Dunkl et al analyze 6 Earth System Models and asses the predictability of interannual variability in gross primary production on land as a function of temperature, soil moisture, and radiation.

Overall the writing in the paper is clear, and the authors describe an appropriate quantity of analysis and results. But there was very little in the way of contextualizing the results from their analysis. The authors do tie their findings frequently to the literature (see further comment #2 below), but as a reader I was left wanting to know more about why this mattered and how to go forward. These revisions can be made by the authors, although I think they are more substantial than minor revisions.

We thank the reviewers for their constructive and helpful comments and appreciate their careful reading of the manuscript. The reviewer makes a general comment expressing the need for putting the findings of the study in a broader context. This also reflects the view of Reviewer No. 1. We will address this by extending the conclusions section of the manuscript to include comments on the implications of the findings and an outlook on the topic.

Major comments:

1. It seems like there is a section missing with further discussion of implications. There is a description of the results, but then no discuss that puts these results in context. As a reader I wondered so what? What are the implications? Right now there are only a few sentences in the last paragraph of the conclusions.

We agree with the reviewer and will elaborate on this topic in the conclusions.

2. In general the results frequently appeals to other papers about why different models do or don't do something without much inclusion of those explanations here. As a reader I needed more help knowing what those previous papers had found to put these results into context. Specific examples:

A major objective of this work is to determine what is causing the differences among the analysed models. Therefore it is especially important to elaborate on these findings. We will review the manuscript for our references to these previous publications and add more details on the mechanisms that lead to the discovered differences.

line 231-232: "due to a misrepresentation of photosynthesis (O'Sullivan et al., 2020)"

This needs more explanation. Misrepresented how? I skimmed the O'Sullivan paper and I didn't find a specific "misrepresentation of photosynthesis" described, just that it matches poorly with observational products (but not *why*). This statement implies that we know why, and I'm guessing that we do not.

Here we are referring to the lines:

"Further, the general lack of in situ observations in tropical latitudes (Cleveland et al., 2015; Schimel, Pavlick, et al., 2015), limits a realistic presentation of photosynthesis in DGVMs, impacting the TRENDYv6 GPP estimates also."

O'Sullivan et al., 2020 found that ESMs have a poor representation of tropical GPP because of an uncertainty in observations. These are because of a lack of flux towers in these regions and because remote sensing products rely on reanalysis data of solar radiation, which has large uncertainties in the wet tropics. We will add these arguments to our reasoning.

line 246 "complex phenological scheme"

What do you mean specifically by complex? What is different about it compared to the other models? How do you know that the performance is better because of the complexity? That evidence isn't shown here.

We agree that the manuscript would benefit from elaborating what this refers to. The cited article uses "*complex phenological scheme*" to describe models which use a plant functional type dependent parametrization of phenology. We will add this information to the manuscript.

line 307 "poor representation of soil moisture"

The paper cited (Qiao et al. 2022) uses reanalysis as their "truth" for soil moisture. Reanalysis is just another modeled product so this is a somewhat misleading statement. Better would be "poor match to other modeled products of soil moisture".

The reviewer points out that the cited study compares the performance of ESMs not with observations but with reanalysis products relying on other models. We thank for this correction and will adapt the section accordingly.

Further I don't see any obvious indication in that paper that CANESM5 is worse than other models. Soil moisture is notoriously poorly constrained by lack of observations and widely varying between models.

With our statement on soil moisture in CanESM5, we are referring to Qiao et al. 2022 Figure 8. This plot shows, that soil moisture has a low variability in CanESM5 compared to other ESMs. As the reviewer has pointed out, it does not necessarily mean a worse performance, since the uncertainty in observations does not allow to make this statement with certainty. We will correct this section to only refer to a lower variability of soil moisture in CanESM5 and the possible implications on predictability.

3. There is a lot of describing baseline climatic regions (i.e. “semi-arid” or similar) that 1) assume that readers will know what regions the authors are referring to, and 2) fail to take into account that there are background climate biases that could shift those locations across models. I suggest that the authors come up with a way to display results that allows for consistency in background climate across models (mean annual T vs mean annual P space being one option). Or that the authors show analysis that is specific to one “climate” region to demonstrate their point. Just staring at maps it was hard to translate back and forth from the statements in the text to the figures.

We thank the reviewer for this helpful insight. We will add a figure showing maps of an aridity index to the manuscript. This will not only allow to better communicate the mentioned regions to the readers but also highlight the differences among the models as the reviewer noted.

4. I found the explanation of the predictable component and predictable fraction challenging. The language wasn't hard to read, but it is a way of quantifying something that I am not familiar with and it was hard to wrap my head around what it should tell me and why. I don't have specific suggestions here but encourage the authors, who know the method well, to consider if they can make it more intuitive to a reader encountering this way of quantifying predictability for the first time. Why is this the type of predictability the authors want to know? What is the interpretation of it?

We introduced the concept of predictable component and predictable fraction to be able to describe different aspects of carbon flux predictability. This concept describes a new way to look at predictability as a multidimensional problem. We acknowledge that the introduction of this concept should take up a more prominent role in the manuscript. We will introduce a section on this issue in the introduction and add an additional subfigure to Figure 2. This subfigure will be dedicated solely to depict the difference between predictable fraction and component, and explain the need for two metrics.

Minor comments:

line 155: "For the lead years five to ten, the effects of initialization are assumed to be negligible." I assume you mean atmospheric initialization? Carbon cycle initialization would sure matter still on those time scales! Leaves probably take 20 years to equilibrate in some of these models.

Yes, we are indeed referring to climate initialization, and will adapt the text accordingly.

line 239: "GSS and GSE" There are already a lot of acronyms in this paper. I suggest just writing these out.

We are removing these acronyms.

Figure 4 and discussion of Fig. 4.

Soil moisture to what depth? Please specify. Total column? Integrated to a similar depth? Surface? Ideally you would want root-zone weighted soil moisture. It isn’t mentioned how the soil moisture shown here compares to root zone.

We will elaborate on the limitations of this approach.