Author responses are embedded below the specific referee comment in green color.

Response to Anonymous Referee #1

This is an interesting study that compares several gridded datasets of land energy, water and carbon fluxes with in situ observations over Europe. Simulations from the CLM5 land surface model are examined in more detail using two categories of simulations, one for each plant functional type and another aggregated to the grid cell level. It is found that CLM5 tends to underestimate the variability of water and carbon fluxes.

A list of possible reasons is given in the Discussion section, but nothing is said about the representation of leaf area index (LAI) by CLM5 and how a misrepresentation of the LAI seasonal cycle and interannual variability could affect model performance. In the introduction, the authors give a very broad definition of phenology without ever mentioning LAI. In reality, LAI is certainly more directly related to phenology than any other variable considered in this work. Moreover, LAI strongly controls land surface fluxes and the evaporative fraction. How is LAI represented in CLM5? Because LAI responds to environmental conditions, it can exhibit large interannual variability. Failure to represent this variability would reduce the ability of the model to represent land surface fluxes.

We are thankful for the referee's comments. We agree that the leaf area index (LAI) is an essential indicator of phenology and ecosystem function and co-varies with carbon uptake (GPP) and evapotranspiration (ET). We did not initially include LAI in our study for two main reasons:

1) In-situ LAI measurements are typically low-frequency, only providing a handful (at best) of data points per year. Therefore, the high-resolved GPP observations are more informative and robust for higher-order statistical analyses.
2) LAI is calculated from the leaf carbon (a partition from GPP), and it controls the upscaling of the carbon uptake from the individual plant to the canopy. Because of this tight relationship between GPP and LAI, we assumed that the analysis of only one variable is sufficiently informative.

However, as the referee rightly pointed out, exploring the phenology and variability of LAI could indeed be beneficial. Therefore, in our revised manuscript, we will comprehensively include the evaluation of CLM5 LAI and in-situ measurements.

Recommendation: major revisions.

Particular comments:
- L. 105-106 (warm winter 2020): Explain why it is called "warm winter".

This curated data-set supported research regarding the record warm winter 2020 in Europe. We will include this information in the revised manuscript.

- L. 143-144 (single soil column): This means that PFT-scale simulations are influenced by other PFTs. This weakens the rationale for PFT-scale simulations. It should be noticed than in other models, each PFT has its own soil column within a grid cell. This should be mentioned in the Discussion section.

We disagree with the referee. PFTs sharing a single soil column compete for water, potentially introducing water stress for less competitive PFTs. PFTs sharing a soil column better represent real-world conditions in heterogenous and water-scarce ecosystems — where competition for water does occur — than PFTs having separate soil columns. Thus, PFT-scale ecosystem processes should compare better to in-situ measurements, strengthening the rationale for PFT-scale evaluations. We will include a respective section in the discussion of the revised manuscript.

- L. 151: How does soil moisture affect stomatal conductance? Given the scope of this work, this should be clearly and completely explained.

A significant soil moisture deficit introduces water stress and down-regulates the stomatal conductance, carbon uptake and transpiration accordingly. We will include this aspect in detail in the Method section of a revised manuscript.

- L. 227 (warm winter 2020): For which time period are data available in the WARM-WINTER 2020 dataset? Only the 2019-2020 winter?

This varies on a site-basis. While the longest available time-series are from 1996, some sites provide only data for a single year. We only consider the simulation data where the observations are also available at the respective location. Please refer to the Supplementary Table S1 for the number of 8-daily data points available for each station. In a revised manuscript, we will include the available years for each site in this table.

- L. 263 (1995-2018): Clarify the link to the WARM-WINTER 2020 data set.

See the reply above. We will include the respective information also in this section in the revised manuscript.

- L. 300: Fig. 1c is not readable as many symbols overlap. This could be improved.

Indeed. We will make the scatter points more distinguishable in a revised manuscript.

- L. 326 (Table 2): units are missing ; what is the meaning of the symbol of column 2, lines 6 and 11?

Sorry for that inconvenience. The units will be included in the revised manuscript. We will also change the symbol indicating the average (Ø).

- L. 340 (Table 3): units are missing ; what is the meaning of the symbol of column 2, lines 6 and 11?

Please see the reply above. We will include the units here, too. The symbol is in some regions commonly used to indicate the average.

- L. 356 (Fig. 2): For a given PFT, is it a mean value across sites?

Correct. It is a mean value across the sites belonging to the respective PFT (and across the years available at those sites). We will include a clearer explanation in the caption in the revised manuscript.

Author responses are embedded below the specific referee comment in green color.

Response to Anonymous Referee #2

CLM5 ET and GPP are compared to ICOS sites in Europe, with RMSE and percent bias metrics. Model ET is often closer to the observations than remote sensing data, but model GPP is underestimated, particularly in deciduous forests.

Generally, the methods in this study seem robust, sources of uncertainty are carefully considered (Section 4), and the aims of the study are worthwhile. I certainly agree with the recommendations, especially with respect to optimizing PFT parameters and co-location of biodiversity and other data with the ICOS sites. The RMSE and bias metrics are well explained and appropriate.

Thanks, we appreciate the recognition of our study objectives and methods.

However, this study would be more accessible to a broader readership if metrics re the phenology and data distributions were better explained, with far less text given to describing the many details of the results and more to interpretation. Data-model comparisons (including for seasonal effects) should be quantified where possible, rather than just assessed by eye.

We agree that quantified indices of phenology would be more accessible than the descriptions in written text. In the revised manuscript, we will provide values for key aspects of the phenology curve (inflection points and modes), and shorten the written descriptions substantially.

The authors present the RMSE and bias results in tables 2 and 3, and also in the text (with some mistakes; e.g. in Section 3.2). Please consider displaying these results in a single diagram, such as a modified Taylor diagram.

Again, we agree and will provide a Taylor diagram in a revised manuscript.

Seasonal effects are shown in Figures 2 and 3, but they are then only discussed qualitatively in the text. There is no attempt to quantify differences (or variability) between model and observed peak ET or GPP timings. For example, model vs observed phase lag or estimated day of max ET or GPP (calculations clearly explained, with appropriate error bars) could be plotted and assessed. In any case, it would be helpful if the second "hypothesis" at the end of the introduction states how goodness-of-fit for phenology is to be quantified. Likewise, for the third "hypothesis", briefly state how the variability will be quantified.

Thanks. Similar to the quantification of modes and inflection points mentioned in the reply above, we will quantify the time difference between these key points in the simulations and the observations in a revised manuscript to support the hypothesis outline in the introduction.

The introduction states that the statistical distributions can help "contextualize" model drought responses, but there is no analysis or discussion about interpretation of the higher moments, responses to drought, or the apparent bimodality seen in Figures 4 and 6 in this article. Do the ICOS data suggest drought conditions at any time at any station? If not, are there any other climate-related

factors that could be discussed or quantified here? Please analyze/quantify/discuss drought, or another factor appearing in the ICOS data. In any case, it would be useful to know how drought (or other factor) affects skewness and kurtosis, and more broadly, what these moments will actually tell us or why we should care about them. For example, would we use the kurtosis to indicate changes in the frequency of extreme values, given kurtosis is a measure of the "heaviness" of the tails of a distribution?

We thank the referee for the critical remark. Investigating drought (e.g., soil moisture deficit) and a drought signal in the carbon uptake or evapotranspiration is highly complex due to the differences in the drought response functionality. For example, plant water stress might occur due to different magnitudes of water deficit in the soil, on different aggregation time scales of the water deficit, and with variable lead time (lag) when propagating from the soil to the vegetation. Given that, investigating whether and when a drought signal is present in the ecosystem processes is out of scope for this study. However, in a revised manuscript, we will shortly discuss how characteristics of the distribution moments, particularly the skewness and kurtosis, could evaluate the representation of observed extreme events in the model data.

Specific remarks

Abstract
The second sentence sounds odd. CLM5 quantifies fluxes and estimates the carbon and water budgets, potentially allowing for a better understanding of how climate change impacts ecosystems.

We will improve this sentence accordingly in the revised manuscript.

Line 30: reanalyses of what?

We will include information on the included reanalysis data in the abstract in the revised manuscript.

Figures and tables

Figure 1c (map of flux towers): Please show the extent of the CLM5 grid (1544x1592 gridcells), perhaps by using a different/lighter grey or white outside the grid area. Please state the number of stations shown in the caption.

The extent of the European CLM5 model corresponds with the complete shown map box. We will include the requested information in the image caption in the revised manuscript.

Figures 2 and 3 (seasonality curves): Define the ICOS, GLASS and model acronyms, and clarify that these curves are means and standard deviations of data covering X-X years during the period 1995 through 2018. It is difficult to see the ICOS curves in some of the panels; please bring them to the foreground in these plots to make them more obvious.

We will adapt the figure according to your suggestions in the revised manuscript.

Figures 4 and 6 (statistical distributions): It is rather difficult to see alignment of the main peaks in some of the panels, which is a point of discussion in the text.

We will quantify the alignment and shifts of modes of the distributions and improve the graph in the revised manuscript.

Figures 5 and 7 (moments): Clarify that these are moments from the distributions shown in figures 4 and 6. The kurtosis appears an "excess" kurtosis, given the normal distribution has kurtosis=3. Please clarify.

Thanks. We will include the requested information in the revised manuscript.

Tables 2 and 3: Ideally, the model and remote sensing acronyms should be defined in the caption; this may be more important than those of the PFTs which are defined in the previous figure and table. Please also explain PFT ⌀ in the final rows for the RMSE and PBIAS sections.

Again, the requested information will be included in the revised manuscript. The ⌀ stands for the average in some regions, but will be written out in the revision.

2.1.2 Setup of the European CLM5
Line 178: did you mean sub models for ice rather than "stub" models for ice?

No. We refer to a stub model as a method that represents a system compartment simplistically.

2.2.1 Station data

Line 229: Table S1 lists a single PFT for each ICOS station; please clarify here that this the dominant PFT as indicated in the last sentence of sec 2.3 ending line 267.

We will correct the information in the revised manuscript accordingly.

Line 234: Please state how many of the 73 stations were kept after wetlands, mixed forest, shrublands and indeterminate-land-cover stations were excluded; I assume 42 (the sum from Table 1).

Correct, 42 stations were retained for the analyses. We will include the information in the text of the revised manuscript.

3.2 General model performance
Line 333 bottom of page 16: The absolute value of PBIAS is smaller for CLM5PFT than for CLM5grid but the actual PBIAS is lower, being more negative. Please clarify; at least replace the word "lower" with "smaller".

We will correct the wording in the revised manuscript.

Line 336: In Table 2, ERA5L and GLASS RMSEs are largest for ENF and DBF as stated, but their RMSEs are lower than those of the CLM for GRA, rather than "similarly" as stated. Their PBIAS values are also much closer to zero than those of CLM5.

Thanks. We will correct the sentence in the revised manuscript.

Section 3.3.1 ET
Line 371: Please refer back to Table 2 when discussing the PBIAS and RMSE for ET. "Conversely" is better than "Oppositely"; the latter sounds weird.

We will improve the wording in the revision.

After this point, I stopped attempting to compare the text to the tables and figures. Rather than summarizing key points of a story, the text rambles on far too much about almost every detail of the results, and it is not always easy to see those details in the figures.

This will be improved in the revised version. We will provide quantified values that were described in the text before, shorten the text, and only summarize the most important points in the text.