The manuscript focuses on changes in the output of the dynamic global vegetation model (DGVM) LPJ-GUESS due to increased accuracy of the simulated timing of spring and autumn leaf phenology. Spring and autumn leaf phenology mark the start and end of the growing season for deciduous trees and shrubs, thereby affecting biosphere-atmosphere interactions such as the carbon and water cycles simulated by DGVMs. Accurate DGVM simulations under projected future climate are pivotal for adequate climate mitigation policies, which justifies and substantiates the present study. However, I believe (1) that the study and the manuscript need to be completed, (2) that the conclusions and language need to be more precise, (3) that the readability and comprehensibility need to be improved, and (4) that the discussion needs to be deepened. First, the study compares a newly and currently implemented leaf phenology module (hereafter referred to as new and current LPM, respectively). While the new LPM was specifically calibrated the current LPM was not. Thus, before evaluating the effect of the module structure, it must be isolated from the effect of the specific calibration or it will be distorted and probably overestimated. Further, the study stops at the calculation of the difference between LPJ-GUESS simulations based on the current and new LPMs. I feel that these differences should be analyzed further (e.g., by testing the significance of the difference and comparing the differences between regions). The manuscript fails to present all the data used in the study and to present the software used to analyze the data. Second, the conclusion is compromised by the distorted effect of the module structure (i.e., due to the specific calibration likely increasing the accuracy of the new LPM). Certain results are labeled ‘significant’, but the study does not apply a measure and corresponding level for statistical significance. Third, the readability and understandability are affected by imprecise language and long sentences as well as by mistakes in grammar and syntax. Fourth, while some results are already discussed in the Results section, I would like to see a more focused and in-depth Discussion section. Below some suggestions.

Major comments

1. Completeness

Two phenology models (Caffarra et al., 2011; Delpierre et al., 2009) are calibrated and constitute the new LPM implemented in LPJ-GUESS. One of these phenology models simulates the start of the growing season (SOS; Caffarra et al., 2011), while the other simulates the end of the growing season (EOS; Delpierre et al., 2009). Simulated SOS and EOS are outputs of LPJ-GUESS, which further include simulated gross primary productivity (GPP), foliar projection cover (FPC), and actual evapotranspiration (AET).

1.1. According to my understanding of the manuscript, SOS and EOS according to LPJ-GUESS are directly taken from the new versus current LPM, and are evaluated against the same data with which the new LPM was calibrated (i.e., the NDVI of the GIMMS data set; L. 113–121). If this is the case, the results regarding SOS and EOS simulated by LPJ-GUESS (L. 302–310) are technically a comparison of the new versus the current LPM (rather than an evaluation of LPJ-GUESS, which should be clearly stated; see below). Moreover, the new LPM was specifically calibrated with the GIMMS data set (L. 251–254), whereas the current LPM was not (i.e., the module parameters were taken from the current LPJ-GUESS code). Because the accuracy of both LPM in simulating SOS and EOS was also assessed with the GIMMS data set (L. 302, not explained in the Data and Methods section), the increased accuracy of the new LPM (L. 302–310) is expected. It cannot be determined, to what degree this increase in accuracy is the result of the specific calibration or the formulation of the new LPM. To really compare the new and current LPM, I suggest to also specifically calibrate the current LPM (i.e., calibrated constants *a*, *b*, and *k* as well as calibrated longevity for the currently implemented models; L. 181–197), using the same calibration sample that has been used to calibrate the new LPM.

1.2. While GPP, FPC, and AET simulated by LPJ-GUESS were compared between the LPJ-GUESS running with the new and current LPM, GPP simulations were further compared with the (not introduced; see below) VPM GPP product. These comparisons include the results of the LPJ-GUESS running with the new versus current LPM as well as the difference between these results. Here, I would like to see more, such as (1) a comparison of the spacial distributions of GPP, FPC, and AET when simulated with LPJ-GUESS running with the new versus current LPM and (2) an evaluation against observational data.

1.3. The authors refer to some results as ‘significant’ (L. 372, 386, and 486) but mention neither a significance level nor a method with which the statistical significance was determined. I am aware that ‘significant’ has the literal meaning of ‘notable’ and may be used in that sense. In my opinion however, the term ‘significant’ usually refers to the value of a significance metric (e.g., the *p*-value) in scientific studies, why I urge the authors to also use it in this latter sense.

1.4. Following data was used but not introduced (therefore needing introduction in the Data section): (1) CRU NCEP v7 gridded climate data (L. 271), (2) VPM GPP products (L. 354–355).

1.5. The software used for data preparation, model calibration, data analysis, and result visualization is omitted and should be mentioned at the end of section 2.

2. Precision

2.1. The study implements the SOS model by Caffarra et al. (2011) in LPJ-GUESS. This model is called D**OR**MPHOT and not D**RO**MPHOT, which must be corrected throughout the manuscript (e.g., L. 87, 207, 213, etc.).

2.2. I feel that the authors sometimes used ‘developed’ and ‘constructed’ where ‘implemented’, ‘adopted’, ‘extended’, ‘improved’, etc. would be more appropriate. For example, did this study really develop/construct SOS and EOS models (L. 20–21 and 479–480) and LPJ-GUESS (L. 24)? Because all these models were taken from the literature, the EOS and SOS were probably rather ‘implemented’ and LPJ-GUESS was rather ‘improved’.

2.3. Lines 393–395 and 492 mention both ‘water stress’ and ‘legacy effects’, which must be defined in the Methods section. Moreover, the statement made in lines L. 393–395 seems not justified by any results.

2.4. I have difficulties with the conclusion that “LPJ-GUESS using the modified phenological module substantially improved […] (the) accuracy of spring and autumn phenology compared to […] (when using) the original phenological module” (L. 483–485). In my opinion, the study rather shows that the timing of SOS and EOS was simulated more accurately by the new versus current LPM implemented in LPJ-GUESS, which may partly be because of the module formulation. However, in contrast to the currently implemented phenology module, the new LPM was specifically calibrated (see above). The study cannot untangle the effect of the specific calibration from the effect of the module formulation. In consequence, the results do not allow to conclude on the isolated strength of either one of these effects. I strongly urge the authors to specifically calibrate both the new and current LPM with an identical calibration sample before comparing their accuracy based on an identical validation sample.

3. Readability and understanding

3.1. The grammar and syntax needs serous improvement to increase readability and understanding of the manuscript. Examples are: «Vegetation phenological shifts impact [...], and affect~~s~~» (L. 14), «we developed and coupled ~~the~~ spring and autumn phenology models into [...] LPJ-GUESS»(L 20–21), «These process-based phenology models **are** driven by temperature and photoperiod, and are parameterized for deciduous trees and shrubs **by** using remote sens~~-ing-bas~~ed phenological ~~observations~~ and reanalys~~is~~**ed climate dataset** ~~ERA5 land~~» (L. 21–24), and «the simulated RMSE for deciduous tree**s** and shrubs» (L. 26–26).

3.2. The manuscript contains many long sentences (e.g., L. 61–71, 101–106, 213–216, 271–278, and 443–446), which arguably can only be understood with additional effort. To increase the readability of the manuscript, I suggest to shorten most of the long sentences, for example by splitting the sentences.

4. Discussion

4.1. I do not understand the relevance of section 4.1 (L. 406–436) for this manuscript. Models to simulate SOS and ESO have been calibrated with remote sensed data before (e.g., Keenan & Richardson, 2015; White et al., 1997). Moreover, in my opinion, because the study does not assess the accuracy of vegetation indices derived from remote sensed observations, this procedure does not need to be discussed here.

4.2. The second paragraph of section 4.1 (L. 420 – 436) does not contain any references to the literature. In addition, I felt that this paragraph is rather an opinion than a discussion of results. Please refer to your results and corresponding literature or consider omitting the paragraph.

4.3. The way advancing spring phenology is discussed now, it appears that an advancement always results in an advantage for the concerned species at high elevations (L. 443–446). I doubt that this is true. Many studies have shown that earlier spring phenology also relates with an increased risk in damaged tissue and shoots due to late frost and the weight of late snow fall (e.g., Augspurger, 2009; Bigler & Bugmann, 2018; Drepper et al., 2022). This aspect of advancing spring phenology should be mentioned in the discussion.

Minor comments

1. To my understanding, particle swarm optimization was used to calibrate the newly implemented phenology models DORMPHOT and DM (L. 257–258). Therefore, the yellow box 'Particle swarm optimization' in Figure 2 (L. 209) should probably come after each of the yellow boxes 'DORMPHOT' and 'DM' rather than between the grey boxes for 'SOS'/'EOS' and 'GLC 2000'.
2. In my opinion, the result regarding the leaf area index (L. 318–322) is unrelated to the results regarding GPP, FPC, and AET, and therefore irrelevant for this study. I suggest omitting it.
3. While the references for Figures 4 and 5 in the text match the figure captions, the actual figures are mixed up.
4. Some results are already being discussed in the Result section (e.g., L. 392 – 395). Please move all discussion the Discussion section.

Refernces

Augspurger, C. K. (2009). Spring 2007 warmth and frost: Phenology, damage and refoliation in a temperate deciduous forest. *Functional Ecology*, *23*(6), 1031–1039. https://doi.org/10.1111/j.1365-2435.2009.01587.x

Bigler, C., & Bugmann, H. (2018). Climate-induced shifts in leaf unfolding and frost risk of European trees and shrubs. *Sci Rep*, *8*(1), 9865. https://doi.org/10.1038/s41598-018-27893-1

Caffarra, A., Donnelly, A., & Chuine, I. (2011). Modelling the timing of Betula pubescens budburst. II. Integrating complex effects of photoperiod into process-based models. *Climate Research*, *46*(2), 159–170. https://doi.org/10.3354/cr00983

Delpierre, N., Dufrene, E., Soudani, K., Ulrich, E., Cecchini, S., Boe, J., & Francois, C. (2009). Modelling interannual and spatial variability of leaf senescence for three deciduous tree species in France. *Agricultural and Forest Meteorology*, *149*(6–7), 938–948. https://doi.org/10.1016/j.agrformet.2008.11.014

Drepper, B., Gobin, A., & Van Orshoven, J. (2022). Spatio-temporal assessment of frost risks during the flowering of pear trees in Belgium for 1971–2068. *Agricultural and Forest Meteorology*, *315*, 108822. https://doi.org/10.1016/j.agrformet.2022.108822

Keenan, T. F., & Richardson, A. D. (2015). The timing of autumn senescence is affected by the timing of spring phenology: Implications for predictive models. *Glob Chang Biol*, *21*(7), 2634–2641. https://doi.org/10.1111/gcb.12890

White, M. A., Thornton, P. E., & Running, S. W. (1997). A continental phenology model for monitoring vegetation responses to interannual climatic variability. *Global Biogeochemical Cycles*, *11*(2), 217–234. https://doi.org/10.1029/97gb00330