Author's answer (AC) to comments of first referee (RC1)

General comments:

RC1: This manuscript reports a very comprehensive assesment of current process-oriented models of leaf senescence in temperate deciduous trees. It considers aspects that are rarely addressed in phenological modelling, and often overlooked or reported quite superficially in other manuscripts, despite their potential strong influence on the interpretation of the research. The considered aspects are related to the model calibration and evaluation (namely the scale of calibration site- vs. species-scale, the choice and parameterization of optimization algorithms, the cal/val sampling strategy), and their effect on model projections.

In essence, this manuscript has the potential to become a vademecum for phenological modellers, and possibly beyond this community (i'm thinking here of modellers working with models simple/fast enough to allow large numbers of computations), providing an example on how to rigorously design cal/val and projection studies. The downside is that the manuscript is very long, and sometimes difficult to follow due to the comprehensiveness of the tests performed and the results reported. This is not prohibitive to me, and I would like to read more often phenological modelling studies conducted in such a rigorous way. Hence I do not ask for a general reduction of the manuscript length. However, I strongly recommend the authors to provide a section (such sections are called "boxes" in some journals) highlighting what they identified from their work as good practices for phenological modelling.

This section would ideally list items regarding to the aspects they deal with (e.g. "how to rigorously sample a phenological database for cal/val of a phenological model", "which optimization algorithm to choose" etc.) and giving practical numbers / orders of magnitude / rules of thumbs useful to other modellers. This may require including in this "box" some definitions located here and elsewhere in the manuscript, possibly with examples (e.g. what is a "stratified" sampling etc.). I think most of this is already present in the manuscript, but it is dispersed and quite difficult to find. To put it more bluntly, currently the manuscript has a strong potential but no strong take-home message, and leaves the reader wrung out after an avalanche of valuable informations.

In other words, I would like this manuscript to offer two levels of reading: the very detailed one that is currenctly presented. And another, more synthetic one, which would help spread good practices in phenological modelling.

> AC: Thank you for your nice summary and general comments. We are happy to hear that you see the potential of our manuscript becoming a vademecum. We considered your suggestion of adding a section/box that synthesizes our methods and results as well as provides practical guidelines to modelers. Further, your specific comments were duly considered and answered here below. Corresponding changes in the marked-up version of our revised manuscript are highlighted in yellow, whereas green and blue signify the changes in response to the comments of the 2nd referee and other changes of the authors, respectively. Line numbers refer to the marked-up version of our revised manuscript.

Specific comments:

RC1: L25-26: should this sentence be understood "in general" for all leaf senescence models?

> AC: Yes. We made this clearer with the expression "current models". (L28)

RC1: L50: In French, it is customary to call this scientist "Réaumur" (not "De Réaumur")

AC: We changed this accordingly. (L116)

RC1: L63: meaning of "was not given" ? Clarify

AC: We clarified the sentence. (L127-128)

RC1: L88-89: about space-for-time approach, the paper by Jochner et al. is interesting, and not appearing in the reference list: Susanne Jochner, Amelia Caffarra, Annette Menzel, Can spatial data substitute temporal data in phenological modelling? A survey using birch flowering, Tree Physiology, Volume 33, Issue 12, December 2013, Pages 1256–1268, https://doi.org/10.1093/treephys/tpt079

AC: We were not aware of this publication and cited it now. Of course, the impossibility of site-specific models to project to new sites remains, why we did not alter our statement but referenced the study you mentioned above. (L155)

RC1: L113 "in contrast": I do not understand the logical link with preceding sentence here

AC: We clarified the sentence. (L179-180)

RC1: L120-121: site-specific vs. species-specific calibration: behind this is the question of tree populations local adaptations, that is not considered here. It was in Delpierre et al. 2009 (and Chuine et al. 2000). Mention it somewhere.

AC: We briefly mentioned local adaption and the indicated references in L150-151 and L153-154.

RC1: L153: the assertion "the proper order (e.g. the date for leaf coloration was before the date for leaf fall)" is wrong: leaf fall can occur before leaf coloration. Or at least, part of the leaf fall can occur before reaching BBCH94 (= 40% of leaves colored or fallen) considered in this paper (L159). Which BBCH code did you consider for leaf fall?

AC: We checked the order according to the BBCH codes, i.e., BBCH10 had to occur before BBCH11 (i.e., leaf separation before leaf unfolding) and BBCH94 had to occur before BBCH95 (i.e., 40% of leaves colored or fallen before 50% of leaves colored or fallen). To our knowledge, these orders cannot be reversed. However, the description of the codes was certainly misleading. We changed this accordingly in L219-220 and L225-226.

RC1: L154: "After corresponding correction..." is unclear. Rephrase.

AC: We clarified the sentence. (L221)

RC1: L174: why using Tmin as a "general" driver? Models often use the daily average temperature.

AC: Indeed, except for the models TPMt, TPMp, PIAGSI, PIA+ and PIA- (Lang et al. 2019, Zani et al. 2020), all compared models use daily average temperature, whereas Jibran (2013) and Lim et al. (2007) describe cold stress to promote leaf senescence.

Since cold stress likely relates more to daily minimum temperature, we used the latter in all models. This is also what was done in the currently most recent model comparison (Zani et al. 2020), which can now directly be compared with our comparison. However, we agree that this adaptation of the models CDD, DM1, DM2, SIAM, TDM1, TDM2, PDM1, PDM2, TPDM1, and TPDM2 must be pointed out and discussed, both of which we did in sections 2.2 and 4.5.2 (L298-299 and Sect. 4.5.2, L1058-1068).

RC1: L178-179: 0.25° is a quite coarse spatial resolution, notably when it comes to mountainous areas. Any correction of temperature with altitude (through lapse rate)?

AC: To not further complicate and lengthen the study, we did not correct temperature with elevation. However, we discussed this in the newly added section 4.5.1 (Sect. 4.5.1, L1034-1057).

RC1: L183: what is a "climate model chain"? Is one CMC corresponding to one particular climate model run under a particular RCP?

AC: A CMC is a particular combination of a global and regional climate model. (Clarified in L249-250)

RC1: Table 1: From Suppl Mat 2, it is unclear how you implemented the relatively complex responses to GSI and Anet described in Ziani et al. 2020. Did you actually code those responses? What cast doubts to me is your use of the term "apparent photosynthesis" in Suppl Mat 2, though it seems from Suppl Mat 1 that you indeed computed photosynthesis. This should be clarified in Suppl Mat 2, e.g. with a mention to Ziani et al. 2020 (their suppl. mat.).

AC: Yes, we coded these responses according to the Supplement S3 (Sect. S2.2 and S2.3). Following your comment, we clarified that GSI, Anet and Anet–w were calculated by ourselves. (Supplement S2: p. 2 and Supplement S3: pp. 1 and 4)
Unfortunately, we are unable to fully understand your doubts as “Suppl Mat” 1 and 2 probably point to Supplements S1 and S2 but these supplements explain the used data (S1) and phenology models (S2). However, we believe you referred to the Supplements S2 and S3 instead, which explain the used models and model drivers, respectively, and answer your comment accordingly. To our understanding of Egle (1960) and Wohlfahrt and Gu (2015), for example, the apparent photosynthesis (A_{net}) equals the difference between the real (or gross) photosynthesis (A_{gd}) and light (or daytime) respiration. And this is what we calculated in Eq. S32 of Supplement S3, defining light respirations as the daytime fraction of apparent respiration (R_d):
$A_{net} = A_{gd} - R_d \times L / 24$.
Our understanding from Zani et al. (2020) is that they also did the following. While they described their models in the main part as dependent on photosynthesis, Eqs. 20 and 22 in their supplement clearly point to the apparent photosynthesis.
Therefore, we strongly agree with you that it must be clarified if the apparent or real photosynthesis was used. We now have done so in the changes in section S2.3 of Supplement S3. In addition, we calculated these drivers ourselves, which we also specified in section S2 of Supplement S3.

RC1: L248: were the models tested in their ability to simulate trends in observed data (if any over 20-65 years?)

AC: Model accuracy was assessed solely based on the RMSE and we did not perform an additional assessment of the model's ability to simulate past trends. However, this would certainly be an interesting study, which we may conduct in the future.

RC1: L259-260: Does this mean that the parameters used for model evaluation and model projections were possibly different? Why so?

AC: Yes, in site-specific calibration, the parameters were possibly different. The site-specific models were evaluated with a 5-fold cross-validation, which leads to five sets of parameters. Which of these sets should be used for projections? Maybe the set that led to the smallest RMSE or the average parameters calculated from all five sets? On the one hand, different RMSE are likely related to differences in the observations used for validation and thus do not justify the use of one parameter set over another. On the other hand, average parameters are unlikely to lie in any local optimum or even the global optimum. Compared to these alternatives, the selection of parameters derived from the calibration with the whole data appeared most reliable to us.

RC1: L440-441: I'm unsure we are here talking about model external validation. If this is the case, it is not particularly surprising that site-specific calibration can sometimes yield, at the very same site, unrealistic results when the model is used to predict unknown data. Indeed, site-specific parametrisation are more prone to over-fitting (few data points over which to fit the model), as compared to species-specific parametrisation (which includes many sites).

AC: We are not sure, if we understand you correctly. With "external validation" we refer to the fact, that the models were validated with observations that have not yet been used for their calibration. These "external" observations, however, were recorded at the same site as the observations used for calibration. This may lead to unrealistic results due to overfitting, if the number of observations left for calibration is low, which is exactly what we wanted to study (i.e., effects of the sample size). We agree that such unrealistic results should be expected, why we deleted the word "surprisingly" in L440 (now L535).

RC1: Fig. 3b: inversion of the x-axis is confusing. It took me several minutes to understand this subplot (as compared to the text description), because i intuitively tried to interpret the x-axis with negative values on the left of the vertical dashed line.

AC: We felt that the plot is easier to understand when more accurate models are visualized with a dot further to the right, but maybe we were wrong. Since we cannot be sure, what our readers will prefer, we changed the plots such that the x-axes are not inversed anymore (L581).

RC1: L531: how possible is this? Considering that NA-producing and non-converging runs are assigned high RMSE values.

AC: We agree that this finding for species-specific models is counterintuitive. However, on the one hand, the LMMs are not perfect (adjusted R^2 range from 0.41 to 0.54) and the ranks of the reference model CDD, whose RMSE is specified on L618 and which has confused you, changed from 14 to 18 and 14 for species-specific models validated within sample and population, respectively. On the other hand, large samples led to most NA-containing calibration runs (Supplement S6: Fig. S1). At the same time, GenSA large samples also led to the lowest RMSE when NA-values were

replaced with 170 d. It may well be that many of these NA-producing runs based on large samples led to lower-than-average external RMSEs, especially since the punishing effect of a particular replacement is weaker in large samples. We added a paragraph in section 4.3 (L930-942), in which we discuss this issue.

RC1: L534-548: a question relative to optimization algorithm is: is the identity of the algorithm involved, or is it more the design of the simulation: in other words we see in SM4, Table S1 that the "normal" vs. "extended" runs of the calibration procedures include less vs. more iterations. I do not see an analysis considering the influence of this number of iterations per algorithm. Apparently, the influence of iteration number is small (Fig 3b: points are close whether considering "norm" or "extd" for one algorithm). If the "extd" simulation number was extended further, would that modify the results?

AC: The influence of the number of iterations per algorithm can be seen in Fig. 3a, for example. More iterations led to better results in most cases except for CMA-ES in site-specific models and for PSO in species-specific models. Since the number of iterations affects the step size with which the global optimum is searched (exploration-exploitation trade-off; L875), we assume that the step size became too small in the afore mentioned exceptions. This led us to the conclusion that careful tuning of the algorithms is very important (L875-878).
We agree that there was no specific analysis of the influence of this number. Analyzing the effect of only two states, i.e., “few” vs. “many” iterations is certainly not enough. A profound study of the effect of the number of iterations together with a suggestion of how to set the corresponding parameters of the algorithm would be helpful for the modelers.

RC1: Fig 4b: if CMC numbers point to particular climate models, one sees the effect of the climate model can be huge ! Were these climate models unbiased (against local observed climate data)? Climate model bias can influence process-based model simulations strongly, see e.g. Jourdan et al. 2021

AC: Yes, CMC numbers point to particular models. For time reasons, we refrained from bias correcting these data, but we now discuss this in section 4.5.1 (L1034-L1058).

RC1: L621-623: I do not see this on Fig. 4b (i.e. dot coefficients of DM2 and DM2Za20 are not remarkable relative to other models)

AC: Yes. Here we refer to Supplement S6: Fig. S6; Supplement S6: Table S27 (L718).

RC1: L698: rewrite to "phenology"

AC: Done (now L793)

RC1: L711-712: recall which models lead to the best results in Zani et al. 2020

AC: We already discuss the best models in sect. 4.1 and mention the corresponding original studies. We now referred to Lang et al. (2019) and Zani et al. (2020) as studies with which our findings do not agree (L834-836).

RC1: L741 "other models": recall briefly their characteristics here

AC: Done. See L834-836.

RC1: L750: rephrase to "local adaptation (Peaucelle et al. 2019)"

AC: Done. See L844.

RC1: L752: "the less such consideration is possible" : unclear, rephrase.

AC: Done. See L848-850.

RC1: L758-759: and/or that the observed data are more prone to observation bias, that can magnify if same observer is operating across years at a given site (in practice, observers inter-calib are rare). See Liu et al. 2021

AC: We added corresponding thoughts to Sect. 4.1 (L845-846, L855-856).

RC1: L788-789: interesting result. Where does this "17 sites" come from? I do not remember seeing that earlier in the manuscript.

AC: 17 is the number of sites in stratified samples based on the average timing of autumn phenology (L348). The result is mentioned in L549–L560, visualized in Figure 3d, and quantified in Supplement 6: Table S2.

RC1: L819: Cochran (1946) is missing from the reference list

AC: Done. See L1230-1231.

RC1: L852: remove "but see"

AC: Done. See L961-962.

RC1: L889-891: we touch here the question of local adaptation again

AC: Yes. We now mentioned this explicitly in L991.

RC1: L897: "Different models altered the reference shifts by -12 to +2 days", recall the average.

AC: Done. See L1007.

RC1: L970: "... and found our data to strongly encourage further research" is unclear.

AC: We tried to refer to the exploratory nature of our study. Since we mentioned this in L1125, we simply deleted it here (L1115).

Author's answer (AC) to comments of second referee (RC2)

RC2: Autumn leaf phenology impacts the biochemical and biophysical feedback of forests to climate. Modelling and projecting autumn leaf phenology of deciduous trees is therefore important and timely. Several studies have proposed and compared various modelling approaches. This study is different in the way that does not focus on a new modelling approach or only comparing existing approaches, but integrate model comparison with an analyze of the impact of different calibration procedures (e.g. site vs species), optimization, data sampling procedure etc considering their impact on model performance and model projections. For the latter aspects, analyses of the different scenarios is also considered. I find the study important and well done. The manuscript is also easy to read and very nicely synthesizes an huge amount of data. Practical useful recommendation are made in conclusions. I have however, some suggestions for improvement.

> AC: Thank you for your nice summary of our study. We are happy to hear that you liked the manuscript. Your suggestions for improvement were duly considered and answered here below. Corresponding changes in the marked-up version of our revised manuscript are highlighted in green, whereas yellow and blue signify the changes in response to the comments of the 1st referee and other changes of the authors, respectively. Line numbers refer to the marked-up version of our revised manuscript.

RC2: 1.while the text is clear, a scheme of the Methodology, thus a schematic synthesis of the different analyses performed, performance indicators used, etc, would be useful.

> AC: Done (Fig. 2, L484-495).

RC2: 2.I realize the analysis of the different formulation of the models considered is not the main focus of the study; yet, the different models are discussed and they will sure attract interest. So, I would add in Methods (not only in supplementary) a paragraph with a general description of the different type of model used (e.g. only driven by current temperature and photoperiod, or modulated by summer conditions, or by budburst timing), their key drivers etc. In practice, a description of Table 1.

> AC: Done (L273-L285).

RC2: 3.in Abstract and the entire text, I would not stress too much the modelled data of growing season length, rather focus on the date of autumn phenology. In fact, the data on growing season length are crucially affected by the spring phenology, which was only very coarsely estimated here.

> AC: We deleted several references to the changes in the growing season (L19, L20, L22, L1009, L1018, L1020, L1026, 1139 and L1140)

RC2: 4.the authors does not consider in fully another source of uncertainty, which is the quality of the observational data, comprising past climate data. For example, is the biases associated with considering climate at 25 km resolution negligible? (L79) I'm worried particularly for larix sites, which are often found on mountain regions. Similarly: what about the spatial match between LAI and soil water characteristics used when compared to data on phenology from PEP? Could large biases (at site level) be introduced?

AC: The resolution is coarse when it comes to simulate leaf phenology of a couple of trees at a particular site. We discuss this now in sect. 4.5.1 (L1048-1057). While there are finer gridded datasets available, the finer grid does not necessarily make the data more accurate. Alternatively, one may bias-correct and interpolate the data oneself. However, without meteorological measurements at the site of interest, one can only make sure, that the past and future data match, i.e., are equally inaccurate. Because this already increases the accuracy of projections, it is certainly a necessity when this accuracy is assessed. The main interest of our study, however, was to identify the relative importance of choices made during calibration for the resulting model performance and projections. This relative importance should remain largely unaffected by the degree of accuracy of the input data.

RC2: 5.autumn leaf phenology is actually made up by several phenological events (e.g. onset of chlorophyll degradation, 50% leaf coloration, leaf fall), with timing varying of several weeks (e.g. Marien et al 2019 New Phytologist, doi: 10.1111/nph.15991); are the models simulating the same exact event? (which one?)

AC: We applied the models to simulate BBCH94 (L219-226), defined as “40% of the leaves have colored or fallen” (Hack et al. 1992, Meier 2001) or “leaf colouration” (http://www.pep725.eu/; accessed on April 13, 2022). In their original publications, the models were used to simulate:

- “leaf fall / yellowing” (Dufrêne et al. 2005)
- “90% of the trees show yellow leaves over 20–50% of their crowns” (Delpierre et al. 2009)
- “more than 50% of leaves have changed color” (Keenan and Richardson 2015)
- “the day when almost all green leaves have colored” (Liu et al. 2019)
- “the day when about 5% of canopy leaves turn from green to yellow or red on more than half of the observed trees” (Lang et al. 2019)
- “the date when 50% of leaves had lost their green color (BBCH94) or had fallen (BBCH95)” (Zani et al. 2020)

RC2: L164: to my knowledge, beech does not growth at site with MAT below 6-7 degree C. A beech site at 0.6 degree MAT (subarctic conditions) is quite unrealistic.

AC: We agree and interpret this outlier as a consequence of inaccurate weather data due to spatial and elevational differences between a particular site and the center of the corresponding grid cell. Thus, we now discuss this inaccuracy in sect. 4.5.1, where we also mentioned this example (L1053-1054).

RC2: L843-845: the explanation based on severity of extreme is questionable; see Marien et al 2021 Biogeosciences (doi.org/10.5194/bg-18-3309-2021), and for a more fundamental impact of drought on autumn phenology see Marchin et al 2010 Oecologie (DOI 10.1007/s00442-010-1614-4).

AC: As we understand Mariën et al. (2021), there is an important difference between an observation of autumn leaf phenology based on canopy greenness vs. chlorophyll content when it comes to discuss the effect of drought. In our study, we worked with observations of canopy greenness. Therefore, we believe that our explanation holds if we specify that we talk about canopy greenness (L952).

RC2: L906-907: "… all analyzed models are based on the same process …". I do not agree: models based on current autumn conditions (temperature and daylength) are different than models considering also the impact of, for example, summer (e.g. implying legacy of tree growth on senescence) or budburst (e.g. implying constraint on leaf longevity).

> AC: We altered our conclusion slightly (L1016-L1017), while remaining convinced that effects other than temperature and day length remain under-considered by current models.

List of relevant changes

Change	As reaction to	Line number*
Insertion of summary (i.e., box) of our study	RC1	35 – 98
New paragraph in Sect. 2.2 that briefly describes the used models	RC2	273 – 285
New Fig. 2, which gives an overview of the applied methods	RC2	484 – 495
Change of x-axis in Fig. 3b such that is not inverted anymore	RC1	581
New paragraph in Sect. 4.3 to discuss the effect of NA substitution vs. the exclusion of NA-yielding runs on the RMSE	RC1	930 – 942
New Sect. 4.5.1 to discuss uncertainty in driver data	RC1 and RC2	1034 – 1057

*) Line numbers refer to the marked-up version of our revised manuscript.

References:

Delpierre, N., E. Dufrene, K. Soudani, E. Ulrich, S. Cecchini, J. Boe, and C. Francois. 2009. Modelling interannual and spatial variability of leaf senescence for three deciduous tree species in France. Agricultural and Forest Meteorology **149**:938-948.

Dufrêne, E., H. Davi, C. Francois, G. le Maire, V. Le Dantec, and A. Granier. 2005. Modelling carbon and water cycles in a beech forest Part I: Model description and uncertainty analysis on modelled NEE. Ecological Modelling **185**:407-436.

Egle, K. 1960. Apparente und reelle Photosynthese, Gaswechselgleichgewicht, Lichtatmung. Pages 182-210 *in* A. Pirson, editor. Die CO2-Assimilation / The Assimilation of Carbon Dioxide: In 2 Teilen / 2 Parts. Springer Berlin Heidelberg, Berlin, Heidelberg.

Hack, H., H. Bleiholder, L. Buhr, U. Meier, U. Schnock-Fricke, E. Weber, and A. Witzenberger. 1992. Einheitliche Codierung der phänologischen Entwicklungsstadien mono-und dikotyler Pflanzen – Erweiterte BBCH-Skala, Allgemein. Nachrichtenbl. Deut. Pflanzenschutzd **44**:265-270.

Jibran, R., D. A. Hunter, and P. P. Dijkwel. 2013. Hormonal regulation of leaf senescence through integration of developmental and stress signals. Plant Molecular Biology **82**:547-561.

Keenan, T. F., and A. D. Richardson. 2015. The timing of autumn senescence is affected by the timing of spring phenology: implications for predictive models. Glob Chang Biol **21**:2634-2641.

Lang, W., X. Chen, S. Qian, G. Liu, and S. Piao. 2019. A new process-based model for predicting autumn phenology: How is leaf senescence controlled by photoperiod and temperature coupling? Agricultural and Forest Meteorology **268**:124-135.

Lim, P. O., H. J. Kim, and H. Gil Nam. 2007. Leaf senescence. Annual Review of Plant Biology **58**:115-136.

Liu, G., X. Q. Chen, Y. S. Fu, and N. Delpierre. 2019. Modelling leaf coloration dates over temperate China by considering effects of leafy season climate. Ecological Modelling **394**:34-43.

Meier, U. 2001. Growth stages of mono-and dicotyledonous plants. 2. Edition edition. Blackwell Wissenschafts-Verlag.

Wohlfahrt, G., and L. Gu. 2015. The many meanings of gross photosynthesis and their implication for photosynthesis research from leaf to globe. Plant, cell & environment **38**:2500-2507.

Zani, D., T. W. Crowther, L. Mo, S. S. Renner, and C. M. Zohner. 2020. Increased growing-season productivity drives earlier autumn leaf senescence in temperate trees. Science **370**:1066-1071.