Dear Reviewer 1, dear Reviewer 2, and dear Shilong Ren

Thank you for your detailed response to the initial submission of our manuscript. We are pleased to provide a strongly revised version, based on your comments and our own further revisions. All changes are highlighted in yellow, whereas light yellow indicates shifted but unchanged text. Further, we provide our detailed responses to your comments here below (**bold** text). Note that references here below without DOI also appear in the manuscript, where they are listed including DOI.

Best regards,
Michael Meier, Christof Bigler, and Isabelle Chuine

Response to reviewer 1

(https://egusphere.copernicus.org/preprints/egusphere-2025-460#RC1)

This manuscript introduces a new process-oriented leaf senescence model considering the three leaf development processes (young, mature, old). The new model doesn't seem to outperform the previous models or the Null model, most probably because the noise in the calibration and validation data is pushing the model simulation closer to the mean observation of the calibration sample.

Essentially, this manuscript highlights the need for modelers to consider the frequently overlooked uncertainty in underlying data, which could have profound implications and would be inspiring to be published here. Yet several main concerns remain.

First, the manuscript does not adequately address the DP3 model or its relevance to the conclusions. That is, the discussion and conclusions would remain unchanged even without the development of the advanced DP3 model. This would possibly weaken this research's merit for publication in this journal. This includes but not limited to:

- The introduction (L75-90 mainly) fails to convey the deficiencies of earlier models or justify the development of this new model. It is not apparent how the 3-phase model is an advancement.

  To make the need of for our DP3 model clearer we have restructured the introduction (former L75-90 are now L46-62) as well as have inserted Figure 1 and an additional paragraph (L94-106).
- Discussion for model accuracy and model error would be nice to focus more on DP3 model and provide more statistics.

  Model accuracy and model error are evaluated and discussed based on a model comparison. Therefore, the DP3 model and the models used for the comparison have to be mentioned. As the DP3 model behaves like the other models (i.e., including the Null model) we only see two key messages here, which have been clearly stated: (1) the accuracy of the DP3 model is within a reasonable range, and (2) the model error mainly depends on data structure, which implies noisy data. Moreover, rather than

introducing additional statistics, we decided to include the results regarding the senescence induction date (Fig. 7; L352–359).

- As there is no improvement in accuracy perspective, would be necessary to show the advances in formulation. Yet section 4.1 lacks the comparison with the previous models, as well as the scientific evidence to support DP3 model findings. And absolutely lacks more implications from model development (see below).

  We have now completed section 4.1 with that regard (L406–478).
- Model accuracy and model error sessions seems to elaborate the same issue, might as well thinking about making them more concise.

  The sessions elaborate different issues: Model accuracy focuses on the justification of the model, as probably both better and similar accuracy as previous models would justify the use of the DP3 model. Model error focuses on model behavior, illustrating that all the DP3 model as well as previous models behave similarly as the Null model. We hope to have clarified this through our revision (see below).

Second, the DP3 model development resembles more like a data analysis exercise. It lacks a solid theoretical foundation or a comprehensive scientific interpretation of the model's outcomes.

- Regarding the DP3 model development, need to justify assumptions first by providing enough evidence and references. For example (if I understand correctly):
  - Stresses act as a compound event instead of several individual events to trigger leaf senescence.

    Stressors act as individual events, but add up and accumulate as one (Eqs. 1 & 4). If this assumption is true or if each stressor accumulates individually, inducing senescence when either stressor-specific threshold was breached, is, according to out knowledge not known yet. However, because the referenced literature clearly mentions stress induced senescence in general rather than senescence induced by either cold stress or photoperiod stress in particular, we summed the stress events before accumulation.
  - Legacy of stresses (all of them) accumulated from the very early spring on tree leaf senescence.

    Current knowledge states that stress accumulates and may induce leaf senescence during the mature leaf phase (Fig. 1 in Jibran et al., 2013; https://doi.org/10.1007/s11103-013-0043-2). This assumption cannot be justified further and is based on all evidence we are aware of and which we have referenced (L46-63).
  - Within leaf lifespan the relationship between age and stress effects remains unchanged for triggering senescence.

    Non of the current studies (referenced in L46-63) implies a change in the relationship between aging and stress. Therefore, and by applying Occam's razor, we implemented the simpler formulation of a constant aging-stress relationship.
  - Reasons for choosing the three main stresses (especially for dryness) and three additional stresses.

The reasons for our choice have been given (including references) in lines 46-63.

- The discussion does not sufficiently cover the scientific importance of model selection, model formulation, or model parameter outcomes. In general, I would like to see more interpretations regarding them in the discussion part. I care about this because, as your manuscript indicates that no matter how much improvements you make for the model structure, you would fail to 'predict'. Therefore, prediction accuracy might not be supposed to be the only goal. Might be necessary to focus more on what the development will bring us scientifically. If makes sense, might be interesting to know:
  - What is the implication of 'more accurate' model. Does it really represent a model with better science? Or simply a model with less noise? This would be the foundation for the followings.

    This is an interesting question. At first, a «more accurate» model seems to be a model that predicts more accurately. However, it may well be a model that is formulated more accurately, which could benefit predictions under changing climatic conditions. We believe that an accurate formulation is important for valid predictions as well as for the research of the processes that the model simulates.
  - The simpler the model [g(x), sudden response, rather than h(x), gradual response], the better the performance. Is it a victory for science or for statistics? Also, the case for 'product' outperforms 'exponential' function.

    We are afraid, but feel that this question cannot be answered yet. While the true stress responses are likely gradual, steep stress gradients may well be approximated with sudden changes in stress. Moreover, because more parameters strengthen compensation effects between them, which we discuss in lines 441–446, responses with less parameters may yield a more stable model.
  - What is the implication for aging doesn't have much influence (better presented by 1) for senescence, which is a contrast to some previous research?

    This likely is an artifact from the calibrated threshold for photoperiod stress, which results in stress of 1 being added on almost each day in the second half of senescence (i.e., the period between senescence induction and LS_{100}). Thus, photoperiod stress acts almost like an age count during senescence (Fig. S3; L373–376).
  - Additional factors not important for leaf senescence prediction, why?

    While these additional factors were not important for leaf senescence prediction within the climate envelope used for model calibration and validation, they may be important within a wider climate envelope. Moreover, model selection based on more precise data with a clearer climate signal of leaf senescence may result in some of these additional factors being included (L558–575).
  - Three-phases model surpasses two-phases model, any implication?

    The young leaf phase, which does not answer to stress, becomes important as soon as the state of senescence must follow a path laid out by at

least two stages of leaf senescence. This exemplifies the need of such data (L564–566).

- Yet in table 3 it shows that the best $Y_{Aging,1}$ is only 1.57d. It is pretty short that the new model can basically regarded as a 2-stage model, which undermines the formulation test from the second iteration. Might be helpful if there is a sensitivity plot. Also I wonder if mature leaf span of around 70 days is realistic?

  The original manuscript listed the parameters of the DP3 model calibrated with the LS_{50} sample in Table 3. However, model selection was based on the LS_{50}-LS_{100} sample. We now included the parameter of both models in Table 3, and those of the DP3 model calibrated with the LS_{50}-LS_{100} sample appear more realistic. While this illustrates the compensation effect mentioned above (L441–446), it also shows the need for data that contains more than one stage of leaf senescence (L564–566).
- Downside of this model.

  We now discuss how the DP3 model could be improved in section 4.4 (L543–557).

In summary, we have now discussed the DP3 model thoroughly in section 4.1 (L404-478), together with way to improve the model in section 4.4 (L543–557).

Specific comments:

L1: There is a lack of 'the latest findings' for this research, or model development, also doesn't appear to be the focus of your research. Might be a better title if concentrate more on data quality.

While we have changed the title (L1–3), the latest findings have been summarized in the introduction (former L75-90, now L46-62) and are now illustrated with figure 1.

L30-34: Please be careful about the suggestion of using 'as few sites as possible' as you don't remind the difficulties of application at larger scales. And the last sentence is pretty hard to grasp the meaning.

We have rephrased our suggestion (L32-34).

L40: 'nutrient resorption' instead of 'nutrient retraction'?

We have changed this accordingly (L41).

L90: Yet what is the problem with the current progress to draw you developing this new model, instead of testing the existing models?

To clarify the need of for our DP3 model, we have restructured the introduction (former L75-90 are now L46-62) as well as have inserted figure 1 and an additional paragraph (L94-106).

L97: Confused by the exact meaning of 'relationship'. And not mentioned in the introduction session.

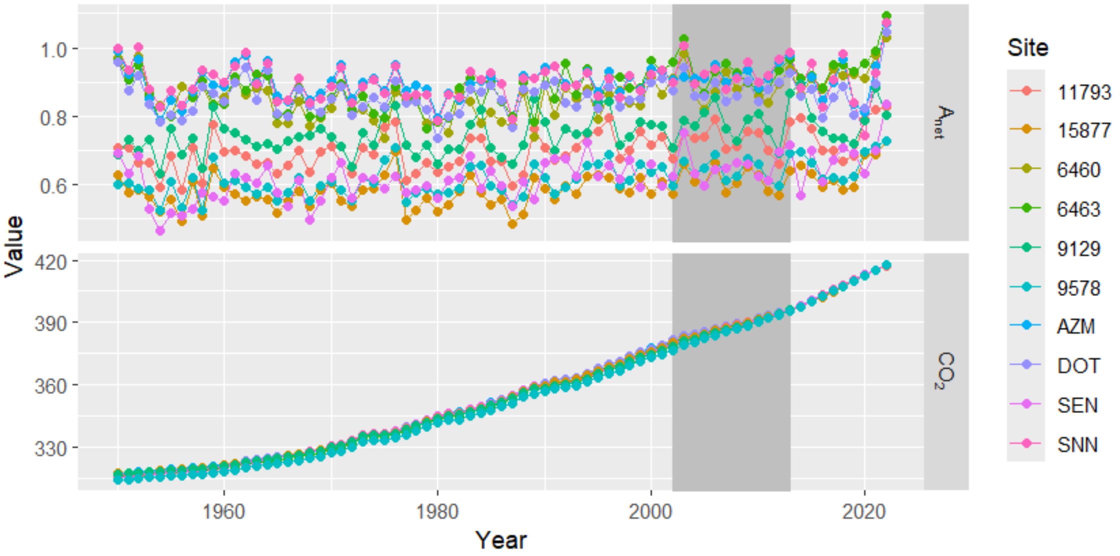
We have rephrased research question 1 (L113).

L105: Please indicate the phase id if possible.

We have done so in L121-122.

L132: I wonder why not taking one dataset as a reference and correct two CO2 datasets at a same level? This might bring a change to year 2013-2022.

We have chosen this procedure as we have been unable to decide, which dataset should be used as reference. Also, we are confident that the here applied procedure did not add an artifact to the calculated A_{net} that drives the PIA model (this was our only use of the CO_2 data; see the figure here below for ten randomly selected sites).



L135: I am a bit confused by what missing observation (variable) you mean here? And what do you mean by 'weighted' for the average LAI?

We clarified, which modeled data was used (L154–155) and what was meant by weighted average LAI, which is now explained at the very beginning of section 2.2 (L137–139).

L140: How do you calculate the 'day length, daily photosynthetic activity…' in absence of LAI for the past (1950-1981 at least)?

We have clarified this in L149–150.

L180, 185: what is the difference between the definitions of 'cold days' and 'frost days'?

We have clarified the difference in the text (L206 and through the new Table S3).

L184: The seasonal cycles of nutrient depletion will be represented by LAI. Yet I wonder if LAI would be a proper metric here if spectrum products would reflect more directly the nutrient supplies for the plants.

While LAI represents the seasonal cycle of photosynthetic productivity (A_{net}; see Sect. S1.2.2), nutrient depletion is modeled as a function of the accumulated A_{net} since the day of leaf unfolding. This is of course a rough approximation, which we have used in the absence of any better suited (soil) data.

L262: what is 'cold degree-days day length'

We have corrected this typo (L287).

L230-240: Model development part is a bit hard to follow. It would be far easier to follow this part if you could relate the texts with a figure similar to Fig. 4.

We have inserted Figure 5.

L232-234: The definition of the senescence rate is quite hard to understand. Would you please rephrase it into short sentences maybe?

We have rephrased and split the sentence (L254-257).

L234-235: The manuscript doesn't present the 2-phase development settings, which leaves me puzzled about the justification of your iteration design. Please adding more details about the 2-phase model in section 2.3, ideally, at least, with a adding figure like figure 2, and an explanation of the 2-phase model's structure.

We have done so in the first paragraph of section 2.3 (L164-171) and in Figure 3 (panels b and d).

Also, please provide the source code or at least instructions for the 2-phase model implementation in the DP3 model source code.

We have provided the corresponding code and mentioned so in line 605–606.

L239-240: Hard to understand the 'subsequent iterations' settings.

We changed "In subsequent iterations" to "In iteration 6" make the phrase easier to understand (now L263–264). Moreover, we included Figure 5 to illustrate the procedure.

L256-257: Including analysis along ELV considering the dry stress remains valid here. Yet it shows in Fig. 4 that the nutrient item has been omitted, and Fig. 5 clearly associates ELV more with photoperiod stress. Could you please provide more insights into this mismatch between your assumption and your findings in the discussion?

We have elaborated these issues in sections 4.1 and 4.4 (L404-478 and L543-557).

Fig. 4:

We have revised Figure 4 (now Figure 6) to clarify the development of the DP3 model. Your corresponding comments have been answered specifically here below.

- Could be related to previous comments. For the 1st iteration, please explain more for the setting of 2 phase model. And why here choose to go with 2-phase instead of directly 3-phase if you wouldn't add more discussions later.

  We started with the 2 phase model, which is now clearly illustrated in Figures 5 & 6. The setting of the 2 phase model is explained in Figure 3 and in lines 165-166.
- For the 4th iteration results, what is 3_D_gCgLgDdP_P? Is it 3_D_gCgLgDgP_P.

  Yes, it is. This has been corrected (now Fig. S1).
- For the 5th iteration tested formulations, what is '…fP_P'? Why not testing hP?

  This was a typo and we meant hP, which we have now corrected (now Fig. S1).
- For the 6th iteration results, same as above.

  This was the same typo, which we have now corrected (now Fig. S1).
- For the 6th iteration, does it mean you are adding the additional factors one by one until you test all factor combinations from 3 factors to 7 factors? If so, show it clearly in the figure. If not, please describe more clearly how you did it in the method session, and explain why you

are not testing all 7 factor combinations together.

We started with three stressors (i.e., the most probable stressors cold days, shortening days, and dry days) and continued with a forward selection of additional stressors, always considering each stressor through the response functions *g(x)* and *h(x)*. We have clarified this by revising section 2.5 (L251-271) and the revised Figure 6 as well as by including the new Figure 5.

Table 3: If possible, would you please add in the note that which day '-0.0016h' corresponds to?

We have done so in Table 3 as well as mentioned the corresponding dates in the not to the table.

Table S3: What does 'both' mean? Like exactly the same date for the causes of stress and aging to happen? I wonder if you could visualize the table by histograms to show how causes act differently regarding the variables you consider?

Here, 'Both' refers to the site-years during which aging and stress reached their thresholds on the same date. We have clarified this in the table (now Tables S5-S8). To illustrate these causes better, we have revised Figure 7 (former Figure 5).

L334: How much is the importance of the dry stress? At least show the results in the supplementary materials please.

We have now mentioned this in the text (L 367) and in Tables S5-S8.

Fig. 5a: Could you please label the number of site years for each sample?

We have altered the figure (now Figure 7) and visualized the number of site-years in the top row of each panel.

Table 4: 'P spatial'

We have corrected this.

L245: Would you please compare with the other research so we could see if the model error is within a normal range?

We are unaware of any study the evaluated the model error for stages of leaf phenology with one exception: Meier et al. (2023; https://doi.org/10.1111/gcb.17099) assessed the strength of the bias to the mean in leaf senescence models through the relationship between model error and phenological difference (i.e., the difference between a given observed date of leaf senescence and the average leaf senescence in the calibration sample). Thus, they showed that the model error depended strongly on this difference. In other words, the model error may be positive or negative and the absolute model error may be large or small, depending on the phenological difference. In consequence, the mean model error depends on the mean phenological difference, which varies between calibration and validation sample pairs and thus between studies. Therefore, to effectively compare the model error between studies, we would have to know the phenological difference between the calibration and validation sample in each study, which we do not.

L351: What does 'evidence' mean?

Evidence refers to the signal in the data against the null hypothesis, i.e., an effect of zero. Here, we have used the minimum Bayesian factor, which expresses the most optimistic change of odds between the null hypothesis and alternative hypothesis together with the *p*-

value. This has been explained in Sect. 2.7, L320–323). In addition, we have reformulated the sentence in the former line 351 (now L393–397) to make this clearer.

L365: As a quite important application, would you please give more details about it? Also the ecological importance of the timing transiting from mature to old leaf.

We have revised section 4.1 completely (L404–478).

L379-384: Can't understand the logical link between the two sentences. And please make it clear if it is cold stress or frost events (frost stress).

This has been rephrased (L457–464).

L390-L409: This part seems to lack a focus on the DP3 model. Would you consider delving deeper into the distinctions between your model and the other models more thoroughly?

The other models were introduced in section 2.6 (L284–295). Here, we rather focused on the match between the used model and the research question / task. We have now clarified this by altering the paragraph (L480–494).

L421: It is such a big gap from the proceeding sentence to this one, quite hard to follow the reasoning here.

We have rephrased the two sentences (L517–519).

L426-429: Would you elaborate on the explanations and implications of this phenomenon?

We referenced to the possible reasons, namely unrealistic model formulations, poor model calibrations, and noisy data, all of which were previously discussed in section S4.2). In addition, we referenced to Meier et al. (2023), who focused their study on this phenomenon. This has now been clarified through rephrasing (L523–530).

L455-458: I don't quite follow your suggestion for the 'revised observation protocols'. Would you please rephrase and make it clearer how to implement the new protocols?

We have rephrased this part (L564–569).

L459: It might need to be more careful in conveying the usage of 'as few as possible' sites for model development. Instead, I would like to know more possibilities if we could select sites with the help of current knowledge of inter-annual and inter-site variabilities.

We have extended our discussion of this subject (L569–575).

And considering that model calibration/validation and evaluation are distinct aspects of dataset application, I wonder if treating these two parts of the dataset separately (since involving different tools) could lead to improvements in this field? Also, could provide different insights for model accuracy and error assessments?

This is a very interesting thought indeed. As Meier et al. (2023; https://doi.org/10.1111/gcb.17099) have shown, the RMSE depends also on the choice of validation sample. Therefore, to enable inter-study comparisons, the validation samples should be selected as thoughtfully and sorrowfully as calibration samples. Again, this selection should be based on the research question the study focuses on, yielding different samples for, say, a study of the underlying process versus accurate projections under scenarios of future climate (L571–575).

L462: Please specify ‘three subsequent phases’.

The phases have been specified (L576).

L462: Please include more scientific indications from the DP3 model, in addition to the ‘structural strength’ of this model.

We have done so in lines 582–586.

L484-486: Same as the above-mentioned comment – this outlook is unclear to me.

We have clarified our idea of revised protocols (L600–603).

Response to reviewer 2

(https://doi.org/10.5194/egusphere-2025-460-RC2)

General Comments

This study presents an interesting and innovative approach to modeling leaf senescence, which remains a challenging process to simulate. The work offers two key contributions:

Unlike previous process-based models that focus solely on leaf senescence, the DP3 model attempts to represent the entire leaf development process from spring to autumn (that means from leaf unfolding to leaf senescence).

The authors analyze the influence of leaf senescence data quality on model performance, which often overlooked in modeling studies.

While the proposed model introduces a novel structure, the results indicate that it does not yet simulate leaf senescence dynamics well. The authors attribute this primarily to data quality limitations. However, given the complexity of leaf development processes (from unfolding to senescence), model performance may also depend on how well these processes are represented. Phenology data derived from camera observations (e.g., PhenoCam) are less susceptible to observer bias and sampling uncertainty compared to traditional ground observations. Have the authors considered using such datasets to further evaluate the model structure?

While we thoroughly discuss the representation of the processes in the model formulation (L404–418, L485–494, L521–530) as well as the uncertainty in visually recorded phenology data (L511–519, L531–541), we did not consider to reevaluate the model with new data from PhenoCams. Although this is a very compelling idea and would likely yield valuable results, we have no such data right now to try this and would be very interested in accompanying such a new study with researchers who would have extended enough time series of leaf senescence dates evaluated with images.

The hypothesis that aging and stress drive leaf development is compelling. A discussion comparing this approach with more conventional growing-degree-day-based models would strengthen the manuscript. For instance, what are the advantages of using aging and stress as drivers instead of accumulative growing season temperature?

We wrote a new paragraph to discuss this (L404–418) and tried to include this thought in the introduction to newly derived hypotheses from the DP3 model (L427) to discuss it later (L454–464).

Interestingly, the results highlight cold, daylength, and dry stress as key drivers—similar to the Growing Season Index (GSI) model, which relies on minimum temperature, photoperiod, and vapor pressure deficit (VPD). Did the authors test VPD as an alternative drought stress indicator?

Unfortunately, we did not test any alternative drought indices. However, we now discuss this shortcoming in section 4.4 (L545–550).

Specific Comments

Line 43, 105: Please clarify the definition of leaf senescence. Autumn phenology typically distinguishes between leaf coloring and leaf fall as separate stages. How is senescence defined in this study based on both events?

While we use leaf senescence as collective term for leaf coloring and leaf fall (now stated accordingly in L44–45), we based our study on the autumn phenology stages BBCH95 and BBCH97 for pome and stone fruit according to Meier (2018, https://doi.org/10.5073/20180906-074619) as now specified in L121–122.

Figure 1C: The delayed leaf senescence at higher latitudes appears counterintuitive, as senescence usually occurs earlier in such regions. Could the authors provide insight into possible causes for this pattern?

The regression through the function geom_smooth in the R package ggplot2 was calculated separately to each response variable (i.e., average LS_{50} and average LS_{100}) as well as separately to each explanatory variable (latitude, longitude, and elevation. Thus, a positive relationship emerged between both LS_{50} and LS_{100} and latitude probably because the more northern sites are generally lower elevated. This is misleading, as you have pointed out, and we have corrected it accordingly. In the revised version of the manuscript, we fitted a linear regression to combined latitude, longitude, and elevation for each response variable (Sect. S1.1.2). These regressions indicate a negative relationship between the response variables and each explanatory variable. Figure 2c (former Figure 1c) was adjusted accordingly by plotting the results of these regressions for each explanatory variable, while keeping the other explanatory variables constant (i.e., set to the mean).

Table 1:

Please include details on the spring phenology (LU, leaf unfolding) data.

The additional information was included.

The dataset combines observations from PEP725 and SPN. Were these collected using the same protocols? If not, how might protocols' differences affect model performance? Have the authors tested the model using only one dataset to assess potential improvements?

We had no access to the precise protocols used to collect the data. As these protocols were established by different institutions from different countries, they likely differ (Menzel, 2013, https://doi.org/10.1007/978-94-007-6925-0_4). However, the same stages were visually observed among countries (i.e., the stages BBCH15, BBCH95, and BBCH97 according to Meier, 2018, https://doi.org/10.5073/20180906-074619). While we did not use data from only one country, the example script we provided together with the code for the DP3 model (Meier, 2025, https://doi.org/10.5281/zenodo.14749339) runs on data from only three sites. In consequence, the accuracy of the predictions for both the observations in the calibration sample and validation sample is considerably improved. This emphasizes our suggestion that the heavy noise in the used data blur the signal of leaf senescence. Thus, comparing the DP3

model with current models based on observations that do not contain any sudden changes in the mean (Auchmann et al., 2018) is an important way forward, which we now suggest in lines 576–580.

Line 121: Please briefly describe the E-OBS dataset.

We now do so (L139–141).

Lines 124–125: Could the authors elaborate on the temperature correction method applied?

We had done so in the lines 127–130 of the original manuscript, but probably did not emphasize this enough. We now restructured the paragraph a bit, such that the temperature correction method is now easily identified (L143–148).

Line 135: Which remote sensing dataset was used?

Here, we referred to the remote sensed CO_2 dataset. However, this was unclear, so we recited the dataset (L156).

Line 139: Since LAI and CO_2 concentration are provided as monthly data, how was daily photosynthetic activity derived?

These data were combined with daily values of surface shortwave down welling radiation, day length, and mean temperature. We now have clarified this in L160–161.

Line 141: The Keetch-Byram Drought Index (KBDI) was selected as the drought metric. Were other indices (e.g., SPEI, PDSI) tested? If so, how did they compare?

Unfortunately, we did not test any other drought indices. Considering the many drought indices there are, such a comparison would have inflated our manuscript too much. However, we totally agree with you that such a comparison would certainly be very valuable and believe that it would yield an entire study by itself. Nevertheless, we now briefly discuss this in section 4.4 (L545–550).

Line 152: Does "several" refer to 34 sites? If so, please specify for clarity.

Yes, in the end, we constructed and tested 34 formulations. However, this number of formulations is a result rather than a component of the method applied. Therefore, we mention it in the first line of the results section (L328). However, in order to avoid confusion, we simply omitted the word «several» in the method section (L173).

Lines 200–204: The parameters *a* and b0 are set to 0.01 h. Could the authors justify this choice?

These are examples. The calibrated values are listed in Table 3. We now made this clearer in the text (L224).

Lines 218–219: Please define "extreme conditions" (e.g., hottest temperatures >30°C, coldest below a certain threshold).

Rather than using a threshold to identify these conditions, we selected the site-years that contained the hottest 10-day period during the growing season observed in the dataset. We did so, because we wanted to select exactly 250 site-years. We have now specified this in lines 240–241.

Figure 4 (3rd iteration): Does *f* represent h(x)? If so, please clarify in the caption.

Yes, it does. We have now corrected this and revised the figure (now Fig. 6) completely.

Table 2:

dm→o: Should this be interpreted as the simulated transition timing from mature to old leaves?

Yes, it should. We have now corrected the definition (Table 2).

SAging,I / SStress,i: Do these states accumulate since LU (leaf unfolding)?

No, these states are the accumulated corresponding rates since the transition from young to mature leaf. We have now clarified this in the definition column of Table 2.

Line 380: The authors associate cold stress with spring frost events. Could the importance of cold stress after midsummer also be examined?

Not directly. An additional assessment would be necessary to examine cold stress accumulated before and after summer solstice. Rather than including such an additional assessment, we now included some results regarding the relative importance of cold stress during senescence (i.e., the period from senescence induction to leaf senescence; L372–376; Fig. S3; Tables S5–S8) and further discussed effects of cold stress (L454–478).

Line 388: "This maybe" → Please revise for clarity (e.g., "This may be due to...").

We modified the sentence as «This may be explained by unrealistic model formulations, poor model calibrations, and noisy data to drive and calibrate the models, all of which we discuss here below» (L480–484).

Line 400: The sentence structure could be improved for readability.

We adjusted the sentence structure (L497–500).

Response to Shilong Ren (community comment 1)

(https://doi.org/10.5194/egusphere-2025-460-CC1)

The authors developed a new and systematic model of the senescence process of woody plants and analyzed the impact of data quality on the simulation of autumn phenology. The research perspective is unique. After previous revisions, I think there are no issues with the article's structure and writing quality. I have only one question: According to the multi-model comparison results in the article, it is not difficult to find that the DP3 model does not outperform the PAI and DM2 models in terms of simulation accuracy. It seems that the improvement significance of the new model is not significant. Could it be that the model process is too complex and has too many parameters?

We have now included the results for the DP3 model calibrated with one and with two stages of leaf senescence (i.e., with the stage when 50% of the leaves have turned color or fallen, LS_{50}, as well as with LS_{50} and the stage when 100% of the leaves have turned color or fallen, LS_{100}). These two calibrations led to different results. For example, the young leaf phase lasts 41 days when calibrated with LS_{50} and LS_{100} but only 1 day when calibrated with LS_{50} only. This illustrates a compensating effect between the different model parameters (i.e., different parameter sets yield similar results), which we now discuss in L441–446). However, while the probability for such compensating effects arguably increases with the number of parameters, it should be irrelevant for the accuracy of the predictions, provided that these remain in the space of the calibration conditions. This accuracy may only suffer by a high number of parameters when the calibration algorithm cannot handle them. Here, we selected the algorithm generalized simulated annealing, which has been used to successfully calibrate models with up to 30 parameters (Xiang et al. 2013, https://doi.org/10.32614/RJ-2013-002).

Moreover, we carefully tuned this algorithm to the models (Sect. S2.2), why we are confident that the complexity of the model did not have any adverse effects on the accuracy of the predictions.