

Reviewer #2

The paper presents a very simple phenology model that is only based on temperature. They calibrate the model on modeled phenophases based on modis observations and then calculate start and end of the growing season based on the temperature simulated by the UKESM. They compare there model to two very simple semi-parametric phenology models both of which were not developed for global coverage. The presented model does not clearly outperform these models.

We thank the reviewer for the thoughtful criticisms. We revised the paper and further justified our methodologies following the highlighted concerns. Here, a summary of the main changes may be found, while detailed responses to each of the raised points may be found below.

Responses are in red. Cited parts of the revised article are in green.

The main changes are:

1. We reframed the introduction goals. In the current form, the main goal is to offer a product that can surpass others in term of spatial extent, species-specificity and robustness under climate change. At the same time, we clarify that this product, constituting a primarily data-based, long-term reconstruction can support both impact studies and the benchmarking of terrestrial ecosystem models. To better support the redefined goals, we are also now considering downscaled, bias-corrected temperature fields developed within ISIMIP3b to infer phenological stages.

Specifically, for the introduction goals are now thw following:

The primary task of this study is to determine the Start and End of the Growing Season (SGS and EGS) for generic forest ecosystems at a global scale, using a novel probabilistic, temperature-based model (GS-P) developed within a machine-learning framework. To assess the model's performance and robustness, we conduct a comparative analysis against established deterministic models and a latitudinal benchmark (GS-Lat), using ERA5 reanalysis and MODIS remote-sensing data for calibration and validation. The ultimate outcome is the creation of an open-source dataset that reconstructs, with quantified uncertainty, the trajectories of SGS and EGS from 1850 to 2100 under different climate change scenarios. These continuous estimates of phenological transitions, spanning from preindustrial to extreme climate warming periods, may support future research and applications. Particularly, because our framework is primarily data-driven, the resulting product can provide a long-term, relatively independent benchmark for terrestrial ecosystem and vegetation models. This study also aims to overcome previous model limitations in spatial extent, species specificity, and robustness under climate change scenarios, while providing, at the same time, a generalizable framework for prediction of phenological stages.

2. The four considered models were trained and validated again using a newly defined dataset split following the suggestion of the reviewer:

The MODIS product is used for calibration and validation of the models, hence linking the SGS and EGS inferred from EVI-2 (phenological season) with temperatures (meteorological season) from ERA5. Of all available years (2001-2023), the last five years (2019-2023) are withheld for validation, while the rest (2001-2018) are used for calibration. Moreover, to further validate the models spatially, the globe is divided into 12 uniform blocks (four longitudinal and three latitudinal splits), among which four are randomly withheld for validation every six years.

The entire manuscript has been updated accordingly to the new results.

3. We added photoperiod to the model predictors to account for an important seasonal constraint on phenological transitions that was absent from the original GS-P formulation. This feature has broadly improved GS-P performances. GS-P now outperforms all other models globally and in most climate classes and performance metrics (see updated Table S3 below).

All sections have been updated accordingly to the new results.

We acknowledge that one of the reviewer's main requests was related to the direct comparison with more sophisticated models. In this regard, we agree that a direct comparison with other machine-learning based approaches would be ideal. While advisable, such comparison is hardly feasible mostly because other machine learning-based studies often do not provide the associated code (Sun et al. 2025), or would require to retrain the model from scratch due to the different spatial coverage (Garnot et al. 2025), which, with a dataset as large as ours, would be too resource and time demanding. Despite this, we now clarify that, while being relatively simple, GS-BC2 has been found not only to outperform traditional models (Mo et al. 2023), but also to perform just slightly worse than Machine Learning frameworks (Garnot et al. 2025). As such, benchmarking against that model is a suitable choice for our study. More details in this regard may be found in the replies below.

Table S3. MB (Model – MODIS), MAE, and percentage of failed instances in predicting SGS and EGS over each KG class, calculated over the validation dataset. The metrics reported here are adjusted similarly to the loss function (Eq. A2) using the percentage of failed instances: MAE and MB are multiplied by the penalization factor, while ρ is divided.

KG classification		C*a	C*b ∨ C*c	D*a	D*b	D*c	D*d	Global
Validation dataset size		133,961	76,653	37,512	240,019	328,195	7,176	823,516
Metric	Model							
SGS								
Failed pred (%)	GS-Lin2	18.9	20.2	0.0	0.2	0.4	0.0	5.2
	GS-BC2	0.0	0.0	0.0	0.0	0.1	0.0	0.0
	GS-P	0.3	0.4	0.0	0.0	0.1	0.0	0.1
	GS-Lat	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MB (d)	GS-Lin2	65.3	60.5	7.7	9.8	0.6	1.2	14.7
	GS-BC2	1.8	-0.1	0.1	2.0	-1.6	0.6	0.2
	GS-P	-15.9	-11.0	0.1	0.4	-3.5	-1.1	-4.9
	GS-Lat	-3.8	22.3	3.2	12.7	-10.4	-3.6	1.1
MAE (d)	GS-Lin2	82.2	83.8	12.9	13.9	11.6	6.1	24.7
	GS-BC2	16.6	21.5	5.9	7.5	10.4	3.2	11.3
	GS-P	21.6	20.4	5.4	6.0	7.5	3.1	10.4
	GS-Lat	19.0	34.9	9.2	15.7	13.4	5.0	16.7
ρ	GS-Lin2	-0.15	-0.01	0.57	0.51	0.31	0.67	0.44
	GS-BC2	-0.29	0.37	0.64	0.66	0.38	0.62	0.87
	GS-P	0.31	0.68	0.76	0.79	0.79	0.72	0.91
	GS-Lat	-0.15	0.47	0.2	0.22	0.54	0.58	0.76
KG classification		C*a	C*b ∨ C*c	D*a	D*b	D*c	D*d	Global
Validation dataset size		97,204	56,081	25,105	168,072	230,479	4,960	581,901
Metric	Model							
EGS								
Failed Pred (%)	GS-Lin2	18.7	20.1	0.0	0.2	0.4	0.0	5.1
	GS-BC2	1.5	0.0	0.0	0.0	0.0	0.0	0.2
	GS-P	0.3	0.4	0.0	0.0	0.1	0.0	0.1
	GS-Lat	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MB (d)	GS-Lin2	-3.4	-17.9	-0.3	0.4	11.5	4.9	4.0
	GS-BC2	1.9	-7.3	-0.3	-1.5	1.8	2.1	-0.1
	GS-P	15.9	12.2	-0.1	0.9	1.7	1.8	4.6
	GS-Lat	-2.4	-11.6	-13.7	-16.1	13.9	17.6	-1.1
MAE (d)	GS-Lin2	37.8	45.7	11.6	11.7	15.4	7.3	18.6
	GS-BC2	29.1	24.8	6.7	7.7	8.3	3.3	12.8
	GS-P	26.4	23.0	5.8	6.2	5.2	3.1	10.5
	GS-Lat	23.6	30.6	14.3	16.6	16.2	17.6	18.8
ρ	GS-Lin2	0.18	0.02	0.43	0.59	0.35	0.21	0.66
	GS-BC2	0.36	0.30	0.56	0.62	0.21	0.38	0.81
	GS-P	0.78	0.60	0.73	0.78	0.79	0.39	0.91
	GS-Lat	0.67	0.45	0.43	0.37	0.44	0.45	0.72

In the introduction, the authors claim (lines 78 - 88):

- Several reasons motivate this effort. First, such a product can provide continuous and consistent phenological estimates for forest ecosystems, allowing us to infer changes in phenological transitions during the considered two and a half centuries. This product may help reduce uncertainties in terrestrial carbon cycle models and could support adaptive forest management strategies in a changing climate.

But this would only be helpful if the estimations are faithful. The UKESM has an LAI estimate itself. I do not see how producing an additional estimate based only on temperature is helpful especially if it does not outperform very simple models.

We agree with the reviewer that our product is helpful only if estimations are faithful. In the new configuration, GS-P outperforms all models directly evaluated in our study, at the global level and in most climate classes and performance metrics (see updated Table S3 above). GS-P improvements over the other approaches have been highlighted throughout the manuscript. Particularly, in the new discussion:

The GS-Lin2 model, arguably the simplest among the three considered, performs on-par with or worse than GS-Lat in terms of prediction errors. Instead, both GS-P and GS-BC2 substantially outmatch GS-Lat; at the same time, GS-P surpasses GS-BC2 in most instances. In terms of MAE, GS-P achieves globally lower errors than GS-BC2 by 0.9 d for SGS and by 2.3 d for EGS. This is reflected in the better performances of GS-P across all KG classes, with the only exception of SGS predictions over C*a. Despite this specific shortcoming, GS-P exhibits substantially higher correlation across all climate classes, including C*a. Beyond predictive performances, a further advantage of GS-P is that computing CIs on the parameters is straightforward, whereas this task is not trivial for GS-Lin2 and GS-BC2 and requires computationally expensive workarounds. This allows for a more grounded interpretation of GS-P results, whereas any interpretation on the parameters for GS-BC2 and GS-Lin2 would be mostly speculative.

We also agree with the reviewer that there are other available models and products, but these approaches present limitations in spatial extent, species-specificity, or robustness under climate change conditions. We now clarify limitations in previous approaches within the introduction:

While process-based models can predict phenological stages by simulating plant physiological responses to environmental conditions, they are often more computationally intensive than their statistical counterpart and require detailed parameterization, with uncertainties that are often difficult to quantify (Chuine et al. 2013; Meier and Bigler 2023; Fu et al. 2020). Alternatively, several studies frequently linked meteorological data to phenological observations to estimate GS using empirical models and often validating model outputs against ground-based phenological records or satellite-derived indices (for an intercomparison study, see e.g. Y. Mo et al. 2023). However, these models are often limited to specific

geographical regions, or to individual vegetation species (e.g., Linderholm et al. 2008; Blümel and Chmielewski 2012). In the more recent years, machine learning algorithms have been increasingly applied to ecological modelling, demonstrating not only their improved predictive performances, but also their valuable contributions to process understanding and hypothesis testing (Reichstein et al. 2019). Specifically, these algorithms have also been applied to phenological stages, demonstrating their superiority over traditional models, but remaining limited to specific regions, species, or ecosystems (Garnot et al. 2025; Dai et al. 2019; Czernecki et al. 2018; Liu et al. 2024; Sun et al. 2025).

Regarding the LAI product cited by the reviewer, cited limitations to process-based models apply (computational limitation, uncertainties difficult to quantify). Moreover, the raw LAI output from UKESM has a monthly temporal resolution (as many other ESMs, <https://metagrid.esgf-west.org/search>), and cannot be used to reduce uncertainties in terrestrial ecosystem models due to circularity issues.

Regarding the comparison with relatively simple models, we partially agree with the reviewer. While all traditional models employed for comparison in our study are less sophisticated than other approaches, they are not necessarily underperforming. In fact, GS-BC2 has been shown in other study to surpass both traditional and machine-learning models, even while being relatively simple. Particularly, Mo et al. (2023) evaluated the capabilities of 17 models in predicting the start of the growing season; among these models, they also considered the model proposed by Blümel and Chmielewski (2012) (indicated as “M1” within their manuscript) and found that it ranked first in predicting the SGS over deciduous needleleaf forests (Table 3 in their manuscript). While not ranking first over Mixed Forests, this model showed performances comparable to the best one (RMSE of 6.6 d versus 6.4 d; correlation of 0.75 versus 0.76). In another study, Garnot et al. (2025) employed a deep neural architecture (PhenoFormer) to predict spring and autumn phenological transitions over Switzerland. While their overarching purpose is to demonstrate PhenoFormer skills, they also evaluated the Blümel and Chmielewski (2012) model, and found its performance to be only slightly worse than PhenoFormer.

We clarify the choice of this model in the methods (Section 2.2.2 “GS-BC2”):

This model was selected because it was found to perform remarkably well against both traditional models (Mo et al. 2023) and machine learning-based frameworks (Garnot et al. 2025).

In the discussion, we now contextualize the intercomparison work of Mo et al. (2023) and the GS-BC2 comparison within Garnot et al. (2025) in relation to our study:

Y. Mo et al. (2023) compared the predictive ability of 17 different phenological models in predicting the SGS in a variety of vegetation types along a north-south transect of Northeast Asia. In their study, model performance over mixed and deciduous needleleaf forests (the only two forest types considered) yielded RMSE values between 4.9 and 8.2 d and a correlation coefficient between 0.54 and 0.76.

Over the same sub-domain, and evaluated over our validation dataset, GS-P achieves a comparable RMSE (6.4 d) and a much higher correlation coefficient (0.89). Garnot et al. (2025) employed a deep neural architecture to predict spring and autumn phenological transitions using site-level observations across Switzerland. Their model achieved a MAE of 6.4-7.2 d for the spring phenology, and of 9.5-10.5 d for autumn phenology. While a direct comparison with GS-P is not possible due to differences in spatial extent and data sources, Garnot et al. (2025) also compared their model with the model of Blümel and Chmielewski (2012) and found relative improvements broadly comparable to that observed in our study between GS-P and GS-BC2.

We now highlight throughout the manuscript that, compared to existing products, our framework is more general (not species- or ecosystem-specific) and, being mostly data-based, allows the quantification of changes in phenological transitions from preindustrial times to the end of the 21st century in a relatively independent way from other approaches.

- Second, we seek to improve the capability of phenological modelling in predicting the SGS and EGS, and to assess the potential advantages of machine-learning approaches over more traditional methods. Namely, we aim to overcome previous limitations in spatial extent, species specificity, and extrapolation skills under climate change scenarios, by providing a robust, generalizable framework for prediction of phenological stages.

I also do not see this point. First, other work using machine learning to predict phenology (phenophases, LAI, greenness, NDVI or EVI) exists and second, I do not see from the results how this work overcomes these limitations as the authors do not compare to state of the art models and do not beat models clearly behind the state of the art.

We respectfully disagree with the reviewer in this aspect. It is true that other machine learning-based frameworks exist, but ours overcomes the limitations cited in the introduction.

For instance, Garnot et al. (2025) employed a deep neural architecture (PhenoFormer) to predict spring and autumn phenological transitions, but only over Switzerland. They found a MAE of around 6.4-7.2 d for the spring phenology and of 9.5-10.5 d for the autumn phenology. Because they predict at the observational site level, while we do so at the grid scale, their errors are expectedly larger than ours (over continental climates, the MAE of our model is 3.1-7.5 d for SGS, and of 3.1-6.2 in EGS). However, as mentioned above, they also performed an evaluation of their model against GS-BC2. While PhenoFormer was found to perform better, improvements in terms of metrics did not appear substantial. Because in our study GS-P outperforms GS-BC2, results suggest that GS-P and PhenoFormer may perform comparatively. Additionally, our framework is global, and we projected our results up to the end of 21st century under strong warming trends – on the other hand, the model of Garnot et al. (2025) was trained and tested only over Switzerland, and was not tested for global end-of-century projections. As mentioned above, the Garnot et al. (2025) is now contextualized within the discussion.

Another relevant study in the field of machine learning phenology prediction is the one from Sun et al. (2025), which employs an approach similar to ours (XGB), with multiple additional features, but only applied to spring phenology, and only for the boreal region. The performances of the model by Sun et al. (2025) were already discussed in section 4.2 of the previous manuscript version, and the comparison has been updated with the new validation dataset (different dataset split):

GS-P can be readily compared with the similar model proposed by Sun et al. (2025): their XGB model was evaluated to have a MAE smaller than 6 days in 41% of the nodes (across all vegetation types), whereas for GS-P the same threshold is met by 53% of the nodes over the forested boreal region.

Furthermore, our analysis indicates that Sun et al. (2025) predicted shifts in phenological phases that might not be consistent with observations reported in literature so far. This has been better highlighted in section 4.1:

Sun et al. (2025), using XGBoost with 18 temperature-related predictors, estimated SGS timing up to 2100 for boreal vegetation in the NH (which broadly corresponds to D*c and D*d climate class in our study). They found much smaller shifts than those obtained in this study; from the beginning to the end of the 21st century, they estimated an advance of only 0.5–1.2 and 6.8–8.3 d under SSP1-2.6 and SSP5-8.5 respectively, depending on forest ecosystem type. These shifts would correspond to rates well below 1 d decade⁻¹, which appear substantially lower than those generally reported in the literature across the last few decades. In contrast, across the subarctic climates of the NH, our estimates indicate, throughout the 21st century, earlier SGS shifts of 22.6–27.1 d under SSP1-2.6 and of 32.4–40.8 d under SSP5-8.5.

Another existing approach is the one employed in DeepPhenoMem V1.0 (Liu et al. 2024). They employed a long short-term memory network to estimate green chromatic coordinates as obtained by the PhenoCam network over United States. While the overall performances of the model in estimating GCC were found to be overall satisfactory, Fig. S4 of that manuscript reveals some shortcomings in estimating the start and the end of the growing season, especially for the deciduous broadleaf site. Because this study focused mostly on GCC estimates and there is no detailed consideration of phenological transitions, we did not include this study in the main manuscript.

We believe that these comparisons are sufficient to show that our approach overcomes previous limitations in either spatial extent, species specificity, or robustness, while retaining competitive predictive skills. We kindly ask the reviewer to cite other studies that we may compare our results to, should they have further suggestions.

- Third, using temperature as the sole predictor allows us to clarify its role as a driver of phenological transitions at global scale. In this context, model uncertainty is not only a measure of predictive performance but may also help to identify regions where additional controls on phenological timing (e.g., precipitation seasonality) are likely

to be more important than temperature alone, and where future climate change may influence forest phenology and, more broadly, forest ecosystem functioning.

I disagree about this point as well. Using a model that predicts only from temperature ignores that temperature may co-vary with other inputs. Predictability from temperature is not the same as reaction to an intervention. Ignoring the other factors does exactly the opposite. It turns a blind eye and does not allow for understanding the importance of temperature.

We agree that temperature may co-vary with some of the other inputs, and it is true that to fully understand temperature as a driver of phenology one also needs to consider other factors (precipitation, Körner et al. 2023; but, in the southern hemisphere, also topography, elevation and species composition Kosztor et al. 2022). The areas with larger errors/model residuals do not provide definitive proof that there are additional underlying mechanisms or environmental drivers that were not considered, but they simply suggest that this might be the case. We removed this goal from the introduction, alongside the parts of the discussion that referred to the interpretation of model residuals.

To solve these problems the main thing the authors need to do is compare their results to the state of the art and demonstrate a more specific gap they want to close. I understand that the authors claim that their model is using only temperature and hence they compare only to other temperature based models but, as discussed above, I do not see how this is advancing the state-of-the-art.

As mentioned at the beginning of this reply, while we agree that a direct comparison with other machine-learning based approaches would be ideal, it would be hardly feasible, mostly because the cited ML-based studies either do not provide a model file or code (Sun et al. 2025) or because they would require to re-train the model from scratch, which would be too resource and time demanding (Garnot et al. 2025).

Despite this, we have highlighted in the replies above and throughout the manuscript that GS-BC2 is far from being considered an underperforming model. Instead, it should be considered a top-performing model among traditional models (Mo et al. 2023), and it may be only slightly sub-optimal compared to ML-based models (Garnot et al. 2025). In the new configuration with photoperiod added as a predictor, GS-P now surpasses GS-BC2 across most climate classes and performance metrics. By comparing our product with what is available from the literature, we also provided sensible arguments that our model GS-P has competitive predictive skills compared to state-of-the-art models.

In addition to the chosen baseline model, also the data-split is not chosen well. Since the target is to generalize in time and to unknown climate, the authors should also reflect this in the data-split. They should ideally cut a test-set from the end of the time series and not from the middle as well as cutting out a region with a different climate to check the generalization.

In addition, the data split is not well described as it is not clear what constitutes an observation.

We thank the reviewer for the suggestion, and agree that cutting a test-set from the end of the timeseries and from specific regions is more appropriate. We trained and validated again the model with the modified dataset split cited in the reply to the main comment, and updated the manuscript accordingly.

As described above, I do not think the interpretation of the model parameters for this very simple and under performing model is meaningful.

We agree that a direct physiological interpretation of the GS-P parameters would be too strong. While the relationship between the GS-P parameters and their uncertainties are, at their core, unvaried with respect to the previous version, we have substantially revised Section 4.3 to provide a more nuanced interpretation of the parameters:

4.3 Ecological interpretation of GS-P parameters

GS-P parameters can be interpreted from an ecological perspective to understand the physiological features captured by the model. Their interpretation should nevertheless remain primarily statistical, as the probability thresholds and associated variance components do not represent direct physiological parameters, but rather summarize the relationships learned by the model from temperature dynamics, photoperiod, and MODIS-derived phenological transitions.

First, our results suggest that GS-P requires approximately symmetric probability levels to signal transitions into and out of the GS, as shown by the mean probability thresholds (μ) associated with each KG class, which represent the average probability level at which the model identifies a phenological transition. However, the lower uncertainty associated with SGS thresholds indicates a stronger temperature-phenology signal in spring than in autumn. This feature is broadly consistent with previous literature, which reported well-defined responses to temperatures in spring phenology, while autumn phenology often showed more variable and less predictable responses (Menzel et al. 2006; Richardson et al. 2013; Chmielewski and Rötzer 2001). Second, GS-P suggests that conditions for the SGS are more spatially heterogeneous than EGS conditions (between-node variance, τ^2). Such spatial variability may be consistent with spring phenology being more sensitive to microclimatic conditions, species compositions and local adverse events such as late frosts, although the present model cannot distinguish among these possible mechanisms. Third, GS-P shows that phenological transitions are more easily detected in continental (D) than in temperate (C) climates. This is suggested by the sharper probability profiles in the D climates than in the C climates (Fig. S3) and, consequently, by the larger values of the within-node variance (σ^2) of the probability thresholds. Over continental regions, relatively small temperature changes lead to large changes in the probability of GS presence. On the contrary, in temperate climates, noisy probability curves and lower predictive skill indicate that temperature

provides a weaker signal for identifying GS transitions, even after accounting for the seasonal modulation of the photoperiod.

We attach below the updated Fig. S3.

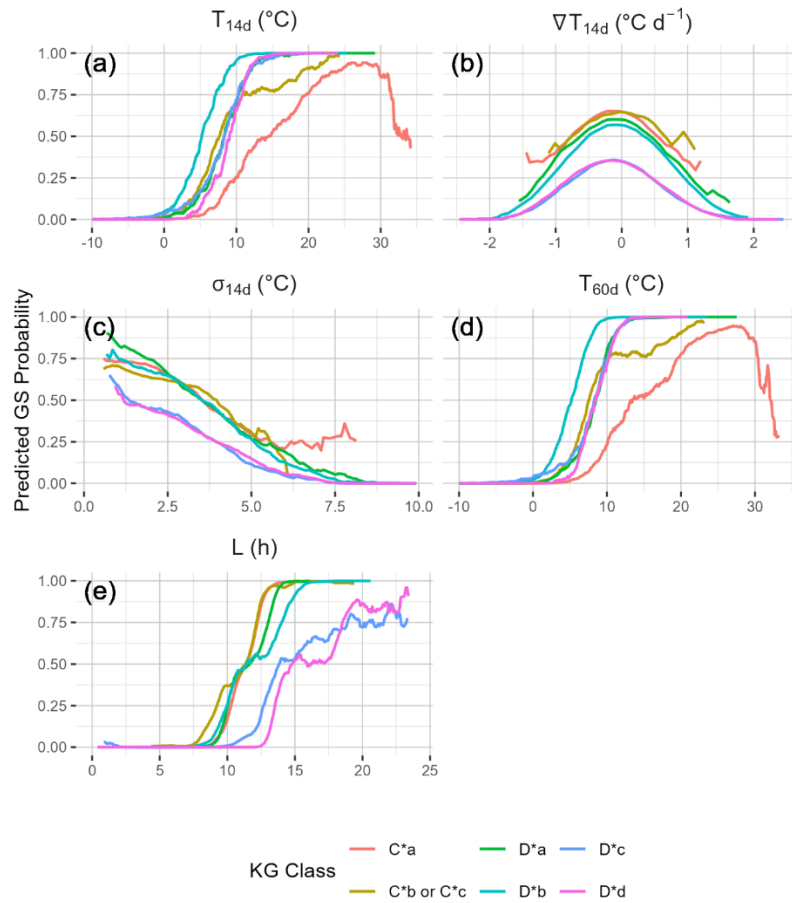


Figure S3. Probability profiles of GS predicted from the binary classifier over each KG class. Profiles are obtained by first computing probabilities over 10,000 samples from the ERA5 dataset, then binning these probabilities, and finally smoothing the binned probabilities using a moving average.

I also think the results for the growing season in the southern hemisphere are [not] believable. Since they go against the current literature, they should be discussed thoroughly.

We thank the reviewer for the suggestion. While results for the SH in the current version of the manuscript are associated with uncertainties, we do not think that these results contradict the current literature. After including the photoperiod among the predictors, we obtain a different outcome for the SH, compared to the previous version. Particularly, over this hemisphere (Section 3.3):

[...] shifts are relatively small from the preindustrial to the present-day period. Here, at future times [(2071-2100 period)], SGS and EGS are both estimated to anticipate by 4.8-8.1 d and 18.8-35.6 d, respectively; the larger early shift in EGS than in SGS yields GS durations shorter by 13.7-28.1 d on average, depending on the scenario.

The new Figure 3 (attached below) reflects the current outcome. Consequently, we now present the rates of change in yearly mean SGS and EGS in a modified Table 3 (attached below) which shows results separately for each hemisphere. We find that (section 3.3):

For the SH, at the present time, global rates are statistically significant for SGS ($-1.4 \text{ d decade}^{-1}$), but not for EGS. Over this hemisphere, at future times, rates become significant only for EGS under SSP5-8.5.

Then, we compare our findings in relation to the current literature in the discussion (section 4.1):

For the SH, our results indicate a negative trend of both SGS and EGS, with a significant global rate of $-1.4 \text{ d decade}^{-1}$ ($p=0.027$) for the SGS, but a non-significant rate of $-0.5 \text{ d decade}^{-1}$ ($p=0.720$) for the EGS. Despite the larger uncertainties in the SH, comparatively to the NH, our results at the global scale are broadly compatible with the findings of Burton et al. (2025). Using NDVI-derived phenology across multiple Australian ecosystems over the 1982-2022 period, they found small non-significant average trends of -0.52 and $-0.69 \text{ d decade}^{-1}$ for the SGS and the EGS of forest ecosystems, respectively, corresponding to an approximately stable GS duration. Few other studies address vegetation phenological phases at relatively large scales in the SH, as research more commonly focus on flowering dates rather than on SGS and EGS timing. From the few long-term studies performed for the SH, evidence of responses to global warming seems weaker and uncertain. For instance, while the review by Chambers et al. (2013) reported an average advance in flowering time of $-5.6 \text{ d decade}^{-1}$, it should be also noted that 72% of the considered datasets lacked statistically significant trends. Similarly, Everingham et al. (2023) investigated species-specific responses in the shifts of flowering dates of 37 herbaceous taxa and found significant shifts in only 12 species

Within the same section we also discuss future projections and related uncertainties:

In the SH, our results project a decrease in GS duration of 13.7-28.1 d by the end of the century, depending on the emission scenario. This decrease is mostly due to

phenological shifts in subtropical climates (C*a), where both SGS and EGS are estimated to anticipate, but EGS is projected to anticipate more than SGS. However, these results are characterized by a relatively large uncertainty, as GS-P predictions are more uncertain over temperate climates compared to continental ones, and the SH domain considered in this study mostly consists of C-class climates throughout the considered two and a half centuries. These uncertainties could be attributed to a number of factors. In Australian ecosystems, for example, Moore et al. (2016) highlighted the strong heterogeneity of vegetation phenology across biomes, and the importance of rainfall and water availability as phenological drivers in many regions, particularly in water-limited ecosystems. Elevation, topography and species composition – which are also not explicitly accounted for in GS-P – were found to be key controls of phenology over *Araucaria-Nothofagus* forests in Chile (Koscior et al. 2022). Furthermore, uncertainty in phenology derived from remote sensing data could itself be an uncertainty source in the SH: Zhang et al. (2017) found that absolute differences among phenological transition dates derived from three remote-sensing products were smaller than 10 d in less than 60% of SH pixels, compared with more than 70% in the NH, with values as low as 40% in South America.

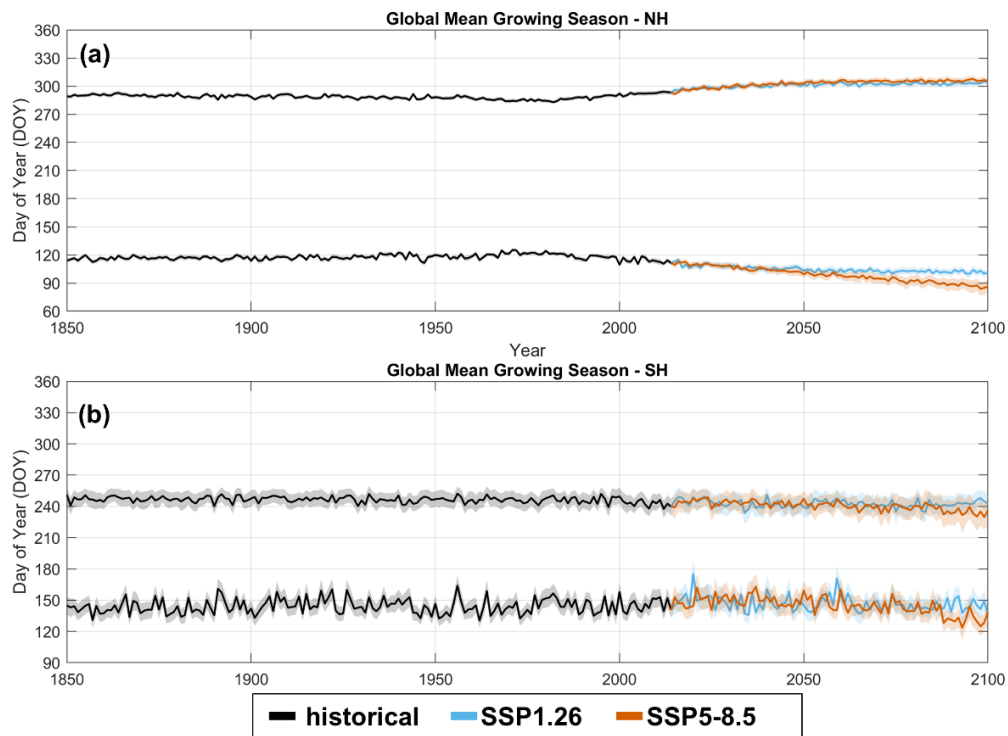


Figure 3. Area-weighted yearly averages of SGS and EGS from 1850 to 2100, for the Northern (a) and the Southern Hemisphere (b), as simulated by GS-P over biased-corrected and downscaled UKESM1-0-LL temperatures, only considering non-KG transition nodes. In the NH, the bottom lines refer to the SGS, while the top lines refer to the EGS. In the SH, the order is reversed. Shaded areas indicate the 95% confidence interval.

Table 1. Rates of changes ([d decade⁻¹]) in yearly mean SGS and EGS over different KG classes and separately for the two hemispheres during the present-day period (1985-2014), and during the end of the 21st century (2071-2100) under different scenarios. Starred numbers indicate $p < 0.05$ for testing non-null mean rates.

	Period	Experiment	NH			SH		
			C	D	All	C	D†	All
SGS	1985-2014	historical	-0.8	-2.3*	-1.8*	-1.3*	0.3	-1.4*
	2071-2100	SSP1-2.6	0.2	-0.3	-0.2	0.7	-6.7*	0.7
	2071-2100	SSP5-8.5	-2.4*	-2.1*	-2.9*	-1.1	-1.4	-1.1
EGS	1985-2014	historical	0.9*	2.8*	2.3*	-0.8	0.9	-0.5
	2071-2100	SSP1-2.6	0.2	0.5*	0.5*	0.6	0.7	0.5
	2071-2100	SSP5-8.5	-1.9*	1.9*	0.9*	-4.7*	4	-4.6*

†While D class values in the SH are reported for completeness, there are very few nodes predominantly belonging to these classes over the selected periods and across this hemisphere (1985-2014: 66 nodes; 2071-2100: 20 nodes for SSP1-2.6 and three nodes for SSP5-8.5)

In the Discussion, the authors mention multiple different models that according to some picked results compare bad to their model. I think incorporating these models into the main analysis or at least picking the best model from these reviews and compare to them instead of the two random models the authors selected.

In Summary, I would recommend to either demonstrate that we can learn something from the temperature only model or to compare to real baselines.

Please see replies above

References

- Blümel, Klaus, and Frank-M. Chmielewski. 2012. 'Shortcomings of Classical Phenological Forcing Models and a Way to Overcome Them'. *Agricultural and Forest Meteorology* 164 (October): 10–19. <https://doi.org/10.1016/j.agrformet.2012.05.001>.
- Burton, Chad A., Sami. W. Rifai, Luigi J. Renzullo, and Albert Van Dijk. 2025. 'Comprehensive Reassessment of Australia's Land-Surface Phenology Trends (1982-2022) Using Circular Statistics and a Harmonised Ndvi Dataset'. Preprint, SSRN. <https://doi.org/10.2139/ssrn.5134345>.
- Chambers, Lynda E., Res Altwegg, Christophe Barbraud, et al. 2013. 'Phenological Changes in the Southern Hemisphere'. *PLoS ONE* 8 (10): e75514. <https://doi.org/10.1371/journal.pone.0075514>.
- Chmielewski, Frank-M., and Thomas Rötzer. 2001. 'Response of Tree Phenology to Climate Change across Europe'. *Agricultural and Forest Meteorology* 108 (2): 101–12. [https://doi.org/10.1016/S0168-1923\(01\)00233-7](https://doi.org/10.1016/S0168-1923(01)00233-7).
- Chuine, Isabelle, Iñaki Garcia De Cortazar-Atauri, Koen Kramer, and Heikki Hänninen. 2013. 'Plant Development Models'. In *Phenology: An Integrative Environmental Science*, edited by Mark D. Schwartz. Springer Netherlands. https://doi.org/10.1007/978-94-007-6925-0_15.
- Czernecki, Bartosz, Jakub Nowosad, and Katarzyna Jabłońska. 2018. 'Machine Learning Modeling of Plant Phenology Based on Coupling Satellite and Gridded Meteorological Dataset'. *International Journal of Biometeorology* 62 (7): 1297–309. <https://doi.org/10.1007/s00484-018-1534-2>.
- Dai, Wujun, Huiying Jin, Yuhong Zhang, Tong Liu, and Zhiqiang Zhou. 2019. 'Detecting Temporal Changes in the Temperature Sensitivity of Spring Phenology with Global Warming: Application of Machine Learning in Phenological Model'. *Agricultural and Forest Meteorology* 279 (December): 107702. <https://doi.org/10.1016/j.agrformet.2019.107702>.
- Everingham, Susan E., Raymond A. J. Blick, Manon E. B. Sabot, Eve Slavich, and Angela T. Moles. 2023. 'Southern Hemisphere Plants Show More Delays than Advances in Flowering Phenology'. *Journal of Ecology* 111 (2): 380–90. <https://doi.org/10.1111/1365-2745.13828>.
- Fu, Yongshuo, Xinxi Li, Xuancheng Zhou, Xiaojun Geng, Yahui Guo, and Yaru Zhang. 2020. 'Progress in Plant Phenology Modeling under Global Climate Change'. *Science China Earth Sciences* 63 (9): 1237–47. <https://doi.org/10.1007/s11430-019-9622-2>.

Garnot, Vivien Sainte Fare, Lynsay Spafford, Jelle Lever, et al. 2025a. 'Deep Learning Meets Tree Phenology Modelling: PHENOFORMER versus Process-based Models'. *Methods in Ecology and Evolution* 16 (7): 1489–506. <https://doi.org/10.1111/2041-210X.70037>.

Garnot, Vivien Sainte Fare, Lynsay Spafford, Jelle Lever, et al. 2025b. 'Deep Learning Meets Tree Phenology Modelling: PHENOFORMER versus Process-based Models'. *Methods in Ecology and Evolution* 16 (7): 1489–506. <https://doi.org/10.1111/2041-210X.70037>.

Körner, Christian, Patrick Möhl, and Erika Hiltbrunner. 2023. 'Four Ways to Define the Growing Season'. *Ecology Letters* 26 (8): 1277–92. <https://doi.org/10.1111/ele.14260>.

Koscior, E., M. Forkel, J. Hernández, D. Kinalczyk, F. Pirotti, and E. Kutchartt. 2022. 'Assessing Land Surface Phenology in Araucaria-Nothofagus Forests in Chile with Landsat 8/Sentinel-2 Time Series'. *International Journal of Applied Earth Observation and Geoinformation* 112 (August): 102862. <https://doi.org/10.1016/j.jag.2022.102862>.

Linderholm, Hans W., Alexander Walther, and Deliang Chen. 2008. 'Twentieth-Century Trends in the Thermal Growing Season in the Greater Baltic Area'. *Climatic Change* 87 (3–4): 405–19. <https://doi.org/10.1007/s10584-007-9327-3>.

Liu, Guohua, Mirco Migliavacca, Christian Reimers, et al. 2024. 'DeepPhenoMem V1.0: Deep Learning Modelling of Canopy Greenness Dynamics Accounting for Multi-Variate Meteorological Memory Effects on Vegetation Phenology'. *Geoscientific Model Development* 17 (17): 6683–701. <https://doi.org/10.5194/gmd-17-6683-2024>.

Meier, Michael, and Christof Bigler. 2023. 'Process-Oriented Models of Autumn Leaf Phenology: Ways to Sound Calibration and Implications of Uncertain Projections'. *Geoscientific Model Development* 16 (23): 7171–201. <https://doi.org/10.5194/gmd-16-7171-2023>.

Menzel, Annette, Tim H. Sparks, Nicole Estrella, et al. 2006. 'European Phenological Response to Climate Change Matches the Warming Pattern'. *Global Change Biology* 12 (10): 1969–76. <https://doi.org/10.1111/j.1365-2486.2006.01193.x>.

Mo, Yunhua, Jing Zhang, Hong Jiang, and Yongshuo H. Fu. 2023. 'A Comparative Study of 17 Phenological Models to Predict the Start of the Growing Season'. *Frontiers in Forests and Global Change* 5 (January): 1032066. <https://doi.org/10.3389/ffgc.2022.1032066>.

Moore, Caitlin E., Tim Brown, Trevor F. Keenan, et al. 2016. 'Reviews and Syntheses: Australian Vegetation Phenology: New Insights from Satellite Remote Sensing and Digital Repeat Photography'. *Biogeosciences* 13 (17): 5085–102. <https://doi.org/10.5194/bg-13-5085-2016>.

Reichstein, Markus, Gustau Camps-Valls, Bjorn Stevens, et al. 2019. 'Deep Learning and Process Understanding for Data-Driven Earth System Science'. *Nature* 566 (7743): 195–204. <https://doi.org/10.1038/s41586-019-0912-1>.

Richardson, Andrew D., Trevor F. Keenan, Mirco Migliavacca, Youngryel Ryu, Oliver Sonnentag, and Michael Toomey. 2013. 'Climate Change, Phenology, and Phenological Control of Vegetation Feedbacks to the Climate System'. *Agricultural and Forest Meteorology* 169 (February): 156–73. <https://doi.org/10.1016/j.agrformet.2012.09.012>.

Sun, Zhe, Jianjun Zhao, Hongyan Zhang, et al. 2025. 'Predicting the Start of the Growing Season in Boreal Forest Under High and Low Emission Scenarios'. *Earth's Future* 13 (8): e2024EF005622. <https://doi.org/10.1029/2024EF005622>.

Zhang, Xiaoyang, Lingling Liu, and Dong Yan. 2017. 'Comparisons of Global Land Surface Seasonality and Phenology Derived from AVHRR, MODIS, and VIIRS Data'. *Journal of Geophysical Research: Biogeosciences* 122 (6): 1506–25. <https://doi.org/10.1002/2017JG003811>.