Earth Syst. Sci. Data Discuss., referee comment RC2
https://doi.org/10.5194/essd-2021-25-RC2, 2021

Comment on essd-2021-25

Anonymous Referee #2

Referee comment on "A global map of root biomass across the world's forests" by Yuanyuan Huang et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-25-RC2, 2021

Like the previous reviewer, I commend the authors for a well-written and comprehensive manuscript describing a new, impressive and important dataset representing the biomass stocks embodied in global tree roots. I particularly appreciated the thorough retrospective in the introduction describing the history of root biomass estimation at the global scale, the simple ecological analyses the authors include in their discussion, and the transparency with which they describe [most of] their methods. Combined, I think these aspects make the manuscript an accessible descriptor that will appeal to readers across many disciplines.

My only substantial concern pertains to the way in which the author's 'central' estimate was determined distinct from their uncertainty estimate. The 'central' estimate (142 Pg) was generated using models trained and executed using covariates each represented by a singular data source. The corresponding uncertainty estimate, though, was generated using an ensemble approach that used multiple [alternated] data sources to represent those covariate values. It's unclear to me why the authors favored the data sources they did when generating their 'central' estimate and, having gone the lengths of creating an ensemble, why they didn't instead just use the ensemble mean or median as the 'central estimate? I ask that at the very least, the underlying rationale be made clear in the text.

In addition, I have several minor suggestions (largely related to clarity) that I ask the authors to consider and address:

Line 31: change "root plays the" to "roots play a"

Lines 55-80: This is a really nice summary of the discipline to date.

Line 85: please add "n = " before 10307.

Lines 90-91: Consider rephrasing to emphasize that *after comparing the results of all three of the candidate techniques*, you chose the RF approach because it performed best and only used it (not the others) for subsequent mapping and analysis. Right the text seems to abruptly drop any reference to the other two approaches (ANN and MARS).

Line 98: "Combining _____ with tree density..." please fill in the blank I've added to the

text.

Lines 103-105: Could these data be available from the authors? Was an attempt made to find out?

Section 2.3: Text explaining why you chose the covariates you chose and why you selected particular datasets (over others) to represent those covariates is needed in this section.

Lines 124-129: Specifying that the "BIO_" variables are simply the WorldClim bioclimatic indicators help clarify why these [otherwise] seemingly odd abbreviations are used.

Line 129: Why not also include slope and/or aspect?

Line 130: Personally, I think it would be useful for table S1 to be included in the main text.

Line 134-136: This sentence is unclear to me. Are you saying that the two input datasets by Baccini and Santoro are the most reliable sources or that the layer you derive from them is? Please clarify in the text.

Line 138: Did you also use the Baccini dataset? From the preceding text and references I was led to believe that you somehow combined the GlobBiomass and Baccini maps but here it seems like you only use GlobBiomass. Whichever is true, please clarify in the text.

Line 145: Please include an explicit reference to the canopy height map, here.

Line 148: Please clarify: I don't believe Hansen reports a tree count, just area. So how can the consensus dataset give the same tree count as Hansen?

Line 152: Given your amendments and modifications described in this paragraph, can you state explicitly here what definition of 'forest area' your map adopts? This'll be important to facilitate future comparisons much like the comparison you make here.

Lines 153-159: Do all of these age maps use the same reference year? I.e. Age in/as-of what year? Please clarify in the text.

Line 163: Does this imply that one candidate map was made by simply applying allometric equations to a map of forest height (and presumably stratified by taxa)? If so, that exercise isn't yet clearly explained in the methods text, here. Please do so.

Line 163: Please also provide the long-form names when introducing these acronyms.

Line 190: It's hard to know how you actually chose your final model given that you considered three distinct criteria. Can you elaborate here a little bit? Presumably you considered a hierarchy in these criteria? For example, if one model had the highest MAE and another had the highest R2, which did you choose? Why? How?

Line 190: From an ecology perspective, I can understand that minimizing the number of covariates can help more clearly explain the drivers of predicted patterns (e.g. the variable importance assessment described below). But, from a mapping perspective, where you ultimately rely on RF and where I would think accuracy is the ultimate goal, I wonder if it would be more appropriate to retain all predictors? Can you at least explain in the text your reasoning for culling covariates/favoring parsimony?

Line 193: Consider including a table with the validation stats of each approach to clearly

illustrate why RF was chosen over the others.

Lines 200-201: This sentence is vague. I assume you mean you combined it with the Crowther map? Either way, please make this step clearer in the text.

Section 2.6: Above you used a single dataset for each covariate to generate your 'central' estimate of root biomass. Here you describe using a separate ensemble approach to generate your error estimates. Why not instead use the mean/median of the ensemble approach as your 'central estimate'? In other words, why did you decide to prioritize the datasets listed in table S1 over those in S2? At the very least, this should be clearly explained in the text.

Line 230: I suggest noting that this is Pg of dry biomass so that it's not wrongly compared to estimates by others that might be in units Pg C.

Lines 231-232: Is this the actual definition of forest cover you used? Above in the methods, it seemed a little more convoluted than this?

Discussion/Figure 1: Both Mokany et al. 2006 and the 2019 refinement to the IPCC guidelines report mean/median R:S's that are larger than 0.50 for some ecotypes -- namely some dry woodlands, savannas, oak forests, and boreal forests. Visually it looks like your predicted patterns generally agree with these sources (at least in a relative sense) but because the colour scale is capped at 0.50, it's hard to know if the magnitude of the estimates in these areas are comparable? Are they? Perhaps an R:S comparison with one or both of these widely used sources would be useful to include in the discussion?

Table 1: The new maps by Spawn et al. (2020) may also be worthy of comparison here in Table 1. Near the end of their manuscript, they report 122 Pg C in global root biomass with 28.3 Pg C embodied in herbaceous (i.e. grass) roots. So the comparable value is likely 94 Pg C in tree biomass which, if you simply assume wood biomass contains 50% C, equates to 188 Pg biomass. Note too that they appear to use a more liberal definition of tree cover than you.

Lines 326-330: A similar comparison with the 2019 IPCC refinement and or Mokany would likely be well-cited but is not required.

Lines 333-363: Nice point and supporting analysis.