend a lot of prose talking about how the results fit their general model when there are multiple interpretations, and the droughts themselves are fairly different.

This relies on the authors having confined their inference to lowland Amazonian forests proper. If they have not, and include non-forest areas and other habitats listed above, their inferences do not match their hypotheses and their interpretations of the results of the analyses would be suspect, if not invalid.

F. Suggested improvements: experiments, data for possible revision

GOSIF and EVI are argued by the authors to give similar information but their own data have spatial differences, and also the anomalies are poorly correlated with each other (Fig 1 vs. Fig S2, Fig S10). This is discussed in more depth below, but given the known variability in what differently derived vegetation indices inform us about, and how they fail, why not use multiple indices in a composite or latent variable analysis of drought?

G. References: appropriate credit to previous work?

Yes

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

Clarity and writing is good. The results are oversold given the strength of the data.

Additional Comments, General and Specific:

In general, the paper does a good job of establishing the hypothesis that there are non-climate factors influencing drought response, and that these should lead to regional variability in drought responses. The evidence they present for the predictors being important is relatively weak, but that is fine in general—this feels more like a molecular biology paper where there is an overview of some observed correlations/responses, and then a model/explanatory framework is developed post hoc, setting the stage for further testing.

The Other Side of Drought focuses on water table depth and posits that forests on shallow water tables should be resilient to drought with lower brown-down or even green up, the latter due to less anoxia (presumably from decreases in water saturated surface and near-surface soils) and decreased cloud cover. A difficulty in this is the spatial resolution necessary in the predictors (forests can change height substantially within 1 km2 resolution of the sampling, including both low HAND and high HAND areas), as well as the temporal resolution of the sampling (one would need to show a decrease in cloudiness from the long term average in times of year when species can take advantage of it, or when it is particularly cloudy).

The Soil Fertility Hypothesis posits that in fertile forests with higher turnover rates investment in traits for drought resistance are reduced as trees will be quickly replaced, and these forests will have greater drought susceptibility. The authors write this as if it was an individual species making these allocations, or if trees were all one species, when the paleo and modern literature would both show that if a community were to be drought resistant it would be due to environmental sorting favoring the assembly of communities of species/taxa with different trait combinations. Forests 'adapt' to large changes in environmental conditions through changes in species composition.

The third hypothesis is that trees with drought tolerance or drought avoidance traits are more drought tolerant, which seems to be hard to argue with (again, I think it is important to clarify that forests only have trait spectra due to the performance of individual species possessing those traits). Here the traits chosen to reflect drought tolerance are debatable in many cases—tall trees, or at least large trees, are chosen by the authors, when literature on drought, particularly Nepstad's drying experiments, show that large trees are the first to succumb to experimental drought, which was interesting. Another reason you have large trees that have resilience in nutrient poor areas is just that growth rates are low and stability is favored—this leads to traits like higher wood density as well. Of course, when water stress gets even stronger, trees decrease in stature. The hypothesis that long-term episodic drought resistance is favored by large, densely wooded species is interesting and could be tested with mortality data within demographic plots.

The argument that these interact with climatic gradients is good, though the mechanistic interpretation is not supported—it is phenomenological, correlated with aggregate traits.

Drought depth and duration is known to have distinct effects on species in research from the temperate zone (in the major Texas, US droughts of the 1950s and 2010s, one killed junipers and the other killed oaks, with one drought being very deep but short, the other being exceptionally long). Do these two axes of drought have distinct effects in the Amazon? Looking at figure S4, the three droughts are not the same treatment applied to different areas—they are very different droughts that also happened in very different biogeographic areas. It would seem that the important approach would be to find forests of different characteristics (variability in HAND, fertility, traits) within the regions affected by the different droughts and focus on that comparison.

A difficulty of this paper is the attribution of causes to the mechanisms inferred. For example, the southern Amazon is referred to as high fertility, but it is only the western and a few parts of eastern S. Amazonia that is so. The much of the east is Brazilian shield, and has relatively low diversity forests, and corresponding differences in tree demography and productivity.

A general issue with studies such as these is understanding the true meaning of what the vegetation indices (VIs) are revealing, and this is a longstanding debate in the field, particularly with respect to the Amazon, and several of the authors are key participants in this debate, particularly with the key measurement used in this paper, EVI. The recent review by Zehng et al. in NEE 2022 is a good overview. For the drought response, it would be more compelling to use a composite, or multivariate, analysis using different satellite-derived indices. The authors include GOSIF as an alternative analysis, but it gives different information on drought effects (direct PS measurements) that correlate better with GPP, and that is not redundant with EVI. Comparing figure 1 and figure s2 shows the spatial locations of the drought anomalies are similar but not the same, and if the same ellipses were put on Fig. S2 that would be more apparent. I'm sure the last thing the authors want to encounter is another reviewer talking about issues with satellite-derived measures of forest performance, but I think that acknowledging the differences in information given and why the occur might strengthen the mechanistic case they are making. You could think of drought response as a latent variable with many inputs, particularly as the paper goes on to make demographic and functional connections. Also, for example, using NVI for statements like L112

Another interesting part of this manuscript are the regional generalizations and the generalizations about the droughts. For reasons listed above and below, the regional generalizations, particularly about the S. Amazon, are inaccurate. Also, looking at the droughts, they are qualitatively different, not just spatially different. For example, the 2015 drought is centered on N. SA and from figure S4 looks to be an exceptionally dry wet season rather than a dry dry season as in areas experiencing the 2010 and 2005 drought (peak rainfall in Boa Vista is in May/June with a sunlight minimum in June, the exact inverse of Rio Branco, or even Iquitos and Manaus. Given that light limitation is a primary hypothesis, and that light limitation is much more likely to be a wet season phenomenon, it would make sense to place more emphasis on how the timing with respect to the wet/dry season affects interpretation.

Specific Comments

Figure 2. HAND—it looks like HAND is also calculated for the Andes, when the paper does not treat Andean forests. If forests above 500m (a traditional definition that separates the Andes from Amazon) were removed, it would give much better discrimination among forests. Also, in several of the authors' original conception of HAND it captured hydrologically relevant, very local features like flow patterns on hillslopes to drainage and had good correspondence to local measurements of soils (Nobre et al. 2014). Here the application resolution is very much coarser--looking at the map it appears HAND is not saying much about drainages per se on the scale that trees experience water

tables, but rather looking at position with respect to major geomorphological features—the Guinan and Brazilian shields, the Fitzcarrald/Purus arch, etc. (e.g. figure S11) as opposed to how it was conceived.

Also, within the HAND results, the effect is largely confined to forests with HAND ~>10m. These would include many forests that are seasonally inundated, large wetlands, etc. How was seasonally inundated (varzea) forest dealt with? In normal years these may have lower greenness just because they are underwater for large portions of the year, giving large positive anomalies in years where they are not flooded, etc. Integrating the GOSIF measurements, or other VI measurements, may be a way around this.

L31: A tipping point of what?

L45: Is it undermining ecohydrological strategies? Or is it just exceeding system capacities for resilience, and those capacities depend on the predictors included, plus any unobserved capacities/adaptations for drought resilience? The latter is an important factor—forests in more seasonal areas of the Amazon have tree floras with a higher proportion of individuals and species from families with dry forest affinities than areas in everwet forests (Gentry 1988, Pitman et al. 2000). The portfolio of adaptations to drought contained within lineages is clearly important in setting distributions and ecotypes, and is much more than simple measures of height, etc.

L91. It doesn't seem from figure 5 like three really emerge—it is a highly complex, geographically localized set of predictions that don't square easily with the three defined regions in figure S3. Particularly in the southern Amazon, which is highly heterogeneous and that defies easy classification. The Everwet Amazon shows much more heterogeneity than is described as well.

More generally, the S. Amazon is highly heterogeneous in soil fertility, as figure 5 alludes to, etc.

L115. HAND would seem to be much more of a measure of topography than water table at the scales measured—how does inference about HAND change with different scaling? Or is EVI binned at such a coarse scale that it can't be resolved? See comments above as well.

L146. Again, were PAR anomalies higher in 2015 because that drought occurred during the wet season in many areas (though with some dry season affect in S. Amaz)?

L162. This result about the RAINFOR plot AGB mortality and HAND is compelling, but it, and the general link of demographic effects to EVI and HAND are tempered when one looks at figures S8-S10. ANPP rates are explained poorly by EVI, accounting for only 25% of the variation, the plots in figure S9 showing little predictive ability for mortality from EVI (B) and that looks to be influenced by the highest mortality plot. Similarly, the three highest recruitment plots have the highest EVI anomaly values (C), but the rest of the points cover the full range of EVI anomaly—it is likely nonrandom, but how much information does it convey beyond the extremes, particularly if one were predicting a full explanatory biogeography? The mortality:recruitment vs. water table depth relationship appears to rely on 1-2 plots (D), and the mortality WTD plot is similarly dependent, but NS. Also, why not plot recruitment in a panel (F)?

Figure S10 shows a poor relationship between EVI and GOSIF anomalies—what is the explanation for this? It undermines the argument made that these are telling similar stories and have good predictive value. Three-quarters of the variation in one is not explained by the variation in the other (A). Panels (B) and (C) are heavily influenced by 1-2 data points. It isn't fatal, but it should temper the argument, and the authors might use this to present this part of the argument as the basis for a hypothesis and to point for the need of sites with additional demographic data, or other measures of forest structural change, etc., rather than key support.

L185 and on. This is one interpretation of the drought effects, but what about other interpretations? There are many explanatory degrees of freedom here, so this story becomes one of many. If the authors are their own worst critics, what would they say about the coarseness of HAND as used in this paper, the scale of measurements of canopy height, which does not necessarily equal average tree height in many areas of the Amazon, particularly given that much of the area measured is degraded or otherwise human altered forest.

Figure S11. The authors' assertions about the simplicity and sufficiency of the three-area classification is belied by the complexity in figure S11. The authors state that the S. Amazon has high fertility and short trees, but it has both high fertility and low fertility and very tall and short trees (and a lot of degraded and disturbed forest). Puzzlingly, cerrado vegetation—a mix of trees and grass—is included in this analysis, which is qualitatively different habitat type from Amazonian forest. With low tree height, and with trees interspersed among large areas of grass, the canopy heights will be low. Similar savanna in Bolivia is not masked out, and I did not find in the SI where the authors did this or mask out degraded or deforested areas. The analysis and conclusions should certainly be confined to forest only. S3B shows the problems with this classification. The paper has forest in the title but a large percentage of the inference area is not forest, but rather cerrado, edaphic savanna, deforested areas, agriculture, and degraded forests. Within what is commonly thought of as Amazonian lowland forests, apparently, forests in the Andes and Guianan tepuis are included.

L240. Given the above, this doesn't generate much confidence—the relationships aren't strong, and the similarities are not assessed in a rigorous way, but rather by visual inspection, and when they are looked by plotting one against the other (GOSIF, EVI) they don't have a great relationship. The authors would need to better support these statements before the narrative explanation of their results would be easily believable, and also be their own worst critics in terms of what else could be giving them this pattern.

Through L258, this all is argumentative, but I don't see it backed up by the actual results, which are not compelling, and when there are problems with the definition of ecotype that, even if the results were true, could give substantially different interpretations.

Referee #3 (Remarks to the Author):

The manuscript addresses a very important question with important implications for vegetation-climate predictions : why have some Amazonian drought datasets reported forest green-up, or reduced mortality), while others observe forest brown-down, or enhanced mortality?

The manuscript represents a very innovative approach (comprehensive functional biogeography) and hypothesis framework to discriminate potential effects of water table depth, soil fertility/forest type; and tree rooting depth as estimated by proxies of forest functional composition.

The manuscript includes very strong datasets: EVI and GOSIF datasets, together with mapping products and a comprehensive analysis of published and novel demographic data from permanent forest plots over the three major recent Amazonian droughts.

And the manuscript reports potentially groundbreaking results; notably, a strong positive correlation between the extent of 'brown-down' and the depth of the water table as inferred by HAND, particularly along a stretch of south-central to south-west Amazonia. Although all regions did not react to drought events in a similar way, tree height and forest soil fertility appear to explain differences in EVI anomalies both within and across regions. As a result, the proposed GAM model not only predicts the pattern of the 2015 response, it also establishes a foundation for predicting forest responses to future drought events based on location and specific forest properties.

Overall, I find this to be a solid contribution that will promote valuable discussions and advance the field. However, I would like to see a bit more fine tuning of the presentation, including the consideration of a few additional analyses.

I agree that there are clear regional differences. Nevertheless, I struggled to be convinced that the evidence presented can be interpreted in the functional biogeography framework suggested by the authors, and that the contribute demonstrates a 'common basin-wide functional biogeography of forest drought response'.

A first major comment for improvement regards the integration of some metrics of land use. If I understand correctly, the current approach assumes essentially equal risk of forest degradation across the studied areas, and we know this not to be the case. I would appreciate an integration of some metrics of land use and land use change that might in fact be correlated with the other metrics presented here. Indeed, the areas of south-central Amazon with deep WTD and shorter trees also correspond to an arc of deforestation and fire risk. A proper treatment of land use among the other factors considered here can only refine the GAM approach.

A second major comment for improvement regards the characterization of functional composition across these forests and the interpretations that are then drawn. Clearly, there are fewer data to address at this scale than we would like to have. However, I argue that we do need to clearly present the uncertainty around these data and limitations that can be drawn on any generalizable inferences, particularly when an objective is to achieve a 'common functional biogeography'. I would suggest some clarification on some of the datasets and some tempering of the presentations drawn. I believe that the uncertainties of existing data may actually render the GAM model a conservative test of what we may be able to achieve with this approach. I would just like to see this contribution establish a path forward to improve datasets and refine our interpretations.

A first comment regards the definition of regions. I am not completely convinced by the number of regions nor by the justification for their discrimination. The first axis in the PCA (Fig S3a) is pulled by the correlations between the fertility maps (derived from species composition data of Zuquim et al.) and some extremely coarse estimates of community-weighted means of wood density and proportion of Fabaceae (based on some very coarse genus-level attributions of forestry data reported in terSteege et al.). Tree height does not appear to map onto any axis and is not strongly correlated with other mapped indices (Table S2; I would have appreciated as well to have axis loadings and eigenvalues in this table), even though later the Guiana region is defined as having taller trees (L93). Essentially, the proposal appears to establish regions based on climate and soil fertility (as inferred primarily from plant distributions). I can accept that, especially since the GAM refines this along a more robust continuous gradient; but I would like to understand if the Southern Amazon defined here might be nuanced into two or three sub-regions that behave differently. Was this approach attempted. Was an analysis such as PermAnova applied to justify the three region discrimination?

A second, related comment regards the definition of functional strategies, especially relative to drought avoidance/tolerance traits. The tree height Lidar dataset represents the best we have for the region, but is tree rooting depth significantly correlated with tree height for tropical trees of this size? Data from cited studies show correlations between shoot size and rooting depth overall, but across the narrow range captured by most of the Lidar dataset, this may not be applicable. I would appreciate more careful discussion that calls for better metrics of drought tolerance/avoidance than the proxies used here.

MINOR ISSUES

L103 remove qualifier 'rigorously'

Fig4 typo Partial EVI prediction

FigS3, typos/formatting of minimum and Fabaceae

Referee #4 (Remarks to the Author):

The manuscript addresses one of the majoring ongoing debates in tropical biology - how does Amazonia respond to drought events and what are the implications of this for the global carbon cycle. Contrasting results/inferences ranging rom green-up to widespread mortality have been published in high impact outlets. Resolving this issue is immensely important to a wide range of disciplines and to humanity.

The "answer" provided by the analyses in this manuscript is, like many things in ecology, it depends on the context. Specifically, there is an important interaction between water table depth, drought duration, soil fertility and the functional ecology of trees.

The team primarily uses a diverse assemblage of remotely sensed data products. Several important and nuanced conclusions are reached. For example, the previously debated "green-up" or "brown-down" of the Amazon during 2005 is linked to intra-regional differences in water table depth such that it wasn't uniformly a green-up or brown-down. Rather, water table depth is of great significance and spatially variable. This is logical, but not well-documented and greatly helps resolve the debate on how Amazonia will respond to future droughts. Similarly, comparing differences in responses to the 2005 vs 2010 droughts is valuable in the context of light and water availability (i.e. PAR, VPD) and drought duration. Lastly, while I am sure it will cause further debate and consternation, there is compelling evidence that the field collected data by RAINFOR occurred was spatially biased (i.e. occur in regions with deeper water tables and therefore more likely to brown-down).

In sum, this appears to be compelling evidence that both the green-up and brown-down perspectives were flawed and that both are occurring and (logically) dependent upon environmental context. I should note that I am not an expert in all of the remote sensing data products used in this work, but it appears robust and could be easily reproduced. The functional biogeography of the work is somewhat weaker, from my perspective. I do not doubt the gradients in mean height and wood density. These trends are found in the datasets used and in other studies. Similarly, soil fertility (%P in particular) have strong gradients in this region. That said, using tree height as a proxy for rooting depth is quite rough. Indeed, if you plot root depth vs height on log-log axes you get a strong relationship, but the magnitude of the scatter around the trend line is probably ecologically relevant. That is, two trees 20m tall that differ in their rooting depth by 5m is likely very important in this system. I do not think there is much the authors can do about an issue like this given the data available, but I don't think the height result is as obviously robust as the authors present. The mean wood density trend is robust, but vegetation is far more than just a mean. There is, of course, a lot of variation around that mean. It is unclear whether there is an obviously an east-west trend in variation in wood density in this region (e.g. Swenson et al. 2012 GEB). However, it would be, obviously, incorrect to lump all vegetation in one part of the soil fertility gradient as having a uniformly "fast" or "acquisitive" functional strategy while at the other end they are uniformly slow or conservative. All forests in the region studied likely have species with wood densities ranging from at least 0.4 to 0.8 (i.e. the vast majority of the global range is found locally). What changes is the relative abundances of values across space and some very light and very density wooded species dropping in or out. In sum, the "functional biogeography" here is a rough caricature of region and I would not consider the functional mechanism and biogeography to be pinned down by this work. If a revision were invited, I would encourage the authors to be a bit more clear with respect to how topical/overly-simplified the functional aspects of this investigation are and that within sub-region functional diversity is likely important. Beyond the overselling and simplification of the functional aspects of this work, I find it to be compelling and a quite important step forward towards out understanding of how this critical ecosystem responds to more frequent and devastating drought events in the region that have profound global consequences.

Referee #5 (Remarks to the Author):

Key results:
- This paper demonstrates that the spatio-temporal patterns of forest drought response in Amazon basin can be explained by a combination/interplay of three different groups of factors - hydrologic environments, edaphic environments, and tree drought resistance traits, rather than any of them taken separately. The paper clarifies a topic that has been debated in the literature.

- I find the Lines 152-166 especially interesting, as paper succeeded to reconcile virtually disparate data about effects of the 2005 drought.

Validity:

[All below mentioned flaws are not enough to prohibit publication, however:]

- The first sentence of the article, which is an important and motivating sentence, states: "Three 'once in a century' droughts (Fig. S1) occurred in the Amazon over a single decade..." However, the Figure S1 shows that, although the drought formally occurred "in the Amazon," it drastically occurred in different places within the Amazon each time. Given that individual trees or smaller groups of trees are experiencing drought effects at their actual locations, it appears that the locally-sensed drought was not as frequent as implied by the first sentence. The sentence may be unnecessarily dramatic and sensationalistic. Overlaying the three panels from Figure S1 clearly shows that drastic negative water anomalies were experienced only once in that decade by every individual tree in the Amazon. The authors later hinted at this in lines 177-179.

- It is not typical to start a manuscript by referencing a figure from the supplementary material, yet the first sentence starts with a reference to Figure S1.

- No link to the github repository with the codes used in the analysis are provided. The methods section contains clear description of the approach at algorithmic level, and ENVI software package (in Python), yet the readers do not know whether the entire development (e.g. PCA) was done in Matlab, R, Python, or something else.

- Table S2 shows the "Correlation among climate and ecotype variables." I assume that the measure of correlation used here is Pearson's correlation coefficient. However, it is not the only measure available, and the name of the measure used should be specified. Additionally, the correlations between environmental variables are often non-linear, so the findings based on Pearson's correlation coefficient might be misleading.

- In Figure S10, the correlation between EVI and SIF is shown to have an R2 of 0.23. It is well established that the average correlation between GPP and GOSIF is close to 0.80, and the correlation between EVI and GPP has been found to vary across biomes but typically lies around 0.5. The value of 0.23 appears to be too low. One possible explanation could be that the drought's impact on activity is greater than its impact on capacity, disrupting the typical relationship. However, I have concerns beyond this. If the value of 0.23 is the mean correlation across all three droughts, would it be possible to provide individual R2 values for each drought, given that the figure presents individual data points for each drought? Additionally, could you calculate the same relationship for a reference period outside of the drought? This is important because the GOSIF data was produced using the Cubist model, which is a single model for the entire Earth and has a cost function based on global mismatch. This non-biome-specific model may sacrifice accuracy for Amazonian biomes to more accurately predict SIF for other biomes. If the relationship outside of the study period is much better, it is likely that the low R2 during the drought is a result of the drought. However, if it remains low, the issues may run deeper. This should have been discussed in the paper.

Originality and significance:

- The conclusions drawn in the paper are indeed original, but they are also specific to the Amazon ecosystem and may not be easily transferable to other forest biomes. As a result, the paper may be more suitable for publication in a specialized journal. The idea of combining multiple factors to explain patterns in forest drought response, instead of a single one, is not necessarily innovative enough for publication in a top-tier journal like Nature. However, the quality of presentation, the flow and the style keep readers involved. The paper is entertaining and interesting for the wider audience.

- The paper lacks immediate practical significance, as the main contribution relates to processes that are beyond easy and direct human control and are happening over log time periods. Although the article lists two potential outcomes - conservation decision-making and improved predictions for future climate change, it does not provide concrete details on how these outcomes will be achieved or who will implement them. In terms of immediate academic interest, the paper is well-written and meets the standards of top journals in the field.

Data & methodology:

- The quality of presentation is very high, the techniques are well-established and mature. The quality of data is not questionable, except the concern I expressed earlier around low R2 in EVI/GOSIF. The study offers enough details to allow reproduction.

- The authors have put in a significant amount of effort to gather, standardize, and analyze the various data used in the analysis. The techniques used, such as machine learning, statistics, and geostatistics (variography), showcase their high level of expertise in data processing and analysis.

Appropriate use of statistics and treatment of uncertainties:

- Principal component analysis (PCA) used for clustering of the Amazon regions and Generalized additive models (GAM) used for the main analysis are well-established techniques, appropriate for this type of research.

Conclusions:

- I find the conclusions and data interpretation are robust, valid and reliable.

Suggested improvements:

- The title of the article could benefit from simplification to increase its readability.

- Consider exploring more advanced modeling techniques, such as NODE-GAM or neural networks, which have been shown to outperform GAMs. Utilizing turn-key solutions, such as Datarobot, may simplify the application of different modeling strategies. (e.g. https://medium.com/@chkchang21/interpretable-deep-learning-models-for-tabular-data-neural-gams-500c6ecc0122; https://www.datarobot.com).

- Consistent labeling of subpanels in plots using letters (e.g. a), b), c), etc.) is necessary for clarity, but this was missing in Figure S10.

- The title is the Figure S10 is:" Comparison of GOSIF to EVI drought anomalies and to ground-based demography during the 2015/2016 drought". Yet, the plot on the panel A (I guess) shows data for years 2005, 2010 and 2015/2016. The title of overall Figure S10 needs to be aligned with the plot and legend, as there is a discrepancy between the caption and the data shown in subpanel A.

- The statement in Line 33 that "Here, we combine remotely-sensed photosynthetic vegetation indices" is misleading as only one index (EVI) was used, and GOSIF is not considered a vegetation index in the literature (although, functionally, it is).

- In Line 239 authors say: "First, the broad consistency between MODIS and GOSIF remote sensing analyses (Fig. 1 and 3A versus fig. S2; fig. S7; fig. S9 versus S10)…". However, Figure S9 does not show any data related to GOSIF. The claim of "broad consistency" between MODIS and GOSIF data is not supported by the R2 value of 0.23 between GOSIF and EVI anomalies (Fig S10).

- The Method section is overly lengthy with over 4500 words and could be significantly reduced to meet the recommended size for publication in Nature (3000 words maximum).

References: The manuscript references previous literature appropriately. However, top tier journals are mainly referenced in the beginning, to set up the context of the study, while other sections mainly referenced other, more specialized journals. It points to the fact that either the work presented in the paper is original up to the point that nothing similar has ever been published in the top tier journals, or that topic-wise it belongs to more specialized journals.

Clarity and context:
- Abstract is clear, accessible and contains concise and relevant information and appropriate references.
- Introduction and conclusions are appropriate, at least within the research framework and choice of methods that have been applied to the problem.

Inflammatory material: Does the manuscript contain any language that is inappropriate or potentially libelous?

- No.

Author Rebuttals to Initial Comments:

Referees' comments:

> *We thank all the referees for an abundance of constructively critical comments. See our point-by-point responses below, in italicized blue text, offset from referee comments. We have highlighted referee text in gray or in yellow that we see as the main substantive points needing response. Underlined text highlights the specific changes made in the manuscript to address the referee comment being discussed. Copies of the manuscript are provided: a clean copy in pdf and a word version showing additions and deletions (note that line number references below refer to the clean copy). Citations below are the same numbers that are in the manuscript.*

Referee #1 (Remarks to the Author):

Comments

In this work, the authors combine valuable complementary sources of data – remotely sensed photosynthetic indices, ground-based demographic data, climate, soil, and topographic information – with the aim to better understand the processes underlying the spatially heterogeneous nature of Amazonian forest drought responses. The overarching goal of the study is to use a mix of community features (e.g. tree height) and abiotic factors (topography, climate, soil fertility) to predict the photosynthetic indices during three major droughts in the past two decades, to then define a biogeography of Amazon drought resilience and vulnerability. A key result of the study is the improved explanation of drought response variation by accounting for (a proxy of) water table depth in southern Amazonia, highlighting regions of deeper water table depths as being the ones corresponding to the increased browning and tree mortality.

Similarly, highlighting the interaction between water table depth and the duration of drought events to understand forest vulnerability spatio-temporal variations importantly draws attention to the urgent necessity to measure and better account for hydrological and edaphic/topographic data to improve the critical forecasts of future paths of carbon dynamics in the Amazon under alternative climate change scenarios.
The study is novel, well written, its hypotheses clearly stated, and its objectives are of utmost importance given the urgency of getting a better mechanistic understanding of Amazonian forest responses to extreme weather events such as droughts.

> ***RESPONSE R1.1****: We thank the reviewer for the comprehensive summary of the manuscript and its importance.*

I have a series of minor suggestions and questions, listed below, as well as a small number of bigger concerns – all of which addressable, in my opinion.
These bigger points relate to 1) the predictive statistical approach used in combination with causal inference objectives and language, 2) the definition of resilience and how it is assumed to be related to the remotely-sensed photosynthetic vegetation indices, and 3) the absence of soil texture in the study, in spite of its potential relevance during droughts for soil water retention capacity.

> ***RESPONSE R1.2:*** *We appreciate the reviewers constructive suggestions, both minor and bigger concerns. We hope the reviewer agrees we have used them to significantly improve the manuscript. To summarize our response to the "bigger points" (more detail of which can be seen below):* ***(1)*** *we have adopted your suggestions on the use of causal statistical modeling via Directed Acyclic Graphs (see R1.4, R1.5);* ***(2)*** *we have now included an explicit definition and rationale for "resilience" (see R1.6, R1.7);* ***(3)*** *we have incorporated effects of soil texture into the analysis (see R1.8).*

Note that line numbers only appear every 5 lines. I will therefore refer to line numbers as the closest previous line number followed by a dot and the line number from there (e.g. 42.3 for the third line of the 42th line chunk).

> ***RESPONSE R1.3:*** *Yes, every 5th line is numbered, so line 43 would be three lines after the numbered line 40 (which we interpret the reviewer to be calling line 40.3)*

Major points

I. Data analyses and causal inferences

I have two points to raise regarding the thinking behind the statistical approach of the study, as well as two suggestions.

1) Predictive versus causal approaches:
While having PAR, VPD, P and MCWD together in the same model is not a problem for the purpose of statistical prediction, it may hinder/obscure interpretations when it comes to drawing causal inferences from the fitted model coefficients (e.g. Pearl et al. 2016, Rohrer 2018, McElreath 2020, Arif and MacNeil 2022). Here the authors include all predictors in a single model (as well as all pairwise interactions between them) irrespectively of the effects these might have the ones on the others (or at least the authors do not address their consideration of this point). For example, I would assume that PAR influences temperature, which influences

VPD, hence creating a causal relation between PAR and VPD. Similarly, P partly causes MCWD.
Having a causal model (e.g. a directed acyclic graph, DAG) of the assumed causal relations among explanatory variables and with the response variable is a necessary condition to be able to draw causal inferences from the coefficients of a fitted statistical model (McElreath 2020, Addicott et al. 2022, Arif and McNeill 2022). Without such causal model to accompany a statistical model, interpreting all predictors' coefficients as if their magnitude were the reflect of their ecological importance for the response makes the (hidden) assumption that all predictors are independent of one another – which is rarely the case, and probably not here – and leads to mixing total causal effects, partial causal effects, and potentially confounds (i.e. non-causal paths relating an explanatory variable to the response of interest). Simple rules such as the "back-door criterion" used in light of a causal diagram including all important variables are very helpful to guide the careful selection of the variables to include/exclude to correctly assess the causal effect of a given predictor (Addicott et al. 2022, Arif and MacNeil 2002).

> ***RESPONSE R1.4:*** *We thank the reviewer for the helpful suggestions about including Directed Acyclic Graphs (DAG). We have now incorporated a DAG analysis in our modeling (see new methods section 2.6 and detailed Response **R1.5**, below), with the overall result that the analysis has increased confidence in the strength of the novel conclusions that can be inferred from the predictive GAM modeling. (note that the main conclusions were not much altered, but they are now more solidly grounded)*

2) Model selection: My second "statistical point" is regarding the use of a "dredging"' model selection approach, consisting of exploring all possible combination of the predictors (and their interactions) to then select the few ones with the lowest AIC and R2. If this approach ensures the best statistical prediction, it does not ensure a correct causal model, and can lead to wrong causal interpretations (Addicott et al. 2022, McElreath 2020), hence potentially confusing the objective to understand the mechanisms underlying the spatial heterogeneity in drought responses of the Amazon.

While using a model selection approach aimed at prediction is not a problem, it is very difficult to do so without confounding "best" resulting models, once selected, with "models providing the best causal understanding of the data". Note also that replacing "causal" by "association" does not really solve things, as two variables can for example be very associated because they have a common cause (but manipulating one would never change the other), or because of a collider bias (see Rohrer 2018 for an example). It quickly becomes difficult to interpret the results of a predictive model while totally avoiding implications of causality, in the absence of a causal model explicitly laying out the causal assumptions underlying the relations among the variables of the system.

I appreciate that the above is a common practice, but it is also a risky – or wrong – one when combined with a causal interpretation language as is often the case, and is the case here. I would appreciate it if the authors could take the above into account, one way or the other, and if they decide to keep their current analytical approach, to then justify how or why they consider they can do so, in light of the above considerations.

Suggestion to address this while keeping the changes to the current analytical framework to a minimum:

a) As explained above, good predictive models are not necessarily good causal models (Addicott et al. 2022, Arif and MacNeill 2022), and while I can imagine that "forest ecotypes" may indeed be important – it would make sense to expect so, to some extent –, drawing causal inferences, both qualitatively and quantitatively, requires making explicit ecological assumptions regarding causal relations among the variables of the system (for example, using directed acyclic graphs, DAGs), to guide a careful choice of variable to include or not to include together in a model, for inference about a covariate of interest (see backdoor criterion; Arif and MacNeill 2022).
If the aim is to draw causal inference, as is suggested by the objectives and the language used in the results regarding ecotypes (e.g. lines 215 and below), I recommend generating a DAG, or a small set of sound alternative DAGS if necessary, and using the backdoor criterion (and if necessary frontdoor criterion) to justify inclusion/exclusion of control variables for the models aiming to assess the causal effect of each covariate (HAND, tree height, etc). This would allow testing and quantifying the effects of each of these in terms of total causal effects (hence also allowing comparing their importance). Different models including different control variables may be necessary, depending on the causal and non-causal paths highlighted by the backdoor criterion, to capture the causal effects of different explanatory variables.
Note also that the functional form of the relations among variables is not included in a DAG, which therefore would not hinder the use of GAMs. In practice, little change to the current analyses; only an explicit causal justification for the inclusion/exclusion of covariates in each model, to ensure causal interpretability of fitted coefficients.

> ***RESPONSE R1.5(a):*** *We appreciate the reviewer's critique of predictive modeling, and agree with the recommendation that additional, complementary analysis (as provided by a DAG approach) could increase confidence in the inferences that associations between drought response and environmental factors are likely causal. We have therefore implemented a DAG analysis to test the assumptions that associations are causal (see methods section 2.6: "Causal approach: Directed Acyclic Graphs)", and integrated it with GAM for prediction (see main text, line 124; methods section on GAM (section 2.7).*

b) The predictive approach – AIC-based model selection following "dredging" – used here could be maintained for the purpose of prediction only, and to compare the predictive power of a model considering these variables and the concept of ecotype (regardless of their exact causal relations / creation of confounds) with a model only considering climate.
In this sense, the predictive model would not especially serve the interpretation of the detail inside the model, but would show the added predictive value of HAND, tree height etc, while the models structured based on a causal model (e.g. a DAG) would allow drawing causal inferences in terms of mechanisms (line 215.5).

> ***RESPONSE R1.5(b):*** *Yes, we agree--the joint DAG-GAM analysis maintains predictions as before but also gives greater confidence to the inferred mechanisms. With respect to the predictive (GAM) analysis, we also agree that comparing a model with ecotypes (now "ecotopes", see R1.13, below) to a climate-only model illustrates the value of ecotypes in predicting drought response (Table S1 and revised text, line 245:*
>
> > ***"the models which includedforest ecotopes ... along with climate had significantly more predictive power (higher R2 while selected by lower AIC) than climate-only models (table S1)."***

While this may look like much, I believe that the above could relatively easily be added to the present analytical framework and would provide a correct and more justifiable/robust basis for the interpretation and implications of the findings – that is, a crucial point given the importance of the objectives of the study.

> ***RESPONSE R1.5(c)****: We agree that the recommended addition of DAG modeling increases confidence in the GAM-based inference of causes, and in the strength of the paper's novel findings and conclusions. We thank the reviewer for suggesting it.*

References

Arif and MacNeil 2022, Ecological Monographs: https://doi.org/10.1002/ecm.1554
Arif and MacNeil 2022, Ecology Letters: https://doi.org/10.1111/ele.14033
Addicott et al. 2022, Frontiers in Ecology and the Environment: https://doi.org/10.1002/fee.2530
McElreath 2020. Statistical Rethinking, Second Edition.
Rohrer 2018, Advances in Methods and Practices in Psychological Sciences: https://doi.org/10.1177/2515245917745629

II. Definition of resilience

1) The concept of resilience has a central role in the study, yet it is only defined in brackets in a sentence centred on something different (line 110.3). Moreover, there are many definitions of resilience out there (e.g. Grafton et al. 2019, Nikiinmaa et al. 2020). Resilience needs to be clearly conceptually defined and explicitly related to the measurements taken as proxies of resilience in the paper (with justifications). I suggest starting the paragraph where the concept appears for the first time (close to line 110) with this definition and links to the data of the study.

> ***RESPONSE R1.6:*** *We agree with the reviewer that a resilience definition and rationale was missing, and is necessary. We have now included this definition and rationale (see methods section 2.4, renamed Drought Resilience and Vegetation anomalies, line 881):*
>
> > *"We defined drought resilience as a forest's ability to increase (or relatively better maintain) photosynthetic capacity or activity during a perturbation -- that is, by its tendency to exhibit more positive/less negative anomalies in vegetation indices (relative green-up) during drought... We chose relative green-up here for conceptual and practical reasons. Conceptually, greater relative green-up implies relatively more photosynthesis and hence, all else equal, more carbon resources to respond to stress... making it a logical general metric of resilience. Practically, greening has been widely cited and discussed in the literature, and, notably, is predictive of outcomes on the ground commonly associated with resilience at the individual tree and plot scale (lower mortality, greater growth, and greater xylem embolism resistance, see methods 1.6, figs. S17 and S18)."*
>
> *Notably, with Tavares et al 2023, we are now able to meet the reviewer request to relate our remote sensing derived resilience to independent ground proxies of specific mechanisms of resilience. In particular, we find that our remote sensing resilience significantly predicts an independent and common measure of resilience, the hydraulic safety margin (HSM) in tree xylem tissue, taken on individual trees and scaled to the plot level (see new main text at 275, methods section 1.6 text at line 741, both citing the Fig.*

5A and S18). We hope the reviewer finds it compelling, as we do, that this independent dataset (not published until 6 months after our initial submission), well validates our metric of resilience. We understand that different definitions of resilience are possible, and even if we disagree with the reviewer on the best definition conceptually, we hope that the reviewer is persuaded that our rationale (and its success in predicting Tavares et al data) makes "green up" a practical, meaningful, and compelling definition.

2) Following up on the previous point: Why would an increase in photosynthesis/greening correspond to a higher photosynthetic resilience? If resilience defines the capacity of a system to return to its stable or equilibrium state following a disturbance, then how does an increase in photosynthesis *during* a disturbance indicate resilience? It surely is a notable response, but not one that I would think to relate to the concept of resilience at first. In addition, would a forest that would not change its photosynthesis rate during the drought (i.e. no greening, but no browndown either) not be as resilient as a forest increasing photosynthesis (if resilience were related to photosynthesis activity during the stress)? The former forest would simply maintain its functions without showing either greening or excess mortality. Would resilience not be as much about recovery *following* a drought (resilience sensu stricto) as it is about photosynthesis response *during* the drought ("resistance" dimension of resilience sensu latto)? It appears that the authors may be referring mostly to the *resistance* property of the multidimensional notion of resilience (e.g. Grafton et al. 2019, Nikiinmaa et al. 2020). Is that correct? If so, properly defining this upon first use in the main text (and in more details in the Methods) would help conceptual clarity, by defining a more precise framework to formulate and think of the questions and results.

RESPONSE R1.7: *We see our resilience definition (in methods 2.4, line 881, quoted above in R1.6) as complementary to definitions of resilience cited by the reviewer. E.g., trees that can maintain greater carbon resources (relative to trees that cannot maintain as much photosynthesis) should have better capacity for both tolerating drought while it happens and for recovery following drought (one of the alternative key metrics of resilience).*

References
Grafton et al. 2019, Nature Sustainability: https://doi.org/10.1038/s41893-019-0376-1
Nikiinmaa et al. 2020, Current Forestry Reports: https://doi.org/10.1007/s40725-020-00110-x.

III. Soil texture

The authors aim to address key drivers of forest sensitivity and resilience to drought, and focus their efforts on a proxy of water table depth, a proxy of peak dry season water deficit, and soil fertility, to go beyond climate data alone.
I would expect a key missing ingredient determining soil hydrological stress to be soil texture (and ideally structure, e.g. approximated by bulk density). Droughts can be attenuated or amplified depending on the clayey/sandy nature of the soil, which relates to its useful water retention capacity; not only the water table depth.
Could this be included in the analyses, even if only for a subset of the whole dataset? This should at least be acknowledged and discussed in relation to the results. How much soil texture variation is there within the three major biogeographic regions defined by the authors, and how is this expected to affect the biogeography of drought as perceived through the current

analyses? Would interaction terms among soil texture, HAND, and/or MCWD not be expected to be necessary to obtain a more complete picture of the spatial distribution of drought responses?

***RESPONSE R1.8:** Yes, we agree with the reviewer about the potential importance of soil texture, and we should not have excluded it. <u>Soil texture is now included in the modeling (See main text around line 191, method section 2.7 (ii), fig. S10F, 11)</u>; we find an important soil texture effect: more clayey soils, which bind water more tightly (making it harder for roots to access), are associated with forests that are more vulnerable to drought (<u>fig. S10F)</u>, as suggested by process model studies (Levine et al 2016; Longo et al. 2018). But again, this depended on water table depth, and deep water table forests were more vulnerable at low clay/high sand relative to intermediate levels (fig. S11), perhaps because in the absence of a shallow water resource, sandy soils drained water too quickly. There are also intriguing interactions with water table depth and tree height.*

We note that our original GAM analysis would have selected soil texture, but for the moderate positive correlation (+0.38, Table S2) between forest height and clay fraction, leading to exclusion of models with both forest height and soil texture; AIC favored forest-height as a better predictor, so among the two, forest height was in the end selected. <u>However, the new DAG analysis implies a causal link for both forest height (fig. S10D) and soil texture (fig. S10F)</u>. In addition, <u>we now include consideration of Variable Inflation Factors (VIF) in building the models (Table S2, top row)</u>, and these are modest, even for tree height and soil texture, suggesting that our previous criterion for variable exclusion was probably too strict. Without the restriction, AIC selection clearly favored including both forest height and soil texture in basin-wide GAM predictions for 2015-2016 drought. Thus considering evidence that both variables are independently causal despite moderate correlation (DAG) and that both are important to prediction (GAM AIC selection), we include both in our final basin-wide model (table S1D).

We note that <u>inclusion of soil texture in the main model (Table S1D) was propagated into everything that the basin-wide model does, which meant regraphing Fig. 4 and 5 and associated figures in the supplement</u>. The difference between graphs with and without soil texture (in our original submission vs in the current resubmission) are detectable if one looks closely, soil texture did not change any conclusion of the previous analysis, but the model is now better and the interpretation enriched (see Fig. S11).

Minor comments

- First sentence of the introduction: I'd remove the parentheses for the dates. *<u>DONE</u>*

- The beginning of Fig. 1 legend could be shortened quite a bit: instead of repeating three times what maps 1,2,3 represent, only changing the year they correspond to, I would suggest writing this only once (like for the insets). Something like: "Standardized anomalies of drought Enhanced Vegetation Index (EVI, a proxy of photosynthetic capacity) in drought-affected pixels (see Fig. S1) for the droughts of 2005 (A), 2010 (B), and 2015 (C), with ellipses highlighting green-up and browndown patterns that correspond to shallow and deep water tables (see Fig. 2A)." And no need to explain what a standardised anomaly is in the legend I'd say.

***Minor Response R1.9:** We have shortened Fig 1. caption accordingly.*

- Still in Fig. 1 legend: If I am understanding correctly, the delta EVI are statistically estimated, is that right? If so, it would be useful to provide some measure of uncertainty around those estimates (e.g. confidence intervals). Note also that the closing parenthesis for the delta EVI of 2005 is missing.

> ***Minor RESPONSE R1.10****: Delta-EVI is simply an observation (at the per-pixel level), quantified as a deviation (or anomaly) relative to a long-term mean within that pixel (see derivation, method section 2.4 at lines 895-908). We appreciate the reviewers request for measures of uncertainty, and we hope our approach addresses that request: E.g When per-pixel observations of delta-EVI are aggregated to mean values (e.g. in bins of many pixels that fit within a water-table depth interval) uncertainty is represented as standard error of the mean (e.g. Fig. 2B). When Delta-EVI is predicted by a statistical model, predictions are shown with 95% confidence intervals (e.g. Fig. 3A).*

- Line 50.4: "a strong decrease" of what? Greenness I suppose, but please make it explicit. ***Minor RESPONSE R1.11*** *Yes, done (see line 53).*

- Line 60.1: "exhibiting green-up, or reduced mortality": Are we to understand that these are one and the same, or two alternatives corresponding to resilient scenarios? I am asking as it is stated further up that the 2005 drought saw simultaneous greening up and increased mortality from ground-based data, which seems at odd with the content of the parenthesis here.

> ***Minor RESPONSE R1.12****: Yes, since this manuscript indeed shows that green-up is associated with reduced mortality (Fig. 3A vs 3D), we hope that readers understand that this association supports remotely sensed green-up as an index of resilience that is also associated with reduced tree mortality (Fig. 3D).*
>
> *And yes, the previous literature reports that* ***both*** *"green-up" (Saleska et al 2007)[9] and "increased mortality" (Phillips et al 2009)[17] were observed during the 2005 drought. And yes, before one examines the divergent spatial patterns (which show that the excess green-up happened in different places than the excess mortality) these two observations do seem at odds. We argue that our new results explain the seeming contradiction of the previous results. As we now write (lines 160-169), our results*
>
> > *"reconcile the much-discussed apparent disagreement between remote sensing studies showing 2005 drought-associated green-up on average[9,10] ... and ground-based plot studies of the RAINFOR network showing 2005 drought-associated excess in tree mortality on average[17]... Our more fine-grained analysis suggests, however, that the excess greening and the excess mortality were not in the same places; it is the locales with shallow water table forests that were benefited by drought, while deep water table forests are vulnerable, ..."*

- Line 60.3: I would remove "well-" in "well-validated" (unnecessary), or provide a more specific definition than "well" about what is special about the indices. ***RESPONSE:*** *OK (see line 62).*

- Line 65.2: I suggest clearly defining "forest ecotype" at the end of the sentence line 65.2, as it is very unusual to use the term "ecotype" – typically reserved for *intraspecific* levels of *genetic* variation related to different habitats – to characterise *community*-level differences mostly related to *abiotic* factors.
Even better might be to use a different word altogether (e.g. "habitat", simply), to avoid

confusion using “ecotype” for a definition not corresponding to its common one in biology/ecology.

> ***Minor RESPONSE R1.13****: The reviewer makes a good point about potential confusion caused by our generalizing the meaning of “ecotype” beyond it’s different use in evolutionary ecology, so we have adopted the alternative term “ecotope”, which though less well-known has long been used in the ecosystem ecology literature to characterize ecosystem types. Introduced by Tansley (1939)[19], and later used by Whittaker (1973)[20], it encompasses precisely what we mean here, characterizing ecosystem types by both biotic and abiotic factors, including water tables, soil types, tree heights, drought resistance traits, etc. We now introduce the term in the main text (line 66), define it in more detail in the methods section 2.4), and use it throughout.*

- Line 75.3-4: I am not sure to follow the parenthesis explanation for the hypothesis. I would understand that fast turnover due to a dominance of fast growers would cause a strategic disadvantage to slow growers (with drought resistance traits) on fertile soils (if one forgets about soil texture and potential edaphic droughts, regardless of fertility). The authors could probably rephrase/clarify their justification for this hypothesis.

> ***RESPONSE R1.14****: Yes, the reviewer is correct (“I would understand... a strategic disadvantage to slow growers...)”. See text revisions (lines 77-82).*

- Line 95.2-5: This part lacks some details to be clear.
1) I am guessing the authors mean “Drought-affected regions *for each drought year* (...)”. Correct? If so, please clarify.
2) “The 20-year mean” is mentioned here for the first time: specify what are the reference years (e.g. “for a 20-year baseline based on XXXX-XXXX”).
3) Regarding drought duration: lines 95.3-5: Please clarify “as the *number of months* that their *monthly* CWD remained below the *20-year* mean MCWD”, if this is indeed how duration was calculated.

> ***Minor RESPONSE R1.15****: (1) yes, drought-affected regions for each drought year (now stated, line 95); (2) the reference years are now noted (at line 97); (3) drought duration is now clarified to be number of months (line 98).*

- Lines 100.4-5: Does the “across multiple droughts” mean that the data for the three droughts were used in a single analysis, or were combined a posterior from separate analysis on each drought? Please clarify what the input data was, in this sentence.

> ***Minor RESPONSE R1.16****: The text here has been clarified, including to indicate that data for the three droughts were used in a single analysis (line 131).*

- Line 110-115: see resilience point raised earlier, with respect to definition and justification of hypothesis.

> ***Minor RESPONSE R1.17****: Point taken (see RESPONSE R1.6, above).*

- Line 115.1: The GOSIF acronym was not introduced before. *Fixed (see line 64).*
- Line 115.3-4: I would rephrase slightly to replace positive “greening” anomalies and negative “browning” anomalies by positive vegetation index anomalies (i.e. “greening”) and negative anomalies (i.e. “browning”). To avoid confusion with a negative “browning” anomaly being less browning than the baseline, for example. *Agreed, fixed (line 118).*

- Line 120.1-3: The “because” in the sentence suggests a capacity to draw causal conclusions from the observation of higher solar radiation anomalies during periods of greening. This sentence warrants either more caution in the phrasing, replacing “apparently” by “potentially” and “which were particularly advantageous” by “which may have been ...”, or requires a stronger justification for the causal interpretation (probably the former; see also comment with respect to Fig. 2C).

> ***RESPONSE R1.18**: “apparently” is intended to convey the appropriate caution, with the new DAG & GAM analysis addressing the issue of more rigorous inference. See also the adjustment to the text in caption Fig. 2C (detailed in Response R1.20, below).*

- Fig. 2B: No legend provided for the empty symbols.

> ***RESPONSE R1.19**: FIXED by removing the empty symbols (they represented seasonally inundated varzea forests, at negative water table depths, and were otherwise excluded from the analysis, but showed a pattern of distinctly less greenup, compared to terra firme forests). As we don’t have space to discuss them, we now exclude them to avoid confusion in a paper that we think already has quite enough material already in it!*

- Fig. 2C: Comment directly related to the one of line 120 about solar radiation:

1) The legend mentions a figure of EVI anomaly *sensitivity* to PAR anomalies, but it is unclear whether Fig. 2C presents a model output, or plots the data simply. I am guessing it is the former, as I see error bars, but the error bars are not mentioned in the legend, and it would be helpful to specify the model this result came from (even if briefly).

2) The legend does not say whether the upper panel of Fig. 2C is for the relation from all drought severities or one in particular. Is this a model fit across all drought severities?

3) Solar radiation has a very high value at 1.5 of PAR anomaly for severe droughts, that has nothing even close to it for the modest and medium droughts. How robust is the association between EVI anomalies and PAR without this extreme value, in the model? Does it remain after removing it, or is it entirely driven by this value?

4) Regarding the causal inference implied by “sensitivity of EVI to PAR”: Unless justified by making the underlying causal assumptions explicit somewhere (i.e. having a causal model or diagram underlying the statistical model), implying a cause from an association is delicate. Solar radiation also causes temperature, which causes VPD, both of which also potential drivers of a greening/browning, for example. See more about this in the main points raised further above.

> ***RESPONSE R1.20 , Fig 2C**: points 1-4: **(1)** To clarify, Figures 1 & 2 consist of simple observations of EVI anomalies, with no modeling (other than that done by NASA to create the publicly available EVI product, and the simple mean removal and residual standardization done by us to create the EVI anomalies, as described in methods section 2.4; Statistical modeling is only introduced beginning with Fig. 3). Error bars are standard errors of the mean, from bin-averaging (by both water table depth and drought severity, in panel B, and by both water table depth and PAR anomaly, in panel C), of the observed anomalies that are plotted per pixel in Fig. 1.*
>
> *We have moved the introduction of statistical modeling to later in the text (to lines 123-130), after the presentation of Fig. 2, to avoid reader misperception that statistical*

modeling was used to create Fig. 2. We also revised the Fig. 2 caption (now refering to "Observed EVI anomalies", and hope these changes make the presentation clearer.

(2) We hope it is now clearer from above (and the revised caption wording) that the bin averaging in Fig. 2C (with no statistical modeling beyond that) is across all pixels in Fig. 1A (i.e. across all severities).

(3) Yes, the most severe drought has a high density of observations (taller bar representing more area) at relatively high PAR (this is mechanistic: more extreme drought is more extreme usually because fewer clouds lead to both less rain and more sun). As mentioned above, there is no model here in Fig. 2, beyond bin averages of EVI anomaly observations, so the association between EVI anomaly and PAR is not here statistically quantified, it is just an apparent pattern. If the extreme drought pixels were removed, the number of points available to be averaged at high PAR would become less, and so the error bars on the mean estimates would become larger, but (since there are still high PAR pixels present even in the modest and medium drought regions) the apparent pattern would likely not differ beyond the variability defined by that larger error. We hope this is now more clear from the revised Fig. 2 caption.

(4) We agree, and have replaced "EVI anomaly sensitivity to PAR" by "Observed EVI anomalies bin-averaged by PAR anomalies" in the caption. These more nuanced issues of inference are left to be better addressed by the combined GAM and DAG analysis.

- Line 124.4-125.5: Lots of information in a 7-line sentence. I suggest putting up front the main message, and dividing this in shorter sentences to help the reader.

RESPONSE: *Done (see lines 132-135).*

- Line 130: "less negative (less browning) anomalies" could simply be written "less browning", to ease the reading. (The same sort of thinking applies to many long sentences throughout the manuscript). Note also that since the structuration of anomalies with water table depth has already been presented at this stage, the whole "with less negative (...) over deep water tables" ending the paragraph is redundant and could simply be removed.

RESPONSE: *Done.*

- Line 180.5: The authors cite 31-33 to support that tree height is a proxy for rooting depth. Chitra-Tarak et al. 2021 (31) does not show this association however (unjustified citation, to be removed). Brum et al. 2019 (32) indeed show a relation between DBH and oxygen stable isotope ratio in the xylem used as proxy of rooting depth, based on 12 species in a seasonal Amazonian forest, while Tumber-Davila et al. 2019 (34) show a relation between shoot height and rooting depth in woody species, but mostly across biomes, with no clear information (that I could find) regarding such a relationship in the moist tropics (most likely a minor proportion of their dataset). On top of removing the citation of 31, I would advise acknowledging how little is still understood (or even available in terms of data) regarding rooting depth and its determinants in the (neo)tropics. A brief discussion regarding this source of uncertainty here would be a valuable addition to a perspectives paragraph. This could go together with the perspectives addressing soil texture (in case it could not be included in the analyses).

RESPONSE 1.21*: We removed the citation to Chitra-Tarak here (but kept it as a basis for hypothesis 3). We now say "forest canopy height (line 193), which we use as a proxy* *for rooting depth, consistent with observations (Christina etal 2011)[45], (Brum et al)[35] (Tumber-Dávila et al)[36]". We now also comment on the uncertainty of rooting depth in the methods in section 1.3, where the canopy height measurement are explained (see at line 646:*

> *"observations of the tree height-rooting depth allometry are limited, especially in tropical forests (although one study cited here[35] is directly relevant, as it is from central-eastern Amazon uplandforest, conducted during the 2015 drought); this limitation remains a key uncertainty in our ability to confidently attribute variations in drought response to rooting depth, as opposed to canopy height itself, or other (as yet unidentified) correlates of canopy height."*)

- Lines 209.5-210.1: See previous comment regarding the use of "ecotype" in the study. Note also that ecotype has been used in previous sections but is being defined here, in parenthesis. The concept of ecotype or whatever other term is used should be clearly defined upon first use.

***RESPONSE**: As noted above (R1.13), "ecot**ype**" has been replaced by "ecot**ope**".*

- Line 210.2: See conceptual comment regarding actual match between greening up and resilience, as well as regarding the definition of resilience.

- Lines 215.4-5: The attribution language and causal inference drawn in this paragraph relate to my concerns regarding the absence of a causal model to accompany the statistical models used. The authors use a predictive approach, selecting models based on an information criterion mostly, and then interpret the output in causal terms (causal language of mechanisms, sensitivity to X or Y, etc, throughout the results). See related point at beginning of the review.

***RESPONSE**: Yes, as noted above, the reviewer's comments about attribution and causal inference are well taken (see RESPONSE R1.5, above).*

- Lines 475-480: Consider having two sentences here. *Done (see lines 531-541).*

- Lines 525.3-530: Speaking of results only in terms of them being "statistically significant" or not is insufficient to highlight their importance or the relevance of the related signal. Some of the figures in Fig. S8 to Fig. S10 are apparently driven by a few extreme points, potentially leading to significant p-values given the relatively low number of observations, and there is a substantial and/or asymmetric variation around the regression lines in some of these figures. The R2s and these considerations should be acknowledged in the text to provide a more complete depiction of these relations.

***RESPONSE**: We have adjusted (now) fig. S17 to include consideration of the effects of potential outliers on significance and nature of the relationship (see details in R2.24(3))*

- Line 565.4: Ecotype variable: A key element that should be added at the beginning of this section is the definition of what the authors call "forest ecotypes", before diving into the different elements they considered to make up the variable. This is particularly necessary as the term "ecotype" is typically defined as concerning geographically distinct groups of individuals within a same species (i.e. populations) that have genotypically-defined adaptations to certain environments, but is not often used to characterise different communities irrespectively of their species composition. This would ensure a clear understanding of what the authors mean.

Similarly, given the structure of Nature papers, I would re-define "forest ecotypes" upon first use in the main text (outside the Methods section).

***RESPONSE**: See R1.13 discussion of changing terminology from ecotype to ecotope.*

Line 595: Does "community-averaged" mean a community-weighted mean? If so, please specify.

***RESPONSE**: Yes. It means a community-weighted mean. Fixed. (see line 656)*

- Lines 605 to 615.5: How do they authors expect the very different sample size for shallow and deep water table plots (i.e. 25 vs 321) to affect their results and conclusions, or the uncertainty around these?

***RESPONSE R1.22**: We don't expect different sample sizes in ground plots to affect our main results and conclusions, which are based on remote sensing analyses (not ground based forest plots). Remote sensing "sample size" (pixels across the basin) essentially captures directly the proportion of shallow versus deep ground waters, and is not related to the different sample sizes for ground plots over shallow versus deep water tables.*

*Indeed, one of our study's highlighted conclusions from (relatively unbiased) remote sensing is that ground plot observations **are** biased towards deep water table forests (see Fig. 3E, showing, bias of the 321 plots towards deep, not, as the reviewer implies, that they are all deep), and that this can skew conclusions from ground-based studies, depending on the question, if this bias is not accounted for (see lines 170-174).*

Our validation results (methods 1.6) depend on direct comparisons of remote sensing to observations across a range of ecosystem fluxes, demographic factors (growth, recruitment, mortality), or hydraulic resilience. Except if bias is so extreme as to remove a substantial part of the range being tested (which does not apply here), the quality of the validation should not be much influenced by the bias of the datasets.

We edited the section referred to by the reviewer (methods section 1.5, now starting at line 669) to make the purpose of the ground plot data more clear.

- Line 615.2: What were the demographic baselines used to define 'demographic anomalies'?

***RESPONSE**: The long-term mean was used as the baseline; this was stated in the caption to Fig. 3E, but has now also been clarified in the methods (e.g. line 674: "*forest mortality demographic anomalies (% deviation from long term mean)*").*

- Line 635: closing parenthesis to be removed. *Done (see line 807).*

- Line 640.4: Please remove "tree height" from the variables contributing the most to the first PCA axis. Fig S3A clearly shows a minimal contribution of this variable to the first axis. Checking the variable contributions would confirm this.

***RESPONSE**: Done (see line 818).*

- Line 640.5: You could name this second axis something more specific than "climate", since the two variables contributing to it are both related to water deficit, or water stress.

***RESPONSE**: Changed to "water".*

- Line 645: Fig. S3A does not show "three nearly uncorrelated axes" (as in PCA axes), since only two axes are displayed in that plane. What appears is correctly interpreted for PCA axes 1 and 2 (or will be, once "tree height" will have been removed from the important variables of axis 1), but soil fertility is a variable contributing mostly to the first PCA axis, and somewhat to the second too. In that sense, it would be okay to maintain that it captures some complementary information to that of [Fabaceae proportion + wood density] and [MCWD var. and minimum precip.], but it is incorrect to speak of uncorrelated axes. Or change terminology to speak of "three complementary dimensions of forests"?
I also suggest acknowledging that the "vegetation axis" is also a "soil fertility axis", negatively correlated to Fabaceae abundance and wood density.

***RESPONSE**: We have accepted the reviewers suggested revision (see line 817).*

- Line 750.1: Add reference for the R package mgcv. Done *(see line 989).*

- Lines 740.3-760.1: see two statistical points made at the beginning of the review, regarding predictive vs causal approaches, and model selection. *See our response to those points.*

- Line 830.2: specify if "concentration" refers to effective cationic exchange capacity. *Done (see line 1089).*

Line 970.5: Location of the code to reproduce the analyses and graphs is missing. *The code will indeed be curated and linked in time for publication.*

This is an impressive and important work, and I hope the authors will be able to address the above comments.

David Bauman

Referee #2 (Remarks to the Author):

A. Summary of the key results

This paper looks at the association between indices of forest vegetation responses and drought in response to landscape predictors, and incorporates demographic data from forest tree plots. The authors reveal a heterogenous response—termed a biogeography of forest drought responses—and suggest a mechanism for its occurrence based on regional variation in average forest canopy height ("tree height"), coarse measures of height above major drainages, taken to be measures of water table, and soil fertility. They use heterogeneity in responses to 2005, 2010, and 2015 droughts, the latter particularly strong in the N. Amazon, as tests of the mechanisms. The authors suggest that a predictive understanding heterogeneous responses to drought are important for multiple factors, from conservation to understanding resilience.

B. Originality and significance: if not novel, please include reference

The ideas and hypotheses generated bring together a large amount of disparate information to answer questions of broad relevance and interest. Each of the pieces has been discussed, but the inclusion in one effort to generate a biogeography of drought with a mechanistic interpretation is new.

> ***RESPONSE R2.1**: we appreciate the reviewer's recognition of novelty brought by a biogeography that integrates disparate factors into a single unified framework.*

C. Data & methodology: validity of approach, quality of data, quality of presentation

I have a number of problems with the approach, even though I am sympathetic to the exigencies of large-scale remotely sensed data, aggregated and interpolated structural and soils data, etc. and detail a number of those below. This ranges from the remotely sensed indices of drought response and their concordance, what the derived variables are really giving information about (which affects attribution of effects observed).

> ***RESPONSE R2.2**: we hope that the clarifications, revisions and responses below address the bulk of the reviewer's concerns. To summarize, changes include responses to these key multiply-emphasized reviewer themes (details in specific sections below):*

1. *We address a key misperception that many different ecosystem types were included by clarifying that the scope of our analysis is indeed limited sensu stricto, to terra firme forests only (see response **R2.3**, below).*
2. *Regarding concerns about "overselling" and the referee's recommendation to be our own toughest critics, we offer both additional support from statistical modeling (by adding a causal model, via Directed Acyclic Graphs, **R2.6**, point 4) as well as consideration of potential complications and alternatives (see **R2.6** & **R2.10** for a comprehensive summary of these, **R2.12** about different droughts and effects of drought dynamics, and **R2.13** on attribution of causes to inferred mechanisms).*
3. *Regarding concerns about EVI versus GOSIF, see **R2.7** and **R2.25.***
4. *Regarding concerns about our PCA-based three-region classification and "inaccurate" characterizations of S. Amazon as "high fertility", see **R2.15** (edits on soil fertility), **R2.20** (does Fig. 5 biogeography "easily square" with 3 regions?); **R2.27**(1) (de-emphasis of 3-region classification; independence of biogeography)*
5. *Regarding advice that composite or multi-variate analysis using different indices would be compelling, and their validation, see **R2.14** and **R2.25**.*
6. *Regarding validation relationships "poorly explaining" tree demography, see **R2.24** (including a more thorough diagnosis of validation regressions supporting their use).*

Though the title and the paper refers to Amazonian forests, it appears that the authors did not constrain the analyses to only forests, but rather included large areas of tropical savanna including Brazilian cerrado and edaphic savanna in Boliva, as well as deforested areas/agriculture, and high elevation Andean grasslands. Within forested areas, Andean forests (albeit within the Amazon drainage), higher elevation forests on Guianan teupis, and the large areas of degraded forests in the Amazon, are also included. This could lead to having greater than half of the inference space for the paper not being in what is traditionally thought of as Amazonian forests, with the accompanying influence on the GAM fitted models and predictor variable influence. If the authors did extensive masking, exclusion of non-forest habitat, etc., this should be stated in the paper, and these areas should be excluded from the figures.

RESPONSE R2.3*. We realize that we need to correct a clear misunderstanding here (the impression that we did not constrain the analyses to forests only), and thank the reviewer for pointing out that this important constraint was not sufficiently clear in the main text. To be clear, our analysis was conducted only after masking the dataset to remove flooded forests, deforested or burned forests, and non-forest vegetation (Andean highlands, cerrado, agriculture, etc). While we had stated the domain of analysis as terra firme forest pixels in the methods section (section 1.4, "Amazon Basin Land cover", around line 600 in the original manuscript), we realize from the reviewer's comment that it may not have been clear to readers who look only at the main text, or who look at the unmasked HAND map (Fig. 2A), or who didn't notice the missing/masked pixels in the forest drought response maps (Fig. 1). We hope the revised main text (addition of "focusing on intact terra firme forests", line 67), the revised title ("***Identification of terra firme Amazon basin forests using land cover maps"**) *and first line of methods 1.4, the greater detail in methods 2.1 clarifying how coarse resolution pixels containing both "mixed" forest and non-forest regions were "purified" to represent intact forest only (lines 660-668) and the revised figures (e.g. all maps – not just the vegetation anomaly maps -now mask out high elevation Andean ecosystems) make it unmistakably clear that the analysis indeed applies, sensu stricto, to forests only. We think that our clarification here will address many of the reviewer concerns that emerged from the misunderstanding that non-forest vegetation or deforested areas were included.*

The figures are well done, but reach a level of complexity in some cases that makes them opaque. It would be good to think about how to simplify the presentation in the more complex cases.

RESPONSE R2.4*: We thank the reviewer for concern about figure complexity, and would be happy to try to simplify presentation of any figure the reviewer finds opaque.*

D. Appropriate use of statistics and treatment of uncertainties

Following on point C, many of the statistics are appropriate, but there are areas for improvement noted in the detailed comments below, particularly in attribution of results and inference. More diagnostics of the GAM fit and partial plots would be important.

RESPONSE R2.5*. We thank the reviewer for these comments; please see our responses below under each of the sections.*

E. Conclusions: robustness, validity, reliability

In many cases the authors make very strong conclusions from weak relationships, or data open to multiple interpretations (many cases are noted below). The authors need to be their own harshest critics and present alternative hypotheses, and alternative interpretations. The authors spend a lot of prose talking about how the results fit their general model when there are multiple interpretations, and the droughts themselves are fairly different.

This relies on the authors having confined their inference to lowland Amazonian forests proper. If they have not, and include non-forest areas and other habitats listed above, their inferences do not match their hypotheses and their interpretations of the results of the analyses would be suspect, if not invalid.

RESPONSE R2.6: *With regard to reviewer's 2nd point, please see **R2.3** above, clarifying that our analyses are indeed confined to "lowland Amazonian forests proper".*

*With regard to the reviewer's first point about alternative hypotheses and being our own harshest critics: we appreciate the advice to consider alternative interpretations; we have acted on this advice by reporting more fully our efforts to test (and where possible, rule out) some of the key alternative hypotheses. This has helped to strengthen the manuscript tremendously. Our response to this concern is two parts: **first,** we now explicitly report in the ms on tests of some of the most important alternative hypotheses (<u>see new methods section 3, "</u>*<u>Testing alternative interpretations and considering caveats")</u>*. To summarize, some of those key hypotheses are:*

1. *(response at **R2.16**): that the primary spatial scale of analysis (~40km, in order to achieved statistical independence) is too large and does not reflect an aggregation of the dynamics at the smaller scales where the hypothesized mechanisms operate (e.g. hill to valley variation in water table depth happens at the scale of 100 m or less) (<u>rejected</u>, insofar as we are able to test based on comparison of water table depth vs forest drought response at 40km, 1km, and 30m scales, <u>see our new methods section 3, Hypothesis 1, and fig. S12</u>).*
2. *(response at **R2.15**): that drought responses are different in dry and wet seasons or that droughts are different with different severities/durations (<u>partly supported</u>, but with an effect size that does not much affect overall conclusions, <u>see methods section 3, Hypotheses 2 and 3, and fig. S13-14</u>)*
3. *(response at **R2.27(4)**): that, although only forests are included in the analysis, some included forests near deforested regions may be partially degraded and inclusion of these partly degraded forests may substantially affect the patterns of drought response (<u>supported</u> in the sense that there is a detectable effect of degradation, but <u>rejected</u> in that it does not make much difference in the basin-scale interpretation, <u>see methods section 3, Hypothesis 4, fig. S15</u>)*
4. *that the factors that produce good predictions of anomaly response to drought (in GAM models) are not the actual causal factors (<u>rejected</u>, to the extent tested by the new DAG analysis (<u>see method section 2.6, fig. S9 and S10,</u> and discussion in R1.5, above) and by additional tests (e.g testing the effect of dry vs wet season droughts, see point #2 above, and the effects of soil texture, fig. S11)*

Second *(as we stated in the "summary of key changes" in the cover letter), we urge the reviewers to appreciate that part of the novelty of our study is that it takes a hypothesis-testing approach, the gold-standard in much of science, but rare in remote sensing studies which often start by looking for patterns and propose explanations after-the-fact. This study, instead, focuses on tests which may either falsify or support the predictions of <u>a priori</u> mechanistic hypotheses proposed previously in the literature based on ground plot studies. When hypotheses are clearly supported or falsified (as they are in this analysis) that is a mark of a successful investigation, and it is appropriate to state the results of a conclusive test strongly and clearly, as we do here. We don't think these statements are too strong or oversold (especially now, in light of the extensive additions suggested by reviewers that we were able to act on), but accurate characterizations of the results of the tests. (<u>We have now inserted text and references</u>*

at the beginning of the methods section, line 517 to clarify that we use a hypothesis-testing framework in this study.)

We also agree that this is a separate matter from the reality that although one can falsify a wrong hypothesis, a correct hypothesis can never be proven (even when a hypothesis is strongly supported there are always alternative hypotheses); we agree it is nonetheless incumbent on authors to discuss (and ideally rule out, where possible) the most obvious or potentially problematic alternative hypotheses. We hope the reviewer agrees we have now made a strong effort to meet this standard (summarized above, in part 1 of this response R2.6), and that this paper may lead to novel understanding that creates the foundation for a new round of hypothesis tests that could not have been conducted before, but which provide an agenda for productive future work.

F. Suggested improvements: experiments, data for possible revision

GOSIF and EVI are argued by the authors to give similar information but their own data have spatial differences, and also the anomalies are poorly correlated with each other (Fig 1 vs. Fig S2, Fig S10). This is discussed in more depth below, but given the known variability in what differently derived vegetation indices inform us about, and how they fail, why not use multiple indices in a composite or latent variable analysis of drought?

RESPONSE R2.7**: We agree with the reviewer that there are spatial differences in the correlations between EVI and GOSIF and that EVI and GOSIF represent different dimensions of photosynthesis. Given this, we realized that former Fig. S10 was potentially misleading and have removed it (see more detail below, response* **R2.25). Focusing on finescale differences between capacity (EVI) vs. activity (GOSIF) (including developing an improved index of drought response) may indeed be helpful in more refined future analyses.*

G. References: appropriate credit to previous work?

Yes

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

Clarity and writing is good. The results are oversold given the strength of the data.

***RESPONSE R2.8**: We hope that with the significant new revisions, including mechanistic tests (see response R2.6), statistical checks, and notable new validation efforts (see expanded remote sensing validation section 1.6), the reviewer is persuaded that the reported results are well-supported are not excessively oversold.*

Additional Comments, General and Specific:

In general, the paper does a good job of establishing the hypothesis that there are non-climate factors influencing drought response, and that these should lead to regional variability in drought responses. The evidence they present for the predictors being important is relatively weak, but that is fine in general—this feels more like a molecular biology paper where there is an overview of some observed correlations/responses, and then a model/explanatory framework is developed post hoc, setting the stage for further testing.

> ***RESPONSE R2.9:*** *We respectfully disagree with the characterization that this work provides an "overview of some observed correlations/responses, and then a model/explanatory framework is developed post hoc". Though common in remote sensing studies, such an approach is the opposite of our approach here (see also Response R2.6, point 2). We hope that the revisions make more clear to the reviewer and readers that a more accurate characterization is that the work adopts a hypothesis-testing framework: we start first with mechanistic hypotheses independently generated by previously published local scale plot studies, and then follow with our analysis of validated remote sensing observations that test whether and how predictions of these local site-specific hypotheses are supported (or not) at the basin scale. Further interpretation of course leads to an additional round of refined hypotheses, which ideally become a priori hypotheses for future work to build on the insights brought here -- but this is quite different from the "patterns-first, then post-hoc explanations" model.*
>
> *As stated in R2.6, we have inserted text and references (at the beginning of the methods section, line 517) to clarify the hypothesis-testing approach taken in this study.*

The Other Side of Drought focuses on water table depth and posits that forests on shallow water tables should be resilient to drought with lower brown-down or even green up, the latter due to less anoxia (presumably from decreases in water saturated surface and near-surface soils) and decreased cloud cover. A difficulty in this is the spatial resolution necessary in the predictors (forests can change height substantially within 1 km2 resolution of the sampling, including both low HAND and high HAND areas), as well as the temporal resolution of the sampling (one would need to show a decrease in cloudiness from the long term average in times of year when species can take advantage of it, or when it is particularly cloudy).

The Soil Fertility Hypothesis posits that in fertile forests with higher turnover rates investment in traits for drought resistance are reduced as trees will be quickly replaced, and these forests will have greater drought susceptibility. The authors write this as if it was an individual species making these allocations, or if trees were all one species, when the paleo and modern literature would both show that if a community were to be drought resistant it would be due to environmental sorting favoring the assembly of communities of species/taxa with different trait combinations. Forests 'adapt' to large changes in environmental conditions through changes in species composition.

The third hypothesis is that trees with drought tolerance or drought avoidance traits are more drought tolerant, which seems to be hard to argue with (again, I think it is important to clarify that forests only have trait spectra due to the performance of individual species possessing those traits). Here the traits chosen to reflect drought tolerance are debatable in many cases—tall trees, or at least large trees, are chosen by the authors, when literature on drought, particularly Nepstad's drying experiments, show that large trees are the first to succumb to experimental drought, which was interesting. Another reason you have large trees that have resilience in nutrient poor areas is just that growth rates are low and stability is favored—this leads to traits like higher wood density as well. Of course, when water stress gets even stronger, trees decrease in stature. The hypothesis that long-term episodic drought resistance is favored by large, densely wooded species is interesting and could be tested with mortality data within demographic plots.

> ***RESPONSE R2.10**: As ecologists first and remote sensing scientists second, we heartily agree with much of what the reviewer has written here about underlying ecological mechanisms of the three hypotheses, and we don't believe that anything we have written for our ecosystem-scale analysis is at odds with a more mechanistic ecological understanding of forests as communities, e.g. that "forests adapt to large changes in environmental conditions through changes in species composition," or that trait spectra can emerge from assemblages of species each containing particular species-specific traits. Indeed a purpose of this study is to connect these scales by testing ecosystem-scale predictions derived from locally observed ecological mechanisms, thereby supporting with evidence the idea that those ecological mechanisms do regulate ecosystem scale patterns (even though, in this large-scale remote sensing study, those ecological mechanisms can't be directly observed).*
>
> *To address the specific issues mentioned by the reviewer, we have made the following changes/additions:*
>
> 1. *with respect to spatial scaling: as noted elsewhere (R2.6, point 1 and R2.16) methods section 3 now tests the scalability of the forest drought response to WTD. To our surprise (and contrary to our expectation of increasing sensitivity at finer scales, as we wrote in the first version of the manuscript) we found that the sensitivity of forest drought response to water table depth appeared to be scale-invariant (due, we think, to an approximately canceling "regression to the mean" effect of averaging on distributions of both HAND and drought response).*
> 2. *With regard to the statement that "*one would need to show a decrease in cloudiness from the long term average in times of year when species can take advantage of it, or when it is particularly cloudy": we believe the first point already follows, at least implicitly: previous work (eg. Huete et al (2006)[89]; Wu et al (2016)[96]) shows that leaves flush and grow in dry seasons (taking advantage of extra sunlight). This current ms documents drought-induced increases in sunlight (due to decrease in clouds) in the dry season for 2005 and 2015 (though not for 2010, due apparently to especially strong smoke then). Thus, the increase in response to drought, where it happens, is indeed likely due to a "decrease in cloudiness" (increase in sunlight) at "times of the year when species can take advantage of it" (i.e., the dry season).
>
>   The second point about "when it is cloudy": as we discuss also in R2.15, our new methods section 3, hypothesis 3 (and fig .S14), responds to the reviewer's suggestion that forests may be more responsive to extra sunlight during normally cloudy (wet season) conditions. Our GAM modeling framework helped to address this directly by showing that forests are indeed also more sensitive to sunlight in wet season droughts vs dry season droughts (just as the reviewer hypothesized), but that this is within the context of forests also being more negatively impacted on average.
>
> 3. *With respect to drought traits:*
>   a. ***Tree height:** We thank the reviewer for reminding us of the Nepstad et al[127] (and incidentally da Costa et al. 2010[46]) results about tall tree sensitivity to drought, which motivated us to look more closely at our results. We address this now by*

*__**(i)**__ clarifying that according to our model, forest canopy height can be either a risk or a protective factor against drought, depending on the context (line 215:*

> *"Meanwhile, taller forests performed worse than shorter tree forests in shallow water table areas (Fig. 4D andfig. S10D, red vs blue lines for low HAND values), consistent with findings that when lacking a deep root advantage, tall trees may suffer higher drought mortality due to greater exposure to atmospheric drought (high VPD)46"),*

*and **(ii)** by adding a fuller discussion, as "hypothesis 5", in methods section 3 (line 1176:*

> *"Observations of drought responses in the RAINFOR network[48] and drought experiments[46,47] report that tall trees were more vulnerable to drought. One of the drought experiments was above a moderately shallow water table (7-10m) and the vulnerability of tall trees there could be explained by our result[46], but the forest of the Nepstad et al drought experiment[47], and many of the plots in the RAINFOR network, are over deep water tables, raising the question as to whether the results reported here might be inconsistent with those.")*

*We concluded this section by hypothesizing that **(i)** ongoing multi-year droughts (as in the experiments) may make tall trees more vulnerable than in intense but transient <1 year droughts (those analyzed here), and **(ii)** that tall **forests** may be protective on average even as relatively tall **individuals** may be at greater risk.*

b. ***Hydraulic traits**: Hydraulic safety margins, assuredly a trait spectra that emerges from individual species performance, also have community-weighted means at plot scales that can affect ecosystem properties, as suggested by the recent Tavares et al 2023[49] study that we now show (see methods 1.6(b)) validates our "relative green up" metric of resilience in Fig. 5A (with inset graph predicting plot-level basal-area weighted mean hydraulic traits (equivalent to community means)[49]). This provides a clear example of a trait that is species specific, but whose aggregate ecosystem consequences can also be detected at large scales.*

The argument that these interact with climatic gradients is good, though the mechanistic interpretation is not supported—it is phenomenological, correlated with aggregate traits.

> ***RESPONSE R2.11**: we would respectfully disagree that the underlying mechanistic interpretation is not supported. We hope the reviewer may be persuaded that these mechanistic interpretations are precisely what **is** supported (to varying degrees of strength) by an a priori hypothesis-testing framework when the evidence supports an ecosystem prediction made by a hypothesis derived from underlying mechanisms.*

Drought depth and duration is known to have distinct effects on species in research from the temperate zone (in the major Texas, US droughts of the 1950s and 2010s, one killed junipers and the other killed oaks, with one drought being very deep but short, the other being exceptionally long). Do these two axes of drought have distinct effects in the Amazon? Looking at figure S4, the three droughts are not the same treatment applied to different areas—they are very different droughts that happened in very different biogeographic areas. It would seem that the important approach would be to find forests of different characteristics (variability in HAND, fertility, traits) within the regions affected by the different droughts and focus on that comparison.

> ***Response R2.12***: *We agree on the importance of exploring sensitivity to drought characteristics. The reviewer proposes a method different from ours (one focusing on selecting subsets of forests and comparing across gradients of interest-- droughts, etc). However, the variations that the reviewer points to are **already present** in the data we analyze in the larger modeling framework, hence **their effects are already captured** by that framework (at least to some extent), and the information sought can be directly extracted from our existing analysis. For example, analyzing further the model of Table S1A, we can directly explore (see figure at left, now in methods 3 on "Testing alternative interpretations", as fig. S13) how modeled photosynthetic drought responses depend on drought length and depth (or severity, defined as in methods 2.3, by the MCWD anomaly). Results show (as expected) that droughts that are both severe and long have the most negative effects (though intriguingly, droughts of intermediate length do not much differ in severity). We agree with the reviewer that whether these two axes of drought have other distinct effects in the Amazon (e,g, on mortality of different types of trees) is ripe for further investigation, and we hope our study both motivates such further work and provides a tool useful for conducting it.*

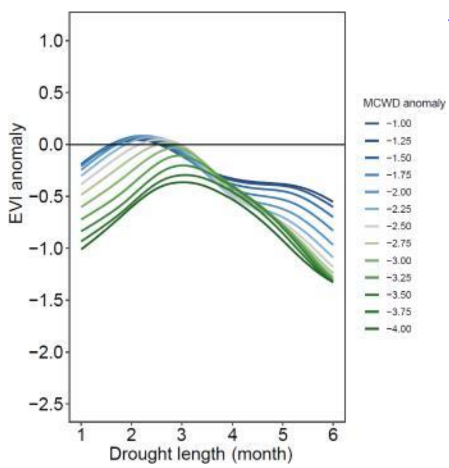


Figure. Amazon forest drought response (EVI anomaly) verses drought length, for different drought severities (ranging from 1=blue to 4=green SD below the mean MCWD), across all three droughts (based on the model of Table S1A).

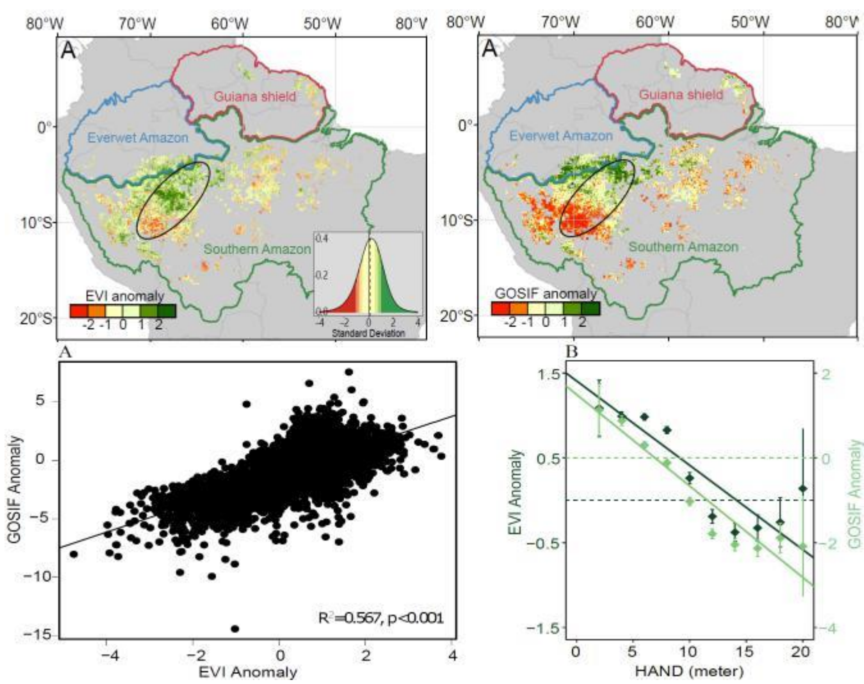
A difficulty of this paper is the attribution of causes to the mechanisms inferred. For example, the southern Amazon is referred to as high fertility, but it is only the western and a few parts of eastern S. Amazonia that is so. The much of the east is Brazilian shield, and has relatively low diversity forests, and corresponding differences in tree demography and productivity.

> ***Response R2.13***: *We hope that new analyses (including DAG analysis, see method 2.6) helps address the attribution concern. We hope it is now clear that references to S. Amazon "high fertility" (because observations show higher average fertility, but with high variability, Fig. 4A margins, fig. S3B) do not have implications for inferences from statistical models (now both DAG and GAM), which use pixel-specific values (both the high and low fertility pixels in the S. Amazon) to make inferences/predictions.*

A general issue with studies such as these is understanding the true meaning of what the vegetation indices (VIs) are revealing, and this is a longstanding debate in the field, particularly with respect to the Amazon, and several of the authors are key participants in this debate, particularly with the key measurement used in this paper, EVI. The recent review by Zehng et al. in NEE 2022 is a good overview. For the drought response, it would be more compelling to use a composite, or multivariate, analysis using different satellite-derived indices. The authors include GOSIF as an alternative analysis, but it gives different information on drought effects (direct PS measurements) that correlate better with GPP, and that is not redundant with EVI. Comparing figure 1 and figure s2 shows the spatial locations of the drought anomalies are similar but not the same, and if the same ellipses were put on Fig. S2 that would be more apparent. I'm sure the last thing the authors want to encounter is another reviewer talking about issues with satellite-derived measures of forest performance, but I think that acknowledging the differences in information given and why the occur might strengthen the mechanistic case they are making. You could think of drought response as a latent variable with many inputs, particularly as the paper goes on to make demographic and functional connections. Also, for example, using NVI *[we think the reviewer may mean NIRv]* for statements like L112

> ***RESPONSE R2.14**: Yes, we quite agree with the reviewer about the long-standing debate! We appreciate the reviewer for highlighting the distinction between reflectance based (EVI) and fluorescence based (SIF) indices and accept the advice to use the differences in information given to strengthen the case they are making (see **R2.25**).*
>
> *We share the reviewer's desire to see useful remote indices of drought, but respectfully decline the invitation to introduce yet another composite remote sensing index (we have already seen a cottage industry devoted to new remote sensing indices!). We do point to Shekhar et al 2022 (https://doi.org/10.1016/j.rse.2022.113282) (now also cited in the paper) that exhaustively and systematically reviews relationships among the various existing indices (including NIRv) and tower-derived GPP in multiple biomes worldwide.*
>
> *In contrast to a new index, what is needed (and which we offer, see revised/new section 1.6 on "Remote sensing validation) " is a compelling case that existing indices are meaningfully tied to relevant ground observations – especially our new remote sensing-derived resilience map (Fig. 5), now validated with ground-based tree hydraulic resilience to drought (Fig. 5 inset). This is perhaps the most important/relevant validation, as the resilience map, the novel development of the paper, is untested.*
>
> *For the record, we follow reviewer's suggestion to put the Fig. 1A 2005 ellipse on Fig. S2 (see below). We hope the reviewer agrees that this (panel B 2nd row) shows concordant support for a hypothesis (notwithstanding low R^2 between EVI and GOSIF) and this is indeed what we look for in a confirmatory test of a hypothesized pattern. We now include panel B from row 2 below, showing consistency between indices in support for hypothesis 1 (negative slopes with respect to water table depth), as panel D in Fig. S2.*

Figure. Manuscript Fig. 1 and Fig. S2 A panels showing EVI (left) and GOSIF (right) 2005 anomaly maps. Panels below show the relationship between EVI and GOSIF anomalies (left) and between the vegetation indices (on the vertical axes) and HAND (right) in the 2005 drought ellipse region (depicted in the panels above).



Another interesting part of this manuscript are the regional generalizations and the generalizations about the droughts. For reasons listed above and below, the regional generalizations, particularly about the S. Amazon, are inaccurate. Also, looking at the droughts, they are qualitatively different, not just spatially different. For example, the 2015 drought is centered on N. SA and from figure S4 looks to be an exceptionally dry wet season rather than a dry dry season as in areas experiencing the 2010 and 2005 drought (peak rainfall in Boa Vista is in May/June with a sunlight minimum in June, the exact inverse of Rio Branco, or even Iquitos and Manaus. Given that light limitation is a primary hypothesis, and that light limitation is much more likely to be a wet season phenomenon, it would make sense to place more emphasis on how the timing with respect to the wet/dry season affects interpretation.

> ***Response R2.15**: We appreciate the reviewer's concerns on regional generalizations. (1) Like the reviewer, we recognize many heterogeneities within the three regions, but respectfully disagree that that makes the regional classification somehow "inaccurate". The reviewer seems to be referring, e.g., to characterizations like "Highly fertile areas (with greater frequency in the southern Amazon)..." (line 196) But that is not inaccurate, it is simply a fact of the dataset -- as is the high variability in the S. Amazon fertility distribution that also includes low fertility areas. This is prominently evident in Fig. 4A, as in the fertility map (fig. S3); so we are not "hiding the ball" on S. Amazon soil fertility!*
>
> *Nonetheless, we have made text changes to avoid overstatement or misunderstanding, including: at current line 34 ("In high**er** fertility southern Amazonia", instead of "high fertility)"; at around current line 92 (deleting entirely the sentence at former line 91 that included reference to "the Southern Amazon (with the most fertile soil))"; and at current line 848 (adding highlighted words "...with soil fertility that was both higher on average but also more variable) "*
>
> *(2) We agree that the three droughts are different, and that some investigation into these differences is important. To that end, we (1) note that effects of inter-drought variations (being present in data used to fit models) are already part of our object of study and are, at least in part, captured by the modeling (as seen in Fig. 3 showing different responses to different droughts); and (2) investigated how the hypothesis of light limitation might be structured by the timing of different droughts with respect to seasons (see added methods section 3 hypotheses H3). The initial results (fig.S14) (possible followup paper!) suggest both that the reviewer is right about the greater sensitivity of wet season droughts to sunlight, and confirm general trends reported in Fig. 3 about interdrought sensitivities to PAR and drought length, thereby increasing confidence in that analysis.*

Specific Comments

Figure 2. HAND—it looks like HAND is also calculated for the Andes, when the paper does not treat Andean forests. If forests above 500m (a traditional definition that separates the Andes from Amazon) were removed, it would give much better discrimination among forests. Also, in several of the authors' original conception of HAND it captured hydrologically relevant, very local features like flow patterns on hillslopes to drainage and had good correspondence to local measurements of soils (Nobre et al. 2014). Here the application resolution is very much coarser--looking at the map it appears HAND is not saying much about drainages per se on the scale that trees experience water tables, but rather looking at position with respect to major

geomorphological features—the Guinan and Brazilian shields, the Fitzcarrald/Purus arch, etc. (e.g. figure S11) as opposed to how it was conceived.

> ***RESPONSE R2.16**: **(1)** Re the Andes: we hope the reviewer is now satisfied that Andean forests and non-forested regions are not included in, and do not affect, analysis of shallow vs deep water table forests (see also R2.3, above).*
>
> ***(2)** Re the scale of HAND: Indeed this is an important question; we thank the reviewer for motivating a more careful analysis of scaling effects on drought response to HAND (as in R2.6 and R2.10(1), we cite our new methods section 3 and the new Fig. S12). We hope the reviewer agrees that this scaling analysis strengthens our findings.*

Also, within the HAND results, the effect is largely confined to forests with HAND ~>10m. These would include many forests that are seasonally inundated, large wetlands, etc. How was seasonally inundated (varzea) forest dealt with? In normal years these may have lower greenness just because they are underwater for large portions of the year, giving large positive anomalies in years where they are not flooded, etc. Integrating the GOSIF measurements, or other VI measurements, may be a way around this.

> ***RESPONSE R2.17:** We hope it is now clear (e.g. see **R2.3**) that seasonally inundated (varzea) forests and wetlands are excluded in the modeling analysis.(note that in original Fig. 2, we displayed innundated forests as hollow diamonds, in a separate bin, with negative WTD, to show that they were distinctly less likely to "green up" than shallow terra firme forests, even though they were excluded in the modeling analyses; we have since removed these to avoid distracting readers with another variation in forest type).*

L31: A tipping point of what?

> ***Response R2.18**: FIXED: "...tipping point towards forest collapse."*

L45: Is it undermining ecohydrological strategies? Or is it just exceeding system capacities for resilience, and those capacities depend on the predictors included, plus any unobserved capacities/adaptations for drought resilience? The latter is an important factor—forests in more seasonal areas of the Amazon have tree floras with a higher proportion of individuals and species from families with dry forest affinities than areas in everwet forests (Gentry 1988, Pitman et al. 2000). The portfolio of adaptations to drought contained within lineages is clearly important in setting distributions, and is much more than simple measures of height, etc.

> ***Response R2.19**: Yes, we agree that droughts are exceeding system capacities, but our point is also that some of these capacities are associated with particular strategies. For example, the strategy of growing in valleys where water tables are shallow (consistent with observations of changing species assemblages and functional compositions along the gradient of water table depth from plateau to valley) might be undermined in higher fertility regions. In any case, we changed the phrasing (to allow for both possibilities) to "undermining multiple ecohydrological strategies and capacities for Amazon forest resilience" (line 44) and also "undermining different strategies and capacities for drought resilience." (at line 291)*

L91. It doesn't seem from figure 5 like three really emerge--it is a highly complex, geographically localized set of predictions that don't square easily with the three defined regions in figure S3.

Particularly in the southern Amazon, which is highly heterogeneous and that defies easy classification. The Everwet Amazon shows much more heterogeneity than is described as well.

More generally, the S. Amazon is highly heterogeneous in soil fertility, as figure 5 alludes to, etc.

> ***RESPONSE 2.20****: We are gratified that our Fig. 5 analysis predicts what the reviewer emphasized all along: that Amazonia exhibits a rich complexity of responses. That we can now predict this complexity from a small discrete set of scientifically sensible rules (the three hypotheses) is, we think, a signal accomplishment of the paper (it is indeed a goal of science generally, to understand how complexity is generated from simple rules).*
>
> *Re. whether Fig 5 "easily squares" with the three regions: We hope it is clearer from our revisions (which now de-emphasize the regional classifications), and, e.g. R2.13, that <u>the predicted biogeography (Fig. 5) does not emerge from (and is independent of) the three-region classification (depicted in former Fig. S3, currently S6)</u> being discussed here at line 91. <u>To avoid this impression, the sentence there (lines 89-95 in original text), "three distinct Amazonian forest regions emerge ... (fig. S3)" has been removed.</u>*
>
> *The rationale for the 3-region classification is <u>now discussed in greater detail (methods 2.2, 822-26, 850-55)</u> for its purposes (different from the functional biogeography): to emphasize Amazonian heterogeneity (it is not just "the" Amazon forest as in many papers, but consists of functionally distinct forests); and second, to tractably organize the exposition of results (e.g. via the 3 regions delineated in the Amazon maps).*
>
> *Perhaps it is also helpful to clarify that clustering in a multidimensional space (as in the classification analysis of fig. S6 (formerly fig. S3)) does NOT mean that the regions that emerge are necessarily homogeneous (as would be the case in one dimension). Regions *may* cluster because they are similarly homogenous, or because they are highly heterogeneous in similar ways. With regard to the soil fertility issue that concerns the reviewer, for example, its distribution in the S. Amazon (Fig. 4A, on the right margin axis) is not only higher (on average), but more variable than the other two regions.*

L115. HAND would seem to be much more of a measure of topography than water table at the scales measured—how does inference about HAND change with different scaling? Or is EVI binned at such a coarse scale that it can't be resolved? See comments above as well.

> ***Response R2.21****: Yes, scale is an important issue. Please see* ***R2.10(1)****.*

L146. Again, were PAR anomalies higher in 2015 because that drought occurred during the wet season in many areas (though with some dry season affect in S. Amaz)?

> ***Response R2.22****: Partially yes, our biogeography model provides a systematic tool to test this (and other) hypotheses (<u>see methods 3, fig. S14</u>, as discussed in* ***R2.15****).*

L162. This result about the RAINFOR plot AGB mortality and HAND is compelling, but it, and the general link of demographic effects to EVI and HAND are tempered when one looks at figures S8-S10. ANPP rates are explained poorly by EVI, accounting for only 25% of the variation, the plots in figure S9 showing little predictive ability for mortality from EVI (B) and that looks to be influenced by the highest mortality plot. Similarly, the three highest recruitment plots have the highest EVI anomaly values (C), but the rest of the points cover the full range of EVI anomaly—it is likely nonrandom, but how much information does it convey beyond the extremes, particularly if one were predicting a full explanatory biogeography? The mortality:recruitment vs. water table depth relationship appears to rely on 1-2 plots (D), and the mortality WTD plot is similarly dependent, but NS. Also, why not plot recruitment in a panel (F)?

> ***Response R2.24**: **(1)** We appreciate that the reviewer finds the HAND-RAINFOR mortality compelling, and hope that we can persuade them that the links (with R^2=0.25-0.35) between plot demography and remote sensing is also compelling. First, with respect to RAINFOR ANPP (current fig. S16): only one-half of the NPP signal is influenced by GPP in the first place (the balance being autotrophic respiration), so we should not expect an R^2 of more than, say, 50% based on first principles alone. Second, since ANPP is also affected by belowground processes (also not detected by remote sensing); and that forest plots are smaller than the pixel scale observations of remote sensing (adding unexplained variability), we think that the explained variation in ANPP and other demographic factors is about as much as should be expected by a remotely sensed photosynthetic index (and we would be skeptical of claims of correlations much higher than that!). <u>As stated in the revised manuscript methods section</u> (lines 727-730):*
>
> > *"We expect that skill in predicting demography to be weaker than for predicting photosynthetic fluxes, because demography emerges, not from photosynthesis alone, but from the balance ofphotosynthesis and autotrophic respiration..."*
> >
> > *and (around line 737): "The R2 values of 0.25 to 0.35 for remote detection of demography (figs. S16-S17) are consistent with our expectation that they should be about half of the remote detection R2 for GPP (0.5 to 0.6, discussed in 1.6(a)), since GPP is about one-half the determinant of the NPP driver of demography."*
>
> *(2) With respect to the water table depth panels D-E of original fig. S9 (now fig. S17): considering that this figure is intended to focus on remote sensing validation, and that panels D-E did not address validation, and that Sousa et al[30] more fully addresses these effects (with a more complete analysis than we can include here due to our different scope), <u>we decided to remove these panels to avoid distracting readers, and simply cite Sousa et al 2020[30] for those who wish to examine forest plot dynamics in more detail</u>.*
>
> *(3) Regarding concerns (potential skewing by high leverage points, "likely non-random" distributions of plot data) about robustness of validation regressions between remote sensing and demography in (now) Fig. S17 (originally figs. S9-S10): we did diagnose those, with metrics we resurrect here for the reviewer's benefit (Fig. 1-6 below); <u>we now reference the Cook's distance diagnoses in the revised version (Fig. S17)</u>.*
>
> *The diagnostics do not appear to us to show much evidence of the concerns that worried the reviewer: there is little evidence of "non-random" data distributions, beyond occasional modest deviations from normality (a tendency toward light-tailed residual distributions in EVI vs recruitment (Fig. 2), EVI vs mortality:recruitment (Fig. 3), and GOSIF vs recruitment (Fig. 5)). The EVI plots (but not the GOSIF plots) do contain high leverage points (Cook's distance > 4/n), but the regression relations and statistics are all robust to the removal of these high leverage points (in the figures below, the dotted regression lines and second set of statistics exclude the high leverage point, shown as a hollow circle). (<u>the sensitivity analysis to leverage point influence is now shown in the ms Fig. S17</u>, to inform the reader's discretion on interpreting these graphs).*

Fig 1: EVI v Mortality diagnostics

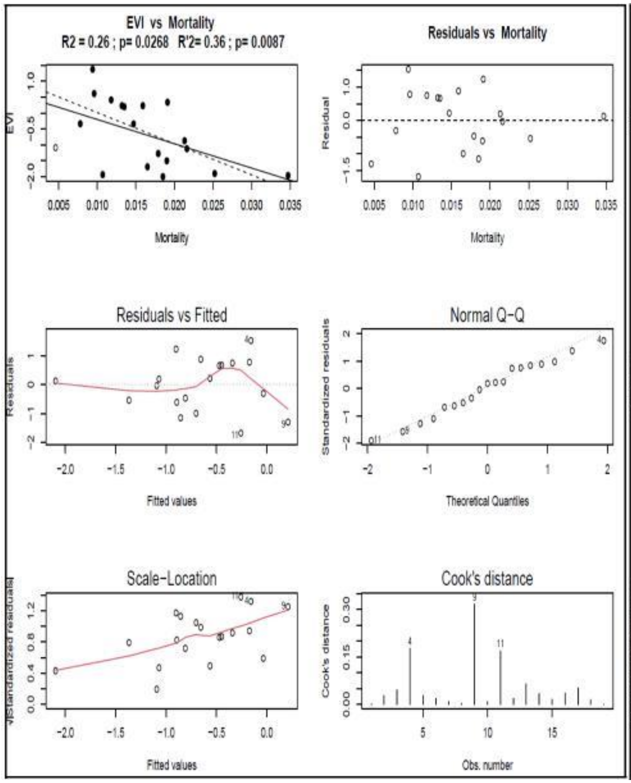


Fig 2: EVI v Recruit. diagnostics

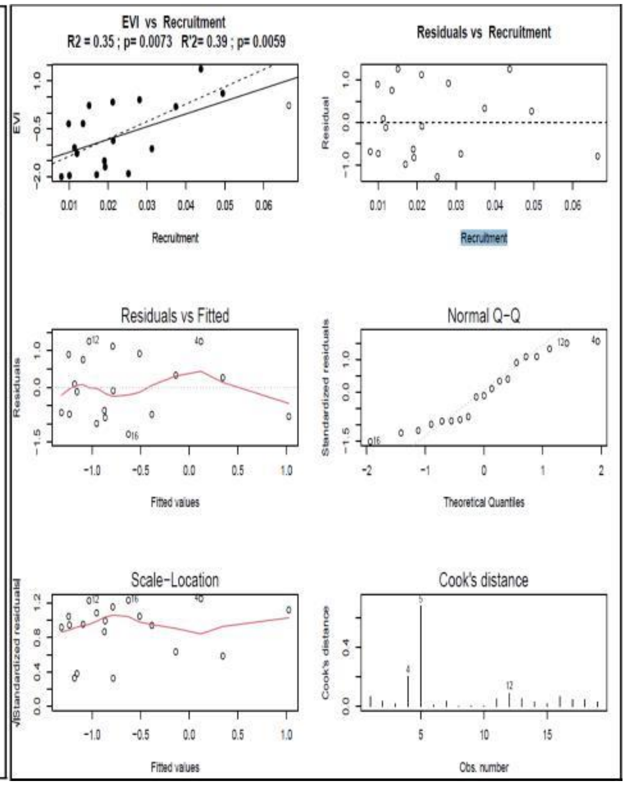


Fig 3: EVI v Mortality:Recruitment diagnostics

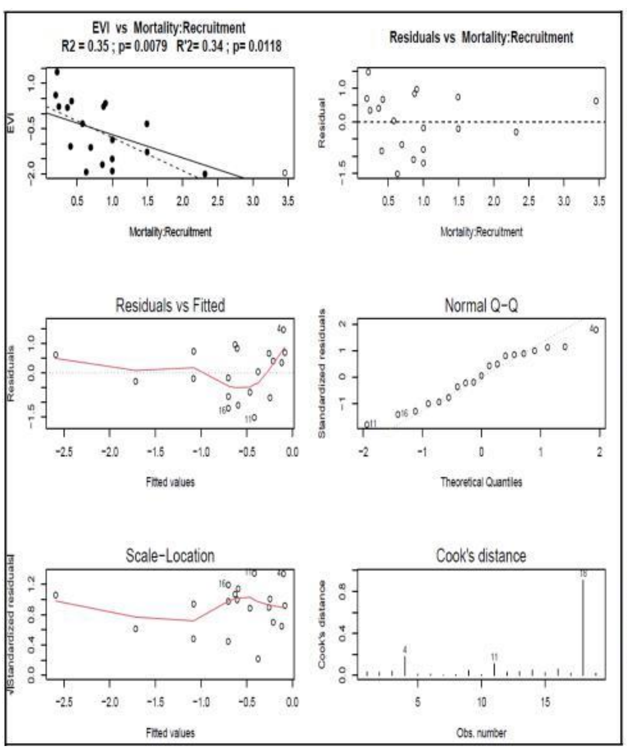


Fig 4: GOSIF v Mortality diagnostics

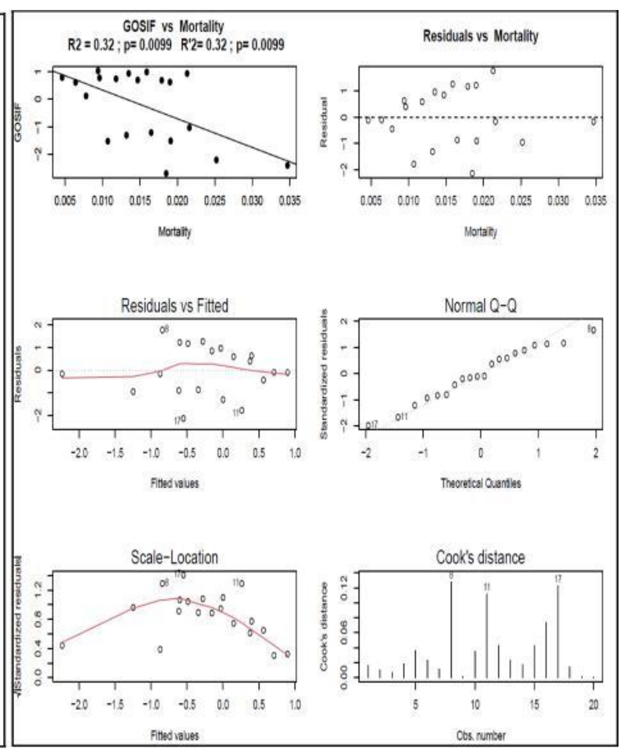


Fig 5: GOSIF v Recruit. diagnostics

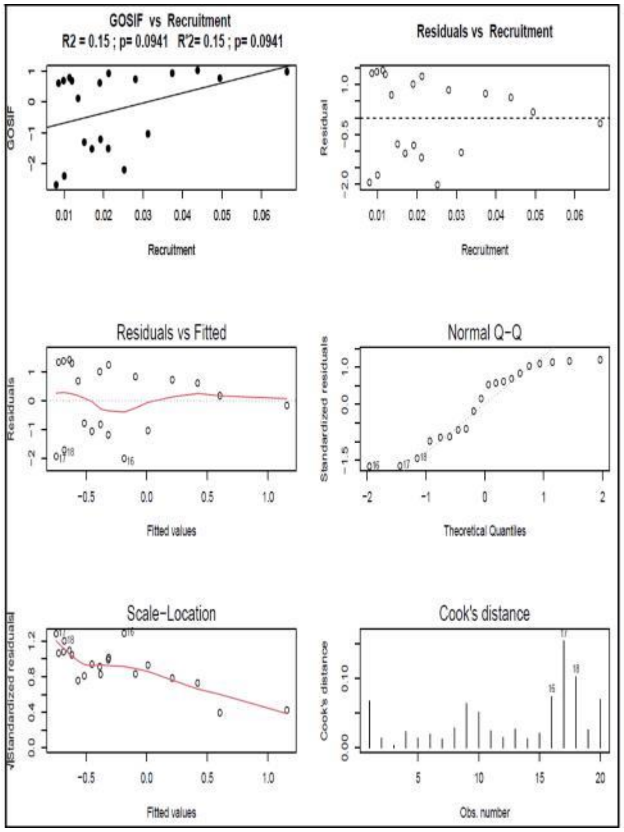


Fig 6: GOSIF v Mortality:Recruitment diagnostics

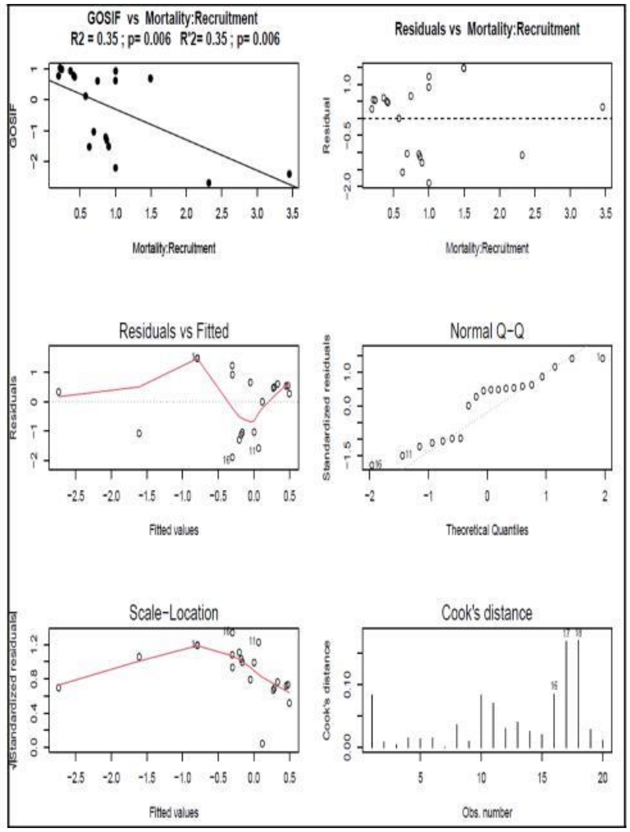


Figure S10 shows a poor relationship between EVI and GOSIF anomalies—what is the explanation for this? It undermines the argument made that these are telling similar stories and have good predictive value. Three-quarters of the variation in one is not explained by the variation in the other (A). Panels (B) and (C) are heavily influenced by 1-2 data points. It isn't fatal, but it should temper the argument, and the authors might use this to present this part of the argument as the basis for a hypothesis and to point for the need of sites with additional demographic data, or other measures of forest structural change, etc., rather than key support.

> ***RESPONSE R2.25****: As stated in R2.14, we are accepting the reviewer's advice to use the differences in information given by EVI and GOSIF to strengthen the mechanistic case they are making: <u>we now better clarify (methods 1.1. Remote sensing indices) the distinction between EVI and GOSIF and how using them both strengthens our case</u> (see excerpt below). This distinction (as also mentioned by the reviewer preceding R2.14) explains why the GOSIF-EVI relationship in the former fig. S10 is not stronger.*
>
> *We hope the added text clarifies this distinction and that we should not expect (and we hope that readers do not expect), a relationship between EVI and GOSIF based on a duplicative standard that expects them to be the same. <u>To avoid this unfortunate expectation (perhaps inevitably fostered by our original use of a bivariate scatter plot of EVI vs GOSIF in fig. S10) we have removed that graph from the paper</u>. As a justification for why GOSIF and EVI are in fact consistent (adjusting for their being metrics of*

different aspects of photosynthesis) we point the reviewer to the original GOSIF paper (Li and Xiao 2019)[27] which shows that SIF is well-predicted (R^2=0.79) by a model of EVI + climate (sunlight, VPD, temperature). (think of GOSIF as EVI+climate; to the extent that GOSIF and EVI differ, it is due to the fact that climate is not a part of EVI) (see also response to reviewer 5 below at R5.6)

We further call attention, in particular, to text (methods Lines 770-778 in the revised manuscript) that we hope clarifies for a reader who might share the reviewer's concerns:

> *"More important, we ask whether there is consistency in terms of support for or rejection of hypotheses that are the focus of this analysis -- for example, whether the "other side of drought" prediction that drought response anomalies should decline with water table depth, and here we do see broad support for this hypothesis from both EVI and GOSIF...(fig. S2D)...(Fig. 3A vs fig. S2E) ... (fig. S17). Together, these comparisons increase confidence that forest drought response hypotheses are robustly supported by the two indices."*

L185 and on. This is one interpretation of the drought effects, but what about other interpretations? There are many explanatory degrees of freedom here, so this story becomes one of many. If the authors are their own worst critics, what would they say about the coarseness of HAND as used in this paper, the scale of measurements of canopy height, which does not necessarily equal average tree height in many areas of the Amazon, particularly given that much of the area measured is degraded or otherwise human altered forest.

***Response R2.26**: The reviewer may be exaggerating the explanatory degrees of freedom here, as much of the interpretation is in fact highly constrained by the modeling results (especially now re-inforced by the DAG analysis). With regard to the specific concerns the reviewer raises: Re coarseness of scale, please see analysis in the new fig. S12 (and methods 3, hypothesis 1), which shows the robustness of sensitivity to water table depth across spatial scales of aggregation, from 40km, to 1km, and then, down to 30 m, using Landsat data from around Manaus. Re tree height: as mentioned earlier in R2.3, we have carefully excluded non-forest areas from our analysis--and when deforestation encroaches within the 40km grid cell, we represent the grid cell only by those sub-pixels that are not deforested (see the additional clarification to methods line 798), down to the native resolution of the remote sensing products (for canopy height, 1 km), so we expect the analyzed pixels to indeed represent true average canopy height in targeted areas.*

Figure S11. The authors' assertions about the simplicity and sufficiency of the three-area classification is belied by the complexity in figure S11. The authors state that the S. Amazon has high fertility and short trees, but it has both high fertility and low fertility and very tall and short trees (and a lot of degraded and disturbed forest). Puzzlingly, cerrado vegetation—a mix of trees and grass—is included in this analysis, which is qualitatively different habitat type from Amazonian forest. With low tree height, and with trees interspersed among large areas of grass, the canopy heights will be low. Similar savanna in Bolivia is not masked out, and I did not find in the SI where the authors did this or mask out degraded or deforested areas. The analysis and conclusions should certainly be confined to forest only. S3B shows the problems with this classification. The paper has forest in the title but a large percentage of the inference area is not forest, but rather cerrado, edaphic savanna, deforested areas, agriculture, and degraded forests. Within what is commonly thought of as Amazonian lowland forests, apparently, forests in the Andes and Guianan tepuis are included.

RESPONSE 2.27.

(1) *We did not intend to make assertions suggesting that the three-region classification is sufficient in a general sense, but upon reading comments by this and other reviewers we agree that the initial version of our paper may have over-emphasized the classification analysis in a way that magnified its significance out of proportion to its role (especially since the main results and conclusions of the regional modeling do not depend on it at all!). Thus, we have reduced the prominence of the classification analysis in the main text (removing the sentence beginning "Indeed, three distinct Amazonian regions emerge..." from its prominent placing, lines 89-95, in the original submission), and have more precisely described its purpose in the methods section, emphasizing first that its goal was to determine "whether the distribution of factors defining forest types (ecotopes) across Amazonia could lead to a coherent clustering of different forest ecotopes into different regions" (lines 806-809) and second (conclusion of classification section, lines 850-855) that*

> *"This three-region classification (which we use to define the regions depicted in the main text figures) is independent of the results (Figs 3-4) of the basin-wide modeling investigation (described in sections 2.7-2.8 below) because model predictions depend on pixel-pixel variations of environmental factors regardless of what region they are in. However, the three region Amazon is useful for presenting model results because it illustrates how different functional responses emerge from different ecotype regions (as shown in Fig. 4E)."*

(2) *Regarding "high fertility" and "short trees": Observations show that the southern Amazon's mean soil fertility is modestly higher than the other regions and that it's mean tree height is lower. Though we report these observational facts from the existing datasets, we also agree with the reviewer that the distributions encompass both higher and lower values (and as the reviewer notes by referring to fig. S11 we show these distributions plainly -- see also the marginal distributions in Fig. 4A.B which we included by design in order to make the distributions in the different regions plain to the reader). We have also emphasized more strongly the high variability of soil fertility in the Southern Amazon wherever we mention that its average is high (e.g. in methods section 2.2. line 848: "The southern Amazon is then differentiated further from the Guiana shield as slightly dryer, with soil fertility that was both higher on average but also more variable.) "*

(3) *As clarified above (e.g. **Response R2.3**) our domain is intact terra firme forests only (which, as the reviewer points out, is quite complex enough as it is!)*

(4) *The reviewer's concerns, and those of reviewer 3, however, motivated us to look more deeply into the question of degraded forests. Though deforested regions are excluded, the mask still includes forests in proximity to deforested regions, that, though not deforested, are likely experiencing degradation. We conducted a sensitivity test to address this issue. The new methods section 3 (motivated by reviewer comments about the need to consider alternative interpretations), now*

includes a sensitivity test that suggests that partially degraded forests likely have greater drought sensitivity, but also shows that the overall patterns in the results do not depend on whether these forests "in between" deforested and intact regions are included or excluded. ***Fig. S15*** *shows the functional relationships of pure intact forests, which do not meaningfully deviate from the main Fig. 5A,B (which includes some forests classified as impacted). We thank the reviewer for stimulating us to make this clarifying analysis that we hope increases confidence in the results.*

L240. Given the above, this doesn't generate much confidence—the relationships aren't strong, and the similarities are not assessed in a rigorous way, but rather by visual inspection, and when they are looked by plotting one against the other (GOSIF, EVI) they don't have a great relationship. The authors would need to better support these statements before the narrative explanation of their results would be easily believable, and also be their own worst critics in terms of what else could be giving them this pattern.

Through L258, this all is argumentative, but I don't see it backed up by the actual results, which are not compelling, and when there are problems with the definition of ecotype that, even if the results were true, could give substantially different interpretations.

> ***Response R2.28****: We thank the reviewer for stimulating us to engage in multiple new statistical analyses to strengthen causal inferences and rule out alternative hypotheses (See* ***R1.5*** *and* ***R2.13*** *and the new methods section 3, "testing alternative interpretations"), to include additional validation efforts (fig. S16); to show that what the reviewer originally saw as weak relationships are in fact exhibiting the expected strength, and to further and better clarify the role and differences between EVI and GOSIF, see* ***R2.25*** *and* ***R2.14****). We hope the reviewer feels we have taken to heart their advice to be our own "worst critics"; we certainly think that the reviewer's tough comments have helped us to make a much stronger manuscript aimed at better supporting the statements in the main text; we hope that this process has helped as well to increase the reviewer's confidence in the overall analysis.*

Referee #3 (Remarks to the Author):

The manuscript addresses a very important question with important implications for vegetation-climate predictions : why have some Amazonian drought datasets reported forest green-up, or reduced mortality), while others observe forest brown-down, or enhanced mortality?

The manuscript represents a very innovative approach (comprehensive functional biogeography) and hypothesis framework to discriminate potential effects of water table depth, soil fertility/forest type; and tree rooting depth as estimated by proxies of forest functional composition.

The manuscript includes very strong datasets: EVI and GOSIF datasets, together with mapping products and a comprehensive analysis of published and novel demographic data from permanent forest plots over the three major recent Amazonian droughts.

And the manuscript reports potentially groundbreaking results; notably, a strong positive correlation between the extent of 'brown-down' and the depth of the water table as inferred by

HAND, particularly along a stretch of south-central to south-west Amazonia. Although all regions did not react to drought events in a similar way, tree height and forest soil fertility appear to explain differences in EVI anomalies both within and across regions. As a result, the proposed GAM model not only predicts the pattern of the 2015 response, it also establishes a foundation for predicting forest responses to future drought events based on location and specific forest properties.

Overall, I find this to be a solid contribution that will promote valuable discussions and advance the field. However, I would like to see a bit more fine tuning of the presentation, including the consideration of a few additional analyses.

> ***Response 3.1:*** *We appreciate the reviewer's recognition of this work as a solid contribution, of potentially groundbreaking results. We appreciate the reviewer's constructive suggestions and we hope the reviewer finds we have carefully addressed them to notably improve our manuscript.*

I agree that there are clear regional differences. Nevertheless, I struggled to be convinced that the evidence presented can be interpreted in the functional biogeography framework suggested by the authors, and that the contribute demonstrates a 'common basin-wide functional biogeography of forest drought response'.

> ***Response 3.2:*** *We appreciate the reviewer's struggle to feel confidence in the functional biogeography framework. We hope that our response to the reviewer's "major points" (more detail of which can be seen below) helps to build such confidence. In sum: (1) we have adopted your suggestions on the integration of some metrics of land use; (2) we have added statements regarding data uncertainty and inference limitations; (3) we also added hypothesis tests, statistical checks and new validation efforts to improve datasets confidence and the interpretations drawn (see details below in specific responses);*

A first major comment for improvement regards the integration of some metrics of land use. If I understand correctly, the current approach assumes essentially equal risk of forest degradation across the studied areas, and we know this not to be the case. I would appreciate an integration of some metrics of land use and land use change that might in fact be correlated with the other metrics presented here. Indeed, the areas of south-central Amazon with deep WTD and shorter trees also correspond to an arc of deforestation and fire risk. A proper treatment of land use among the other factors considered here can only refine the GAM approach.

> ***Response 3.3:*** *Thanks for pointing out a very important issue about the differentiation of forest and degraded forest. As also presented above in our response to reviewer 2 (**R2.3**), our analysis is intended to apply to undisturbed <u>terra firme</u> forests and to EXCLUDE deforested or burned forests, inundated forests, and non-forested vegetation (Andean highlands, cerrado, agriculture, etc). <u>In case this was not clear, we have revised the text (see line 67, main text, and the first line (660) of methods 1.4) to make this clearer.</u>*
>
> *The reviewer's concerns, however (along with analogous ones of reviewer 2, see above **R2.27**(4)) motivated us to look more deeply into the specific question of degraded forests. We realized that although deforested regions are excluded, our forest mask still includes forests in proximity to deforested regions, that, though not deforested, are likely experiencing degradation. Taking advantage of a recently published land cover*

classification that now includes a category for "degraded" forests *<u>(Vancutsem et al. 2021, updated to 2022)</u>*[84]*<u>, we have</u> addressed this issue as "additional Hypothesis 4" (in the new section 3 on "Testing alternative interpretations)". This suggests that partially degraded forests indeed likely have greater drought sensitivity, but also shows that the overall patterns in the results do not depend on whether these "in between" forests are included or excluded:* ***Fig. S15*** *shows the drought responses of pure intact forests, which do not meaningfully deviate from the main Fig. 5A,B (which includes some forests classified as impacted).*

A second major comment for improvement regards the characterization of functional composition across these forests and the interpretations that are then drawn. Clearly, there are fewer data to address at this scale than we would like to have. However, I argue that we do need to clearly present the uncertainty around these data and limitations that can be drawn on any generalizable inferences, particularly when an objective is to achieve a 'common functional biogeography'. I would suggest some clarification on some of the datasets and some tempering of the presentations drawn. I believe that the uncertainties of existing data may actually render the GAM model a conservative test of what we may be able to achieve with this approach. I would just like to see this contribution establish a path forward to improve datasets and refine our interpretations.

Responses 3.4: *Yes, we agree, and we hope that <u>our additions -- e.g. of causal modeling analysis in methods section 2.6; of the new methods section 3 to test additional hypothesis about alternative interpretations; of the additional quantification of uncertainty (as in the regressions of fig. S16-S17); of the new additional validation (e.g. with Tavares et al hydraulic vulnerability data, fig. S18), and caveats to inferences in the main text (see line 139 about anoxia, ~206-211 about a cause for effects of sandy soils, the sequence of questions posed around line 227)</u> - bring the improvement the reviewer asks.*

A first comment regards the definition of regions. I am not completely convinced by the number of regions nor by the justification for their discrimination. The first axis in the PCA (Fig S3a) is pulled by the correlations between the fertility maps (derived from species composition data of Zuquim et al.) and some extremely coarse estimates of community-weighted means of wood density and proportion of Fabaceae (based on some very coarse genus-level attributions of forestry data reported in terSteege et al.). Tree height does not appear to map onto any axis and is not strongly correlated with other mapped indices (Table S2; I would have appreciated as well to have axis loadings and eigenvalues in this table), even though later the Guiana region is defined as having taller trees (L93). Essentially, the proposal appears to establish regions based on climate and soil fertility (as inferred primarily from plant distributions). I can accept that, especially since the GAM refines this along a more robust continuous gradient; but I would like to understand if the Southern Amazon defined here might be nuanced into two or three sub-regions that behave differently. Was this approach attempted. Was an analysis such as PermAnova applied to justify the three region discrimination?

Response 3.5: *We thank the reviewer for the specific suggestions, which we accept: <u>we now include eigenvalues (as a scree plot added as a new panel to fig. S6) and PermAnova statistics (in the legend of fig. S6)</u> to support the three region classification; we hope the reviewer agrees that the most important loadings are already adequately* *represented visually by the original factor axes (of e.g. wood density, soil fertility, etc) plotted on top of the PCA axes.*

As far as the justification for the classification analysis, please see our response to a similar concern of reviewer #2 (Response R2.27, above, especially point 1). With regard to the number of regions: We agree that the Amazon could likely be divided into a greater number of regions, but we chose to classify it into three -- as we now discuss in the methods section 2.2 on classification, lines 823-826, "Given their diversity, Amazon forests could likely be classified into more than three, but we judged that three would be sufficient to capture substantial functional variation, without being so complex as to prevent intuitive understanding.". We hope the reviewer may be persuaded that we here strike the right balance in light of the existing remote sensing literature.

We further agree with the reviewer, that it is the GAM modeling that does the main work of refining this analysis along a robust gradient. We emphasize that the GAM brings the main results of a unified functional biogeography without dependence on or inference from the three region classification analysis. As we now state at the end of methods section 2.2 (lines 850-855) (and clarified also in R2.27)

> *"This three-region classification (which we use to define the regions depicted in the main text figures) is independent of the results (Figs. 3-4) of the basin-wide modeling investigation (described in sections 2.7-2.8 below) because model predictions depend on pixel-pixel variations of environmental factors regardless of what region they are in. However, the three region Amazon is useful for presenting model results because it illustrates how different functional responses emerge from different ecotype regions (as shown in Fig. 4E)"*

A second, related comment regards the definition of functional strategies, especially relative to drought avoidance/tolerance traits. The tree height Lidar dataset represents the best we have for the region, but is tree rooting depth significantly correlated with tree height for tropical trees of this size? Data from cited studies show correlations between shoot size and rooting depth overall, but across the narrow range captured by most of the Lidar dataset, this may not be applicable. I would appreciate more careful discussion that calls for better metrics of drought tolerance/avoidance than the proxies used here.

Responses 3.6: *Yes, Brum et al[35] do in fact show this correlation between tree size and rooting depth in our domain of the central Amazon for trees of this size -- from adult canopy trees to understory and gap trees also represented in the Lidar distributions. We have added more discussion about caveats and nuances of tree height (methods section 1.3, lines 646-655 and methods 3, Hypothesis 5, lines 1173-1196). We have also added explicit drought tolerance traits with the incorporation of HSMs (main texts, lines 275-280; methods section 1.6(b), lines 741-755, fig. S18)..*

MINOR ISSUES

L103 remove qualifier 'rigorously'

Response 3.7*: At this point in the paper (former line 103, currently 123), we are transitioning from straightforward data presentation (Figs. 1-2) to a more systematic and rigorous quantification of sensitivity and prediction, with uncertainty, using the detailed* *combined framework of DAG and GAM statistical modeling (presented in Figs. 3-4). In this context, we think "rigorous" is an appropriate and informative qualifier demarcating the difference in approach, and have elected to keep the word "rigorous" here.*

Fig4 typo Partial EVI prediction	*FIXED.*
FigS3, typos/formatting of minimum and Fabaceae	*FIXED.*

Referee #4 (Remarks to the Author):

The manuscript addresses one of the majoring ongoing debates in tropical biology - how does Amazonia respond to drought events and what are the implications of this for the global carbon cycle. Contrasting results/inferences ranging from green-up to widespread mortality have been published in high impact outlets. Resolving this issue is immensely important to a wide range of disciplines and to humanity.

The "answer" provided by the analyses in this manuscript is, like many things in ecology, it depends on the context. Specifically, there is an important interaction between water table depth, drought duration, soil fertility and the functional ecology of trees.

Response 4.1: *We appreciate the reviewer's recognition of the importance of this work and the concise summary of the findings presented therein.*

The team primarily uses a diverse assemblage of remotely sensed data products. Several important and nuanced conclusions are reached. For example, the previously debated "green-up" or "brown-down" of the Amazon during 2005 is linked to intra-regional differences in water table depth such that it wasn't uniformly a green-up or brown-down. Rather, water table depth is of great significance and spatially variable. This is logical, but not well-documented and greatly helps resolve the debate on how Amazonia will respond to future droughts. Similarly, comparing differences in responses to the 2005 vs 2010 droughts is valuable in the context of light and water availability (i.e. PAR, VPD) and drought duration. Lastly, while I am sure it will cause further debate and consternation, there is compelling evidence that the field collected data by RAINFOR occurred was spatially biased (i.e. occur in regions with deeper water tables and therefore more likely to brown-down).

Response 4.2*: We appreciate the recognition of the diversity of our datasets, the compelling nature of the evidence and the importance of the conclusions.*

In sum, this appears to be compelling evidence that both the green-up and brown-down perspectives were flawed and that both are occurring and (logically) dependent upon environmental context. I should note that I am not an expert in all of the remote sensing data products used in this work, but it appears robust and could be easily reproduced. The functional biogeography of the work is somewhat weaker, from my perspective. I do not doubt the gradients in mean height and wood density. These trends are found in the datasets used and in other studies. Similarly, soil fertility (%P in particular) have strong gradients in this region. That said, using tree height as a proxy for rooting depth is quite rough. Indeed, if you plot root depth vs height on log-log axes you get a strong relationship, but the magnitude of the scatter around the trend line is probably ecologically relevant. That is, two trees 20m tall that differ in their rooting depth by 5m is likely very important in this system. I do not think there is much the authors can do about an issue like this given the data available, but I don't think the height result is as obviously robust as the authors present. The mean wood density trend is robust, but vegetation is far more than just a mean. There is, of course, a lot of variation around that mean. It is unclear whether there is an obviously an east-west trend in variation in wood density in this region (e.g. Swenson et al. 2012 GEB). However, it would be, obviously, incorrect to lump all vegetation in one part of the soil fertility gradient as having a uniformly "fast" or "acquisitive" functional strategy while at the other end they are uniformly slow or conservative. All forests in the region studied likely have species with wood densities ranging from at least 0.4 to 0.8 (i.e. the vast majority of the global range is found locally). What changes is the relative abundances of values across space and some very light and very density wooded species dropping in or out. In sum, the "functional biogeography" here is a rough caricature of region and I would not consider the functional mechanism and biogeography to be pinned down by this work. If a revision were invited, I would encourage the authors to be a bit more clear with respect to how topical/overly-simplified the functional aspects of this investigation are and that within sub-region functional diversity is likely important. Beyond the overselling and simplification of the functional aspects of this work, I find it to be compelling and a quite important step forward towards out understanding of how this critical ecosystem responds to more frequent and devastating drought events in the region that have profound global consequences.

Responses 4.3: *We are grateful that the reviewer finds the work compelling and important. We agree with the caveats raised by the reviewer, and have added more attention to those as the reviewer suggests:*

(a) with respect to assumptions about tree-height/rooting depth correlation, we have added text (section 1.3, lines 646-655),

"However, observations of the tree height-rooting depth allometry are limited, especially in tropical forests (although one study cited here[35] is directly relevant, as it is from central-eastern Amazon uplandforest, conducted during the 2015 drought); this limitation remains a key uncertainty in our ability to confidently attribute variations in drought response to rooting depth, as opposed to canopy height itself, or other (as yet unidentified) correlates of canopy height."

(b) with respect to wood density: wood density did not end up being selected in the GAM statistical model, so our functional biogeography did not implicate assumptions about wood density distributions directly, though the effect of wood density and associated fast-slow spectrum dynamics are likely partly captured by soil fertility which as the reviewer implies, is correlated with wood density (Table S2). On soil fertility, we certainly agree--not all vegetation in a region is uniformly "fast" or "slow", but as an ecosystem ecology study more than a community ecology study, we naturally focus on mean predictions at the pixel scale, even though we understand that those means emerge from sub-pixel scale local or community scale variations (which are of course the focus of our three core hypotheses).

(c) with respect to functional biogeography being a "rough caricature" and overselling/simplification: Although we would not use the term "caricature", we agree that the functional biogeography developed here is broad-scale and simplified, in the positive sense that it identifies simple rules that might underlie and generate the complex drought response patterns we observe at the basin-wide scale. That said, we argue that

an important feature of this study is that the functional biogeography connects these broadscale patterns and predictions to locally observed ecological mechanisms.

We hope that with the addition of evidence supporting our assumptions and conclusions, including:

(i) *For the accuracy of our scaling approach with respect to the effect of water table depth, we now show consistency of relationships between drought response and water table depth across multiple scales, from 40 km down to 30 m (see methods section 3, H1 and fig. S12 and associated text, lines 1099-1115).*
(ii) *We now show that a recently published hydraulic drought tolerance trait dataset (on individual tree hydraulic safety margins -- clearly a highly diverse species associated functional trait that varies tremendously even within the individual plot scale) also propagates to ecosystem scales and supports our Amazon resilience map (see Fig. 5A inset, fig. S18, and methods section 1.6(b) on validation).*

... AND by highlighting caveats that may limit current understanding but also motivate future work (see the new methods section 3, "Testing alternative interpretations and considering caveats", added to more fully address potential reader concerns like those raised here),

... that the reviewer is persuaded that the functional biogeography is not oversold or over-simplified with respect to the scale of prediction it provides, and that it now more strongly offers what the reviewer called a "compelling and a quite important step" towards improved understanding of this critical ecosystem.

Referee #5 (Remarks to the Author):

Key results:
- This paper demonstrates that the spatio-temporal patterns of forest drought response in Amazon basin can be explained by a combination/interplay of three different groups of factors - hydrologic environments, edaphic environments, and tree drought resistance traits, rather than any of them taken separately. The paper clarifies a topic that has been debated in the literature.

- I find the Lines 152-166 especially interesting, as paper succeeded to reconcile virtually disparate data about effects of the 2005 drought.

Response 5.1: Yes, it was gratifying to be able to reconcile the long-standing puzzle of seemingly discrepant responses seen by remote sensing and ground based studies.

Validity:

[All below mentioned flaws are not enough to prohibit publication, however:]

- The first sentence of the article, which is an important and motivating sentence, states: "Three 'once in a century' droughts (Fig. S1) occurred in the Amazon over a single decade..." However, the Figure S1 shows that, although the drought formally occurred "in the Amazon," it drastically occurred in different places within the Amazon each time. Given that individual trees or smaller groups of trees are experiencing drought effects at their actual locations, it appears that the locally-sensed drought was not as frequent as implied by the first sentence. The sentence may be unnecessarily dramatic and sensationalistic. Overlaying the three panels from Figure S1 clearly shows that drastic negative water anomalies were experienced only once in that decade by every individual tree in the Amazon. The authors later hinted at this in lines 177-179.

> ***Response 5.2****: Yes, we agree about the spatial variation in drought locations. However, we are speaking in the same sense as much of the broader literature (including the cited references), which is that the Amazon basin as a whole has not historically seen droughts of this magnitude so frequently. We changed the reference from "the Amazon" to the "the Amazon basin" to make this clearer.*

- It is not typical to start a manuscript by referencing a figure from the supplementary material, yet the first sentence starts with a reference to Figure S1.

> ***Response 5.3****: We agree it might be helpful to combine Fig. S1 with Fig. 1 into a single figure of two rows, one of drought patterns, and one of drought responses-- but we are cognizant of pushing the limits for number of figures in a Nature publication, so we kept main text figures focused on the new results presented here. If the manuscript is accepted and the editors are amenable we would be happy to consider a modification.*

- No link to the github repository with the codes used in the analysis are provided. The methods section contains clear description of the approach at algorithmic level, and ENVI software package (in Python), yet the readers do not know whether the entire development (e.g. PCA) was done in Matlab, R, Python, or something else.

> ***Response 5.4:*** *We will include a repository for codes on publication according to journal guidelines. We have added clarification of methods -- e.g. the PCA was done using R (see line 816).*

- Table S2 shows the "Correlation among climate and ecotype variables." I assume that the measure of correlation used here is Pearson's correlation coefficient. However, it is not the only measure available, and the name of the measure used should be specified. Additionally, the correlations between environmental variables are often non-linear, so the findings based on Pearson's correlation coefficient might be misleading.

> ***Response 5.5****: We have now revised the Table. S2 caption to clarify that the measure used is Pearson's correlation coefficient, and we have also added "variable inflation factors (VIFs)" as a measure ofjoint multicollinearity. As the reviewer notes, both of these measure linear correlations; we now include, in the separate Directed Acyclic Graph (DAG) analysis (methods section 2.6), non-linear dependence checking in DAG analysis that considers non-linear correlations.*

- In Figure S10, the correlation between EVI and SIF is shown to have an R2 of 0.23. It is well established that the average correlation between GPP and GOSIF is close to 0.80, and the correlation between EVI and GPP has been found to vary across biomes but typically lies around 0.5. The value of 0.23 appears to be too low. One possible explanation could be that the drought's impact on activity is greater than its impact on capacity, disrupting the typical relationship. However, I have concerns beyond this. If the value of 0.23 is the mean correlation across all three droughts, would it be possible to provide individual R2 values for each drought, given that the figure presents individual data points for each drought? Additionally, could you calculate the same relationship for a reference period outside of the drought? This is important because the GOSIF data was produced using the Cubist model, which is a single model for the entire Earth and has a cost function based on global mismatch. This non-biome-specific model may sacrifice accuracy for Amazonian biomes to more accurately predict SIF for other biomes. If the relationship outside of the study period is much better, it is likely that the low R2 during the drought is a result of the drought. However, if it remains low, the issues may run deeper. This should have been discussed in the paper.

> ***Response 5.6**: We thank the reviewer for motivating us to clarify the EVI vs GOSIF usage -- see our revised methods section 1.1 and our discussion above in response to reviewer 2 (**R2.25** and **R2.14**).*
>
> *We would like first to offer a clarification important for responding to this reviewer concern: the correlation (R^2 "close to" 0.80) the reviewer is referring to (between GPP and GOSIF) is indeed the global number (reported as 0.73 in the Li and Xiao (2019)[27] paper that developed the GOSIF product). But the evergreen broadleaf forest (EBF) biome (including sites in the Amazon) are reported to have a lower biome specific R^2 of 51% -- not because of sacrificing accuracy in that biome to optimize a global model fit (Li and Xiao developed both the global model to which the reviewer refers and a biome-specific GOSIF, and the EBF R^2 is not different between the global model and the biome specific model), but because evergreen systems without a dormant season (tropical forests) are intrinsically harder to capture compared to systems with a dormant season (dormancy, when GPP '-0, increases the total variability, making it easy to achieve high R^2 in seasonal datasets). With both EVI and GOSIF capturing eddy flux GPP at about 50% in tropical systems (and capturing differing aspects of it), we might even have predicted an EVI-GOSIF R^2 of 0.50x0.50 = '-0.25, approximately what we in fact see (and consistent with the EVI-GOSIF relationships reported in Shekhar et al 2022[61]). We now add a paragraph ("**Note on lower R^2 for Tropical vs Temperate forest GPP detection"**) to the validation section 1.6(a) (around line 717) clarifying this:*
>
> > "*Although both indices (GOSIF and EVI) capture GPP comparably in deciduous broadleaf (temperate) versus evergreen broadleaf (tropical) forests within active growing seasons, most statistical assessments are offull annual cycles, which typically show substantially better statistics (R^2 > 0.8) for temperate zone forests, simply because temperate forests include easily detectable dormant periods when GPP ~ 0, which make total annual variability (hence R^2) higher, while tropical evergreen forests are active year round.*"
>
> *Another way to consider the EVI-GOSIF R^2 of 0.23 is this: Li and Xiao (2019)[27] created GOSIF (to extend the spatially coarse and temporally limited OCO-2 SIF observations to higher resolution and longer time periods) by modeling it as a function of photosynthetic capacity (represented by EVI) and its interactions with climate ((sunlight, VPD, temperature) plus a factor to represent different landcover types (different biomes). In other words, within the EBF, GOSIF = f_{EBF}(EVI, PAR, VPD, temperature), by definition. If R_2 between GOSIF and EVI alone is 0.23, it is just a mathematical consequence of leaving three of the GOSIF defining variables (the climate) out of the analysis (i.e that the remaining 0.77 is due to climate variations independent of EVI).*

In light of these considerations, we don't see this issue (of the modest R^2 between EVI and GOSIF) as an important or deep puzzle in need of explanation, at least not within the context of this paper which has a quite different scope and purpose (though perhaps it may be productively explored in a future remote sensing paper). We ourselves of course first raised this issue by including the EVI vs GOSIF scatterplot (former fig. S10) in the paper, but, considering the reviewer responses, we think we have created a distraction. We apologize!

Therefore, we would like to decline the reviewer's invitation to discuss this matter further in this paper, and instead propose the opposite: that we remove the EVI-GOSIF scatterplot from the paper altogether to avoid similar reader distraction (thus we have deleted the former fig. S10), and to focus instead on our reasons for including both EVI and GOSIF: it is simply to use two different but widely accepted and cited indices of photosynthetic function, and (once some basic validations make clear they are relevant, now shown more clearly in the validation section 1.6), to use them to investigate ecological hypotheses about Amazon forest and drought, with the implication that if both support those hypotheses, that the confidence is greater than ifjust one of them did. We now more clearly emphasize in methods section 1 that GOSIF and EVI represent different aspects of photosynthesis that "bracket the plausible range of forest photosynthetic response to drought" (see section 1.1) and that our finding (see section 1.6(c)) is that both of them do.

Originality and significance:

- The conclusions drawn in the paper are indeed original, but they are also specific to the Amazon ecosystem and may not be easily transferable to other forest biomes. As a result, the paper may be more suitable for publication in a specialized journal. The idea of combining multiple factors to explain patterns in forest drought response, instead of a single one, is not necessarily innovative enough for publication in a top-tier journal like Nature. However, the quality of presentation, the flow and the style keep readers involved. The paper is entertaining and interesting for the wider audience.

***Response 5.7**: We respectfully disagree with the idea that because this analysis focuses on the Amazon, it is not suitable for publication in Nature. First, we disagree that the results are necessarily only specific to the Amazon. We use general ecological and biogeographic principles that should in no way be limited to a particular biome, even if it is beyond the scope of this paper to test our new approach beyond where it is now first being developed, in the Amazon (it certainly suggests a path for follow up work to test the generality of this approach in other biomes). Second, Amazon forests represent a critical, globally impactful biome; analyses of Amazon forest climate change, carbon and water biogeochemistry, biological diversity, forest degradation and other topics are regularly published in Nature and other high profile journals when those analysis reach a level of importance and originality commensurate with those journals.*

- The paper lacks immediate practical significance, as the main contribution relates to processes that are beyond easy and direct human control and are happening over log time periods. Although the article lists two potential outcomes - conservation decision-making and improved predictions for future climate change, it does not provide concrete details on how these outcomes will be achieved or who will implement them. In terms of immediate academic interest, the paper is well-written and meets the standards of top journals in the field.

> ***Response 5.8****: We appreciate the recognition of the clarity and the academic interest of our study. We do not think that the level of practical significance or the identification of who might implement policies is relevant for publication (Nature publishes both fundamental science and practically relevant policy research), but we do agree <u>our study has practical implications which we have now taken more care to highlight (see the second last paragraph beginning around line 304). Fig. S20 highlights the overlap between the most vulnerable forests and (1) the arc of deforestation and (2) the forests most important to sustaining hydraulic recycling relevant to tipping point sensitivity.</u> The policy implications for conservation would be that these particular most-vulnerable forests (important for sustainable function of the biome) should be prioritized.*

Data & methodology:

- The quality of presentation is very high, the techniques are well-established and mature. The quality of data is not questionable, except the concern I expressed earlier around low R2 in EVI/GOSIF. The study offers enough details to allow reproduction.

- The authors have put in a significant amount of effort to gather, standardize, and analyze the various data used in the analysis. The techniques used, such as machine learning, statistics, and geostatistics (variography), showcase their high level of expertise in data processing and analysis.

> ***Response 5.9****: We appreciate your recognition of the presentation's quality, capability of reproduction, as well as the data analysis, and hope our weighing of the EVI-GOSIF issues is well taken.*

Appropriate use of statistics and treatment of uncertainties:

- Principal component analysis (PCA) used for clustering of the Amazon regions and Generalized additive models (GAM) used for the main analysis are well-established techniques, appropriate for this type of research.

Conclusions:
- I find the conclusions and data interpretation are robust, valid and reliable.

> ***Response****: we appreciate the reviewer's assessment.*

Suggested improvements:

- The title of the article could benefit from simplification to increase its readability.

> ***Response 5.10****: <u>We proposed a new title "A biogeography of Amazon forest drought response predicts basin-wide resilience and vulnerability"</u>*

- Consider exploring more advanced modeling techniques, such as NODE-GAM or neural networks, which have been shown to outperform GAMs. Utilizing turn-key solutions, such as Datarobot, may simplify the application of different modeling strategies.

(e.g. https://medium.com/@chkchang21/interpretable-deep-learning-models-for-tabular-data-ne ural-gams-500c6ecc0122; https://www.datarobot.com).

> ***Response 5.11**: We hope the addition of new DAG causal modeling (see methods section 2.6) satisfies the spirit of the reviewer request, even if we did not implement the modeling approach suggested.*

- Consistent labeling of subpanels in plots using letters (e.g. a), b), c), etc.) is necessary for clarity, but this was missing in Figure S10. *FIXED.*

- The title is the Figure S10 is:" Comparison of GOSIF to EVI drought anomalies and to ground-based demography during the 2015/2016 drought". Yet, the plot on the panel A (I guess) shows data for years 2005, 2010 and 2015/2016. The title of overall Figure S10 needs to be aligned with the plot and legend, as there is a discrepancy between the caption and the data shown in subpanel A.

> ***Response 5.12**: The portions of fig. S10 that survive (now in fig. S17) have been redesigned and, we think, are more clearly labeled.*

- The statement in Line 33 that "Here, we combine remotely-sensed photosynthetic vegetation indices" is misleading as only one index (EVI) was used, and GOSIF is not considered a vegetation index in the literature (although, functionally, it is).

> ***Response 5.13**: We agree with the reviewer that "functionally, it [GOSIF] is" an index of photosynthesis, and we have therefore carefully revised references to GOSIF to refer to "photosynthetic index", distinguishing it from the traditional remote sensing reflectance-based "vegetation index" (see line 31) .*

- In Line 239 authors say: "First, the broad consistency between MODIS and GOSIF remote sensing analyses (Fig. 1 and 3A versus fig. S2; fig. S7; fig. S9 versus S10)...". However, Figure S9 does not show any data related to GOSIF. The claim of "broad consistency" between MODIS and GOSIF data is not supported by the R2 value of 0.23 between GOSIF and EVI anomalies (Fig S10).

> ***Response 5.14**: Figures S9 and S10 (now S17) show EVI and GOSIF correlations, respectively, with demography in forest plots on the ground. Our intent is to show the evidence that EVI-demography relations are consistent with GOSIF-demography relations. These plots have now been redesigned to be more parallel and combined into one figure to more clearly show the consistency (fig. S17).*
>
> *Regarding the scatterplot between GOSIF and EVI: following the discussion above (**R5.6**) and in response to a previous reviewer (see **R2.25 and R2.14**), we have removed this scatterplot (in former Fig S10).*

- The Method section is overly lengthy with over 4500 words and could be significantly reduced to meet the recommended size for publication in Nature (3000 words maximum).

> ***Response 5.15**: If the manuscript is accepted for publication, we will be happy to follow the guidance of the journal on this point (e.g trimming or moving more details to the supplement). In the meantime, for purposes of convenient review of the many points raised by reviewers, we have kept a more detailed methods section.*

References: The manuscript references previous literature appropriately. However, top tier journals are mainly referenced in the beginning, to set up the context of the study, while other sections mainly referenced other, more specialized journals. It points to the fact that either the work presented in the paper is original up to the point that nothing similar has ever been published in the top tier journals, or that topic-wise it belongs to more specialized journals.

> ***Response 5.16****: We suspect it probably points to the fact that the beginning (introduction) of an article is typically where broad questions are framed (often the focus of top-tier journal articles), whereas other sections are where specific issues or methods are addressed (often the focus of more specialized journals).*

Clarity and context:
- Abstract is clear, accessible and contains concise and relevant information and appropriate references.
- Introduction and conclusions are appropriate, at least within the research framework and choice of methods that have been applied to the problem.

Inflammatory material: Does the manuscript contain any language that is inappropriate or potentially libelous?

- No.

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

The three main concerns that I raised in my initial review related to 1) the predictive analytical approach used in combination with causal inference objectives and language, 2) the lack of a clear definition of “resilience”, and how it is assumed to be related to the remotely-sensed photosynthetic vegetation indices, and 3) the absence of soil texture in the study, in spite of its potential relevance during droughts for soil water retention capacity.

In the revised manuscript, the authors went to great lengths to address these three points by using a DAG and adopting a causal framework to ground the scientific models, by defining resilience and presenting clear physiological mechanisms to support their definition, and including soil texture in their analyses. I also appreciate the details with which they responded to each comment.

I still have one important concern, regarding what seems to me like an incompatibility between the new causal analysis framework and the use of AIC and a “dredging” model selection approach (see details below). I also present a number of minor and addressable points, mostly regarding the clarity and communication of the data analysis throughout the manuscript. Below, my review structured around the initial three major points I had raised.

I. Data analyses and causal inferences

Comments about the content of the new text in section 2.6 of the Methods:
I am very pleased to see that the authors made important and substantial changes to their analytical methodology to follow the suggestion to use a structural causal modeling approach (what they refer to as “DAG analysis”) to justify a causal interpretation of the mechanisms the study aims to understand.
They created a causal diagram (DAG) of their initial ecological assumptions regarding the studied system, tested the DAG-data consistency (causal discovery) to improve the initial DAG, then used the backdoor criterion to define a set of control variables necessary to allow a causal effect quantification for each environmental variable of interest.
This truly makes for a more formal and now justified causal interpretation – conditional on the DAG being true. Presenting the DAG will allow readers to not only take the authors’ word, but to also assess (and potentially improve in future works) the causal assumptions upon which the study’s conclusion rely. An important step towards a more reproducible and transparent science.

I still have a few comments to further improve / correct some more minor aspects of the DAG-related changes in the revised version:

Page 5 – Lines 18-22 (and section 2.6 of the Methods): A few corrections are needed. I suggest speaking of Causal Structural Modeling (SCM) approach instead of “DAG analysis”, which is vague and does not encompass all the steps that were used here. In addition, I would be careful with the

wording when saying “to identify likely causal variables and relations”, as this is not an exploratory analysis where one arrives without knowing what will pop up as being a causal variable, and finishes having discovered them. Rather, the authors decide the variables for which they want to know and quantify the causal effect (reflecting the scientific questions and objectives of the study), and they build (and improve, using causal discovery) a causal diagram to represent the whole system – including their variables of interest – to then define the control variables needed in a statistical model in order to block all non-causal paths and allow a causal interpretation of the coefficient for one variable of interest.
While the authors went through the right steps in their analysis, the explanation they provide needs improving to reflect that their approach aims to justify a causal understanding of variables of interest. And that this approach – Structural Causal Modeling approach – consists in 1) creating a DAG, 2) testing the DAG-data consistency and improving the DAG using field-specific knowledge, 3) backdoor criterion to define control variables needed for each variable of interest.

I also advice to clearly state in the methods as well as in the main text that causal inferences in the study are conditional on the causal diagram being true. While this is always the case, it will not be obvious to many readers, and making this explicit helps stressing the importance of the DAG and associated causal assumptions to the conclusion of the work, and for future works that would aim to build on assumption and results generated here (McElreath, 2020; Barnard-Mayers et al., 2022).

Citations are missing and need to be added when introducing the structural causal model framework used in the study, mostly Judea Pearl’s work (Pearl, 2009a,b), ideally also the recent extensive work on that team, applied to Ecology for observational data like this one (Arif & MacNeil, 2022a).

Two points related to the above, i.e. the way the causal framework is introduced and referred to throughout the text:

- (1) In Fig. S9 legend, figure C, the authors wrote “Final DAG constructed from the complete set of detected causal links”. This needs rephrasing, as causal links are not detected (or detectable) based on the causal discovery step that the authors applied to the initial DAG. Rather, causal links are hypothesized based on field-related theory and observed associations among variables. The DAG of Fig. S9C therefore is the DAG corrected based on ecological considerations coupled to the conditional independency inconsistencies of the initial DAG.
Regarding this, the authors say they reconciled the initial DAG with the inconsistencies in the DAG-data consistency test by adding direct causal links or associations through common causes, however they never mention whether they based these added associations in the updated DAG based on existing theory and scientific mechanisms. This is crucial to do, and if done to explicitly say, because simply adding links to a DAG so that all its conditional independencies (highlighted by the causal discovery step) are coherent with the DAG is not a valid way of reaching a final DAG (McElreath, 2020; Ankan et al., 2021). Rather, one needs to try and reconcile the inconsistencies between the initial DAG and the causal discovery step by thinking of mechanisms and scientific assumptions. Here, it is unclear whether this is indeed what the authors did or not.
Ideally, all arrows in a DAG would be justified. If not done, at the very least, it is essential to know that the authors used scientific thinking – not just solving problematic statistical associations – to reach a final DAG.

- (2) Methods, section 2.6: “DAG analysis” is too vague a term, and is not “a form of structural equation modeling”. A DAG is a heuristic graphical tool used by structural equation modeling, and more broadly by structural causal modeling (Pearl, 2009a; Arif & MacNeil, 2022). Both “analyse” the DAG, but more specifically they use its structure to guide decisions regarding the statistical models to use (use of backdoor or frontdoor criterion, for example). Please correct.

Section 2.7 of Methods:
The “DAG-inferred causal network” is an odd and incorrect wording. The DAG is the causal network, and the reason for the “inferred” part here is unclear.

More importantly (and this refers to the remaining major concern I mentioned at the beginning), while the authors begin the section saying the inclusion of covariates in the model was grounded in the use of the backdoor criterion applied to the DAG, to allow causal inference from the model output, they then continue to use a dredging approach (which is the exact opposite of a causal framework) and R2 and delta AIC to “select a model”. This is clearly problematic, as one either puts causal inference first, or the selection of a best predictive model (that does not allow causal inference) (Arif & MacNeil, 2022b,a). If one defines a set of control variables necessary for causal inference of a variable X on Y, then using AIC to change that set of control of variables may improve out-of-sample expected predictive accuracy, but is likely to not allow the causal inference anymore, in case the AIC selects a different set of covariates than the one obtained by applying the backdoor criterion. And cases where the backdoor criterion and an information criterion lead to different subsets of covariates are very common (e.g. McElreath, 2020; Arif & MacNeil, 2022b).
How do the authors respond to that? Did I misunderstand something in that section? If so, then at the very least, some clarifications are likely needed in the explanation of the section. If the objective for this analysis is prediction regardless of any causal interpretation, that would be fine, but this needs to be very clearly stated both in the main text prior to presenting those results, as well as in the corresponding Methods section. My understanding for now is that the authors added the causal framework to the study but maintained part of their previous approach, without realizing they were incompatible. If causal inference is the objective, the authors can simply use the set of variables defined by their DAG and the backdoor criterion to each region, keeping models a, b, c, and d, but removing the model selection part. The selected model is selected by the backdoor criterion, simply. If maximizing prediction is the objective, then there is no need to mention the DAG and resulting sets of covariates from the backdoor criterion. I stress the important, if both causal inference and maximizing predictions are used in different analyses in the same study, that the results in the main text should very clearly differentiate the two and clarify the objective of the two approaches, to avoid confusion for the reader.

Thought regarding the DAG:
More clayey soils tend to have higher effective cationic exchange capacity or base cations, leading to higher soil fertility. I would say this warrants a causal link from SoilTexture to SoilFertility, wouldn’t it?

I would also expect precipitation anomalies to partly cause VPD anomalies (moister conditions (negative precip. anomalies) will cause higher relative humidity and therefore lower VPD anomaly). I see that this ended up in the final DAG, and while this is not presented I am guessing the causal thinking behind that arrow is along the line of what I mention here?

Suggestion: I would mention the use of a causal analysis framework in the abstract. This is a strong point of the study, which many surely will greatly appreciate.

In summary, the authors did a great job improving the data analyses of the work and ensuring to be able to use them to answer the scientific questions of the study. An important remaining concern for them to address is the incompatibility between their causal framework and causal inference objective, and the use of the "dredging" and maximization of an AIC to select a model (see above). Finally, a last extra effort is still needed to improve and correct the explanation of their approach in the main text, associated methods section, and figure legend.

II. Definition of resilience
The authors now very clearly define "resilience" while acknowledging other possible definitions, which greatly clarifies the scope of their objectives and questions. I like the new paragraph they wrote in Method section 2.4, where they further lay out the physiological thinking behind their definition.

III. Soil properties
Soil fertility needs to be defined upon first use in the introduction (mention that is refers to exchangeable cations), to avoid needing to go the Methods to known what "fertility" here refers to (e.g. not related to nitrogen or phosphorus).

In the Methods, please specify the range of soil depths used for the variables downloaded from SoilGrids.

I suggest removing the use of "soil types" terminology, as it suggests discrete soil categories, but actually refers to two continuous variables considered independently – soil texture and fertility. I would simply mention "soil texture and fertility". It will allow only using "types" when referring to the ecotopes (which here consists of actual categories).

I hope the above comments and suggestions will help the authors further improve and clarify their manuscript. This is a very valuable, interesting, and topical subject and set of results.

David Bauman

References

Ankan A, Wortel IMN, Textor J. 2021. Testing Graphical Causal Models Using the R Package "dagitty". Current Protocols 1: e45.
Arif S, MacNeil MA. 2022a. Applying the structural causal model framework for observational causal inference in ecology. Ecological Monographs: 1–18.
Arif S, MacNeil MA. 2022b. Predictive models aren't for causal inference. Ecology Letters 25: 1741–1745.

Barnard-Mayers R, Kouser H, Cohen JA, Tassiopoulos K, Caniglia EC, Moscicki A-B, Campos NG, Caunca MR, Seage GRS, Murray EJ. 2022. A case study and proposal for publishing directed acyclic graphs: The effectiveness of the quadrivalent human papillomavirus vaccine in perinatally HIV Infected girls. Journal of Clinical Epidemiology 144: 127–135.
McElreath R. 2020. Statistical Rethinking (Second Edition): A Bayesian Course with Examples in R and Stan. Abingdon: CRC Press.
Pearl J. 2009a. Causal inference in statistics: An overview. Statistics Surveys 3: 96–146.
Pearl J. 2009b. Causality: models, reasoning, and inference. Second Edition. Cambridge University Press.

Referee #1 (Remarks on code availability):

I was not able to review the R code underlying the study's analyses as it was not provided.

The revised manuscript states "Code for reproducing the modeling analysis and graphs is posted at XXX.".

I therefore can't evaluate whether the code content and its clarity (adapted level of comments and explanation to accompany the coding) are sufficient to ensure the sufficient transparency and reproducibility of the study.

Referee #2 (Remarks to the Author):

In the revision the authors have done a thorough job in responding to the points raised in the previous reviews and have taken new approaches to bolster their arguments. I found that the responses to the rebuttal to be more convincing than what appeared in the actual revision.

The presentation and writing in the main body of the paper makes the information difficult to evaluate, with the writing and presentation being inconsistent, and full of multi-clause sentences with much of the information being stuck into parenthetical remarks. They hypotheses are not well stated and they are inconsistently used and referred to throughout the text.

A lot of methods that are specific are still left in the text, while a very clear, crisp overview of the study design that might go well in the introduction is in the methods. In particular L516-527, and really L516-544 are important. In the main body of the paper (before the methods) the analysis and sources of data that the conclusions rely on are multiple and complex, and it is not easy to follow the authors' reasoning as to why one approach was chosen over multiple others. Having that first section of the methods early on would really help the reader to understand what is being done, the overarching design, etc. The writing is terse and informative too.

It may be that this paper is just too complex for adequate presentation in the Nature short form. Sometimes studies are elegant and convincing in total, but can't be convincingly boiled down to short format without losing the key elements of the argument and support for the choices made. It might be easier to publish this in a long format, still top-tier journal, and do the work full justice.

As it stands, the first section still seems a bit haphazard in what it includes and what it leaves out, compared to the nice, full discussion of the ecology and analysis choices that reveals it that comes in the methods. If I were to make a recommendation I might say to move even more detail to the methods and focus hard on clear, concise statement of the results and discussions of their implications. The writing in the methods is clear and rock solid. The main body of the paper needs to follow this tone and directness, leaving out anything that is unessential for the main points, keeping detailed methods in the methods section. Go for the story in its clearest, most streamlined way. The authors have worked on this, but it still required editing and tightening.

Specific comments:

L26: Capitalize E in Earth (referring to planet)

L34: Cut "importantly"

L36: Maybe response rather than responsiveness for clarity?

L39-40: This sentence is unclear--is the cause hardier trees or tall forests with deep roots? Are you talking about the same area with alternative explanations or different areas with different explanations?

L61: End with period, maybe "increased mortality"

L65: "...and test". More generally, that sentence is a monster with many clauses and ideas--maybe break into a couple sentences for clarity?

L69: The other side of drought hypothesis needs to be introduced--move to end of sentence so the reader has some idea what it means? It is not self-evident. You can even say, "This hypothesis is termed the other side of drought hypothesis because...."

L80: rewrite that clause for clarity

L83: and throughout. The manuscript could still use a good editing to streamline and reduce wordiness/increase clarity. Especially strongly realized-->strong? Or Especially strong, prevalent, etc.?

L86: : --> .

L88: Large seed mass doesn't seem to be important per se for canopy response to drought

L110: I would put the idea that you are using three droughts to test three spatially segregated regions with distinct ecotypes up in the paragraph ~L62 where you lay out what you are going to do in the paper.

L123: In the writing a lot of important, stand-alone information is put into parenthetical comments within clauses. I'd suggest not using this construction and integrating the information into the sentences proper, and in some cases breaking sentences for clarity. In this paragraph it seems like some of the detailed methods should be moved to methods.

The sentence on L195 is another example of this—important information for evaluating or even understanding the full hypothesis is put into a multi-clause sentence with parenthetical inclusions.

L151: Set up the importance of sunlight anomaly and define

L233: Citation? Tall trees are more resistant to drought? I thought it was the opposite.

L260: I would change the OR to AND as the sentence is talking about how these patterns emerge when the three datasets are combined.

Figure 3: You might put some interpretation into the legend as well. The technical description is good, but this is still a very complicated figure with many assumptions going into the choices of what is shown, so the reader would benefit by a clause at the end of each section saying something like, "....showing" and then saying what the authors summarize as the important points for the reader to extract in order to evaluate their hypothesis. Figure 4 accomplishes this with the additional in-figure labels that help orient the reader, though it is still complex.

L301: This is a big result! Should be in abstract.

Referee #3 (Remarks to the Author):

I will lead with an apology if delays in reviewing were due to my holding the revision for an extended period. I will underline that a large part of this delay was due to the comprehensive work accomplished by reviewers as well as the very detailed and comprehensive responses provided by the authors. It was a pleasure and an educational experience to read through and digest the work that has been accomplished.

I am extremely satisfied with responses to all of my inquiries, and I appreciate the inquiries and responses from all of my colleagues. We could continue dialogue over some of these issues, but I believe that is best continued in the public literature, where others can profit from the value of the work achieved. I hope that we all find this experience to be a model for peer review (even if it involves delays), and I am very pleased with the solid result, which I believe will have a great deal of impact.

I do provide a couple of minor points for polishing the revision below by clarifying some language and tempering some framing.

Best wishes to all involved,
Chris Baraloto

L65 - might be worth underlining early on that these are not exclusive

L95/L128 - suggest adding something about the spatial scales at which these three hypotheses and their synergies may behave and the scales at which you will conduct the analyses across the landscape

L111 - suggest tempering the language of 'distinct' regions, as described in response to reviewers - they are distinct in large part because they permit simplification of interpretations

L123 - bravo to the articulate Bauman suggestion of DAG, but not sure this merits the employ of 'rigorous'

L193 - suggest tempering this proxy (a convenient generalization) as my read of these references does not justify a broad generalization of tree height-rooting depth for such a strong inference. For example, I can think of a few large-statured taxa that are relatively common in deep water tables and yet in my experience have relatively low belowground biomass.

L206 - soil texture integration is useful, but this reads as an add-on response rather than integration. what about vegetation effects - extreme sandy soil areas have lower statured forest, and Guianan forests are highly sandy but support high-statured, high biomass stock forests (see L214 after)?
L222 - bis for soil texture, what happens in deep water tables on more sandy soils? what about extreme sands?

L294 - this can be rephrased for clarity 'strategy of growing trees fast…'

L370 - change ecotype to ecotope for consistency (also thereafter in Fig legends and annexes)

L611/L784- it might be useful to add a small table on the spatial resolution of each of the different datasets integrated herein; at least add for all when they first appear in text (eg, soil texture).

L810 - was classification analysis re-run using soil texture during revision?

Referee #4 (Remarks to the Author):

I was, I believe, Rev 4 from the last submission. As I voiced last time, I find this to be an interesting and important advance regarding green-up/brown-down dynamics in the Amazon and provides a logical and compelling insight.

In their revisions, the authors addressed most of my concerns fairly well. I was not totally satisfied with falling back on this being an "ecosystem" study and not community ecology as a reason why the functional biogeography was so basic. Functional diversity is key to functional biogeography and not simply a community ecology issue. So, I still kind of find the functional biogeography aspect of the study to be more window dressing than a key advance. It will be interesting to hear if Rev 3 still feels similarly. All of that said, I don't think my concern about this issue is super critical as I think the core results and findings are important enough on their own.

I think the team largely addressed the concerns by the other reviewers, but I will leave that up to those reviewers to weigh in. I will note that DAGs are important and outstanding tools, but I think we are fooling ourselves to think they mean we can now say our studies are truly causal. They aren't. Our DAGs are only as good as what is in them in the first place and we are still dealing with large (and inherently messy) datasets. I think adding DAGs was a wise suggestion and I'm pleased the team included this approach, but I don't consider this work (or any similar work using DAGs) to have causal certainty.

Referee #5 (Remarks to the Author):

You substantially revised the paper, addressing both my comments and comments by other reviewers. I believe your manuscript now meets the standards required for publishing.

Author Rebuttals to First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

The three main concerns that I raised in my initial review related to 1) the predictive analytical approach used in combination with causal inference objectives and language, 2) the lack of a clear definition of “resilience”, and how it is assumed to be related to the remotely-sensed photosynthetic vegetation indices, and 3) the absence of soil texture in the study, in spite of its potential relevance during droughts for soil water retention capacity.

In the revised manuscript, the authors went to great lengths to address these three points by using a DAG and adopting a causal framework to ground the scientific models, by defining resilience and presenting clear physiological mechanisms to support their definition, and including soil texture in their analyses. I also appreciate the details with which they responded to each comment.

I still have one important concern, regarding what seems to me like an incompatibility between the new causal analysis framework and the use of AIC and a “dredging” model selection approach (see details below). I also present a number of minor and addressable points, mostly regarding the clarity and communication of the data analysis throughout the manuscript. Below, my review structured around the initial three major points I had raised.

I. Data analyses and causal inferences

Comments about the content of the new text in section 2.6 of the Methods:
I am very pleased to see that the authors made important and substantial changes to their analytical methodology to follow the suggestion to use a structural causal modeling approach (what they refer to as “DAG analysis”) to justify a causal interpretation of the mechanisms the study aims to understand.
They created a causal diagram (DAG) of their initial ecological assumptions regarding the studied system, tested the DAG-data consistency (causal discovery) to improve the initial DAG, then used the backdoor criterion to define a set of control variables necessary to allow a causal effect quantification for each environmental variable of interest.
This truly makes for a more formal and now justified causal interpretation – conditional on the DAG being true. Presenting the DAG will allow readers to not only take the authors’ word, but to also assess (and potentially improve in future works) the causal assumptions upon which the study’s conclusion rely. An important step towards a more reproducible and transparent science.

RESPONSE R1.1: *We thank the reviewer for motivating the addition of causal modeling to our analysis and are gratified that the results make more justified causal interpretation.*

I still have a few comments to further improve / correct some more minor aspects of the DAG-related changes in the revised version:

Page 5 – Lines 18-22 (and section 2.6 of the Methods): A few corrections are needed. I suggest speaking of Causal Structural Modeling (SCM) approach instead of "DAG analysis", which is vague and does not encompass all the steps that were used here. In addition, I would be careful with the wording when saying "to identify likely causal variables and relations", as this is not an exploratory analysis where one arrives without knowing what will pop up as being a causal variable, and finishes having discovered them. Rather, the authors decide the variables for which they want to know and quantify the causal effect (reflecting the scientific questions and objectives of the study), and they build (and improve, using causal discovery) a causal diagram to represent the whole system – including their variables of interest – to then define the control variables needed in a statistical model in order to block all non-causal paths and allow a causal interpretation of the coefficient for one variable of interest.
While the authors went through the right steps in their analysis, the explanation they provide needs improving to reflect that their approach aims to justify a causal understanding of variables of interest. And that this approach – Structural Causal Modeling approach – consists in 1) creating a DAG, 2) testing the DAG-data consistency and improving the DAG using field-specific knowledge, 3) backdoor criterion to define control variables needed for each variable of interest.

RESPONSE R1.2: *We are gratified that we are in agreement with the reviewer that the right steps were taken in the analysis. We appreciate the reviewer's suggestions to further improve the presentation with suggested corrections. <u>We now introduce the causal modeling approach by the name Structural Causal Modeling (SCM) (rather than "DAG analysis"), and removed language implying that causes are identified or detected by SCM</u> (see paragraph beginning at line 115). We also <u>restructured our statistical methods sections (now all section 2.6) to better clarify these issues (See also Response R1.8, below).</u>*

I also advice to clearly state in the methods as well as in the main text that causal inferences in the study are conditional on the causal diagram being true. While this is always the case, it will not be obvious to many readers, and making this explicit helps stressing the importance of the DAG and associated causal assumptions to the conclusion of the work, and for future works that would aim to build on assumption and results generated here (McElreath, 2020; Barnard-Mayers et al., 2022).

> ***RESPONSE R1.3****: We understand and agree with the reviewer that causal inferences (e.g. of effect sizes) and model results (whether from structural causal models or predictive regression models) are conditional on the models being true. We believe this is implied by the structure of the hypothesis testing framework which pervades the logic of the paper, and for this reason -- and because of the pressing need for brevity -- we were not able to include a separate statement to this effect in the main narrative, but <u>we now also include such an explicit statement in the methods section</u>, where it is appropriate to spell out underlying assumptions in greater detail. See lines 957-958.*

Citations are missing and need to be added when introducing the structural causal model framework used in the study, mostly Judea Pearl's work (Pearl, 2009a,b), ideally also the recent extensive work on that team, applied to Ecology for observational data like this one (Arif & MacNeil, 2022a).

> ***RESPONSE R1.4****: We have adjusted the references to SCM to be both complete for sourcing the analysis we do, and parsimonious. We have newly added Arif & MacNeil, 2023 and Pearl J. 2009a; we agree we should have cited this previously, as it was a particularly helpful guide for our own DAG analysis. We also continue to cite the following references:*
>
> - *Pearl 2009 (which we also cited in our previous submission, endnote 113) as a foundational contributor to the field of SCM*
> - *Textor et al 2016 and Ankan et al 2021 (previously cited, endnotes 114, 115) -- as references for the R package 'dagitty' which we used to implement the DAG analysis*
> - *Arif & MacNeil, 2022 (cited in previous endnote 111, to highlight biases that may arise in inferences from predictive models;*
> - *Addicot et al 2022 (previously cited in endnote 112) -- which augments Arif & MacNeill on the hazards of inference from predictive models, and suggests a "solution space" where both approaches have value.*

References:

- *Addicott ET, Fenichel EP, Bradford MA, Pinsky ML, Wood SA. 2022. Toward an improved understanding of causation in the ecological sciences Front. Ecol. Environ.*
- *Ankan, A., Wortel, I. M. N. & Textor, J. Testing graphical causal models using the R package 'dagitty'. Curr. Protoc. 1, e45 (2021)*
- *Arif S, MacNeil MA. 2022. Predictive models aren't for causal inference. Ecology Letters 25: 1741–1745.*

- *Arif S, MacNeil MA. 2023 (nb formerly cited as 2022). Applying the structural causal model framework for observational causal inference in ecology. Ecological Monographs: 1–18.*
- *Textor J, van der Zander B, Gilthorpe MS, Liskiewicz M, Ellison GT. 2016. Robust causal inference using directed acyclic graphs: the R package 'dagitty', Int. J. Epidemiol.*
- *Pearl J. 2009a. Causal inference in statistics: An overview. Statistics Surveys 3: 96–146.*
- *Pearl J. 2009b. Causality: models, reasoning, and inference. Second Edition. Cambridge University Press.*

Two points related to the above, i.e. the way the causal framework is introduced and referred to throughout the text:

- (1) In Fig. S9 legend, figure C, the authors wrote "Final DAG constructed from the complete set of detected causal links". This needs rephrasing, as causal links are not detected (or detectable) based on the causal discovery step that the authors applied to the initial DAG. Rather, causal links are hypothesized based on field-related theory and observed associations among variables. The DAG of Fig. S9C therefore is the DAG corrected based on ecological considerations coupled to the conditional independency inconsistencies of the initial DAG.

> ***RESPONSE R1.5:*** *We have rephrased the Fig. S9C (now Extended Data Fig. 9C) legend (and elsewhere) to avoid implying that causes are detected or detectable by SCM.*

Regarding this, the authors say they reconciled the initial DAG with the inconsistencies in the DAG-data consistency test by adding direct causal links or associations through common causes, however they never mention whether they based these added associations in the updated DAG based on existing theory and scientific mechanisms. This is crucial to do, and if done to explicitly say, because simply adding links to a DAG so that all its conditional independencies (highlighted by the causal discovery step) are coherent with the DAG is not a valid way of reaching a final DAG (McElreath, 2020; Ankan et al., 2021). Rather, one needs to try and reconcile the inconsistencies between the initial DAG and the causal discovery step by thinking of mechanisms and scientific assumptions. Here, it is unclear whether this is indeed what the authors did or not.
Ideally, all arrows in a DAG would be justified. If not done, at the very least, it is essential to know that the authors used scientific thinking – not just solving problematic statistical associations – to reach a final DAG.

> ***RESPONSE R1.6:*** *We agree and have added text and illustrative references to justify arrows added to the DAG.*

- (2) Methods, section 2.6: "DAG analysis" is too vague a term, and is not "a form of structural equation modeling". A DAG is a heuristic graphical tool used by structural equation modeling, and more broadly by structural causal modeling (Pearl, 2009a; Arif & MacNeil, 2022). Both "analyse" the DAG, but more specifically they use its structure to guide decisions regarding the statistical models to use (use of backdoor or frontdoor criterion, for example). Please correct.

RESPONSE R1.7: *Corrected*

Section 2.7 of Methods:
The "DAG-inferred causal network" is an odd and incorrect wording. The DAG is the causal network, and the reason for the "inferred" part here is unclear.

More importantly (and this refers to the remaining major concern I mentioned at the beginning), while the authors begin the section saying the inclusion of covariates in the model was grounded in the use of the backdoor criterion applied to the DAG, to allow causal inference from the model output, they then continue to use a dredging approach (which is the exact opposite of a causal framework) and R2 and delta AIC to "select a model". This is clearly problematic, as one either puts causal inference first, or the selection of a best predictive model (that does not allow causal inference) (Arif & MacNeil, 2022b,a). If one defines a set of control variables necessary for causal inference of a variable X on Y, then using AIC to change that set of control of variables may improve out-of-sample expected predictive accuracy, but is likely to not allow the causal inference anymore, in case the AIC selects a different set of covariates than the one obtained by applying the backdoor criterion. And cases where the backdoor criterion and an information criterion lead to different subsets of covariates are very common (e.g. McElreath, 2020; Arif & MacNeil, 2022b).

How do the authors respond to that? Did I misunderstand something in that section? If so, then at the very least, some clarifications are likely needed in the explanation of the section.

RESPONSE R1.8: *We have adjusted the wording -- see lines 115-124 (main text), and 262-264 where conclusions from the models are made -- to make it clear that we now use the two different approaches, completely independently*: ***(1)*** *AIC-selected GAM predictive modeling (for both inference and to identify best predictive models) -- since making predictions is also an important goal of the analysis AND,* ***(2)*** *SCM Directed Acyclic Graph modeling (which also uses GAM analysis to test causes hypothesized by the DAG, blocking confounding variables and their associated causal backdoor paths).*

To better clarify these issues, <u>we have also restructured our statistical methods sections, combining the former sections 2.6 (causal modeling) and 2.7 (predictive modeling) into one statistical section with three sub-sections: 2.6.1 (GAM modeling), 2.6.2 (SCM modeling), and 2.6.3. (comparing the two)</u>. The intro paragraphs (940-961) to the modeling section 2.6, aimed at users of either approach, now clarifies the need for both approaches, and clarifies how AIC-selected predictive models can also be used for inference when used within an appropriately designed hypothesis testing framework.

We agree with the reviewer that the approach and presentation needed to be improved to avoid implying that AIC-selected regression modeling is linked to the DAG analysis (as mentioned above, it isn't), but we don't think there is any conceptual error in independently conducting both analyses -- and indeed, <u>we now state more explicitly in the modeling methods introductory paragraph (section 2.6)</u> that including both approaches is valuable because: (1) our goals include both causal inferences and prediction; and (2) if both approaches agree on their inferences about the results of hypothesis testing, that increases confidence in the outcome, especially considering that within our community are statistical modelers who are more inclined to/familiar with one or the other approach, and we hope to be compelling to both audiences.

In particular, <u>we find that both the SCM-DAG-GAM causal models and AIC-selected GAM regression models imply consistent conclusions about the hypotheses being tested</u> (method section 2.6.3, and main text lines 262-264). We hope the reviewer will agree that using both approaches to modeling increases overall confidence in the outcome.

If the objective for this analysis is prediction regardless of any causal interpretation, that would be fine, but this needs to be very clearly stated both in the main text prior to presenting those results, as well as in the corresponding Methods section. My understanding for now is that the authors added the causal framework to the study but maintained part of their previous approach, without realizing they were incompatible. If causal inference is the objective, the authors can simply use the set of variables defined by their DAG and the backdoor criterion to each region, keeping models a, b, c, and d, but removing the model selection part. The selected model is selected by the backdoor criterion, simply.
If maximizing prediction is the objective, then there is no need to mention the DAG and resulting sets of covariates from the backdoor criterion. I stress the important, if both causal inference and maximizing predictions are used in different analyses in the same study, that the results in the main text should very clearly differentiate the two and clarify the objective of the two approaches, to avoid confusion for the reader.

RESPONSE R1.9: *As indicated above (R1.8), including both approaches, separately, is indeed the approach we have adopted here, <u>adjusting the wording to "clearly differentiate" the two, clarifying that the objectives of SCM (testing of causal hypotheses) and AIC-selected GAM prediction (identifying best models for predicting drought response, as well as testing hypotheses)</u>. (lines 116-120) We understood this to also be the recommended approach of the reviewer in his first review, where he has written (Reviewer 1, section I.(2)(b)):*

> "The predictive approach – AIC-based model selection following "dredging" – used here could be maintained for the purpose of prediction only...
>
> In this sense, the predictive model would not especially serve the interpretation of the detail inside the model, but would show the added predictive value of HAND, tree height etc, while the models structured based on a causal model (e.g. a DAG) would allow drawing causal inferences in terms of mechanisms."

From this most recent review, It appears the reviewer may be of the view that AIC-selected prediction models should never be used, at all, for hypothesis testing. Without taking a side in such a debate, we suggest that the best practice for this paper is to include *both* approaches, implemented separately (as indicated above). The reviewer and the literature he cites makes a compelling case that SCM is a sound, and perhaps better-founded, approach to causal inference with observational data, and we are happy to have learned about it and to use it. At the same time, there is a large and well-developed literature on using AIC-selected regression models in hypothesis testing (not just prediction), at least some of which is well-versed in the hazards of blind application of AIC selection to 'kitchen sink' models (Clark 2007; Burnham et al 2011). There is a community that uses and understands that approach.

By using both approaches for testing causal hypotheses (and finding them consistent) we may better speak to both communities, and hence reach a broader audience, with our analysis. If it is true that inferences derived from AIC-selected regressions are not well-founded, then we are right to rely on the SCM inferences. At the same time, those who use AIC-regression and are unfamiliar with the newer SCM will learn about the causal approach and in the meantime not be far wrong in taking the AIC-regression inferences at face value (since they are consistent with those of SCM).

Thought regarding the DAG:
More clayey soils tend to have higher effective cationic exchange capacity or base cations, leading to higher soil fertility. I would say this warrants a causal link from SoilTexture to SoilFertility, wouldn't it?

RESPONSE R1.10: *We have indeed considered this causal link between soil texture and soil effective cation exchange capacity, though the association between them is quite weak (r=0.12~0.16). We have added this link to the DAG (Extended Data Figs. 9-10E).*

I would also expect precipitation anomalies to partly cause VPD anomalies (moister conditions (negative precip. anomalies) will cause higher relative humidity and therefore lower VPD anomaly). I see that this ended up in the final DAG, and while this is not presented I am guessing the causal thinking behind that arrow is along the line of what I mention here?

> ***RESPONSE R1.11***. *Yes, that was exactly our thinking (except correcting the sign: <u>positive</u> precip anomalies -> higher relative humidity, etc), and we have updated the text there to clarify this (Line 1066-1067).*

Suggestion: I would mention the use of a causal analysis framework in the abstract. This is a strong point of the study, which many surely will greatly appreciate.

> ***RESPONSE R1.12:*** *We are excited by the opportunity to add SCM to this paper, but are inclined to respectfully decline the reviewer's invitation here, only because the abstract is already slightly in excess of the suggested word limit, and we think it important to prioritize for Nature the substantive conclusions from the statistical analysis rather than the analysis methods themselves.*

In summary, the authors did a great job improving the data analyses of the work and ensuring to be able to use them to answer the scientific questions of the study. An important remaining concern for them to address is the incompatibility between their causal framework and causal inference objective, and the use of the "dredging" and maximization of an AIC to select a model (see above). Finally, a last extra effort is still needed to improve and correct the explanation of their approach in the main text, associated methods section, and figure legend.

> ***RESPONSE R1.13:*** *The reviewers' suggestions are much appreciated, and we believe we have addressed all the remaining concerns (see responses above).*

II. Definition of resilience
The authors now very clearly define "resilience" while acknowledging other possible definitions, which greatly clarifies the scope of their objectives and questions. I like the new paragraph they wrote in Method section 2.4, where they further lay out the physiological thinking behind their definition.

> ***RESPONSE R1.14:*** *We thank the reviewer for helping us to much improve this section of the paper!*

III. Soil properties

Soil fertility needs to be defined upon first use in the introduction (mention that is refers to exchangeable cations), to avoid needing to go the Methods to known what "fertility" here refers to (e.g. not related to nitrogen or phosphorus).

> ***RESPONSE R1.15:*** *We agree that this should be clarified in the main text, and have added this clarification where specific soil fertility values are first mentioned (see line 177, referencing the fertility map in Extended Data Fig. 9). We decided not to define fertility as "exchangeable base cations" at the point of first use of the word "fertility" in the introduction, because it is in the context of explicating hypothesis 2, where specific nutrients are not important to the more general fertility hypothesis, and so referencing them would be inappropriately, and perhaps confusingly, specific.*

In the Methods, please specify the range of soil depths used for the variables downloaded from SoilGrids.

> ***RESPONSE R1.16:*** *Done*

I suggest removing the use of "soil types" terminology, as it suggests discrete soil categories, but actually refers to two continuous variables considered independently – soil texture and fertility. I would simply mention "soil texture and fertility". It will allow only using "types" when referring to the ecotopes (which here consists of actual categories).

> ***RESPONSE R1.17:*** *We agree, we have made the suggested change (e.g., lines 33, 205, and 550).*

I hope the above comments and suggestions will help the authors further improve and clarify their manuscript. This is a very valuable, interesting, and topical subject and set of results.

> ***RESPONSE R1.18:*** *We again thank reviewer Dr. Bauman for comments which we think have much improved the manuscript. With the reviewer's permission, we have included your name in the acknowledgements.*

David Bauman

References

Ankan A, Wortel IMN, Textor J. 2021. Testing Graphical Causal Models Using the R Package "dagitty". Current Protocols 1: e45.
Arif S, MacNeil MA. 2022a. Applying the structural causal model framework for observational causal inference in ecology. Ecological Monographs: 1–18.

Arif S, MacNeil MA. 2022b. Predictive models aren't for causal inference. Ecology Letters 25: 1741–1745.
Barnard-Mayers R, Kouser H, Cohen JA, Tassiopoulos K, Caniglia EC, Moscicki A-B, Campos NG, Caunca MR, Seage GRS, Murray EJ. 2022. A case study and proposal for publishing directed acyclic graphs: The effectiveness of the quadrivalent human papillomavirus vaccine in perinatally HIV Infected girls. Journal of Clinical Epidemiology 144: 127–135.
McElreath R. 2020. Statistical Rethinking (Second Edition): A Bayesian Course with Examples in R and Stan. Abingdon: CRC Press.
Pearl J. 2009a. Causal inference in statistics: An overview. Statistics Surveys 3: 96–146.
Pearl J. 2009b. Causality: models, reasoning, and inference. Second Edition. Cambridge University Press.

Referee #1 (Remarks on code availability):

I was not able to review the R code underlying the study's analyses as it was not provided.

The revised manuscript states "Code for reproducing the modeling analysis and graphs is posted at XXX.".

I therefore can't evaluate whether the code content and its clarity (adapted level of comments and explanation to accompany the coding) are sufficient to ensure the sufficient transparency and reproducibility of the study.

> ***RESPONSE R1.19****: the code has now been posted (see code ocean: https://codeocean.com/capsule/2432086/tree) as part of this resubmission.*

Referee #2 (Remarks to the Author):

In the revision the authors have done a thorough job in responding to the points raised in the previous reviews and have taken new approaches to bolster their arguments. I found that the responses to the rebuttal to be more convincing than what appeared in the actual revision.

The presentation and writing in the main body of the paper makes the information difficult to evaluate, with the writing and presentation being inconsistent, and full of multi-clause sentences with much of the information being stuck into parenthetical remarks. They hypotheses are not well stated and they are inconsistently used and referred to throughout the text.

> ***RESPONSE R2.1****: We have simplified the presentation of the hypotheses, made references to the hypotheses more consistent, and simplified sentence structure throughout (see marked texts changes).*

A lot of methods that are specific are still left in the text, while a very clear, crisp overview of the study design that might go well in the introduction is in the methods. In particular L516-527, and really L516-544 are important. In the main body of the paper (before the methods) the analysis and sources of data that the conclusions rely on are multiple and complex, and it is not easy to follow the authors' reasoning as to why one approach was chosen over multiple others. Having that first section of the methods early on would really help the reader to understand what is being done, the overarching design, etc. The writing is terse and informative too.

> ***RESPONSE R2.2****: We have revised the first paragraph after the introduction, modeling it on the beginning of the methods section, to now read (line 59):*
>
> > "*Here, we used satellite indices of forest photosynthesis to test whether three non-exclusive ecological hypotheses that go beyond climate-only explanations, developed from forest plot-scale observations, can also predict regional scale responses to these recent droughts across intact terra firme forest types of the Amazon basin.*"
>
> *We have lightly trimmed methods detail from the main text -- but, mindful of reviews of the earlier version that asked for more methods detail, we have done a minimalist trimming; hopefully we now have the balance right!*

It may be that this paper is just too complex for adequate presentation in the Nature short form. Sometimes studies are elegant and convincing in total, but can't be convincingly boiled down to short format without losing the key elements of the argument and support for the choices made. It might be easier to publish this in a long format, still top-tier journal, and do the work full justice.

As it stands, the first section still seems a bit haphazard in what it includes and what it leaves out, compared to the nice, full discussion of the ecology and analysis choices that reveals it that comes in the methods. If I were to make a recommendation I might say to move even more detail to the methods and focus hard on clear, concise statement of the results and discussions of their implications. The writing in the methods is clear and rock solid. The main body of the paper needs to follow this tone and directness, leaving out anything that is unessential for the main points, keeping detailed methods in the methods section. Go for the story in its clearest, most streamlined way. The authors have worked on this, but it still required editing and tightening.

***RESPONSE R2.3**: We don't think the paper is too complex for the Nature format, but appreciate the reviewer-recognized challenge of presenting a rich, multifaceted story clearly and concisely. To this end, we have worked to follow the reviewer's advice to simplify the presentation, move details to the methods, and tighten the wording (specific examples of this are in the reviewer 2 responses below). We thank the reviewer for constructive suggestions to this end.*

Specific comments:

L26: Capitalize E in Earth (referring to planet) ***RESPONSE:** Fixed.*

L34: Cut "importantly" ***RESPONSE:** Done.*

L36: Maybe response rather than responsiveness for clarity? ***RESPONSE:** "response" used.*

L39-40: This sentence is unclear--is the cause hardier trees or tall forests with deep roots? Are you talking about the same area with alternative explanations or different areas with different explanations?

***Response 2.4**: This is an alternative explanation. We have changed the text to improve clarity; it now reads (underlined word added) 'hardier trees (or <u>alternatively</u>, tall forests, with deep-rooted water access)'.*

L61: End with period, maybe "increased mortality"

***Response 2.5**: We have now changed the text so that this sentence is a proper question (Lines 57-58). We feel that 'enhanced mortality' is clear.*

L65: "...and test". More generally, that sentence is a monster with many clauses and ideas--maybe break into a couple sentences for clarity?

***Response 2.6**: We have simplified this sentence to address this concern (the revised sentence is quoted in R2.2).*

L69: The other side of drought hypothesis needs to be introduced--move to end of sentence so the reader has some idea what it means? It is not self-evident. You can even say, "This hypothesis is termed the other side of drought hypothesis because...."

Response 2.7: *To maintain parallelism across hypotheses, we have retained the writing style of giving each hypothesis a nickname label at the beginning of the sentence posing the hypothesis (Lines 63, 70, and 76).*

L80: rewrite that clause for clarity. ***Response***: *rewritten to clarify (Line 71 and on)*

L83: and throughout. The manuscript could still use a good editing to streamline and reduce wordiness/increase clarity. Especially strongly realized-->strong? Or Especially strong, prevalent, etc.?

Response 2.8: *We have deleted the sentence at former line 83 to reduce the wordiness. We have edited throughout to streamline.*

L86: : --> .

Response 2.9: *Done. The colon was replaced by a period amidst a re-phrased hypothesis.*

L88: Large seed mass doesn't seem to be important per se for canopy response to drought

Response 2.10: *Large seed mass is given as an example of drought tolerance traits according to ter Steege et al., 2006, but agree with the reviewer that it is potentially confusing here, and have deleted it.*

L110: I would put the idea that you are using three droughts to test three spatially segregated regions with distinct ecotypes up in the paragraph ~L62 where you lay out what you are going to do in the paper.

Response 2.11: *We have incorporated this sense into the one sentence "lay out" paragraph (the sentence being the one quoted above, in R2.2)*

L123: In the writing a lot of important, stand-alone information is put into parenthetical comments within clauses. I'd suggest not using this construction and integrating the information into the sentences proper, and in some cases breaking sentences for clarity. In this paragraph it seems like some of the detailed methods should be moved to methods.

Response 2.12: *We have moved detail out of parenthesis and moved some detail to the methods in this revised paragraph (now at lines 115-124):*

> *"In order to rigorously quantify the sensitivity of forest response across multiple droughts, we implemented two separate statistical approaches in sequence: non-linear multiple regression (using Generalized Additive Modeling, GAM), to test hypotheses and predict basin-wide drought anomalies, using AIC selection to identify the best predictive models (methods §2.6.1) and Structural Causal Modeling (SCM)*

(using Directed Acyclic Graphs, DAGs) to more systematically evaluate the causal relations suggested by the GAM analysis (methods §2.6.2)... We focus on the multiple regression GAM results below, and report comparisons with SCM results in methods §2.6.3."

The sentence on L195 is another example of this—important information for evaluating or even understanding the full hypothesis is put into a multi-clause sentence with parenthetical inclusions.

***Response 2.13**: We have simplified the sentence structure here (see current line 190).*

L151: Set up the importance of sunlight anomaly and define.

***Response 2.14**: We rephrased here to better pick up on the existing setup in the previous paragraph: "*Despite the even greater sunlight increases in 2015 than in 2005 (Fig 3B…" *(Line 146)*

L233: Citation? Tall trees are more resistant to drought? I thought it was the opposite.

***Response 2.15**: It was this reviewer's comment on the seeming contrast between greater risk for individual tall trees (as seen in the Nepstad dry down experiment) and the result reported here of a protective effect of collectively tall canopy heights, that led us to address this issue in detail last time around. See our previous response to this reviewer (December 2023 resubmission, response **R2.10 (3a)** about tree height), and the discussion added to the paper in consequence (now at lines 1237-1245):*

> *"Recalling that the satellite-derived canopy heights are not individual tree heights but overall mean heights of forest canopies over a 1km pixel, we hypothesize that both results are true: that deep water table forests that are tall on average (and presumed to have on average deeper roots that bring greater collective access to deep water resources) are more resilient than forests that are on average shorter, but that individual tall trees, subject to greater atmospheric drought stress from higher VPD, may be individually more vulnerable than their average-height neighbors…."*

L260: I would change the OR to AND as the sentence is talking about how these patterns emerge when the three datasets are combined. ***RESPONSE:** fixed (OR changed to AND).*

Figure 3: You might put some interpretation into the legend as well. The technical description is good, but this is still a very complicated figure with many assumptions going into the choices of what is shown, so the reader would benefit by a clause at the end of each section saying something like, "....showing" and then saying what the authors summarize as the important points for the reader to extract in order to evaluate their hypothesis. Figure 4 accomplishes this with the additional in-figure labels that help orient the reader, though it is still complex.

RESPONSE 2.16: *We have edited the Fig. 3 caption to be both more concise and give more interpretive guidance.*

L301: This is a big result! Should be in abstract.

Response 2.17: *We agree, and have now included this result in the abstract, at a net cost of about 10 additional words.*

Referee #3 (Remarks to the Author):

I will lead with an apology if delays in reviewing were due to my holding the revision for an extended period. I will underline that a large part of this delay was due to the comprehensive work accomplished by reviewers as well as the very detailed and comprehensive responses provided by the authors. It was a pleasure and an educational experience to read through and digest the work that has been accomplished.

I am extremely satisfied with responses to all of my inquiries, and I appreciate the inquiries and responses from all of my colleagues. We could continue dialogue over some of these issues, but I believe that is best continued in the public literature, where others can profit from the value of the work achieved. I hope that we all find this experience to be a model for peer review (even if it involves delays), and I am very pleased with the solid result, which I believe will have a great deal of impact.

I do provide a couple of minor points for polishing the revision below by clarifying some language and tempering some framing.

Best wishes to all involved,
Chris Baraloto

Response R3.1: *We thank reviewer Dr. Baraloto for constructive and helpful reviews. We are glad that they are satisfied with our responses and that they have found this to be an enjoyable and educational experience.*

With the reviewer's permission, we have included your name in the acknowledgements.

L65 - might be worth underlining early on that these are not exclusive

Response R3.2: *hypotheses have now been clarified to be non-exclusive (line 60).*

L95/L128 - suggest adding something about the spatial scales at which these three hypotheses and their synergies may behave and the scales at which you will conduct the analyses across the landscape

Response R3.3: *Implications for the large spatial scale of our analysis (relative to the plot scale data from which the hypotheses were generated) is now introduced in the opening statement of goals (see lines 59-62).*

L111 - suggest tempering the language of 'distinct' regions, as described in response to reviewers - they are distinct in large part because they permit simplification of interpretations

Response R3.4: *We have deleted the word "distinct"*

L123 - bravo to the articulate Bauman suggestion of DAG, but not sure this merits the employ of 'rigorous'

Response R3.5: *As we mentioned last time around (see our response R3.7 in the December 2023 resubmission): At this point in the manuscript, we are transitioning here from a straightforward presentation of observational patterns (Figs. 1-2) to a more systematic quantification of sensitivity and prediction, with uncertainty, using two different statistical modeling approaches (SCM and GAM regression, presented in Figs. 3-4) to increase confidence in the resulting interpretations. In this context (not just SCM/DAG but the whole modeling approach), we think "rigorously quantify" is an appropriate and informative qualifier demarcating the shift in approach.*

L193 - suggest tempering this proxy (a convenient generalization) as my read of these references does not justify a broad generalization of tree height-rooting depth for such a strong inference. For example, I can think of a few large-statured taxa that are relatively common in deep water tables and yet in my experience have relatively low belowground biomass.

Response R3.6: *We changed the wording in the main text from "focusing here on forest canopy height, a proxy for rooting depth when water tables are deep" to "**using** forest canopy height **as** a **rough** proxy for rooting depth when water tables are deep" (highlighting the added words). Please note that in the corresponding methods, there is further exploration of the caveats to this assumption (at line 643-649):*

> *"observations of the tree height-rooting depth allometry are limited, especially in tropical forests (although one study cited here[35] is directly relevant, as it is from central-eastern Amazon upland forest, conducted during the 2015 drought); this limitation remains a key uncertainty in our ability to confidently attribute variations in drought response to rooting depth, as opposed to canopy height itself, or other (as yet unidentified) correlates of canopy height. We also note that shallow WTD limits rooting depth such that canopy height correlations to rooting depth in these forests may be diminished."*

L206 - soil texture integration is useful, but this reads as an add-on response rather than integration. what about vegetation effects - extreme sandy soil areas have lower statured forest, and Guianan forests are highly sandy but support high-statured, high biomass stock forests (see L214 after)?

> ***Response R3.7***: *By "integration" we mean that soil texture is now incorporated into the integrated model which generates predictions from texture and all of the other factors (climate, Water-table depth, canopy height, and fertility) considered together. (That is, when we test effects of soil fertility, it is with the effects of texture also accounted for). Because soil texture is not one of the three core hypotheses were are testing, we moved the soil texture paragraph from the main text to a (main text referenced) updated discussion (including of the interactions with height raised by the reviewer, see Extended Data Fig. 11) in the methods (see lines 1024-1048).*

L222 - bis for soil texture, what happens in deep water tables on more sandy soils? what about extreme sands?

> ***Response R3.8***: *Please see the updated more extensive discussion (also referenced in **R3.7**) about soil texture in the methods section (lines 1024-1048), specifically the discussion around Fig S11.*

L294 - this can be rephrased for clarity 'strategy of growing trees fast...'

> ***Response R3.9***: Rephrased *(line 285).*

L370 - change ecotype to ecotope for consistency (also thereafter in Fig legends and annexes) ***Response R3.10***: *Fixed.*

L611/L784- it might be useful to add a small table on the spatial resolution of each of the different datasets integrated herein; at least add for all when they first appear in text (eg, soil texture).

> ***Response R3.11***: *We now give the resolutions of the different datasets in a bulleted list (at lines 792-801), and update the first description of each dataset to include its resolution as well. .*

L810 - was classification analysis re-run using soil texture during revision?

> ***Response R3.12***: *No.*

Referee #4 (Remarks to the Author):

I was, I believe, Rev 4 from the last submission. As I voiced last time, I find this to be an interesting and important advance regarding green-up/brown-down dynamics in the Amazon and provides a logical and compelling insight.

In their revisions, the authors addressed most of my concerns fairly well. I was not totally satisfied with falling back on this being an "ecosystem" study and not community ecology as a reason why the functional biogeography was so basic. Functional diversity is key to functional biogeography and not simply a community ecology issue. So, I still kind of find the functional biogeography aspect of the study to be more window dressing than a key advance. It will be interesting to hear if Rev 3 still feels similarly. All of that said, I don't think my concern about this issue is super critical as I think the core results and findings are important enough on their own.

I think the team largely addressed the concerns by the other reviewers, but I will leave that up to those reviewers to weigh in. I will note that DAGs are important and outstanding tools, but I think we are fooling ourselves to think they mean we can now say our studies are truly causal. They aren't. Our DAGs are only as good as what is in them in the first place and we are still dealing with large (and inherently messy) datasets. I think adding DAGs was a wise suggestion and I'm pleased the team included this approach, but I don't consider this work (or any similar work using DAGs) to have causal certainty.

> ***Response R4.1****: We thank the reviewer for the insightful comments and constructive suggestions!*

Referee #5 (Remarks to the Author):

You substantially revised the paper, addressing both my comments and comments by other reviewers. I believe your manuscript now meets the standards required for publishing.

> ***Response R5.1****: We thank the reviewer for the insightful comments and constructive suggestions!*

Reviewer Reports on the Second Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

Third review of the manuscript “Amazon forest biogeography predicts resilience and vulnerability to drought” by Chen et al.

Line numbers refer to the “merged” document of resubmission.

I am overall satisfied with the clarifications and revisions the authors made to the work.

In my last review, I was very pleased to see that the authors raised to challenge, by assuming the causal nature of their questions and objectives, and therefore using appropriate causal inference frameworks to formerly justify causal interpretations of statistical models built by applying the backdoor criterion to a causal diagram, that is, using causal structural modelling. The authors embarked on this “causal journey”, drawing a causal diagram (DAG), confronting it to the data to verify conditional independencies (DAG-data coherence), update the DAG based on incoherent conditional independencies, then for each exposure (predictor whose causal effect on the response one wants to quantify) apply the backdoor criterion to define a minimum set of control variables necessary to allow a causal interpretation of the exposure’s coefficient fitted by the statistical model.

My remaining concern was that while the authors went to great lengths to do all this, the results presented in the manuscript were still based on an approach aimed at optimising predictive capacity (dredging approach combined with information criterion to select models).

I suggested – both in the first and the second review – that the authors could either replace their initial approach by the models resulting from the structural causal modelling framework (option A), or (option B) that they could use both a causal framework to address the causal questions, in addition to a predictive statistical approach (their initial analytical approach) to optimise prediction, i.e. a complementary – though different – valuable objective.

In their latest rebuttal, and in the revised manuscript, the authors now clarify much of what was not clear to me in the previous version, in terms of what they chose as analytical approach, what they present in the main text, and what is done but presented as supplementary material.

The work is now much clearer and has greatly improved over the successive revisions, including this last one. It will be great to see this important work published.

I have a few remaining minor suggestions/edits, mostly of rephrasing for clarity and sometimes accuracy.

Most of these relate to their referring to the AIC predictive modelling approach to also being “to test

hypotheses”.
I understand what the authors mean now, but it took me a while, and I think a minor reformulation would help remove any potential ambiguity. The way I understand it, the authors here refer to their three competing hypotheses, the ones they present in the introduction. What their predictive models do is that it indeed *tests those three hypotheses in terms of their capacity to predict the outcome*. The ambiguity is that predictive models in general are precisely not for “hypothesis testing” in the broader sense of “hypothesis testing” in Ecology, when meaning “testing the existence of a causal effect or mechanism” (see many refs from previous comments on the topic).To say it differently, if y = beta*X, where y is the predicted outcome, X is a matrix of covariates, and beta is the fitted coefficients for those covariates, the AIC approach allows comparing hypotheses in terms of what they do to y (optimised prediction of outcome; regardless of beta), whereas SCM allows focusing on beta to test and compare hypotheses (regardless of y). The authors then say that the Structural Causal Modelling approach is *also for hypothesis testing*, as if both approaches had that goal in common. I get what the authors want to say, but think that a minor reformulation would clarify the above.
I’d suggest saying explicitly that the AIC approach tests the three alternative hypotheses in terms of their capacity to predict forest drought resilience, whereas the SCM approach tests causal effects and quantifies the support to the three hypotheses. This, or something along those lines, to me better explains and stresses the complementarity (and complementary strengths) of the two approaches. This is a detail, but it is an important one I think, mostly in a context where there is still unfortunately a major lack of distinction between predictive and inference goals and analyses in Ecology. Such clarity will go a long way for future works (Tredennick et al. 2021, https://onlinelibrary.wiley.com/doi/abs/10.1002/ecy.3336).
I make some specific (and minor) suggestions below:

Line 117: I suggest specifying “to test our three hypotheses” (to make it clearer that you refer to these specific hypotheses)

Lines 115-124: I suggest rephrasing a few parts of this important paragraph for improved clarity (key points between asterisks): “… we implemented *statistical analyses based on two complementary analytical frameworks* in order *to both optimise predictive capacity, and quantify specific mechanisms and causal effects: (1) non-linear multiple regression (using …) to test our three working hypotheses based on their capacity to best predict basin-wide drought anomalies, using AIC-based model selection (methods …), and (2) a Structural Causal Modelling framework to formally assess and quantify causal effects of the different relevant environmental predictors on forest drought resilience (refs)*. Both modelling approaches were conducted (…). *Because both analytical approaches mostly converged, we hereafter focus on the predictive modelling approach and results here, but provide comparisons with the Structural Causal Modelling framework and its results in the methods* ….”

Lines 187: Close “hypothesis 3” parenthesis.

Line 215: I suggest adding the bit between the comas: “Our GAM modelling framework, in particular guided or accompanied by a formal causal inference analytical framework, powerfully allows further

…” as this combined approach with complementary strengths is clearly one of the points where the study stands out.

Line 263: Slight rephrasing needed here, for the content of the parentheses: “our causal effect models – Structural Causal Modeling – suggest that the GAM predictive models (Figs. 3, 4) and their predictors’ coefficient here reflect causal effects)”. If the initial phrasing of the parenthesis content is kept unchanged, then change “more rigorously” to “rigorously”, as “more rigorously” implies that AIC-based approaches informs causality too.

Line 958: “Model” could mean “statistical model” too. Please clarify “is conditional on the causal model (diagram) being true, and on the inference about a given predictor being justified by the backdoor criterion”.

Line 1054: Close parenthesis after reference 120.

Line 1069: “no conditional independence” instead of “no independence”

Lines 1078-1081: Complicated sentence. I suggested cutting in two and slightly rephrasing.
For example: “Completing these steps, we obtained the final DAG (Extended Data Fig. 9c) laying out our assumptions regarding the data-generating processes for forest drought response. We used this final DAG by, for each causal predictor of interest, applying the backdoor criterion to derive a set of control variables to be used in our statistical models (GAMs), as reported in Extended Data Fig. 10. Note that because the backdoor criterion often yields different sets of control variables (e.g. Arif and MacNeil 2022a,b), for different pairs of causal predictor and outcome of interest, we ran separate models to quantify the causal effects of the different predictors of interest.”

Line 1442: Correction needed in my name: David (not Dave) Bauman.
And thank you for the inclusion in the acknowledgments. I am happy I could help.

Best of luck for the continuation.

David Bauman

Referee #2 (Remarks to the Author):

The authors have done a good job of increasing the clarity of the manuscript and responding to the reviews. A few specific comments below.

L118: The clause about AIC selection can be cut and moved to methods--it is too much detail for this part of the paper, particularly given how routine AIC use is in model selection.

L215: Change "Our GAM modeling framework powerfully allows further investigation of additional questions, generating a rich suite..." to "Our GAM modeling framework generates a rich suite..." Sentence keeps the same meaning with less hype.

L217: Sentence would benefit from numbering or otherwise demarcating the hypotheses beyond just the semicolons.

L273: lower-case b in biogeography. And, maybe, "Implications of a functional biogeography"?

Author Rebuttals to Second Revision:

RESPONSE TO REVIEWERS

Referee #1, Specific Comments:

Line 117: I suggest specifying "to test our three hypotheses" (to make it clearer that you refer to these specific hypotheses)

DONE.

Lines 115-124: I suggest rephrasing a few parts of this important paragraph for improved clarity (key points between asterisks): "… we implemented *statistical analyses based on two complementary analytical frameworks* in order *to both optimise predictive capacity, and quantify specific mechanisms and causal effects: (1) non-linear multiple regression (using …) to test our three working hypotheses based on their capacity to best predict basin-wide drought anomalies, using AIC-based model selection (methods …), and (2) a Structural Causal Modelling framework to formally assess and quantify causal effects of the different relevant environmental predictors on forest drought resilience (refs)*. Both modelling approaches were conducted (…). *Because both analytical approaches mostly converged, we hereafter focus on the predictive modelling approach and results here, but provide comparisons with the Structural Causal Modelling framework and its results in the methods* …."

Accepted with editing for brevity. See the revision at current line 117-126.

Lines 187: Close "hypothesis 3" parenthesis. *DONE.*

Line 215: I suggest adding the bit between the comas: "Our GAM modelling framework, in particular guided or accompanied by a formal causal inference analytical framework, powerfully allows further …" as this combined approach with complementary strengths is clearly one of the points where the study stands out.

Accepted with editing for brevity. See the revision at current line 217.

Line 263: Slight rephrasing needed here, for the content of the parentheses: "our causal effect models – Structural Causal Modeling – suggest that the GAM predictive models (Figs. 3, 4) and their predictors' coefficient here reflect causal effects)". If the initial phrasing of the parenthesis content is kept unchanged, then change "more rigorously" to "rigorously", as "more rigorously" implies that AIC-based approaches informs causality too.

Accepted with editing for brevity. See the revision at current line 265.

Line 958: "Model" could mean "statistical model" too. Please clarify "is conditional on the causal model (diagram) being true, and on the inference about a given predictor being justified by the backdoor criterion".

See the revision at current line 962. (note: the reviewer seems to be referring only to the SCM models, but the condition on the model being true applies to both modeling approaches (including GAM). Reworded to clarify, and citing the DAG as an example).

Line 1054: Close parenthesis after reference 120. *DONE*

Line 1069: "no conditional independence" instead of "no independence" *DONE (Line 1079)*

Lines 1078-1081: Complicated sentence. I suggested cutting in two and slightly rephrasing. For example: "Completing these steps, we obtained the final DAG (Extended Data Fig. 9c) laying out our assumptions regarding the data-generating processes for forest drought response. We used this final DAG by, for each causal predictor of interest, applying the backdoor criterion to derive a set of control variables to be used in our statistical models (GAMs), as reported in Extended Data Fig. 10. Note that because the backdoor criterion often yields different sets of control variables (e.g. Arif and MacNeil 2022a,b), for different pairs of causal predictor and outcome of interest, we ran separate models to quantify the causal effects of the different predictors of interest."

We realized the sentence starting with "Completing these steps…" was mostly redundant with the prior sentences and the problem of complex sentence construction here could be better addressed by deleting this sentence and adjusting the prior sentences to accommodate the reviewer comment. See the revision at current line 1084-1088.

Line 1442: Correction needed in my name: David (not Dave) Bauman.
And thank you for the inclusion in the acknowledgments. I am happy I could help.

Accepted! (following Nature style it is now "D. Bauman")

Best of luck for the continuation.

David Bauman

Referee #2 (Remarks to the Author):

The authors have done a good job of increasing the clarity of the manuscript and responding to the reviews. A few specific comments below.

L118: The clause about AIC selection can be cut and moved to methods--it is too much detail for this part of the paper, particularly given how routine AIC use is in model selection.

DONE (Line 119)

L215: Change "Our GAM modeling framework powerfully allows further investigation of additional questions, generating a rich suite..." to "Our GAM modeling framework generates a rich suite..." Sentence keeps the same meaning with less hype.

DONE (Line 217)

L217: Sentence would benefit from numbering or otherwise demarcating the hypotheses beyond just the semicolons.

DONE (Line 219)

L273: lower-case b in biogeography. And, maybe, "Implications of a functional biogeography"?

DONE (Line 274)