DeepPhenoMem V1.0: Deep learning modelling of canopy greenness dynamics accounting for multi-variate meteorological memory effects on vegetation phenology

Guohua Liu, Mirco Migliavacca, Christian Reimers, Basil Kraft, Markus Reichstein, Andrew D. Richardson, Lisa Wingate, Nicolas Delpierre, Hui Yang, Alexander J. Winkler

Response to Reviewers

Reviewer #1

Based on in situ observation data of plant phenology at 50 sites distributed across Northern American, the authors trained a deep learning model LSTM which has the potential to capture and model the meteorological memory effects on vegetation phenology. The topic of this study is very important and interesting. It provides a new pathway to simulate and investigate the complex impacts of environmental factors on plant canopy dynamics. In addition, the manuscript is overall well organized, and the methods used in this study have been introduced detailedly.

Author Response (AR): We thank the reviewer for the thoughtful evaluation of our study and for the valuable feedback. Below, we address each comment provided by the reviewer in our responses point by point.

Nonetheless, I still have some questions on the method and results of this study:

• The DeepPhenoMem model in this study is trained at 45 sites, and evaluated at 5 sites. Some random factors (e.g. the specific climate and species in the 4 validation sites compared to the training sites) might strongly affect the evaluation results. To give a more robust evaluation on the model, I would suggest the authors to do a 10-fold (or 5-fold) cross validation.

AR: Thank you for your valuable suggestion. We recognize the importance of robust model validation and appreciate your recommendation for cross-validation to ensure the generalizability of our DeepPhenoMem model.

In response to your concern, we would like to clarify that during the development of DeepPhenoMem, we indeed implemented a cross-validation strategy to assess the model's performance accurately. Specifically, we employed a leave-one-site-out cross-validation approach, which varies in folds depending on the PFT: 25-fold for the DB, 12-fold for the EN, and 8-fold for the GR. This method was integral to our process for identifying the most effective model architectures and hyper-parameters, ensuring that our model is robust across different sites.

Furthermore, beyond the leave-one-site-out cross-validation, we also designated an additional test set, entirely excluded from the model development phase. This test set, drawn from 5 separate sites, was not involved in any part of the training or validation processes, thereby eliminating the possibility of influencing the model's development.

We acknowledge the importance of the point raised in your review and realize that our initial explanation might not have been sufficiently detailed. In our revised manuscript, we will elaborate on our validation strategy.

• There are 3 unseen sites for deciduous broadleaved forests were used to test the trained model in this study. Why not show the evaluation results at all of these 3 sites in Figs. 5 & 6. The readers might wonder that only the site with best model performance for PFT DB was showed in Figs. 5 & 6. The evaluation results might thus be biased.

AR: Thank you for your feedback. In our study, we chose to present the results from the site with the longest time-series in Figs 5 and 6. This decision was made to demonstrate the model's ability to capture changing trends over time. However, we acknowledge your concern regarding the potential for biased interpretation of the model's performance. To address this, we will include the evaluation results from the two additional unseen sites in the supplementary materials.

• The deciduous broadleaved forests (DB) generally show stronger seasonal variability compared to evergreen needle-leaved (EN) forests. I am wondering why the DeepPhenoMem model performs better for EN, compared to DB (Table 1, Fig. 3). If it is only because the BD has been tested at three sites, while the EN was only tested in 1 sites? In theory, the observed and simulated GCC for DB and grassland could be low to 0. Why the GCC for DB and grassland are all higher than 0.3. Are there any evergreen plants living in the DB and grassland sites. In addition, the EN keeps being green across the year, is it accurate enough to extracted the GCC from the digital images photographed by automated and high-frequency digital cameras. Even in the end of growing season, the GCC for EN should still be high.

AR: Thank you for your insightful questions, which have prompted us to refine our explanations. Here is the summary response to them.

1. **Better performance of DeepPhenoMem Model in EN compared to DB**

   The superior performance in estimating GCC for the EN compared to the DB might be attributed to two main factors:

   1) Spring and autumn variability in GCC: The EN exhibits more gradual changes in GCC during spring and autumn, facilitating more accurate model simulations of GCC. This contrasts with the DB, where abrupt environmental changes lead to more volatile GCC values.

   2) Memory effect: The EN may have a longer memory effect, meaning its GCC values are influenced by past conditions over a longer period. In contrast, the GCC for DB respond more immediately to current environmental factors.

2. **GCC values**

   We would like to clarify that the statement "GCC for DB and grassland could be as low as 0" is untrue. GCC is calculated as the relative intensity of green compared to the total intensity from red, green, and blue (RGB) channels. A GCC value of 0 would imply the absence of green light, which is not representative of any natural vegetation state. In reality, a balanced mix of RGB, resulting in a

"grey" appearance, corresponds to a GCC value of approximately 0.33 (below). Therefore, a GCC of 0 (below), indicating the presence of only red, only blue, or a mix of red and blue with no green, does not accurately describe the appearance of barren or dormant vegetation.

GCC ≈ 0.33, corresponding to "grey":

RGB = 100, 100, 100

RGB = 50, 50, 50

RGB = 200, 200, 200

GCC = 0:

RGB = 100, 0, 0

RGB = 100, 0, 100

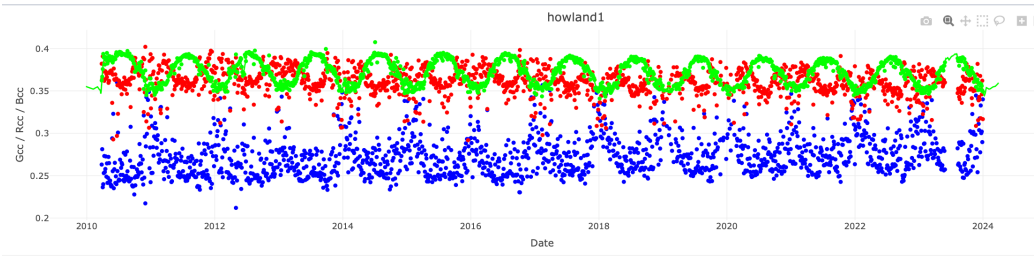
RGB = 0, 0, 100

3. **Seasonal changes of GCC in EN forest in Phenocam data**

While PhenoCam data are not perfect, the ability of GCC and other derived indices to parallel EN canopy changes in canopy photosynthetic capacity, pigment ratios, and other vegetation indices, suggests that these data are sufficiently "accurate enough to extracted the GCC from the digital images photographed by automated and high-frequency digital cameras". See Seyednasrollah et al. 2019 for more information.

The PhenoCam dataset, despite its limitations, effectively captures changes in canopy photosynthetic capacity, pigment ratios, and other vegetation indices. This suggests that GCC and similar indices derived from high-frequency, automated digital camera images are reliable for tracking vegetation phenology across diverse biomes. For more detailed information, we refer to the study by Seyednasrollah et al. (2019), which supports the accuracy of PhenoCam data in monitoring vegetation changes.

The seasonal changes in pigment ratios in EN forests result in a canopy that is much less green in the winter than it is in the summer, in terms of GCC. As shown in the plot below, when GCC decreases during the winter season (data for Howland AmeriFlux site), the relative "blueness" and "redness" of the canopy are increasing, reflecting photoprotective pigment changes (more carotenoids, change in xanthophylls). As demonstrated in our response to the question about L126 (see Reviewer #2), human eyes appear to be much more sensitive to the overall intensity of a particular (e.g., green) color channel, than the relative intensity of that color.



• Based on the observed data, the authors can actually calculate the sensitivities of SOS and EOS to warming using linear regression (e.g. Fu et al., 2015, Nature). I am wondering if the temperature sensitivities simulated from the trained DeepPhenoMem model in this study are comparable to the values calculated based on observations. I would suggest the authors to do a comparison/evaluation.

AR: Thank you for your suggestion. We did not compare the sensitivities of SOS/EOS to warming with Fu et al. 2015, because we also consider that 1) Fu et al. 2015 is based on leaf unfolding data, our study used SOS/EOS from PhenoCam, which is a more integrated measure of land surface phenology. Given that PhenoCam data may cover multiple individuals within the region of interest, the signal may not be as distinct as that from leaf unfolding data; 2) Fu et al. 2015 used very long archives of data, whereas PhenoCam data, with the exception of few long term's sites, has shorter data record. The length of the period can affect the estimated SOS/EOS sensitivity to temperature, which can lead to spurious differences in temperature sensitivity of phenology (Keenan et al., 2020).

Specific comments:

L147-148: Is it accurate enough to use the sum of precipitation over the previous month as a proxy of daily SW? Is there any reference of Eq. 3. Maybe it is better to simply say the sum of precipitation over the previous month has been included in the model, without mentioning the SW.

AR: We agree with the reviewer that the use of the term "soil water available" might lead to misunderstanding. Given the unavailability of direct soil moisture data, we have used a 30-day backward running mean of precipitation, assigning decreasing weights (ranging from 0 to 1) to days further back in the preceding month, to serve as a proxy for soil moisture. This index has been demonstrated to serve effectively as a proxy for soil moisture where direct soil moisture measurements are lacking (Migliavacca et al. 2011). Therefore, we have adopted it as a feasible workaround for estimating soil moisture levels within our model. We acknowledge the necessity of elucidating this methodology in our revised manuscript.

L163-164: Why not conduct a cross-validation?

AR: We apologize for any confusion caused by our initial description in the methods section. To clarify, we indeed conducted a leave-one-site-out cross-validation, which was implemented as a 25-fold for DB, a 12-fold for EN, and an 8-fold for GR. This procedure is detailed in lines 178-179 of our manuscript. We will amend the methods section to ensure this is clearly communicated.

Table 1: Please add RMSE of each model for each vegetation type

AR: Thank you for your advice. We will include the RMSE values for each model and vegetation type in Table 1.

Fig. 4: What does the rhombus represent here? The significance of difference between two versions of the model? If no, please provide a significance test between results from the M0 and Mfull. The sub-plots a, b, c show results for DB, EN, GR, respectively?

AR: In Fig. 4, the rhombus represents the outliers, which are defined as the points beyond 1.5 times the interquartile range (the difference between the 75th and 25th percentiles). You are correct that the sub-plots labeled a, b, and c correspond to the results for DB, EN, and GR, respectively. We will ensure to include this information in the caption of Fig. 4 for clarity.

Regarding the significance of the difference between the results from models M_0 and M_{full}, we will conduct a significance test and provide the relevant statistical analysis. Thank you for bringing this to our attention.

Figs. 5 & 6: There are 3 unseen sites for DB (Fig. 1), why only results for harvardbarn2 was presented?

AR: We appreciate your observation. The results for the two other unseen sites for DB will indeed be included in the supplementary materials. Please also refer to our response to the second general comment for further clarification.

L376-378: I did not find what result in this study indicate the cumulative thermal summation, rather than daily temperature alone, determines vegetation phenology

AR: Our findings indicate that the full-memory-effect model (M_{full}), which accounts for the memory effect of temperature, outperforms the no-memory-effect model (M_0), which only considers the instantaneous effect of daily temperature. This suggests that temperature memory, or cumulative thermal summation, plays a crucial role in driving vegetation phenology, rather than relying solely on daily temperature. We will improve the formulation in the revised manuscript.

L406-411: Not fully true. The phenology module of Earth System models (ESMs) indeed only focuses on a few specific phenological events (e.g. start and end of the growing season). However, the ESMs also simulate the whole time series of canopy development/evolution across the whole growing season, by mechanically simulating the photosynthesis, autotrophic respiration, carbon allocation, etc. To have a closed mass balance of carbon, the canopy evolution has to be simulated mechanically, rather than using an empirical model or machine learning model.

AR: In many Earth system models, such as CLM 4.5 and LPJ (Peano et al. 2021), phenology is modeled as a function of climatic drivers only using PFT-specific thresholds for chilling, growing degree days, etc. Some models also more realistically connect to the carbon cycle, i.e., leaves grow at a carbon cost. For the first type of ESM, where phenology is derived only as a function of climate, our data-driven approach can be directly used as a substitution for the empirical formulations. For the second type of ESM, where phenology also is dependent on the available carbon resources, one would have to extend our data-driven approach to a hybrid approach (e.g., ElGhawi et al. 2023 for land-

atmosphere fluxes) where the carbon resources are also considered in the input and the loss function, i.e., leaf growth depletes carbon resources from the reserve pool and dropped leaves are added to the humus pool, satisfying carbon mass balance. The carbon mass balance then acts as a constraint on the data-driven phenology model.

Reviewer #2

General comments:

This study aims to train an LTSM model to simulate the temporal evolution of a measure of canopy greenness observed with in-situ repeat, digital photography across three plant functional types, testing performance against the final year of observations at the ~50 training sites, and across multiple years at several sites not used in training. This is an interesting proposal and, given the current limitations in long established phenology models, potentially very useful to a wide community of vegetation modelers. The primary conclusions that the LTSM model seems to capture some of the underlaying controls on phenology, that incorporating meteorological memory effects improves performance, and the models exhibits plausible relationships are sound.

However, the results as presented are difficult to interpret beyond this – the reliance on R2 statistics is potentially misleading considering the marked biases exhibited by the model. Incorporating RMSE or some other measure of bias throughout the results and discussion is required to better understand where and when the model performs well or otherwise. Contrasting and explaining (lack of) model performance across space/time/PFT would potentially be more informative than the current approach which tends towards endorsing model performance in very vague terms. At a minimum, this would include adding RMSE values to Table 1 and including in Figure 4 and moving Figure 9 from the Discussion to the Results and maybe incorporating more discussion of Figure S2, as well as adding additional interpretation and discussion of these results related to the bias.

AR: We appreciate your detailed review and valuable feedback. Recognizing the significance of including more comprehensive indices, such as RMSE, for enhanced evaluation and interpretation of model performance, we have revised Table 1 and Figure 4 to incorporate RMSE values. In the following sections, we will address each of your specific comments in detail.

Specific comments:

L126: Need some additional information that explains the limited and very similar dynamical range in GCC across the three PFTs. Why is there not more annual variation in deciduous v. evergreen trees in particular?

AR: Thanks for your question. Firstly, GCC ranges observed are as follows: 0.30-0.46 for Deciduous Broadleaf (DB), 0.32-0.43 for Evergreen Needleleaf (EN), and 0.30-0.43 for Grassland (GR). This data indicates a relatively higher variation in GCC for deciduous trees compared to evergreen trees, approximately 50% more.

However, it's important to note that "evergreen trees", particularly those in seasonally cold conifer forests, exhibit significant seasonal variation in canopy color. This variation is largely due to changes in photoprotective pigments, which cause the canopy to appear "more red" (indicating a lower GCC) in winter compared to summer. This phenomenon is supported by the work of Seyednasrollah et al. (2019) and further illustrated by Keenan et al. (2014), which demonstrates how canopy color in Deciduous Broadleaf Forests can be modeled as a nonlinear mixing model with two endmembers, one the color of the leafless canopy and the other the color of individual leaves.

In winter, the canopy of an EN forest is slightly "more green" compared to the grey branches of a DB forest. Conversely, in summer, the EN canopy is "less green" than the bright emerald green of new foliage in DB. Therefore, the dynamical range in GCC is not identical across these forest types.

Additionally, the GCC index does not necessarily correspond to what human eyes would perceive as the brightest green, but rather characterizes the relative brightness of green relative to other colors. For instance, an RGB signature of 0, 50, 0 results in a GCC of 1.0 (left panel), while an RGB signature of 50, 150, 50 has a GCC of 0.60 (right panel). To the human eye, the right panel probably looks "more green" due to its brighter intensity, despite having a lower GCC value. This discrepancy highlights the importance of considering both the quantitative GCC values and the qualitative aspects of color perception in analyzing canopy dynamics.

GCC = 1.0 **GCC = 0.6**

RGB = 0, 50, 0 **RGB = 50, 150, 50**

L153: It doesn't seem appropriate to call weighted mean monthly precipitation "soil water" – it isn't and should be renamed. Also, it was an unfortunate choice of variable to include when trying to tease out the difference between models with and without memory effects as by design it will capture an approximation of soil moisture memory that won't be removed by the shuffling in the M0 model and leads to a mixture of instantaneous and time integrating variables in the regression model. If this can be included, then why not some cumulative temperature term, or day of year etc. which are known to strongly influence phenology.

AR: We acknowledge the concern regarding the terminology used for weighted mean monthly precipitation and its designation as "soil water availability". To clarify, this index is calculated by determining the weighted mean precipitation, with the values decreasing to zero one month before the date of interest. Given its calculation method and application, we will rename the terminology in our revised manuscript. This index has been demonstrated to serve effectively as a proxy for soil moisture where direct soil moisture measurements are unavailable (Migliavacca et al. 2011).

In our paper, the inclusion of soil moisture's influence is essential. However, due to the lack of measured soil moisture data, we opted for this proxy to represent soil moisture

conditions. We agree that this proxy can have the memory effect, which is like the memory effect of soil moisture.

Furthermore, we acknowledge the recommendation to consider additional factors such as cumulative temperature, which also significantly influence phenology. In theory, the LSTM approach is capable of inherently assimilating this temperature-related memory information. Therefore, we directly incorporate temperature as an input.

L251: Figures 1, 4 and Figure S1 indicate there are three deciduous test sites? It's unclear which (maybe all?) are being shown in Figure 3

AR: You are correct that there are 3 unseen sites and the data in 2018 from each site as the test data for DB. In Fig. 3, we indeed present the results for all test data, including those from the three deciduous sites. We will clarify this point in the manuscript to avoid any confusion. Thank you for your feedback.

L270: Here, and elsewhere in several places in the manuscript, reference to figure numbers is incorrect. Please check all these carefully.

AR: Thanks for your kind reminder. We have renumbered the figures throughout the manuscript to ensure they are consecutive.

L276: There is evidently zero/minimal skill in simulating daily anomalies, whilst overall R2 values indicate some skill in seasonal variability/monthly times scales. Is attempting to simulate what are presumably noisy daily data a valuable test? Can some smoothing be applied to investigate if there is any model skill between daily and monthly time scales. Or is there only really skill in seasonal variability - making additional analysis of variation of a few days in SOS and EOS difficult to interpret?

AR: We acknowledge the challenges in simulating daily anomalies, as indicated by the minimal skill observed in our LSTM models. Despite the difficulty associated with modeling daily variability, our findings reveals that the full-memory-effect model (M_{full}) performs better in predicting daily anomalies compared to the no-memory-effect model (M_0). This finding indicats the significance of memory effects in enhancing the model's capability to simulate daily anomalies.

As you mentioned there might be noise in daily data, in this study we have applied a locally weighted scatterplot smoothing method to reduce noise in the daily data.

L280: Here is a clear example of the bias that needs to be quantified and examined more thoroughly throughout the whole analysis.

AR: We acknowledge that our models have limitations in simulating absolute GCC values accurately, as discussed in lines 415-428. However, it's important to note that the bias in absolute values is less significant compared to the seasonal dynamics of GCC which is used for detecting phenology.

Reference:

ElGhawi, R., Kraft, B., Reimers, C., Reichstein, M., Körner, M., Gentine, P., and Winkler, A. J.: Hybrid modeling of evapotranspiration: inferring stomatal and aerodynamic resistances using combined physics-based and machine learning, Environ. Res. Lett., 18, 034039, https://doi.org/10.1088/1748-9326/acbbe0, 2023.

Keenan, T. F., Darby, B., Felts, E., Sonnentag, O., Friedl, M. A., Hufkens, K., O'Keefe, J., Klosterman, S., Munger, J. W., Toomey, M., and Richardson, A. D.: Tracking forest phenology and seasonal physiology using digital repeat photography: a critical assessment, Ecological Applications, 24, 1478–1489, https://doi.org/10.1890/13-0652.1, 2014.

Keenan, T. F., Richardson, A. D., and Hufkens, K.: On quantifying the apparent temperature sensitivity of plant phenology, New Phytologist, 225, 1033–1040, https://doi.org/10.1111/nph.16114, 2020.

Migliavacca, M., Reichstein, M., Richardson, A. D., Colombo, R., Sutton, M. A., Lasslop, G., Tomelleri, E., Wohlfahrt, G., Carvalhais, N., Cescatti, A., Mahecha, M. D., Montagnani, L., Papale, D., Zaehle, S., Arain, A., Arneth, A., Black, T. A., Carrara, A., Dore, S., Gianelle, D., Helfter, C., Hollinger, D., Kutsch, W. L., Lafleur, P. M., Nouvellon, Y., Rebmann, C., Da ROCHA, H. R., Rodeghiero, M., Roupsard, O., Sebastià, M.-T., Seufert, G., Soussana, J.-F., and Van Der MOLEN, M. K.: Semiempirical modeling of abiotic and biotic factors controlling ecosystem respiration across eddy covariance sites, Global Change Biology, 17, 390–409, https://doi.org/10.1111/j.1365-2486.2010.02243.x, 2011.

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.

Seyednasrollah, B., Young, A. M., Hufkens, K., Milliman, T., Friedl, M. A., Frolking, S., and Richardson, A. D.: Tracking vegetation phenology across diverse biomes using Version 2.0 of the PhenoCam Dataset, Sci Data, 6, 222, https://doi.org/10.1038/s41597-019-0229-9, 2019.