Reply rev 1

Dear reviewer,

We would like to thank you for the constructive comments on our manuscript. We have revised the manuscript to accommodate all your concerns. Below, we have replied to all your comments in detail providing concrete reference to the sections that have been changed. We have also marked the main revised parts of the manuscript in yellow.

We hope that this new version accommodates all your concerns.

Comments

(1) The language needs a thorough revision. There are numerous grammar errors, typos and spelling errors. Additionally, there are obvious flaws in several sentences which are either incomplete or include remnants of earlier versions.

We have thoroughly revised the paper. If we revised the paper according to the comments of the second reviewer we will check the language again.

(2) While the general structure of the model description part is adequate, its implementation is not consistent. The model structure section includes also processes. The water processes in the process section are weakly described.

We have cleaned the structure and renamed the chapter into "model overview". The water processes are described shortly in 2.1.2.5 as part of a quick model overview and in the detailed description (chapter 6).

(3) The model structure section needs a thorough revision in its own. The basic and essential structural features of the model needs to be described clearly and unambigously. For instance, the role of the tree cohorts is never clearly described. There are some statements about the inter-cohort competition. But nothing is said about intra-cohort relationships of trees. Area size of the forest that can be simulated is also not clear.
We have reformulated the section and added information about intra-cohort relationships. We now write:
"It can be applied for patches of various sizes (varying from 100 m² to several hectares) and mono- and mixed-species forests."
and
"All trees within a cohort share the same characteristics which are species, age, tree dimensions (height, height of crown base (or bole height), and diameter at breast height), biomass differentiated into various compartments (foliage, fine roots, sapwood, and heartwood) and stage of phenological development. This allows simulating a representative tree of each cohort instead of each tree of the stand. The model is distance independent and so the trees within a cohort are horizontally evenly distributed and their position unknown. There are no differences in the growth behaviour of the trees of a cohort and there is no competition between the trees within a cohort. In contrast, the tree cohorts compete for light, water and nutrients."

(4) In a manuscript like this one I expected an overview on past model applications and related literature. In Table 2 past evaluation experiments are listed, but no general overview about 4C-related literature is provided.

We added in chapter 2.2. (Previous model evaluations and applications) a short description of past applications and in the ESM a detailed table with an overview of past 4C applications.

(5) Main part of the manuscript is a new evaluation study that includes four sites across Europe. What has then been gained from these new experiments compared to earlier evaluation studies is never clearly presented.

We inserted a reason for the use of PROUFOUND data for our evaluation in chapter 2.3:

"Within PROFOUND, several other forest models are using the same data so that comparisons between those models are possible."
Therefore, we decided to use these data for this evaluation paper.

(6) The results section is quite long and includes too many graphs and tables. Some of the material could be moved to annex or supplementary material. In general there is the tendency, that without much context the results for individual stand and tree attributes are listed in no obvious order. Summarizing, it is very demanding to get the essential information from the vast amount of numbers.
Thank you for this advice. We shifted some figures to the supplement (the time series in figure 3-5, Figure 7), and some additional for the final revision. Former evaluations based on similar data /Reyer etal 2014) were not extensive or consider only fluxes. We think that the evaluation based on all available data allows identifying the deficits or benefits of the model.

(7) In Figures 3 and 4 the initial values for observed biomass and simulated biomass show a rather huge mismatch. Please explain why. Labels of y-axes are also not consistent. Please check.

The missing match for the initial values is explained by the initialization process. The model used averaged data of DBH, H and stem numbers to generate a stand with these characteristics. However, the initialization does not exactly match the DBH, H and biomass. In chapter 4.1 we tried to discuss this problem. Only if single tree data are available a better correspondence between data and model in the initialization is possible.

We used different scales for biomass and diameter for the two stands Peitz and Solling because of the very clear differences in the amounts of these values between the stand caused mainly by the species (pine versus spruce) and site conditions. Peitz is a very poor site with dry condition whereas Solling is a wet site with better soil conditions.

(8) Tables 7, 8 and 9 are not well structured. How sites are depicted is not consistent in formats.

We depicted the sites with available flux data (Hyytiälä and Soroe) and have checked that he formatting and layout of tables is consistent. We try to combine Tbale 6 and 7 but it leads to confusing and large table.

(9) Correlating time series data of simulated and observed stand level attributes will usually produce impressive R2 values without having too much meaning

Yes, we agree and yet it is a widely used model evaluation metric. In our examples you see clear difference between annual, daily and monthly correlation analysis. However, exactly because of the limitations of correlations in mind, R2 is not the only criteria for comparison we are using. We are also using other criteria, especially the model efficiency.

(10) The discussion deals mainly with the new evaluation experiment. No in-depth well-founded linkages to other earlier studies are provided.

This is correct because the focus of our paper was the evaluation of the recent model version with the PROFOUND data. However, we mentioned the paper by Reyer et al. (2014), which provide a detailed model evaluation) and in 4.2 (Evaluation of carbon and water fluxes and 4.3 (Evaluation of soil water content and soil temperature) we inserted some remarks about how these 4C results compare to earlier evaluation studies..

(11) There are not that many models available that are applicable over a broad range of tree species and site conditions that could be used for climate change impact studies. Thus, provision of the context to other similar models would definitely improve the manuscript.

We appreciate the reviewer noticing the assets of 4C being and available and applicable over a wide range of environmental conditions
The main focus of the paper was evaluation of 4C. The comparison or description in the context to other similar models is planned in other ongoing papers.

(12) The conclusions are generally positive despite partly rather weak results of the new evaluation experiments. This may leave an interested reader also quite puzzled.

We revised the conclusion

(13) The abstract should be revised, too, based on an improved version of the manuscript.

We revised the abstract.

Further comments:

L 45: corrected

l 77: features described in the above mentioned papers

l 86: in the very extensive model description

l 88: model structure → Overview

l 96: all elements in a cohort are identical, there nothing happens in a cohort, but the number of elements can be reduces by mortality

l 105: see new title of this chapter

l 111: replaced

l 112: revised

l 115: the top line is not clear! Please explain. We explained it

l 151: revised

l 152: heartwood is the correct name, the basal area of a tree is divided into sapwood area and heartwood area

l 161: reference inserted

L 181: revised

L 188: water content and soil temperature

L 195: revised in chapter 2.1.1.

L 213: revised

L 217: revised

L 224: reviewer: take care when using this form:

Nature recommends the followings https://www.nature.com/nature-research/for-authors/write

"Nature journals prefer authors to write in the active voice ("we performed the experiment...") as experience has shown that readers find concepts and results to be conveyed more clearly if written directly. We have also found that use of several adjectives to qualify one noun in highly technical language can be confusing to readers. We encourage authors to "unpackage" concepts and to present their findings and conclusions in simply constructed sentences."

L 227: revised

L 270 reviewer: not necessarily. that the smallest model time step is daily doesnt mean that you rely on daily climate input.
We revised it
The references indicate the sources of the CO_2 scenarios for the RCP climate scenarios, which are not used in this study but available in the model.

L 296: revised

L320 revised with explanation of PROFOUND

L 331: revised

L 334: arithmetic mean/ average and geometric mean are clear mathematical terms; stem biomass is sapwood plus heartwood

L 347: reviewer: why cant you say monthly? i.e. the monthly variation.????
We think that inter-monthly is more comprehensible.

L 359 reviewer: what is the relevance of this statement. there will always be a site where the fit is better than at another site.
Here, we compared Peitz and Solling.

L 360 reviewer: worse; not as good as, we revised it.

L 365: revised

L 367: the ME values do not indicate poor results. the results are actually poor.
The statistical measures indicate the quality of the results.

L 368 the quality of model results is measured by ME, which indicate poor or good model results in comparison to measurements

L 369 reviewer: not a good idea to regress time series data of accumulating variables. will always show high R2 values. you should better use differences (i.e. growth).

Annual differences in biomass or diameter from simulation results are influenced by mortality, which differs clearly from the observation data. We intended analyzing the times series of diameter and biomass and not the increments.

Similar analyses were shown in:

Seidl, R., Lexer, M. J., Jager, D., and Honninger, K.: Evaluating the accuracy and generality of a hybrid patch model, Tree Physiology, 25, 939-951, 2005.
Miehle, P., Battaglia, M., Sands, P. J., Forrester, D. I., Feikema, P. M., Livesley, S. J., Morris, J. D., and Arndt, S. K.: A comparison of four process-based models and a statistical regression model to predict growth of Eucalyptus globulus plantations, Ecological Modelling, 220, 734-746, 2009.
Grote, R., Kiese, R., Grünwald, T., Ourcival, J.-M., and Granier, A.: Modelling forest carbon balances considering tree mortality and removal, Agricultural and Forest Meteorology, 151, 179-190, 10.1016/j.agrformet.2010.10.002, 2011.

L 391: revised

L472: revised

L483: well, not very convincing model test.

This is not a model test, it should only show, that the initialization is reasonable

L 492: revised

L 496: revised

L 500: revised

L 629 reviewer: what is sufficient accuracy? you need to define and explain your definition.
We deleted this expression.

L 665 reviewer: in this form this sentence is not very usefull.

We revised it:

Using the PROFOUND database enables the comparison with simulation results of other process-based models using the same database. This includes the potential to gain new insights into the process understanding of forest growth on the base of such model-intercomparisons.

Reply rev 2

Dear reviewer,

We would like to thank you for the constructive comments on our manuscript and the possibility to revise our paper. We have now uploaded a completely revised manuscript to accommodate all concerns the reviewers have articulated. Below, we reply to all comments in detail providing concrete reference to the sections that have been changed. We have also marked the main revised parts of the manuscript in green and the revised parts according to the other reviewer in yellow.

We hope that this new version accommodates all your concerns.

General comments: In “Description and evaluation of the process-based forest model4C at four European forest sites” the authors does exactly what the title states. Al-though GMD could be a suitable outlet for such a study, the current manuscript is not ready yet for publication:

(1) the model description skips over most of the numerical details which are the key interest of GMD readers;

Reply: We agree that this is very important. The numerical details of 4C are described very extensively in the technical model description (in the 4C repository https://gitlab.pik-potsdam.de/foresee/4C/tree/master/descriptions) and this is now clearly stated in the manuscript in L80/98. Moreover, we have rewritten the Sect. 2.1.3. (L190-469) to provide more details about the numerical details, model equations etc..

(2) given the model has been extensively evaluated before, the objective of this study needs to be clarified; and

Reply: Thank you for spotting this slight inconsistency in motivating our work. Indeed, 4C has been evaluated within almost every study we used (as is good practice). However, in parallel to the development of this manuscript, we have published 4Cv2.2 as an open source tool in late 2018. Our intention is to publish, together with the model and its description in the GitHub repository an evaluation of the model with the best available data, which are also open-source. This should allow potential model users to not only download and use the model but theoretically also reproducing exactly the same results as we did using freely available data. Moreover, the PROFOUND Database is currently being used as data basis for a forest model comparison study within the framework of the Intersectoral Impact Model Intercomparison Project (ISIMIP, https://www.isimip.org/) and hence our results are now available to be compared to other models run on exactly the same data. We have clarified these points in the manuscript in L80-94.

(3) the discussion is superficial and largely speculative. Although the simulation set-up and evaluation are sound but routine, moving beyond the current level of speculation would require more demanding simulation experiments.

Reply: We thank the reviewer for the critical advice. We have largely rewritten the discussion to bring in more depth and breadth, e.g. with regard to ground vegetation and carbon and water fluxes. We have also designed and including more complex experiments with regard to the soil water content (see reply to comment #10). However, we have not included too many new experiments as, in line with the now clarified objective of the paper (see comment 2)) we really see this paper as a background paper that should provide the potential users of 4C with the main approaches to evaluate the model, showing them current deficiencies and possible ideas to stir model development.

(4) The objectives of the manuscript need to be clarified. Given the absence of new model developments and extensive previous model evaluation at the site level, the objective of this study remains unclear. The model has already been evaluated against many more sites (including 100s if not 1000s of sites of BWI) with similar data so what was the hypothesis that justifies running the model against just four more sites. What do

you expect to learn from 4C by using the PROFOUND data that hasn't been revealed by the previous model tests? L474-L476 may contain a hint but this sentence needs to be rephrased as it is not at all clear to me what is being proposed.

Reply: We agree that the objective should have been spelled out much more clearly. In line with the comment 2) above, we have clarified the objectives arguing why evaluating the model at the 4 sites adds value. Moreover, many of the previous model evaluations were carried out with data that are not freely available and hence could not easily be reproduced by potential users. With the data from the PROFOUND database we have the possibility to be much more transparent.

(5) The writing is often imprecise and fails to satisfy the interest of a readership rooted in the modelling community. For many of the interested details the readers are referred to previous work or have to do with a vague description. I listed some examples in the detailed comments addressing the first 15 pages but the team of authors seems sufficiently experienced to revise the whole manuscript by adding the missing information either in the main text or an appendix without detailed editing from my side. The results section contains many vague classification such as: good, better, underestimated, less, smaller range, . . . the reader wants to see the numbers (some are given in the tables but they should be integrated in the text to substantiate such statements).If you insist on using qualitative language you will have to make a table showing that, for example, an R2 of 0.5 is considered average, R2 of 0.75 is considered good, . . . it would also be nice if there is a scientific ground for such a classification. Given that this is a stand-level model that comes with simplifications, what would the authors consider good enough? One cannot expect a 100% match but what would give you confidence in the model?

Reply: We have rewritten the entire manuscript trying to be as precise and detailed as possible and your comments have helped us a lot to do so. We have also clarified the link to the most up-to-date technical model description which contains more than 100 pages of technical details (Lasch-Born et al. 2018). Moreover, we have added more statistical details to table 1 wherever possible and have ensured that the statements in the results and discussion are backed up by the actual value of the statistical metric. We have refrained from classifying R^2 values or so into classes as we believe adding the actual value has more value for the primarily scientific audience of this paper.

Lasch-Born, P., Suckow, F., Badeck, F.-W., Schaber, J., Bugmann, H., Fürstenau, C., Gutsch, M., Kollas, C., and Reyer, C. P. O.: 4C model description, PIK, Potsdam, 133, https://dx.doi.org/10.2312/pik.2018.006, 2018.

(6) The figures contain some inconsistencies: in Fig 7 the obs are shown in blue and the sim in red, Fig 8 shows the obs in black and the sim in red, and Fig 9 shows the 25-75th percentile in black and the 10-90th percentile in red. Changing the meaning of the colors does not make it easy on the readers. Nine tables and 13 figures is a about two times too much display items for a research paper, especially because it looks like several of the tables could be merged and several of the figures are rather trivial or they simply complement the table; they show how the ME, R2 and NRMSE reported in the table looks like in data/simulation space. Such a figure could be useful to give the reader an idea how to interpret the tables but it becomes a burden to read when there are too many.

Reply: We thank the reviewer for these observations. We agree that the many figures and tables are sometimes breathtaking and therefore we have

- moved former Fig. 7 and Fig. 13 to the supplement.
- removed former table1 and table 5 and integrated its content into the text.
- Revised former table 6 and 7 to make them clearer. The tables help to get an overview of the long term annual means (table 5) and the statistical measures an daily, monthly and annual scale (table 6).

(7) The discussion (4.1, 4.2) refers to relationships, equations and parameters that are not presented in the model description. Given the readership of GMD consists of modelers, show the equations that are essential to understand the discussion.

Reply: We agree and we have revised the methods section by adding much more detail and the key equations as well as referring to the technical model description available online and ensuring that the points brought up in the discussion are properly introduced in the methods.

(8) Much of the discussion in 4.1 should have been presented in the results as it compares the simulation with other data sources.

Reply: We have thought about this for quite some time and discussed this amongst the authors. Finally, the editor recommended us not to include the discussion of how 4C compares with other data sources in the result section:

```
Editor: So you may move such discussion part from result section to
discussion section, and make a new paragraph or subsection about the
performance of your result in comparison with other data sources.
```

(9) The sentence “Collalti et al. (2016) also found a better performance for their 3D-CMCC-FEM model on a monthly scale for these sites.”(also see L540-541). Provokes more questions than it answers: what is the probability that this happens by chance? Which other sites did Collalti analyze? If a different set of sites where analyzed, thus this render this comparison meaningless? If you stand to your position and consider this a good comparison, what makes these sites stand out or which site properties makes them more easy to model their behavior?

Reply: Collati analysed 10 European forest sites including Hyytiälä and Soroe. We cites his publication and Grote (simulated Hyytiälä with his model) too because we assume that a qualitative comparison of our model results with their results underline the difficulties of sufficient simulations measured with statistical indicators. Model inter-comparisons (quantitatively) at the same sites with the same data are in progress in the framework of PROFOUND.

We also added here our ideas on the aspect of different model performance depending on time scale which was found in Collalti et al. (2016) and in our study and now write:

“The main reason here is the strong dependence of daily and monthly water and carbon fluxes on the daily and seasonal course of temperature and radiation. […]. Therefore, the relative importance of other variables, besides the meteorological parameters, as leaf area dynamics, transpiration

limitation due to water shortage the length of the growing season and, the ground vegetation increases at the annual scale, thus rendering these simulation results more uncertain."

(10) The discussion remains speculative and the model experiments do not allow to isolate causes. Isolating causes would have been required to result in new insights. The model could have been forced to use observed soil water contents, when compared with the available simulations, such an approach could have isolated the role of the soil water simulations in the simulation of latent heat and would have enabled the authors to move beyond the current speculation.

Reply: We would like to thank you for your criticism and have revised the paragraphs regarding the deviations of carbon and water fluxes to be less speculative. We have also taken up your idea of a 4C-simulation with measured water contents instead of simulated ones but were only able to carry it out in the case of Hyytiälä. Unfortunately, this model exercise was not possible on the stand in Sorø, as the measured values were only available for one soil layer (8 cm). The results of this model exercise and the more detailed analysis of the AET deviation are now discussed more clearly and focused:

"4C simulates acceptable AET values on daily and monthly time scales (ME≥0.65) but not on the annual scale. […].Like for GPP and NEE, the strong systematic bias at the Sorø site is a result of neglecting the observed ground vegetation in 4C. In the model we assume that there is no transpiration when there are no leaves."

and

"Unfortunately, in Sorø only the water content at 8 cm…[…]. So, this model exercise also did not yield any further results than that the soil water does not play a role for the deviations from the measured AET at Hyytiälä for all time scales."

(11) L641-642 demonstrate this point as this conclusion is not based on the results. As far as I could tell none of the simulations show the possible contribution of an understory. A literature search on GPP and transpiration of the understory could help the authors to make this point but at current there is no evidence in support of this claim.

Reply: We agree that more detail is needed here and we added more literature citations to strengthen our argumentation and now write:

"Based on reported transpiration values for the ground vegetation comparable to our pine sites Hyytiäla (56 to 76 mm year-1, Launiainen 2011) and Peitz (173 to 185 mm year-1, Lüttschwager et al. 1999) also the ground vegetation in a beech stand as Sorø explains the simulated deviation of 10 to 20 mm month-1 from February to May (Fig. S11)."

(12) The discussion on L598-594 is trivial (all observations come with uncertainties). The authors should use this measurement uncertainty to set a clear target for the model-data comparison. Either the measurement errors are small enough to justify a comparison and then a "good match" (defined by the target) would give confidence to the model or the measurement errors are considered too large and then the data should not be used. The current approach of using the data but casting doubt over their quality when there is a mismatch between the model and the data is unfair. Same measurement errors exist at

the sites where the model performs well. What does this tell you of the model given that the data come with considerable uncertainty?

Reply: We agree that this point could be misinterpreted and have removed the sentences altogether.

(13) Note that on L629 the term “sufficient accuracy” is used. What is considered sufficient in this context?

Reply: We have removed/placed this statement.

Detailed comments: Title: version number. Give all the information that is required to link this study to a specific model code. Make sure that when you update the code, the code underlying this manuscript can still be downloaded. I thought GMD required at least a doi referenced code.

Reply: The model version is included in the title 4c2.2, and a DOI is now available. The model code (open source) is available in the repository:

https://gitlab.pik-potsdam.de/foresee/4C/tree/master/source_code

L26: abstract start with describing the objectives of the model, not its age.

L34: delete one “on”

L33-37: Split in two sentences to enhance readability.

L46: delete “and”

Reply: The abstract was revised and all these points addressed.

L50-58: No need to be complete but a more structure overview of what is currently available and what makes FORESEE a unique model would be informative. How does it differ from some relatively recent developments in land surface models (see Naudts etal in GMD (the ORCHIDEE model), Fisher et al in GMD (CLM model with cohorts based on ED)) and how does it differ from individual based models such as iLAND (Seidl etal). In this respect, the introduction lumps stand-level and individual-based models in one group. This seems a bit rough to me. Highlight some similarities and differences between all these models. I really hope there are some differences in functionality and/or underlying principles. If not, the community is wasting a huge amount of efforts and resources by repeating the same work over and over again.

Reply: Thank you for this comment. We have rewritten the relevant section in the introduction (now **L49-67**) and also clarified the role of individual- and stand-based models. We have also identified one possible point of confusion: 4C is a model applied to the stand-scale, yet, because it operates with individual cohorts (that theoretically could also be individual trees) it can also be considered as an individual-based model. We have clarified and better structured the discussion of similar and different modelling types now explicitly referring to the differences to land-surface models and land-scape models. The text now reads:

“For this reason, stand-scale process-based forest models (PBM) have been developed over the past 30 years that try to explain forest growth and development based on an ecological understanding (Fontes et al., 2010; Landsberg, 2003; Mäkelä et al., 2000a; Medlyn et al., 2011). These models can be

stand-based or individual-based and many of these models were developed to study climate change impacts on forest productivity (see review by Reyer (2015)) or matter dynamics (water, carbon, nitrogen) (Cameron et al., 2013; Constable and Friend, 2000; Kramer et al., 2002), or the effects of forest management (Fontes et al., 2010; Porte and Bartelink, 2002; Pretzsch et al., 2008). These models are typically operating at stand scale and yet include similar process detail as Land-Surface models (Fisher et al., 2015; Naudts et al., 2015) that are typically applied to larger scales. They can also be applied to larger spatial scales but typically without considering interactions among landscape patches, as opposed to landscape models that place particular emphasis on the processes connecting different patches of forests such as dispersal or propagation of disturbances ((Seidl et al., 2012).”

L68: “respects the principle of parsimony”. Several modelling groups embrace this principle but how was it applied. How is it reflected in the code? Which parameters were removed after test runs because they did not helped explaining additional variation? What are the thresholds for parsimony. Please detail how the parsimony principle has been applied throughout the 20 years of developments.

Reply: Thank you for these questions. We have deleted the parsimony sentence in the introduction and inserted a section in the overview section where we explained how we applied the parsimony principle in the model:

“*Physiologically-based as well as empirical functions were selected and implemented in order to provide general but also as simple as possible solutions (law of parsimony (Coelho et al., 2019)). As an example, the empirical relationship between foliage mass and height used in the model is described with one single function that uses only three species-specific parameters (see Sect. 2.1.3. equation (13)). This function was selected after analysing the general applicability across species as well as simple species-specific parameter estimation. Another example is the reduction of the number of parameters in the soil temperature model of 4C. Analyses showed that it is sufficient to use the air temperature of the last three days for the calculation of the surface temperature of the ground and to determine the corresponding three parameters (Suckow, 1985). It should also be noticed that the temporal resolutions of process descriptions is selected specifically for medium- and long-term analyses (several hundred years). Therefore, a coarser resolution is preferred for processes such allocation that may vary at short time scales but still obey general rules on the longer term.”*

L82: The introduction does not describe a problem with the previous version of 4C or a hypothesis for which new model functionality is required to test it. Why did you believe it was worth to make the effort to conduct and write up this study? It could be that in the past the data used for model evaluation were spatially and/or temporarily incoherent (for example, NPP from MODIS but biomass from NFI or NPP from 1970 through IGBP and biomass from 2010 through NFI) and that the PROFOUNF database is the first consistent database for forest models in Europe. In that case you put the model to a new test that is more challenging than any of the other evaluations done with 4C.Likewise it could be that new model code was developed since the last manuscript and that this code is now evaluated.

Reply: Thank you for this observation. As outlined above (see replies to comments 2) and 4) we have revised the objectives and the motivation for this paper (4C is open access, PROFOUND Data are freely available etc.).

L115: Add a caption to figure 1. What do grey boxes represent? What is the difference between a grey and a white box? What is shown by an arrow?

Reply: We have added a caption and the figure 1 is better explained now.

L118-124: List the four methods. The readers want to learn about 4C without having to read other papers/manuals/websites. This manuscript can be concise but it should be complete. The target audience of GMD are modelers, they want to learn how processes were modelled. It is frustrating to read a manuscript without getting the information one is expecting. These different approaches for the same model is a rather unique feature of 4C. It should be detailed so that the reader can better appreciate it.

Reply: thank you for inciting us to add more detail about this specialty of 4C. We revised the description of the light competition, production and allocation, water balance and soil and have provided clear reference to the full technical model description, which is the one document to contain all details.

L125-130: List the different methods. See previous comment.

Reply: thank you, we have also substantially revised this section by adding more detail.

L139: How is plant water supply formalized? Does it account for root water conductivity? State the assumption that is being made on root distribution and how it is calculated. If I understand it correctly, the plants take up water in proportion to the root mass in that soil layer. If the top soil layers dry, the plant will thus experience water stress. How does this assumption holds against data?

Reply: Thank you for this comment. The water supply in 4C is formalized in a way that all water of rooted soil layers can be extracted up to the wilting point. The water is extracted layer by layer up to the calculated demand. Therefore, if upper soil layers are dry but deeper soil layers can provide the water demand, trees experience no water stress. The proportionality to the relative share of fine roots is relevant to distribute the water of one soil layer between the cohorts but this only leads to noticeable effects in extreme drought events. 4C does not account for root water conductivity on a physiological basis but on an empirical water resistance function which can be found in the full model description Lasch-Born et al. 2018 on page 74. Due to the fact that the water supply is mainly driven by the calculated water demand (which we implemented in more detail in the revised paper version) we consider the water uptake by roots as not so important for the model overview and let the sentence in the paper unchanged.

L147-153: How is N calculated in the plant? Is the C/N ratio fixed? Is the C/N ratio dynamic? How does the plant C/N ratio interact with the soil C/N. How does it interact with (long term) environmental changes? Which nitrogen-species are distinguished (more meaningful for the description of the soil)?

Reply: Thank you for the comment. We added in the model overview the main equations to understand the principles of the approach regarding C and N turnover. Concerning the C/N ration within the trees we added a paragraph at the end of Sect. 2.1.3.*Production, allocation and growth*. We now write here:

"The nitrogen content in the tree results from constant species-specific C/N ratios separated for fine and coarse roots, twigs and branches, stem, and foliage. The C/N contents are used to calculate the nitrogen demand of the tree and interact with the soil, vegetation and atmosphere due to its influence on the mineralisation of the plant litter (see below)."

L205-206: First time specific species are listed. Given the model aims to be generic, I assume that different species relate to different parameter sets. List the different species in the general description if you want to keep line 205-206. Once you mention one species, the reader wants to know for which other species the model can be used.

Reply: We shifted and revised the section about tree species and their parametrization. It now provides much more detail about the parameterization philosophy of 4C.

L217-218: Figure 2 repeats Fig 1. Remove fig. 1. Add a caption. Boxes, arrows and colors need to be explained.

Reply: We agree the layout was confusing. To address this issue we have not removed Figure 2 but revised it so that the details of Figure 1 are now not shown in Figure 2 anymore. We think that it is clearer if we do not give all model interactions and sub-models in one figure.

L220: the typical disturbances are fire, wind, drought and pests. Following the description 4C only deals with pests. The title of this paragraph should be "pests". Based on the description, the implementation of pests seems very empirical. A mechanistic approach should probably simulate its own pest dynamics as a function of the environ-mental conditions (which is key in the climate change discussion and the question of whether forest will suffer from more frequent and more intense pest outbreaks).

Reply: Thank you for your comment. We basically agree with you however, our philosophy for including disturbances in 4C is slightly different: Because there are already available modeling approaches to simulate disturbance in a process-based way, we have decided to enable 4C to make use of that information without calculating it "online". Instead, each disturbance agent affects the physiology in 4C in a distinct way and hence we can simulate effects of disturbances in a mechanistic way without having to simulate the disturbances ourselves which would add a huge amount of complexity to the model. We believe we have found a very elegant compromise to include mechanistic detail about how disturbances affect plant physiology while still being able to use existing disturbance data or scenarios from much more adequate tools/data. We have now reformulated Sect. 2.1.3. *Disturbances* to better present our description and motivation for that approach.

L256: the least that should be reported is the minimum and maximum number of pa-rameters to give the reader some idea of the number of parameters

Reply: We agree and we have revised the tree species parameterization section (now 2.1.2).

L263-268: Rephrase "Calibration of the parameters is therefore not usually carried out when setting up the model for a new site." It is not clear when the parameters are calibrated. I fail to see "Therefore, in recent studies, 4C has also been calibrated using a

Bayesian framework (van Oijen et al., 2013; Reyer et al., 2016)" fixes the problems described above. Bayesian uses a prior and an uncertainty of the prior but it cannot distinguish acclimation from ecotypes. Rewrite this section as the logic seems to be flawed. Is there a atypical sequence that is followed (first photosynthesis, then mortality, . . .)? How many parameters are calibrated at once? Where do the priors come from? How was the uncertainty on the priors determined? Was heterogeneity used a proxy for uncertainty of the prior? This section lacks lots of basic information that is required to understand the model parameterization procedure

Reply: Thank you for this comment and indeed the entire section was very unclear. We have rewritten the entire section replying to your comments and referring to the key papers (van Oijen et al. 2013 and Reyer et al. 2016 for further details. The section now reads as follows:

„The philosophy of 4C is to rely on processes as close as possible to the underlying principles of forest growth, demography, carbon and water cycling, heat transport etc.. Covering the most important of such processes, one parameter set for each species can be chosen that reproduces species' growth, water and carbon cycling under a wide range of environmental constraints and hence can be kept fix over time and space without need for calibration. Therefore, the values of the 4C parameters were derived from the scientific literature, by expert knowledge or from other published models. If a variety of values were found in this way, the value set for 4C was determined detailed testing and sensitivity analyses (not published). Calibration of the species-specific parameters is therefore not carried out when setting up the model for a new site.

In recent years, more and more evidence has accumulated that different physiological parameters have been determined in different environments (Kattge et al., 2011), or are dependent on stand density or site fertility (e.g. (Berninger et al., 2005)). To address these issues of parameter uncertainty, we have tested calibrating 4C in a systematic Bayesian calibration studies (van Oijen et al., 2013; Reyer et al., 2016). The main goal of these studies was to analyse effects of parameter uncertainty on simulated net primary production (NPP) and forest growth. Reyer et al. (2016) used uniform priors for 42 parameters varying by +/-25 and 50% around their standard value and data from different Scots pine stands throughout Europe to calibrate 4C. The different calibrations showed that even though the output uncertainty induced by the parameter variations is large when projecting climate impacts on NPP, Bayesian calibration in the historical reduced those uncertainties. Most importantly, these tests showed that the direction of NPP change is mostly consistent between the simulations using the uncalibrated, standard parameter setting of 4C and the majority of the simulations including parameter uncertainty. Following a similar simulation set-up but examining results in a multi-model context, 4C was found to be the most plausible out of six established forest models (van Oijen et al. 2013)."

L277-278: "From these data tree cohorts are generated using distribution functions. "Setting the initial conditions is an essential step in a forest model. This description is too vague. Again, the readership are persons who want to know exactly that kind of details. If you feel this may overload the manuscript, add it into an appendix.

Reply: We give some more details about initialization (Sect. 2.1.4) and a reference to the detailed description.

Table2. is nice as gives the reader a concise overview of which aspects of the model has already been evaluated but the wording in the results column is too vague. What is good? What is called “underestimation” and when is an underestimation so large that it becomes unacceptable or just bad? Those statements should be quantified. Percentage deviations, RMSE or even correlation coefficients could do the job here.

Table 2. Add version numbers to the different tests. Given the range of publication years, I assume this table refers to different model versions.

Reply: Very good points! Thank you. We added the model version numbers. If statistical measures are available in the mentioned papers we added these as well. In some cases only visual inspections with graphics were done in the papers and statistical analyses were not executed. The wording (e.g. satisfactory, good correspondence) was used in such papers without statistical details.

L300: the manuscript does not describe any changes to the model. Without model version numbers, table 2 suggest all these test have been performed on the current model version. This provokes the question what you expect to learn from the new sites that you haven’t learned from the extensive previous test. The objective of this study and manuscript should be better described.

Reply: We have revised the objectives (see above). In Sect. 2.1 we inserted:

“This actual model version differs from its predecessors by a variety of model extensions and revisions. Starting from the first model version we enlarged the number of species and species parameters, we included new management methods (e.g. short rotation coppice), we revised the calculation of the effect nitrogen availability on growth and implemented the effects of pests on stand dynamics of among others.”

L305. Sorø needs to be introduced first. The reader has no idea yet that Sorø was one of the test sites.L308. I suppose you mean “recalibrate” because the parameters were calibrated earlier for other sites. Now it reads as if the parameters were never calibrated which sounds unrealistic for the totality of the parameter set (I expect the most sensitive parameters were calibrated, others were just based on literature values and others were just guessed and/or tuned because they cannot be observed).

Reply. We revised this section in Sect. 2.3 and it now reads:

“To evaluate the current version of 4C regarding long-term growth, as well as water and carbon fluxes we selected the four sites Peitz, Solling, Sorø, and Hyytiälä representing the main central European tree species from the PROFOUND database that allows to test forest models against a wide range of observational data (developed by the COST Action FP1304 PROFOUND; (Reyer et al. (in preparation), Reyer et al. (2019)). Additional data sources (Table 3, Supplement Table S3) for the sites were applied. In the scope of PROFOUND several other forest models are using these data and comparisons between the models regarding the results are ongoing.”

I stopped making detailed comments as I trust that the authors have enough experience in scientific writing to revise the remainder of the manuscript along the lines suggested by the above comments. Most of these comments require careful editing rather than an expert review.

Reply: We thank the reviewer for the time and the very valuable comments and for trusting in our experience. We have substantially revised the manuscript along the lines you suggested, clarifying the objectives and adding detail to make the model description easily accessible within one document as well as linking the key discussion points more to the actual process descriptions in the model.