Response to the reviewer#1's comments

We thank Reviewer #1 for dedicating significant effort to the careful review of our rather extensive paper. In the following, the reviewer's comments are presented in standard black font, while our responses are in italic and blue.

The submitted manuscript 'The new plant functional diversity model JeDi-BACH (version 1.0) in the ICON Earth System Model (version 1.0)' describes a new land surface model that merges functionality of the JSBACH model, which is the standard land surface scheme of the ICON ESM, with a more complex representation of terrestrial vegetation, based on the JeDi model, thereby replacing the plant functional type (PFT) approach with a higher resolution of functional traits.

The idea behind this manuscript is interesting and relevant, as effects of biodiversity on climate have rarely been studied so far. However, the manuscript has several substantial deficiencies (see also detailed comments):
1) The text is overly long, most of the methods and the description of the extended analyses should be moved to the appendix. Apart from that, the text is mostly well written.

Following the reviewer's suggestion we will go through the paper and see where we can shorten the text and/or move passages to the appendix.

2) The study substantially lacks validation. It is heavily based on the work by Pavlick et al. (2013), but, unlike Pavlick et al., the authors omit evaluation of global patterns of vegetation properties (e.g. NPP, Biomass). At least it should be demonstrated that the model is able to predict a realistic distribution of vegetation cover and spatial pattern of trees and grasses. It would be straightforward to add a stand-alone simulation to compare JeDi-BACH to the original JeDi model.

Indeed we do not validate our new model in the paper. But this is by purpose: The paper is a model description to be used for future reference. Maybe the impression that the paper would need validation is evoked because we have included also simulation results. But the simulation results shown address only aspects of the model that arise from qualitatively new features that should be independent of the quality of the simulation results compared to observations. We feel that a demonstration of such — partially unexpected — model behavior makes the peculiarity of this setup more clear than a mere documentation of the implemented concepts and formulas could do, thereby also illustrating more convincingly why this new implementation is scientifically of interest. Obviously, our short comment that our "first land-atmosphere simulation experiments with this new model to explore its behaviour" (L65) made this point not sufficiently clear — we will remedy this point.

Concerning the reviewer's suggestion to add validation by offline simulations: Currently our new implementation of JeDi is only a prototype that would indeed need tuning and validation when aiming at quantitative results. As Pavlick et al. (2013) have demonstrated, the JeDi concept is upon proper tuning capable to produce realistic results when driven with realistic climate. Being interested only in a coupled setup, there is in our opinion no point in investing work in demonstrating this anew. Moreover, investment in tuning of an uncoupled setup would be a waste of time, because the coupled setup had to be tuned differently to cope with the climate biases and the peculiar internal variability of the atmospheric host model. But tuning

the coupled model is currently beyond our capabilities because of the enormous computational resources needed, likely available only with one of the next generation of mainframes (see the discussion in section 3.2 of our manuscript). We will make this clear in the resubmitted manuscript.

3) The conclusions of the study are not supported by the outcomes of the model experiments. Specifically, the 'stabilization' of climate at higher values of initialized functional types (higher potential diversity) is, in my opinion, rather a statistical effect and not the result of an ecological mechanism (see detailed comments below).

We thank the reviewer for this comment. The questioned 'stabilization' appears in the manuscript in two different contexts. We are not sure, which of those two cases is meant here. Therefore we give here separate answers to both of them.

First, we diagnose a 'stabilization' from our sensitivity experiments (see L1192), where we conclude that ecosystem adaptability is the reason why there is no significant change in regional climate despite plant-related parameter changes. This conclusion was questioned by the reviewer in the detailed comments arguing that our statistical significance test is not sound. We respectfully disagree and suspect that the reviewer's critique may be caused by a misunderstanding regarding how the significance test was conducted (see also comments #31 and #32 below). Specifically, we tested at each grid point whether changes in climate variables between the different sensitivity simulations and the control simulation were statistically significant. For this we performed a Mann-Whitney test, a widely used method for assessing whether two distributions differ significantly, was applied. We suspect that a misunderstanding happened because we did not make explicit that for the test we use the whole time series of simulation data at each grid point instead of "only one sample" as understood by the reviewer (see reviewer's comment #31 below). Therefore in our opinion the test is sound and indeed shows that the parameter changes cause only insignificant differences in climate, what we diagnosed as a 'stabilization' by high diversity. The mechanism underlying this 'stabilization' is different from being simply 'statistical' as evidenced by contrasting shifts in ecosystem composition between a grid cell in the Congo and one in the western Sahel (explained in section 5.3 of the manuscript). See also our more detailed answers below (#31 and #32). — In a revised version of the manuscript we will make explicit on what simulation data the significance test is based.

Second, we use the term 'stabilization' for the convergence of terrestrial climate seen when increasing potential diversity in our first set of diversity experiments (section 4, particularly line 1017). The reviewer argues that this convergence (or "stability") in climate is merely a sampling effect (see the reviewer's detailed comment #29). To some extent, we agree—greater sampling of strategies does indeed lead to convergence in simulation results. However, we believe this is not the whole explanation. We observe a systematic tendency for the terrestrial climate to shift toward wetter and cooler conditions as diversity increases (Figs . 10 and 11 in the manuscript). A mere sampling effect cannot explain why temperature develops towards a low value and precipitation towards a high value with increasing diversity, from mere sampling one would expect convergence towards a value somewhere in the middle of the range of values seen at low diversity. We think that this behaviour is explained by the biomass-scaling hypothesis implemented in JeDi: Strategies with larger biomass are weighted stronger in the calculation of ecosystem fluxes of a grid cell. Accordingly, as random sampling expands with increasing diversity, those strategies with high biomass density increasingly dominate

ecosystem fluxes, leading to those systematic shift in terrestrial climate. Understanding the biomass-scaling as competition for space, this is the ecological mechanism explaining the convergence. Nevertheless, we agree that the term "stabilization" is misleading here, denoting this behaviour as 'convergence' will be sufficient and we will thus change the manuscript accordingly. "

1. L30 It would be good to clearly define 'diversity' from the start.

   Reply: We agree with the Reviewer's comment #1 and # 22. We will move the definition of 'diversity' employed in the paper (see L886) to the introduction:

   'In the model world of JeDi, "functional diversity" is meant to be represented by the number of PGSs in a simulation.'

2. L126 Describing leaves of temperate trees as low-cost is questionable. Also, following this argument, there should be no needle-leaf trees at high latitudes. Usually, leaf traits follow a trade-off between long-lived but slowly growing and short-lived but fast growing (the "leaf economic spectrum", which is mentioned later in the manuscript). This should be clarified here.

   Reply: Indeed, the explanation for the emergence of deciduousness in temperate regions is incomplete. We will replace this sentence by a more complete formulation:

   "In temperate regions the multi-annual carbon balance between carbon investment into the growth of leaves, the resulting photosynthetic carbon gain, and the maintenance costs of leaves turns out to be better for a deciduous growth strategy by developing thin and fast-growing leaves and to shed them in winter, while in boreal regions, because of the shorter period for photosynthesis, carbon investment into more expensive leaves kept for several years is more advantageous (Kikuzawa and Lechowizc, 2011). "

   The purpose of this passage is to give examples for the environmental selection of growth strategies, referring to the leaf-economic spectrum would not suit this purpose — even though the reviewer is surely right in noting the close relation to this topic.

   Kikuzawa, Kihachiro, and Martin J. Lechowicz. Ecology of leaf longevity. Springer Science & Business Media, 2011.

3. L136 This makes no sense to me: The PGS are defined as differing in at least one trait value, albeit by a small amount. How can they be functionally redundant when they differ in a functional trait?

   Reply: The term "functional redundancy" refers to an ecosystem property—namely, "the degree to which the loss of an individual species impacts overall ecosystem structure and function" (Biggs et al., 2020). In contrast, a "functional trait" is a property of a PGS. Hence, it seems quite natural to assume that different PGSs with only slightly differing trait values may contribute similarly to ecosystem structure (e.g., in terms of carbon and water fluxes), and may therefore be considered functionally

redundant. To avoid confusion, we will include an explanation of the term "functional redundancy" in the text.

Biggs, Christopher R., et al. "Does functional redundancy affect ecological stability and resilience? A review and meta-analysis." Ecosphere 11.7 (2020): e03184.

4. L138 How is 'sufficiently complex' defined? Trade-offs should limit the number of PGS, so increasing the number or complexity of trade-offs should reduce the number of strategies, not increase them.

   Reply: We thank the reviewer for this comment. After reading the reviewer's remark, we realize that our formulation was not sufficiently clear and may give rise to two opposing interpretations.

   First, we agree with the reviewer that the number of trade-offs determines the number of possible growth strategies: More trade-offs means that the trait space has a higher dimension (more trait values). Hence, the number of surviving strategies found for a larger trait space is smaller than the number found for a less dimensional trait space when considered at equal potential diversity. This happens because for the same size of randomly sampled strategies it becomes less likely to sample the "suitable" trait combinations for survival in a higher dimensional space. We assume that this reasoning underlies the reviewer's comment.

   However, if we do not consider a fixed number of sampled strategies, increasing the number of trade-offs in fact expands the range of possible new behavioral strategies that can emerge. Roughly speaking, each trade-off introduces an alternative in the behavior of a PGS. With N trade-offs, there are approximately 2^N possible behavioral combinations. Adding another trade-off therefore doubles the number of potential growth strategies.

5. L140 It is unclear why the distinction between grasses and trees is given so much weight here. There are many more possible categorizations of plants, e.g. shrubs, which do not have a stem, in contrast to trees, but do have woody roots. What about different photosynthesis pathways, leaf phenology etc.? This focus on woody tissue seems arbitrary to me.

   Reply: We implemented this distinction between trees and grasses because it is eminently important when coupling a vegetation model with a climate model: Forests and grasslands have very different surface properties. For instance, the albedo values particularly in the presence of snow, lead via the albedo-snow feedback to rather different surface climates in boreal winter (e.g. Loranty et al., 2014). This reason for implementing an explicit grass type is indeed not mentioned in the paper and will be added.

Loranty, M. M., Berner, L. T., Goetz, S. J., Jin, Y., & Randerson, J. T. (2014). Vegetation controls on northern high latitude snow-albedo feedback: observations and CMIP 5 model simulations. Global change biology, 20(2), 594-606.

6. Tab.1 I suggest to remove the term "suck up water" for root water transport, and replace it by an appropriate wording, throughout the text.

   Reply: We will modify the text accordingly.

7. L151 It is unclear what a conceptual parameter is and how it differs from an actual plant trait.

   Reply: Indeed, this distinction is not well explained. What we intended to clarify is that not all trait parameters have a directly measurable counterpart that can be obtained from trait databases, but that some trait parameters are model-specific ('conceptual'). We will revise the text accordingly to make this distinction clearer.

8. L156 'five traits t1 to t4' ?

   Reply: We thank reviewer for spotting this error. We will revise the text accordingly.

9. L159 Above, the authors write that survival is determined by the carbon storage level, here a trade-off to reproduction and growth is mentioned. This seems to be inconsistent.

   Reply: This remark refers to the sentence "Second, to imitate life history strategies, a trade-off among growth, reproduction, and survival is introduced." This indeed causes confusion because its unclear what a trade-off with "survival" should mean. Therefore we will revise the sentence and be more clear by writing "... trade-off among growth, reproduction, and allocation to storage assuring survival is introduced."

10. L156-164 The description of the trade-offs is not well justified. Why were these trade-offs chosen, is there a body of literature to support their relevance over others, and how are they parametrized?

    Reply: The implemented trade-offs are those from Pavlick et al. (2013), who slightly extended the original set of trade-offs used in the JeDi implementation by Kleidon & Mooney (2000). We recognize that the selection of trade-offs can influence model behavior. However, given that the re-implementation of JeDi into ICON-ESM was already a substantial task in itself, we took a pragmatic approach and adopted the existing set, knowing from Pavlick et al. (2013) that it had produced convincing results in previous simulations. That said, it is indeed conceivable to explore alternative sets of trade-offs.

11. L230 It is not clear in which way the biomass is computed, such that it can be compared among the PGS. Is it estimated per m2 ground covered by a PGS, i.e. specific biomass? The term 'biomass density' sounds like it, but a unit for M should be provided. Without a common reference unit, summing up biomass values per grid cell makes no sense. If it is biomass per area of ground, the biomass ratio hypothesis as implemented in JeDi-BACH may lead to biased estimates in ecosystems that consist of a mixture of trees and grasses. In Savannas, trees may have a high specific biomass but low abundance in the ecosystem. Thus, the biomass ratio hypothesis would predict a large share on ecosystem biomass and productivity, but this is not consistent with observations.

    Reply: The unit of biomass is mole carbon per square-meter ground (see table B3 in the appendix of the manuscript). As noted in lines 231-239 of the manuscript, we fully agree to the reviewers remark on a potential bias between trees and grasses when using (as we do) biomass density in combination with biomass ratio theory. And the savanna example given by the reviewer is indeed well illustrating the problem. It is known that savanna ecosystems are shaped by fire disturbances that affect woody vegetation differently than grasses. A similar problem arises at high latitudes where wind break takes the role of fire to shape the tundra vegetation. Model components for fire and wind break disturbances are available in JSBACH, but we currently did not take them over into JeDi-BACH. From the reviewer's comment we now recognize that it's indeed not immediately clear how to merry the JSBACH disturbance concept with biomass ratio theory — obviously a task for further model development. We will add some remarks on this in the next version of the manuscript.

12. L467 C4 grasses are common, but C3 grasses, too, in many regions of the world. The exclusion of C3 grasses seems arbitrary to me, and needs to be justified. From a conceptual viewpoint, this introduces a systematic bias in the prediction of PGS, since the growth form grass/tree is linked without need to photosynthesis pathways.

    Reply: In Section 2.5, the implementation of the JSBACH C3 and C4 photosynthesis modules are described but obviously we did not provide sufficient clarification on how they are applied within JeDi-BACH. In practice, we use the C3 photosynthesis module for all tree-type PGSs. For grass-type PGSs, a fraction is assigned the C3 module, while the remaining fraction is assigned the C4 module. The specific C3/C4 fraction is prescribed during model initialization. This information will be added to the manuscript for clarity.

13. L482-487 This sounds as if the model was not able to simulate self-shading, so introduction of fapar_max is necessary. However, JSBACH includes canopy layers, so LAI values larger that 10 should already lead to no further gain in light. Even in a big leaf approach, high LAI values lead to very small additional carbon gain due to the exponential extinction. I do not see why fapar_max is necessary, unless the leaf construction costs are generally too low in JeDi-BACH. This needs to be better explained.

    Reply: In JeDi, the carbon allocation among different tissue pools is static (although different for the different strategies) while in reality the distribution of photosynthesized

carbon to the different plant organs changes during the different growth phases. Hence, even if at high LAI additional allocation to leaves gets completely uneconomic, in JeDi further carbon is allocated to leaves because the fraction of carbon allocated to the different tissue pools is static. Thereby, strategies can reach LAI even up to 50 in some cases. A proper correction would be to implement a dynamic allocation scheme. But this would be a quite fundamental modification of the JeDi concept. Instead we introduced the parameter fapar_max to slowdown wasteful investment. The consequence is that, when the fraction of absorbed sunlight (fapar) approaches fapar_max, strategies slow down growing new tissues by an overall down-scaling of the allocation to all tissue pools and more carbon is kept in the storage pool. In this way we can reduce the amount of unrealistic growth strategies resulting from random sampling in JeDi-BACH. Overall, we think that the problem with the unrealistic large LAI values is the static allocation scheme and that higher leaf construction costs — as suggested by the reviewer — are not the source of this problem, even though an increase of such costs to (likely) unrealistically high values may also suppress growth strategies with unrealistically high LAI. But this had the additional downside to generally reduce the chance for survival, even for strategies that by their set of randomly sampled trait values have low LAI by construction.

14. L520 The well-documented value of 2.1 (Wullschleger, 2013) could be used, or it should be justified why 1.9 is more appropriate here.

Reply: Unfortunately, we didn't find your reference (Wullschleger, 2013), but we are aware of Wullschleger (1993), from which the ratios of Jmax/Vcmax—ranging from 1.5 (conifers) to 2.1 (tropical trees) and 2.2 (hardwoods)—are obtained (see Table 3 therein). From Kattge and Knorr (2007), values between 1.5 and 2.7 for broadleaved and coniferous trees, varying with growth temperature, can be found in their Fig. 3. A quick linear regression of the logarithmic model from Walker et al. (2014), based on the parameter values given in their Table 4, yields a value of 1.6. Accordingly, our chosen value of 1.9 falls within the range of suitable values, as does the value of 2.1 preferred by the reviewer. We will add a remark on this in the manuscript.

Wullschleger, S. D. (1993). Biochemical limitations to carbon assimilation in C3 plants—a retrospective analysis of the A/Ci curves from 109 species. Journal of experimental botany, 44, 907-920.
J. Kattge and W. Knorr (2007), Temperature acclimation in a biochemical model of photosynthesis: A reanalysis of data from 36 species, Plant, Cell and Environment 30, 1176-1190.
Walker, A. P. et al. (2014). The relationship of leaf photosynthetic traits–Vcmax and Jmax–to leaf nitrogen, leaf phosphorus, and specific leaf area: a meta-analysis and modeling study. Ecology and evolution, 4, 3218-3235.

15. L527 This further increases the bias in tree vs. grass PGS. I suggest to make either the grasses C3 or at least test the sensitivity of the model to this setup, i.e. use the C3 photosynthesis scheme for the grasses, too, and test to what extent this affects global biomass distribution and NPP.

Reply: We suspect this comment is related to the reviewer's comment #12. The 21 simulations conducted in the diversity experiment include both trees, C3 grasses and C4 grasses. Please refer also to our reply in Comment #12.

16. L540 This statement definitely needs to backed up with appropriate references and discussion. The alternative view is that plants thrive as long as leaf water potential does not drop to critical levels. In that case, low transpiration does not matter, since stomata can simply remain open for diffusion of CO2. It has even been argued that the water transport from the roots to the leaves is not essential for the provision of nutrients, hence further reducing potentially negative impacts of saturated air.

    Reply: Indeed, further explanation is needed here and shall be added to a revised manuscript. We agree to the reviewer's comment that even at low transpiration photosynthesis may continue because under such conditions stomata are typically widely open so that CO2 uptake through them continues. But we think that if the situation of low transpiration prevails, other mechanisms get relevant that lead to a reduction of photosynthesis. One such mechanism may be the provision with nutrients that slows down with a reduction of xylem water transport at low transpiration. The other mechanism may be that photosynthesates (sugars) accumulate in the cells, because the transport to other plant organs slows down with decreasing transpiration. That the accumulation of sugars leads to a down regulation of photosynthesis is well known (Paul and Foyer, 2001). The way we model the effect of down regulation of photosynthesis is an attempt to account for such non-stomatal effects that likely happen at longer time scales (days, weeks) than the time scale at which stomata operate (minutes). Because models as ours are intended to be used at time scales of months to years, the inclusion of down regulation of photosynthesis at high humidity is a decision to give more emphasis to the behaviour at longer time scales than the shorter ones of stomatal operation. Some evidence for such behaviour can be found in (Chen et al., 2022), where the authors document down grading of photosynthesis at low vapor pressure deficit in Amazonia, but we agree that our way of modelling is experimental and needs further testing upon future model development.

17. *Paul, M. J., & Foyer, C. H. (2001). Sink regulation of photosynthesis. Journal of experimental botany, 52(360), 1383-1400.*
    *Chen, R., Liu, L., & Liu, X. (2022). The negative impact of excessive moisture contributes to the seasonal dynamics of photosynthesis in Amazon moist forests. Earth's Future, 10(1), e2021EF002306.*L645 This is a relatively arbitrary choice and the sensitivity of the global biomass distribution and NPP to this choice needs to be tested.

    Reply:It is indeed an arbitrarily chosen value intended to ensure that at least a certain proportion of randomly sampled strategies experience reduced water stress. This threshold value can also be adjusted as necessary. Nonetheless, when conducting production runs with the updated model, the sensitivity of the simulation results to variations in this value should be assessed.

18. L794 While this adaptation saves computational time, it introduces an imbalance into the competition for water and light, if I understand correctly. There seems to be

no impact of LAI of a PGS on the absorption of light by another PGS, but the root length of a PGS will affect water availability of other PGS through the shared soil water pool. This should at least be discussed.

Reply: There is no competition for light nor for soil water in the original JeDi. By the usage of a single soil water pool, we introduce competition for soil water in JeDi-BACH. We do not understand what imbalance the reviwer refers to. Similar to other models JeDi-BACH also lacks other types of competition such as competition for nutrients. We will add the following paragraph in section 2.9 in the resubmitted manuscript hoping that the reviewer's concern is addressed thereby:

"Note that one difference arises from the modeling setup concerning soil water accessibility in JeDi-BACH that is different to the JeDi-DGVM (see also section 3.2). At each grid cell, all strategies share the same soil water bucket so that the soil water availability of one strategy will be influenced by the others. "

19. L803 As the main outcome of this study is the effect of vegetation on the global water cycle, this is a substantial limitation regarding the interpretation of the model results. A stand-alone setup with prescribed meteorological fields is required and the simulated global distributions of biomass and NPP need to be compared to the coupled model run.

    Reply: As mentioned in the second main reply, the simulations in this model description serve only to demonstrate qualitatively interesting model behavior. Accordingly, we do not consider our paper to be a "study" with a defined "main outcome." We are using an untuned model and therefore only explore the qualitative behaviour of the results. Conducting additional simulations to estimate the magnitude of errors would run counter to the purpose of this paper.

20. L808-836 This part is overly long and should be shortened by at least 50%.

    Reply: We will shorten the text in the revived manuscript.

21. L820 Better: 'To address this issue...'

    Reply: We will modify the text accordingly.

22. L886 Indeed, it would be better to make this clear from the start of the manuscript, please change this.

    Reply: We will move the definition of diversity to the introduction as suggested in Comment #1.

23. L908 This distribution looks substantially different from Fig. 8a in Pavlick et al. (2013), and, more importantly, it does not show much similarity to the global species

richness pattern of plants, in contrast to the original JeDi model. This needs to be addressed somewhere.

24. L919 I do not think that this pattern agrees well with Barthlott et al (1996), in particular since the stand-alone JeDi performed better.

    Reply:

    Comments #23 and #24 are closely related, and we therefore address them together. We agree that, in comparison to the diversity patterns reported by Barthlott et al. (1996) and Pavlick et al. (2013), JeDi-BACH primarily captures the general latitudinal gradient, with higher relative diversity emerging in regions characterized by wetter and warmer climates. Due to substantial precipitation biases resulting from the atmospheric model operating at a relatively coarse resolution (see Fig. A1), the current untuned version of JeDi-BACH tends to overestimate diversity in subtropical regions where precipitation is overestimated, and to underestimate diversity in the Eurasian region where precipitation is underestimated. We will make this clearer in the revised manuscript.

25. L922 I strongly disagree with this statement. The value of actual diversity in this model is arbitrary, you only need to increase the number of initial PGS to increase absolute diversity. It is the pattern of relative diversity which is ecologically meaningful.

    Reply: We thank the reviewer for this remark. We fully agree that the absolute values of diversity obtained in our simulations do not have direct meaning when compared to observed diversity. We now see that our formulations in the manuscript are misleading. Because the point we wanted to make with them is only of minor importance we will remove this part in the revised version of the manuscript instead of going into lengthy explanations.

26. L926 I do not think that the higher vegetation coverage in the high diversity simulation is convincing in general. A relative diversity of up to 10% in the Sahara is highly unrealistic, even if these are all grass PFTs. There is no extensive vegetation cover in the Sahara. In contrast, no PGS seem to survive in Western Siberia, which is probably due to bias in the climate model. Again, I recommend to run at least one stand-alone simulation to test if these unrealistic patterns result from the uncalibrated climate model or from deficiencies in JeDi-BACH.

    Reply: As previously discussed, we do not consider offline simulations to be informative at this stage: A coupled model needs to be tuned differently than an offline model to cope with climate biases and different climate variability. Hence comparing offline and online simulation results would give no clear answer whether "the unrealistic patterns result from the uncalibrated climate model or from deficiencies in JeDi-BACH". The only thing that we can do at the current stage of model development is to document the unrealistic biases in simulation results with the current model version.

27. L966 While the convergence is expected at a sufficient number of initialized PGS, what surprises me is the low standard deviation of the distribution. For the high initial

diversity ensembles, it looks like more than 50% of all surviving PGS have a CWM value around 0.5, which means that there is little selection towards certain trait values happening in the model. It is clear that the variation of climate at the land surface likely covers a large part of the potential trait range, and summing up CWMs from different regions of the world will drive the distribution to the mean value. However, climate regions are not equally distributed, which means that CWMs from regions that contain many grid cells should have a larger influence on the distribution. I would like to see the global pattern of the CWM(t3).

Reply: *(1) The magnitude of the standard deviation has no intrinsic meaning here; whether it is small or large depends on the choice of the scaling factor applied to t3 in Eq. (13), which can be selected almost arbitrarily. The value of 20 used in the equation was chosen to be sufficiently large so that t3 values in the range of 0 to 1 fully cover the realistic range of temperatures at the start of the growing season. Any larger value would serve equally well. As such, the standard deviation could be made arbitrarily small depending on the scaling factor. The same holds for the mean value: If we would choose a different upper and/or lower value in Eq. (13) for the scaling range, the mean value would shift away from 0.5.*

(2) Regarding the question of whether trait selection actually occurs: we would like to emphasize that the distribution shown is that of CWM(t3), not of t3 itself. These community weighted means are single values for each grid cell, while within each grid cell there are many different t3 values that have contributed to CWM(t3). Therefore the CWM(t3) distribution contains only limited information about the spectrum of t3 values found among surviving PGSs globally.

(3) In view of these remarks, we do not agree that the presented distributions justify the conclusion that there is only "little selection towards certain trait values happening in the model." As requested by the reviewer, we show the global

distribution of CWM(t3) of each ensemble member in the following figure.

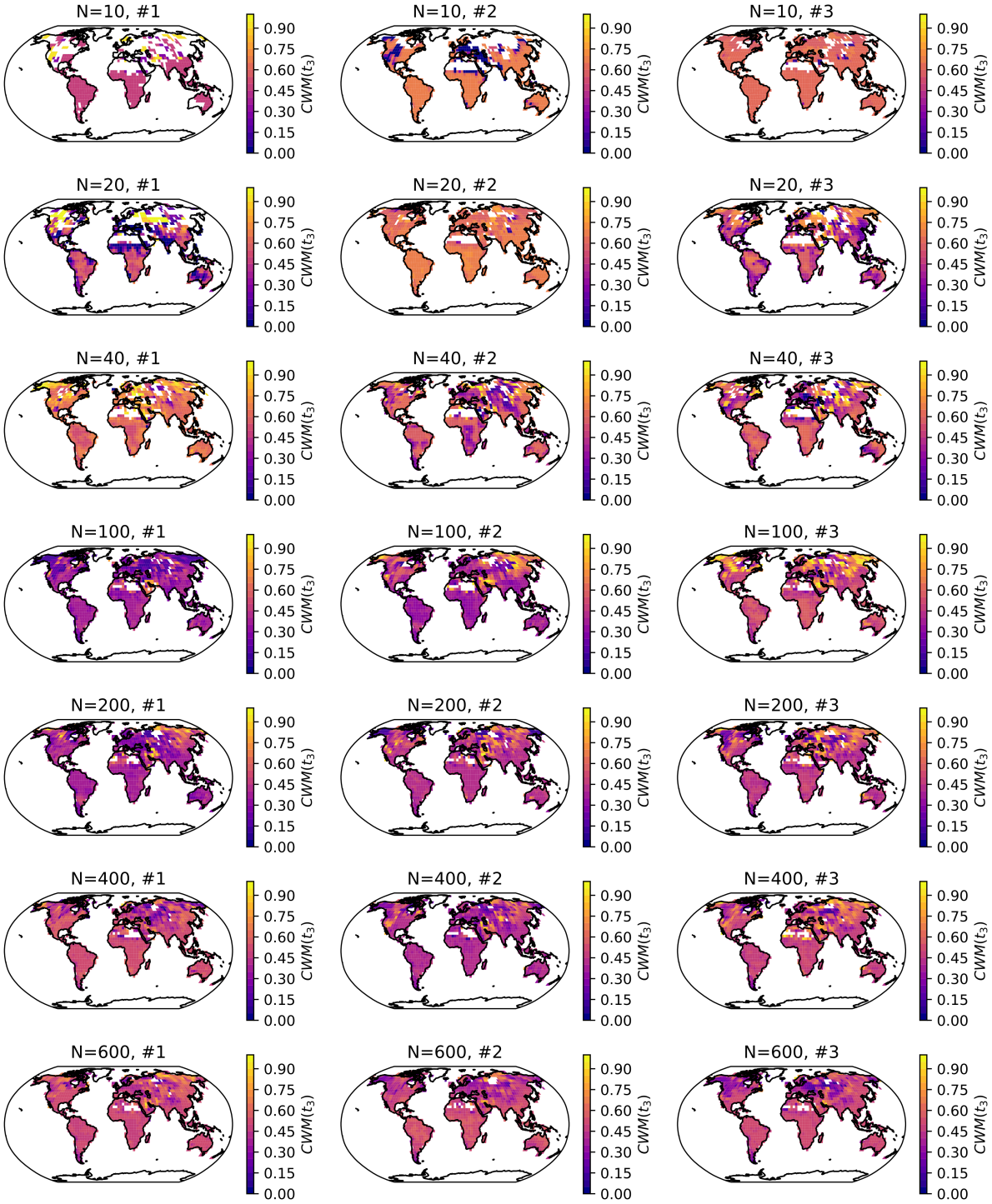

Figure 1: Global distribution of CWM(t3) of tree strategies for the three ensemble members at different potential diversity. A value of t3 closer to zero indicates that a strategy can start the growing season at lower air temperatures compared to one with a t3 value closer to one (see Eq. (13)).

28. L982-999 It is not clear to me why the idea of 'high biomass islands' is necessary here. For any given environment, there will be a certain number of locations in the multidimensional trait space that allow survival of the corresponding PGS. As soon as

the number of initialized PGS is large enough to sample all these regions, the results of the ensemble simulations will converge, as a more frequent or 'dense' sampling of these regions will not result in further functional variation, but simply a higher number of similar PGS. If there is more to that model outcome than a purely statistical effect, it should be more clearly described.

Reply: The description of the reviewer is also how we understand at an abstract level how JeDi operates except for one point. Insofar we think that this abstract description of the operation of JeDi is helpful.

The point where we disagree concerns the "idea of high biomass islands" criticized by the reviewer. By the biomass ratio hypothesis those PGSs with a high biomass density dominate ecosystem functioning. This hypothesis is implemented by considering community weighted means. To understand why its not the islands of surviving PGSs that are relevant for ecosystem functioning (denoted as "locations .. that allow survival" by the reviewer), but those islands of surviving PGSs that in addition have high biomass density, one must consider the effect of the community weighted means on the frequency distribution of PGSs properties: To dominate the distribution of community weighted means it is not sufficient to have a high property value, in addition the associated PGSs must have a high biomass density.

This is particularly important when considering community fluxes. Here we take the NPP "property" as an example: a high biomass density is typically a result of a high NPP. Hence, by the community weighted mean, those PGSs with high NPP dominate the distribution of NPP not only because they have high NPP, but also because of the weighting with their high biomass. Thereby biomass roughly enters quadratically when calculating community weighted mean values. In the trait space of all PGSs it is therefore not sufficient to cover the regions of surviving PGS (as suggested by the reviewer), for ecosystem functioning it is important to cover the regions of surviving PGS that in addition have high biomass density. We think that once the random sampling covers such regions of high biomass density, those regions with low biomass density become less relevant in the sense of community weighted means. Hence there is also a non-statistical effect that matters here.

The following figure demonstrates the effect of biomass scaling at high diversity (potential diversity of 600): The left plot shows the abundance distribution of CWM(t3) across all grid cells, where t3 is weighted by biomass. Conversely, the right plot illustrates the abundance distribution without this weighting, depicting the distribution of the mean of t3 in the grid cells. While Mean(t3) resembles a Gaussian distribution, as expected from purely statistical behavior, the distribution of CWM(t3) is non-Gaussian, appearing broader and skewed due to biomass scaling.

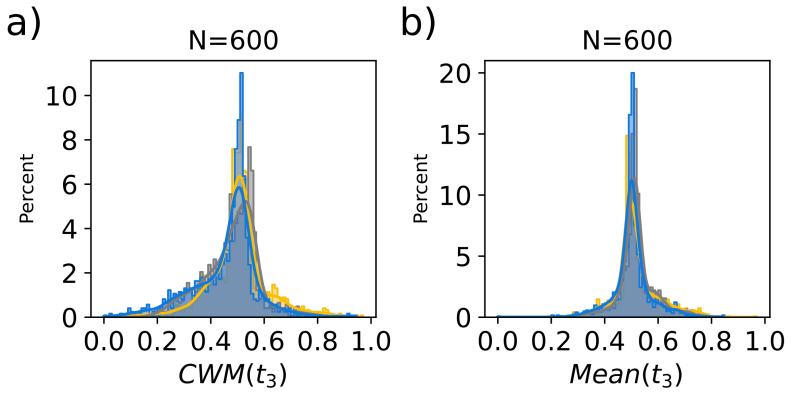

Figure 2 The abundance distribution of (a) CWM(t3) and (b) Mean(t3) from all grid cells at 600 strategies for each ensemble member.

29. L1016 I do not think that this conclusion is justified. The authors write above that trees have a lower chance of survival than grass PGS, and they also note that at low initial diversity a larger part of the global land surface remains free of vegetation. This alone would lead to lower evapotranspiration in the simulations with low initial diversity, as bare soil evaporation is lower than transpiration, and grasses on average are less efficient in accessing water in deeper soil layers. The other effects are simply the result of the well-known moisture recycling effect over the land surface (e.g. Zemp et al., 2017, Nat.Comm.). The term 'stabilization' is misleading here. As I wrote in the previous comment, the convergence of the diversity estimates and, consequently, the global vegetation properties, can be interpreted as outcome of a sufficient sampling of the trait space. Stabilization, however, suggests that there is some mechanism that leads to a certain value of a flux, such as evapotranspiration, at high diversity. If the authors identified such a mechanism, it should be better described.

Reply: *As described in our response to the comment #28, the biomass scaling approach boosts the contributions of PGSs with higher biomass, and meanwhile such PGSs are increasingly sampled at higher levels of potential diversity. Since PGSs with high biomass density typically have higher transpiration fluxes, we expect that community-level fluxes are dominated by these PGSs. As potential diversity increases, PGSs with progressively higher biomass are found, leading to a tendency toward higher ecosystem fluxes. This effect is different from a simple convergence arising from higher sampling.*

However, we agree with the reviewer's comment that in low-diversity simulations, the absence of surviving PGSs in many regions may also contribute to the observed "stabilization" effect. Given this, our claim is at the current state of analysis too speculative to be kept in the paper. Therefore, we will remove the corresponding remark from the manuscript.

30. L1029 In principle, I agree with the authors that the low number of PFTs used in many land surface models may promote the simulation of climatic conditions that are

too sensitive to the parametrizations of the PFTs. However, an important detail is not mentioned here: Fig. 8 clearly shows that the majority of grid cells exhibits CWM values around 0.5 for trait t3. If this is the case for other traits, too, then the selection algorithm in JeDi-BACH mainly chooses 'average' PGS as survivors in most regions of the world. Otherwise, the frequency distributions of CWMs across all grid cells should be broader or more skewed. PFTs are typically parametrized according to 'average' plants, so the mismatch to the JeDi-BACH estimates and the consequences for simulated climate feedbacks may be much smaller than expected.

Reply: As noted in the reply to the comment #27 we do not agree to the reviewer's conclusion that there is only " little selection towards certain trait values" so that the model "mainly chooses 'average' PGS as survivors in most regions". Insofar we think that our conclusion that "the high sensitivity reported in PFT-based modeling studies likely stems from the poor representation of ecosystem complexity" remains valid.We think that comparing the magnitude of the simulated climate feedback to that of standard PFTs offers limited insight. A key difference between the PFT-based approach and the JeDi approach lies in how the 'average' or community-weighted mean (CWM) vegetation is determined. In JeDi, global vegetation emerges from environmental filtering rather than being predefined. The adaptability of ecosystems simulated by JeDi allows vegetation to respond dynamically to environmental changes. Due to this adaptive behavior, we argue that climate feedbacks simulated with diversity change provide a unique framework for investigating how a biosphere capable of adaptation interacts with the atmosphere — something that is difficult to capture using conventional PFT-based models.

31. L1106 A Mann-Whitney-U test is used to check if values drawn from two distributions show a tendency to differ from each other. It is not clear to me how the presented sensitivity analysis justifies the application of such a test, since for each distribution (characterized by ctrl, increased, and decreased parameter value), only one sample is drawn, so there is no basis to characterize the distributions. This should be either removed from the manuscript or better explained.

Reply: We disagree with the reviewer's comment. The Mann-Whitney test—like the Student's t-test—is a commonly used method for assessing whether two distributions differ significantly (see, e.g., Wilks (2011, Section 5.3.1); von Storch & Zwiers (1999, Section 6.6.11)). Regarding the concern that "only one sample is drawn, so there is no basis to characterize the distributions," we suspect this refers to a possible misunderstanding that we would be working with only a single data point per grid cell. However, this is not the case. For each experiment, we have a time series at every grid cell, and each time series defines a frequency distribution of values. This allows us to test, at each grid cell, whether the distributions from the sensitivity simulations differ significantly from those of the CTRL simulation.

Wilks, D. S. (2011). Statistical methods in the atmospheric sciences (Vol. 100). Academic press.

Von Storch, H., & Zwiers, F. W. (1999). Statistical analysis in climate research. Cambridge University Press.

32. L1190 I think the statistical basis for such a conclusion is not sufficient here (see previous comment). While it is logical that a strong change in vegetation structure may affect regional climate via impacts on transpiration and the hydrological cycle, Figs. 13 -15 show marked effects of the parameter variation in all regions of the world, also those with high diversity. Since I do not follow the explanation of the test for statistical significance, I also consider these regions affected by the parameter changes, meaning that high diversity may not necessarily promote a stabilization of regional climate.

Reply: As noted in our reply to comment #31, we think that our analysis for significantly different behaviour upon parameter change is sound. If we understand the reviewer correctly, the remark "I also consider these regions affected by the parameter changes" refers to the regions in Figs. 13–15 that we have marked with dots as not significantly different. Viewed without the context of our statistical test, the interpretation suggested by the reviewer would be reasonable. However, for most of the world, we find that the parameter changes are insignificant (the dotted regions make up most of the world in Figs. 13–15). Accordingly, we still think that our conclusion questioned by the reviewer—namely, that diversity stabilizes regional climate in our simulations—is a valid one.