Authors' 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 manuscript (following after your comments and our answers) are marked 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.

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.

This Figure 2 is:

Concept

- Model performance (RMSE) — affect —
  - Phenology model
  - Optimization algorithm
  - Sampling procedure
  - Sample size ($N{:}q$, $\bar{N}{:}q$, and $s{:}S$)
  - Climate scenario (RCP&CMC)
- — affect — Model projections (Δ_{100})

Quantification of effects (RQ I & II)

Site-specific calibration: $\ln(\text{RMSE}) \sim Model + Algorithm + Sampling\ prc. + \ln(N{:}q) + (1|Site) + (1|Species)$
Species-specific calibration: $\ln(\text{RMSE}) \sim Model + Algorithm + Sampling\ prc. + \ln(s{:}S) + \ln(\bar{N}{:}q) + (1|Species)$ → ANOVA

Site-specific calibration: $\Delta_{100} \sim RCP\&CMC + Model + Algorithm + Sampling\ prc. + \ln(N{:}q) + (1|Site) + (1|Species)$
Species-specific calibration: $\Delta_{100} \sim RCP\&CMC + Model + Algorithm + Sampling\ prc. + \ln(s{:}S) + \ln(\bar{N}{:}q) + (1|Species)$ → ANOVA

Relation between model perfomance and model projections (RQ III)

Site-specific calibration: $\Delta_{100,Model} \sim RCP\&CMC + Algorithm + Sampling\ prc. + \ln(N{:}q) + (1|Site) + (1|Species)$
Species-specific calibration: $\Delta_{100,Model} \sim RCP\&CMC + Algorithm + Sampling\ prc. + \ln(s{:}S) + \ln(\bar{N}{:}q) + (1|Species)$ → ANOVA

Rank correlation of factors within group
(e.g. model rank$_{\ln(\text{RMSE})}$ ~ model rank$_{\Delta_{100}}$)

Rank correlation of model performance and relative influence on $\sigma^2_{\Delta_{100,Model}}$
(e.g. model rank$_{\ln(\text{RMSE})}$ ~ RCP&CMC rank$_{\Delta_{100,Model}}$)
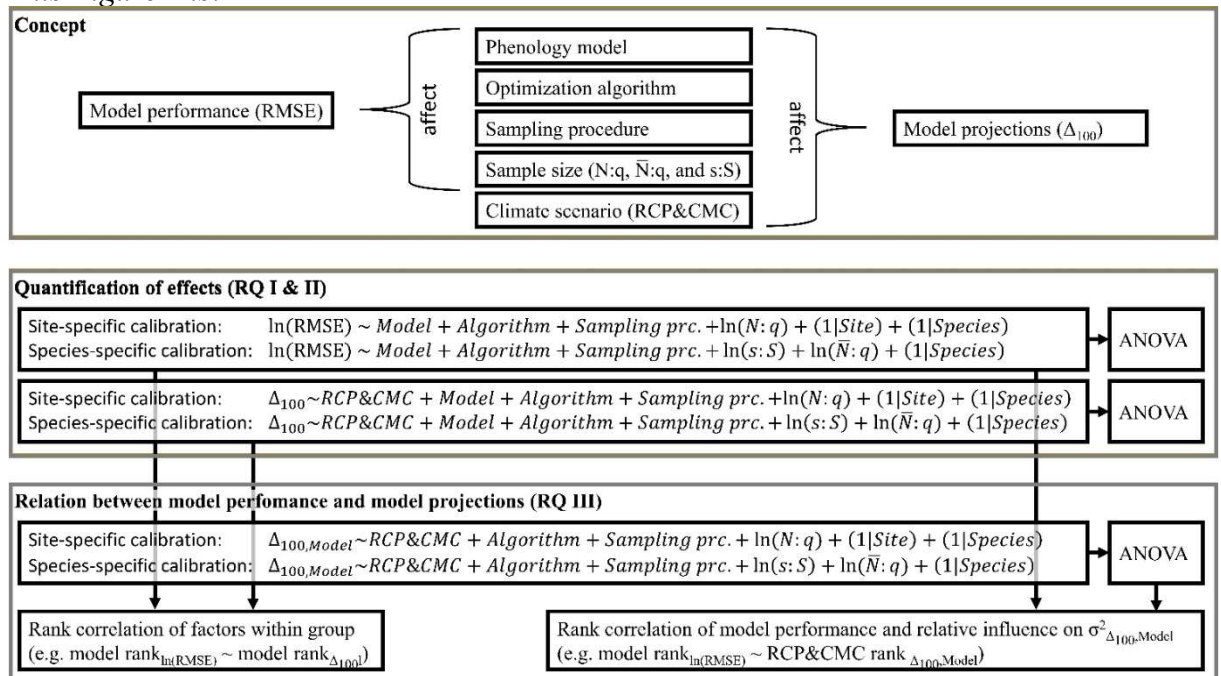
==**Figure 2: Concept and methods applied.** This study assumed that, in addition to phenology models and climate scenarios, the choice of optimization algorithm and calibration sample (i.e., sampling procedure and sample size) affect model performance and model projections (i.e., the root mean square error, RMSE, and the shift between autumn leaf phenology in 2080–2099 and 1980–1999, Δ_{100}, respectively). To answer research questions I and II (RQ I & II), the effects of these factors on the RMSE and==

$\Delta 100$ were quantified with linear mixed-effects models. Subsequently, the relative influence of the factors (e.g., all phenology models) on the explained variance (σ^2) of RMSE and Δ_{100} were quantified with type-III ANOVA. To answer RQ III, the effects on the RMSE were related to the effects on $\Delta 100$ and the influences on $\sigma^2_{\Delta_{100}}$ by calculating the Kendall rank correlations (e.g., between the effects of the phenology models on the RMSE and Δ_{100} or between the effect of the phenology models on the RMSE and the influence of each model on $\sigma^2_{\Delta_{100}}$. The phenology models were calibrated site- and species-specifically (i.e., one set of parameters per site and species vs. one set of parameters per species, respectively). Sample size was quantified by the number of observations relative to the number of free parameters in the phenology model (N:q), the average number of observations relative to the number of free parameters ($\overline{\text{N}}$:q), and the number of sites relative to the 500 sites of the entire population (s:S).

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.

Thus, the 2nd paragraph of Sect. 2.2 now reads:
While all models differ in their functions and drivers considered, they can be grouped according to the formulation of the response curve of the senescence rate and of the threshold function (Table 1). Models within a particular group differ by the number of free parameters, by the determination of the initial day of the accumulation of the senescence rate, or by the seasonal drivers of the threshold. The difference in the number of free parameters is relevant for the groups Mon− (Co) and Mon+ (Co). These groups contain two models each, which differ by the two exponents for the effects of cooler and shorter days on the senescence rate. Each of these exponents can be calibrated to the values 0, 1, or 2 in the models with more parameters, whereas the exponents are set to 1 in the models with fewer parameters. The initial day of the accumulation of the senescence rate is either defined according to temperature or day length in the two models of the group Sig (Co). The one or two seasonal drivers considered by the models of the groups Mon− (Li), Mon+ (Li), and Sig (Li) are site-specific anomalies of the timing of spring phenology, the growing season index, and daytime net photosynthesis accumulated during the growing season ignoring or considering water limitation constraints, as well as the actual leafy season or growing season mean temperature, the low precipitation index averaged over the leafy season, or the adapted low precipitation index of the growing season. All models are explained in detail in Supplement S2).

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.

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.

The 2nd paragraph of Sect. 4.5.1 reads:
Spatial and elevational differences between a particular site and the centre of the corresponding grid cell, from which the meteorological data were extracted, affect the input data. Gridded data may poorly represent the conditions at a particular site due to spatial and elevational differences. For example, precipitation and temperature can change in response to different terrain and the lapse rate, respectively, while the leaf area index and plant-available water capacity can change due to different vegetation and soil conditions. These effects of spatial and elevational differences were not considered in this study and may have led to inaccurate input data (e.g., average MAT for the site Grossarl, 47.2° N / 13.2° E at 900 m a.s.l., in the Austrian Alps was ~0.6° C, which makes beech growth unlikely; Holtmeier and Broll, 2020). The degree of inaccuracy probably differs between sites, which inflated the site effects on model performance and model projections. In contrast, the effects of models, sampling procedures, and optimization algorithms were probably unaffected by the inaccurate input data (cf. above), so these data most likely had a neglectable effect on our results.

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, 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 (see 2nd paragraph of Sect. 4.5.1 inserted further above).

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.

Thus, the 1st paragraph of Sect. 4.4 now reads:
Overall, the climate projection scenarios were the primary drivers of the projected shifts in autumn phenology, with the warmer scenario causing later autumn phenology than the cooler scenario, which is consistent with the currently observed main effect of climate warming. Having the largest influence in two out of three projection modes, climate projection scenarios explained between 46% and 64% of the variance in the 100-year shifts of autumn phenology. On average, the projected autumn phenology occurred 8–9 days later when projected with the warmer RCP 8.5 than with the cooler RCP 4.5 scenarios, which corresponds to the observed main effect of warming. Past climate warming was found to mainly delay autumn phenology (Ibáñez et al., 2010; Meier et al., 2021), but slight forward shifts or a distribution around a temporal change rate of zero have also been observed (Menzel et al., 2020; Piao et al., 2019). Such inconsistent past trends may be explained by the fact that autumn phenology (i.e., observed with canopy greenness rather than chlorophyl content; cf. Sect. 2.1.1 and Mariën et al., 2021) depends more on the severity than the type of weather event, with, for example, moderate heat spells causing backward shifts but extreme heat spells and drought causing forward shifts (Xie et al., 2015). Since the number and severity of heat spells is related to sites (e.g. warmer lowland vs. cooler highland sites; Bigler and Vitasse, 2021), such opposing effects of weather events may explain the large influence of sites on projected shifts in autumn phenology, as discussed below. In addition, the length of the growing season is affected by shifts in spring and autumn phenology for deciduous trees. Our projections were based on spring phenology that advanced by 20 days within 100 years. Subsequently, the projected growing season lengthened by 7–32 days (RCP 4.5) or by 16–40 days (RCP 8.5), even when autumn phenology shifted forward, as projected with some models and discussed further below. Therefore, our study supports a general lengthening of the growing season due to projected climate warming, as also suggested by Delpierre et al. (2009), Keenan and Richardson (2015), and Meier et al. (2021), in contrast to Zani et al. (2020).

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.

Thus, the 4th paragraph of Sect. 4.4 now reads:
The influence of phenology models on projected autumn phenology was relatively low and the range of projections relatively small. The largest influence of phenology models was 11% and occurred in projections based on site-specific models and hence was almost six times smaller than the influence of climate projection scenarios. While the underlying processes differ between each model (Delpierre et al., 2009; Keenan and Richardson, 2015; Lang et al., 2019; Liu et al., 2019; Zani et al., 2020), the influence of these differences on the projected autumn phenology did not affect the projected lengthening of the growing season: Different models altered the reference shifts of +8.2 to +11.6 days by -12 to +2 days, which resulted in some forward shifts in autumn phenology with the cooler RCP 4.5 scenarios, but never in a shortening of the growing season because the latter is calculated in combination with the -20 days shift in spring phenology. Moreover, the difference between the models lay within 14 days (i.e., −12 to +2 days), which is less than the uncertainty attached to recordings of autumn phenology based on human observations (i.e. due to small sample sizes and observer bias; Liu et al., 2021). In other words, the different process-oriented models led to differences in the length of the growing season that were smaller than the uncertainty in the data upon which we based our projections. Therefore, our results justify the assumption, that the examined phenology models do not differ fundamentally in their underlying processes, even if we acknowledge that the TDM, PDM, and TPDM models (Liu et al., 2019) behaved differently than the other models (i.e. they resulted in the largest forward or smallest backward shifts of autumn phenology). Rather, we suggest that the effects of temperature and day length, which all analyzed models simplify in different ways, mostly suppress the effects of other concerned drivers.

References of authors' answer:

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.

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.

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.

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.

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.