Authors Response to Reviewer's Comments

ECOSPHERE Manuscript ID (ECS23-0342) Tree regeneration in models of forest dynamics: a key priority for further research

Olalla Díaz-Yáñez*, Yannek Käber, Tim Anders, Friedrich Bohn, Kristin H. Braziunas, Josef Brůna, Rico Fischer, Samuel M. Fischer, Jessica Hetzer, Thomas Hickler, Christian Hochauer, Manfred J. Lexer, Heike Lischke, Paola Mairota, Ján Merganič, Katarina Merganičová, Tobias Mette, Marco Mina, Xavier Morin, Mats Nieberg, Werner Rammer, Christopher P.O. Reyer, Simon Scheiter, Daniel Scherrer, Harald Bugmann

2023-10-31

Table of contents

1 Response to the Editor

1.1 General Comments

Your revisions should address the specific points made in the review comments. Please note that Reviewer #2 made extensive comments directly on the manuscript.
I should note that personally I can hardly disagree with your basic conclusion that models of forest dynamics need to do a good job of capturing regeneration and recruitment processes. I have devoted most of my research over the past 40 years to developing the field methods to allow incorporation of accurate seedling, sapling and canopy recruitment in the models SORTIE and SORTIE-ND, in systems ranging from the tropics to boreal forests. The specific ecological processes incorporated in those models depended on the goals of the modeling and the ecology of the system, and included processes such as masting cycles, effects of disturbance on seedbed substrate dynamics, seed and seedling predation by small mammals, browsing by ungulates, Janzen-Connell effects, and, of course, detailed analyses of shading and neighborhood competition. More recently, this includes developing statistical models of seedling recruitment from the US forest inventory network that have been directly incorporated in the model for forests of the entire eastern US. To my mind, the difficulties the models you reviewed encountered in capturing canopy recruitment speak more to the limitations of the approaches used in model development than in any inherent difficulties in capturing regeneration and recruitment in models of forest dynamics.

Response to editors comments

Thank you for taking the time to review our manuscript and for your comments. Our main take-away from this manuscript is that while it would certainly be possible to calibrate all 15 models against the EuFoRIa data, it is important to ensure that this aligns with the study's goals. If the goal is to provide extrapolation into a future climate, calibrating against data under the current climate may not be reliable, as shown by the two models in our set that *were* calibrated against such data (i.e., SIBYLA and xComp). This is why we believe it is valuable to test the performance of the models "as they are", i.e. as we have done in this study.

We have carefully reviewed and addressed each of the specific points made in the review comments in this revised version, as explained below.

2 Response to reviewer 1

2.1 General Comments

The manuscript compares how regeneration and recruitment is simulated in 15 forest models, using a large network of forest reserves in Europe as the observed data. I want to start by saying there are many strengths of the paper. I was impressed with the comparison of 15

models, that the observed data was not used to parameterize or fine-tune the models, and the high quality of the writing. The introduction and discussion were easy and enjoyable to read, and the topic is critical and relevant. Some of the details about the models and methods were placed in the very substantial appendix, and I agree with their choices. However, the manuscript is still quite long with lots of multi-paneled figures. If the journal does not have a page limit, then this is not an issue. If there is a page limit, then some tough choices will have to be made for which figures to keep. Figures 2, 3, 6 and 7 are the most important (in my opinion).

I have a few major comments. One of which requires a different statistical analysis – which is why I do not consider this a 'minor revision'. However, there are no fatal flaws, and all of my comments are fixable. I hope they serve to improve the manuscript.

Our general comments to reviewer 1:

Thank you for taking the time to review our manuscript. We have made changes to the manuscript based on your feedback and believe that these changes have improved the overall quality of our manuscript.

We appreciate the concern about the length of the manuscript and the number of figures included. We have made an effort to reduce the text where possible and have verified that we are following the journal guidelines. Despite our best efforts, however, it has been difficult to significantly reduce the length of the manuscript.

Our responses to each of the comments and the concrete actions we have taken are detailed below.

2.2 Major comments

2.2.1 COMMENT 1

'regeneration' versus 'recruitment'. I personally have had this conversation multiple times over the years, so I understand that there is variation in the definitions and uses. However, the authors do define these terms in the methods (L167-171) and I completely agree with them. Regeneration is for flowers, seeds, and seedlings while recruitment is for reaching a certain size threshold. However, the entire manuscript refers to regeneration as "passing of trees across a specific diameter threshold ("ingrowth") (L142), either reaching a size of 7 or 10 cm DBH. This is recruitment. Since this is the response variable for the manuscript, I would suggest replacing 'regeneration' with 'recruitment' throughout the manuscript.

Response to comment 1:

Thank you for pointing this out. We discussed terminology several times and basically agree. We have followed your suggestion and have replaced the erroneous use of 'regeneration' by 'recruitment' where we refer to ingrowth into a certain DBH class. However, in some places

the term 'regeneration' is more accurate, especially in many parts of the introduction and discussion sections. 'Regeneration' is a broader term that starts from flowering and ends with ingrowth into a given size class; that is, recruitment is the result of successful regeneration. Thus, where we refer to the entire process of regeneration, it would not be appropriate to use 'recruitment', and therefore we maintain the former term where it is fitting.

2.2.2 COMMENT 2

The difference between 7 cm and 10 cm DBH in the observed data. From the methods, there is one dataset that has sites where recruitment was defined as 7 cm DBH, and then another where recruitment was somewhere between 7 – 10 cm DBH. It was not clear how these were treated and compared in the results. I assume that the models simulated all 200 sites, but then the recruitment results were compared like-to-like. So, if a site defined recruitment as DBH = 8.0 cm, then all 15 models would compare simulated recruitment at that site only at 8 cm. Since the data are boxplots, I am not able to tell. For example, Figure 2 – there should be 165 points for all of the 7 cm bars (observed and simulated) and 35 data points for the 10 cm bars (observed and simulated). If true, then please clarify. If not true, then please explain. Also, how was the data analysis performed when looking at the change in mortality and species diversity between 7 to 10 cm? I assume only the 165 stands were used. Please clarify.

Response to comment 2:

We realized that there is an issue with the way we communicated our diameter thresholds approach in the manuscript. The observed data had varying diameter thresholds for reporting tree recruitment, and, for the purposes of this study, we merged the recruited trees values into two thresholds of 7 and 10 cm. That is, in all plots ($n = 165$) that had a calipering limit <7 cm, only the successful recruitment events at 7 cm were considered; and analogously for plots ($n = 35$) with a callipering limit between 7 and 10 cm, where only recruitment at 10 cm was considered. Thus, the 7 cm dataset ($n = 165$) actually is a subset of the 10 cm dataset ($n = 200$), as all data from the 7-cm threshold could be re-calculated for the 10-cm threshold. Thus, there was no 'like-to-like' comparison for any other diameter threshold, only for 7 and 10 cm.

When examining the change in mortality between 7 and 10 cm, or the difference in species diversity between observed and simulated data (Figure 4), we could only use the 165 sites that contained both pieces of information at the 7 and 10 cm thresholds in the observed data. This is reported in the supplementary materials software, script "'Figures.R"' line 536 and L284-349 and now better explained in the manuscript: we have better explained the difference on the observed data in the section 3.2 Observed data L[225-230] and we have added a new text in the caption of figure 4 L[1085-1086].

2.2.3 COMMENT 3

The t-tests. The results in Table S2 and S3 (comparing at 7 cm and 10 cm DBH) should be at least by paired t-tests not two-sided t-tests (L319), where you are comparing the difference within the same stand at 7 and 10 cm. The results Table S5 and S6 should not be a t-test at all. Performing 15 individual t-tests is a classic Type-1 error situation (technically this applies to Tables S2 and S3 as well, a comment on that below). For Tables S5 and S6, I would suggest a general linear mixed effects model. You can take the difference between the observed, and the simulated output and make this the response variable (Y). Then, include model complexity, feedback, model type and scale as fixed explanatory variables. And include the stand ID as the random effect. This will more accurately answer your question about if any of these fixed explanatory variables can predict over or under representation of recruitment. For Tables S2 and S3, you could do the same thing (i.e., calculate the difference between the observed and the simulated output for each model) and make this your response variable. But the only fixed effect would be the actual model, and then you could do an ANOVA (again, stand ID as the random effect) and then a post-hoc test to see which models were or were not different from observed. If you do want to keep the 15 paired t-tests, then you need to account for Type I error rates and adjust your p-values accordingly.

Response to comment 3:

The objective of running the t-tests was to allow us to discuss differences between mean values per site between models, and observations and simulations between the 7 and 10 cm diameter thresholds. The t-test, as employed in our study, is not intended to serve as 'proof' of our findings, and we have carefully avoided an over-interpretation of their outcomes. Rather, they serve as a practical tool for handling the extensive number of comparisons in our study. To clarify this, we have included a statement in the manuscript, cf. L[336-337].

Following the reviewer's comment, we have included a correction for multiple tests, although we are aware that all corrections have pros and cons. We used the Bonferroni correction to make sure that the actual Type I error probability for the combined H0 is smaller than or equal to the nominal test level.

We have also looked into the alternative approaches suggested by the reviewer. We cannot do a paired t-test as for paired samples it is necessary to have the same number of subjects in both groups (observed and simulated), which is not the case here, as for example some models did not simulate all the sites. Yet, we tested the possibility of creating a GLME using as random effect the site and as explanatory variables all the features that we were interested in (model complexity, feedback, model type and scale), as suggested by the reviewer. However, this approach does not provide the comparison of observed vs. simulated data that we are interested in. Following the reviewer's comments, we conducted further tests to see if we could better explain differences between observed and simulated values using other approaches. We tested a mixed model to account for the randomness induced by to site (plot) differences.

For example, a GLME with recruitment values as the response variable and models as the explanatory variable, with observed data as the reference (with a Poisson distribution because we are dealing with count data). This worked well, and the patterns were the same as in the t-test (except for the model 4C). We also considered pursuing the same approach with the Shannon index and the differences in mortality, but in this case we think it is more relevant to understand the mean differences across sites, and therefore a t-test is more suitable. We furthermore tested a linear mixed model to assess differences between the 7 and 10 cm thresholds regarding the Shannon index, by creating a new parameter where the explanatory variable was a combination of the model and the diameter threshold. We then ran pairwise multiple comparisons to determine which means are different. Also here, the patterns here were quite similar to the ones resulting from the t-test except for two landscape-level models (TreeMig and Landis-II) that were significant in this test, contrary to the t-test.

After considering all the options that the reviewer suggested and testing others, we have made some adjustments to the t-test analysis, with the correction for multiple testing and cleaner output tables, as suggested by the reviewer. In addition, we have explained the purpose and interpretation of these tests better on lines [336-337].

2.3 Minor comments:

2.3.1 COMMENT 4

The abstract is not clear if the models do a good or poor job in simulating regeneration (L59, that they match observed ranges, and then L66, considerable mismatch between simulations and observed data, but that they do capture tree regeneration). Please clarify the main message in the abstract.

Response to comment 4:

We have changed the abstract based on feedback from both reviewers, including clarifying certain parts for better understanding. The models were able to capture some aspects of regeneration more effectively than others. One example, is the sentence highlighted by the reviewer that was referring to the diversity of regeneration.

2.3.2 COMMENT 5

L123-138: this paragraph has a lot of "on the other hand" type of statements, and it does not lead nicely into the next paragraph about the current study. It was a bit of a surprise (i.e., I did not expect the main purpose of the study when I started to read the following paragraph). It seems like the main issue is that we just don't know how uncertain regeneration and recruitment methods are, because we lack the comparison across models and a comprehensive empirical recruitment data set that we can use to compare. And that is what your study is doing.

Response to comment 5:

We have made several changes in this paragraph to improve its readability and flow of ideas towards explaining the main purpose of the study.

2.3.3 COMMENT 6

Table 1. Please clarify what the letters are for (a, b, c and d in the table headers). For models that start from saplings or ingrowth, please add in the size as well (i.e., start from saplings with DBH = 1.0). Could also add in the classification scheme mentioned in the methods (L168, regeneration or recruitment) to Table 1.

Response to comment 6:

We have added a new column classifying the models into recruitment or regeneration models; and the letters a, b, c, d, e and f in the table are now better described. Regarding the suggestion of adding the size of the saplings or ingrowth: we are concerned that this would be more confusing than helpful. For example, there are models that start from saplings without a diameter threshold, as they are biomass-based (e.g. LandClim), and providing all information for each model would make this table rather overloaded, without a clear benefit, as we believe. Therefore we have decided not to include this information. We would also like to note that the 15 model reports in the Supplementary Materials provide more details how regeneration is simulated.

2.3.4 COMMENT 7

L218 – do I understand this correctly, that regeneration rates are a mean of 56 trees per ha per decade? This seems really low (assuming that this is for >10 cm DBH).

Response to comment 7:

Yes, this is the mean value across all observations and sites. There are two major reasons why this number appears low. First due to the zero-inflated nature of the data, and second because these are forest reserves, typically being characterized by denser forests in which regeneration can be expected to be low. This value is in agreement with further observed data from reserves (Käber, Y., et al (2023). Sheltered or suppressed? Tree regeneration in unmanaged European forests. Journal of Ecology, https://doi.org/10.1111/1365-2745.14181).

2.3.5 COMMENT 8

L264 – 274: how will this be accounted for later, that not all models simulated the same set of species?

Response to comment 8:

For those models that simulated more than the eleven species, we created a different category "others" were we lumped those extra recruited species. These were counted when we calculated the diversity index as well as for total recruitment values. We have added this information in the simulation protocol section L[278-281].

2.3.6 COMMENT 9

L340: "regeneration basal area of that species" is mentioned twice, please delete the second one.

Response to comment 9:

Following the reviewers comment we have modified the text L[354-357].

2.3.7 COMMENT 10

L364: so please remove the 4C model from the figure as it is not comparable. Do you need to also remove other models that don't simulate the same number of species too?

Response to comment 10:

We acknowledge the point that the reviewer is making, but even though it is not directly comparable, we still believe to be insightful for the reader that we share the findings related to the 4C model. Note that in those models that simulate more species, these additional species were lumped in the category 'others' (as explained further above), thus with little impact on the comparison.

2.3.8 COMMENT 11

Table S2 and S3: What are the 3 columns for p-values? Please report just the one you used. No need to have a column to state 't-test' – just move this to the table caption. Then add in a column for df, the test statistic (t) and the effect size (was it higher or lower and by how much?). NOTE, this comment may become irrelevant if you change the statistical analyses.

Response to comment 11:

We have cleaned all the output tables.

2.3.9 COMMENT 12

Figure S1 and S2 – I can't tell the difference between observed and simulated. Are they different colours? The legend should be larger. What is the black line for?

Response to comment 12:

We have increased the font size of the legend, and now the difference between observed and simulated is visible. We have also explained the meaning of the black line in the caption.

2.3.10 COMMENT 13

Figure 4 (and Figure S3) are interesting in that there are no patterns with overstory diversity, and that this was not explained by feedback with seed production. What happens if you colour the points, by those from models with the feedback? Do you address in the discussion why this feedback makes no difference to recruitment?

Response to comment 13:

We agree that it is interesting that the models with feedback did not capture better the species diversity of the regeneration compared with those without, with the exeption of ForCEEPS. Figure 4 shows three representative patterns observed across models using 6 of the models as an example, and Figure S3 provides the details for each model. We believe it is not very informative to color the points from the models with feedback in Figure 4 as this will only color all the points from some of the panels and it is possible to identify the models with feedback in table 1 We have addressed this point in the discussion section L[563-570].

2.3.11 COMMENT 14

All figures in the Appendix, please make the X and Y axis labels larger and centered. I actually thought labels were missing initially and it took a while to find them.

Response to comment 14:

We have corrected these figures.

2.3.12 COMMENT 15

L400 – 402: where is this conclusion about mixed species or monocultures coming from? I don't see that in the figures or the results.

Response to comment 15:

This paragraph is addressing the comparison with the theoretical Reineke value which is usually calculated for even-aged, single species stands. We have change this text to clarify this aspect L[324-330].

2.3.13 COMMENT 16

Figure 6 is excellent.

Response to comment 16:

Thank you!

2.3.14 COMMENT 17

Table S4: please clean up. No need to report 7 places after the decimal point. Please clarify in the table caption what the slope represents (maybe mention Figure S5 too, as I assume it is illustrating the slopes). Another suggestion, would be to make the slopes solid lines if significant, or dashed lines if not significant. Not sure if that would help, but if it does then it could eliminate the need for this table.

Response to comment 17:

We have cleaned this table and changed the table caption including a reference to Figure S5 as suggested by the reviewer.

2.3.15 COMMENT 18

Not sure both Figure 7 and Figure S7 are necessary (I admit I spent a few minutes going back and forth between them trying to understand how they differed). I would just keep Figure 7 and note that the scales differ.

Response to comment 18:

We have removed figure S7.

2.3.16 COMMENT 19

L547-548: predicting stand diversity, did I miss this? I don't believe Figure 4 shows this, as if you put all data points on the same panel it would just be a random scatter of points that covers the entire space (not necessarily closer to the 1:1 line).

Response to comment 19:

Figure 4 shows three examples of the patterns observed across models (overestimation intermediate and underestimation) using one exemplary model of each category. The actual patterns per model can be found in Figure S3. In the results, we showed that most models capture reasonably well or overpredicted species diversity at the stand level, as most of the models are following the patterns presented in plot 1 and 2 (part A of Figure 4). We have decided to keep the referenced statement regarding this finding in the discussion.

2.3.17 COMMENT 20

L575: that most models did not deviate exceedingly from observations seems in contrast to what the previous section was stating. What is this based on?

Response to comment 20:

In this statement we were referring only to regeneration levels. Although we agree with the reviewer that we did state that most of the models overestimate regeneration, here our point is that even though they did overestimate regeneration, they did not deviate exceedingly, especially given that they have never been confronted with this type of data.

2.3.18 COMMENT 21

L584: What does this mean, about degrees of freedom for modelling regeneration?

Response to comment 21:

With this statement, we were referring to the multiple potential combinations of approaches of modeling tree regeneration. We have modified this for clarity L[596-598]

2.3.19 COMMENT 22

L624: Missing the end of the sentence.

Response to comment 22:

This was a glitch, which now is corrected L[636-638]

2.3.20 COMMENT 23

L647: Completely agree that observed data is a snapshot of a stochastic and dynamic system and care should always be taken with comparing to simulation data, however I disagree about the following stamen (L648). The models should be able to capture broader patterns along climatic or environmental gradients.

Response to comment 23:

We have removed these statements.

2.3.21 COMMENT 24

L655: But the landscape models generally did about the same (i.e., not consistently worse) then the rest, correct?

Response to comment 24:

Yes, the landscape level models were not performing significantly worse than the others. However, they were not performing better either, potentially because they were used in "single site" rather than "landscape" mode. Thus, the explicit consideration of horizontal space would possibly allow the landscape level models to better represent regeneration processes, as they are designed for that. This, however, was completely out of the scope of the present analysis, as there are no landscape-level data of the surroundings of the 200 plots.

2.3.22 COMMENT 25

L663: But in general, the models were more sensitive to water balance and the observed data did not have any real pattern. Why would this coarser representation of water balance in the model, cause it to be more sensitive?

Response to comment 25:

Our point here is that the models, although many of them have detailed schemes for the water balance, may not have performed well because of the absence of site-specific (truly reliable) soil information.

2.3.23 COMMENT 26

L675: what would a 'comprehensive regeneration dataset" look like? There are large data sets for seed production (MASTIF, MASTREE, etc.) and seedling data from national forest inventory programs. Does EuFoRIa (L677) have everything you need? Is it just a matter of expanding the spatial and temporal extent? Please explain what sort of data that the models would need (imagine an ideal world).

Response to comment 26:

We have changed this text L[684-688]. Although there are multiple data sources (such as MASTIF, MASTREE, National Forest Inventories, Forest Reserve plots, etc.), usually these data do not refer to the same locations and thus cannot be used in a site-specific manner for calibrating or testing models that intend to follow the "regeneration" chain of processes.

3 Response to reviewer 2

3.1 General Comments

It is valuable to demonstrate that leading forest dynamics models are unable to characterise regeneration processes very well. I also thought the "blind trial" approach was clever. Huge effort must have gone into running 15 models under various starting conditions and synthesising the outputs into a coherent paper. Choices have been made about which details to include in the main body of the text vs supplementary information and I thought a good balance was struck, although I have made several suggestions on the pdf.

Our general comments to reviewer 2:

We would like to thank Reviewer 2 for this positive overall assessment and the level of detail of the review. We have taken care to go through all suggestions made in the PDF, and have changed/improved the manuscript in most instances. Our responses to each of the comments and the concrete actions we have taken are detailed below.

3.2 Major comments

3.2.1 COMMENT 1

Where the paper falls short, in my opinion, is in providing a clear path forward. It shows that the models are too simplistic to simulate regeneration processes accurately but provides few insights into how we can improve their performance. The paper reports that creating more complex models is not valuable in and of itself, but that's not to say that intelligent improvement of specific subcomponents may hold the key. For instance, deer browsing and

competition with herb layer plants may well contribute to lower recruitment being observed than simulated, but how can we demonstrate that, and if the inadequate representation of these processes in the models is the problem, what should be done to improve the models? The review by Hanbury-Brown et al. (2022) on "future forests within Earth system models: regeneration processes critical to prediction" provides a good summary of current knowledge and future directions, including thoughts on 1. reproductive allocation and seed production; 2. dispersal; 3. seed survival, germination, and resprouting; and 4. seedling survival and growth. I would encourage the authors to frame their discussion in a similar way. Advanced statistical approaches that bring together simulation models and Approximate Bayesian Computation to estimate parameters of recruitment submodels may hold the key (e.g. https://onlinelibrary.wiley.com/doi/full/10.1111/ecog.04824). To me, a key question is how we can make better use of field measurements and statistical analyses to refine forest dynamics models, i.e. not only parameterise them but also to identify which submodels need refinement and the function that need inclusion.

Response to comment 1:

In this first assessment we were not aiming at identify what specific subcomponents may hold the key and therefore based on the current approach we could not provide more concrete recommendation for model development which most likely be specific for each of the 15 models. Maybe we need more such studies, including calibration as a first step, and then application against a withheld part of the data; or application (and comparison) under climate change scenarios to evaluate the robustness of the signal under future conditions.

3.2.2 COMMENT 2

I felt the readability of the paper would be improved by introducing a "road mapping" paragraph at the end of the introduction or start of the methods section that sets out the approach in broad terms.

Response to comment 2:

We have added a "road map" at the end of the introduction L[159-163].

3.2.3 COMMENT 3

The words "ingrowth", "regeneration" and "recruitment" are used interchangeably to describe the recruitment of new trees into the 7-cm diameter class. I would encourage the authors to use the term "ingrowth" or "recruitment" to this to describe this process, and keep "regeneration" as an overarching term describing the entire process of seed production, dispersal, early establishment, and onward growth.

Response to comment 3:

Thank you for pointing this out. We discussed terminology several times and basically agree. We have followed your suggestion and have now use recruitment to discuss the process and regeneration as the overarching term.

3.2.4 COMMENT 4

The article is written entirely by Europeans, unless I'm mistaken, and very much focused on the central/eastern European forestry literature. I would encourage a deeper dive into the relevant North American and Asian literature when introducing and discussing their work.

Response to comment 4:

Indeed, most of the people involved in this manuscript have a central European focus in their current work, and so did the study setup. We have included global references or references with a focus outside Europe when they were relevant. We have evaluated whether this could be further "internationalized", but felt our choices were appropriate.

3.3 Minor comments:

These comments have been extracted from the reviewer 2 comments on the PDF document of the manuscript.

3.3.1 COMMENT 5

L51-52 "However, an assessment of their ability to accurately represent tree regeneration is lacking." Not true.

Response to comment 5:

We have modified this sentence L[52-53]

3.3.2 COMMENT 6

L56 "The results are evaluated against comprehensive data from unmanaged forests." comprehensive?

Response to comment 6:

We have changed the word "comprehensive" to "extensive" to express that we used a wide-range data that considers many important elements of regeneration L[57]

3.3.3 COMMENT 7

L56 “Models simulating higher species diversity at the stand level do not feature higher regeneration diversity.” feature?

Response to comment 7:

We have kept “feature” in this case.

3.3.4 COMMENT 8

L76-78 “A wide range of models of forest dynamics were developed over the past decades considering the impacts of climate”

Response to comment 8:

We have changed the word “were” with “have been” L[78].

3.3.5 COMMENT 9

L79 K. Vanclay and Skovsgaard 1997;

Response to comment 9:

We have corrected the reference style L[80].

3.3.6 COMMENT 10

L83-84 “Which is a clear research gap in the context of climate-induced forest disturbances and forest resilience.” incomplete sentence

Response to comment 10:

We have changed this sentence L[84-85].

3.3.7 COMMENT 11

L85-86 “Tree regeneration arises from multiple processes such as pollination, fruit maturation, seed production, dispersal, germination, juvenile growth and survival” including

Response to comment 11:

We have changed “such as” with “including” L[86].

3.3.8 COMMENT 12

L90-92 “Currently, tree regeneration processes in dynamic forest models are handled in a multitude of ways (König et al. 2022; Bugmann and Seidl 2022): from 1) entirely ignoring it (as done in classical forest growth models, e.g., Pretzsch et al. 2002),” it.

Response to comment 12:

We have changed “it” with “them” L[92].

3.3.9 COMMENT 13

L96-98 Particularly given your conclusion that the models are poor at predicting regeneration processes, should we now return to statistical parameterisation afresh and find better ways of extrapolation/interpolation across space/time?

Response to comment 13:

It is indeed interesting to account with the increasing data availability to better understand tree regeneration processes. However, we believe that the main take away it is not that it would be impossible to calibrate all 15 models against e.g. the EuFoRIa data to better capture regeneration, but that the focus should be in the study goals. And when the goal is to provide extrapolation into a future climate, then calibration against data under current climate may be unreliable. With the approach we have followed in this study where we have asked the models to simulate forest dynamics and capture regeneration without statistical parametrisation we can better assess their behavior and understand how their current model structures and inclusion of exclusion of certain ecological processes impact the outputs we are seeing.

3.3.10 COMMENT 14

L100-101 “Overall, models are needed to…(3) identify the most important processes that are shaping ecological patterns.” doesn’t that depend on what goes into the model?

Response to comment 14:

We are not completely sure of the reviewers point here. If the reviewer is referring to the fact that the most important processes will be those included in the model, it is true that you can only asses those processes included or how including or not another processes impact regeneration. This was also the idea with considering a large pool of models, that we could assess how their structural complexity, or the inclusion/exclusion of a certain process affects the results we are seeing. This also goes very much in line with our research recommendations presented in the discussion where we proposed that improvement of the regeneration modules is implemented as additional features that can be traced back, as done here for the variants of

ForClim and ForCEEPS, and that model complexity and structure must always be connected with modeling objectives.

3.3.11 COMMENT 15

L101-106 "Given the current strategies that are used in models of forest dynamics to represent tree regeneration, model behavior often is prone to problems, such as very high levels of tree regeneration that necessitate excess mortality at early stages of tree life to simulate correct stand structure and composition." ref? tens of thousands of seeds...

Response to comment 15:

We have added a reference here.

3.3.12 COMMENT 16

L112-114 "A related issue is the excessive reduction of species diversity due to positive feedback effects, such that eventually just single-species stands remain." ref

Response to comment 16:

We have added a reference here L[105].

3.3.13 COMMENT 17

L127 I suggest making better use of the hanbury-brown 2022 review.

Response to comment 17:

We currently use the exciting Hanbury-brown 2022 review in several places across our manuscripts, following the reviewer's comment here and the general comment 2 we have also get inspired by their structure by creating a road map to help the reader to navigate the manuscript content.

3.3.14 COMMENT 18

L145-147 "Due to the large variability in tree regeneration patterns in nature and the large number of factors driving this process including some that are not incorporated explicitly in most models, such as deer browsing—"

Response to comment 18:

We believe the reviewer highlighted the words “models, such as deer browsing” because they think this is not correct. From our sample pool only 4 models incorporated explicitly deer browsing.

3.3.15 COMMENT 19

L185 “(mean regeneration formulation complexity across all processes >0, Table 1).” Meaning

Response to comment 19:

This statement relates to Bugmann and Seidl (2022) where complexity values were provided. The exact details of how we calculated this mean regeneration formulation complexity value are available in the software supplementary materials to this paper: “‘figure.R“‘ line 706-750. This value represents the mean complexity of regeneration formulations per model.

3.3.16 COMMENT 20

L186-188 “ForClim variant 1 (Bugmann et al. 1996) is based on a recruitment module that adheres closely to the concept introduced by Botkin, Janak and Wallis (1972)” which is…

Response to comment 22:

We have added information to briefly explain the general approach L[196-197].

3.3.17 COMMENT 21

L197-198 “Regeneration data covering a wide range of environmental conditions are hard to obtain, and this is one of the reasons why most models of forest dynamics have never been confronted with a dataset” not sure that’s true. Some of the large forest inventories include small seedlings and saplings in nested plots, which are a step in the right direction.

Response to comment 21:

We agree with the reviewer that permanent plots such as those present in National Forest Inventories could potentially be a good source of information to obtain regeneration data. However these often are collected from managed sites, do not capture ingrowth below 10cm, and are hard to harmonize between different inventory strategies.

3.3.18 COMMENT 22

L200-202 “The observations used here are derived from a novel and unprecedented network of sites in forest reserves that represent the range of environmental gradients in temperature and precipitation in Central Europe as compiled in the framework of the EuFoRIa network (EuFoRIa 2019)” I don’t feel “novel and unprecedented” adds anything to this sentence. There are plenty of networks outside Europe providing this sort of data

Response to comment 22:

We agree with the reviewer’s comment that there are other forest inventory data networks outside Central Europe even on forest reserves that could provide this sort of data. However, within Europe no such data existed before. The research network (EuFoRIa) was established in 2019 and therefor is in fact novel, and within Europe such data is also without precedence.

Yet we acknowledge that our statement needs to be put in a European context much clearer. We hope that our adjustments to the text resolve the exaggeration of our initial formulation L[211-213].

3.3.19 COMMENT 23

L208-209 “We selected 200 sites from this network as the benchmarking dataset for the simulation to be representative of the environmental variation contained in the data.” COOL

Response to comment 23:

We agree :)

3.3.20 COMMENT 24

L211-212 “exposition (i.e.,slope and aspect).” unfamiliar term to me

Response to comment 24:

We have changed “and exposition (i.e.,slope and aspect).” with “, slope and aspect.” L[223].

3.3.21 COMMENT 25

L212-213 “Regeneration thresholds for these sites differed between diameters of 0 and 10 cm” minimum size threshold? and stem diameters

Response to comment 25:

Following the reviewer’s comment we have changed this sentence L[224-225].

3.3.22 COMMENT 26

L215 “another 35 sites with diameter thresholds between 7 and 10 cm.” OK, so we’re talking about v large trees here

L218 “featured 30,900 newly established trees” so they are trees….

Response to comment 26:

The minimum threshold that could be standarized across reserves was of at least 7cm.

3.3.23 COMMENT 27

L215 “on this unique dataset” not so unique

Response to comment 27:

We have changed this text L[235].

3.3.24 COMMENT 28

L221 “nformation, cf. Käber et al. (2023).” in review and not available

Response to comment 28:

This paper have recently being published and the reference has been updated accordingly.

3.3.25 COMMENT 29

L223 “The overarching goal of the experiments was to assess the tree regeneration as it arises” simulation experiments and remove “the”

Response to comment 29:

Changed as suggested L[238-239].

3.3.26 COMMENT 30

L225 “tree regeneration as the passing of a breast height diameter threshold of 7 or 10 cm, respectively” I suggest “tree recruitment” not “tree regeneration” here. respectively??

Response to comment 30:

We have removed “respectively” L[240].

We have re-assessed the terms used across this manuscript and change it across the text accordingly.

3.3.27 COMMENT 31

L 227-228 “simulations, providing input variables” provided with

Response to comment 31:

We have removed “providing” L[242].

3.3.28 COMMENT 32

L229 “Neither were further site information (except for the data specified below) nor any data” These were “blind trials”: modelling group were not provided with site information ...

Response to comment 32:

The reviewer is correct that we did not provide e.g. site coordinates (with the exception of iLand that needed this for a model specific soil input, this is described in the manuscript), but we did provide topography, climate and soil site specific variables.

3.3.29 COMMENT 33

L230-231 “That is, the models were run in”blind flight” mode.” Remove

Response to comment 33:

Changed as suggested.

3.3.30 COMMENT 34

L237-238 “Soil quality data were provided as continuous values between 1 and 5” ?hydrological quality?

Response to comment 34:

The soil quality value is based on a random forest that was trained to predict expert knowledge based soil quality on a scale from 1 to 5 from the SoilGrids250 data set and the WISE data set. The most influential variables in the random forest were total depth to bedrock and sand content. Thus the variable is a good predictor for plant available water storage capacity. This information has been provided in the protocol of this study.

3.3.31 COMMENT 35

L247 “The exact length of the simulation was also decided by the modeling teams” seems odd not to constrain that.

Response to comment 35:

We understand the reviewers concern on not constraining the length of the simulations, but we decided to leave it to the assessment of the individual modelers, as we wanted to obtain simulation that were run from bare ground in the absence of management to a simulated equilibrium (“potential natural vegetation”) with current climate. Each modeler know best how to obtain this in each model and different model uses its own logic to get to the equilibriium that can require different lengths.

3.3.32 COMMENT 36

L250 “The simulations were run in the absence of management to a simulated equilibrium” The absence of disturbance and running to equilibrium set quite specific conditions; suggest this is emphasised more in abstract / intro / discussion.

Response to comment 36:

We agree with the reviewer that the simulations set up regarding absence of management and to a simulated equilibrium are quite specific. We found challenging to incorporate this in the introduction as we see this a more of a methodological specification, and in the abstract due to space constraints as we though is more interesting to focus in the results and meaning of our findings than in methodological aspects. However we currently have a lengthy discussion on the equilibrium assumption and also about the importance of considering disturbances, which we think it would be a great further research direction to better understand regeneration from models of forest dynamics.

3.3.33 COMMENT 37

L252-253 “This entails the assumption that (1) the observations from the forest reserves reflect no traces of forest management,” That’s seldom the case in Europe where the legacies of management are evident centuries on

Response to comment 37:

The reviewer is correct and management legacies might be evident for long periods of time. The data were collected in forest reserves where no management has taken place for long periods of time which makes our assumption of an equilibrium between forest properties and environmental drivers reasonable. This is the best data we have and we have addressed its limitations, as this one, in the discussion because we are also aware that is not ideal.

3.3.34 COMMENT 38

L275 “Each of the models reported the regeneration number by sampling 200 times in a 10-year interval” regeneration number = number of “recruits” entering the 7-cm size class?

Response to comment 38:

Following the reviewer’s comment we have changed this sentence L[290-291]

3.3.35 COMMENT 39

L295 “(2) regeneration species diversity” species diversity of recruitment

L295 “(3) regeneration mortality” mortality of recruits

Response to comment 39:

We have changed the terminology.

3.3.36 COMMENT 40

L296 “ingrowth gradients along the regeneration niches.” ?? not clear to me. How is regeneration niche defined??

Response to comment 40:

We defined the regeneration niche as the passing of a diameter threshold of 7 and 10 cm and we focused both in the width of the regeneration niche (i.e., in environmental space) as well as the intensity of the regeneration process (i.e., the number of ingrowth trees per area and per unit of time) were of interest. This is currently defined in the material and methods section and in the protocol of this study.

3.3.37 COMMENT 41

Equations 1-5. not convinced these formulae need to be shown as the Shannon index is very well known. Consider converting the Shannon index to the effective number of species, by taking the negative of the SI and exponentiating it. This is gaining ground in ecology as it's more intuitive to interpret than SI (see paper by Jost).

L1046 Such converting to effective no of species using Jost formula

Response to comment 41:

We agree with the reviewer that the Shannon index is very well known index and its equation as well. However we think is still important that we define how we did calculate the Shannon index based on the available data we had as we had some internal discussion on how to do this and the meaning of this index in this context, we thought the reader will benefit from a more detailed explanation. In any case the exact details of how we did calculate this index based on the simulation outputs are also available in the software supplementary materials, in the script figures.R L42-78. We also thank the reviewer for the suggestion of using the effective number of species instead of the Shannon index, it is indeed a good alternative, however we have decided to leave it with the SI.

3.3.38 COMMENT 42

L308 "Mortality in tree regeneration was assessed based on the ratio of regeneration between the 7 and 10" of recent recruits. Now I understand the 7 and 10 cm thresholds mentioned earlier; please clarify earlier mention.

Response to comment 42:

Following the reviewer's comment on the lack of clarity of the selected thresholds explanation we have changed at the beginning of the section simulation protocol the first sentence where we explain that we considered tree regeneration as the passing of a breast height diameter threshold of 7 and 10 cm. We have also reported later on in this section (but before this line) that each of the models reported the number of trees crossing the two diameter thresholds L[290-291].

3.3.39 COMMENT 43

L311-313 "The Reineke self-thinning rule is usually calculated for even-aged, single species stands and assumes a fixed relationship between the number of stems and the quadratic mean diameter in fully stocked pure stands." more to the point, it's always (I think) applied to all stems in the stand, not just regenerating trees, as it's fundamental concept is that entire canopies are space filling . Why would we expect it to apply to a narrow cohort? I think not.

equation 5 Suggest writing N7 / N10

Given my comment above, is there a strong justification for the -1.605 exponent? Or just keep it at one?

Response to comment 43:

Following the reviewer recommendation we have changed the equation formulation to better represent the Reineke relationship. The reviewer is correct that the Reineke is applied to all stems in the stand, in our case the N value is only changing due to the available regeneration at the two diameter thresholds as the adult trees number is the same. By comparing our expected relationship of recruited trees at 7 and 10cm with the Reineke value we are just comparing to a theoretical value in fully stocked stands. We have use the coefficient -1.605 because Reineke attributed a general validity to this allometric coefficient for fully stocked, even-aged forest stands, regardless of tree species and site.

3.3.40 COMMENT 44

L317 "model type (empirical or process based)," if statistical analyses are used to estimate the coefficients of complex functions that underpin a process based model, is that model empirical or process based?

Response to comment 44:

The reviewer's comment is a very valid one as it is true that many processes in process based models have been defined based on empirical approaches. In this study we have categorized the models in process based and empirical based on their general approaches to define forest dynamics. This categorization is available in table 1. It is also possible to find further details on each of the models in the supplementary materials.

3.3.41 COMMENT 45

L319-321 - overly complex sentence structure

Response to comment 45:

We have improved this sentence L[334-337].

3.3.42 COMMENT 46

L323-324 presumably, recruits make up a very small proportion of stand basal if the models have been run to equilibrium?

Response to comment 46:

Yes the reviewer is correct, the recruits represent a small proportion of the total stand basal area.

3.3.43 COMMENT 47

L327 of each species?

Response to comment 47:

Added as suggested L[343].

3.3.44 COMMENT 48

L332-335 This give a matric of species-level variation in recruitment across sites, but how does this relate to the regeneration niche concept of Grubb 1977?

Response to comment 48:

We have changed the text to clarify this point. L[350-351].

3.3.45 COMMENT 49

L496 “confronted with a unique dataset from unmanaged”

L498 “The EuFoRIa data (Käber et al. 2023) are exceptional, particularly”

I think there are plenty of other datasets that could be used and have better information on smaller trees than EuFoRIa, so suggest a more nuanced sales pitch here.

Response to comment 49:

We agree with the reviewer’s comment that there are other forest inventory data networks outside Central Europe even on forest reserves that could provide this sort of data. However, within Europe no such data existed before, e.g. from unmanaged forests and collecting such a large number of records.

3.3.46 COMMENT 50

L525-527 “There are multiple constraints to the regeneration niche of tree species (Price et al. 2001), and therefore the absence of regeneration is likely to be common (Fortin and DeBlois 2007), even over larger areas” YES - good point. This should be featured more prominent I think

Response to comment 50:

Thank you for this remark, we agree with the reviewer that this is an important point and specially how different models are capturing or missing to represent the absence of regeneration. We have now captured this aspect also in the introduction in the sentence, L[147-148], on top of the current whole paragraph on this issue present in the discussion.

3.3.47 COMMENT 51

L530-533 “This substantial difference may be due to the fact that the simulation results were drawn from equilibrium forests, whereas in reality many of the forest reserves are recovering from past management activities and have become denser over the past decades (e.g., Heiri et al. 2009), leading to less regeneration than in an equilibrium situation.” But also --- browsing, establishment sites, competition with herb layer all reduce recruitment

Response to comment 51:

We have included this point by adding into this paragraph a new sentence: L[544-546].

3.3.48 COMMENT 52

L552 “there is no evidence that models with feedback from the canopy captured better the species” could you explain what you mean by “feedback” here?

Response to comment 52:

We have modified this sentence to clarify what do we mean here with feedback L[563-567]

This had also been defined in the material and methods where we stated L[180-182].

3.3.49 COMMENT 53

L586-587 “Our study showed that increasing complexity in the regeneration modules is not linked with a higher accuracy of the projections of regeneration levels,” Surely more complex models WOULD be better if they captured the relevant processes! Arguing for simple vs complex models is a distraction here... we need models that are as simple as possible but not too simple..

Response to comment 53:

We agree with the reviewers statement that we need the “right amount” of complexity which considers the most relevant processes. However there are limited studies looking at this aspect in models of forest dynamics and even more limited looking at this aspect in the regeneration related processes in models of forest dynamics. We still think it is important to discuss this issue in the context of our findings even though this was not the solely focus of our research.

3.3.50 COMMENT 54

L596-598 “Competition for light as a strong filter for tree regeneration has been widely documented (Messier et al. 1999; Collet and Chenost 2006; Berdanier and Clark 2016), but the models examined here did not reproduce this expectation.”, because ... remind us of the evidence you are putting forward

Response to comment 54:

Following the reviewers comment we have added the a sentence in this part of the text L[611-612].

3.3.51 COMMENT 55

L603-606 “This made it impossible to evaluate the regeneration for the extremes of the stand density ranges in some models. For example, regeneration levels at low stand densities are relevant to assess how well forests are recovering e.g. after gap creation due to disturbance” Agreed ... and this paper isn’t testing performance of models in predicting regeneration in large gaps. It’s possible that the models do a better job at this, but it’s not something you tested. Suggest making that point e.g. in the abstract.

Response to comment 55:

We have decided not to include this aspect in the abstract due to lack of space.

3.3.52 COMMENT 56

L656 “Yet, the global models should not be at a disadvantage due to the limited spatial consideration” Don’t quite follow your point here. what does “at a disadvantage” mean?

Response to comment 56:

In this sentence we are discussing how the sampling strategy presented in the study protocol might have represented a disadvantage for landscape level models due to the limited spatial scale, but how this might not be the same for global level models as they usually lack dispersal between cells.

3.3.53 COMMENT 57

L676-677 “Therefore, we further recommend that more effort should be invested into collecting harmonized datasets on tree regeneration.” regeneration in what sense? Seedlings/ ground layer/ …. What is really needed?

Response to comment 57:

We have changed this text L[687-688].

3.3.54 COMMENT 58

L685-687 “However, this will require an entirely different set of observed data, and potentially not all models of forest dynamics would be able to assess the relationship of these aspects on tree regeneration, e.g. due to the lack of disturbance or appropriate management modules.” more akin to data already available in national inventories I think

Response to comment 58:

We agree with the reviewer that a more systematic kind of inventory could be useful to detect regeneration patterns under management or disturbances impacts.

3.3.55 COMMENT 59

L696 “Exercises like the one presented here, where the models are operated in”blind flight” mode.” Definition of blind flight “To do something based on guesswork, intuition, or without any help or instructions”….is that what you mean?

Response to comment 59:

With blind flight we refer to doing the simulations without having important information about the expected regeneration values.

3.3.56 COMMENT 60

Table 1 caption: A useful table. I'd encourage the authors to write a lengthier table heading, which briefly summarises the methods section, so that readers can know what "start from", "runtime for sampling" mean without having to delve into the text. Also I'd like to see an explanation of "formulation complexity" in the text and here. The supplementary information gives more details about model assumptions, and I do wonder whether those could be included here; in particular, seed production and dispersal is clearly important for regeneration, so could columns be introduced to compare assumptions of model with respect to these. Also deer browsing, herb-layers and nurse logs all have an influence of regeneration, but not incorporated into these models; I suggest the table heading makes that clear.

Response to comment 60:

Following the reviewers comment we have expanded the caption of table 1. We have decided not to include further variables on the model assumptions as we believe the current information provides a good general overview of the models approach and further details are included in the appendix.

3.3.57 COMMENT 61

L831 Add publisher

L939 Correct doi

Response to comment 61:

Changed as suggested.

3.3.58 COMMENT 63

L1041 suggest sticking with "ingrowth" (as used in the figure) or "recruitment"

Response to comment 63:

We have changed the terminology.

3.3.59 COMMENT 64

L1050 Overperdicted

Response to comment 64:

Changed as suggested.