Author responses to Reviewer 1 of egusphere-2025-1534

We sincerely thank Reviewer 1 for the thorough and constructive evaluation of our manuscript. We greatly appreciate the valuable comments and suggestions, which have helped us improve the quality and clarity of the paper. Our detailed responses to each comment are provided in red font below. All revisions made to the main manuscript are indicated by line numbers and refer to the new revised version. Additional analyses, figures, and tables have been incorporated into the revised manuscript and Supplementary Material (hereafter SM) as appropriate.

We substantially revised and extended the manuscript. Specifically, we *(i)* clarified the role of "influential outliers" and added new analyses quantifying model performance for the pre-2018 period (Results §3.1 and 3.2, SM 3.3.1); *(ii)* deepened the mechanistic interpretation of the post-2017 mortality surge by analyzing the relative importance of predisposing, inciting, and contributing factors (Results §3.1 and 3.2, new Fig. S 3.3.2, and new Discussion §4.3); *(iii)* conducted a targeted sensitivity analysis for Norway spruce key parameters (kDrSc and P_{bark}) to assess model robustness (new SM §3.3.7, Table S 3.3.7); *(iv)* examined spatial patterns of observed and simulated mortality using spatial autocorrelation metrics (new SM §3.3.3, new Figs. S 3.3.3.1–S 3.3.3.2); and *(v)* improved figure readability (Figures 2,3,4), clarified methodological descriptions (soil parameterization, model setup), and corrected minor errors and typos throughout.

We believe these changes strengthen the manuscript's transparency, reproducibility, and mechanistic interpretation of drought-induced tree mortality.

RC1 comments

In this study, the authors present a refined version of the DVM ForClim (version 4.2) with the aim to more accurately simulate drought-induced mortality of Norway spruce and European beech. In particular, they implement predisposing, inciting, and – in the case of spruce - contributing factors, which they termed a PI(C)-scheme. Importantly, the authors do not per se calibrate their model against observations in order to test whether their model implementation mechanistically captures mortality.

The presented results indicate, that for both species the stark increase in mortality observed during and after the extreme 2018 drought is reproduced. Yet, absolute mortality rates were largely overestimated for beech whereas the ongoing high mortality of spruce was not captured by the simulations. Based on 105 AWC-simulations, the authors moreover conclude a high importance of soil properties and soil heterogeneity (particularly for beech). Finally, for spruce only the simulations including a bark-beetle component were able to reproduce recent mortality rates. Eventually, the authors advocate for incorporating such mechanistic schemes into DVMs rather than striving for statistical/empirical models while also stressing the importance of an actual calibration of the model if simulating mortality under future conditions. As such, the study touches an important topic in context of dynamic vegetation models, namely the incorporation of drought-induced mortality which to date remains a major challenge. Consequently, the study can be considered very suitable to the general audience and scope of GMD.

Yet, before being publishable some major aspects have to be considered.

Firstly, while I particularly appreciate the approach not to calibrate their model against observations, I wonder to what degree the deployed 105 different AWC-scenarios in combination with the observed mortality rates (which feature a stark increase after 2017) do not result in similar problems arising from classic empirical models. In particular, I wonder to what degree the high error squares introduced by the high mortality rates after 2017 act as 'influential outliers' which have the potential to largely boost the main evaluation metric of r^2. Given this, it seems likely that those AWC-models are selected which match best the observed mortality increase after 2017. But do they also represent the models with the best mechanistic mortality implementation? In other words: how would your models perform if only simulating the years 2000-2017? In particular for spruce this seems to play a role since – based on supplementary figure 3.2.2.2 – the peak of mortality in some simulations occurred in 2021 (instead of 2018) or sometimes even already in 2017. Since all of the 105 simulations are based on similar model parameters, I wonder how such different mortality peaks can be achieved (stochasticity?) and to what degree the mortality-implementation really can be considered robust. I believe these points needs to be clearly highlighted when interpreting the results, since I am guessing that the r^2 will largely drop if not applying the model to the full period. At least, the authors should – in addition to results representing the full period – show how consistent their model evaluation is if excluding the years after 2017 to avoid the influence of these extreme years. This would then provide a better picture on how the mortality implementation performs under less dry conditions (for instance 2003 and 2015 were also pretty dry in Germany). And it would tackle my concern, that the model-selection procedure is biased by influential outliers, i.e. the extreme impact of the 2018 drought.

Author response: The reviewer raises important aspects here. Indeed, the R^2 is boosted by the 'influential outliers', i.e. the mortality wave post-2017. When the performance pre-2017 is considered, the skill of the model is much lower. Yet, it has to be taken into account that observed mortality rates are very low in this period (for beech even more so than for spruce), which implies that the uncertainty in the observed data is large, among others due to the very small sample size. Furthermore, the simulated mortality in this period is mostly 'background' mortality, i.e. it cannot be attributed to any particular cause and thus should best be viewed as being stochastic. Thus, there is no expectation that the model would match the pattern of the pre-2017 mortality well, and we prefer not to add such analyses to the manuscript. Regarding the temporal placement of the peak of the mortality as a function of the AWC scenario, it is important to note that about 40% of the scenarios feature the peak in 2018 or 2019 for beech, and about 30% feature the peak in 2019 for spruce. Yet, we set up the 105 scenarios to cover a truly broad suite of AWC conditions, including scenarios that feature very low (i.e. unrealistically low) values, and it is these scenarios that give rise to erroneous mortality peaks e.g. in 2003 and in other years. We now explain this in the revised manuscript at lines 372-378 for beech and lines 447-453 for spruce in the Results section, as well as in the Discussion section (lines 488-495). Additionally, we have added a figure in the supplementary material (SM 3.3 *"Mechanistic understanding of mortality: periods, PIC, spatial patterns"*, Figure S 3.3.1).

Secondly, while the authors conclude that incorporation of a bark-beetle component as well as soil properties (mainly AWC) appear as major drivers of tree mortality, the model-mechanisms causing the stark increase of mortality after 2017 are barely discussed. To provide the full picture, the authors should more deeply explore their model output in order to understand which environmental driver variables are responsible for the strong increase in mortality. Is this simply related to the extraordinary drought of 2018 or are there also predisposing factors (e.g. the dry year 2015) that contribute to this increase? This might also help to explain, why the observed ongoing high spruce mortality after 2018 is not captured by the DVM and it would also provide a better understanding of what may be simulated if applied to climate projections.

Author response: This is an important aspect indeed. In order to answer the reviewer's questions, we have added the analysis of the stand-level average (i.e., what fraction of the dead trees is experiencing the predisposing factors (slow-growth vs. drought-memory induced), as well as the inciting factor (fraction of trees affected). This is now shown in SM3, Figure S 3.3.2. We have also added an explanation on these results for beech (lines 379-387) and spruce (lines 468-479).

Thirdly, I understand why and generally agree with the authors not wanting to calibrate their model against observations. Yet, some of the model parameterizations appeared somewhat arbitrary to me and I wonder whether a sensitivity analysis of specific parameters wouldn't be meaningful to gain a better understanding of model behavior, which I think should be more emphasized in a model development framework. For instance, the classification of base annual probability for the bark-beetle outbreak classes as well as the factor of 2/3 applied for the inciting factors seem arbitrary but likely have an impact on the model outcome. For future implementations of the PIC-scheme it would be very helpful to know how sensitive the model reacts to these metrics. In other words: would it be possible to achieve models with similar or even higher performance if choosing different outbreak classes or a different factor? And to which of the two factors is the simulated mortality more sensitive? This information would provide readers with more guidance on how to implement comparable mortality mechanisms in other DVMs.

Author response: We appreciate the reviewer's suggestion and agree that targeted sensitivity analysis is important in a model development framework, especially when parameters are motivated by process understanding rather than being calibrated. We therefore have performed a local sensitivity analysis for spruce around the two best scenarios (i.e., no. 29 and 42) for the two parameters highlighted by the reviewer, i.e. the drought scaling factor (*kDrSc*) and the baseline outbreak probability (P_{bark}).

Specifically, we have perturbed one factor at a time by ±10% and ±20% and summarized the fit using MAE (i.e., lower is better) and adjusted R^2. We quantified sensitivity as a central-difference slope (change in the fit metric per 1% parameter change, *par*) with 95% bootstrap confidence intervals obtained by resampling years. The simulation runs from the originally submitted manuscript served as the baseline. Across both scenarios, simulated mortality is more sensitive to *kDrSc* than to P_{bark}. Point estimates of $\left|\frac{\partial \mathrm{MAE}}{\partial} par\right|$ are ~10× larger for *kDrSc* than P_{bark} in scenario 29 and materially larger in scenario 42 (cf. new Table S 3.3.7). Increasing *kDrSc* tends to reduce MAE and increase ${R_{adj}}^2$, while P_{bark} yields near-flat responses within a

change of ±20%. Scenario 29 shows that *kDrSc +20%* improves MAE by ~2.4% and increases R_{adj}^2 by ~0.008. The response under scenario 42 is flatter: MAE changes are ~±0.2–0.3% and small R_{adj}^2 gains are possible at negligible MAE cost. We have added the above-described sensitivity analysis to the manuscript lines 287-295 (Methods), lines 459-467 (Results), and its supplement (cf. new SM Section 3.3.7 and Table S 3.3.7). Overall, these findings demonstrate that the model behaves robustly with respect to both parameters and that the specific parameter choices made in this study do not critically affect model performance or interpretation.

Finally, while the comparison of model output and observations is based on mean mortality across all sites, the spatial scale is largely ignored. I understand why this is the case (only few mortality observations and stochasticity of the model likely result in spatially varying patterns) but it nevertheless deserves a mention in the discussion and maybe 1-2 display items in the supplementary to visualize the spatial patterns of simulated mortality. For instance, I wonder whether simulated mortality shows a spatial pattern or a rather random structure. If the former (spatial patterns) this might also point at the environmental drivers being mostly responsible for the mortality increase (see my second point above).

Author response: We thank the reviewer for this comment. We decided to assess the spatial patterns of observed and simulated mortality separately, rather than only their differences (cf. SM section 3.3.3. Spatial mortality). Specifically, we quantified global and local spatial autocorrelation for both European beech and Norway spruce plots to detect spatial clustering or dispersion. This approach provided a quantitative and spatially explicit assessment, besides comparing observed and simulated maps (cf. panels A of Figures S 3.3.3.1 and S 3.3.3.2 respectively), and the results confirm that both datasets exhibit very weak or absent spatial structure, consistent with the reviewer's expectation. This consideration has been added to the main MS, lines 370-371 (beech) and line 446 (spruce).

Only if these additional aspects have been taken into consideration the manuscript will transparently show how the suggested PIC-scheme may enhance the accuracy of mortality simulations in DVMs which I believe should be the major aim of the study. And even if some of the currently very convincing model performance evaluations (r^2 of 0.72 for beech!) would drop under a corresponding reanalysis (e.g. if adding validation metrics representative of the period 2000-2017 only) this is valuable and important information to the readers, since it would reflect that matching extreme patterns not necessarily means that mortality is generally well implemented (i.e. under less dry conditions). This in turn would also indicate the necessity to only very carefully interpret model output if applied to future climate projections. And finally, understanding what exactly drives the enhanced mortality after 2017 within the model may shed more light on the actual mechanisms driving tree mortality although inference from mechanistic models should be undertaken carefully.

Please find more detailed comments referring to specific sections of the manuscript below.

Abstract

Line 16: isn't hypothesis 3 a logical consequence of 2, i.e. if soil properties have a strong influence, local soil heterogeneity will automatically have a modulating impact.

Author response: We have incorporated H3 into H2 as suggested (cf. lines 17-20).

Line 21: please quantify 'hundreds'. How many plots in total?

Author response: Done (line 23).

Introduction:

In contrast to the abstract, you here combined the second and third hypothesis from the abstract. I personally prefer this combination, since H2 and H3 in the abstract are closely related. I suggest to adapt the 3 hypotheses also in the abstract (see also my point above).

Author response: Done (cf. lines 17-20).

Methods:

Line 100: What is the reason for the large gap in northeastern Germany. Aren't there any sites with beech or spruce? I would at least expect a couple of beech sites here and there. If not, please briefly mention the reasons for this geographic gap.

Author response: We selected only sites with documented drought-induced tree mortality (line 111), similarly to Knapp et al. 2024 and we referred to the author's same mortality observations, which explains the absence of sites in northeastern Germany. We have clarified this in lines 111 and 122.

Line 147: It seems that Marano et al., 2025 is currently under review. This obviously hampers the inspection of details as suggested. Would it make sense to show these details in the supplementary?

Author response: As the reviewer requested, the main equations of Marano et al. (2025) have been included in the new "*SM 4 | ForClim 4.1: Predisposing and Inciting factor scheme*".

Line 150: If I understand correctly, the heterogeneity was artificially generated. I wonder, how this reflects actual soil heterogeneity. And it isn't fully clear to me, whether you actually used existing soilmaps to characterize the soil properties which eventually determine AWC (I later learned this information comes below). Since soil properties are quite crucial for drought-related mortality (as you claim yourself), the soil parameterization is a quite crucial step, which deserves a more detailed description.

Author response: We agree that soil characterization is a critical aspect of modeling drought-induced mortality. As described in the submitted manuscript (lines 226-231, old version), we initially attempted to derive site-specific soil water holding capacity (AWC) and its spatial heterogeneity from soil maps (BGR, Meinel 2015), but found these data to be ecologically implausible, as described in the original submission. Therefore, rather than relying on such products, we resorted to a scenario-based sensitivity analysis, spanning a plausible yet admittedly expert-based range of AWC values and heterogeneity levels. This approach allowed

us to systematically test how variations in soil water availability and heterogeneity modulate drought-induced mortality, acknowledging that this constitutes an exploration of plausible conditions rather than a reflection of true soil heterogeneity. We have clarified this rationale in Section 2.3.4 (lines 231-233).

Line 182: please reword: 'This heuristic approach we used combined'

Author response: Done, line 185.

Line 182: did you run a sensitivity analysis to see how these somewhat arbitrary boundaries affect your model outcome? Might be worth a try to see how influential this classification is and whether a different classification might provide better/different results.

Author response: We agree that the classification boundaries for bark beetle outbreak probabilities introduce a degree of arbitrariness. Yet, we used this classification just to derive the value of the parameter P_{bark}, based on averaging the outbreak map of Hlasny et al. (2021). As explained in our response to the general comments (above), we conducted a targeted sensitivity analysis showing that this parameter choice has only a minor influence on model performance. Thus, the specific classification scheme does not critically affect the model outcomes.

Line 189: how is the stress status of trees defined/quantified? Please elaborate.

Author response: The sentence was ill placed and has been deleted and reformulated (line 190). The explanation is following in the text just below. The term "stress status" used in the manuscript refers to the algorithmic check of whether a tree meets predefined environmental stress thresholds, in this case, drought stress indicated by the annual drought index (mDrAn). When this threshold is crossed, the tree is considered a candidate for the application of stress factors in the PI scheme, including potential bark beetle infestation.

Line 196: this factor (2/3) is again somewhat arbitrary and would require a sensitivity analysis to quantify its impact on the model outcome.

Author response: As per the reviewer's request, we have performed a targeted sensitivity analysis to quantify the impact of the 2/3 scaling factor (kDrSc) on model outcomes for Norway spruce. The results, summarized in Table S2, show that simulated spruce mortality is moderately sensitive to this parameter: increasing kDrSc by +20% improved model fit (MAE −2.4%, R^2_{adj}+0.008), while decreasing it by −20% had the opposite effect. In contrast, changes to the baseline bark beetle probability (P_{bark}) produced negligible effects within the same range. Overall, these results indicate that model behavior is primarily governed by drought-related processes, and that the current parameterization is robust, meaning that small variations in either factor do not materially alter model performance or interpretation. This is clarified in lines 459-467.

Equation (5): according to equation 3, gGen can only reach values between 0 and 1 or exactly 2. I wonder whether this abrupt jump from (less than) one generation to 2 generations isn't arbitrary or maybe if there is a typo in equation 3 based on the query of 1.5 here. In any case, the threshold of 1.5 generations is again somewhat arbitrary. Please verify and potentially elaborate.

Author response: We thank the reviewer for this helpful comment. Indeed, the linear term in Eq. (3) was missing a factor 2. This has now been corrected (line 193). The choice of the 1.5-generation threshold is not arbitrary but reflects observed elevational phenology patterns of *Ips typographus* in Switzerland as in Jakoby et al. (2019). In their work, the authors demonstrated a strong negative relationship between voltinism and elevation and reported that at approximately 1,000 m a.s.l. the mean number of generations is ~1.5, i.e. "in about 50 % of all years two, otherwise one generation" (their Fig. 3b and accompanying text). Our corrected equation yields $g_{\mathrm{Gen}} = 1.5$ at $mDD_{An} = 1600$ degree-days, which corresponds closely to this elevational band under current Swiss climatic conditions. Consequently, the criterion $g_{\mathrm{Gen}} > 1.5$ identifies years or sites whose thermal regime consistently favors bivoltinism, a biologically important tipping point for mass-attack risk both in reality and also within our PIC framework. We have clarified this rationale in the Methods (lines 206–209) and now explicitly cite Jakoby et al. (2019) as the empirical basis for the 1.5-generation cutoff.

Line 219: that's the information I was expecting above. Maybe briefly mention above and refer to this section.

Author response: Done

Line 236: I suggest to show a supplementary display item which depicts the original data and - in comparison - the min and mean AWC values achieved by your approach to reflect how much of the original spatial variance in AWC is retained in your data. At current it is not clear to me how well your AWC-scenarios actually mirror reported AWC.

Author response: As explained on l. 226-227 of the submitted manuscript, the range of values for the 105 scenarios was taken from the mapped BGR product (cf. Figure S 2.1.1 in which the values are evident from the legend), whereas we did not use the spatial variance of the BGR product because we deemed it unreliable (as explained on l. 227-230 of the submitted manuscript, see also Figure S 2.1.2); we therefore do not think that comparing spatial patterns or the frequency distribution of the BGR data would be useful. Note that in our approach, at each ICP Level I plot the entire range of the scenarios is explored.

Line 239: what is the reason for choosing this specific period, i.e. 2000-2022?

Author response: We selected this period as it encompasses the most recent, well-documented drought events in Central Europe, notably the droughts of 2003, 2015, and 2018–2022, which have been extensively studied for their ecological impacts. Our aim was to assess model

performance against mortality patterns associated with these droughts. Starting in 2000 allowed us to capture pre-drought baseline conditions, enabling a comparison of forest dynamics before, during, and after these critical events, and particularly checking whether the new mortality scheme is capable of simulating the 2018ff. mortality wave while not simulating mortality in 2003 or 2015. An explanatory sentence has been added (lines 248-249).

Line 243: A subtraction can lead to negative mortality rates. Did you encounter this? If so, how did you treat this?

Author response: We did not encounter negative mortality rates. This is because the simulations are averaged across multiple plots, resulting in simulated background mortality values that are very close to the theoretical ones (0.921% and 0.495%). The subtraction is performed as a post-processing step (i.e., not within the simulation itself), which ensures that negative values do not occur.

Line 269: From equation 9 it seems you only used Msim, so why are you concerned about overfitting? Or did I miss something? Well, if you're concerned about overfitting, variance inflation should be considered e.g. by computing VIF for the predictor variables and excluding highly co-linear predictor variables (but again, if only using one predictor variable this does not make sense). So, I wonder which predictor variables you've been using at all. Please clarify and – if necessary - elaborate.

Author response: Indeed in our analysis we used only one predictor variable M_{sim} in the linear regression of Eq. (9), meaning overfitting due to multicollinearity or excess predictors is not an issue here. This was an error in the text, and the sentence has been corrected (line 280).

Line 275: I assume your R^2 adj values do not follow a Gaussian distribution since ranging from 0 to 1. Did you account for this in your GAM? Which datatype/family did you specify in your GAM? binomial? Please elaborate.

Author response: We thank the reviewer for this insightful comment. Although adjusted R^2 values are indeed bounded between 0 and 1, their distribution in our dataset was approximately symmetric and unimodal, without values close to the boundaries. Therefore, we fitted the GAMs using a Gaussian family with identity link, which is appropriate for continuous response variables that are roughly normally distributed within this range. Diagnostic checks confirmed that residuals were homoscedastic and approximately normally distributed. Consequently, a transformation or alternative family (e.g., binomial or beta) was not required. We have added the specification the GAM family that we used to the caption of Table S 2.2.

Results:

Fig. 3, panel A: the dark end of the color scale for R^2 does not always allow for depicting the size of MAE. Please adjust. Same for Fig. 4

Author response: We have revised the color palette of all main figures, namely Figures 2, 3 and 4 to enhance readability and ensure all elements within each plot can be recognized easily.

Lines 325-336: I wonder to which degree the overall variance of the data affects your r^2. It would be interesting to compute r^2 for the period before 2018 only to see how well the 'average' mortality under less dry conditions is captured by the models. It seems, that your model parameterization is able to capture the stark increase in mortality after 2017 but I wonder to what degree the model mechanistically captures mortality or whether it simply reacts to one extreme year. This aspect deserves more careful thinking and interpretation, particularly if using the model later on to predict mortality rates based on projected climate data. This does not require to rerun the simulations but only to evaluate their performance for a sub-period which is a common procedure when evaluating model performance.

Author response: Following our first response to the General Comments, we have included an additional breakdown of this evaluation in the Supplementary Material (cf. SM Section 3.3.1) and in the main MS in the Results section for beech (lines 372-378) and for spruce (lines 447-453) as well as in the Discussion (cf. section 4.1, lines 488-495).

Fig. 4: to avoid misinterpretation I suggest to use the same range of r^2 values in the legend as for beech to visually highlight that r^2 is much lower for spruce.

Author response: Following the reviewer's recommendations, the scale in Figures 3 and 4 has been adjusted accordingly to improve the readability of the two figures.

Line 386: Again, I wonder to what degree the extreme years after 2017 affect your model-selection process. Moreover, while I agree that the bark-beetle model is important to incorporate, it yet seems to require some improvements, given the inability to capture prolonged impacts of the 2018 drought. Also, from Fig. 3.2.2.2 in the supplementary it seems that some model runs obtained quite different mortality peaks (some in 2017, some in 2021). Since – if I understood correctly – the only difference in these runs was the AWC implementation, I wonder which circumstances have driven such temporally inconsistent mortality peaks. As suggested above, I suggest to gain a deeper understanding of the actual climatic forces driving the mortality peaks, since this also would allow for a better mechanistic interpretation of the parameterization.

Author response: We agree that the post-2017 period, characterized by extreme drought events, has a strong influence on the model selection process and interpretation. To quantify this effect, we compared the explanatory power (R^2) of the model calibrated for the full observation period versus the pre-2018 period (Fig. S 3.3.1A–C). For both species, model fits typically improved when including the post-2017 data ($\Delta R^2 > 0$), indicating that the extreme years provide valuable information on system behavior under unprecedented stress conditions rather than distorting the model selection.

For spruce, the inclusion of 2018–2020 particularly improved the fit of simulations that included bark-beetle dynamics (Fig. S 3.3.1 D-F), suggesting that the beetle submodel captures

the abrupt mortality pulse triggered by the 2018 drought. However, we acknowledge that the model still underestimates the legacy effects of bark beetle (line 454), likely because it does not fully account for lagged physiological decline associated with secondary beetle outbreaks following the initial event.

Regarding the temporal inconsistency of simulated mortality peaks (2017 vs. 2021), this variation primarily stems from differences in the available water capacity (AWC) parameterization. Scenarios with lower AWC experience stronger water deficits (low drought index values), leading to earlier and sharper mortality peaks (often in 2003 or 2018). In contrast, higher AWC buffers drought impacts, delaying mortality peaks to later years (e.g., 2021). Thus, these differences are not random but reflect distinct drought sensitivities of the simulated stands, reinforcing that AWC is a critical control on drought-induced mortality timing and magnitude.

In response to the reviewer's suggestion, we have added an explicit reference to the role of drought intensity in both the Results for beech (lines 336-336 and lines 358-360) and for spruce (lines 420-422) and Discussion sections (lines 490-492) respectively and have provided a mechanistic interpretation of model behavior in the new Discussion section 4.3.

Discussion:

Line 446: but it is not yet fully clear what these key drivers are. in other words: which environmental circumstances have led to the stark increase in mortality after 2018? Please elaborate.

Author response: The environmental drivers in question are mainly the drought memory and the annual drought index, which are part of the predisposing (drought memory) and inciting (drought intensity) scheme. As noted above, these considerations have been added new discussion Section 4.3.

Line 464: I generally agree that soil properties are important in mediating drought but some care needs to be taken when interpreting model performance since what you describe here most likely relates to your model-specific parameterization of beech. In reality this small-scale variability might not be as important for a relatively anisohydric species with relatively deep rooting systems.

Author response: We agree that the influence of small-scale soil heterogeneity on drought-induced mortality, particularly for anisohydric and deep-rooting species like European beech, warrants careful interpretation. Yet, beech's capacity to buffer drought stress due to deep rooting is debated (e.g., Gessler et al. 2022). Moreover, in drought-prone environments and on soils with restricted rooting depth (due to bedrock, compaction, or poor structure), even beech may depend on variable water availability within the upper soil profile, thereby rendering microsite heterogeneity relevant (Walthert et al., 2021). Thus, we would like to maintain our argument.

Line 484: Again I do agree, that soil conditions are important but we have to keep in mind that your interpretation relies on model output and thus mirrors how the model was parameterized. This not need to directly mirror reality. Thus, I would be more careful when deriving implications for real systems from model output.

Author response: We agree that model-based interpretations must always be made with caution. Yet, our study relies neither on model calibration nor on empirical fitting, but on ecologically grounded representations of soil-plant interactions. We thus are convinced that the emerging patterns, particularly the role of soil water availability and heterogeneity in modulating drought-induced mortality, are robust, and they are also consistent with empirical evidence (e.g., Walthert et al., 2021). Upon re-reading of these portions of the Discussion, we feel that the statements are cautious enough, and we would like to maintain them.

Line 488: when doing a species-specific calibration, a robust cross-calibration verification should be undertaken to avoid artifacts introduced by influential outliers (as the years after 2017). Please elaborate.

Author response: With this text, we wished to imply that that species-specific calibration should be accompanied by a robust validation strategy, such as cross-validation or split-sample testing, to avoid artifacts caused by influential outlier years, particularly extreme drought periods like those after 2017. We reformulated the sentence and added a cautionary remark (lines 653-655).

Line 539: You stressed to prioritize process understanding. Yet, the processes leading to the increased mortality after 2018 are barely discussed. Is this mostly related to one extremely dry year (2018), ongoing soil-drought, or predisposing factors? Please evaluate your model output accordingly to provide a deeper understanding of the underlying mechanisms.

Author response: The reviewer raises a crucial point, which we now have discussed extensively in the new Discussion section "4.3 Mechanistic interpretation of the post-2017 mortality surges".

References

Gessler, A., Bächli, L., Rouholahnejad Freund, E., Treydte, K., Schaub, M., Haeni, M., ... & Meusburger, K. (2022). Drought reduces water uptake in beech from the drying topsoil, but no compensatory uptake occurs from deeper soil layers. New Phytologist, 233(1), 194-206.

Walthert, L., Ganthaler, A., Mayr, S., Saurer, M., Waldner, P., Walser, M., ... & Von Arx, G. (2021). From the comfort zone to crown dieback: sequence of physiological stress thresholds in mature European beech trees across progressive drought. *Science of the Total Environment*, *753*, 141792.