Reviewer response

Alexey N. Shiklomanov, Elizabeth M. Cowdery, Michael Bahn, Sabina Burrascano, Chaeho Byun, Joseph Craine, Andrés Gonzalez-Melo, Alvaro G. Gutiérrez, Steven Jansen, Nathan Kraft, Koen Kramer, Vanessa Minden Ülo Niinemets, Yusuke Onoda, Enio Egon Sosinski, Nadejda A. Soudzilovskaia, Michael C. Dietze

1 Editor comments

> From my own reading, I am strongly concerned about the very small sample sizes, combined with trait-filling of missing data, that mean many PFTs are quite data poor. I also wonder whether it is really that surprising that relationships observed across broad data ranges break down when the data are aggregated. I think it is well known that any regression will exhibit less strong results if one samples a narrow range of data. This makes me wonder whether the results reflect anything particular about the nature of tradeoffs within PFTs or just the fact that species within PFTs are generally similar so we would expect weaker relationships detected within them. This could be tested quite easily by randomly assigning species to PFTs, either fully at random so they would have small sample sizes but the full spread of the data, on average, or in a stratified random fashion such that species tended to be similar within groups. Would the results be similar? If so, are we learning something about PFTs and tradeoffs, or just something about the statistical properties of breaking a large data set into small groups. Exploration of this question would be a useful 'null model' to back up this study.

There were several reasons for the relatively large number of plant functional types in the original manuscript. For one, such a large number was necessary to explore the different drivers of trait correlation – for instance, to compare the differences in correlations between biomes versus between growth forms and leaf types. Furthermore, we were interested in the ability of our hierarchical modeling approach to support efforts to improve representation of biodiversity in models by increasing the number of plant functional types (e.g. Boulangeat *et al.*, 2012). In particular, we note that some ecosystem models are parameterized for individual species (LANDIS, Scheller & Mladenoff, 2004; e.g. Linkages, Post & Pastor, 2013).

However, we agree the number of PFTs selected and their resulting low sample sizes made it possible that many of our results were a mere statistical artifact rather than ecologically meaningful. To address this concern, and to make the paper more directly applicable to ecosystem modeling, we replaced our PFT scheme with the scheme used by the Community Land Model (CLM 4.5, Oleson *et al.* (2013)). This brings the number of PFTs down to 14. We have provided additional details about how we assigned species to PFTs in the methods (**TODO: Need reference**).

> Finally, I think it is important to be more specific about just how the implications of this work are important for modeling, or for other biological implications. Most importantly, for ecosystem models it is the PFT means that are most important, and it appears that your models suggest only very small changes in estimating means. In that the analysis does not suggest a particular problem with current approaches. Alternatively, if you can point to specific contexts where trait values are estimated or modeled among species within functional groups and how these methods would lead to different estimates, please be more specific about the contexts of such work.

TODO: Respond

2 Reviewer 1

1) My main feedback, is that I think the authors are not careful and consistent to describe the within functional group and across functional group analyses (the latter case having

two versions, the nested version and the unnested version). The authors need to pick on a sensible name for each of these and then use it repeatedly. I have highlighted just some of the places this occurs in the specific issues below, but it is more generally and the paper needs to be edited start to finish with this in mind.

TODO: Respond

2) Line 131 - Leaf Dark Respiration is a fairly rare actor in analyses of LES. Maximum photosynthetic rate is much more common (although I'm sure these are correlated). The authors own introduciton to LES on lines 69-70 mentions Amax but not dark respiration. An acknowledgment and justification for using dark respiration over Amax is needed.

TODO: Respond

3) Line 150 forward - to interpret the methods in this area I really wanted to know how many plant functional groups there are. I can kind of guess it from a later table (but it is not completely clear to me there). Can you spell this out here.

TODO: Respond

4) Line 150 - can you give a justification for assuming MVN across plant functional groups?

TODO: Respond

5) Line 179-184 - I am always suspicious of trait filling methods. Can you show your results are robust to choices here? Especially to an analysis with no trait filling? One would think this should be possible in a Bayesian world especially.

TODO: Respond

6) Line 198, line 225, line 233, line 241, line 243 line 267 - several of the places I am not sure which model is being used with which word

TODO: Respond

7) Discussion in general - I would have appreciated a little more speculation about mechanisms that cause this "scale-break" in LES. Conversely some of the rest of the discussion largely repeats the results without adding a lot of value and could be shortened. I personally was much more interested in the correlation results than the mean estimation results, and I expect most other readers would be too, so you might want to think about adjusting time devoted to each.

TODO: Respond

3 Reviewer 2

1. Why did you use mass-based (instead of area-based) nutrient concentrations and dark respiration rates? You don't justify the choice, but you should, because this makes a crucial difference for interpreting trait correlations and trait distributions among and within different groups. Given the strength and utility of your statistical approach and ideas, I think you are missing useful information by relying on mass-based values. What you gain are pretty pictures because mass-based correlations are tight in most groups of plants by mathematical necessity because they are all inextricably linked to LMA, and LMA varies interspecifically in many ecologically relevant groups of plants. It's just not that interesting, in my opinion, to reproduce strong(ish) mass-based trait correlations, but I'm very interested in your results regarding LL-LMA and look forward to seeing what your straightforward statistical approach will reveal for area-based or LMA-independent trait data (see below).

TODO: Respond

2. Why didn't you include photosynthetic rate as a variable in your analysis? You certainly don't have to, but I was expecting it and was disappointed to discover its absence when I got to the list of your analyzed traits in the methods section. My perception is that data in TRY for maximum net photosynthetic assimilation rate (Amax) is better and more abundant than dark respiration rate data, and it's certainly variable across species and interesting, so it seems odd to me that you didn't analyze it, too. If it wouldn't be a huge lift to add Amax, maybe do it? It's great, by the way, that you use TRY data in your study.

TODO: Respond

1. Introduction

a. Line 63: I loathe the phrase "Plant functional traits provide a useful framework for…" Traits do not "provide" a framework. They may constitute a framework, but what do you mean by "framework"? Please use a more careful description here. I know a huge fraction of ecologists are throwing "framework" into proposals, manuscripts, and talks, but it's an unfortunate trend toward shiny-but-sloppy language, usually used when the person making assertions about the "framework" can't effectively articulate exactly how the thing in question (traits) actually can accomplish the scale transition(s) supposedly in the "framework." I am confident you can introduce the utility of traits more precisely! In fact, I think in the line 63 sentence, you can just say "…traits are potentially useful for linking…" and you've accomplished that.

We have revised this sentence accordingly. **TODO: Provide line number**

b. Line 67: Perhaps specify that plant strategies are life history strategies?

TODO: Respond

c. Line 69: I disagree with the word "efficient" as a description of high photosynthetic rate. Maximum photosynthetic assimilation rate (what you're talking about) is different from photosynthetic efficiency (with respect to what? Water or nutrients or initial carbon investment?). Also, moving to line 70, I think your existing language describes a confusing dichotomy because you use nutrient concentrations and a gas exchange rate on one hand and LMA and longevity on the other. Also, it's a little odd that you use photosynthetic rate in your description of the LES considering you don't use it as a variable in your study. Please revise your description of the leaf economics spectrum (LES). Incidentally, though, I like the word "sturdy" here to describe the high LMA leaves. It made me smile.

TODO: Respond

d. Line 99: Good description of the background material motivating the study

e. The last paragraph of the introduction (starting at line 114) is a great description of your motivation and research approach.

Given that these two paragraphs worked well, we have kept them largely without revision.

2. Materials and Methods

a. Line 130: You listed the inverse units for LMA. The LMA units should be mg mm-2, not mm2 mg-1 (those are the units for specific leaf area (SLA), the inverse of LMA). Looking at the results, I can see that you did in fact use LMA and not SLA, which is great, because I find LMA more intuitive.

We have fixed references to units throughout the paper. However, in the revised analysis, we have switched from LMA to SLA. While we agree that LMA may be the more intuitive unit, the use of SLA allows us to directly compare our estimates to those of CLM.

b. The "Multivariate analysis" section: All the text here is good but is too sparse. The models are described clearly but need an explanation and justification of their use. What does each

model tell you? Why did you make it? How does each model individually and together help you achieve your research goal of testing your hypothesis?

c. Line 144-145: I don't like that x is lowercase in the text and capitalized in the mathematical expression

d. Line 152: need a bar over mu in the text

e. Lines 158-161: Edit the bars in your mathematical expressions and text. I don't think the xi,p in the mathematical expression should have a bar, right? Your mus in the text need bars.

f. The "Model implementation" section:

g. Yay for Stan! But they like you to spell it "Stan" instead of "STAN" (http://mc-stan.org/documentation/), so you'll have to fix that throughout your text. Thanks for putting your code up online, but I wasn't sure where to look to find it on the website provided (https://github.com/ashiklom/np_trait_analysis). Looks like you'll have to clean that up and provide a better description of the files to make this useful for people to find your code. Please do this, though, because you've obviously done a good job, and it this would be a valuable resource.

In the revised version of the text, we have moved from using Stan to a direct implementation of the sampling algorithms. This was done primarily to allow efficient filling of missing data, rather than simply omitting missing values, as we had done in the Stan implementation. However, a side effect of using our own implementation is that the sampling is more computationally efficient, and the underlying code for running the models is somewhat simpler.

To make the code easier to use, we have isolated the code for fitting multivariate and hierarchical models, as well as some associated utilities, into their own R package (`mvtraits`). We have put considerable effort into making this package more user friendly, both by adding documentation and examples to the package and by making the code design more modular and versatile.

ii. I liked your treatment of missing trait values (a rampant problem in trait analysis; I hope Stan someday allows NAs; you might mention these 2 points) and this description. If you don't already, you should list in the Supplementary materials the sample sizes for each pairwise analysis for each of your 35 PFTs. Following line 184, though, should x in the mathematical expression have a bar?

Given our new implementation of the missing data model, this comment is no longer relevant.

g. "Analysis of results" section: Again, the text is fine, but it's too sparse. You need to clearly explain how your treatment of the results of your model fitting help you achieve your objectives.

TODO

i. Line 198 typo: "with-PFT" should be "within-PFT"

TODO

ii. Line 208-209: You should move the R citation up to where you first mention R. I think the TRY data repository information should be moved up to where you introduce the TRY trait data that you used.

TODO

3. Results

a. Section "Constraint on estimates…." and Figure 1: You need to provide a biological interpretation of the results presented in Figure 1, both in the text and figure caption. You show

dashed lines for the sample means of the traits but don't provide any interpretation of what it means for a model joint probability distribution to encompass the sample mean or not.

TODO

b. Section "Trait correlation patterns":

c. Figure 2 looks great! In the figure caption, though, you need to put in a sentence about what you want us to learn from this figure. The hierarchical model results definitely look more like the data than do the multivariate model results. If you wanted to make this figure smaller, instead of printing the correlation value above the diagonal, you could use the ellipses from one model below the diagonal and the other above.

TODO

ii. First paragraph, about global-scale (starting line 225): Shouldn't you mention that the hierarchical model results capture patterns in the data much more strongly than do the results from the multivariate model? Why does the hierarchical model perform so much better?

TODO

iii. Line 233-234: awkward language regarding trait relationships. In many cases, and in the figure caption of Fig. 4, I think you can just say "pairwise trait correlation" instead of "among-trait pairwise correlation."

TODO

iv. Figure 3: Just show panel b; put panel a in the supplement or just say that it's not shown. And put a sentence in the caption about what interpretation you want us to take away from the figure. You also need to label the x-axis and explain your nomenclature for the trait relationships in the figure (x1.x2).

TODO

v. I really like Figure 4! Again, though, provide some kind of interpretation in the caption.

TODO

4. Discussion

The discussion is pretty good. Specific quibbles:

a. Line 270, "…confounding factor in characterizing…: Why?

TODO

b. Line 273: Figure out something to say other than "formal framework."

TODO

c. Line 295: drop the dash in within-PFTs

TODO

d. Line 299: not "incur." Maybe "foster"?

TODO

e. Line 324: LMA was orthogonal to N? Really? I don't recall your emphasizing that in your results, and it's certainly not the case for the results of the hierarchical model.

TODO

f. Paragraphs starting at line 340: this is the strongest part your discussion and is really good

TODO

4 Reviewer 3

1. There are a number of concerns about the PFT classification.

a) With 35 PFTs, many of the sample sizes (Table 1) are far too low for any kind of reliable estimates within PFTs and comparisons among PFTs. Given the authors' interests in PFTs and not in species per se, I would think a minimum cut-off of 10 or 12 or more species would be useful. That means almost one-fourth of the cases in Table 1 are insufficient. It is also not clear how many of the measurements of the five columns with trait data that follow came from each species. I read number of species as the number for ANY data (any column value), not for all. Thus, with species =12, the number of species for LL or Pmass or Rd might be 3 or 11. Authors should put those values in parentheses next to the numbers of observations. If there are 38 observation and 35 came from one species that does not seem a useful test at the PFT scale, even though the total number of observations is good. If number of observations of a given trait in a given PFT is 11, but from 3 species, that also is marginally useful in this context. We need to know more about this as a reader and I think the authors need to eliminate many PFT categories where number of species and/or observations and/or unique species-observation counts are low.

TODO

b) Why is growth form and leaf type missing for PS scheme C4?? I assume we know this information (and likely they are mostly herbaceous, broadleaved, yes?)? Given the low numbers of C4 species in every biome other than temperate, is this a useful class in any case?

This comment is no longer relevant in the context of our revised PFT scheme.

c) Whether a plant is tropical or boreal is a description of where it lives on Earth. It is important and potentially useful, but the framework for how to think about the different aspects of the different components of the PFTs is missing.

We have addressed this by defining climate zones in terms of mean annual temperature cutoffs, and placing species into climate zones based on the average mean annual temperature of all sites that they are present in TRY. We describe this in the methods (see **TODO**).

d) What they call "biome" is really more a "climate zone". These biomes and PFTs don't map out on to traditional biomes or IGBP-DIS biomes (see Woodward et al 2004) or biomes or PFTs used in models such as CLM or CABLE or JULES. The 'arid' biome is strange as usually there is either a desert or a grassland biome or both, and unless I missed it, I don't see how they defined a biome as arid. I have no problem with them making up a new classification but the rationale behind it needs to be stated and the rules clear. Many other PFTs distinguish trees from shrubs. It is ok not to do so, but why not? Why include CAM when the data are so scarce and only in one biome (and missing leaf type, growth form, and phenology). Does any information about CAM and arid climates become hopelessly confounded if this one photosynthetic type is only found in one climate region? In general the classification – so key to the entire paper – is kind of a mess.

See main response to editor comments. In general, we agree that our original PFT classification was not done well, and we have addressed this by following the PFT classification scheme used by the Community Land Model, and by more clearly documenting the rules we used for assigning attributes like climate zone and growth form to specific PFTs.

2. Conceptual concerns

Even more problematic from my standpoint is the lack of hypotheses about biology and ecology. From first principles, the authors need to lay out a foundation regarding our thinking about why a given growth form, leaf type, PS scheme, or phenology should differ (or not) from another in terms of relationships of trait Y to trait X? For example should needle-leaved species have similar

or different LMA-Nmass relationship as broad-leaved species from the same climate zone? How should they differ, and why? This kind of question applies to every kind of contrast (e.g., woody, nonwoody; evergreen, deciduous; needle-leaved, broadleafed; tropical, temperate). By laying out hypotheses about what we might expect to see, testing that with the data, and interpreting how to think about support for or refutation of those hypotheses, the authors would advance the field. Unfortunately this is almost entirely lacking in the paper.

TODO

3. Value for modeling

The authors pitch this work as useful to models, rather than useful from a biological concept framework. But, most earth system models use only a small number of PFTs (e.g., 5 or more recently 9 in JULES, Harper et al. 2016; or 14 in CLM Bonan et al 2011, etc.) and define those PFTs in ways that don't easily match what was done in this paper.

We have addressed this concern by directly adapting the PFT scheme used by CLM. We demonstrate the applicability of the approach by directly comparing our parameter estimates to those listed in the CLM manual (**TODO: where?**) in the discussion, as well as to other studies similarly aimed at providing revised parameter estimates for ecosystem models.

As a result of these issues (particularly 1 and 2), I don't find much of the major contrasts useful in terms of advancing understanding; e.g., I find Figures 3 and 4 useless in their present form.

The Figures for this paper have been significantly altered. **TODO: Finish this!**

They don't advance our understanding of the underlying plant biology or ecology, nor do they advance our quantitative description of how the trends vary, as the 35 choices are far too many to be useful given the sample size and structure (not enough observations, not enough species).

TODO

I also don't think they are useful for modeling, given the extremely high number of PFTs.

TODO

Boulangeat I, Philippe P, Abdulhak S, Douzet R, Garraud L, Lavergne S, Lavorel S, Es JV, Vittoz P, Thuiller W. **2012**. Improving plant functional groups for dynamic models of biodiversity: At the crossroads between functional and community ecology. *Global Change Biology* **18**: 3464–3475.

Oleson KW, Drewniak B, Huang **Maoyi,** Koven CD, Levis S, Li F, Riley WJ, Subin ZM, Swenson SC, Thornton P. **2013**. *Technical description of version 4.5 of the community land model (clm).* NCAR Earth System Laboratory Climate; Global Dynamics Division.

Post WM, Pastor J. **2013**. LINKAGES: An individually-based forest ecosystem biogeochemistry model.

Scheller RM, Mladenoff DJ. **2004**. A forest growth and biomass module for a landscape simulation model, LANDIS: Design, validation, and application. *Ecological Modelling* **180**: 211–229.