Scheiter et al present an interesting approach to delineate biomes based on traits information. They use a combination of a number of extensive databases and interesting approaches. Even so, I have a number of major and minor comments.

We thank the reviewer for the helpful and constructive comments. In response to the comments, we will (1) better explain and motivate the analyses on different PFTs and different trait subsets, (2) focus on trait sets with high performance in the main text while moving additional analyses to the supplement, (3) analyze the relation between traits and biomes in more detail to highlight the ecological aspects of our study, and (4) include a schematic figure to illustrate the different steps of our analysis. More details are provided in our responses to the comments.

Major comments:

- A main concern is that some components of the methods do not logically link to each other, which does not help the story line. This specifically applies to two analyses; analyses described in section 2.4 and 2.6.

a. The aim of the analysis in 2.4 and how it contributes to the story line remains unclear to me. You create, for each of the four cases (?) a trait set to work with. First of all, I don't understand why that is needed because you already tested the performance of different complete trait sets in 2.3. So, why would you continue working with subsets of traits if you already know how the entire combination of traits works? Now, you seem to throw away most of the information presented in 3.1 instead of building upon them. That does not only seem unnecessary, but also creates biases (because subsets of traits are used; an argument you make yourself in l. 275). Moreover, given the different selection criteria and procedure for different trait sets/cases, interpretation of these results is made even harder (not possible). Also, I don't see the need for doing it. You use it in 2.5 (but then only for the case 4), but I think that also could have been done directly on the best performing trait sets from 2.3. To me, that would have been a much more direct way to test it and with fewer biases.

As stated by the reviewer, we selected four different sets of traits and conducted cluster analyses informed by each of the 31 biome maps provided by Fischer et al. (2022). In addition, for each trait set, we used traits of different vegetation types (woody, non-woody). We agree that the reasons for conducting this analysis need to be described better, and that we did not use trait combination with the highest performance in the previous sensitivity analysis in 2.3. However, we think that the analysis on trait subsets adds an important aspect to our study for several reasons:

(1) Previous studies used subsets of traits (for example SLA, height, wood density in Boonman et al., 2022). Using the same subsets of traits allows comparisons with previous studies in clearly demonstrates the benefit of including a higher number of traits. Such comparisons highlight the value of merging TRY and GBIF, because the availability of trait information for many traits at large spatial coverage improves the performance of biome classification compared to using only three or four traits.

(2) The sensitivity analyses in 2.3 are based on randomly selected sets of traits, and the analysis shows that some traits have a high rank, i.e., they are represented in cluster analyses with high kappa (Figs. 2, 3). Including all traits in the cluster analysis would not allow an identification of relevant traits and a trait ranking. We therefore think that using a systematic selection of traits rather than random selection is reasonable. This selection explicitly builds on the findings of section 3.1 and Figs. 2 and 3 by choosing relevant traits. In addition, we used a PCA to select trait subsets as a more subjective and quantitative method.

(3) We did not use complete trait sets in 2.3 as stated by the reviewer, but only up to 12 out of 33 traits. Fig. 1 suggests that kappa saturates as the number of traits increases such that including more traits does not improve the performance of the cluster analysis substantially. This can be explained by the correlation and redundancy of different traits (Fig. S11); adding correlated traits does not improve the performance of the clustering substantially and a smaller number of traits is sufficient.

This approach also agrees with the common practice of variable selection in statistical modelling. Moreover, our results agree with the coordination (redundancy) of traits as found in Wright et al. (2004), Diaz et al. (2016) and Bruelheide et al. (2018).

(4) Using a lower number of traits (i.e., 6 instead of 12 or even more) facilitates the exploration of trait combinations that characterize different biomes (in the analyses in section 2.5) and trait covariation. See also response to comment regarding trait covariation and confusion matrices below.

In the revised version of the ms, we would for the above-mentioned arguments prefer to keep the analysis on the trait subsets and (1) better motivate the analysis based on the points described above, and (2) follow the suggestions made by the reviewer and include trait subsets with higher performance (kappa>0.6 in the sensitivity analysis in 2.3). We will further use this trait set with higher kappa for analyses in 2.5 and 2.6.

b: I don't understand why the analysis in 2.6 is useful to do given that you started with using trait maps to fit biome distributions based on complete biome maps. Why would you then derive additional trait maps, using a different procedure with other input data than used elsewhere in the paper. Moreover, you already have 31 biome maps to test performance against. So, why create another one? I would say the study is about understanding biome distributions based on traits (a story already told by the kappa values) and not about predicting/extrapolating biome distributions. In other words, the aim and position of this analysis in the story line is not clear to me.

We have indeed not been entirely clear on the motivation of this analysis. The biome maps have a spatially continuous coverage, but the trait maps used for the cluster analyses derived by combining TRY and GBIF do have spatial gaps in places where we have no observations (see Fig S1 for spatial coverage of traits). Therefore, the modeled biome maps derived from the trait map-based cluster analyses do not cover the entire land surface and kappa values only refer to areas with trait data. Thus, we think that a trait-based global biome map with continuous coverage is a valuable outcome of our study, and it represents a novel observation-driven biome map. To create trait-based biome maps with continuous spatial coverage, an additional extrapolation step is required; this is done by using niche models and bioclimatic variables in section 2.6. As we conducted cluster analyses for many different trait combinations supervised by 31 different biome maps, we decided pick only one combination of a biome map and trait combination for this extrapolation to global scale.

We will describe and motivate this analysis in more detail in the revision and ad a map showing the spatial distribution of available trait data in the main text. We will further add a schematic figure to illustrate the different steps of our analysis.

- Another methodological issue is that I do not understand why the authors thought it important to first use the traits in combination with species distribution models to make trait maps and then to couple those to the biome maps instead of using the locations of the trait observations (g. using the original TRY data, possibly aggregated to the 0.5 degree pixels of the locations) directly? Coupling individual locations can be done to calibrate and validate biome models and avoid the major uncertainties involved in creating/ extrapolating the trait maps, i.e. i don't see that necessity. I wonder to which extent these uncertainties contributed to the low Kappa-values of the predictions.

We did not use species distribution models to create trait maps (in contrast to other studies that used species distribution models or machine learning methods to extrapolate traits to global scale). Rather trait information from TRY was extrapolated to larger areas by linking observed traits and observed species distributions from GBIF. This method was presented previously (Schiller et al. 2021, Wolf et al., 2022) and showed unprecedented agreement with independent observational data (see section 2.1). The advantage of using these trait maps is that it includes gridded trait data for 33 different traits at much larger spatial scale than the original site data of these traits in TRY (possibly aggregated to 0.5 degree), and all traits are available for the same spatial extent. Further, TRY only

represents trait observations obtained from single plants and these observations are known to be not representative of species distributions and plant communities, and have a large spatial bias (Kattge et al. 2020). The trait maps obtained from coupling TRY and GBIF represent mean trait values of entire plant communities (Wolf et al. 2022). To be able to do our analyses only with the original TRY site data, instead of gap-filled TRY data, we would need to select sites (or 0.5 degree grid cells) where all 33 traits are available. We suspect that the number of sites would be low and not cover all biomes included in the analysis. This would make it difficult or even impossible to generate global biome maps. We are confident that this analysis can only be conducted because trait data were extrapolated as described in section 2.1.

We will better describe the reasons for using the trait maps obtained from TRY and GBIF in the revised version of the manuscript.

- I had hoped that the authors would have focused more on the ecological interpretation, rather than on the methodological aspects.

This is a valid point. The study currently focusses more on methodological aspects. We will revise the manuscript according to the suggestions provided below.

a.: This already starts from the presentation of the data. With slightly different analyses and visualization, potentially a lot more ecological insights might have been gained. For instance, fig 7 presents the mean trait values for some biomes. However, these are individual traits while you tested how different trait combinations allow distinguishing among biomes. However, nowhere we learn about those combinations (their synergies and their redundancies). For instance, I would have liked a so-called confusion matrix to see how for a given trait set, biomes were predicted properly or not and whether this is consistent among the different (best-performing) trait sets: which trait combinations (in addition to single traits) lead to the best predictions? Is that consistent between trait sets? Which trait combinations allow distinguishing e.g. ''subtropical forest'' (just to name one) from other forests? Is that consistent through different trait sets? I think the analysis has the potential to bring such answer and thus ecological understanding, but that remains unexplored.

As suggested by the referee, we will conduct additional analyses to explore trait covariation in different biomes and which trait combinations allow distinguishing biomes. We will prepare figures showing the 2-dimensional trait space of different trait combinations and where different biomes are located within this trait space. As we cannot provide such plots for all trait combinations, we will select traits based on their loadings in a principle component analysis as well as on our trait ranking. In addition, we will provide information on the performance of the clustering per biome and illustrate the results in confusion matrices. Using these analyses, we will answer the questions raised by the referee. We already performed such analyses, but given that we used 31 different biome maps and 33 different traits, it is not possible to provide all results. Single cases need to be selected and further analyzed. Therefore, we constrained our analyses in 2.4 to only 6 traits in the original version of the manuscript to reduce complexity. Using the best models with all traits as suggested in a previous comment makes it almost impossible to provide figures of the 2-dimensional trait space for all combinations such that a trait selection based on a PCA or the trait ranking is necessary.

b: Also the discussion section is now almost entirely focused on the methods instead of the question why the analysis is a useful thing to do and what we learn from it: I would be interested in seeing an ecological interpretation for how and why the results of the analysis help global ecological understanding of biome distribution and functioning. That would have led to a much more interesting story. In the current set-up, the discussion section does not provide much insight.

We agree that the study has a strong focus on methodological aspects and that there is potential to better highlight ecological aspects. In response to the previous comment, we will add analyses on covariation of different traits in biomes and we will revise the discussion to interpret these novel results from an ecological perspective. In addition, we will give a better perspective how our methods and results can advance our ecological knowledge.

Minor comments:

1. *The abstract does not help to tell the story. The method applied is not clearly explained, i.e. which steps were taken. Also, the role of the 31 biome maps was unclear (in the abstract it seemed an input, while it is used to train the data. Also, a clear take home message is missing: ''we can make biome maps'', but it is not explained why that is important or what is the added value of this study.*

We will revise the abstract to address the points raised by the reviewer. The biome maps were used to train the models (which means that they are an input?). Further, the performance of models was tested against those biome maps. While it is often split into training and testing data, we used the entire data set for training and testing, because we did not use the models/cluster analyses for predictions with novel input data sets. The main aim of the study was indeed to show that traits can be used for biome classification and not so much the ecological implications. Based on the additional analyses described in response to the previous comment, we will strengthen the ecological aspects of the abstract and the added value.

We think that the trait data and the classification methods proposed in our study have huge potential to further explore trait covariation in biomes, trait diversity in biomes and to improve the parametrization of PFTs in vegetation models. Yet, we think that such analyses go beyond the scope of the presented study, focusing more on methodological aspects and how this can be used to generate new knowledge. While we will discuss these aspects in more detail in the revised version of the manuscript, a full in-depth analysis of such aspects should be conducted in a follow-up study.

2. *I would have appreciated an explanation/argumentation on why choosing these trait maps instead of other maps, i.e. what is the conceptual advantage to these maps? (Also in light of the discussion section where it Is mentioned that different trait maps could have led to different outcomes)*

The advantage of the utilized trait maps is that by combining TRY and GBIF, they cover an unprecedented number and coverage of traits with global cover than covered in the original TRY data (Wolf et al., 2022). Therefore, our analysis is not constrained to a low number of traits commonly used in trait-based analyses (including height, SLA, leaf N, wood density), which allows a more systematic assessment of relevant traits for biome classification and how the performance of classification is related to the number of traits used for clustering. Previous studies where this approach was developed and presented showed better agreement in presenting global patterns of community-weighted traits with independent trait data (sPlotOpen) than other products (Schiller et al. 2021, Wolf et al. 2022). In addition, TRY data are species- and site-specific (Kattge et al. 2020), while the trait data used in our analysis represent entire plant communities. By using GBIF data, we implicitly assume that the observations in GBIF mirror the actual abundances of the species so that we can calculate community weighted means of traits. With the TRY data alone, we cannot assume that observations mirror abundances. We are therefore convinced that using this novel dataset is suitable for the purpose of our study. We will better explain and motivate our selection in the revised version.

3. *With some self-advertisement; Verheijen et al. NewPhytol 209: 563-575 also evaluated which traits could be used best to distinguish among biomes, albeit which is much smaller dataset and fewer traits (And a different method). Interestingly, similar traits popped up as important.*

Thanks for sharing this resource, which is indeed very valuable to back-up our findings. We will cite the paper as suggested and refer to the results presented in this study. It is indeed interesting, that similar traits are important in both studies.

4. *At some parts in the methods section it is unclear whether all individual 31 biome maps were tested independently or whether an aggregate was used. It seems that you tested each map individually, but a better explanation throughout the methods on this would have been appreciated.*

All maps were tested independently, and no aggregation was used. As the biome maps were created by different methods and using different data sources, aggregation is challenging. We will make this clear in the revised version.

5. *The role and use of PFTs is confusing to me. You derive different trait maps depending on growth form/PFT and clustered them into four cases in 2.2. Then, it seems that trait clusters were made for each of those four cases (if I understand the methods correctly) in 2.3. Then, later in 2.3 you seem to continue with the outcomes of case4 only. This is not clearly described and also the reason why you do this, remains unclear to me. Also the role and use of PFTs is confusing to me. In general, i don't think you tested different combinations of PFTs, but trait maps of woody vs non-woody vegetation.*

Yes, the analysis in 2.3 was conducted as described by the reviewer, and we conducted cluster analyses for four different cases. In further analyses, we continued with case 4 only because our analyses showed that kappa values are higher when traits of both grasses and woody plants are used. For further analyses we aimed at using traits and trait combinations with high kappa values. The PFT-specific traits values were derived by combining only those TRY and GBIF observations that correspond to the defined PFT and represent mean trait values of all grass and/or woody species within the 0.5° grid cells. We will describe this point more clearly in the revised version.

Regarding the last point of the comment: for the first two cases with only grasses or only woody plants, we test indeed woody vs non-woody vegetation. We tested if the performance of cluster analyses using only grasses (or woody plants) are better in grass-dominated biomes (or forests, respectively), but there was no robust pattern (results not shown in ms). For the two other cases, both grasses and woody plants are included but they were aggregated in a single trait value representing all grass and woody species in a grid cell (in case 3) or considered separately (in case 4). Using these two approaches, we could show that separately considering the trait space of woody and non-woody plants enhances the differentiation of biomes.

We will better explain the reasons for using different PFTs in the revised version. We will also move the results from this analysis into the supplement (particularly Fig. 4) to focus on one case (case 4) in the main text.

6. *I understand you have a personal interest to compare to aDGVM2 results, but to me those comparisons do not add much to the story.*

In the discussion, we compare our results to previous results, derived from both empirical analyses and modeling. We prefer to keep this modeling aspect in the manuscript and compare both approaches. Given that the aDGVM2 results represent a process-based approach, while the citizen science approach of this paper is a data-driven approach. Both approaches show that a biome

characterization based on traits is possible but that the data-driven approach still outperforms trait distributions from theory. We think that this finding is very interesting. We will revise this paragraph such that it compares data-driven vs model driven trait and biome maps. We will also refer to the Verheijen et al. (2016) study.

7. *The title of section 4.2 is strange and seems wrong.*

This section describes how the choice of the observation-based biome map used to train the supervised biome classification (here 31 different maps provided by Fischer et al., 2022) influences the performance of the cluster analysis and agreement of the observation-based biome map. Therefore, we think that the title is correct. Nonetheless, we will revise it for clarity.

8. *I am surprised that the kappa does not go beyond 0.6. I had hoped that higher kappas would be feasible. This is not discussed, but I would be interested (e.g. instead of the current 4.3) if such aspects of model performance would be discussed and interpreted.*

Good point. The kappa values below 0.6 in Fig. 4 can be explained by (1) the selection of a subset of traits and (2) the aggregation of all biomes. When using different trait combinations and calculating kappa per biome, higher kappa values can be obtained. For example, the kappa goes beyond 0.6 in the sensitivity analysis in Fig. 1 when at least 8 grass and woody plant traits (i.e., 16 variables in total) are included in the analysis. When considering biomes individually (Figs. 6, S3-S6), kappa values can exceed 0.75 (for example tropical forest or evergreen forest in Fig. 6). In response to a previous comment, we will include an analysis with a trait combination with a higher kappa value exceeding 0.6 and we will better explain in the discussion why the kappa values do not exceed 0.6 considerably.

9. *With respect to figure 3; what does it tell us if ranks of certain traits vary with number of traits (while ranks of other don't)? i.e. what is the message/understanding this figure gives?*

The figures shows that the trait ranking is robust for varying numbers of traits included in the cluster analysis, i.e., the same sets of traits show high or low rank for different numbers. This implies that for biome classification it is more important to select an appropriate set of traits instead of selecting a high number of (randomly selected) traits. Selecting a high number of inappropriate traits can imply low performance of the clustering. Traits are often coordinated and/or correlated and therefore different trait sets may lead to different ranks and performance. The ranking allows systematic selection of appropriate traits for biome classification, and the selection of traits in the analyses described in section 2.4 is based on this ranking. We will clarify the value of the figure and the selection of traits in 2.4 in the revised version of the manuscript.

References

Boonman CCF, Huijbregts MAJ, Benitez-Lopez A, Schipper AM, Thuiller W, Santini L (2022) Trait-based projections of climate change effects on global biome distributions. Diversity and Distributions, 28, 25–37.

Bruelheide H, Dengler J, Purschke O, et al. (2018) Global trait–environment relationships of plant communities. Nature Ecology & Evolution, 2, 1906–1917.

Diaz S, Kattge J, Cornelissen JHC, et al. (2016) The global spectrum of plant form and function. Nature, 529, 167–171.

Fischer JC, Walentowitz A, Beierkuhnlein C (2022) The biome inventory - standardizing global biogeographical land units. Global Ecology and Biogeography, 31, 2172–2183.

Kattge J, Bönisch G, Diaz S, et al. (2020) TRY plant trait database - enhanced coverage and open access. Global Change Biology, 26, 119–188.

Schiller C, Schmidtlein S, Boonman C, Moreno-Mariinez A, Kattenborn T (2021) Deep learning and citizen science enable automated plant trait predictions from photographs, Scientific Reports, 11, 16395

Verheijen LM, Aerts R, Bönisch G, Kattge J, Van Bodegom PM (2016) Variation in trait trade-offs allows differentiation among predefined plant functional types: implications for predictive ecology. New Phytologist, 209, 563-575.

Wolf S, Mahecha MD, Sabatini FM, et al. (2022) Citizen science plant observations encode global trait patterns. Nature Ecology & Evolution.

Wright IJ, Reich PB, Westoby M, et al. (2004) The worldwide leaf economics spectrum. Nature, 428, 821–827.