Response to topic editor:

[Comment 1] Thank you for your resubmission and patience in the rereview process. While both reviewers appreciated the authors' revisions and have a more positive outlook on the study, they still have remaining comments that need to be addressed before we consider the manuscript for publication.

Response: We sincerely appreciate the editor's time and effort in handling our manuscript and thank the reviewers for their thoughtful reassessment of our work, we have now included our appreciation in the acknowledgments section of the revised manuscript. We are grateful for their constructive feedback and are pleased to hear that they have a more positive outlook on our study. We have carefully revised the manuscript accordingly. Below, we provide a detailed response to each comment, outlining the corresponding revisions.

[Comment 2] Both reviewers still have reservations about combining the phenologic information across four datasets that use inherently different approaches. Both have specific thoughts on these points and I agree all should be addressed. At the very least, there needs to be some more discussion that combining these datasets across different methodologies is a limitation, even despite some benefit from the REA method. I noticed one or two sentences on this in the discussion, but given the concerns of the reviewer and myself, it could be expanded on a bit more. Specifically, there are both differences in the data sources, but also the methodology used from these sources. I, however, can appreciate that there are different approaches to estimating start and end of season, with not much agreement or motivation across the community about which method is "best." Integrating across methods is valuable. Indeed, also showing how REA has higher value than arithmetic average is useful and shows some potential mitigation of these issues.

Response: We sincerely appreciate the editor's and reviewers' insightful comments and concerns regarding the integration of phenological information across datasets that employ inherently different approaches.

We have expanded the discussion section to provide more examination of these limitations by explaining the differences in data sources and processing methodologies among the four phenology datasets, clarifying how these variations may influence the REA results. While discrepancies may exist in multi-source data integration, the REA method mitigates these inconsistencies by dynamically weighting datasets based on their interannual variability and agreement with others.

We conducted sensitivity analyses using different dataset combinations and time spans to evaluate the robustness of this method. Our results (see Figures S6 and S7) indicate that the composition of input datasets has a greater influence on fusion results than the length of the time series. The average deviation between different dataset combinations and the long-term four-dataset fusion results is 7.1 days, whereas SOS and EOS variations across different time spans (e.g., 5-year vs. 10-year fusion) are relatively minor (4.1 and 3.2 days, see Figure S6). Moreover, fusion weights derived from shorter time series are largely consistent with those obtained using the full dataset (see Figure S8). In the multi-time span analysis, the fusion results of SOS are more stable than EOS, and the longer time series result is more similar to the REA fusion result (1982-2022).

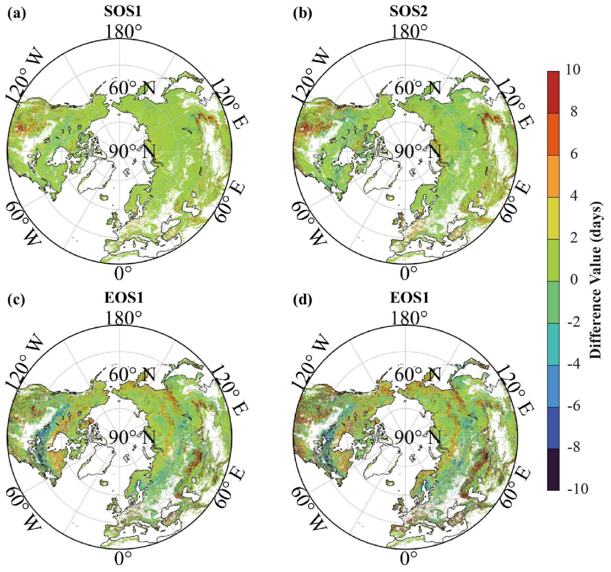

Figure S6: Comparison of fusion results of four groups of data at different time lengths from 2001 to 2010. (a, c) Difference plots between the fusion results of SOS and EOS for the four datasets over the 2001-2010 period and the original long-term REA fusion results. (b, d) Difference plots between the fusion results of SOS and EOS obtained by separately fusing the two sub-periods of data (2001-2005 and 2006-2010) and subsequently concatenating them, and the original long-term REA fusion results.

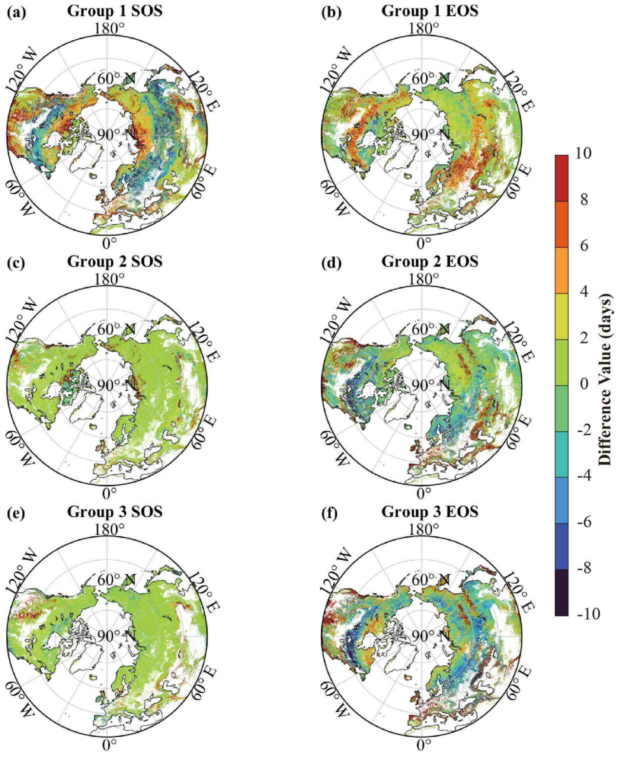

Figure S7: Differences between the SOS and EOS data fusion results from various data source combinations and the fusion results of the long-term four datasets over the 2001-2010 period. (a-

b) Group 1 represents the difference between the fusion of GIM_4g and VIP data and four datasets REA result. (c-d) Group 2 represents the difference between the fusion of VIP data and GIM_3g data and four datasets REA result. (e-f) Group 3 represents the difference between the fusion of GIM_3g and MCD12Q2 data and four datasets REA result.

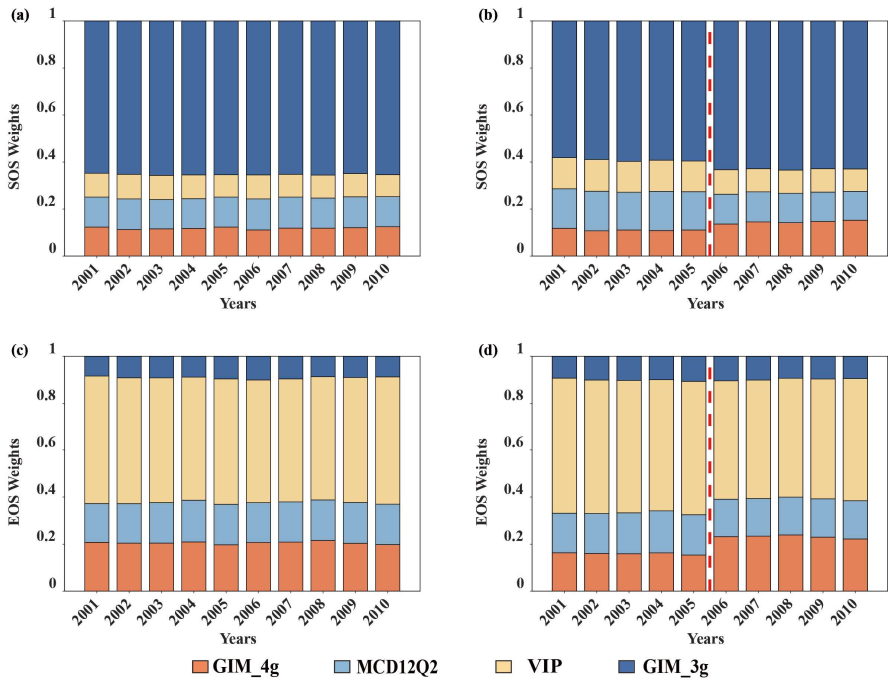

Figure S8: Comparison of fusion data weights for the four datasets over different time lengths from 2001 to 2010. (a, c) Weights of the SOS and EOS fusion data for the four datasets over the entire 2001-2010 period. (b, d) Weights of the fusion data after separately fusing the two sub-periods of data (2001-2005 and 2006-2010) and subsequently concatenating them. The red dashed line separates the two sub-periods.

Our analysis suggests that datasets incorporating more reliable sources could improve the result of data fusion. For example, groups 2 and 3, which include GIM_3g SOS, exhibit smaller differences from the long-term fusion results than group 1 and demonstrate higher consistency with PhenoCam observations (Figure 5d). The EOS fusion results are more stable across different dataset combinations. These results show that multi-source data integration could help improve accuracy. The similarity in weight distributions across different dataset combinations (Figure S9) suggests that dataset selection remains a critical factor influencing final results, and incorporating additional high-quality datasets in the future could further enhance accuracy and robustness.

The effectiveness of data fusion results depends on the complementarity and quality of input datasets. The REA method can effectively integrate different datasets with independent information, enhancing both the accuracy and consistency of the phenological result. Given the lack of consensus on the "best" approach for extracting phenological dates, our study highlights the value of integrating multiple-method datasets rather than relying on a single dataset.

Furthermore, we clarify the advantage of the REA method over a simple arithmetic average by demonstrating its ability to refine SOS and EOS estimates through dynamic weighting. Our results show that REA reduces biases associated with individual datasets, producing a more reliable phenology dataset. We have made these points clearer in the revised manuscript.

Please refer to lines 448-474.

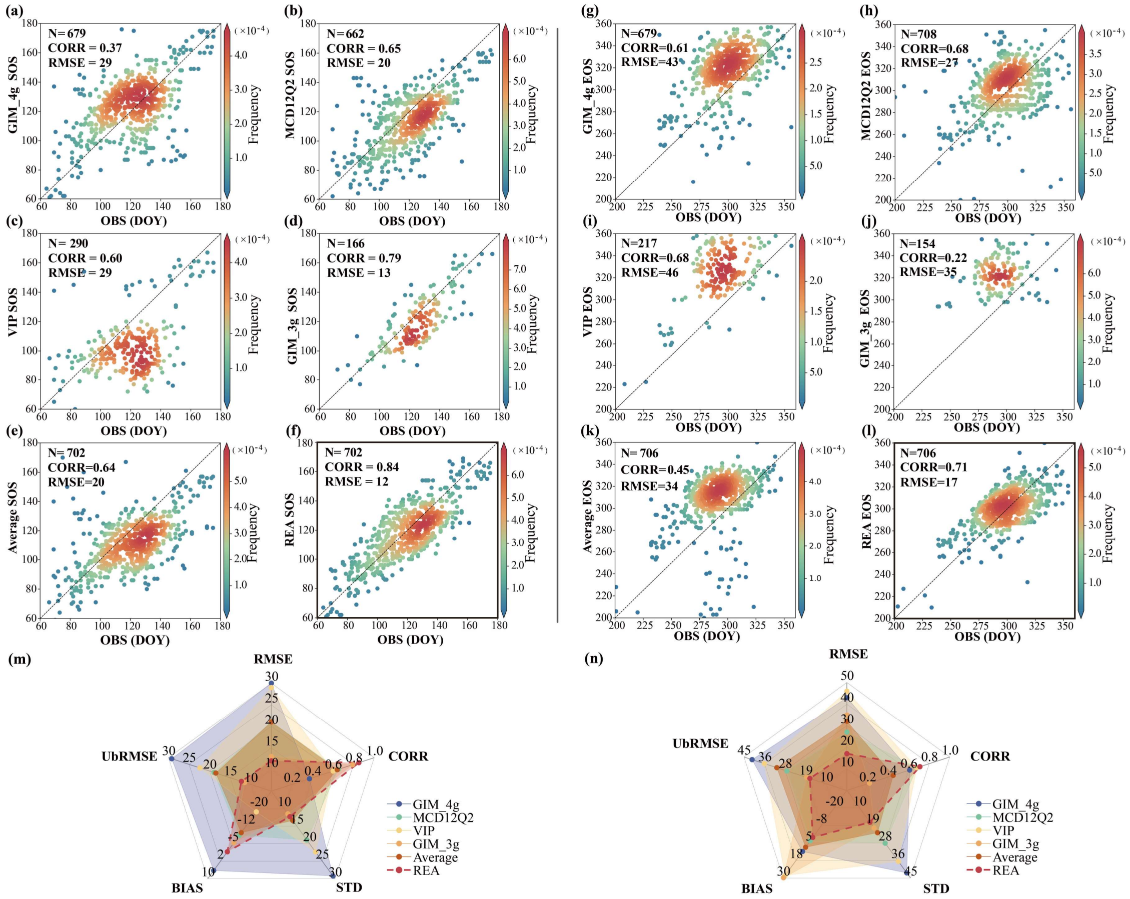

Figure 5: Scatterplots and radar charts of performance for each phenology dataset and the merged phenology dataset obtained using the REA method. (a–f) SOS evaluation results of the GIM_4g, MCD12Q2, VIP, GIM_3g, Average, and REA datasets, respectively, (m) radar chart of the SOS evaluation results, (g–l)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

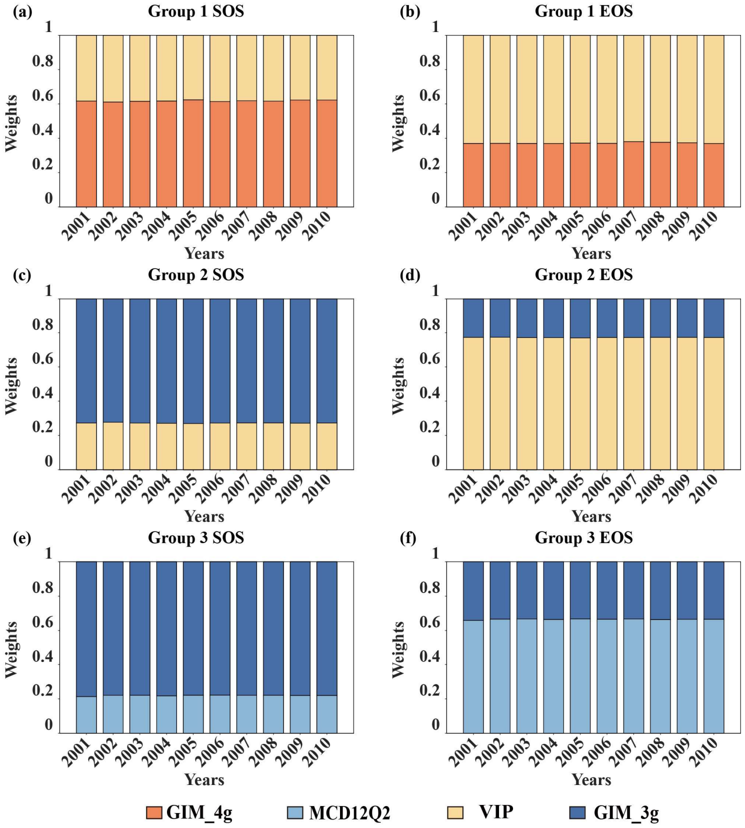

Figure S9: Distribution of fusion weights for SOS and EOS data from different data source combinations over the 2001-2010 period. (a, b) Group 1: Fusion weights for the combination of GIM_4g and VIP data. (c, d) Group 2: Fusion weights for the combination of VIP data and GIM_3g data. (e, f) Group 3: Fusion weights for the combination of GIM_3g and MCD12Q2 data.

[Comment 3] Both reviewers also agree revisions are required in the discussion. Please follow their points, which is not limited to more discussion about the limitations of the REA method itself and limitations combining the datasets that use different methodologies.

Response: We have revised the manuscript following the reviewers' suggestions. Specifically, we have added a discussion on the limitations of the REA method in Section 4.2 (Strengths, Limitations, and Sensitivity Analysis of the REA Approach), expanded the discussion on dataset discrepancies and integration challenges in Section 4.1 (Integrating Multi-Source Phenology Data and Addressing Dataset Discrepancies), and provided further insights into the differences in vegetation phenology trends and key influencing factors in Section 4.3 (Consistency of SOS and EOS Trends Across Studies and Key Influencing Factors). These revisions directly address the reviewers' concerns and enhance the discussion of methodological and dataset-related limitations. Please refer to lines 419-513 in the revised manuscript and we cope the text directly, please see the details below.

4.1 Integrating Multi-Source Phenology Data and Addressing Dataset Discrepancies

This study integrates four widely used remote sensing phenology datasets (VIP, GIM_4g, MCD12Q2, and GIM_3g) using the Reliability Ensemble Averaging (REA) method to improve the consistency of Start of Season (SOS) and End of Season (EOS) estimations. Our analysis *shows substantial discrepancies among these datasets, consistent with previous findings that SOS estimates can vary by 60 days (White et al., 2009) or even exceed 100 days (De Beurs and Henebry, 2010), depending on the extraction method. We find the inconsistencies that arise from both differences in data sources and variations in phenology extraction methodologies by evaluating the spatial distribution of the multi-year mean SOS and EOS across datasets.*

The REA method produces a more harmonized phenology result by assigning different weights to each dataset based on its reliability to mitigate discrepancies. The effectiveness of this approach is demonstrated by the lower deviation of SOS and EOS estimates across merged results and phenocam datasets. As shown in Figure 1, the differences in SOS and EOS among the four datasets across the Northern Hemisphere reach approximately 15 days on average, with some regions exhibiting even larger discrepancies (more than 120 days). These variations might be caused by factors such as sensor characteristics, data preprocessing techniques, and phenological extraction methods. For instance, similarities in spatial distributions between the GIM_4g and VIP datasets likely stem from their common use of AVHRR sensor data, whereas their mean difference (14 days) could be attributed to differences in time-series reconstruction techniques. Differences between GIM_4g and GIM_3g are primarily driven by different data fitting methodologies. The data smoothing techniques aim to preserve vegetation dynamics while reducing noise, but the optimal method is still unclear, as the appropriate method depends on the specific biogeographical characteristics of a given region (Hird and McDermid, 2009).

Additionally, phenology extraction methods exhibit varying effectiveness across different ecological regions (Reed, 2003). Evaluating dataset consistency with PhenoCam observations across different plant functional types (PFTs), we find that dataset performance varies by PFT and time period. For instance (in Figure S2), VIP data show consistency with PhenoCam observations in deciduous broadleaf forests in 2001, whereas GIM_3g outperformed other datasets in 2002. And MCD12Q2 shows better agreement in evergreen needleleaf forests. These findings stress the challenges of integrating datasets with inherently different methodological foundations. We further evaluated the greening trend, which reflects the vegetation growth status during the growing season, as indicated by NDVI trends (Li et al., 2023). The greening trend estimated from VIP ($-10.16 \times 10^{-4}\ yr^{-1}$) *was lower than that from REA* ($-1.14 \times 10^{-4}\ yr^{-1}$) *and GIM_4g* ($3.55 \times 10^{-4}\ yr^{-1}$)*. The proportion of areas exhibiting greening trends was 25.22% (with 41.44% significantly greening) for VIP, 68.49% (with 38.32% significantly greening) for GIM_4g, and 49.83% (with 56.83% significantly greening) for REA. The substantial differences in greenness trends across datasets (in Figure S3) highlight the necessity of an integrated approach for robust phenology estimation.*

4.2 Strengths, Limitations, and Sensitivity Analysis of the REA Approach

Given the lack of consensus on the "best" approach for extracting phenological dates, our study highlights the value of integrating multiple methodologies rather than relying on a single dataset. The REA method improves phenology estimations by weighting datasets based on reliability rather than assuming equal contributions, as in a simple arithmetic average (Lu et al., 2021). The REA method considers the temporal correlation of vegetation phenology data by employing a voting principle (Giorgi and Mearns, 2002), and this approach facilitates convergence of data while retaining differences of the spatial distribution, thereby offering *advantages in multisource data fusion than simple averaging. Our results indicate that REA merged phenology results are more consistent with PhenoCam observations than individual datasets and simple average results, demonstrating its potential to reduce biases between different datasets caused by different extraction methods. However, while REA enhances consistency, it does not fully eliminate methodological discrepancies.*

To evaluate the robustness of this approach, we conducted sensitivity analyses using different dataset and time span combinations. Our results (see Figures S6 and S7) indicate that the combination of different input datasets has a greater influence on fusion results than the length of the time series. The average deviation between different dataset combinations and the long-term four-dataset fusion results is 7.1 days, whereas SOS and EOS variations across different time spans (e.g., 5-year vs. 10-year fusion) are relatively minor (4.1 and 3.2 days, see Figure S6). Moreover, fusion weights derived from shorter time series are largely consistent with those obtained using the full dataset, see Figure S8. In the multi-time span analysis, the fusion results of SOS are more stable than EOS, and the longer time series result is more similar to the REA fusion result (1982-2022).

The effectiveness of data fusion depends on the complementarity and quality of input datasets. Our analysis suggests that datasets incorporating more reliable sources could produce improved fusion results. For example, GIM_3g SOS demonstrate higher consistency with PhenoCam observations (Figure 5d); groups 2 and 3, which include GIM_3g SOS, exhibit smaller deviations from the long-term REA fusion results than group 1. In contrast, EOS estimates are more stable across different dataset combinations. These results show that multi-source data integration could help improve accuracy. The similarity in weight distributions across different dataset combinations (Figure S9) suggests that dataset selection remains a critical factor influencing final estimates. Incorporating additional high-quality datasets in future studies could further enhance accuracy and robustness.

4.3 Consistency of SOS and EOS Trends Across Studies and Key Influencing Factors

Global climate change has significantly influenced vegetation phenology, with a general advancement in SOS and a delay in EOS (Piao et al., 2019a). Our REA-based dataset estimates an SOS advancement of 0.19 days per year during 1982-2020 ($p<0.01$), aligning with previous findings, such as 1.4 ± 0.6 days per decade for 1982-2011 (Wang et al., 2015) and a 5.4-day advance from 1982 to 2008 (Jeong et al., 2011). Similarly, EOS has been delayed at a rate of 0.18 days per year ($p<0.01$), consistent with prior estimates of 0.18 ± 0.38 days per year for 1982–2011 (Liu et al., 2016) and a 6.6-day delay from 1982 to 2008 (Jeong et al., 2011). Deviations between REA-based trends and those from individual datasets (Blunden et al., 2023) suggest that variations in data sources and study regions contribute to observed discrepancies.

Differences in SOS trends across studies arise from multiple factors, including study period, land cover changes, spatial domain, environmental heterogeneity, and methodological differences. Trend estimates are highly sensitive to the selected time window, particularly at the start and end years (Cong et al., 2013). Interannual and decadal climate variability can further influence these trends contributing to discrepancies across studies. Land cover changes, such as wildfires and deforestation (Jeong et al., 2011), can influence the extraction result of phenology, then affect SOS trends. Additionally, spatial differences in studies also influence the *rate of trends in different studies, as phenology is affected by vegetation type and climate sensitivity. For instance, Jeong et al. (2011) analyzed temperate vegetation between 30°N–80°N, and Wang et al. (2015) focused on 30°N–75°N, excluding evergreen forests and managed landscapes. Since phenology responds differently across species and locations (Maignan et al., 2008), variations in study areas contribute to discrepancies. Photoperiod constraints introduce additional latitudinal differences in phenological responses to warming (Meng et al., 2021). Moreover, precipitation and other environmental factors vary spatially, further influencing SOS estimates.*

Different methods of extracting phenology from remote sensing data could introduce uncertainties between different datasets (White et al., 2009; Jeong et al., 2011; Cong et al., 2013). Differences in filling missing data and filtering methods can affect the continuity of vegetation index time series, leading to discrepancies in phenology results. Reducing these inconsistencies among different datasets and producing more harmonized data is a primary motivation for employing the REA method, which reduces uncertainties by integrating datasets based on their reliability. Recognizing these factors also explains the discrepancies in SOS/EOS trends and stresses the need for standardized methodologies in phenological trend analyses.

4.4 Conclusion and Perspective

Shifts in vegetation phenology affect ecosystem structure (Kharouba et al., 2018; Yang and Rudolf, 2010), consequentially affecting biodiversity (Renner and Zohner, 2018), terrestrial carbon and water cycles (Piao et al., 2020), and the climate system (Green et al., 2017; Piao et al., 2020). The Establishment of a comprehensive and reliable vegetation phenology dataset is thus of critical importance. Our study demonstrates that the REA method provides a robust approach for integrating multi-source phenology datasets and produced a vegetation phenology dataset for regions above 30 degrees in the Northern Hemisphere from 1982 to 2022. This dataset could be used for subsequent analyses, such as examining vegetation phenology dynamics and their impacts on the terrestrial carbon cycle and water balance, and providing climatic feedback for global vegetation dynamics modeling. Integrating more vegetation phenology datasets that consider regional characteristics and refining the weight could be done in a future study, which could improve the accuracy and reliability of the merged phenology dataset. Additionally, higher-resolution phenology datasets will also improve the consistency between remote sensing phenology datasets and ground-based results. High-resolution datasets will enhance our ability to assess phenological changes in heterogeneous landscapes and improve local-scale ecological modeling, offering new opportunities to enhance monitoring and prediction of vegetation dynamics in response to environmental change.

Response to Reviewer #1:

[Comment 1] I appreciate the authors' edits in response to my and the other reviewer/editor's comments. The dataset and method descriptions are much improved, including how SOS and EOS are extracted for each dataset. Figure 2 is much easier to interpret now. I also appreciate the author's response about the different SOS/EOS extraction thresholds used across the different datasets. I understand that they are meant to indicate similar phenological stages, so can be combined, though I still have reservations about this method. I think this needs to be stated more clearly in the intro and/or methods so readers understand this potential source of error, and why it still works to combine datasets with differing extraction thresholds. However, given the improvement in the merged REA dataset over the initial 4 datasets, I do think this is an effective method to improve accuracy in extracting phenological transition dates.

Response: We sincerely appreciate the reviewer's constructive feedback and recognition of the improvements made to the dataset and method descriptions, as well as the enhanced clarity of Figure 2. Regarding the concern about combining datasets with different SOS/EOS extraction thresholds, we have revised the Introduction section to explicitly address this issue, please refer to the second paragraph in the introduction revision in the response to comment 2.

[Comment 2] The discussion section still needs considerable work. As I mention in the line edits below, the first couple paragraphs describing the different datasets and data fusion methods fit much better in the introduction and/or methods sections. The discussion section should provide interpretation for the observed results with examples from the existing literature. As it is, there is very little interpretation of the results and what they mean from an ecological perspective. Finally, the authors describe a "greenness trends" analysis at the end of the discussion section, which is not described in the methods section and I'm unsure of the interpretation. How was this analysis done and why? I see that the authors added this in response to my previous comment to "compare entire seasonal trends in vegetation greenness." I meant that it would be helpful to compare annual within-year variability in greenness across the different datasets rather than just the SOS/EOS dates. However, I think this could be saved for a separate paper. I apologize for the confusion.

Finally, the manuscript should generally be edited for clarity. I've suggested some edits below, but they are not exhaustive.

Response: Following the reviewer's suggestions, we have reconstructed the discussion section and moved the discussion of data fusion methods and the explanation of REA methods for data fusion of different threshold extraction methods to the introduction section.

The trend of Growing Season Greenness (GSG) is calculated based on the change in the mean Normalized Difference Vegetation Index (NDVI) values within the growing season (between the SOS and EOS events). The trend analysis method used in Figure S2 (now Figure S3) is the Mann-Kendall test. Greenness, which reflects vegetation growth conditions, is typically represented by NDVI (Myneni, 1997). The differences in GSG among datasets highlight the necessity of an integrated approach for robust phenology results. The specific calculation method for GSG is detailed in the response to Comment 56.

Meanwhile, we have supplemented the limitation discussion of the REA method and influencing factors of differences in the phenology trend, please refer to "4.2 Strengths, Limitations, and Sensitivity Analysis of the REA Approach" and "4.3 Consistency of SOS and EOS Trends Across Studies and Key Influencing Factors".

Introduction revision: (line 58-88)

Data fusion methods generally include unmixing-based, weight-function-based, and Bayesian-based approaches (Gevaert and García-Haro, 2015; Piao et al., 2019a). In vegetation phenology studies, fusion methods based on raw remote sensing data, such as the Spatial and Temporal Adaptive Reflectance Fusion Model (Gao et al., 2006) and the Enhanced Spatial and Temporal Adaptive Reflectance Fusion Model (Zhu et al., 2010), are often influenced by vegetation types, growth conditions, and methodological assumptions (Sisheber et al., 2022). These methods are typically applied to specific regions, and their performance can be affected by nonlinear spectral mixing, where the reflectance of vegetation endmembers (i.e., the pure spectral signatures of distinct land cover types) changes nonlinearly, and the spectral response of a single pixel is no longer a simple linear combination of the endmember spectra (Ma et al., 2015). The nonlinear combination of the ground feature can degrade the accuracy of vegetation phenology extraction. Unlike these approaches, the Reliability Ensemble Averaging (REA) method is not based on the assumption of linear reflectance changes. Instead, it directly merges annual phenology products based on their reliability. Compared to traditional data fusion methods, the REA method shows its advantages by simplicity and efficiency (Lu et al., 2021) while explicitly accounting for dataset reliability, in contrast to simple averaging methods that assume equal reliability across datasets. The simple averaging method treats all datasets equally, despite their uncertainties varying across time and space (Lu et al., 2021; Wang et al., 2019), which leads to potential inaccuracies in the result. The REA method considers the temporal consistency of vegetation phenology data and uses the voting principle (Giorgi and Mearns, 2002), which provides convergence while preserving spatial differences, making it suitable for multi-source data fusion.

Remote sensing vegetation phenology typically reflects key transition dates in the vegetation growth cycle, such as the start and end of the growing season, using various vegetation indices (e.g., NDVI, LAI, and SIF) (Cong et al., 2012; Piao et al., 2019b). Ideally, phenological dates extracted from different methods (thresholds, derivatives, smoothing functions, and fitted models) should accurately capture changes in actual physiological conditions (De Beurs and Henebry, 2010). However, in existing phenology datasets, there is often no "best" definition of these transition dates (White et al., 2009). The effectiveness of different extraction methods can vary across regions and periods, and may not always perfectly reflect true vegetation conditions (Cong et al., 2013; De Beurs and Henebry, 2010). For instance, in high-latitude areas, meaningful observations are relatively sparse. If the smoothing method removes too mach information, it may reduce the ability to extract phenological signals that accurately reflect surface dynamics (Wang et al., 2015). Therefore, integrating multiple datasets based on their reliability, rather than relying on a single dataset or using a simple averaging method, is a more robust approach. Our study stresses the value of integrating different datasets, and phenology results can be improved by applying the REA method, which assigns weights based on dataset reliability. This method can reduce uncertainties and provide a more accurate representation of phenological dynamics across different spatial and temporal scales.

Discussion revision: We have substantially revised the discussion section, please refer to the response to the editor's comment#3.

Major concerns:

Line edits:

[Comment 3] Line 16-17: "Improved substantially" compared to what? The 4 original data sets?

Response: We have revised it into "The evaluation using ground-based phenocam data from 280 sites indicated that the accuracy of the newly merged dataset was substantially improved compared to the four original datasets.". Please refer to lines 15-17.

[Comment 4] Line 18: put "with phenocam data" outside of the parentheses so readers immediately know what it's correlated with: "had the largest correlation with phenocam data"

Response: We have revised it into "The start of growing season and the end of growing season in the newly merged dataset had the largest correlation with phenocam data (0.84 and 0.71, respectively) and accuracy in terms of the root mean square error between phenocam data and merged datasets (12 and 17 d, respectively).". Please refer to lines 17-19.

[Comment 5] Line 20: Rather than saying an "advanced trend", I would put this into context, e.g., "the start of the growing season is occurring 0.24 days earlier per year." (Same for the end of the growing season result)

Response: We have revised it into "Using the new dataset, we found that the start of the growing season is occurring approximately 0.19 days earlier per year ($p < 0.01$), while the end of the growing season is occurring 0.18 days later per year ($p < 0.01$) over the period 1982-2020.". Please refer to lines 19-21.

[Comment 6] Line 25: Perhaps change "implications" to "impacts on"?

Response: We have revised it into "Global change has notably altered the timing of vegetation phenology (Ettinger et al., 2020; Zhang et al., 2022), leading to important impacts on the carbon and water cycles of terrestrial ecosystems.". Please refer to line 25.

[Comment 7] Line 28: Suggested re-wording: "there is large variation in spatiotemporal resolution"

Response: We have revised the sentence into "Various vegetation phenology datasets using remote sensing data have been produced, but inconsistencies and uncertainties arise when comparing these datasets with ground-based phenological observations, and there is large variation in spatiotemporal resolution.". Please refer to lines 26-28.

[Comment 8] Line 32: Perhaps start this sentence with "For example," and remove "as" after PhenoCam. Also, phenology cameras (e.g., PhenoCam) aren't really a ground-based measurement since the cameras are mounted above the canopies. It's usually considered a "near-surface" remote sensing method (see https://link.springer.com/chapter/10.1007/978-3-031-75027-4_20).

Response: We have revised the sentence into "For example, PhenoCam, a near-surface remote sensing method, has been operational for more than 20 years.". Please refer to lines 32-33.

[Comment 9] Line 36-37: This sentence is a little misleading because it implies that PhenoCam data isn't processed using standardized methods, but it actually is. All PhenoCam sites are processed the same way, allowing for consistency and comparability across locations and time periods.
Response: We have revised the sentence into "Additionally, remote sensing datasets and PhenoCam data, are processed using standardized methods that ensure consistency and comparability across different locations and periods.". Please refer to lines 36-37.

[Comment 10] Line 47: Start a new sentence here
Response: We have revised the sentence accordingly by starting a new sentence at this point. Please refer to line 47.

[Comment 11] Line 49-50: Briefly describe and/or provide an example of "different extraction methods". What are you extracting (I assume SOS and EOS?).
Response: We have revised it into "For estimates obtained using different extraction methods, such as varying algorithms or approaches for extracting SOS and EOS from the same satellite data, the discrepancies can exceed one month.". Please refer to lines 49-51.

[Comment 12] Line 50-51: Please explain what it means that the required NDVI phenology extraction threshold varies across biomes.
Response: We have revised it into "Additionally, the NDVI (Normalized Difference Vegetation Index) threshold required for phenology extraction varies across different biomes due to differences in vegetation types, growth patterns, and environmental conditions, which affect how NDVI values correspond to phenological events such as the start and end of the growing season.". Please refer to lines 51-54.

[Comment 13] Line 52-53: Please re-word this for clarity: "method which can choose the best dataset in different time and space among all input datasets"
Response: We have revised it into "Because it is difficult to determine the optimal dataset from the various phenology datasets, producing a merged dataset using a method that selects the most suitable dataset for different times and locations from all input datasets is essential for providing a comprehensive and accurate estimation of vegetation phenology with high spatiotemporal resolution.". Please refer to lines 54-57.

[Comment 14] Line 55: Please define the "simple averaging method", or could re-word to say something like "Calculating the mean value across different datasets…
Response: We have revised it into "The simple averaging method treats all datasets equally by calculating the mean value across different vegetation phenology datasets, despite their uncertainties varying across time and space". Please refer to lines 70-71.

[Comment 15] Line 57: I suggest replacing "whereas" with "but in reality"
Response: We have reconstructed this paragraph and modified the expression. Please refer to the third paragraph of the introduction section.

[Comment 16] Line 59-60: This is confusing. Please clarify the definition of the “vegetation index method”- please clarify. What are “the mathematical formulas”?

Response: We have reconstructed this paragraph and modified the expression, and have revised it into “Remote sensing vegetation phenology typically reflects key transition dates in the vegetation growth cycle, such as the start and end of the growing season, using various vegetation indices such as NDVI (Normalized Difference Vegetation Index) and EVI (Enhanced Vegetation Index) to assess vegetation conditions”. Please refer to lines 75-77.

[Comment 17] Line 60-61: I’m not sure what this sentence means. What “homogenous surfaces”?

Response: We have reconstructed this paragraph and modified the expression and removed the “homogenous” The word "homogeneous surfaces" refers to regions with relatively uniform vegetation cover. Please refer to the third paragraph of the introduction section.

[Comment 18] Line 67: Also provide spatial extent of new dataset here.

Response: The spatial resolution of the new dataset is 0.05°, with a temporal scale spanning 1982-2020, and it covers regions north of 30°N latitude. Please refer to lines 90-91.

[Comment 19] Line 85: Change “was” to “were”

Response: We have revised the sentence (line 107).

[Comment 20] Line 86: Can delete last sentence. Table 1 is already referenced above.

Response: We deleted the last sentence.

[Comment 21] Line 92-93: The definition of the “threshold method” is unclear. (e.g., “growing state of the vegetation as the time when…”). Maybe it would be helpful to include a figure showing this method, even if it’s just in the supplement. For example, see Fig 2 of this paper: https://www.nature.com/articles/sdata201828

Response: According to your suggestion, we have drawn a figure (Figure S1) of the threshold method to help readers better understand this phenological extraction method. Please refer to line 117.

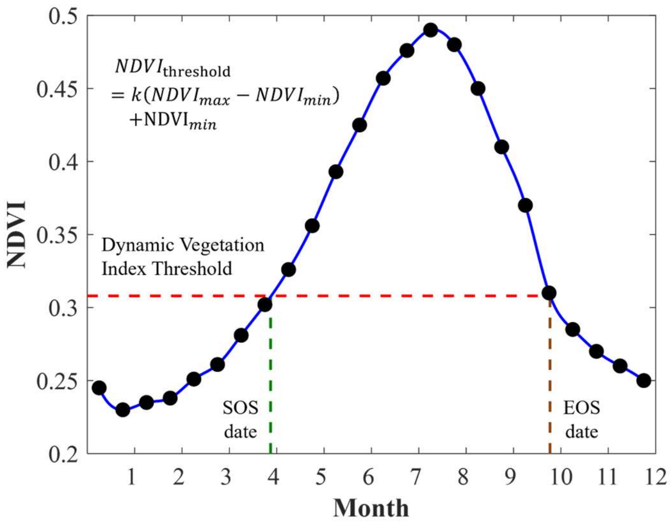

Figure S1: Diagram of vegetation phenology extraction using the threshold method. The red dotted line denotes the dynamic vegetation index threshold, which is calculated by multiplying the given threshold parameter and the maximum and minimum range of vegetation index. The green dotted line denotes the SOS (start of season) date, and the brown dotted line denotes the EOS (end of season) date.

[Comment 22] Line 95: EOS abbreviation has not been defined in the paper yet.
Response: We have defined the abbreviation of EOS, end of growing season, please refer to line 50.

[Comment 23] Line 104-105: "The filtering method based on confidence interval and operational continuity algorithm were used" - What does this mean?
Response: The filtering method, which uses confidence intervals and an operational continuity algorithm, was applied to reconstruct the time series curves. We revised this section as "which uses confidence intervals and an operational continuity algorithm" to make it clear that this is part of the filtering method, please refer to lines 128-129.

[Comment 24] Line 107: Here and elsewhere, does "segment" mean "annual" since each "segment" is growing season? Consider using annual or growing season instead for clarity.
Response: The start (end) of the season is defined using the modified Half-Max method (White et al., 2009) as the date when the NDVI2 time series first (last) crosses 35% of the growing season NDVI2 amplitude. In response to your suggestion, we have replaced "segment" with "growing season" to make it clear that the period refers to the growing season, please refer to line 131.

[Comment 25] Line 114-115: This is a great description of amplitude- I would move this to the first time you describe extracting SOS and EOS dates.
Response: We have moved the description of amplitude to the first time we describe extracting SOS and EOS dates, please refer to lines 117-119.

[Comment 26] Line 130: remove "i.e."

Response: We have removed "i.e.", please refer to line 155.

[Comment 27] Line 131: Add "the" before "PhenoCam Network"

Response: We have added "the" before "PhenoCam Network", please refer to line 156.

[Comment 28] Line 133: Start a new sentence for "For detailed information…"

Response: We have started a new sentence for "For detailed information…", please refer to line 157.

[Comment 29] Line 136: Add "A" before "spline interpolation"; add "the" before "phenocam data"

Response: We have added "A" before spline interpolation and added "the" before "phenocam data", please refer to line 161.

[Comment 30] Line 137: Define "ROI" and "GCC". Add "the" before PhenoCam Dataset v2.0"

Response: We have revised the sentence. "A spline interpolation method was applied to PhenoCam data to extract transition dates for each region of interest (ROI) mask using the green chromatic coordinate (GCC) in the PhenoCam Dataset v2.0." Please refer to lines 161-162.

[Comment 31] Line 140: Can you use "ROI" instead of "AOI" for consistency? Or are they fundamentally different?

Response: We have used "ROI" instead of "AOI" for consistency, please refer to line 165.

[Comment 32] Line 142: Derivative of what? GCC? Another greenness metric? (Also missing a closing parenthesis)

Response: The date of green-up each year was estimated as the day of year (DOY) corresponding to the maximum rate of increase in the 2G-RBi index. This is calculated as the maximum of the second derivative of the GCC time series (the maximum of the second derivative of GCC). We have also corrected the missing closing parenthesis in the revised manuscript, please refer to lines 168-169.

[Comment 33] Line 147-148: "no phenology values in four data sources"- what does this mean? Are the four sources the 4 remote sensing datasets? Does it mean that 90 phenocams weren't in this study's geographic range? Or is there another reason the remote sensing and phenocam data didn't overlap?

Response: The phrase "no phenology values in four data sources" refers to the 90 PhenoCam sites where the four remote sensing datasets (used in this study) did not provide corresponding phenology values. This discrepancy arises because these sites include cropland surfaces, which are often not covered by large-scale remote sensing phenology datasets, and sparse vegetation areas, where low-resolution remote sensing images struggle to detect phenological signals. We have revised it into "For use in this study, we excluded 26 sites that only provided one direction of transition dates (either SOS or EOS) and removed 90 sites where the four remote sensing datasets did not provide corresponding phenology values (due to factors such as cropland surfaces or sparse vegetation areas where low-resolution remote sensing struggles to detect phenological signals). We then selected PhenoCam data from 280 sites over the period 2000–2018, resulting in 1410 site–year combinations.", please refer to lines 173-177.

[Comment 34] Line 159: Be sure to specify up front that the weight of each dataset can shift through time.

Response: We thank the reviewer for their suggestion. To clarify, the weighting method was applied to derive more accurate Start of Season (SOS) and End of Season (EOS) dates from the four vegetation phenology datasets. The weight assigned to each dataset is not fixed but can vary over time, depending on the interannual variability of each dataset and the degree of consistency and offset among the four datasets. This has now been explicitly stated in the revised manuscript, "The weight assigned to each product was based on the interannual variability of each phenology dataset, as well as the degree of consistency and offset among the four datasets. Importantly, these weights can vary over time to reflect changes in dataset reliability and performance.", please refer to line 187-189.

[Comment 35] Line 162: what does "calculated in iterations" mean?

Response: The phrase "calculated in iterations" refers to the iterative process used to determine the final weight coefficients. Specifically, consistency is measured as the difference between each input dataset and the mean value of the four datasets, while the offset is measured as the difference between the REA result and each input dataset. These values are updated iteratively during the weighting process until convergence is achieved. We have revised it into "The consistency is measured as the difference between each input dataset and the mean value of the four datasets, and the offset is measured as the difference between the REA result and each input dataset. These values are iteratively calculated during the process of determining the final weight coefficients.", please refer to lines 189-191.

[Comment 36] Line 169-170: Add "d" to end of "generate". How can the REA method produce data that is "consistent with most of the input phenology products at the pixel level"? Isn't the idea that each of the 4 datasets are more or less accurate at different places and times? So how can the REA dataset be consistent with most/all of them?

Response: The REA method operates on the "voting principle," where the final result is generated by assigning different weights to the input data sources. This approach assumes that the majority of data values at each pixel are accurate, while outliers are down-weighted or excluded. As a result, the REA dataset aligns with the majority of input phenology products at the pixel level. However, it is important to note that the accuracy of the REA data varies across different locations and times. The method dynamically determines the weight of each dataset for specific places and times, ensuring that the most reliable data sources contribute more significantly to the final result. This flexibility allows the REA method to account for the varying accuracy of the input datasets across space and time. We have revised it into "The REA method, based on the 'voting principle' (where the REA result is generated by assigning different weights to the input data sources), produces data that aligns with the majority of input phenology products at the pixel level. This approach assumes that most data values at each pixel are accurate, while outliers are down-weighted or excluded.", please refer to lines 198-200.

[Comment 37] 196: “on” is a typo I assume- “more stringent “on” are required”

Response: We have revised this sentence, please refer to line 226.

[Comment 38] Line 204: What does “data” refer to? One dataset or data point?

Response: In this context, "data" refers to individual data points. If a data point shows significant discrepancies compared to others, it may indicate issues such as improper extraction methods or regional anomalies. We have revised it into “If an individual data point shows significant discrepancies compared to others, potentially caused by improper extraction methods in that region, the $B_{Phe,i}$ and $D_{Phe,i}$ will extract this variance and incorporate it, along with the natural variability ε_{Phe} of the region into the weight distribution process” , please refer to line 234-236.

[Comment 39] Line 230: I would just write out Mann-Kendall I(instead of M-K) since it’s not used much in the manuscript.

Response: We have revised it from M-K to Mann-Kendall, please refer to line 261.

[Comment 40] Line 249: Replace “lower” with “earlier”

Response: We have replaced it from lower to earlier, please refer to line 287.

[Comment 41] Line 304: This is a little confusing. Of course EOS dates are going to be later than SOS dates. By saying “Unlike the SOS data”, are you trying to point out that EOS dates are more uniform in timing that SOS? Consider re-wording to make this clearer.

Response: We have reworded the sentence for clarity. The revised text please refer to lines 342-343."Unlike the Start of Season (SOS) dates, which exhibit greater variability, the End of Season (EOS) dates are more consistent, predominantly occurring within day of year (DOY) 270–330 (80.0%)."

[Comment 42] Line 312: This is unclear: “regarding the coefficient of variation (CV) in the uncertainty range of”

Response: We thank the reviewer for pointing out the lack of clarity in this sentence. To address this, we have revised the text as follows:

"In Fig. 4(c, f), the coefficient of variation (CV) of the uncertainty in Start of Season (SOS) and End of Season (EOS) dates from 1982 to 2020 is analyzed. More than 56% (73%) of regions exhibit a CV below 1, 31% (18%) of regions have a CV between 1 and 1.5, and only 13% (8%) of regions show a CV higher than 1.5. No evident correlation is observed between CV and latitude changes."

Additionally, we have adjusted minor punctuation and wording to improve clarity and readability, please refer to lines 351-354.

[Comment 43] Line 330: First time CORR is used, please define.

Response: We have defined CORR as correlation coefficient, please refer to line 370.

[Comment 44] Line 336: What does “overestimate” mean (used 2 times)- predict dates that are too late?

Response: In this context, "overestimate" means that the GIM_4g dataset predicts End of Season (EOS) dates that are later than the observed dates. Specifically, the GIM_4g dataset demonstrates good performance but tends to overestimate EOS, resulting in a Root Mean Square Error (RMSE) of 43 days. We have revised it to “The GIM_4g dataset shows good performance but tends to overestimate the End of Season (EOS), with predicted dates occurring later than observed, resulting in an RMSE of 43 d. Both the VIP and the GIM_3g datasets also overestimate the EOS due to their spatial and temporal distributions, with RMSEs of 46 and 35 d, respectively.” Please refer to lines 375-378.

[Comment 45] Line 351: I suggest adding this to the end of the first sentence: “across years at a single location.”

Response: As recommended, we have revised the sentence to: "We selected a long-term PhenoCam site (Morganmonroe) from PhenoCam to evaluate the merged dataset across years at a single location."

[Comment 46] Line 352-353: suggested wording: “with data from 2010-2020 for both SOS and EOS”

Response: As recommended, we have revised the sentence to: "We have chosen a US. PhenoCam site characterized by deciduous broad-leaved forest, with data from 2010–2020 for both SOS and EOS." Please refer to lines 393-394.

[Comment 47] Line 354: Largest compared to what? The original 4 datasets?

Response: As illustrated in the time series plot in Figure S4, the consistency between the REA results and PhenoCam data for both SOS and EOS is the highest when compared to the original four individual datasets. Please refer to lines 395-396.

[Comment 48] Line 365: I suggest starting a new paragraph at “Over the study area”

Response: We started a new paragraph at “Over the study area”. (line 407)

[Comment 49] Lines 375- 421: This is all background information that belongs in the introduction section. For the discussion section, I suggest starting with a 3-4 sentence recap of the study’s motivations/methods, and then spend the rest of the discussion interpreting the results and linking them to the literature.

Response: Thank you for your suggestion. We have revised the discussion section by moving the background information to the introduction, and restructuring the discussion section into four subsections 4.1 Integrating Multi-Source Phenology Data and Addressing Dataset Discrepancies, 4.2 Strengths, Limitations, and Sensitivity Analysis of the REA Approach, 4.3 Consistency of SOS and EOS Trends Across Studies and Key Influencing Factors 4.4 Conclusion and Perspective.

Following your suggestions, we have included a brief summary of the study’s motivations and methodology at the beginning of the discussion. Additionally, we have expanded our discussion on the limitations of the REA method, the factors influencing differences in phenological trends across studies, and other key considerations. These revisions ensure a clearer and more structured discussion while addressing the concerns raised.

Please refer to comment 2.

[Comment 50] Lines 382-386: I don’t think this is needed.
Response: We have reconstructed the discussion, please refer to comment 2.

[Comment 51] Line 402: What is the “final result”?
Response: The "final result" refers to the outcome of the Reliability Ensemble Averaging (REA) method, which integrates multiple datasets to produce a more robust and reliable estimate. This has now been clarified in the revised manuscript.

[Comment 52] Line 411: Thanks for adding this explanation, however, please specify what “endmember” means.
Response: We appreciate the reviewer's comment and have clarified the term "endmember" in the revised manuscript. Endmembers refer to the pure spectral signatures of distinct land cover types. In this context, the reflectance of vegetation endmembers changes nonlinearly. Due to the spatial mixing of spectral signals from different land covers, the spectral response of a single pixel is no longer a simple linear combination of the endmember spectra. Please refer to line 63.

[Comment 53] Line 416: What should be available? Multiple years of data?
Response: There is no restriction on the minimum length of the time series, but multiple years of data should be available to capture natural variability and maintain accuracy. We have added a discussion of method limitation in 4.2 Strengths, Limitations, and Sensitivity Analysis of the REA Approach. Please refer to comment 2.

[Comment 54] Line 430: I suggest replacing “in details” with “for example”. Also, how did you determine that SOS “significantly advanced”? Was a statistical method used? If so, please provide p-value. If not, please re-word.
Response: As recommended, we have replaced "in details" with "for example" in Line 430, and the p-value is smaller than 0.01. We have stated in the revised manuscript “Our REA-based dataset estimates an SOS advancement of 0.19 days per year during 1982-2020 ($p<0.01$)” and “Similarly, EOS has been delayed at a rate of 0.18 days per year ($p<0.01$)”. Please refer to lines 475 and 477.

[Comment 55] Line 432: Why do you think there is such large variation in the rate of SOS advance between studies. Speculate on some possible reasons (with citations).
Response: In Discussion 4.3 Consistency of SOS and EOS Trends Across Studies and Key Influencing Factors, we have expanded our discussion to address the reasons for the observed variations in the rate of Start of Season (SOS) advancement across different studies.

Global climate change has significantly influenced vegetation phenology, with a general advancement in SOS and a delay in EOS (Piao et al., 2019a). Our REA-based dataset estimates an SOS advancement of 0.19 days per year during 1982-2020 ($p<0.01$), aligning with previous findings, such as 1.4 ± 0.6 days per decade for 1982-2011 (Wang et al., 2015) and a 5.4-day advance from 1982 to 2008 (Jeong et al., 2011). Similarly, EOS has been delayed at a rate of 0.18 days per year ($p<0.01$), consistent with prior estimates of 0.18 ± 0.38 days per year for 1982–2011 (Liu et al., 2016) and a 6.6-day delay from 1982 to 2008 (Jeong et al., 2011). Deviations between REA-based trends and those from individual datasets (Blunden et al., 2023) suggest that variations in data sources and study regions contribute to observed discrepancies.

Differences in SOS trends across studies arise from multiple factors, including study period, land cover changes, spatial domain, environmental heterogeneity, and methodological differences. Trend estimates are highly sensitive to the selected time window, particularly at the start and end years (Cong et al., 2013). Interannual and decadal climate variability can further influence these trends contributing to discrepancies across studies. Land cover changes, such as wildfires and deforestation (Jeong et al., 2011), can influence the extraction result of phenology, then affect SOS trends. Additionally, spatial differences in studies also influence the rate of trends in different studies, as phenology is affected by vegetation type and climate sensitivity. For instance, Jeong et al. (2011) analyzed temperate vegetation between 30°N–80°N, and Wang et al. (2015) focused on 30°N–75°N, excluding evergreen forests and managed landscapes. Since phenology responds differently across species and locations (Maignan et al., 2008), variations in study areas contribute to discrepancies. Photoperiod constraints introduce additional latitudinal differences in phenological responses to warming (Meng et al., 2021). Moreover, precipitation and other environmental factors vary spatially, further influencing SOS estimates.

Different methods of extracting phenology from remote sensing data could introduce uncertainties between different datasets (White et al., 2009; Jeong et al., 2011; Cong et al., 2013). Differences in filling missing data and filtering methods can affect the continuity of the vegetation index time series, leading to discrepancies in phenology results. Reducing these inconsistencies among different datasets and producing more harmonized data is a primary motivation for employing the REA method, which reduces uncertainties by integrating datasets based on their reliability. Recognizing these factors also explains the discrepancies in SOS/EOS trends and stresses the need for standardized methodologies in phenological trend analyses.

[Comment 56] Line 437- 441: I'm not sure what this analysis is. What are "seasonal trends in greenness"? This analysis needs to be described in the methods, and reported in the results. New results are generally not introduced in the discussion section. Also what are the units of the numbers reported? What is the ecological interpretation of this analysis?
Response: We thank the reviewer for raising this important point. The "seasonal trends in greenness" refer to the trends in vegetation greenness during the growing season, which are quantified by analyzing the changes in the mean Normalized Difference Vegetation Index (NDVI) values between the Start of Season (SOS) and End of Season (EOS) events. The trend analysis method employed here is the Mann-Kendall test, consistent with the approach used in Figure S2 (now Figure S3).

Greenness is a widely used indicator of vegetation growth, typically represented by NDVI (Myneni, 1997). The Growing Season Greenness (GSG) is calculated as the mean NDVI value within the growing season in this study, defined by the period between SOS and EOS:

$$GSG = Mean(NDVI[Date_{SOS}, Date_{EOS}]) \tag{16}$$

$Date_{SOS}$ is the day of year value of SOS data, and $Date_{EOS}$ is the day of year value of EOS value, $Mean$ denotes calculating the mean NDVI value in the corresponding date range.

To address the reviewer's concerns, we have now included a detailed description of this analysis in the Methods section (please refer to lines 271-277). Additionally, we have clarified the units and ecological interpretation of the analysis. Specifically, the GSG values are unitless, as NDVI is a dimensionless index, and the trends in GSG reflect changes in vegetation productivity and health over the growing season.

[Comment 57] Line 441: Start new paragraph with “Shifts in”

Response: We have reconstructed the discussion section, please refer to line 500.

Figures:

[Comment 58] Figure 1: In panel a, the shades of green are hard to distinguish. Maybe consider switching to a color palette with a larger color range.

Response: In response to this comment, we have modified the colormap of the figure. This adjustment significantly improves the distinguishability of the shades, thereby enhancing the overall clarity and interpretability of the figure. (line 290)

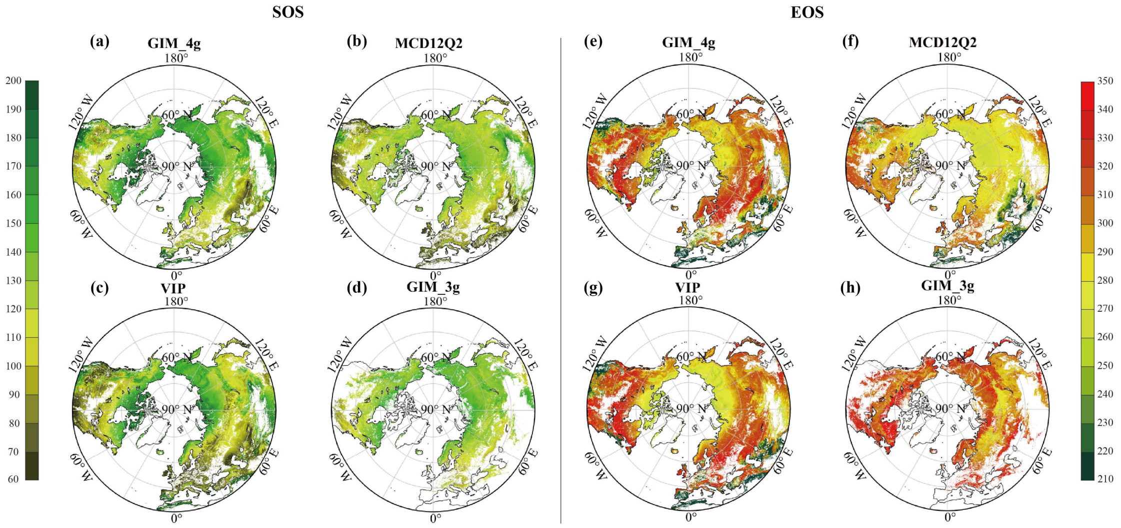

Figure 1: Spatial distribution of multiyear mean SOS and EOS dates from each phenology dataset: (a-d) multiyear mean SOS dates and (e-h) multiyear mean EOS dates derived from the GIM_4g, MCD12Q2, VIP, and GIM_3g datasets, respectively.

[Comment 59] Figure 2 legend: For panels b and d, are the latitudinal weights averaged over the entire time period of the REA dataset? Please state this.

Response: We have clarified this in the figure legend by stating: “For panels b and d, the latitudinal weights are the mean values of each dataset over their respective timespans.” (line 318)

[Comment 60] Figure 4: Consider a color scale with more variation for panel d. For panels c and f, I suggest listing the legend values in increasing order from left to right (instead of in columns) for easier interpretation.

Response: Thank you for your suggestions on the modification of Figure 4. We have adjusted the color scale in panel (d) to enhance variation for better visual distinction. Additionally, we have modified the legend in panels (c) and (f) by listing values in increasing order from left to right for easier interpretation. (line 355)

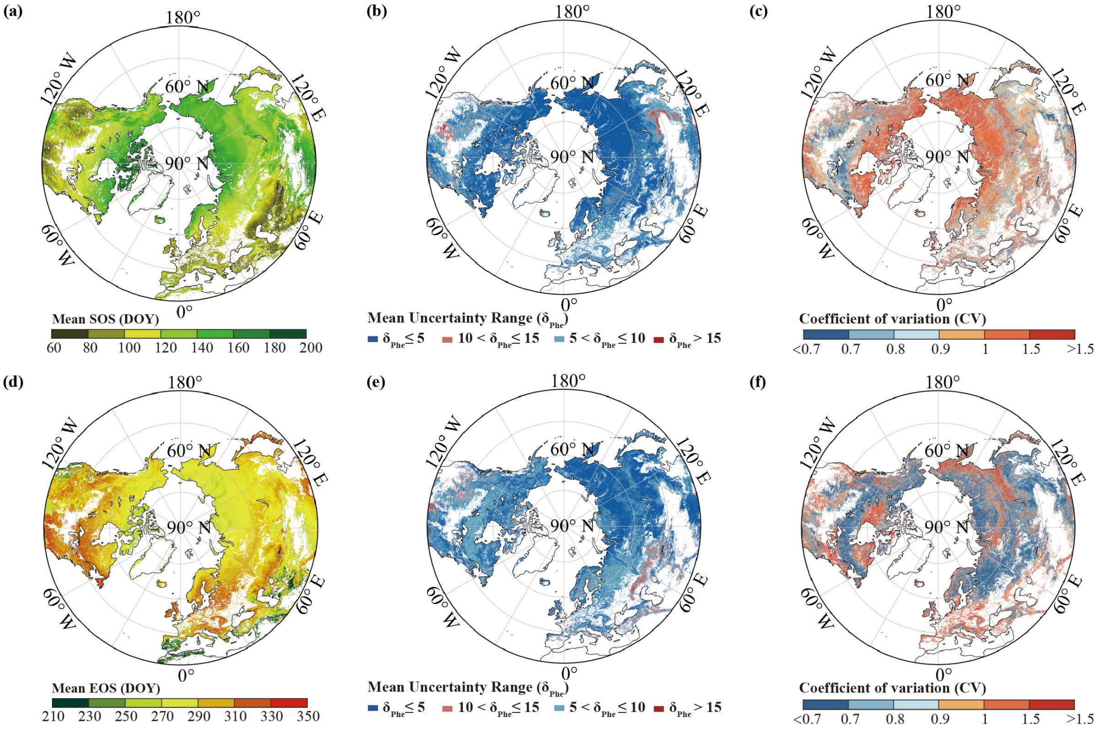

Figure 4: Merged mean (a) SOS and (d) EOS dates (DOY) obtained using the REA method for the period 1982-2020 and the uncertainty in the REA merged data. Mean uncertainty (δ_{Phe}) of SOS dates (b) and EOS (e) obtained using the REA method for the period 1982-2020, and its coefficient of variation (CV)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

[Comment 61] Figure 5 legend: Remove “respectively”. Please briefly describe what a radar chart is.

Response: We have removed “respectively” in the legend. Additionally, we have briefly described the radar chart: "The radar chart is a graphical method used to display multivariate data in the form of a two-dimensional chart with axes starting from the same point. (line 391)

[Comment 62] Figure 6 legend: I recommend using “trendlines” instead of “fitting lines”. For the last sentence, specify that these abbreviations are used in the insets of panels b and d.

Response: We have replaced “fitting lines” with “trendlines”. We also have specified that these abbreviations are used in the insets of panels b and d. (line 414)

The abbreviations DS (significant delay), DN (non-significant delay), AS (significant advance), and AN (non-significant advance) are used in the insets of panels b and d. (lines 415-416)

[Comment 63] Figure S1 legend: RMSE between these datasets and what? The phenocam dataset?

Response: We have stated in the legend "RMSE between the four remote sensing datasets and the PhenoCam dataset for SOS across different PFTs during 2001-2018". (line 10-14 in Supplement material)

[Comment 64] Figure S2: What does the "NDVI trend" measure? In b, what does the frequency distribution show? I'm not clear on what this analysis is.

Response: We have revised the legend for the previous S2 (now S3) "Figure S3: Trends in growing season NDVI derived from (a) the VIP phenology dataset, (b) the GIM_4g dataset, and (c) the REA phenology dataset for the period 1982-2015. Panel (d) presents the frequency distribution of the Sen's slope for the growing season NDVI trend in regions north of 30°N. GS: growing season. Black dots indicate regions where the trend is statistically significant (P < 0.05)." (line 20-224 in Supplement material)

Panel (d) illustrates the frequency distribution of the Sen's slope for the growing season NDVI trend in regions north of 30°N, allowing for a comparative analysis of greenness trends across different datasets.

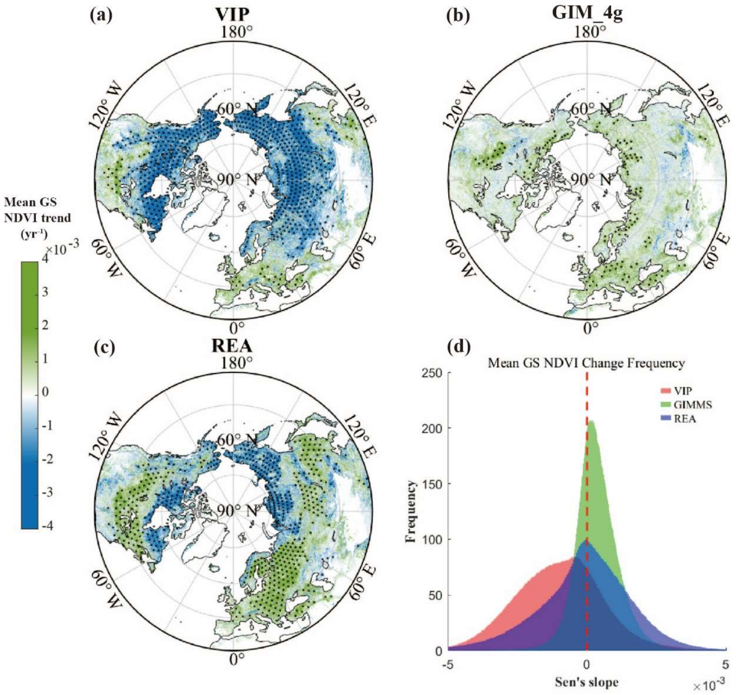

Figure S3: Trends in growing season NDVI derived from (a) the VIP phenology dataset, (b) the GIM_4g dataset, and (c) the REA phenology dataset for the period 1982 – 2015. Panel (d) presents the frequency distribution of the Sen's slope for the growing season NDVI trend in regions north of 30° N. GS: growing season. Black dots indicate regions where the trend is statistically significant (P < 0.05).

[Comment 65] Figure S3 legend: Specify that the simple average is across all 4 datasets.

Response: We revised the legend as "Mean (a) SOS and (b) EOS dates (DOY) obtained using the simple average across all four datasets for the period 1982-2020. RMSE represents the root mean square error (days), and CORR represents the correlation coefficient. The four datasets include MCD12Q2, VIP, GIM_4g, and GIM_3g. For further details on these datasets, please refer to Section 2.1.”. (line 26-30 in Supplement material)

[Comment 66] Figure S4: Specify what that dotted lines indicate (trendlines for each dataset I assume). Specify that “morganmonroe” is a PhenoCam site and add more site details (location? Vegetation type?). For panel a, the green dataset looks like it has a short dashed trendline for the early years, but no data points?

Response: We have added “The dotted lines in a and c represent the trendlines for each dataset. The Morganmonroe site, located in Morgan Monroe State Forest, Indiana, is characterized by deciduous broadleaf forest.” in the legend. (line 31-37 in Supplement material)

In panel (a), the green dataset's data points overlap with the REA data, which is why they may not be clearly visible.

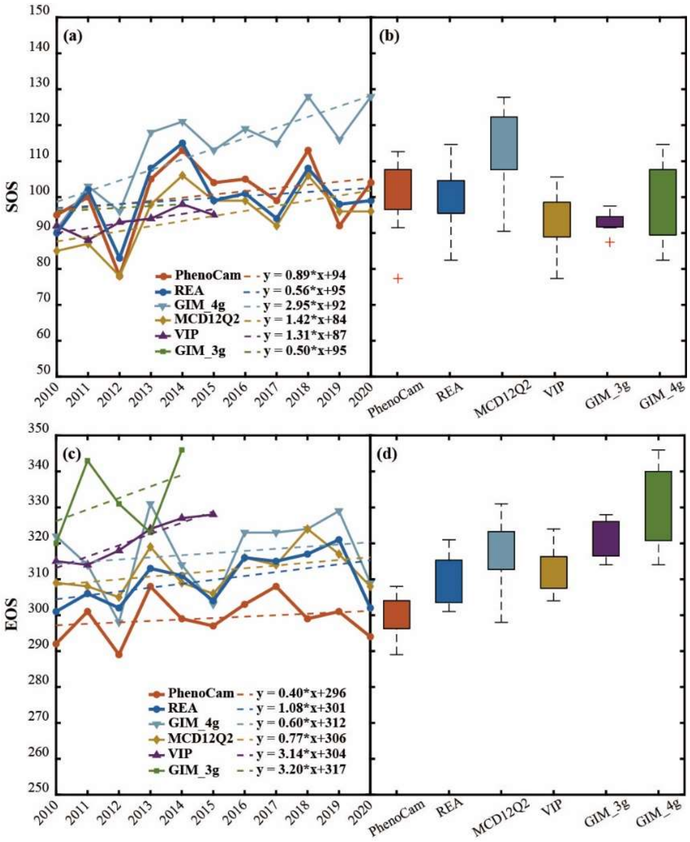

Figure S5: Time series and box plots of Morganmonroe site data with each phenology dataset and the merged phenology dataset obtained using the REA method. (a-b) SOS time series and box plot of the PhenoCam, GIM_4g, MCD12Q2, VIP, GIM_3g, and REA datasets, respectively, (c-d) EOS time series and box plot of the PhenoCam, GIM_4g, MCD12Q2, VIP, GIM_3g, and REA datasets, respectively. The dotted lines in a and c represent the trendlines for each dataset. The Morganmonroe site, located in Morgan Monroe State Forest, Indiana, is characterized by deciduous broadleaf forest.

Response to Reviewer #2:

[Comment 1] Thank you for the detailed point-by-point response. I think most of my comments have been addressed.

Response: Thank you for your feedback. We appreciate your constructive comments and are glad that most of your concerns have been addressed.

[Comment 2] I checked the GitHub codes and feel that it lacks sufficient comments or guidance to ensure reproducibility.

Response: Following your suggestion, we have added comments in the GitHub to ensure the reproducibility of the codes.

[Comment 3] Regarding the response to comments 12 and 18, I recommend that the authors include a more detailed discussion of the limitations of the REA method.

Response: Thank you for your suggestion, we have added more details about the limitations of our method in the discussion section 4.2 Strengths, Limitations, and Sensitivity Analysis of the REA Approach.

Given the lack of consensus on the "best" approach for extracting phenological dates, our study highlights the value of integrating multiple methodologies rather than relying on a single dataset. The REA method improves phenology estimations by weighting datasets based on reliability rather than assuming equal contributions, as in a simple arithmetic average (Lu et al., 2021). The REA method considers the temporal correlation of vegetation phenology data by employing a voting principle (Giorgi and Mearns, 2002), and this approach facilitates convergence of data while retaining differences of the spatial distribution, thereby offering advantages in multisource data fusion than simple averaging. Our results indicate that REA merged phenology results are more consistent with PhenoCam observations than individual datasets and simple average results, demonstrating its potential to reduce biases between different datasets caused by different extraction methods. However, while REA enhances consistency, it does not fully eliminate methodological discrepancies.

To evaluate the robustness of this approach, we conducted sensitivity analyses using different dataset and time span combinations. Our results (see Figures S6 and S7) indicate that the combination of different input datasets has a greater influence on fusion results than the length of the time series. The average deviation between different dataset combinations and the long-term four-dataset fusion results is 7.1 days, whereas SOS and EOS variations across different time spans (e.g., 5-year vs. 10-year fusion) are relatively minor (4.1 and 3.2 days, see Figure S6). Moreover, fusion weights derived from shorter time series are largely consistent with those obtained using the full dataset, see Figure S8. In the multi-time span analysis, the fusion results of SOS are more stable than EOS, and the longer time series result is more similar to the REA fusion result (1982-2022).

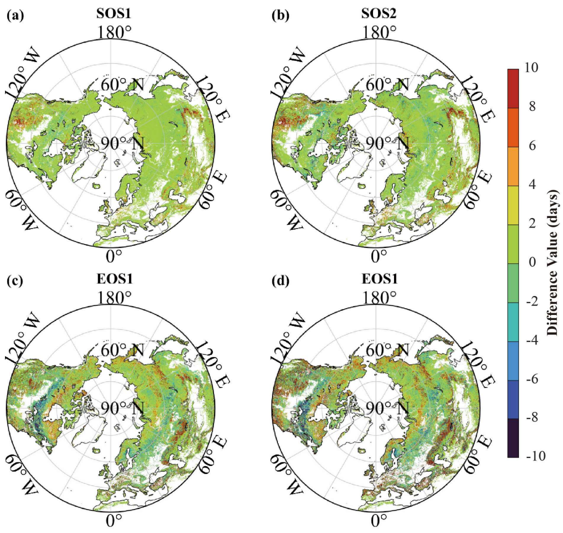

Figure S6: Comparison of fusion results of four groups of data at different time lengths from 2001 to 2010. (a, c) Difference plots between the fusion results of SOS and EOS for the four datasets over the 2001-2010 period and the original long-term REA fusion results. (b, d) Difference plots between the fusion results of SOS and EOS obtained by separately fusing the two sub-periods of data (2001-2005 and 2006-2010) and subsequently concatenating them, and the original long-term REA fusion results.

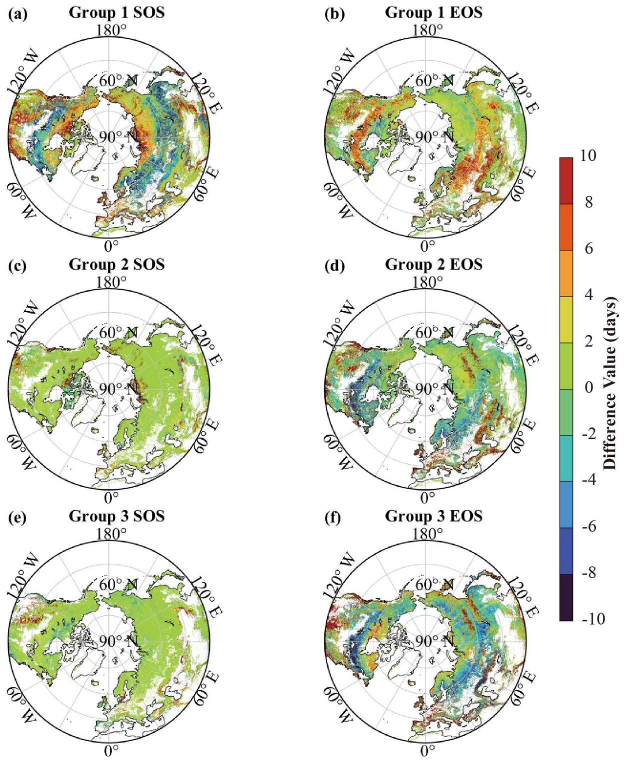

Figure S7: Differences between the SOS and EOS data fusion results from various data source combinations and the fusion results of the long-term four datasets over the 2001-2010 period. (a-b) Group 1 represents the difference between the fusion of GIM_4g and VIP data and four datasets REA result. (c-d) Group 2 represents the difference between the fusion of VIP data and GIM_3g data and four datasets REA result. (e-f) Group 3 represents the difference between the fusion of GIM_3g and MCD12Q2 data and four datasets REA result.

The effectiveness of data fusion depends on the complementarity and quality of input datasets. Our analysis suggests that datasets incorporating more reliable sources could produce improved fusion results. For example, GIM_3g SOS demonstrate higher consistency with PhenoCam observations (Figure 5d); groups 2 and 3, which include GIM_3g SOS, exhibit smaller deviations from the long-term REA fusion results than group 1. In contrast, EOS estimates are more stable across different dataset combinations. These results show that multi-source data integration could help improve accuracy. The similarity in weight distributions across different dataset combinations (Figure S9) suggests that dataset selection remains a critical factor influencing final estimates. Incorporating additional high-quality datasets in future studies could further enhance accuracy and robustness.

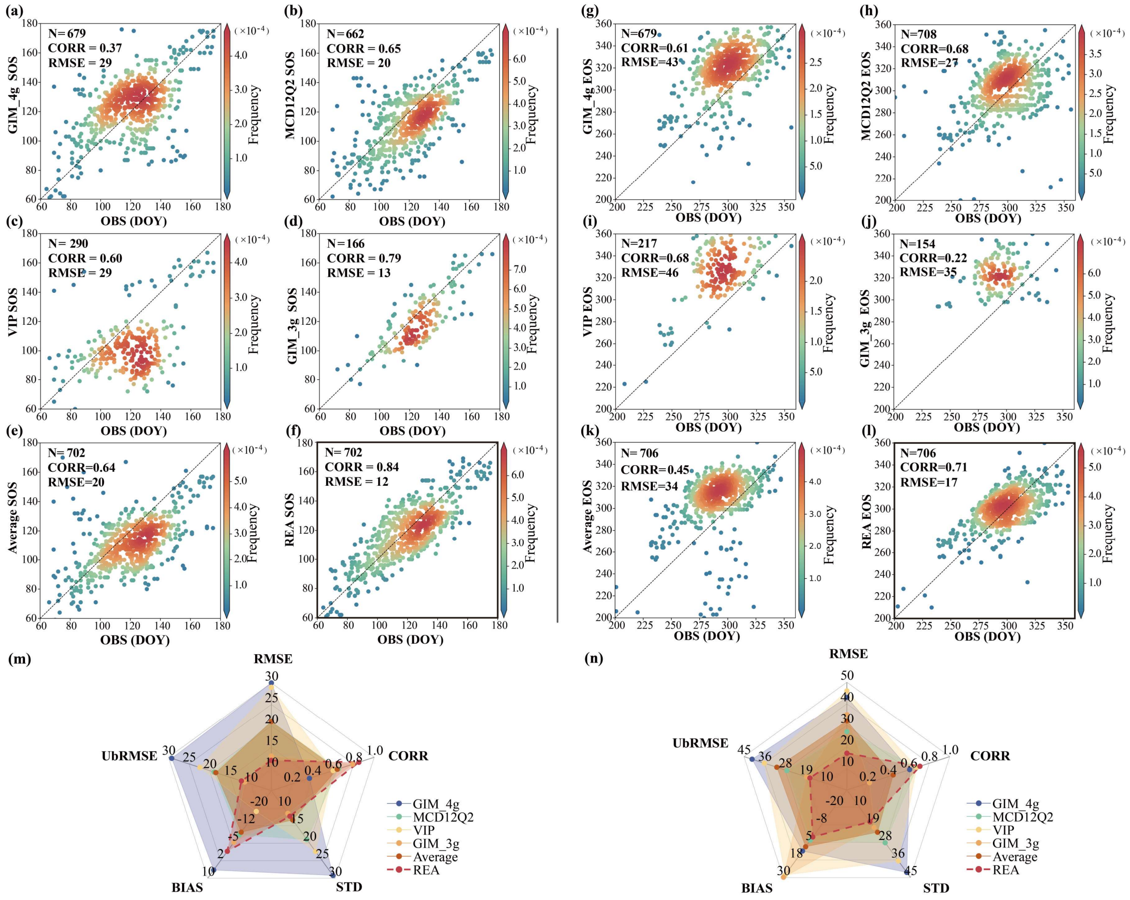

Figure 5: Scatterplots and radar charts of performance for each phenology dataset and the merged phenology dataset obtained using the REA method. (a–f) SOS evaluation results of the GIM_4g, MCD12Q2, VIP, GIM_3g, Average, and REA datasets, respectively, (m) radar chart of the SOS evaluation results, (g–l)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

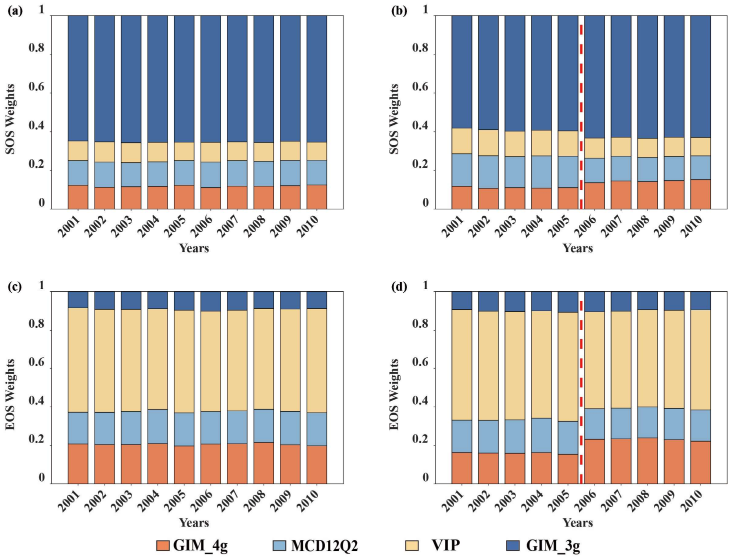

Figure S8: Comparison of fusion data weights for the four datasets over different time lengths from 2001 to 2010. (a, c) Weights of the SOS and EOS fusion data for the four datasets over the entire 2001-2010 period. (b, d) Weights of the fusion data after separately fusing the two sub-periods of data (2001-2005 and 2006-2010) and subsequently concatenating them. The red dashed line separates the two sub-periods.

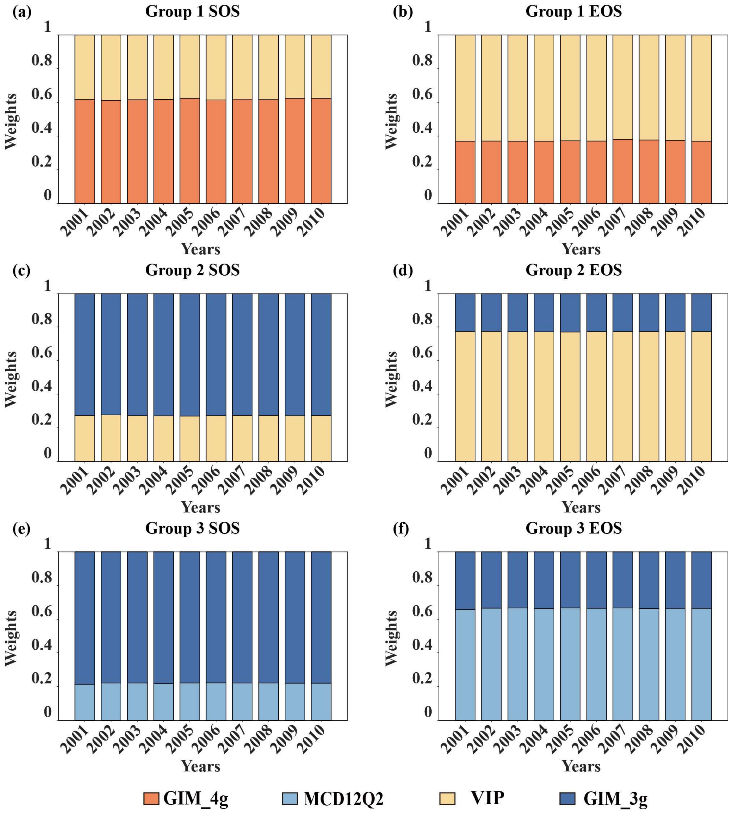

Figure S9: Distribution of fusion weights for SOS and EOS data from different data source combinations over the 2001-2010 period. (a, b) Group 1: Fusion weights for the combination of GIM_4g and VIP data. (c, d) Group 2: Fusion weights for the combination of VIP data and GIM_3g data. (e, f) Group 3: Fusion weights for the combination of GIM_3g and MCD12Q2 data.

[Comment 4] Additionally, were all four datasets available all the time across 1982-2020 to be merged for the new dataset? Would it be possible to assign a quality flag to the new dataset?

Response: Thank you for your constructive suggestions. Based on the number of datasets involved in data fusion at different times, we assigned a quality flag (1-4, Data confidence from low to high) and uploaded it to the data website to enhance the application of the data.