We thank the reviewers very much for their constructive comments, which have been very helpful to improve the manuscript, as well as for their patience in reporting language errors in the original manuscript.

Reviewer #1 – Harald Bugmann

Q1.1. This manuscript tackles the considerable challenge of parameterizing a highly detailed, process-based model of forest dynamics for 217 taxa towards an application at the regional scale (ca. 30'000 km2; Catalonia, Spain). It is a methodologically interesting and valuable contribution that I think should be published, notwithstanding my concerns detailed in the paragraphs below.

R1.1. Thank you very much for the thorough revision and useful suggestions to improve the manuscript and modeling advice. We carefully considered all your suggestions and tried to follow them as much as possible. Please, see our responses to your comments below.

Q1.2. First, the manuscript tries to make the point that highly detailed models that distinguish a very large number of very detailed ecophysiological processes are useful (or perhaps even needed?) for faithful assessments of climate change impacts on forest dynamics. I am hesitant to accept this. The description of MEDFATE in this paper shows that it is based on multiple assumptions that are highly uncertain. Hence, different assumptions (i.e., different formulations) could have been chosen, leading on the one hand to different parameter requirements and on the other hand to different model behavior at least at some 'intermediate' level (i.e., between the ecophysiological processes and what is ultimately shown in Fig. 3, i.e. ecosystem-level NPP and wood volume). Thus, we cannot really tell whether the model is getting the right answer for the right reasons, or whether errors are compensating each other internally, thus actually enhancing the uncertainty in future projections rather than decreasing them. In this context, it is somewhat marring that the simpler variant of the model is generally closer to the measurements in what the authors call model "evaluation" (although the data are not independent of those used for the parameterization; more on that under the "Specific comments" below). I would have appreciated if the authors had used both model variants for the future projections – they mention computational reasons for not doing so, but even a 1728 h (= 72 days) simulation effort might be undertaken as a "production run" for such a paper. It would be quite instructive to see whether the two variants agree under future conditions.

R1.2. We agree with the reviewer's view in that highly detailed models are not necessarily better than simpler approaches. We have experience in developing both empirical and process-based models of forest dynamics, and we are aware that more mechanistic detail does not always lead to a higher predictive value. A simple model capturing the essential key processes and drivers for forest dynamics may be better than a detailed one. Acknowledging the large uncertainty inherent to complex models and their parametrization led us to choose a robust, empirical model (IPM) as benchmark. Our goal was not to demonstrate that a complex ecophysiological model (with either the advanced or basic sub-model) could better predict forest dynamics than an empirical approach, but to illustrate the challenges of parametrizing trait-based mechanistic models and to provide a generalizable strategy to address them. In our opinion, increased model complexity has positive and negative sides. On the negative side, the additional complexity may be unjustified for a given target, here forest dynamics, and model evaluation puts everybody in its place. Here, our better evaluation results with the basic sub-model advocate for its use, and not that of the "advanced" sub-model, when the aim is to predict forest dynamics only (and the empirical model would also be a good option, see L619-620). On the positive side, the additional results of complex mechanistic models can be of interest in their own right. In this sense, models of forest dynamics with a strong ecophysiological background offer opportunities to predict forest dynamics at the same time as other variables of interest, such as water provision, carbon sequestration or fuel moisture dynamics, using a coherent set of assumptions. Of course, we are aware that intermediate results should also be validated, because as you indicate errors may be compensating, and we have done so in previous publications and will be forced to keep doing this effort. Working with ecophysiological models has the advantage that data exists for some of these 'intermediate' results, and we plan to conduct large-scale evaluation of leaf area dynamics predicted by MEDFATE (L647-649). The two sub-models differ in the detail of water balance, energy balance and photosynthesis processes, which can lead to diverging predictions of forest dynamics even if the assumptions taken regarding carbon balance, growth, mortality and recruitment are the same. For this reason, and because several parameters of the basic sub-model do not map easily to measurable traits, we have increased our meta-modelling effort to make the two sub-models more comparable (see changes in Appendix C2), and the new evaluation results reflect this similarity at the level of basal area changes. We did not repeat the future projection with the advanced sub-model, for the computational reasons already mentioned, but performed it for a subset of plots, and we now compare the results with those of the basic sub-model (Fig. A4). The two sub-models do not perform equally in all variables predicted when using long-term simulations, as expected from differences in assumptions and parametrization. These differences will be the subject of study in future model developments.

Q1.3. Second, while I admire the effort and creativity employed by the authors to come up with parameter values (a total of ca. 25'000 parameter values had to be estimated!), I am concerned about the vast number of parameters that had to be estimated by imputation based on partly shaky assumptions (cf. below). And using a default value across all taxa is a somewhat desperate assumption to begin with. Such things induce a lot of uncertainty in the model on top of those mentioned above re. process formulations. So how much signal is actually gained by using a highly resolved approach when the projections are blurred by uncertainties in both process formulations and parameter estimates? It would appear to me (but this is not coming as a surprise) that going to this level of detail is not appropriate, and simpler models (such as the IPM...) should be used.

R1.3. Despite the enormous effort put into developing and parameterizing MEDFATE, we agree with the reviewer that simpler models such as IPM may be better if the aim is to predict forest dynamics alone (although it does not handle understory vegetation, which is very limiting in the Mediterranean), and we acknowledge it in the manuscript (L619-620). We routinely use IPM in contracts with the Spanish and Catalan administrations when the aim is to project forest dynamics. We sometimes take IPM predictions as input for water balance simulations done with MEDFATE, so that we can evaluate functional consequences of structural and compositional changes predicted by IPM. We have also compared IPM and MEDFATE regarding their projected dynamics under climate change and they do not differ substantially. This being said, having a model were both functioning and forest dynamics are found together opens new avenues. For example, the impact of extreme events is more naturally dealt by models of this kind, partly because of their finer temporal resolution (L619-620). We agree in that parametrization of complex models involves many assumptions, and we tried hard to build them in accordance with existing knowledge, but it is obvious that the work is not finished (we have a project to continue this, at the European level). In this sense, our manuscript illustrates well that parametrization is one of the most important research challenges in mechanistic forest models. MEDFATE predictions are not equally sensitive to all parameters, and default values are very often found in complex models. We agree with the reviewer that parameter and input (particularly soil) uncertainties can reduce the value of a highly resolved approach (L621-625), but collaboration with empirical scientists can help to decrease parametrization uncertainties, and better inputs may become available in the future, which make us keep our hope regarding the value of this kind of models.

Q1.4. Third, I think for the manuscript to be more convincing, the authors would need to do a better job re. model validation, particularly regarding the "intermediate" levels in the model, to show that it does *capture ecosystem structure and function and their dynamics reasonably well. It is only then that we could "trust" the regional-scale projections. In the current setup, the manuscript does not provide any model validation in the strict sense, which I think is unfortunate.*

R1.4. In the manuscript we tried to avoid the word "validation", because we are well aware that models are never validated in a strict sense. The evaluation using the SNFI3-4 is of course biased by the lack of temporal and spatial independence between the observations used for model parameterization and those used for model evaluation, but in our opinion this does not imply that the performance evaluation is useless, because there is a large part of model parameters that do not come from observations, and some are only indirectly related to the observed data. For example, we applied the relationship between RGRcambium and annual relative basal area increment at the species level (i.e. using growth averages across plots of the forest inventory), not at the plot level. For other processes, like mortality and ingrowth, the dependency of parameters on forest inventory data is much stronger and, therefore, the evaluation is less useful as a test to the model. We agree that additional evaluation was needed, and in the revision we include an evaluation of basal area changes predicted by the model also for the SNFI2-3 period (10 years) and the SNFI2-4 period (~ 25 years). Both sub-models have been tested (after re-running the meta-modelling and calibration exercises), and we keep the comparison with IPM.

Specific comments (referred to by line number)

Q1.5. 65: It would be relevant to expand a little bit on what "sufficient" actually is - in the present sentence, this is a claim whose scope is hard to assess. It appears that the starting point for the authors is that a detailed treatment of the energy, water and carbon balance is a pre-requisite for any vegetation model, and demography is "nice to have". One could view things exactly the other way around when focusing on the regional scale (cf. the statements on l. 50-51, and the application of the model to project wood volume).

R1.5. You are right. Our use of the word "sufficient" was meant to express a higher level of structural detail than other process-based models. We did not mean that demography was something only "nice to have". We changed the sentence, which now reads (L66-68) "Finally, hybrid models exist that combine a detailed mechanistic approach to energy, water and carbon balances with the ability to represent vegetation structure and simulate demographic processes". We also stressed in the next sentence (L69-70), that this comes at the cost of increased parametrization complexity and computational requirements, so that the reader does not take those models as a panacea.

Q1.6. 72-74: The nature of "trait-enabled" should be phrased in a more comprehensible manner for readers who are not familiar with the concept. I would write something like "...are 'trait-enabled', in the sense that their parameters can be linked quantiatively to easily measurable plant traits for which large-scale data bases exist." - and if you should disagree with such formulation, you would need to explain even better because I wouldn't have understood either…

R1.6. Thanks for pointing out this. We rephrased the definition. The text now reads (L73-74): "… are 'trait-enabled', in the sense that their parameters can be conceptually and quantitatively matched to measurable plant traits".

Q1.7. 87: I think it is not quite fair to state that "all these approaches ignore taxonomic information". I think all of them do consider them, although in widely different form. Your approach differs from these other approaches by being much more consistent in linking species parameters to plant traits at different levels of taxonomic resolution, rather than to impose one level of resolution across the model (as e.g. the 'classical' DGVMs with their PFTs are doing).

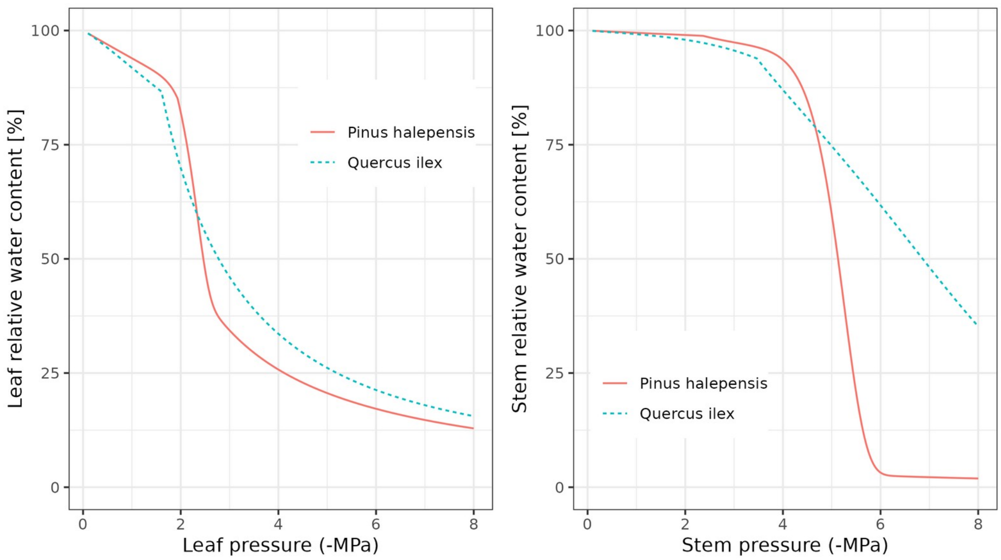
R1.7. We agree with your statement about DGVMs using PFTs, but some of the models reviewed in the paragraph focused so much on sampling parameters from the distribution of trait values across individuals that taxonomic identity was only implicit in the shape of this distribution or in specifying different distributions for broadly-defined types. Our aim here was not to diminish the value of such models, or any other, but to stress that species identity has useful information to be exploited (even if we are then missing intra-specific variation, see comments by reviewer #2). We modified the sentence (and the following one) to decrease the boldness of the statement (L89-90).

Q1.8. 90-91: "trait coordination" and later also "trait syndromes" or "functional syndromes": I presume your are referring to "trade-offs" between traits here? If so, pls use this more common term. And if you mean something else, pls explain better.

R1.8. Here we referred to the covariation between traits, which can be different when comparing species mean values as opposed to comparing intra-specific values. We replaced the word “coordination” with “covariation”, hoping that the statistical term is easier to understand (L90-92).

Q1.9. 279-290: The assumptions about the causes of tree mortality (starvation and desiccation) are a bit shaky. Mortality formulations are highly likely to be pivotal for model behavior, and hence they require careful thought and parameterization. I am not convinced that the assumption of "30% of maximum" for both processes is appropriate. Furthermore, I was surprised to read that turgor loss (to 30% of max symplastic water content) is a key cause of mortality in MEDFATE - to the best of my knowledge, wilting is easily reversible, but plant hydraulic failure is a syndrome that is more difficult to overcome (although I disagree with the authors' interpretation of the Choat et al. 2018 paper in that regard, but this is not the major point here). So why should I trust these formulations and their parameterization?

R1.9. Prescribing tree mortality from physiological thresholds is difficult (and debatable), because empirical evidence of tree mortality causes is often not univocal and empirically-determined thresholds usually come from seedlings or samplings in experimental settings. We justified our choices using recent reviews on the subject in which we have been directly involved (e.g., McDowell et al. 2022), although we acknowledge that future research may point to different solutions. Note that we are trying to simulate the point of irreversible dehydration following hydraulic failure, not simply turgor loss, would happen at relative water contents of ~80% in leaves (see also Martínez-Vilalta et al. 2019 in New Phytol. 223:22-32). In addition, our relative water content threshold refers to the stem compartment, not to the leaf compartment, which has a very different pressure-volume curve, meaning that 30% relative water content is reached substantially later in stems than leaves during drought progression. As an example, we draw below the leaf and stem pressure-volume curves we use for *Pinus halepensis* and *Quercus ilex*. As a result of these differences, lower values of relative water content will be observed in leaves than stems, for the same water potential, and leaf wilting will be predicted by the model much earlier than plant hydraulic failure, in accordance to the hydraulic fuse hypothesis (Hochberg et al. 2017). In the case of carbon starvation, we estimated that mortality thresholds should be determined in relative terms (i.e. as a function of maximum storage capacity) rather than using absolute concentration values, for which the model state variable may be very different from real-world measurements.



Q1.10. 315: The authors later (in the Discussion) acknowledge the problems with using the SoilGrids database. We have tried multiple times to use that database for estimating soil properties for simulation studies, and have always given up because of truly strange results, taking resort to other sources of data. Wouldn't there be a better data product for Catalonia?

R1.10. We also do not trust SoilGrids, and we acknowledge that uncertainties coming from global soil databases are a source of prediction errors (L621-623). However, there is no better digital product for Catalonia as far as we are aware. Soil maps are being produced slowly, for agricultural areas mostly.

Q1.11. 338-341: So the SNFI data were used for model "tuning". This means that what is shown in section 5.1 is not a "model evaluation" in the sense of a validation. These results have been forced, which strongly reduces their usefulness for demonstrating the skill of the model. I think alternative model validation data would be needed. I am aware that this may be everything but easy. However, this is a major weakness of the manuscript.

R1.11. Yes, we tuned (manually modified) one parameter (basal mortality rate), among 217, for some species, according to SNFI data. For the remaining species, the basal mortality rate was set to one third the observed mortality rate, so it is also related to forest inventory data. Any forest model (whether empirical or mechanistic) developed to project transient forest dynamics on forest inventory plots would have parameters being estimated from the data itself, which would mean that "true" (in the sense of using completely independent data) validation would be impossible. Manually tuning the basal mortality rates decreases the generality of the evaluation of basal area decreases due to mortality (and we never tried to hide it), but it improves the model's performance for the target region, providing an example regarding the generality-realism recently studied for Europe (Mahnken et al. 2022). We never intended the evaluation to be generalizable to other target regions (L619), hence prioritizing the realism over the generality of the model to the current exercise. If MEDFATE had to be run in another region (say France), new parameterization and evaluation exercises should be done using the French NFI, most likely ending with a different parameter set, specially for demographic parameters (note that we devote a whole sub-section of the discussion to transferrability of parameter values and parameterization procedures). Moreover, mortality is only one component of the performance evaluation with respect to overall stand-level basal area changes, for which the validity of growth assumptions and parameters are more important. Although we not recommend model tuning (as we stated in the original manuscript), we disagree that using tuning in one parameter becomes a major weakness of the manuscript, which focuses on the challenges of parameter estimation of complex models.

Q1.12. 362: I wouldn't use this example, as RGGS is not explained anywhere in the main paper. At the same time, however, I appreciated this very example because it shows how uncertain the parameterization procedures used here actually are: shade tolerance (sensu Niinemets & Valladares) is a highly aggregated concept that is pragmatically useful in some models, but probably not in highly detailed ones. Why should shade tolerance be related to these traits, and even in a linear manner? We have recently evaluated whether we can relate P88 (or P50) from the TRY database to the species' drought tolerance (according to Niinements & Valladares). The correlation is essentially non-existing, because the two concepts relate to vastly different levels of integration, and actually no relationship should be expected to begin with.

R1.12. Parameter RGGS was introduced when describing model's design (L249-253), so we think it is a valid example. Besides, there were not many other trait mappings involving transformations in MEDFATE's parameterization. We agree that Niinemets & Valladares is not an ideal source for parameter estimation. Regarding plant hydraulics, there are much better sources, that we employed in the development of MEDFATE. This particular parameter, RGGS, is important in the model as it affects the prioritization of storage and maintenance over growth, but it cannot be related to any directly-measured trait. The conceptual relationship between RGGS and shade tolerance (shade-tolerant species suppress growth to favor survival under shade) led us to turn our attention to Niinemets & Valladares. This parameter is by no means representative of the overall parameterization procedures, but rather an exception. Direct clear mappings to measured traits are provided for many parameters in Table B1. We modified the sentence in the manuscript (L365-368) to make it clear that RGGS is an exception, rather than the rule.

Q1.13. 460 (Tab. 3): Why are there different entires under "Mean observed" for the basic vs. advanced model version (and again different for the IPM)? This is not explained anywhere, and the 'naïve' reader would expect the observations to be independent of the model resolution?!

R1.13. It is true that differences in observed values are strange. Actually, this came from decisions in the scripts comparing simulation results with forest inventory changes, which lead to slightly different sets of plots being compared and, hence, different observed values. We completely rewrote those scripts and now they produce the same mean observed values for both model versions (see Table 3).

Q1.14. 578-580: This assertion is only valid if the processes are captured in the right way, which may not be the case (see general comments on structural uncertainties in the model). I suggest being more careful here.

R1.14. We modified the sentence to acknowledge this possibility (L594-596).

Q1.15. 599-600: Even if one follows this conclusion (which I do not really, for all the reasons mentioned above), one easily arrives at the question why a very complicated model like MEDFATE is actually used if the goal is indeed to "only" project regional standing timber volume. Simpler approaches would be equally suitable, would be inflicted with fewer uncertainties, and are even likely to provide more robust results. So why go with MEDFATE in this case? See also lines 635, where many more variables are mentioned that are of (potential) interest for forest management planning - but none of them are shown here, and the question arises how accurately they are projected by MEDFATE.

R1.15. We agree with the reviewer that other models may be equally or better suited if the goal was only to project regional standing volume (see our responses above), and we added a new sentence in the discussion to state this (L619-620). We advocate for MEDFATE because of the additional variables that the model predicts, although we acknowledge that these require additional evaluation exercises.

For example, we plan on using the PROFOUND database for a more complete testing of the model at the stand level. We added a sentence in the discussion to highlight the need to evaluate the model's performance with respect to additional variables (L655-656).

Technical comments
Q1.16. The manuscript is generally easy to read and understand, but it should still undergo a careful linguistic check before it is being published. I am pointing out examples of the kind of issues below, this is not a comprehensive list.
63: "impact these" -> "impact of these".
70: remove "into model parameters" (simply not needed). Then "while it is known that"-> "as".
83: "consists in" -> "consists of".
91: "within- and among-": remove dashes.
R1.15. Thank you for the linguistic checking. All these were corrected and we have carefully checked the whole text to identify and correct additional mistakes.

Q1.16. 110: Not clear how one can add processes to TWO preceding models. Below, two different additions ("basic" and "advanced") are explained that relate to ONE preceding model version. Pls clarify (consider using the terms "version" or "variant" when refering to these entities, maybe this would help).
R1.16. Thanks for pointing that the sentence was not clear. It now reads (L108): "We begin by describing the design and formulation of MEDFATE, which evolved from two preceding models (De Cáceres et al., 2015, 2021) that now constitute alternative sub-models of varying complexity for energy balance, water balance and photosynthesis processes". We hope it is clearer now.

Q1.17. 196: "conforms" -> "constitutes" ???
R1.17. Constitutes seems more appropriate. Thanks!

Q1.18. 222: replace "and" by a comma (note that this is another documentation of high structural uncertainty in the model, cf. comments further above).
R1.18. Corrected (also pointed by reviewer #2).

Q1.19. 271, 272: "feedback" -> "feed back" (the verb is used here, not the noun).272: add a comma before "and" (a full sentence = main clause is starting here).
R1.19. All corrected. Thanks.

Q1.20. 292: remove "but" (grammatically incorrect, not needed).
R1.20. You are right. This was grammatically incorrect. Removed.

Q1.21. 303: "includes a strong" -> "includes strong".
309: remove "different" (simply not needed).
310: number of SNFI plots in inventory #2 is not given here. Why?
369: "aim make" -> "aim to make".
395: "found" -> "find".
R1.21. All these were corrected. We now include the number of plots in SNFI2. Thanks!

Q1.22. 432-433: Why not between the second and the fourth inventory? This would have provided a longer time series and thus better insights into model behavior. The shorter the 'evaluation' period (when the model is initialized with the data from the first inventory, and this is what was done here, I suppose), then the more difficult it is for any model to be wrong - because the initial data is most likely

the strongest predictor of its skill, not the model-internal processes. Hence this 12-year (?) period is a really short snippet for a model test, on top of the non-independence of the data (cf. comments further above).
R1.22. Thanks for this suggestion. We agree that additional evaluation was needed, and in the revision we include an evaluation of basal area changes predicted by the model for the SNFI2-3 period (10 years), the SNFI3-4 period (~ 15 years) and the SNFI2-4 period (~ 25 years). The data used for evaluation is still not independent from the data used for estimation of some parameters, but at least the period covered is longer. Since we are assessing changes in basal area (and not BA values at the end of the simulation), we think that the dependency on initial conditions is not that relevant. In fact, we sometimes get better evaluation results for the longer (SNFI2-4) period than the shorter ones (see Table 3).

Q1.23. 507: "First": there is no "second" further below. Re-consider the structure of this paragraph.
R1.23. True. We changed "First" to "Clearly", which we think better fits the structure of the paragraph.

Q1.24. 554: "account" -> "amount".
560: "intra-specific can" -> "intra-specific variation can" (?).
575: "focus" -> "foci".
596: "prediction forest" -> "prediction of forest".
599: "confidence on" -> "confidence in".
614: "this equations" -> "these equations".
615: "dessication" -> "desiccation" (also elsewhere).
623: "perform under" -> "perform well under".
650: "think MEDFATE can be" -> "think that MEDFATE is".
R1.24. All these text issues were corrected. Many thanks for noting them.

Refs.

Hochberg, U., Windt, C. W., Ponomarenko, A., Zhang, Y. J., Gersony, J., Rockwell, F. E., & Holbrook, N. M. (2017). Stomatal closure, basal leaf embolism, and shedding protect the hydraulic integrity of grape stems. *Plant Physiology*, *174*(2), 764-775.

Martinez-Vilalta J, Anderegg, W.R.L., Sapes, G., Sala, A. (2019). Greater focus on water pools may improve our ability to understand and anticipate drought-induced mortality in plants. *New Phytol*. 223: 22-32.

McDowell, N. G., Sapes, G., Pivovaroff, A., Adams, H. D., Allen, C. D., Anderegg, W. R. L., Arend, M., Breshears, D. D., Brodribb, T., Choat, B., Cochard, H., De Cáceres, M., De Kauwe, M. G., Grossiord, C., Hammond, W. M., Hartmann, H., Hoch, G., Kahmen, A., Klein, T., Mackay, D. S., Mantova, M., Martínez-Vilalta, J., Medlyn, B. E., Mencuccini, M., Nardini, A., Oliveira, R. S., Sala, A., Tissue, D. T., Torres-Ruiz, J. M., Trowbridge, A. M., Trugman, A. T., Wiley, E., and Xu, C.: Mechanisms of woody-plant mortality under rising drought, CO2 and vapour pressure deficit, Nature Reviews Earth & Environment, 0123456789, 41–44, https://doi.org/10.1038/s43017-022-00272-1, 2022.

Mahnken, M., Cailleret, M., Collalti, A., Trotta, C., Biondo, C., D'Andrea, E., ... & Reyer, C. P. (2022). Accuracy, realism and general applicability of European forest models. *Global Change Biology*, *28*(23), 6921-6943.

Responses to reviewer #2 – Nikolaos Fyllas

Q2.1. The manuscript of De Cáceres and colleagues is an interesting study that describes the design of a forest dynamics model that is parameterised based on a diverse range of datasets from forest inventory to allometric and physiological datasets. In doing so the authors provide a systematic presentation of their methodology to derive the parameters needed to run the model at a regional scale. This study supports a model design that maintains the taxonomic resolution of species which is probably more relevant for forest management simulations in contrast to other approaches that use traits distributions to incorporate the range of functional diversity.

R.2.1. Thank you for the assessment and careful revision of our ms. We agree with the reviewer in that maintaining the taxonomic resolution is more relevant for some applications than for others. Maintaining species (or taxonomic) identity allows both the inputs (e.g. definition of management scenarios) and the results of the simulation at the regional scale can take advantage of this resolution.

Q2.2.The ms is well structured and the methods employed are adequately described, taken into account the wide range of work that the authors put in this study. I believe that this study is an important step towards a better understanding and simulation of Mediterranean forest dynamics and should be published. However, I believe that the authors need to make clear that their approach essentially uses trait average values for parameterising a the species under study, thus it does not enable plastic responses especially under climate change conditions.

R.2.2. Thanks for the positive opinion on our ms. We agree that our approach does not address the need to incorporate plastic (and non-plastic) intraspecific trait variation. We already mentioned this limitation (and several ways to overcome it) in the original manuscript (Discussion; L568-581) but we now acknowledge it more clearly already in the Abstract (L40-42) and Introduction (L99).

Q2.3. Also I think that the authors should also expand their simulation outputs under climate change conditions in a more species specific rationale, as for example aggregating all simulations in figures like Figure 3 does not help the reader to evaluate the ability of the model to provide species specific responses (which is also one of the advantages of the MEDFATE).

R2.3. We thank the reviewer for pointing out this shortcoming. We had included species-level results of the regional-level application in Appendix A, but it is true that these were not discussed at all. We moved the figure to the main text (now Fig. 4) and added a few sentences in the revised version (L500-504) to discuss the differences across species in the predicted impacts of extreme droughts (especially under RCP 8.5). We are open to producing new figures with species-specific results, if required, but we estimate the manuscript is at present already a bit long.

Q2.4. L54: year missing from first reference

R2.4. Corrected.

Q2.5. L63: the impact +”of” these

R2.5. Corrected.

Q2.6. L98-L101. Please rephrase, this sentence is not very clear

R2.6. Thanks for pointing this. We simplified the sentence and made it more clear that one takes the (maximum) taxonomic resolution from the forest inventory data (L99-101).

Q2.7. L108-109: You are not only illustrating the challenges that you encountered but also provide suggestions for solving them.

R2.7. Thanks! We added this in the objective description (L110)

Q2.8. L124. Please explain a bit better what you mean by density variable? Is it just the number of individuals per area, per species per area?
R2.8. Thanks for noting this. It is the density of individuals (per hectare) belonging to the plant cohort. We modified the text to make it more specific (L124-126).

Q2.9. L221 "although"
R2.9. Corrected. Thanks

Q2.10. L343 Does each tree species have a species-specific probability of ingrowth model based on the three bioclimatic variables?
R2.10. The ingrowth models are species-specific, based on the three bioclimatic limits plus a (species-specific) non-zero probability of ingrowth within the bioclimatic limits (probability of ingrowth is zero outside the bioclimatic limits). We added some text to make more clear that parameters are species-specific (L343-349).

Q2.11. L369 With the aim "to"
R2.11. Corrected. Thanks

Q2.12. L383 the large amount "of"
R2.12. Corrected. Thanks

Q2.13. L388 "of" leaves and fine roots
R2.13. Corrected. Thanks

Q2.14. 5.1 Evaluation with SNFI data. I do not understand here why you do not compare the predicted BA changes of the different models with the actual observed BA from the plots data. Also why the mean observed BA is different for the two sub-models? This is what you compare against no?. Also shouldn't the units be m2 ha-1 y-1 in Table 3?
R2.14. In each row of Table 3, we compare BA changes predicted by the different models with BA changes observed between surveys of the forest inventory. In the previous version of the manuscript, the units of the response variable were *m2 ha-1* because we were assessing the absolute change in BA across the simulation. Now we evaluate annual BA changes, and units are now *m2 ha-1 y-1*, which is more convenient because there are three simulation periods spanning different numbers of years. It is true that differences in observed values are strange. Actually, this came from decisions in the scripts comparing simulation results with forest inventory changes, which lead to slightly different sets of plots being compared and, hence, different observed values. We completely rewrote those scripts and now they produce the same mean observed values for both model versions (see Table 3).

Q2.15. L451 "…predictions we …" something is missing here (observed?)
R2.15. Corrected. Thanks

Q2.16. L482 Predicted "from/by" the model
R2.16. 'by' seems fine. Thanks

Q2.17. L493 "could/are predicted to" cause widespread drought-induced defoliation
R2.17. 'could' seems more appropriate here. Thanks

Q2.18. L494. I think it would be good to present in separate figures the "evolution" of ingrowth and mortality for each species. You should expect some species to be show higher drought vulnerability based on the parameterisation of your drought response.

R2.18. The "evolution" of mortality for each species was already shown in Fig. A4, but it is true that this could easily be missed when reading the manuscript. As explained in the response R2.3, we added a few sentences to describe species-specific results with respect to mortality patterns (L500-504). We have not inspected effects on ingrowth, but since the model still lacks resprouting (a very important regeneration mechanism in Mediterranean forests) we think it is better to keep the analysis of drought impacts on regeneration for future works.

Q2.19. L507-509. Strange sentence, please rephrase

R2.19. The sentence now reads "The wealth of information in global plant trait databases and forest inventory data facilitates dealing with functional diversity, but our experience shows that model parameterization using these resources has its own challenges". We hope it is clearer now.

Q2.20. L512-514. Agree but also some might expect that the demographic parameters should emerge as properties of a fully developed trait-enabled model. The forest inventory data could then be used to validate the recruitment and mortality submodels?

R2.20. The sentence was misplaced, because forest inventory data is observable, so that it should be mentioned as an alternative for non-observable parameters. We removed the sentence. Besides that, and answering the question, in our opinion if models were fully trait-enabled and mechanistic, then forest inventory data would be used for initialization and validation only, but at present the degree of knowledge in mortality and recruitment processes still requires using inventory data as a source of empirical parameters.

Q2.21. L560 Second, intra-specific… something is missing

R2.21. Yes. 'variation' was missing. Thanks.

Q2.22. L611. Please provide year of publication

R2.22. Done. Thanks

Q2.23. L638 "delete second "estimates"

R2.23. Corrected. Actually, we meant 'taxon-level estimates'.

Q2.24. L644 "an hybrid"?

R2.24. Corrected. Thanks

Q2.24. C.1 Line 697, are the extinction coefficients the same for each taxon? If not please specify.

R2.24. Extinction coefficients can potentially be taxon-specific, but when missing in the taxon parameter table they are derived from leaf shape. We now make it explicit in L702-704.

Q2.25. The approach the authors follow here to infer the parameters of the basic submodel (meta-modelling) is interesting, but it assumes that the advanced model simulations are correct, no? Do you have any data to validate this, or at least show that is behaves relatively realistically?

R2.25. Exactly. The meta-modelling assumes that parameters of the advanced submodel are correct. We evaluated the performance of the advanced model on experimental sites in De Cáceres et al. (2021), and this was already mentioned in the discussion (L539-542). However, we added more information about this in Appendix C.

Q2.26. Table C3. I would make bold only the statistically significant correlation in the upper right part of the table.
R2.26. Done. Only one correlation was statistically significant.