Review of the manuscript

“The growing seasons of global forest ecosystems from 1850 to 2100 estimated with a probabilistic temperature-based model”

General comment

This manuscript presents a relevant and timely contribution to global forest phenology modelling. The authors introduce GS-P, a probabilistic temperature-based model designed to estimate the start and end of the growing season of global forest ecosystems from 1850 to 2100. The study combines MODIS phenology data, ERA5 reanalysis, and UKESM1-0-LL climate simulations, and compares the proposed model with alternative temperature-based approaches and a latitude-based benchmark.
The manuscript is scientifically interesting, generally well structured, and clearly relevant to the readership of *Biogeosciences*.
The main strengths of the paper are the global perspective, the long temporal coverage, the comparison with benchmark models, and the explicit treatment of prediction uncertainty. The GS-P model appears particularly useful as a flexible and interpretable framework for identifying temperature-driven phenological transitions and regions where temperature alone may provide limited explanatory power.

My recommendation is for a **minor revision**. The manuscript is suitable for publication after some revisions aimed at improving clarity and moderating some ecological interpretations.

Detailed comments

Abstract

L 10–25: The Abstract is clear and effective, but some statements could be slightly softened. In particular, the references to “ecosystem vulnerability”, “carbon cycle modelling”, and “climate–ecosystem disequilibrium” may suggest stronger ecological conclusions than are directly demonstrated by a simple temperature-based phenological model.
I suggest rephrasing these claims as potential applications or interpretations. For example, rather than stating that the model identifies ecosystem vulnerability, the authors could state that GS-P could identify regions where phenological predictions are more uncertain and where additional ecological investigation may be needed to preserve ecosystem stability.

Introduction

L 38–55: The discussion of the different definitions of growing season is useful and important. The distinction between phenological, productive, meteorological, and biological growing seasons should be carried more consistently into the rest of the manuscript. The authors correctly state that meteorological growing season represents potential vegetation activity rather than actual growth. At the end of the Introduction, add one sentence clearly defining what GS-P estimates, for example:
“Hereafter, GS-P estimates are interpreted as temperature-conditioned estimates of phenological growing-season transitions, calibrated against MODIS phenological markers.”

L 78–88.
The motivation is convincing, but the sentence suggesting that the product may reduce uncertainties in terrestrial carbon-cycle models and support forest management should be slightly moderated.
Use wording such as “may contribute to” or “may provide useful constraints for”, rather than implying a direct demonstrated reduction of uncertainty.

Data and methods

L 105–113: The use of UKESM1-0-LL is clearly described. However, since future projections presented in the paper rely on one ESM, this limitation should be stated explicitly in the Methods and recalled in the Discussion or Conclusions.
Add a sentence noting that the future projections represent GS-P responses under UKESM1-0-LL forcing and do not sample the full CMIP6 inter-model range.

L 185–189: The calibration/validation split is reasonable, but it is based partly on random sampling. Because phenological and climatic data are spatially autocorrelated, the authors should briefly acknowledge that this validation mainly tests predictive skill within the sampled spatial–climatic domain.
I do not consider additional validation essential for publication, but the manuscript would benefit from a short comment on the possible value of future spatial block validation or forest-type-specific validation.

L 240–257: The uncertainty approach is a strength of the paper, but the interpretation of the prediction intervals should be clarified. The current description explains how the uncertainty estimates are derived, but not whether they are empirically calibrated against validation observations. Add a short sentence clarifying whether the reported intervals should be interpreted as empirical error envelopes rather than fully probabilistic confidence intervals. If available, report the empirical coverage of the intervals on the validation dataset.

Results

L 321–331.
The model performance is clearly presented. GS-P achieves strong global improvements in MAE, but the manuscript also shows that performance is substantially better in continental than in temperate climates. The sentence stating that MAE values of 3.0–7.6 days can be considered "negligible for most applications" should be softened. Such errors may be small for global or hemispheric trend analyses, but they may still matter for local phenological studies, frost-risk assessment, or species-specific applications.

Suggested revision:
Replace "negligible for most applications" with "negligible for many large-scale applications".

L 336–352: The discussion of spatial bias patterns is useful. I suggest adding one sentence emphasizing that regions with large bias or large error spread should be treated cautiously by users of the dataset, especially where other factors such as water availability, monsoon dynamics, evergreen phenology, or topography may influence phenological timing.

Discussion

L 480–491: This section appropriately notes that temperature alone may not provide sufficient explanatory power in the Southern Hemisphere and that additional drivers may be important. This is an important point and should be made slightly more visible, perhaps also in the Conclusions.
Add a concise "recommended use" statement: GS-P is expected to be most reliable where temperature is the dominant seasonal control, while outputs should be interpreted with more caution in regions with weak temperature–phenology coupling.

Lines 503–521. The ecological interpretation of GS-P parameters is interesting and potentially valuable. However, the wording should be slightly moderated. The model parameters are inferred from temperature–phenology relationships calibrated against MODIS observations; therefore, they should not be presented as

directly capturing physiological features of forest ecosystems. Rather, they reveal statistically derived patterns that are consistent with known ecological and phenological expectations, such as stronger temperature control of spring phenology than autumn phenology and clearer temperature–phenology coupling in continental climates. I suggest replacing expressions such as “physiological features captured by the model” with “ecologically interpretable phenological patterns inferred by the model” or “patterns consistent with known physiological and ecological controls.”

“Although GS-P cannot directly resolve physiological processes, its parameters provide ecologically interpretable information on temperature–phenology relationships across climate classes.”

Lines 522–530.
The GS-flipping discussion is one of the most interesting parts of the manuscript, but it also requires careful framing. The authors already note that it may reflect either climate–ecosystem disequilibrium or a limitation of the current framework.

Suggested revision:
Present GS-flipping primarily as a model diagnostic indicating extrapolation beyond well-established temperature–phenology relationships. The ecological interpretation can remain, but should be framed as a hypothesis requiring further testing.

Conclusions

L 533–543.
The Conclusions are generally clear, but the projected extension of the growing season should be explicitly linked to UKESM1-0-LL forcing. Change wording such as “under climate warming” to “under UKESM1-0-LL climate forcing” or “in the UKESM1-0-LL-based projections”.

L 537–543.
The statements that GS-P identifies biological features, climate–ecosystem mismatch, and priority zones for conservation are interesting but should be slightly moderated.
Use “suggests”, “may indicate”, or “highlights regions that need further ecological investigation”, rather than more definitive language.

Line 557.
The data availability statement is useful, but it only mentions the GS-P output.

Suggested revision:
Clarify whether the code used for preprocessing, calibration, prediction, and figure generation will also be made available.

Language and grammar corrections

The manuscript is generally well written, but several small language corrections are needed.

L 42: “which aspect are covered” → “which aspects are covered” or “which aspect is considered”.

L 109: “primarly” → “primarily”.

L 130: “generic forests ecosystems” → “generic forest ecosystems”.

L 149: “comparatively to others” → “compared with others” or “relative to others”.

L 242:
“To compare the performance the different GS model” → “To compare the performance of the different GS models”.

L 264:
“comparatively to D” → “compared with D” or “relative to D”.

L 266:
“only warm spells of only 3 days” → remove one “only”.

L 484:
“substantial higher number” → “substantially higher number”.

L 514: “species compositions” → “species composition”.

L 550: “synchronous climate-GS regimes transitions” → “synchronous transitions between climate and GS regimes”.