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 376-385 and 398-404 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.

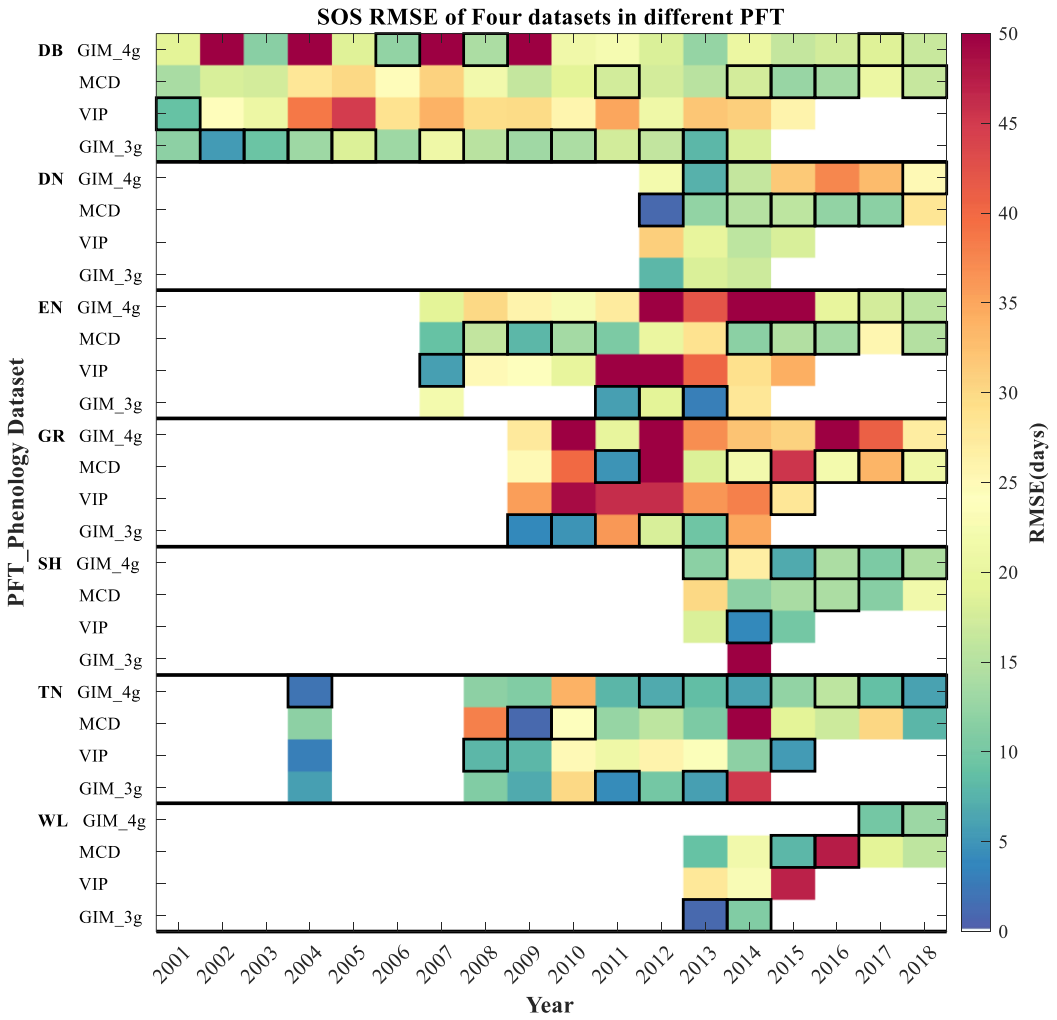

Figure S1: SOS RMSE of four datasets in different PFT for the period 2001–2018. Four datasets refer to GIM_4g, MCD12Q2, VIP, and GIM_3g datasets, respectively. PFT: plant functional type, DB: deciduous broadleaf, DN: deciduous needleleaf, EN: evergreen needleleaf, GR: grassland, SH: shrubs, TN: tundra, 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 329-330 and 336-338.

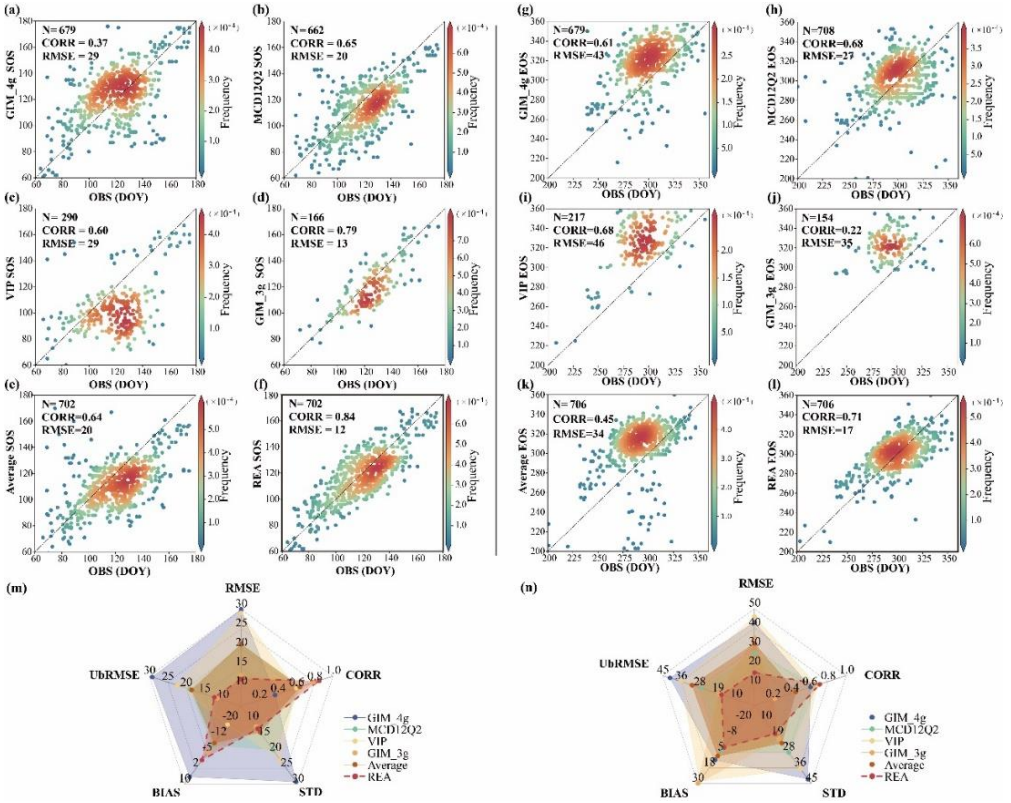

Figure 5: Scatterplots and radar charts of performance for each phenology dataset and the merged phenology dataset obtained using the REA method. (a–f) SOS evaluation results of the GIM_4g, MCD12Q2, VIP, GIM_3g, Average, and REA datasets, respectively, (m) radar chart of the SOS evaluation results, (g–l)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

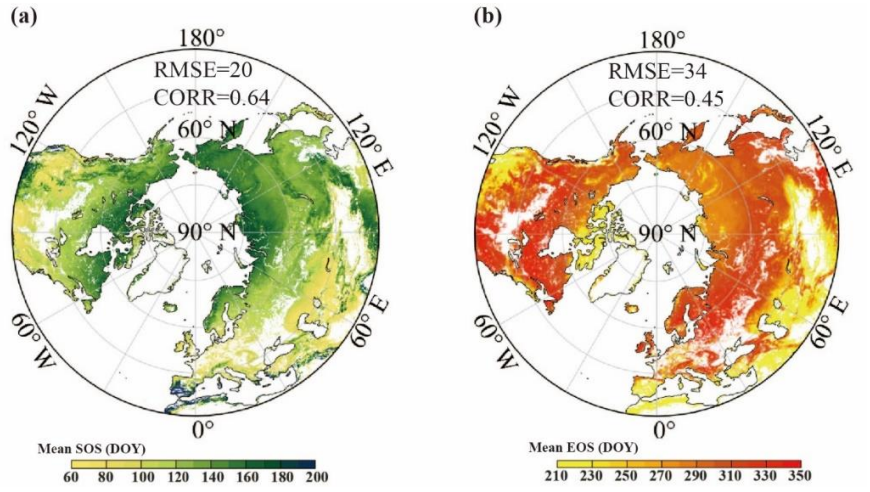

Figure S3: Mean (a) SOS and (b) EOS dates (DOY) obtained using simple average for the period 1982 − 2020. RMSE indicates the root mean square error (day), CORR indicates the correlation coefficient.

[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 415-416 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.

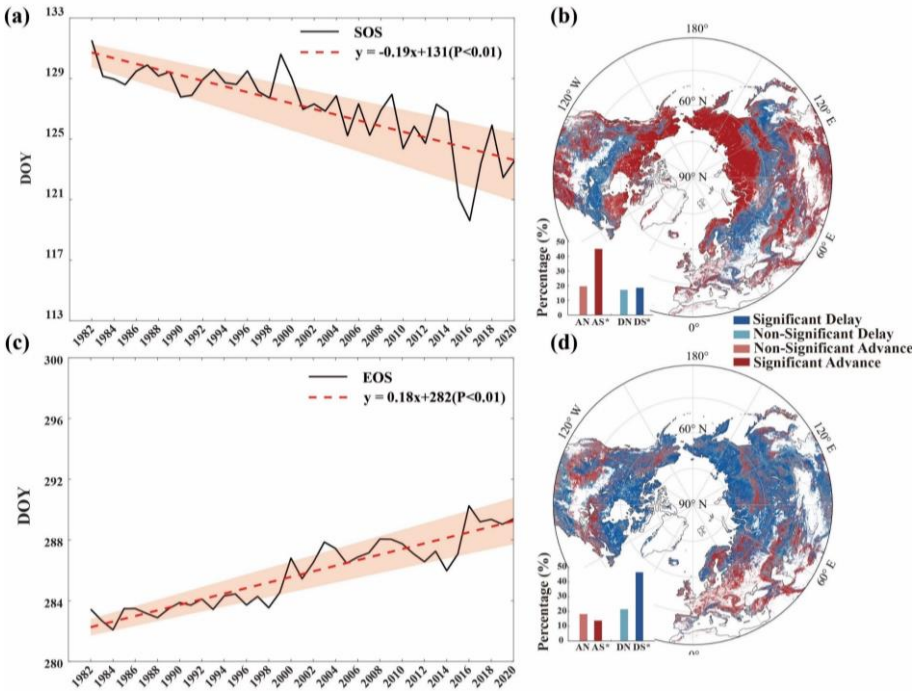

Figure 6: Temporal and spatial trends of the SOS and the EOS over the period 1982–2020 based on the merged dataset obtained using the REA method. (a) Temporal trend of the SOS over the period 1982-2020, (b) Spatial trend of the SOS over the period 1982-2020, (c)Temporal trend of the EOS over the period 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.