Response to Reviewer #1:

[Comment 1] This data paper describes a new vegetation phenology dataset that fuses four existing remotely sensed datasets. It then compares the timing of two phenophases (start and end of the growing season) from this new dataset to the same metrics from three phenology camera networks (phenocams). Considering the discrepancies in accuracy and temporal/spatial coverage between existing phenology datasets, this new fusion dataset, created based on weighed averaging, is extremely useful and appears to have higher accuracy than existing datasets. However, the methods could be explained in greater detail, in addition to a few other concerns, which I highlight below.

Response: We are grateful to the reviewer for recognizing the extremely useful and importance of our research and for the constructive comments. We also thank the reviewer for these thoughtful and constructive comments and suggestions, which have substantially improved our manuscript. We appreciate the reviewer's assistance in the acknowledgments section of the revised manuscript. We have addressed each point and adjusted the manuscript according to the reviewer's comments. Please find below a detailed response to each comment.

Major concerns:

[Comment 2] First, I'm concerned about how the SOS and EOS dates were extracted and compared across the different datasets. Of the 4 datasets fused together, at least several appear to use different SOS and EOS thresholds and methods to extract the dates, which begs the questions – are they directly comparable? For example, if one dataset identifies SOS as 15% of maximum green-up, but another uses 50%, the extracted SOS date will naturally differ between those datasets, but one is not necessarily more or less accurate than the other. The same applies when comparing the fused dataset with the phenocam dataset – how are SOS and EOS identified across the 3 phenocam networks included in the phenocam dataset? Also, in addition to different thresholds, it appears that the datasets use different methods to extract SOS and EOS dates. This could be one reason that there is observed variability across the SOS and EOS dates. If possible, it would be best to standardize the methods and threshold across all the datasets. It might also be helpful to compare entire seasonal trends in vegetation greenness through time to visualize differences between datasets instead of just the seasonal transition dates.

Response: We thank the reviewer for his/her thoughtful comment. We totally agree that different thresholds were used in different extraction methods and may induct large variation in phenological dates, as previous studies reported that there are differences in vegetation phenology results obtained from various remote sensing phenology algorithms (time series data processing methods and phenology extraction methods) (Cong et al, 2012; Wu et al, 2021; Zeng, 2020).

Remote sensing vegetation phenology typically reflects transition dates in the vegetation growth cycle, such as the start (leaf out) and end (leaf senescence) of the growing season, and different vegetation indices, e.g. NDVI, LAI and SIF, were used. The phenological dates that were extracted from different methods were supposed to

indicate changes in actual physiological conditions as accurately as possible. However, the effectiveness of these methods varies across regions and time period, and may not always represent the true vegetation conditions. Such as, different vegetation phenology datasets show different performance across regions and years comparing to the phenocam dates, see the Fig. S1 below. The consistency of VIP and ground phenocam data in the year 2001 of deciduous broadleaf forest is the best, whereas only the consistency of GIM_3g data in the year 2002 is better than that of VIP data. Comparing to the forest types, the consistency of remote sensing based phenological dates and phenocam data is higher in deciduous broadleaf region when using GIM_3g method, but in evergreen needleleaf when using the MCD method. We want to get the results that best reflect the physiological state at different sites and years. Therefore, a method that integrates data from different methods based on reliability to combine the advantages of different data sources is feasible. In the revised manuscript, we clarify this argument, please refer to line 375-385 in the revised manuscript.

In PhenoCam, Spline interpolation method was applied to extract transition dates for each ROI mask in PhenoCam Dataset v2.0. We used the date when the GCC first (last) crosses 25% of the GCC amplitude as the SOS and EOS. The second phenocam dataset is from the Japan Internet Nature Information System digital camera data (http://www.sizenken.biodic.go.jp/) acquired over the period 2002 – 2009. Ide and Oguma (Ide and Oguma, 2010) provided greenup dates for two phenocam sites with areas of interest (AOI) defined at the species level scale. The vegetation types included in their data comprise wetland and mixed deciduous forest. The date of green-up each year was estimated as the DOY of the maximum rate of increasing 2G-RBi (i.e., the maximum of the second derivative. The third dataset consists of phenology data for deciduous broadleaf forests in Japan (Inoue et al., 2014), supported by the Phenological Eyes Network (http://www.pheno-eye.org/), which is a network of ground-based observatories for long-term automatic observation of vegetation dynamics established in 2003 (Nasahara and Nagai, 2015), the start and end of season is defined as the first day when 20% of leaves had flushed and the first day when 80% of leaves had fallen in the given ROI, respectively.

We thank the reviewer for recommending to standardize the methods and threshold across all the datasets, whereas it is difficult to standardize the various methods and thresholds, because standardizing thresholds requires the acquisition of a complete vegetation index time series. This approach may diminish the significance of data fusion, i.e. the selection of different thresholds and smoothing methods for the data source that best reflect the changes in the physiological state of the vegetation.

Following the reviewer's suggestion, we estimated the entire seasonal trends in vegetation greenness, e.g. the NDVI trends, for VIP, GIM_4g and REA method, please see the results in Fig. S2 below. The average greening trend in VIP ($-10.16 \times 10^{-4}/yr$) is lower than REA ($-1.14 \times 10^{-4}/yr$) and GIMMS ($3.55 \times 10^{-4}/yr$). The greening rate in VIP, GIM_4g and REA are 25.22% (with 41.44% significantly greening), 68.49% (with 38.32% significant greening) and 49.83% (with 56.83% significant greening), respectively. Large difference in these greenness trends were found, supporting our findings and an integrated method is thus needed to produce a dataset that combing the advantages of different data sources. In the revised manuscript,

we have added this argument in the discussion section, please refer to line 435-439.

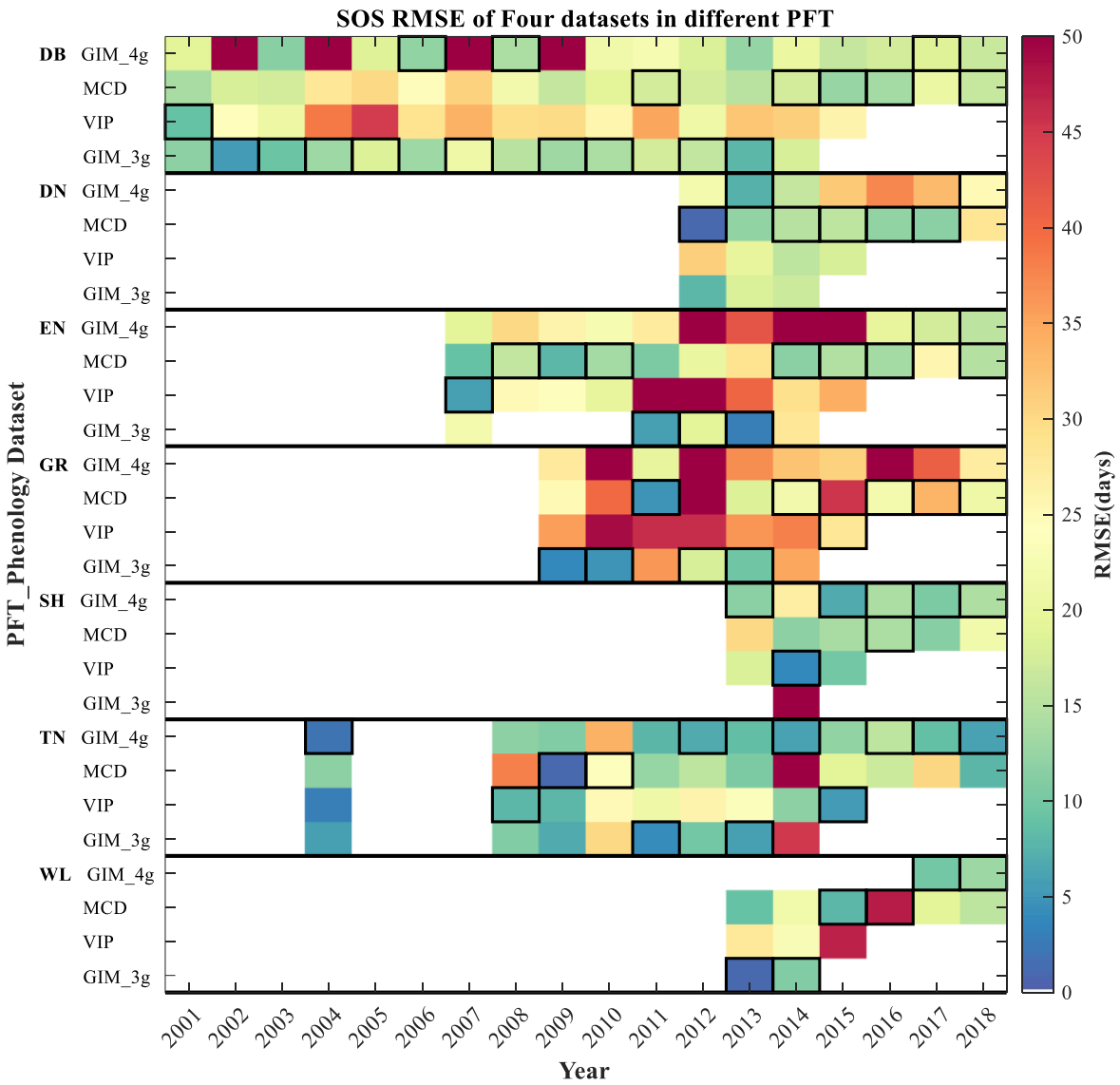


Figure S1: SOS RMSE of four datasets in different PFT for the period 2001 – 2018. Four datasets refer to GIM_4g, MCD12Q2, VIP, and GIM_3g datasets, respectively. PFT: plant functional type, DB: deciduous broadleaf, DN: deciduous needleleaf, EN: evergreen needleleaf, GR: grassland, SH: shrubs, TN: tundra, WT: wetland. The black boxes represent the best data for the year in that PFT.

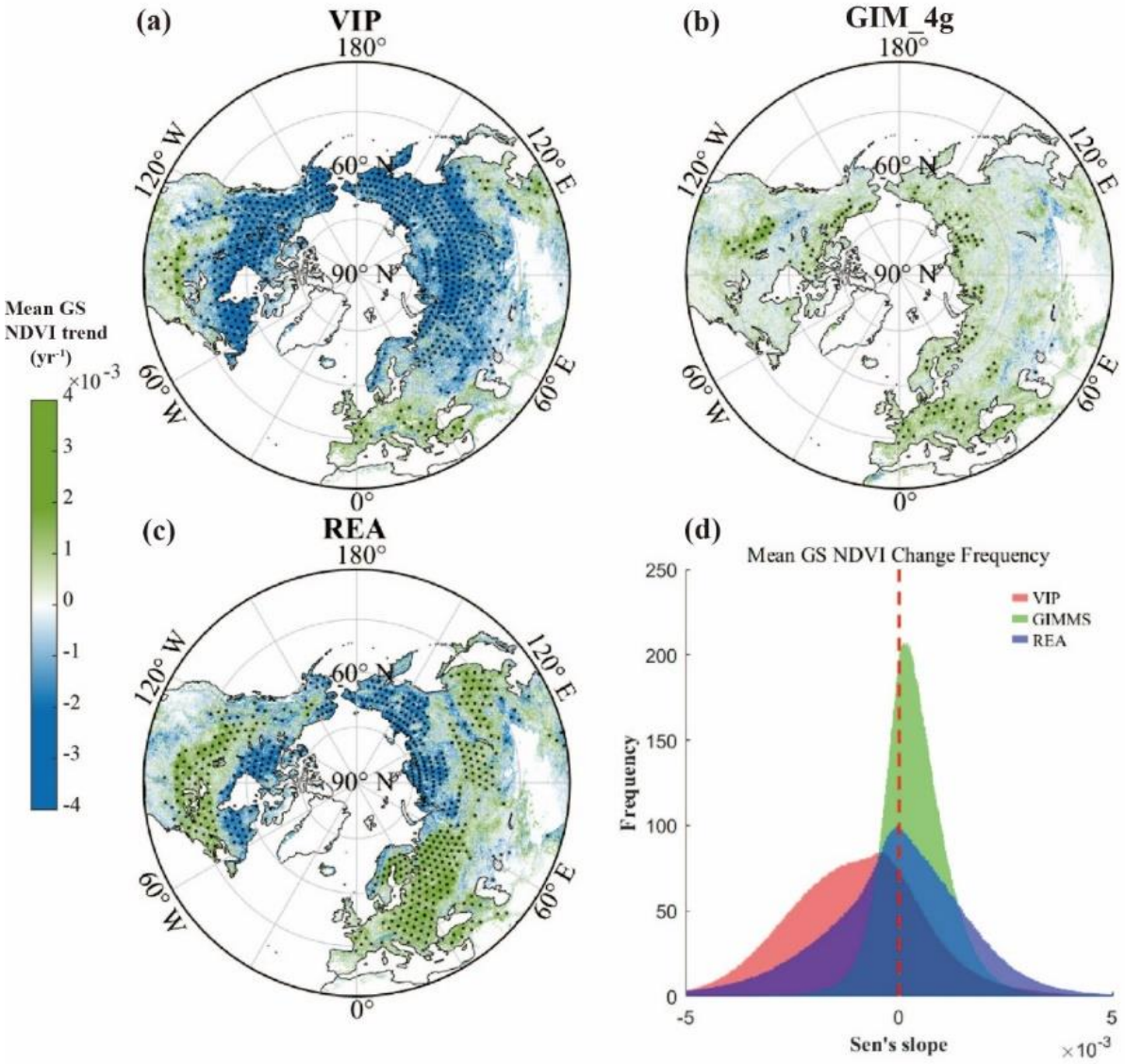


Figure S2: Growing season NDVI trend obtained using (a)VIP phenology dataset, (b) GIM_4g dataset and (c) REA phenology dataset for the period 1982 – 2015 and (d) the distribution of their frequency. GS: growing season. Black dots represent for regions where the trend is significant (P<0.05).

Cong, N., Piao, S., Chen, A., Wang, X., Lin, X., Chen, S., Han, S., Zhou, G., and Zhang, X.: Spring vegetation green-up date in China inferred from SPOT NDVI data: A multiple model analysis, ***Agricultural and Forest Meteorology***, 165, 104 – 113, https://doi.org/10.1016/j.agrformet.2012.06.009, 2012.

Wu, W., Sun, Y., Xiao, K., and Xin, Q.: Development of a global annual land surface phenology dataset for 1982–2018 from the AVHRR data by implementing multiple phenology retrieving methods, ***International Journal of Applied Earth Observation and Geoinformation***, 103, 102487, https://doi.org/10.1016/j.jag.2021.102487, 2021.

Zeng, L., Wardlow, B. D., Xiang, D., Hu, S., and Li, D.: A review of vegetation phenological metrics extraction using time-series, multispectral satellite data, ***Remote Sensing of Environment***, 237, 111511, https://doi.org/10.1016/j.rse.2019.111511, 2020.

[Comment 3] Also, this is mostly semantics, but an important distinction to not mislead readers. "PhenoCam" with a capital P and C is usually used to indicate data from the PhenoCam Network (phenocam.nau.edu) and phenocam (lowercase p and c) indicates a generic camera from any network. Please see Richardson 2023, Box 3 for further explanation (https://www.sciencedirect.com/science/article/pii/S0168192323004410). Since you used digital camera data from 3 different sources, please use "phenocam" in this paper to avoid confusion.

Response: We thank the reviewer for pointing out this mistake, and have corrected the use of "PhenoCam" and "phenocam" in the revised manuscript.

[Comment 4] Finally, I know discussion sections are normally shorter in data papers, but it seems like the authors could connect their results more to other studies. For example, how does the rate of growing season advance/delay compare to the rate that other studies have found? The fact that this analysis is done makes this more than just a data paper, and requires more contextualization with the literature. I would suggest either removing that analyses or adding more sources to support your results.

Response: Following the reviewer's suggestion, we have added more details in the discussion about the rate of phenological changes.

Both advanced spring phenology and delayed autumn phenology were found between REA-based phenological dates and previous studies. But the amplitudes of trends are different among these studies. In details, SOS was found significantly advance at the rate of 0.19 days per year in the REA result during 1982-2020, while the advancing rate of 1.4 ± 0.6 days per decade during 1982-2011 (Wang et al., 2015) and 5.4 days advanced from 1982 to 2008 (Jeong et al., 2011) was also found in previous studies. Similarly, EOS was found significantly delayed at the rate of 0.18 days per year in the REA result over the same period, while the 0.18 ± 0.38 days per year delay was found for 1982-2011 (Liu et al., 2016) and the 6.6 days delay was found from 1982 to 2008 (Jeong et al., 2011) in previous studies. In the revised manuscript, we added these statements in the discussion section in line 426-433.

Piao, S., Liu, Q., Chen, A., Janssens, I. A., Fu, Y., Dai, J., Liu, L., Lian, X. U., Shen, M., and Zhu, X.: Plant phenology and global climate change: Current progresses and challenges, ***Global change biology***, 25, 1922–1940, https://doi.org/10.1111/gcb.14619, 2019.

Wang, X., Piao, S., Xu, X., Ciais, P., MacBean, N., Myneni, R. B., and Li, L.: Has the advancing onset of spring vegetation green-up slowed down or changed abruptly over the last three decades?, ***Global Ecology and Biogeography***, 24, 621 – 631, https://doi.org/10.1111/geb.12289, 2015.

Jeong, S.-J., HO, C.-H., GIM, H.-J., and Brown, M. E.: Phenology shifts at start vs. end of growing season in temperate vegetation over the Northern Hemisphere for the period 1982–2008, ***Global change biology***, 17, 2385–2399, https://doi.org/10.1111/j.1365-2486.2011.02397.x, 2011.

Line edits:

[Comment 5]: Here and elsewhere, I would suggest saying "a ground-based phenocam data" instead of "the ground-based…"- using "the" makes it sounds like you're

referring to a single data source, (such at the PhenoCam Network), but your phenocam dataset actually includes multiple data sources. Also, see my comment for line 109 about using "PhenoCam" versus "phenocam"

Response: Thank you for your comment. Following the reviewer's suggestion, we have revised "the ground-based" as "a ground-based phenocam data" through the revised manuscript.

[Comment 6]: largest correlation and accuracy in comparison with what?

Response: The REA-based phenological dates were compared with the phenocam dates. We updated the statement as "*The start of growing season and the end of growing season in the newly merged dataset had the largest correlation (0.84 and 0.71, respectively with phenocam data)*", please refer to line 17-18 in the revised manuscript.

[Comment 7]: root mean square error between what? (Observed and predicted dates?)

Response: We have completed the statement here as "*accuracy in terms of the root mean square error (12 and 17 d, respectively between phenocam data and merged datasets)*". Please refer to line 18-19 in the revised manuscript.

[Comment 8]: When giving the start and end of the growing season trends, what region are you referring to – the entire globe?

Response: Our study focuses on the region above 30° North latitude. In the revised manuscript, we clarify the region and modified the statement as "*The new dataset has a spatial resolution of 0.05 ° and covers the period from 1982 to 2020, with geographic coverage extending above 30 degrees North in the Northern Hemisphere.*". Please refer to line 14-15 in the revised manuscript.

[Comment 8]: They still are used – I suggest changing to "are commonly".

Response: the terms were updated following the reviewer's suggestion, please refer to line 31.

[Comment 9]: What region was assessed to see trends in SOS and EOS in these studies? The entire globe?

Response: Thank you for your comment. The study region is in the northern hemisphere. Please refer to line 42-45 in the revised manuscript.

[Comment 10]: "merits and demerits" is awkward. Perhaps replace with "advantages and disadvantages"

Response: Following the reviewer's suggestion, we have replaced " merits and demerits" with "advantages and disadvantages" in the revised manuscript, please refer to line 45-46.

[Comment 11]: Perhaps set up/explain the merged dataset a little more here. From this sentence, my initial thought was "Why would a merged dataset be better if the raw datasets that go into it are inaccurate?" I understand it better after the next paragraph when you explain REA, but it's unclear here.

Response: We appreciate the reviewer's suggestions for improving our text. We have modified our text as "Because it is difficult to determine the optimal dataset from the various phenology datasets, producing a merged dataset using method which can choose the best dataset in different time and space among all input datasets is therefore essential for providing a comprehensive and accurate estimation of vegetation phenology with high spatiotemporal resolution." Please refer to line 51-54 in the revised manuscript.

[Comment 12]: change "was" to "is"

Response: Following the reviewer's suggestion, we have change "was" to "is", please refer to line 55.

[Comment 13]: What is the "vegetation index method"? Please explain in the text.

Response: Sorry for the confusing argument. The vegetation index method refers to the utilization of indices such as the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI) to assess and extract the status of vegetation. We have modified our text as "*Alternatively, methods such as weighted functions, the Bayesian approach, and mixed models have been combined with the vegetation index method, which used the mathematical formulas to assess vegetation conditions to integrate datasets with high temporal and spatial resolutions*". Please refer to line 58-60 in the revised manuscript.

[Comment 14]: Please write out all satellite/dataset abbreviations (MODIS, VIP, GIM) the first time (e.g., MODIS = Moderate Resolution Imaging Spectroradiometer)

Response: Thank you for your comment. Following the reviewer's suggestion, the full names of abbreviations were provided when it was first used, and the abbreviation was used when it was mentioned again in the following text. Please refer to line 76-85 in the revised manuscript.

[Comment 15]: I don't think you need the "Note" under the table. The dataset names and abbreviations are clear in the table.

Response: We removed the 'Note' section following the reviewer's suggestion.

[Comment 16]: Write out and define NDVI the first time it's used.

Response: Following the reviewer's suggestion, we have added the full name of NDVI the first time it is used. Please refer to line 50 in the revised manuscript.

[Comment 17]: What is the "threshold method"? You use this term multiple times, so it would be good to explain it in more detail the first time.

Response: We have added an explanation of the threshold method the first time it is mentioned in the text. The threshold method defines the growing state of the vegetation as the time when the vegetation index reaches a certain percentage of the annual amplitude and reflect a specific vegetation physiological growth stage. Please refer to line 91-93 in the revised manuscript.

[Comment 18]: What does "segment EVI2 amplitude" mean? When the time series reaches 15% of the maximum seasonal amplitude? Please add more detail about these methods.

Response: The definition of greenup and dormancy is officially given by the USGS (https://lpdaac.usgs.gov/products/mcd12q2v061/). "Segment EVI2 amplitude" refers to the range of variation in the Enhanced Vegetation Index 2 (EVI2) over a specific growing season segment. It is calculated as the difference between the maximum and minimum values of the EVI2 time series within that segment. To clarify this statement, in the revised manuscript, we added detail information as "*This amplitude is calculated as the difference between the maximum and minimum values of the EVI2 time series within the growing season.*". Please refer to line 94-97 in the revised manuscript.

[Comment 19]: What threshold was used to determine SOS and EOS? If the datasets use different thresholds, this could be one reason they differ in their SOS and EOS dates.
Response: Thank you for your comment. The start (end) of season is defined as the date when the NDVI2 time series first (last) crosses 35% of the segment NDVI2 amplitude. The 35% threshold using for the NDVI2 was found to be more accurate especially in regions with a protracted slow emerging growing season (see Didan, K., Barreto, A., 2016). Please refer to line 103-106 in the revised manuscript. Indeed, different methods with different thresholds result large variation in phenological dates, and thus a reliable integrated method is needed, please see our response to the similar comment#2 for detailed information.

[Comment 20]: This dataset uses a different threshold (20%) than the MCD12Q2 dataset (15%). As I mentioned above, this could contribute to their differences in accuracy. Also, please explain what "segment NDVI amplitude" means.
Response: Please refer to our response to the similar comment#2 and comment#19.

[Comment 21]: Why is SOS 20% and EOS 50%? Usually, the same threshold is used for both the start and end of season
Response: Different algorithm of phenology extraction can cause differences in vegetation phenology results. In their study (Chen and Fu, 2024), SOS 20% and EOS 50% were used as the best reflection of the state of surface vegetation.

Chen, S., Fu, Y. H., Li, M., Jia, Z., Cui, Y., and Tang, J.: A new temperature–photoperiod coupled phenology module in LPJ GUESS model v4. 1: optimizing estimation of terrestrial carbon and water processes [data set], ***Geoscientific Model Development***, 17, 2509–2523, https://doi.org/10.5194/gmd-17-2509-2024, 2024.

The dataset for this study: Chen, S. and Fu, Y.: Vegetation phenology data based on GIMMS4g NDVI from 1982 to 2020, https://doi.org/10.5281/zenodo.11136967, 2024.

[Comment 22]: PhenoCam with capital C is used to indicate data from the PhenoCam Network (phenocam.nau.edu) and phenocam (lowercase p and c) indicates a generic camera from any network, which I would suggest using here to avoid confusion. Please see Richardson 2023, Box 3 for further explanation
(https://www.sciencedirect.com/science/article/abs/pii/S0168192323004410).
Response: Following the reviewer's suggestion, we used 'phenocam' instead of 'PhenoCam' through the revised manuscript.

[Comment 23]: Please add source of data (PhenoCam Network) "includes data

acquired from the *PhenoCam Network*…."

Response: The sources of data were provided in the revised manuscript, please refer to line 127-132 in the revised manuscript.

[Comment 24]: The PhenoCam Network does not use fisheye cameras – they use standard camera lens. They also aren't completely downward-facing- they are tilted slightly downwards, but always include the horizon. See Fig 5: https://phenocam.nau.edu/pdf/PhenoCam_Install_Instructions.pdf

Response: Thank you for your correction. We have corrected our text as "*the PhenoCam Dataset v2.0 (Richardson et al., 2018b; Seyednasrollah et al., 2019a, b), includes data derived from conventional visible-wavelength automated digital camera imagery through PhenoCam Network (Richardson et al., 2018a) over the period 2000–2018 and across 393 sites in various ecosystems, for detailed information, please refer to https://daac.ornl.gov/VEGETATION/guides/PhenoCam_V2.html and https://phenocam.nau.edu/webcam/* " please refer to line 127-132 in the revised manuscript.

[Comment 25]: include full link to dataset (https://daac.ornl.gov/VEGETATION/guides/PhenoCam_V2.html). Please also cite PhenoCam data paper associated with this dataset:

> Seyednasrollah, B., A.M. Young, K. Hufkens, T. Milliman, M.A. Friedl, S. Frolking, and A.D. Richardson. 2019. Tracking vegetation phenology across diverse biomes using PhenoCam imagery: The PhenoCam Dataset v2.0. Manuscript submitted to Scientific Data. https://www.nature.com/articles/s41597-019-0229-9

Response: Following the reviewer's suggestion, we provided the linkage and cited the paper in the revised manuscript, please refer to line 127-132 in the revised manuscript.

[Comment 26]: What geographic area is included in these datasets?

Response: These datasets are located in Japan, and the inclusion of this portion of data is intended to supplement the validation data for Asia. Please refer to line 137 and 141 in the revised manuscript.

[Comment 27]: Is this dataset also collected by digital cameras?

Response: Yes, it is collected by digital cameras.

[Comment 28]: How were these 280 sites selected (for example, the PhenoCam Network has 393 sites alone)? How many sites are from each of the 3 networks?

Response: We use phenocam data with roi_id (a numeric code to distinguish between multiple ROIs of the same vegetation type at a given site) equals to 1000, and delete sites which only have one direction of transition dates, 26 sites was deleted in this step. And then we remove sites with no phenology values in all four data sources (90 sites), therefore 277 sites left remain out of 393, then we added 3 sites in Japan to supplement Asian area. In the revised manuscript, we clarify the selection procedure, please refer to line 145-147.

[Comment 29]: How are SOS and EOS extracted from all the phenocam datasets? In general, the methods of data collection and extraction from all 3 phenocam datasets could be explained in more detail.

Response: Following the reviewer's suggestion, in the revised manuscript, we provided

more details of phenological dates extraction methods as: "*Spline interpolation method was applied to PhenoCam data to extract transition dates for each ROI mask in PhenoCam Dataset v2.0. We used the date when the GCC first (last) crosses 25% of the GCC amplitude as the SOS and EOS. The second phenocam dataset is from the Japan Internet Nature Information System digital camera data (http://www.sizenken.biodic.go.jp/) acquired over the period 2002–2009. Ide and Oguma (Ide and Oguma, 2010) provided greenup dates for two phenocam sites with areas of interest (AOI) defined at the species level scale. The vegetation types included in their data comprise wetland and mixed deciduous forest. The date of green-up each year was estimated as the DOY of the maximum rate of increasing 2G-RBi (i.e., the maximum of the second derivative. The third dataset consists of phenology data for deciduous broadleaf forests in Japan (Inoue et al., 2014), supported by the Phenological Eyes Network (http://www.pheno-eye.org/), which is a network of ground-based observatories for long-term automatic observation of vegetation dynamics established in 2003 (Nasahara and Nagai, 2015), the start and end of season is defined as the first day when 20% of leaves had flushed and the first day when 80% of leaves had fallen in the given ROI, respectively.*"(line 134-145).

[Comment 30]: How was interannual variability used to assign weights to each dataset? Are datasets considered more or less accurate with higher interannual variability? Why?

Response: Thank you for your comment. Interannual variability is measured by ε_{Phe} in equation (2), which is also represents for natural variability. Natural variability changes from region to region, in Equation (1) and (6), ε_{Phe} cancels out under the condition of $B_{Phe,i}$ and $D_{Phe,i}$ greater than ε_{Phe},which based on the assumption that more stringent on are required to increase the reliability over regions characterized by lower natural variability. The natural variability does not work single, it works with $B_{Phe,i}$ and $D_{Phe,i}$ jointly. For the region in lower natural variability, if the phenology data from one dataset also have large difference with other datasets, it is given lower weight for generate the REA phenology at that region, which is thought to be less accurate data at that region. In the revised manuscript, we updated the description to clarify this issue, please refer to line 189-194 and 201-204.

Giorgi, F. and Mearns, L. O.: Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the "Reliability Ensemble Averaging" (REA) Method, ***Journal of Climate***, 15, 1141–1158, https://doi.org/10.1175/1520-0442(2002)015<1141:COAURA>2.0.CO;2, 2002.

Lu, J., Wang, G., Chen, T., Li, S., Hagan, D. F. T., Kattel, G., Peng, J., Jiang, T., and Su, B.: A harmonized global land evaporation dataset from model-based products covering 1980–2017, ***Earth System Science Data***, 13, 5879–5898, https://doi.org/10.5194/essd-13-5879-2021, 2021.

[Comment 31]: How exactly were the datasets compared to get a value of "consistency" and "offset"? Please explain these methods in more detail.

Response: The consistency is measured by $B_{Phe,i}$,which is defined by is defined as the difference between the input dataset and the mean value of the four datasets. The offset is measured by the $D_{Phe,i}$, which is measured by the difference between the REA result and each input dataset. They are calculated in iterations. We used consistency and offset to summarize the process of $B_{Phe,i}$ and $D_{Phe,i}$. Following the reviewer's suggestion, we

revised this description, please refer to line 158-160 in the revised manuscript.

[Comment 32]: How are the methods described in section 2.2 and 2.2.1 different? Isn't 2.1 describing the REA method? (Also missing a 2.2.2 section in the paper).
Response: We are sorry for missing serial number. We have moved the contents of part 2.2 to 2.2.1 and correct the serial number in the following section.

[Comment 33]: What is the "voting principle"?
Response: The result of REA is determined by the interannual variability of each phenology dataset, together with the degree of consistency and offset among the four phenology datasets. Voting principle means that the REA result in one region is depend on all four datasets. Each dataset is given different weight to generate the result, which is determined by the differences among four datasets and its natural variability, and the result will closer to where most of data is. To clarify this issue, we revised the text, please refer to line 166-169 in the revised manuscript.

[Comment 34]: What does BIAS stand for? What is it measuring? Similarly, what is the difference between the RMSE and unbiased RMSE (For both, I don't mean mathematically, but rather contextually in terms of understanding the data).
Response: RMSE is calculated as the square root of the average of the squares of the residuals, which penalizes larger errors than smaller ones and provide an estimate of the magnitude of errors between remote sensing estimated value and phenocam datasets. BIAS is the average difference between remote sensing estimated value and phenocam value, that helps in understanding whether the estimated value is higher or lower than phenocam value. The correlation coefficient measures the linear relationship between two variables. The ubRMSE measures the deviation between two variables without systematic errors. Standard deviation quantifies the variation of the dataset, which measures the deviation between data and the mean value. Please refer to line 217-223 in the revised manuscript.

[Comment 35]: Were all analyses only for above 30 degrees N? If so, that should be stated in the methods.
Response: Thank you for your suggestion, we have stated the spatial range of the dataset in method section, e.g. 2.1 Phenology dataset part. Please refer to line 84-85 in the revised manuscript.

[Comment 36]: Remove the "etc" (not clear what it refers to – just list more places if desired).
Response: We have removed the "etc".

[Comment 37]: Include references for this claim.
Response: Following the reviewer's suggestion, the relevant references were added, please refer to line 304-306 in the revised manuscript.

[Comment 38]: Which site did you chose? From which phenocam dataset? Change "American" to "US".
Response: We chose Morganmonroe site from phenocam in figure 6, and in the revised manuscript we moved it to Fig. S4 since it is just a one site example. Following the

reviewer's suggestion, we have changed "American" to "US.".

[Comment 39]: Remind readers that an "advance" in SOS means earlier dates. The terms "advance" and "delay" can sometimes be confusing to follow.

Response: Thank you for your suggestion, we have added explanation in 3.4 to remind readers "advance" means earlier dates, and "delay" means later dates in the revised manuscript, please refer to line 356 and 360.

[Comment 40]: Some regions also show a significant delay in SOS (Fig 7b) – it would be good to point that out, too.

Response: Following the reviewer's suggestion, a description of the proportion of significant delays has been added in Fig.6 and corresponding text in the revised manuscript, please refer to line 358.

[Comment 41]: Line 329: change "were" to "are"- they're still widely used.

Response: We have changed our text to correct the Syntax.

[Comment 42]: What does the "complexity of surface backgrounds" mean?

Response: Complexity of surface background refers to the intricacy and variability of the land cover type in the surface, which would bring challenges in the process of extracting physiological characteristics of vegetation.

[Comment 43]: Please explain what the "mixed-pixel effect" is.

Response: The mixed-pixel effect refers to the phenomenon where a single pixel in a remote sensing image encompasses various vegetation types. This can lead to significant discrepancies in phenological dates when comparing datasets with lower spatial resolution to those with higher resolution, as the latter can capture finer details and more accurately represent the true vegetation conditions. To clarify this issue, we revised the text in the revised manuscript, please refer to line 389-391.

[Comment 44]: What does the "complexity of surface backgrounds" mean?

Response: Please see the response to comment#42 for the same question.

[Comment 45]: What does the "process of coefficient determination" mean?

Response: It refers that algorithm performance are related with their coefficients which may differs in different place, and the process of coefficient may influence the results. We revised the text to clarify this issue, please refer to line 403-406 in the revised manuscript.

[Comment 46]: Please further explain what the non-linear change in endmembers is and why that would result in poor performance. This is the first time it is mentioned in the paper.

Response: The nonlinear mixing effect of endmembers refers to the phenomenon in remote sensing imagery where the spectral signals of different land covers mix spatially in a nonlinear manner, causing the spectral response of a single pixel to no longer be a simple linear combination of the endmember spectra. The non-linear mixing of spectra caused the changes in remote sensing data reflectance, and then increase the uncertainty of vegetation indexes calculation. To clarify this issue, we revised the corresponding

text, please refer to line 407-410 in the revised manuscript.

[Comment 47]: Please redefine REA acronym the first time it's used in the discussion section.
Response: We have redefined the REA acronym, please refer to line 397 in the revised manuscript.

[Comment 48]: What is the "voting principle"?
Response: Please see the response to comment#33 for the same question.

[Comment 49]: It's not clear what "high processing efficiency" means – how does that relate to the low RMSE of the REA dataset?
Response: It means that the REA method has higher processing efficiency comparing with common data fusion method. We have removed "high processing efficiency", since it is not related to the low RMSE of the REA dataset for the revised manuscript.

[Comment 50]: How does this rate compare to other studies? Include citations.
Response: Please see the response to comment#4 for the same question.

[Comment 51]: I suggest replacing "invaluable" with "accurate" or "reliable" (or something similar)
Response: Following the reviewer's suggestion, we have changed the "invaluable" to "reliable".

[Comment 52]: Consider starting a new sentence here.
Response: Following the reviewer's suggestion, we changed the text as "*The mean uncertainty range of merged SOS and EOS dates, calculated using Equation (6), is presented in Figure 4. This range was determined using the REA method over the period from 1982 to 2020.*" Please refer to line 307 in the revised manuscript.

Figures
[Comment 53] (Fig 2): Why are plots b & d sharing an axis with plots a & c? They don't appear to share axis values. I was confused at first thinking that the weights (a & c) were shown by latitude, but that doesn't seem to be the case, so it is misleading to share an axis.
Also, the vertical figure legend is hard to read- is it possible to place it somewhere where the dataset names can be written horizontally?
Response: It may lead to misunderstanding with the border lines joined together. We separate a/c with b/d by discontinuous the line. And due to the limited space, placing them horizontally might require a smaller font, and we switch the way we placed the legend before in Figure 2.

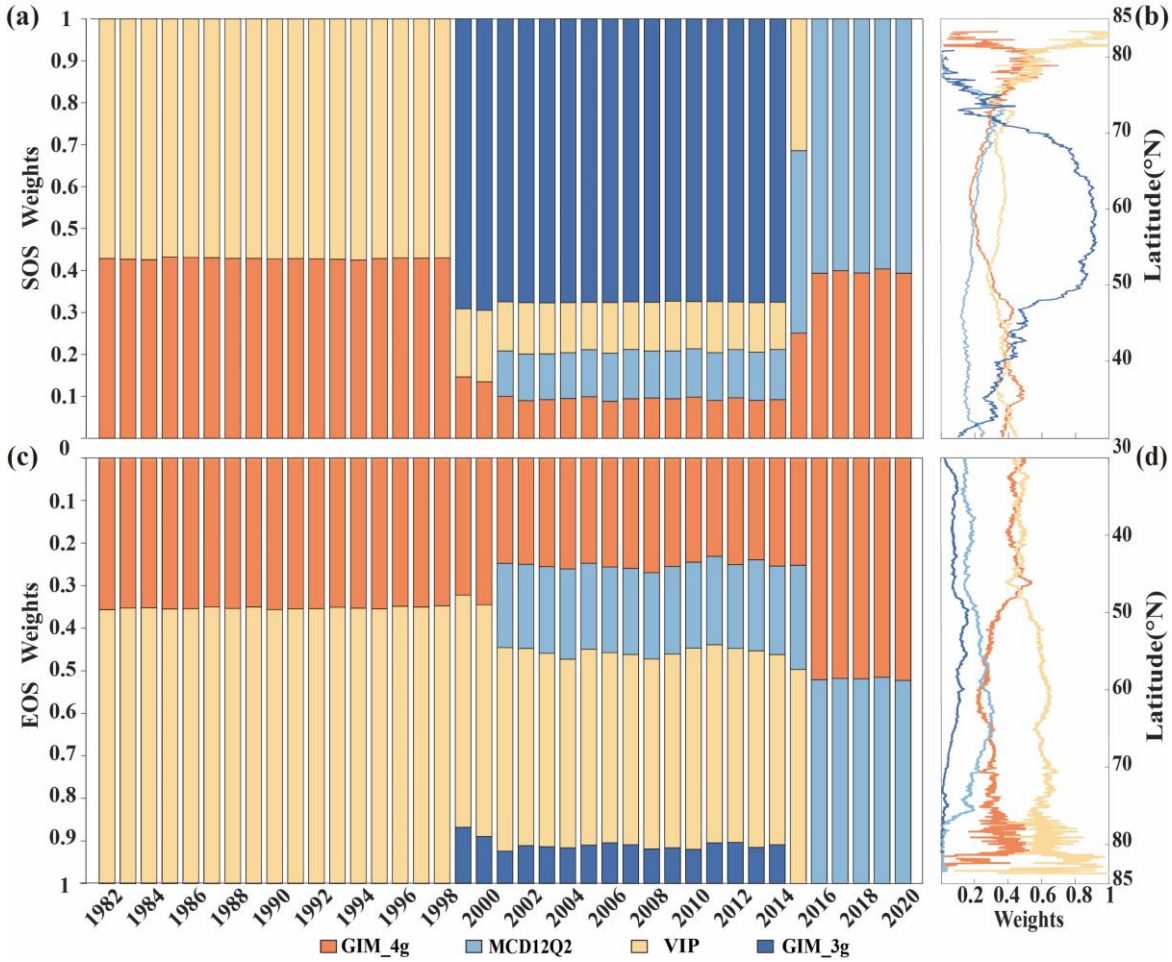


Figure 2: (a and c) Weights of the four phenology datasets during 1982–2020 and (b and d) latitudinal differences for (a and b) the SOS and (c and d) the EOS. The four datasets comprise the GIM_4g, MCD12Q2, VIP, and GIM_3g datasets (for the full names, see Table 1).

[Comment 54] (Fig 3): Please add a label/title to the legend scale bar (something like "Proportion of contribution")

Response: Following the reviewer's suggestion, we added "Rate of Contribution" to the legend scale bar, please see the updated figure3.

[Comment 55] (Fig 4): The legends and text are small and hard to read. In panel a, consider using a color scale with more variation – it's hard to see differences between the shades of green. In the figure legend text, "EOS dates" panel should be "d" instead of "b".

Response: Following the reviewer's suggestion, we have changed the color of Fig 4, and revised the legend text.

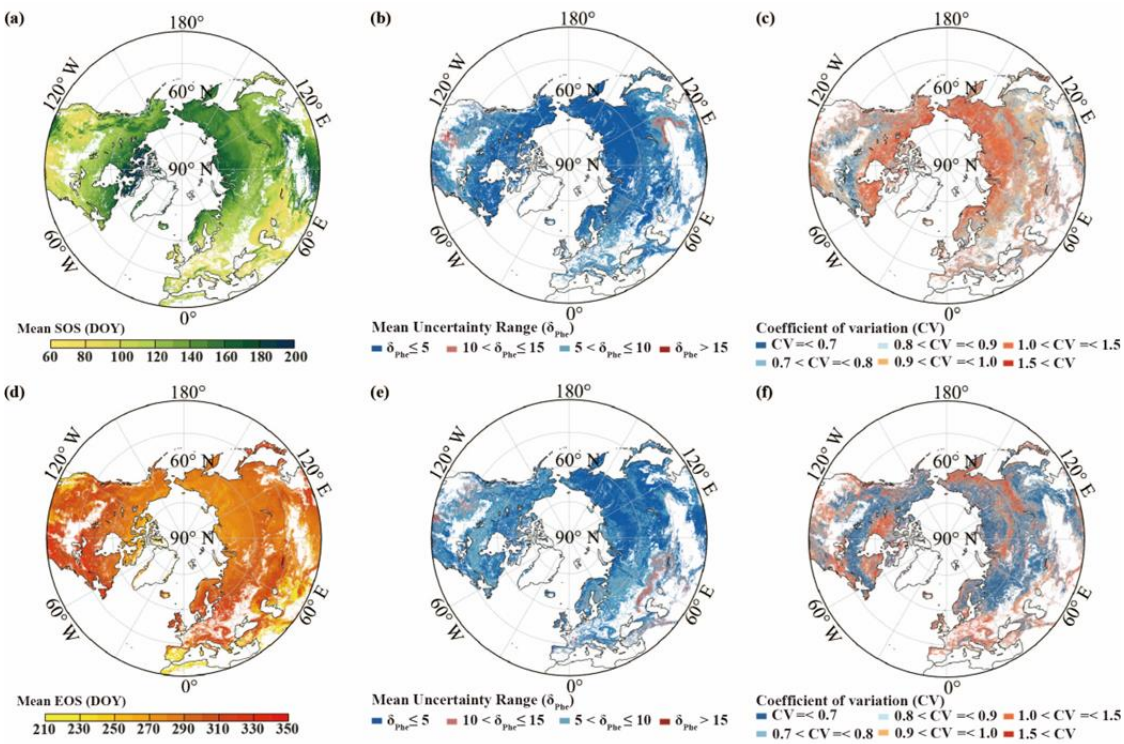


Figure 4: Merged mean (a) SOS and (d) EOS dates (DOY) obtained using the REA method for the period 1982–2020 and the uncertainty in the REA merged data. Mean uncertainty (δ_{Phe}) of SOS dates (b) and EOS (e) obtained using the REA method for the period 1982–2020, and its coefficient of variation (CV) in merged SOS (c) and EOS dates (f).

[Comment 56] (Fig 5): For all panels, include units in axes labels (DOY). In figure legend, remind readers that each data point represents a site year. I suggest moving the radar charts (panels f & k) to a separate figure (they're too small and hard to read) and include an explanation about how to interpret them in the results section.

Response: Following the reviewer's suggestion, we have added the "DOY" in axes labels, and Fig.5 has been redrawn for clarity, please see the updated Fig.5.

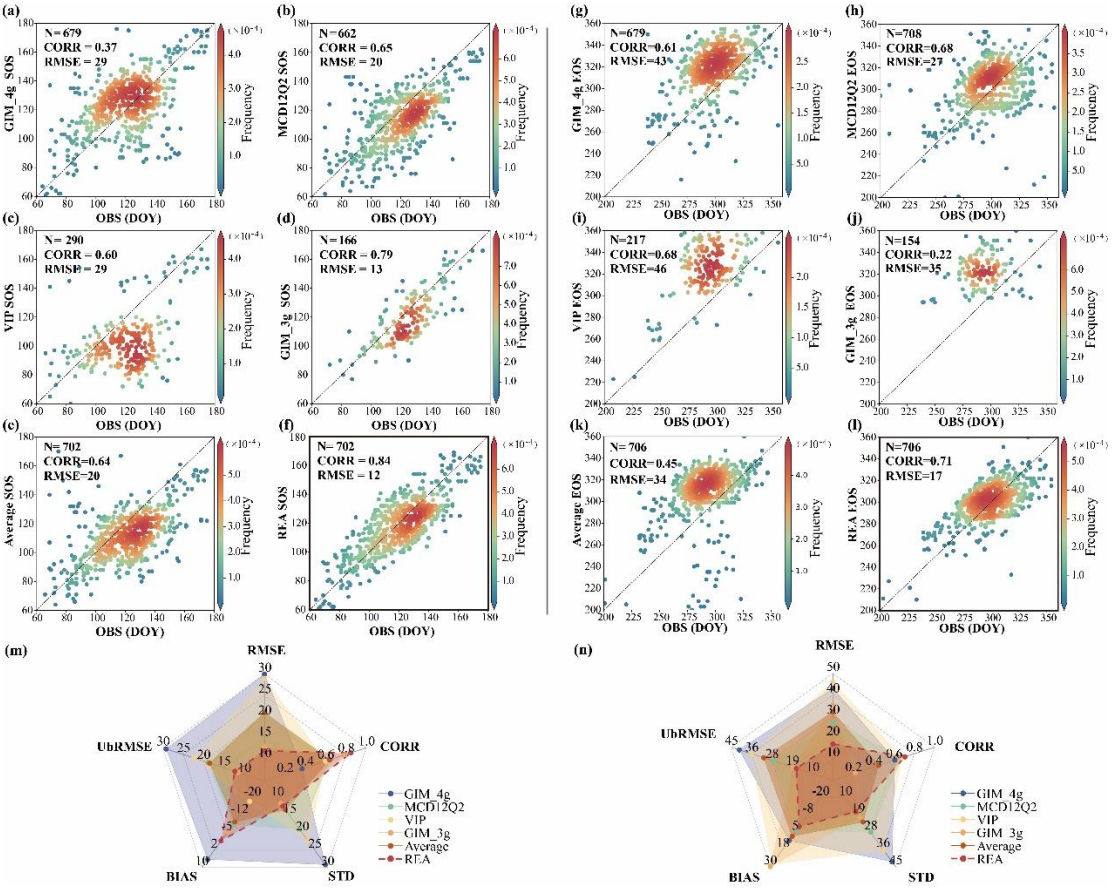


Figure 5: Scatterplots and radar charts of performance for each phenology dataset and the merged phenology dataset obtained using the REA method. (a–f) SOS evaluation results of the GIM_4g, MCD12Q2, VIP, GIM_3g, Average, and REA datasets, respectively, (m) radar chart of the SOS evaluation results, (g–l) EOS evaluation results of the GIM_4g, MCD12Q2, VIP, GIM_3g, Average, and REA datasets, respectively, and (n) radar chart of the EOS evaluation results. Each point represents a site year in the figure. OBS indicates ground-based phenocam phenological dates, RMSE indicates the root mean square error, UbRMSE indicates the unbiased RMSE, BIAS indicates the mean difference between the satellite-based results and the ground-based verification results, STD indicates the standard deviation, and CORR indicates the correlation coefficient.

[Comment 57] (Fig 6): Just an observation - for EOS, REA estimates are consistently late compared to phenocam. It could be related to how the EOS date is determined (which method/threshold used) for phenocam vs REA.

Response: It is true that if the time series data is reconstructed in the same method, the larger the threshold used may result in later SOS and earlier EOS. But datasets we used here are not reconstruct in the same method (see the revised data introduction section for details). The result of REA method is most similar to that of PhenoCam. And the PhenoCam EOS are not always lower than other datasets in all sites. Since it is difficult to observe autumn phenology on a large scale, there may be differences between satellite observation results and phenocam, the phenocam uses high frequency digital camera images to monitor vegetation phenology and is able to capture subtle changes in phenology, and remote sensing data is acquired less frequently and may not capture the exact date of EOS. We change another site and move it to supplementary materials to avoid confusion

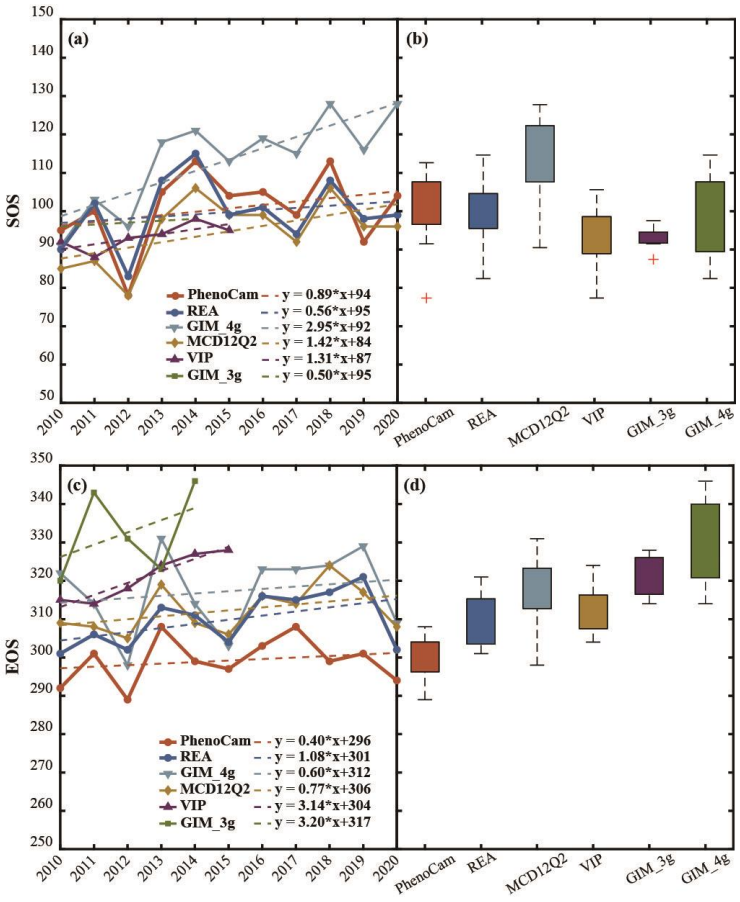


Figure S4 :Time series and box plots of Morganmonroe site data with each phenology dataset and the merged phenology dataset obtained using the REA method. (a-b) SOS time series and box plot of the PhenoCam, GIM_4g, MCD12Q2, VIP, GIM_3g, and REA datasets, respectively, (c-d) EOS time series and box plot of the PhenoCam, GIM_4g, MCD12Q2, VIP, GIM_3g, and REA datasets, respectively.

[Comment 58] (Fig 7): Need to add x-axis label. Are the black lines the average SOS/EOS date for each year? Please note in figure legend what the black and red dotted lines represent. In the percentage insets in panels b and d, I assume the x-axis letters represent the significant/non-significant advance and delays (abbreviations aren't defined)? Perhaps include the abbreviations in the legend with the colors: e.g., significant delay (DS), non-significant delay (DN), etc.

Response: We thank the reviewer for the detailed suggestions, and following the reviewer's suggestion, we have updated the fig 7 (now fig 6). In details, red lines in (a) and (c) are the fitting lines of average SOS/EOS date for each year, blue lines in (a) and (c) are the fitting lines of average SOS/EOS dates during 1982-2020, and black lines are the average SOS/EOS date for each year. In Figure6 (a) and (c) share the x label in the middle, we have added x-axis label to make the figure clear, and added the abbreviations for figure7 (c) and (d) in the figure illustration.

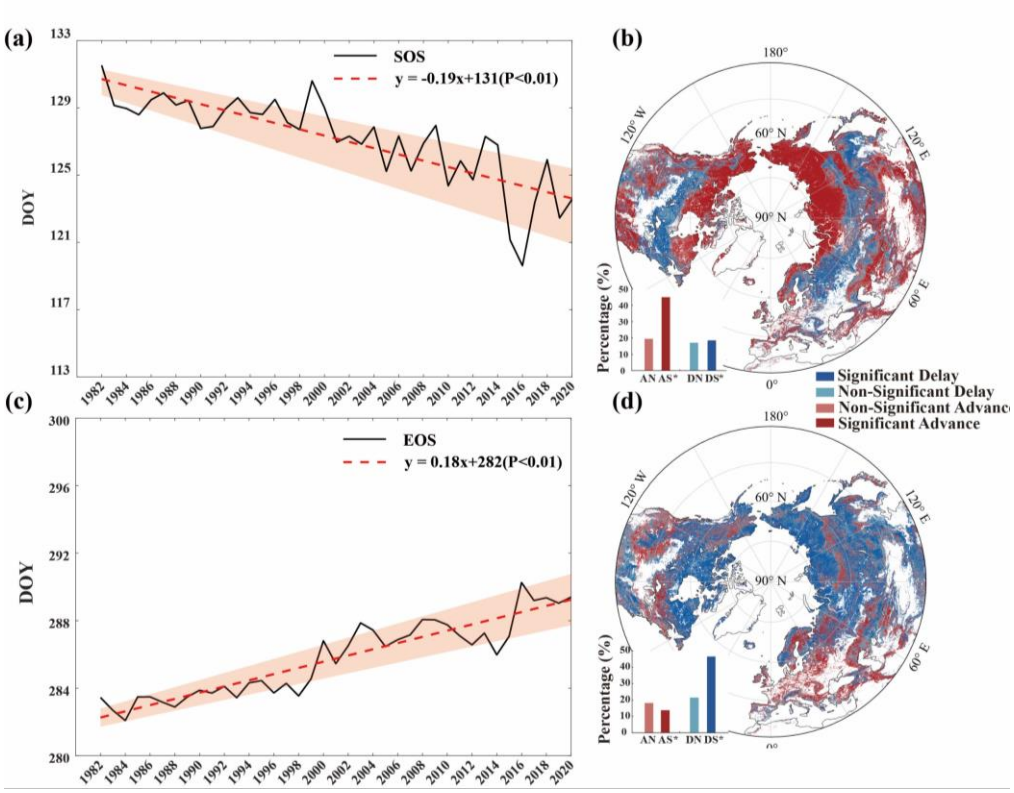


Figure 6: Temporal and spatial trends of the SOS and the EOS over the period 1982–2020 based on the merged dataset obtained using the REA method. (a) Temporal trend of the SOS over the period 1982-2020, (b) Spatial trend of the SOS over the period 1982-2020, (c)Temporal trend of the EOS over the period 1982-2020, (d) Spatial trend of the EOS over the period 1982-2020. The shaded area in (a) and (c) indicates uncertainty at one standard deviation, red lines in (a) and (c) are the fitting lines of average SOS/EOS dates for each year, and black lines are the average SOS/EOS date for each year. Significant delay (DS), non-significant delay (DN), significant advance (AS), non-significant advance (AN).

Response to Reviewer #2:

[Comment 1] The paper integrates four existing phenology datasets by applying different weights to the start (SOS) and end (EOS) of the growing season, as determined by each dataset. Overall, the manuscript is well-written and the dataset presented would be useful for community. However, several issues need to be addressed before it can be accepted for publication.

Response: We thank the reviewer for the supportive and constructive comments, and we appreciate the reviewer's assistance in the acknowledgments section in the revised manuscript. We have addressed each point and adjusted the manuscript according to the reviewer's comments. Please find below a detailed response to each comment.

Major concerns:

[Comment 2] First, the remote sensing datasets are not processed exactly the same, in terms of curves used to fit the time series and threshold used to extract transition dates.

the authors should clarify how these methodological differences might influence the uncertainty or accuracy of the resulting merged dataset.

Response: Thank you for your comment. Yes, indeed different methods were applied in different dataset. In MCD12Q2 dataset, the time series data was fitted by a penalized cubic smoothing spline, and Greenup (dormancy) is defined as the date when the EVI2 time series first (last) crosses 15% of the segment EVI2 amplitude. In VIP dataset the filtering method based on confidence interval and operational continuity algorithm were used to rebuild the time series curves, the start (end) of season is defined using the modified Half-Max method as the date when the NDVI2 time series first (last) crosses 35% of the segment NDVI2 amplitude. In GIM_3g dataset, a double logistic function was applied to fit the NDVI curve, and it uses the date when the NDVI first (last) crosses 20% of the segment NDVI amplitude as the SOS (EOS). In GIM_4g dataset, the NDVI time series data were fitted and smoothed using five fitting methods: the HANTS-Maximum, Spline-Midpoint, Gaussian-Midpoint, Timesat-SG, and Polyfit-Maximum methods, and it uses the date when the NDVI first (last) crosses 20% (50%) of the segment NDVI amplitude as the SOS (EOS). The smoothing method and phenology extraction method differs in these datasets.

Among different methods for vegetation phenology extraction, it is hard to distinct the best method for extracting vegetation phenology. According to previous and the present study, the phenology estimates obtained from different extraction methods show significant variation, with the estimated results differing by up to one month or more than 60 days depending on the method applied across different regions (Cong et al., 2012).

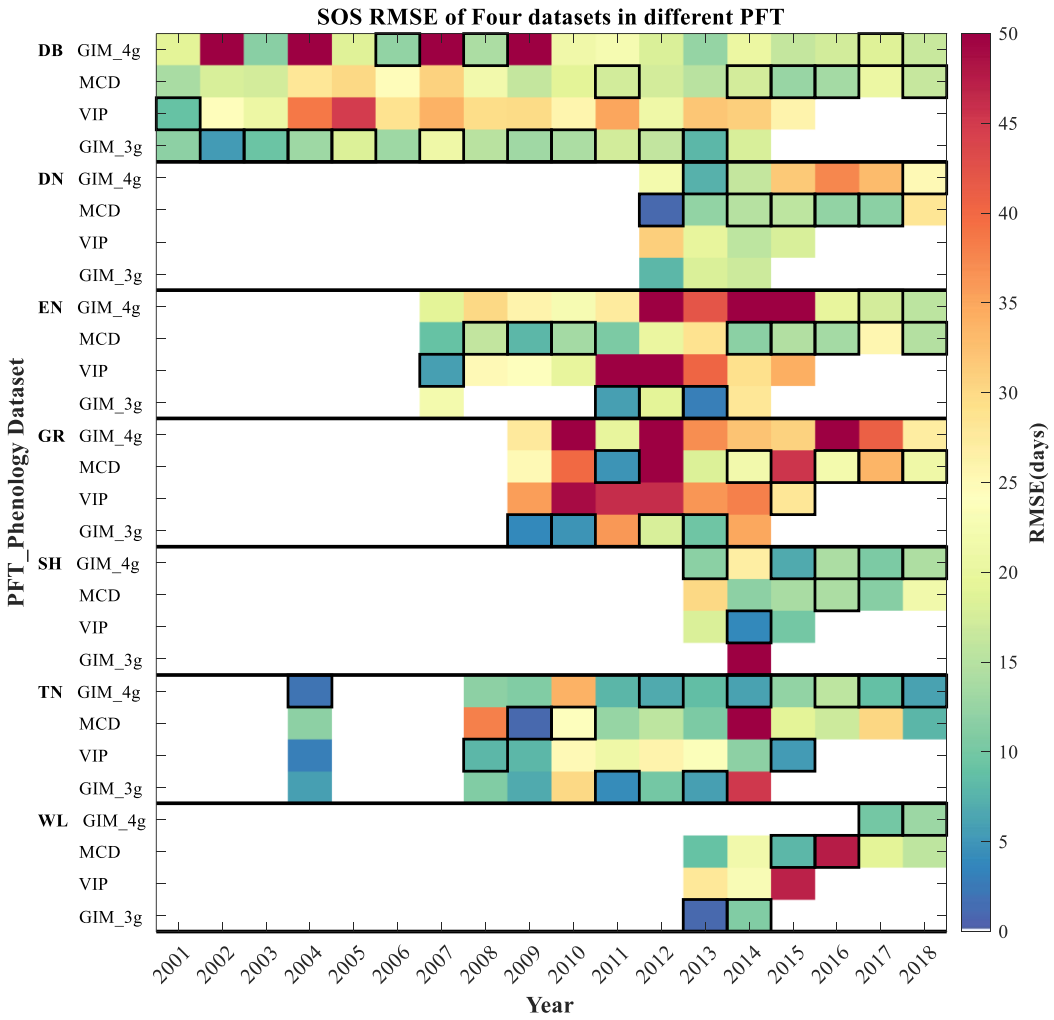
The phenological dates that were extracted from different methods were supposed to indicate changes in actual physiological conditions as accurately as possible, and the average method is often used for the fusion of different datasets, however, the effectiveness of these methods varies across regions and time period even, and may not always represent the true vegetation conditions. Different vegetation phenology datasets show different performance across regions and years comparing to the phenocam dates in the Fig. S1 below, for example, the consistency of VIP and ground phenocam data in the year 2001 of deciduous broadleaf forest is the best, whereas only the consistency of GIM_3g data in the year 2002 is better than that of VIP data. Comparing to the forest types, the consistency of remote sensing based phenological dates and phenocam data is higher in deciduous broadleaf region when using GIM_3g method, but in evergreen needleleaf when using the MCD method. We want to get the results that best reflect the physiological state at different sites and years.

Therefore, we use the REA method to catch the dates which can best reflects the change of the vegetation growing state based on the assumption that there exists a data source capable of reflecting the vegetation conditions at each gridcell, and different weights assigned to each data are calculated based on their reliability to get the final result.

To avoid confusion and clarify this issue, we revised the corresponding text, please refer to line 375-385 in the revised manuscript.

Cong, N., Piao, S., Chen, A., Wang, X., Lin, X., Chen, S., Han, S., Zhou, G., and Zhang, X.: Spring vegetation green-up date in China inferred from SPOT NDVI data: A multiple model analysis, ***Agricultural and Forest Meteorology***, 165, 104 – 113, https://doi.org/10.1016/j.agrformet.2012.06.009, 2012.

Wu, W., Sun, Y., Xiao, K., and Xin, Q.: Development of a global annual land surface phenology



dataset for 1982–2018 from the AVHRR data by implementing multiple phenology retrieving methods, ***International Journal of Applied Earth Observation and Geoinformation***, 103, 102487, https://doi.org/10.1016/j.jag.2021.102487, 2021.

Zeng, L., Wardlow, B. D., Xiang, D., Hu, S., and Li, D.: A review of vegetation phenological metrics extraction using time-series, multispectral satellite data, ***Remote Sensing of Environment***, 237, 111511, https://doi.org/10.1016/j.rse.2019.111511, 2020.

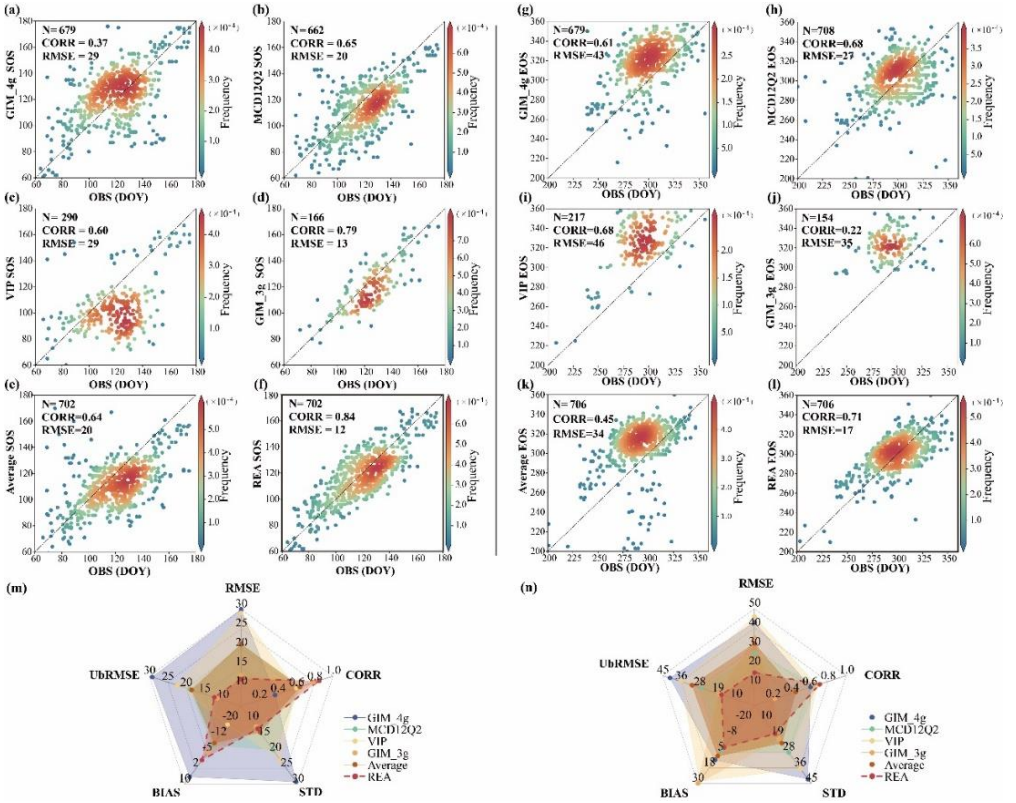
Zhang, J., Zhao, J., Wang, Y., Zhang, H., Zhang, Z., and Guo, X.: Comparison of land surface phenology in the Northern Hemisphere based on AVHRR GIMMS3g and MODIS datasets, ***ISPRS Journal of Photogrammetry and Remote Sensing***, 169, 1–16, 2020.

Figure S1: SOS RMSE of four datasets in different PFT for the period 2001–2018. Four datasets refer to GIM_4g, MCD12Q2, VIP, and GIM_3g datasets, respectively. PFT: plant functional type, DB: deciduous broadleaf, DN: deciduous needleleaf, EN: evergreen needleleaf, GR: grassland, SH: shrubs, TN: tundra, WT: wetland. The black boxes represent the best data for the year in that PFT.

[Comment 3] Second, I question whether the REA method truly outperforms a simple average. I recommend that the authors include additional analysis comparing the results obtained using the REA method with those from a simple average.

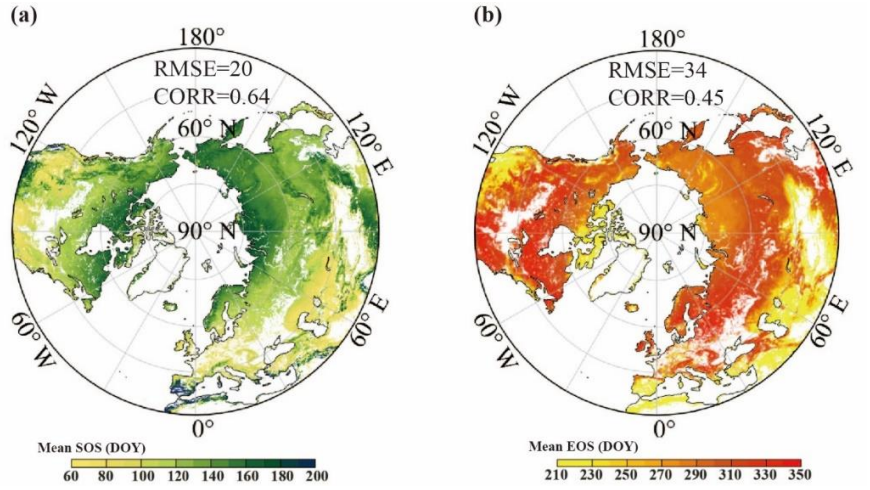
Response: Thank you for your thoughtful comment. Following the reviewer's suggestion, we have added the comparison between simple average and REA method result, see below Fig. 5(e) & (k). Comparing with simple average, the REA-based SOS shows better performance in RMSE (REA and Average, 12d and 21d, respectively), CORR (REA and Average, 0.84 and 0.65, respectively), BIAS (REA and Average, -1.5d

and -9.7d, respectively) and UbRMSE (REA and Average, 12d and 18d, respectively). The REA-based EOS also shows better performance in RMSE (REA and Average, 17d



and 32d, respectively), CORR (REA and Average, 0.71 and 0.45, respectively), BIAS (REA and Average, 1.0d and 8.0d, respectively) and UbRMSE (REA and Average, 17d and 31d, respectively). We also calculated the mean SOS and EOS using simple average for the period 1982-2020 in Fig. S3, there was little difference in the overall spatial distribution patterns of simple average and REA results, but the specific dates differ. We revised the figure and added the new results in the revised manuscript, please refer to line 327-329 and 336-337.

Figure 5: Scatterplots and radar charts of performance for each phenology dataset and the merged phenology dataset obtained using the REA method. (a–f) SOS evaluation results of the GIM_4g, MCD12Q2, VIP, GIM_3g, Average, and REA datasets, respectively, (m) radar chart of the SOS evaluation results, (g–l) EOS evaluation results



of the GIM_4g, MCD12Q2, VIP, GIM_3g, Average, and REA datasets, respectively, and (n) radar chart of the EOS evaluation results. Each point represents a site year in the figure. OBS indicates ground-based phenocam phenological dates, RMSE indicates the root mean square error, UbRMSE indicates the unbiased RMSE, BIAS indicates the mean difference between the satellite-based results and the ground-based verification results, STD indicates the standard deviation, and CORR indicates the correlation coefficient.

Figure S3: Mean (a) SOS and (b) EOS dates (DOY) obtained using simple average for the period 1982 – 2020. RMSE indicates the root mean square error (day), CORR indicates the correlation coefficient.

[Comment 4] PhenoCam data are not properly cited; please check out the fair use data policy here https://phenocam.nau.edu/webcam/fairuse_statement/.
Response: We thank the reviewer for point out this mistake. We have correctly cited the PhenoCam data both in the text and the acknowledgement in the manuscript.

Line edits:
[Comment 5]: PhenoCam, as a ground-based measurement, has been operational for more than 20 years. It should be introduced earlier in the text here.
Response: Following the reviewer's suggestion, we have added the introduction of PhenoCam earlier in the text. PhenoCam, as a ground-based measurement, has been operational for more than 20 years (Richardson et al., 2018a). Please refer to line 31 in the revised manuscript.

[Comment 6]: What is the specific time period over are evaluated?
Response: The trends of vegetation phenology from GIMMS3g and MODIS were evaluated during 2000-2015. This information has been added in the text, please refer to line 44-45 in the revised manuscript.

[Comment 7]: Please provide examples of regions where significant differences in the phenological metrics are observed.
Response: Please see the response to comment#2 for the same question.

[Comment 8]: As you mentioned earlier, the performance of datasets varies across regions. How does the REA method address or resolve these regional performance variations?
Response: Thank you for your comment. The REA method merged different datasets for a better performance globally, we assume that significant deviations are unlikely to occur simultaneously across most data sources within a specific region. Therefore, data containing anomalies will be excluded to achieve more accurate results using the REA method. If there is one data that shows significant discrepancies compared to other data, which may cause by improper extraction methods in that region, the $B_{Phe,i}$ and $D_{Phe,i}$ will extract this variance and combine with the natural variability ε_{Phe} of the region in the weight distribution process. If the natural variability of that region is low, a smaller value is assigned to the weight, and if the natural variability of the region is large, the weight is assigned by both the natural variability and the deviations. This is why the REA method demonstrates robust performance across all regions. To clarify this issue, we updated the corresponding text, please refer to line 201-204 in the revised manuscript.

[Comment 9]: How are SOS and EOS determined in the VIP phenology dataset? Please compare these criteria with the methods used to determine greenup in the MCD12Q2 dataset.

Response: Following the reviewer's suggestion, we added details information about criteria in these methods. In details, the start (end) of season is defined using the modified Half-Max method as the date when the NDVI2 time series first (last) crosses 35% of the segment NDVI2 amplitude in VIP. Greenup (dormancy) is defined as the date when the EVI2 time series first (last) crosses 15% of the segment EVI2 amplitude in MCD12Q2. Please refer to line 103-107 and 94-95 in the revised manuscript.

[Comment 10]: What specific curves are applied for the MCD12Q2 and VIP datasets?

Response: The time series data was fitted by a penalized cubic smoothing spline to rebuild time series curve in MCD12Q2 dataset (line 90). The filtering method based on confidence interval and operational continuity algorithm were used to rebuild the time series curves in VIP dataset. We revised the corresponding text, please refer to line 95 and line 103-104 in the revised manuscript.

[Comment 11]: What threshold is used to extract phenological metrics from the GIM_3g dataset?

Response: This product provides phenology data for the Northern Hemisphere, and it uses the date when the NDVI first (last) crosses 20% of the segment NDVI amplitude as the SOS (EOS). We revised the corresponding text, please refer to line 112 in the revised manuscript.

[Comment 11]: Please provide a link to the PhenoCam dataset for reference.

Response: We have added the website of PhenoCam_V2 (https://daac.ornl.gov/VEGETATION/guides/PhenoCam_V2.html) and the PhenoCam (https://phenocam.nau.edu/webcam/) in the text. Please refer to line 133 in the revised manuscript.

[Comment 12]: Since the method relies on interannual variability in the time series, what is the minimum required length for the time series? Is it possible to use REA to merge only two datasets? A discussion or quick test with shorter time series would be valuable, especially considering the availability of recent Planet data with.

Response: Thank you for your comment and following the reviewer's suggestion, we updated the discussion section. In the method proposed by Giorgi et al. 2001, there is no restriction on the minimum value for this parameter. It can be adjusted multiple times according to the actual data to find an appropriate range for the specific dataset, but it should be as large as possible to reflect the natural variability. It is possible to get the result by merging two datasets with REA, but the accuracy may be less than that of merging with more reliable data sources. Please refer to the revised text in line 413-415 in the revised manuscript.

Giorgi, F. and Mearns, L. O.: Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the "Reliability Ensemble Averaging" (REA) Method, ***Journal of Climate***, 15, 1141–1158, https://doi.org/10.1175/1520-0442(2002)0152.0.CO;2, 2002.

[Comment 13]: For datasets with higher interannual variability, did you assign them lower weights in the REA method? Please clarify.

Response: Interannual variability is measured by ε_{Phe} in equation (2), which is also represents for natural variability. Natural variability changes from region to region, in Equation (1) and (6), ε_{Phe} cancels out under the condition of $B_{Phe,i}$ and $D_{Phe,i}$ greater than ε_{Phe},which based on the assumption that more stringent on are required to increase the reliability over regions characterized by lower natural variability. The natural variability does not work single, it works with $B_{Phe,i}$ and $D_{Phe,i}$ jointly. For the region in lower natural variability, if the phenology data from one dataset also have large difference with other datasets, it is given lower weight for generate the REA phenology at that region, which is thought to be less accurate data at that region. To avoid confusion, we revised the corresponding text, please refer to line 189-194 and 201-204 in the revised manuscript.

[Comment 14]: Open-source code for the REA method should be made available. This would assist the community in merging datasets from various sources and years.

Response: The code is shared on Github (https://github.com/PRqA642/REA).

[Comment 15]: Please provide a brief description of the metrics used and their characteristics.

Response: We have supplemented the description to our manuscript. The RMSE is calculated as the square root of the average of the squares of the residuals, which penalizes larger errors than smaller ones and provide an estimate of the magnitude of errors between remote sensing estimated value and phenocam datasets. BIAS is the average difference between remote sensing estimated value and phenocam value, that helps in understanding whether the estimated value is higher or lower than phenocam value. The correlation coefficient measures the linear relationship between two variables. The ubRMSE measures the deviation between two variables without systematic errors. Standard deviation quantifies the variation of the dataset, which measures the deviation between data and the mean value. Please refer to line 217-223 in the revised manuscript.

[Comment 16]: Provide a specific example of how the M-K test will be applied in the study.

Response: We used M-K test to analyze the trend of SOS and EOS during 1982-2020 in the merged dataset, and we have supplemented this part of content. We have added how we will use the M-K test in the method part. Please refer to line 227-228 in the revised manuscript.

[Comment 17]: The citation of Mao and Sun may not be necessary here—consider removing it.

Response: Following the reviewer's suggestion, we have removed the citation of Mao and Sun from the text in the revised manuscript.

[Comment 18]: I am a bit concerned about the huge deviations in spring 2022 shown by Figure 7, which seems very inconsistent with Figure 2.78 in DOI: https://doi.org/10.1175/BAMS-D-23-0090.1

Response: Thank you for your comment. During 2022, there is only MCD12Q2 left as the data source, since other dataset do not include this time period and large uncertainty may exist, so we remove the data after 2020. In the revised manuscript, we estimated the trends in SOS and EOS until 2020, please see the revised figure 6.

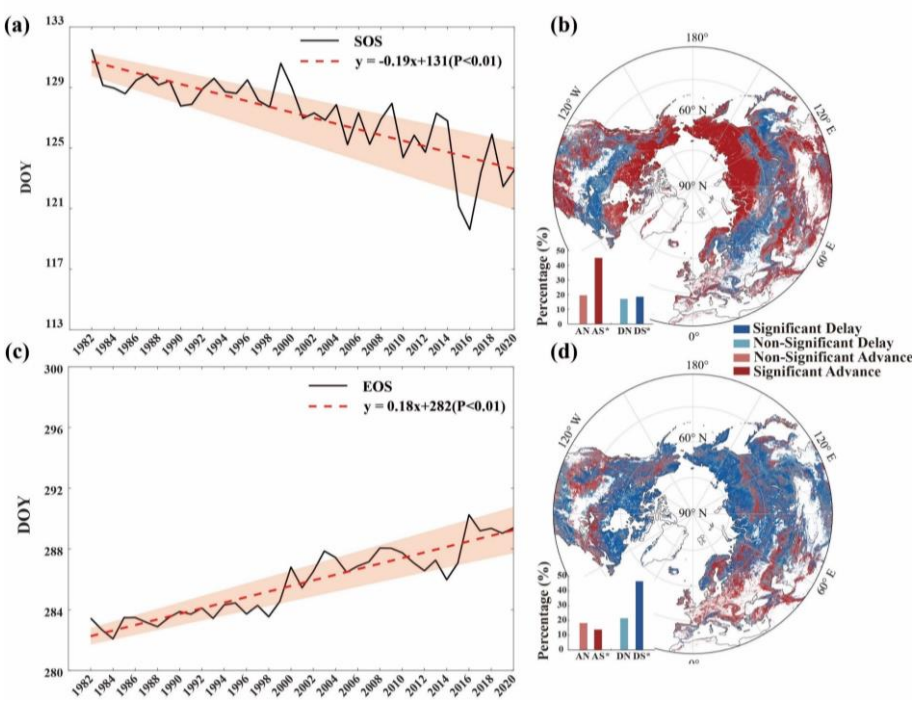


Figure 6: Temporal and spatial trends of the SOS and the EOS over the period 1982–2020 based on the merged dataset obtained using the REA method. (a) Temporal trend of the SOS over the period 1982-2020, (b) Spatial trend of the SOS over the period 1982-2020, (c)Temporal trend of the EOS over the period 1982-2020, (d) Spatial trend of the EOS over the period 1982-2020. The shaded area in (a) and (c) indicates uncertainty at one standard deviation, red lines in (a) and (c) are the fitting lines of average SOS/EOS dates for each year, and black lines are the average SOS/EOS date for each year. Significant delay (DS), non-significant delay (DN), significant advance (AS), non-significant advance (AN).

[Comment 19]: A description of how uncertainty is determined needs to be added to the REA phenology dataset.

Response: The calculation of uncertainty is introduced in the method part, please refer to line 195-200 in the revised manuscript. The uncertainty range is calculated based on the weight of each dataset and the deviation between REA result and data sources, the upper and lower uncertainty limits are measured by REA result and the uncertainty range.