Reviewer 1

Comment on "Plant phenology evaluation of CRESCENDO land surface models. Part II: Trough, peak, and amplitude of growing season" by Daniele Peano et al., Biogeosciences

This article described using global remote sensing based LAI products, including GIMMS-LAI3g, Copernicus Global Land Service LAI version 2 and MODIS MOD15A2H collection 6, to evaluate the simulated peak, trough and duration of LAI (or maybe growing season?) from the historical transient simulation of 7 different LSMs, following TRENDY S3 protocol, during 2000 ~ 2011. The outcome of this work is solid and has important values for improving parameterizations related to phenology and leaf dynamics for terrestrial ecosystems and land surface models (LSMs). But I have several major concerns on the way authors present their work (comment #1 and #2) and the lack of discussion about advancements/limitations of phenology parameterizations in different LSMs (comment #3). Thus, I cannot recommend the paper for publishing before authors can make a major revision to address my following comments.

We thank the reviewer for reading the manuscript and providing their comments.

Major:

1. One of my major concerns is that the purpose of this work is not clear.

It seems like authors want to compare modeled phenology against observation through LAI as an indicator, as suggested by the title "Part II: Trough, peak, and amplitude of growing season". But another paper from authors (Peano et al., 2021) already studied the start and end of the growing season and this "Part II" paper has relatively few interesting conclusions. Moreover, the use of NDVI instead of LAI as indicator for trough, peak, amplitude of growing season has at least 2 advantages: 1) RS-based LAI products are derived fully/partially from NDVI through complicated algorithms, which has more uncertainties and 2) metrics describing growing season, e.g., the onset/offset dates are (or can be) model outputs and can be directly compared to the metrics diagnosed from NDVI annual trace.

I was also thinking another possible purpose of this manuscript is trying to compare the value of modelled LAI amplitude vs. observation directly, which not only contains phenology but also productivity information, but I found no discussion on productivity side.

The third possible purpose is that LAI is directly linked to model parameterizations, thus authors want to utilize LAI as an indicator for diagnosing issues of LSM parameterization. If this is (one of) the focus(es) of this draft, I shall recommend authors to explore how these direct comparisons against LAI can be used to diagnose issues in model parameterizations. We see that all 7 different models show at least 3 different types of phenology and LAI parameterizations from Table 1. You can discuss how your findings can help us understand the spatial and temporal advantage/disadvantage of different phenology parameterization?

In summary, I suggest authors clarify my doubt on the purpose of this paper and provide persuasive reasons on why LAI is a good choice once the purpose is clarified.

We thank the reviewer for raising this concern and we are grateful for the opportunity to clarify the core purpose and details of this work.

The purpose of this study and paper is to evaluate the ability of state-of-the-art land surface models (LSMs), utilised in the EU CRESCENDO project (Coordinated Research in Earth Systems and Climate CRESCENDO - European Commission), to capture the broad-scale timing of specific vegetation phenological events, specifically the timing and value of vegetation peaks and troughs, and the seasonal amplitude, utilizing monitored data from three different satellite-derived Leaf Area Index (LAI) products. The results discussed in this manuscript expand on results from the 'Part 1' companion paper (Peano et al., 2021), which focused on evaluating the ability of the same LSMs to capture the start, end and length of the vegetation growing season using the same satellite LAI products.

We agree with the reviewer that NDVI may indeed be closer to the actually instrument-measured reflectance than the LAI products that take information from NDVI knowledge (or EVI, or other indices). However, LAI products outperform NDVI in dense vegetation. NDVI exhibits well-documented saturation when LAI exceeds 2 m^2/m^2, making it insensitive to further increases in canopy density (Tian Z. et al. 2025; Gao et al. 2023). This saturation effect may result in a weaker relationship between NDVI and vegetation dynamics in high-biomass ecosystems such as forests. In contrast, LAI products already partly correct for the saturation effect of VI using empirical or biophysical modelling or machine learning methods with observed LAI data as input and account for plant biophysical and soil background effects on radiative transfer to enable more accurate LAI retrieval (see Zhu et al. 2013, and Cao et al. 2023 for GIMMS LAI4g product). Thus, using LAI products can largely avoid the limitation of VIs and provide more meaningful information to validate LSM simulated LAI across the full range of vegetation densities. Moreover, LSMs simulate LAI as a prognostic biophysical variable to scale up from leaf-level processes to canopy and wider scales, which makes direct comparison with satellite LAI products more appropriate than using NDVI as a proxy. This avoids introducing additional uncertainty from NDVI-to-LAI conversion relationships. Note that the NDVI-LAI relationship itself is significantly affected by vegetation phenology (Tian X. et al. 2025). For these reasons, LAI was selected as the base data for both of these studies. It is definitely of interest how NDVI-derived phenology (and also FAPAR-derived phenology) compares with the LAI-derived metric used here, and we hope to do this in future work, but it is beyond the scope of this project. Some initial work to compare NDVI with LAI was done under the CRESCENDO project, and this would be a good start for further comparisons in the future.
To clarify this point, we add this text in the introduction:

" […]or evaluating the vegetation response to / influence on the ongoing climate change (e.g. Forzieri et al., 2017; Li et al., 2022a). Satellite-derived LAI data are usually products of biophysical modelling or machine learning methods that relate satellite-derived vegetation indices with ground LAI measurements (Zhu et al. 2013). These methods specifically account for the effects of vegetation types and structural characteristics on radiative transfer. Thus, the resulting LAI products effectively mitigate the saturation problem of vegetation indices directly

constructed from satellite reflectance, such as the Normalised Difference Vegetation Index (NDVI), especially for forest areas with dense canopies (Zeng, Y. et al. 2023, Gao et al. 2023; Cao et al., 2023; Tian Z. et al. 2025). This advantage makes satellite-derived LAI more suitable than NDVI for evaluating model-simulated LAI, particularly when assessing interannual and seasonal vegetation dynamics under climate change. Despite the utility […]"

This study also aims to highlight regions or biomes where LSMs correctly and incorrectly capture both the timing and value of the LAI-derived phenology metrics, which help identify specific reasons why the models are finding it difficult to reflect the satellite observations in these areas, and help provide suggestions for developments to the model parameterizations of phenology.

A specific analysis of each parameterization is limited in this study since it would require outputs at the Plant Functional Type (PFT) level and a mapping between PFTs and phenology scheme as done by Li et al. (2024). This type of output has been requested from modelers for the Coupled Model Intercomparison Project phase 7 (CMIP7) simulations (Li et al., 2025). The availability of that information will provide the possibility to improve our knowledge of the limitations and abilities of each phenology scheme.

To clarify the purpose of this study, we propose to modify the title as follows:

Plant phenology evaluation of CRESCENDO land surface models using satellite-derived Leaf Area Index. Part II: Seasonal trough, peak, and amplitude.

We also propose to rephrase the paragraph on the purpose and aims of the study in the introduction section (lines 57-60) as follows:

"[...] Consequently, the present study follows on and complements the earlier study (Peano et al., 2021) by performing a compound assessment of the amount (amplitude) and time (peak and trough) of leaf production in the same set of LSMs models and satellite-based products. The evaluation of these three variables enriches our understanding of the abilities and limitations of state-of-the-art LSMs gained in the previous study (Peano et al., 2021)."

Finally, we propose to add in the revised manuscript a discussion on the need for PFT level outputs in the discussion section 4.3 as follows:

"[...] Finally, a detailed comparison between phenology parameterizations requires data at the PFT level and a mapping between PFTs and phenology schemes as done by Li et al. (2024a) and requested in the next Coupled Model Intercomparison Project phase (CMIP7, Li et al., 2025). The availability of that information will provide the possibility to improve our knowledge of the limitations and abilities of each phenology scheme."

2. The analysis is performed at a spatial resolution of half by half degree, so I assume most of the grids are covered by mixed plant functional types. It makes sense to use LAI peak, trough and amplitude for the analysis over mixed forest dominated by deciduous or seasonal forest/grassland/cropland since they have relative clear seasonality both from model

parameterization and vegetation spectral signal, but may exaggerate the bias of model parameterization when compared to grids dominated by evergreen forest at high latitudes, where both some of the model parameterizations do ignore the LAI variability, and the vegetation spectral signal is contaminated by snow cover. In fact this is one example of my major concern #1, that using LAI trough, peak and amplitude might not be appropriate for describing growing season and phenology. Good examples to examine land surface phenology on a global scale are the work of Buitenwerf et al., 2015 and the 4GTS method from authors' Part I paper, which combined quantitative metrics with mode detection. I would strongly recommend authors to either justify findings over these grids, or at least explain this point as limitation in their work.

We agree with the reviewer that the grid resolution limits the model evaluation, especially in areas covered by mixed plant functional types. As mentioned by the reviewer, this may influence the results obtained in regions dominated by evergreen forests, such as at high latitudes. As we mentioned in the answer to the reviewer's point 1, above, for this work we did not have availability of data at the PFT level, but we hope this will be available for the CMIP7 simulations and this should improve such evaluation. Also, higher resolution satellite products becoming available will also help improve the scale at which these analyses can be conducted.

For this reason, we discuss this point in the revised manuscript in section 4.3 as follows:

"[...] vegetation parameterisation, crop and plant functional type population, soil characterisation, and initial spatial resolution, as already noted in Peano et al. (2021).

In particular, the discrepancies in model grid resolution and a relatively coarse initial spatial resolution (between about 2° and 0.5°, Table 1) induce differences in the simulated grid vegetation mixture, which may explain the mismatch between LSMs, especially in regions characterised by high biodiversity and areas with evergreen forests. The availability of data at the PFT level would reduce the resolution impact and refine the investigation of differences between LSMs as requested for the next phase of the Coupled Model Intercomparison Project (CMIP7, Li et al., 2025).

In general, the results of this study highlight [...]"

3. Direct comparison of models against observation can provide large amounts but only fragmented pieces of information, which is hard for deriving robust conclusions (e.g., Line 192 - 196). Since authors know the different types of parameterization in models, I'm thinking if authors can link model phenology parameterizations with your model-data comparison. One suggestion from me is that authors can group models with similar parameterization and check if models with different parameterizations can have statistically significant difference in three metrics you used through Student's t-test? You can also think about per-biome or per-zonal analysis if necessary? This can improve the value of the work and suggest the direction for improving the current phenology scheme in these LSMs?

We agree with the reviewer that the global comparison between LSMs and satellite products provides partial information. For this reason, we performed the analysis at the biome scale described in Section 3.3. Based on the results of this section, we will further discuss the

implications for each LSM and their parameterisation by adding these paragraphs in the revised version of the manuscript to address this issue:

"[…] In general, the growing seasons simulated by the LSM show delays in their peaks compared to the satellite estimates, especially in the northern hemisphere. Moreover, LSMs sharing similar phenology parameterisation schemes, such as CLM4.5, CLM5.0, and LPJ-GUESS (Table 1), display discrepancies in phenophases estimates, such as in the southern hemisphere BET biome, where CLM5.0 differs from CLM4.5 and LPJ-GUESS by approximately 6 months (Figures S5i and S6i). This behaviour highlights the influence of models' features beyond the specific phenology schemes in representing the growing season cycle.

[…]

Several LSMs represent LAI values based on the values of specific leaf area and the amount of leaf carbon or biomass content (i.e. CLM4.5, CLM5.0, LPJ-GUESS, and ISBA-CTRIP, Table 1). The implementation of similar parameterisation reflects on reduced differences between LSMs (Figure S7), except for southern hemisphere BDS-dominated areas (Figure S7k), where LPJ-GUESS substantially overestimates the LAI seasonal amplitude compared to CLM4.5, CLM5.0 and ISBA-CTRIP. On the other hand, LSMs primarily driven by temperature, such as JULES-ES and ORCHIDEE, tend to underestimate the LAI seasonal amplitude (Figure S7), which is not the case when also leaf features are considered, as done in JSBACH (Table 1). This comparison, then, underscores the need to incorporate leaf features and leaf carbon content in LAI computation within LSMs."

Finally, a direct evaluation of models sharing similar parameterisation is also discussed in Section 4.3, where the results obtained from two versions of the Community Land Model (i.e. CLM4.5 and CLM5.0) are compared. CLM4.5 and CLM5.0 only slightly differ in phenology parameterisations, showing reduced differences in the timing of peak and trough. On the contrary, the two LSMs differ in the model structure, such as the representation of soil, plant hydrology, and carbon and nitrogen cycles. These differences result in a higher discrepancy in simulated LAI quantity (i.e. LAI amplitude).

This comparison highlights the complexity in attributing the model skill to a specific parameterisation, as already reported in the manuscript. Detailed evaluation of specific parameterisation, then, requires sensitivity studies and evaluation at the PFT level, which are beyond the scope of the present study, but we hope to do these in future work.

Minor:
Did you perform new simulations or analyzed the existing TRENDY model outputs? If you used TRENDY model products, which version did you use? Please clarify.

The outputs evaluated here are from simulations done for the EU CRESCENDO project which followed the TRENDY protocol for simulation S3 by applying the same set of CO_2, Land-use, and climatic forcings. To clarify this point, we rephrase the last sentence of the Introduction as follow:

"[…]to evaluate the CRESCENDO LSMs output when forced with varying atmospheric CO_2 concentrations, climate and land-use changes employed in the international "Trends and drivers of the regional-scale sources and sinks of carbon dioxide" (TRENDY, in particular experiment S3, https://blogs.exeter.ac.uk/trendy/protocol/, last access: 18 November 2024) project (Sitch et al., 2015; Zhao et al., 2016)."

Authors need to clarify whether crop phenology is activated or not in the model simulations, since the TRENDY S3 simulation includes the dynamic land use change and cropland. If so, please add crop phenology schemes used by different models in Table 1.

Unlike the TRENDY simulations, the CRESCENDO ones do not require active crop schemes. Some, but not all, of the LSMs implement separate classifications of crops, as reported in Table 1 and described in the LSMs presentation in Section 2.2 and described in the LSMs' presentation in Section 2.2.

Line 59: "the seasonal timing of trough and peak, and amplitude (trough to peak) of LAI." The authors used "growing season" as title but here used "LAI" instead. It is not a good idea to use both in an exchangeable way since growing season and LAI are different. Suggest to revise and keep consistency.

We agree with the reviewer that the LAI and growing season are not synonymous. We think, for this particular study, that the main confusion arises from the title of the manuscript. As noted in previous responses above, we hope to study and compare different base data (e.g. NDVI, Phenocam data) and other indicators of seasonality in future work. For this reason, we propose to change the title as follows:

Plant phenology evaluation of CRESCENDO land surface models using satellite-derived Leaf Area Index. Part II: Seasonal trough, peak, and amplitude.

Line 88: "Seven LSM" shall be "Seven LSMs"

We thank the reviewer and correct accordingly.

Line 87: since you only have one paragraph describing each LSM, you can remove the subsection titles and merge them into one paragraph.

We thank the reviewer, in the revised version of the manuscript, we will aggregate each model description within Section 2.2 as suggested.

Line 141 and Line 305: "the same spatial land coverage evolution forces them leaving only differences in plant growth and seasonality among them" is not precise. Since TRENDY models have different PFT definitions, the use of the same land use history as forcing only guarantees the consistent total forest/crop/pasture/urban areas across different models. So the land coverage

evolutions for each PFT is not the same across LSMs, and only the relative fraction of total forest land on each grid cell is the same in different models.

We agree with the reviewer that the vegetated areas remain coherent among LSMs. To better expose this point, we rephrase these lines as follows:

"Despite each LSM implementing the LUH2 data differently (e.g. different numbers of PFTs), the same vegetated areas evolution forces them to leave differences in plant growth, biodiversity, and seasonality among them."

And

"Nonetheless, the implementation of a common land-use dataset allows all LSM to reproduce the same vegetated areas evolution, leaving only differences in plant growth, biodiversity, and seasonality among them."

Line 218: How did you calculate the agreement of LAI seasonal amplitude? Since amplitude is a numerical, not a categorical variable? I see that you defined +-0.25 as the threshold (correct me if I'm wrong) to define "agreement" of LAI amplitude. This shall be explained in methodology.

Following the reviewer's suggestion, we will add a description of the agreement evaluation in Section 2.4 as follows:

"[...] Results from the land surface models are also aggregated and evaluated as a multi-model ensemble mean (MME). Finally, the agreement between LSMs and satellite products refers to differences of 0 months in peak and trough (i.e. both LSM and satellite product produce peak and trough occurring in the same month) and of 0.25 m^2/m^2 in LAI amplitude."

Figure 1b: The pattern of LAI trough from CGLS in boreal regions near the arctic circle, such as central Siberia and part of northern Europe, is very similar to the error maps (1d, 1f). This makes me doubt the credibility of CGLS products over these regions, and I wonder what reason causes this similarity? Authors listed several caveats for using RS products in section 4.4, so I feel this can be a good example and suggest authors to check the reason behind and add into this section.

The high-latitude areas present various issues for satellite products due to technical limitations, such as data reconstruction and gap-filling, and environmental conditions, such as cloud and snow coverage and polar nights, as mentioned in Section 4.4. In particular, the CGLS product uses climatology values to extend northward the values in the winter season, which is when troughs occur. On the contrary, MODIS and LAI3g stop at lower latitudes. Consequently, the differences in Figures 1d and 1f above 50ºN are driven by the dissimilar treatment used to gap-fill those areas in the winter season. Following the reviewer's suggestion, we will add an explicit reference to this point in Section 3.1.1 as follows:

"[...] than CGLS or MODIS (LAI3g root mean square error of 2.6 months, Table 2). The differences in boreal regions derive from discrepancies in the gap-filling approaches applied to high-latitude winter values between satellite products, which is a relevant limitation of satellite products, as discussed in Section 4.4. The longest differences in the timing of troughs [...]"

The biome specific results (section 3.3) authors presented are derived from which biome mask? Authors shall display the difference between PFT coverage from models and the biome mask used for analysis as supplemental material to consolidate their conclusions.

The biome mask derives from the ESA CCI Land Cover map, as reported in Figure 8. We will add this information in Section 2.1, which will be titled "Satellite products", as follows:

"[...] Note that the following sections also provide results from the comparison between the three satellite products.

Finally, the land cover distribution from ESA CCI (Li et al., 2018) is used to derive a common Plant Functional Type (PFT) mask to evaluate the differences among LSMs and satellite LAI products at the biome scale."

Finally, we agree with the reviewer that a comparison between the ESA CCI PFT mask and each LSM PFT mask would help in further investigating the source of the difference among models. However, because the model PFT distributions were not among the shared CRESCENDO variables it was not possible for this study. We hope to include such analyses with future work, as noted in previous review responses.

References

Cao, S., Li, M., Zhu, Z., Wang, Z., Zha, J., Zhao, W., Duanmu, Z., Chen, J., Zheng, Y., Chen, Y., Myneni, R. B., and Piao, S.: Spatiotemporally consistent global dataset of the GIMMS leaf area index (GIMMS LAI4g) from 1982 to 2020, Earth Syst. Sci. Data, 15, 4877–4899, https://doi.org/10.5194/essd-15-4877-2023, 2023.

Si Gao, Run Zhong, Kai Yan, Xuanlong Ma, Xinkun Chen, Jiabin Pu, Sicong Gao, Jianbo Qi, Gaofei Yin, Ranga B. Myneni. Evaluating the saturation effect of vegetation indices in forests using 3D radiative transfer simulations and satellite observations.Remote Sensing of Environment,295,113665(2023) https://doi.org/10.1016/j.rse.2023.113665.

Li, W., MacBean, N., Ciais, P., Defourny, P., Lamarche, C., Bontemps, S., Houghton, R. A., and Peng, S.: Gross and net land cover changes in the main plant functional types derived from the annual ESA CCI land cover maps (1992–2015), Earth System Science Data, 10, 219–234, https://doi.org/10.5194/essd-10-219-2018, 2018.

Li, X., Ault, T., Richardson, A. D., Frolking, S., Herrera, D. A., Friedl, M. A., Carrillo, C. M., and Evans, C. P.: Northern hemisphere land-atmosphere feedback from prescribed plant phenology in CESM, Journal of Climate, https://doi.org/10.1175/jcli-d-23-0179.1, 2024a.

Li, Y., Tang, G., O'Rourke, E., Minallah, S., e Braga, M. M., Nowicki, S., Smith, R. S., Lawrence, D. M., Hurtt, G. C., Peano, D., Meyer, G., Hassler, B., Mao, J., Xue, Y., and Juckes, M.: CMIP7 Data Request: Land and Land Ice Priorities and Opportunities, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2025-3207, 2025.

Peano, D., Hemming, D., Materia, S., Delire, C., Fan, Y., Joetzjer, E., Lee, H., Nabel, J. E. M. S., Park, T., Peylin, P., Wårlind, D., Wiltshire, A., and Zaehle, S.: Plant phenology evaluation of CRESCENDO land surface models – Part 1: Start and end of the growing season, Biogeosciences, 18, 2405–2428, https://doi.org/10.5194/bg-18-2405-2021, 2021.

Xianchao Tian, Xingyu Jia, Yizhuo Da, Jingyi Liu, Wenyan Ge. Evaluating the sensitivity of vegetation indices to leaf area index variability at individual tree level using multispectral drone acquisitions. Agricultural and Forest Meteorology, 364,110441 (2025) https://doi.org/10.1016/j.agrformet.2025.110441.

Zeng, Y., Hao, D., Park, T. et al. Structural complexity biases vegetation greenness measures. Nat Ecol Evol 7, 1790–1798 (2023). https://doi.org/10.1038/s41559-023-02187-6

Zezhong Tian, Jiahao Fan, Tong Yu, Natalia de Leon, Shawn M. Kaeppler, Zhou Zhang. Mitigating NDVI saturation in imagery of dense and healthy vegetation. ISPRS Journal of Photogrammetry and Remote Sensing,227,234-250 (2025) https://www.sciencedirect.com/science/article/pii/S0924271625002394

Reviewer 2

The authors did a good job of addressing the reviewers' comments in the revised version of this paper.

We thank the reviewers for their comments and suggestions which have helped to improve the manuscript.

However, the following sentence is difficult to understand: 'On the contrary, the heterogeneity of phenology schemes may improve the ability to capture the correct timings'. Could you please clarify this?
We have rephrased this sentence, as follows:

"*On the contrary, the variety of phenology schemes may improve the ability to capture the correct timings, as done by JSBACH, which distinguishes up to six phenology schemes (Table 1), in peak timings (Figure 2).*"

For the sake of clarity, finally, Table 2 should distinguish between observational datasets and model simulations. The best score values among the model simulations should be written in bold.

Following the reviewer's suggestion, we have changed Table 2 in the revised version of the manuscript, as follows:

Table 2. Root mean square error (in month and m^2/m^2) between CGLS and the other satellite products (first two rows of the table) and land surface models (last eight rows of the table) and the percentage of the region in agreement (green areas in Figures 1, 2, 3, 5, 6) with the CGLS values. Note that the best score values among LSMs are bold.

	LAI Peak Time		LAI Trough Time		Seasonal Amplitude	
	RMSE [months]	Agreement [%]	RMSE [months]	Agreement [%]	RMSE [m^2/m^2]	Agreement [%]
MODIS	1.5	65.5	1.7	54.8	0.7	48.5
LAI3g	1.7	54.3	2.6	26.0	0.7	43.9
MME	2.4	21.3	3.2	11.8	1.4	13.0
CLM4.5	2.9	15.1	3.6	12.3	1.5	28.5
CLM5.0	3.0	10.1	3.9	9.6	1.7	22.9
JULES	3.2	3.4	4.2	2.5	1.4	26.2
JSBACH	**2.2**	19.3	3.5	6.8	1.2	21.3
LPJ-GUESS	2.3	**24.4**	2.2	**42.3**	1.3	17.2
ORCHIDEE	3.3	4.5	3.4	9.9	1.4	22.0
ISBA-CTRIP	2.4	12.9	**2.1**	37.5	**1.0**	**34.7**