Response to Reviewers: Importance of plant functional type, dynamic vegetation, and fire interactions for process-based modeling of gross carbon uptake across the drylands of western North America

RC1: 'Comment on egusphere-2025-2841', Anonymous Referee #1, 30 Aug 2025

General comments

The manuscript ‘Importance of plant functional type, dynamic vegetation, and fire interactions for process-based modeling of gross carbon uptake across the drylands of western North America’ analyses the simulated variability in vegetation carbon uptake by terrestrial biosphere models, and contextualises the results with observational datasets. The authors identify the representation of C4 grasses as a key reason for failure in reproducing interannual variability reported in observation datasets. Understanding interannual variability in the land carbon cycle (here done by analysing GPP) is very topical and relevant, and is becoming increasingly important for monitoring and successfully projecting ecosystem responses to climate variability and extremes.

We thank the reviewer for their thorough and useful review. We are glad they like the overall scope of the study and think that the work is topical and relevant. We have carefully considered all their comments below and have made changes per their suggestions, which we think have considerably improved the manuscript.

Please see below responses to each comment in blue and changes made to the text in orange (with strikethrough showing deleted text and bold showing additional text).

While I like the overall scope of the study, the paper will need revising before it is suitable to be published. I would encourage the authors to shorten the paper significantly, which is, with 45 pages excluding references!, very long and at times verbose and repetitive. I also wonder whether the authors would consider changing the title to clearly communicate their focus on GPP variability but that is just a suggestion and I am also happy with the current title.

We agree that the original manuscript was too long. We have made several major revisions – based on comments below and in response to Reviewer 2 – to make the results more clear and to shorten and streamline the manuscript. These include:

- We have switched from using the slope of the linear regression between the models and DryFlux as a measure of IAV to only using standard deviations (and standard deviation differences) per the Reviewer's general comment below and comments L467-469, L469 and L507 and L552 below. This has allowed us to streamline the results text, especially in Section 3.1, in which we have cut the text by 130 words.
- We have also switched from using the coefficient of variation in the original Figure 5 and now show the relationship between standard deviation in annual total PFT fCover with standard deviation in annual GPP to be more consistent with the change described in the previous bullet. This also resulted in edits to the text in the first half of Section 3.3 that enabled us to shorten and streamline the key results to reduce the word count by 185 words. We also switched the order of Figures 5 and 6 to be more consistent with the order in which we present the burned area results later in Section 3.3. Please see our response to comments L555-558 and L561 for further details, the new Figure 6 and the new section of text describing both Figures 5 and 6.
- We have removed Section 3.2 as it was largely speculative (thus cutting a further 299 words). Therefore the original Figure 4 has been removed.
- We have reordered the hypotheses and results so we discuss the mean fractional cover before discussing the role of dynamic vegetation and fire because this structure is more coherent and flows better. As a result we have moved the original Section 3.4 up to 3.2 to discuss those results first. Figure 9 is now Figure 4 (with the original Figure 4 removed). No other figure numbers have changed except for switching Figure 5 and 6 (see above).
- We have cut text and streamlined the abstract, which allowed us to cut 290 words.
- We have cut most of the last paragraph of the introduction to avoid repetition with Section 2.5.
- We have significantly revised and streamlined Section 2.5 describing analyses performed following the major changes discussed above and in response to comments from both reviewers. This allowed us to cut 175 words.
- All statements in results that refer back to methods have been removed to avoid repetition (partly in response to various comments from reviewers).

- We have removed discussion of hypotheses in the results and ensured that text is addressed in the discussion.
- We have also removed superfluous sentences providing unnecessary detail throughout.
- We have removed all references and links to where to download data from the main text so they now only appear in the Data and Code Availability Statement.
- We have edited sentences throughout for conciseness and clarity.
- We have cut sections of text in the discussion – partly in response to comments and partly to streamline the text.

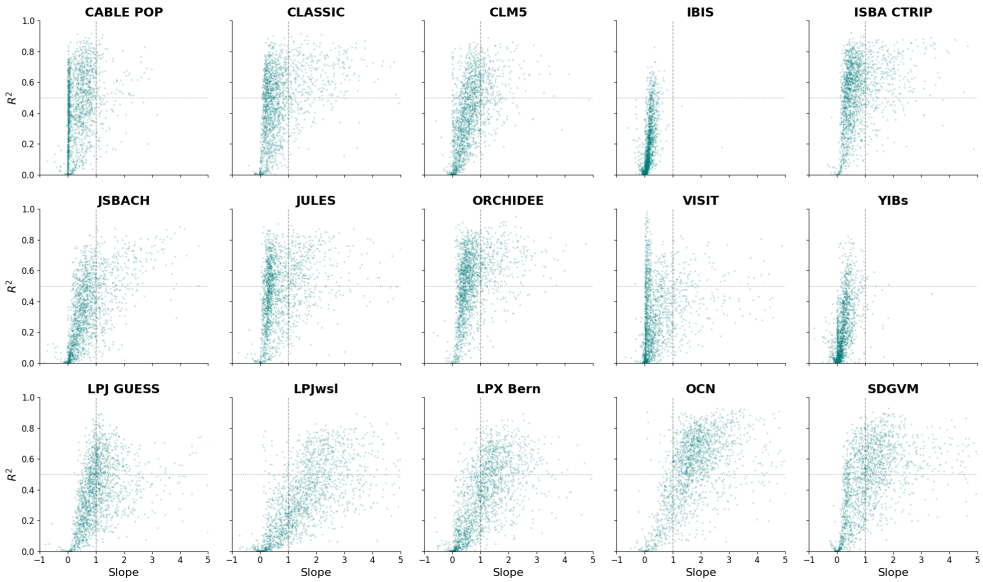
However, per reviewer suggestions we have also added in additional subplots to Figures 1, 2, 5, 7 and 8. At least one sentence was added for each to describe those plots, adding to the manuscript length. We also needed to add two short methods subsections to describe the new burned area reference datasets.

With the exception of smaller sections of text and the results Section 3.1 we have not provided revised larger sections of text in this document. We have made so many changes to the manuscript that we would end up copying most of the manuscript into our response to reviewers. We anticipate that a tracked changes version of the manuscript will be needed to fully understand changes we have made. We have made an exception for Section 3.1, as it is in that section that we have made the key change from using slope as a metric of IAV to using standard deviations (and their differences). We have provided all new figures and their captions in the response to reviewers.

The methods the authors chose are overall relatively easy to understand, but I think they need to be more careful about the interpretation of the results. I especially found the use of the slope to assess the ability of models to capture interannual variability problematic, because the authors did not consider whether those relationships were meaningful, for example by testing the p-value. They also did not consider the correlation coefficient or coefficient of determination to fully understand the relationship between observations and simulated GPP anomalies. This is important because slopes can have values close to one, while not being significant and having low correlation and R2 values when one time series is more variable than another, which is a key aspect of this analysis!

We agree this was a weakness of our original analysis. We have used slope values in previous site-based analyses where it is much easier to grasp whether the slope is meaningful. We did calculate p-values for the slope and we should have put “dots” on

significant grid cells in the original figure. We did also look at R2 values originally, but we did not explore relationships between slope and R2. In considering these comments we have plotted R2 vs slope values to see their relationship (see figure below, which has not been added to the manuscript) and find that R2 generally increases with higher slope values, but all slope values can have both high and low R2. This suggests that – as the reviewer commented – it is not very robust to use slope as a metric of GPP, especially in a wider regional analysis like the one presented in this manuscript.



Given this reviewer also suggested we look at differences in standard deviations of annual GPP between the models and DryFlux, we have decided to remove the slope analysis as a metric of GPP IAV in this paper, and instead we just compare standard deviations in annual GPP timeseries. We agree this is much more simple, clear and defensible as a metric of GPP IAV. Please see our response above with major changes to the manuscript and our response to comment L507 below. We could add the original Fig. 3c and d slope related plots (as well as the scatter plot shown just above) to the supplementary, and add “dots” for significant slope grid cells in original Fig. 3c but given the manuscript is already long and we are also trying to make the text as short and concise as possible, so we have chosen not to do that. If the reviewer thinks those would be good supplementary figures to add then we would be happy to do that.

Finally, we note that we used the “spatial mean” of the slope as a measure of GPP IAV in the original Figure 9 (now Figure 4). We have replaced that with an area weighted mean of the differences in standard deviation, which shows a very similar relationship

and R2 values. Please see our response to L442 below with the new Figure 4 and its caption. We have put these results as mean biased error (MBE) in Table 1 (together with the RMSE to show the spread in standard deviation differences). Please see our response to L507 below.

Lastly, I appreciate that it is challenging to test the impact of PFT distribution, dynamic vegetation and fire on carbon fluxes in the TRENDY dataset, and in consequence results based on the grouping of models according to the (non-)inclusion of dynamic vegetation / fire were at times very speculative. I wonder whether the authors could have come to more robust conclusions if they had picked outputs from the ISIMIP fire sector where a (smaller) number of modelling groups have in fact run simulations with fire switched on and off.

When we started this analysis several years ago the ISIMIP 3a Fire sector (FireMIP simulations were not yet available (and we were not originally planning on exploring the role of fire, only dynamic vegetation). We added in a brief analysis of modeled fire burned area later in the process of developing this manuscript as we thought fire might explain some of the patterns in mean fractional cover and dynamic vegetation that we were seeing. But the downside of adding that analysis at a later stage was that it was not well represented in the original version of this manuscript. We have improved the burned area related analysis in the revised version by including two additional satellite-derived burned area datasets to compare with the model per Reviewer 1's L619 and L801 comments (please see response to this comment for further details). And we have included sections in the methods describing these satellite-derived burned area products.

However, we thank the reviewer for pointing out the potential of ISIMIP simulations and we are strongly considering doing such an analysis in the near future. We have revised the sentence in Section 4.3 that discussed doing factorial simulations with and without fire (lines 839-840 in the original manuscript) to explicitly mention the ISIMIP simulations. Please see our response to comment L828.

I am also curious about the choice of ecosystem variable: This paper is motivated with the importance of drylands for interannual variability in the net carbon balance. Especially in the context of fire, the more natural choice in key variable to me would be

therefore net biome productivity (NBP) to represent the net carbon uptake. Why did you go with GPP?

We definitely agree that NBP is a crucial variable to explore in terms of figuring out how well the models are performing in drylands, and we have done so at site level in previous studies. However, with the exception of atmospheric inversions (which require processing to estimate the natural land C flux), NBP estimates at regional to global scale are lacking. In contrast, several regional to global scale GPP data-derived products are available but a previous study (Barnes et al., 2021 - cited here) has shown most of the commonly used GPP datasets do not tend to perform well in dryland ecosystems. Now we have a GPP product that performs better in drylands from Barnes et al. (2021; especially in the western North America region we focus on in this study) and as GPP has been shown to be the primary control on net CO_2 flux in drylands (Wang et al., 2022), we thought it would be an interesting to explore how DGVMs compare to that dryland-specific GPP product. Previous site based studies have suggested that DGVMs might perform poorly in drylands due to vegetation responses to changing water availability. Therefore we do need to analyse how well models perform in capturing gross CO_2 uptake, and not just the net CO_2 fluxes. But we agree exploring how DGVMs capture NBP is crucial and a comparison with atmospheric inversions in drylands would be interesting to do in future studies.

We have made the following additions to the text to reflect some of what is written here:

- Highlight GPP as a control on NEEin the introduction.
- Mentioned the potential to compare DGVMs with atmospheric inversions to evaluate net fluxes in Section 4.4

Specific comments

In the following I have also pointed out a few typos - I think the manuscript could use more proofreading before re-submission.

We agree and thank you for pointing out these mistakes. We have proofread the manuscript for this revision and have corrected additional typos and mistakes, but we will do so again once we have accepted all the changes we have made..

L52-56 I think this is too much detail for an abstract

We agree, thank you. We have removed or streamlined a lot of the abstract, following both reviewer's suggestions and our own edits. Please see the tracked changes version of the revised manuscript.

L58 Two typos - missing space between 'theDryFlux' and missing 'a' before newly developed

Thank you, we have corrected that.

L60 Typo 'a one'

Corrected, thank you.

L60-62 I would suggest to drop the text in the parentheses here

We agree that the level of detail was too much. We have removed it.

L63-64 Again, drop the text in the parentheses

Thank you. We have removed this.

L73-83 In this section too there is a lot of detail on the results but the abstract should only highlight key conclusions

Thank you for this suggestion. We agree. We have cut a large chunk of the results in the abstract and only highlight key conclusions. Please see the response to Reviewer 2's "Abstract: too long" comment for the revised abstract text.

Introduction :

L85 Would suggest 'water demand (potential evapotranspiration; PET)'

Thank you for the suggestion, we have changed that.

L88 Remove white space between parenthesis and full stop

Corrected, thank you.

L93 Capital E for ‘earth’

Thank you, corrected.

L106-114 I think the description of TRENDY can be shortened

Thank you. We have shortened this description as follows:

“The TRENDY ~~(“Trends and drivers of the regional scale terrestrial sources and sinks of carbon dioxide”)~~ model intercomparison project, initiated ~~began~~ in 2009, compares~~with the goal of comparing a suite of~~ DGVM estimates of ~~for~~ global atmosphere-land CO_2 fluxes and~~(Sitch et al. 2015). The TRENDY model ensemble simulations~~ provides ~~important~~ estimates of the natural land carbon sink ~~(cumulative net biome production)~~ and~~in addition to the impact of~~ land use change emissions ~~on land carbon cycling~~ for the annual Global Carbon Budget (Friedlingstein et al., 2024**; Sitch et al., 2015; Sitch et al., 2024**).;~~ however, outputs from TRENDY have been used in a host of other studies exploring land surface and dynamic vegetation process responses to global change drivers beyond the original TRENDY remit (Zhu et al. 2016; Yuan et al. 2019; Pan et al. 2020).~~”

L116, 118, 122 Missing white space before citation

Thank you, corrected.

L189/190 Given the authors are mentioning an ongoing debate on which biome contributes the most the carbon cycle variability I would suggest to include an additional reference to Ahlstroem et al. that supports the dominant contribution of tropical forests

Thank you for this suggestion. We have included a few more references: Poulter et al. (2015), Zhang et al. (2018) and Bogucki et al. (accepted) (see comment L897 below) – all of which are already in the reference list.

L201 The transition into the objectives is a bit abrupt

Thank you, we have added the following sentence at the end of the previous paragraph to transition more smoothly:

"Therefore, the DryFlux product should serve as a more robust benchmark for assessing dryland GPP simulated by DGVMs"

L201 Couldn't a third hypothesis be that the models match the DryFlux product well? :)

Yes, indeed this is true. And in our reevaluation of the model results we indeed identified that many of the models estimate DryFlux variability well for much of the study region and have updated our description of the results accordingly. Here, we have edited this 1st objective to account for this possibility, so it now reads: "The two main objectives of this study are to: 1) **use DryFlux v1.0** to identify if TRENDY v11 models have **similar,** higher, or lower variability in annual GPP compared to the DryFlux product across dryland regions of western North America;"

L211 Here you say a potential outcome is a 'better' match with DryFlux GPP while before (see my previous comment) you say that models can only be too high or too low

True, we missed this point in the original manuscript. We have altered the 1st main objective to identify we are also expecting a similar match with DryFlux in certain situations. Please see our response to the previous comment L201.

L250 Was the aridity index calculated in this study, if so which method for the calculation of PET was chosen, if no, where is the dataset coming from [...]

Given we are trying to shorten this manuscript we have not included a full description of how PET was calculated. However, we have included the following sentences here (some parts have been moved up from Section 2.4):

"Here we use the aridity index dataset from Trabucco and Zomer, (2022) which is based on the method of Zomer et al. (2022). The aridity index is provided as a long-term mean for the 1970-2000 period and is derived from the WorldClim 2.1 dataset (Fick and Hijmans, 2017). We note that this is not the same climate data as was used to drive the TRENDY model simulations (Section 2.2)."

We have also included the following in the Data and Code Availability statement (that was in Section 2.4 line 383-385 in the original manuscript):

"The aridity index data is based on the method of Zomer et al., 2022 and was downloaded from Global Aridity Index and Potential Evapotranspiration (ET0) Climate Database v3 (Trabucco and Zomer, 2022)."

L256 Would suggest to start a new paragraph

Thank you for pointing out that the paragraph was too long. We added a new paragraph.

L257-258 You never use MAT or MAP again, so no need to define those abbreviations

Thank you for pointing this out. We have removed these sentences to shorten this section (see response to L256-269 below).

L262 missing bracket after October

Thank you. This has been added.

L256-269 I think this description is too detailed. Currently your Figure 1 is relatively large with a single variable, and given the importance of the regional climate I wonder whether Figure 1 could be changed into a multi-panel plot also showing MAT, MAP, and IAV of the relevant climate variables

Thank you for this suggestion. We have added maps of MAT and MAP to Figure 1 and removed the sentence describing MAT and MAP ranges in this paragraph to shorten it. We have also edited the rest of this section to briefly describe the climate in the three different regions and then the main vegetation types, as these are key to understanding the results. Here is the new Figure 1:

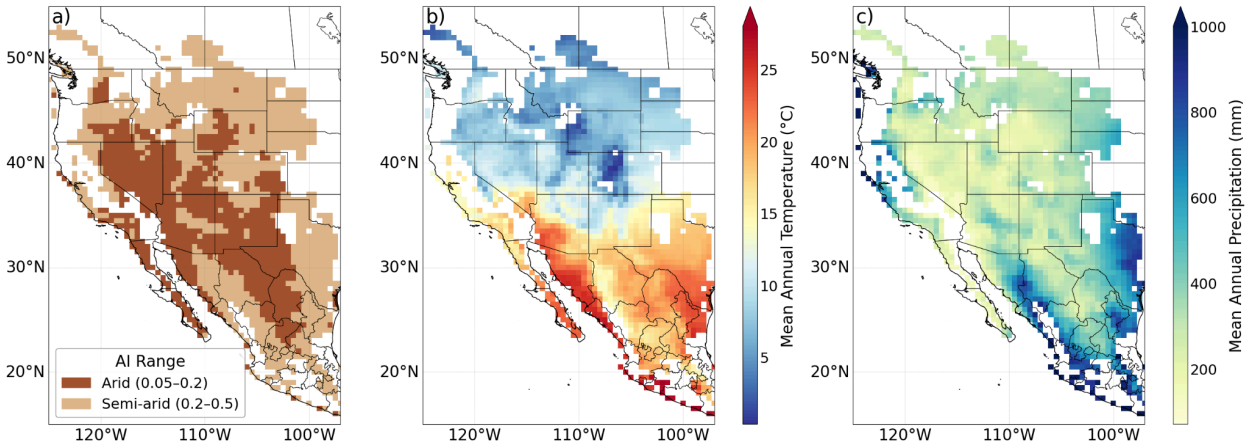


Figure 1: Drylands (excluding modeled cropland PFTs) in western North America. a) Arid regions (aridity index 0.05 - 0.2) are shown in dark brown, semi-arid regions (aridity index 0.2 - 0.5) in light brown**; b) Mean annual temperature and c) Mean annual precipitation for the study period from 2011 to 2016**.

L272 Have you calculated the Aridity index yourself - if so you should mention which reanalysis dataset it is based on, if you have taken someone else's dataset it needs to be cited here.

No, we didn't calculate it ourselves. Apologies for not specifying that in the original manuscript. Please see our response to L250 above for further details and changes to the manuscript text.

L283 In line 255 you say you explicitly want to focus on natural vegetation, yet you take S3 with transient land-use change. Why?

We agree this is confusing. Actually we only excluded *croplands*, not any vegetation that has not experienced any change. This was to remove sources of inter-annual variability that might be more directly due to short-timescale managements, such as changes in crop planting and harvesting and also to focus on non-irrigated vegetation. The other reason for excluding croplands from the models was to make the models more comparable with DryFlux data as the Dryflux random forest model explicitly excluded cropland sites. However, the dryland flux sites that were included in DryFlux "see" effects of land-use change within those flux site footprints, so for this reason we still used the S3 simulations. To clarify this, we replaced "natural vegetation" with "non-crop" vegetation in this sentence.

L285-287 I'm not sure it is necessary to explain why some simulation types were NOT chosen

We agree, thank you, and have deleted this sentence. For the purpose of reducing the paper length we are deleting any unnecessary information or sentences.

L287-288 This doesn't add much, I would just cite Sitch et al., 2024 at the end of L285 like '(see Sitch et al., 2024 for further details on simulation protocol)' or similar

We agree and have followed this suggestion, thank you.

L288 Which method did you choose to regrid your data?

We used normalized conservative resampling for total GPP regridding. This detail has been added to this section of the methods. Please see our response to L376 for why we chose this regridding method for the GPP dataset.

L304-315 Again, this is very wordy. Could an illustration or a table help to summarise the key differences in PFTs in addition to a few sentences?

We agree. We have edited this section to remove the lists of model names and changed the text slightly so we can more easily refer to the information in Table S1. We have also removed the sentence about bare soil in order to shorten the manuscript and because we do not discuss bare soil much throughout the rest of the manuscript. Readers can find that information in Table S1. This section now reads:

"Each of the TRENDY models have their own PFT fCover maps (either prescribed or predicted) based on the PFT types represented in their model (Supplementary Table S1). The common land use change forcing **used in TRENDYv11 (Land Use Harmonization Dataset – LUH2)** is imposed on the models' own PFT maps (or dynamic PFT simulations) (Sitch et al., 2024). Each model has a different number of PFTs and spatial resolution. Some models **represent**~~have~~ shrubs as separate PFTs, **while** ~~such as CABLE-POP, CLM5, IBIS, ISBA-CTRIP, JSBACH3.2, VISIT, YIBs. O~~**others include only tree PFTs or use unspecified woody vegetation categories** ~~but some have combined shrubs with woody tree vegetation~~ **(Supplementary Table S1)** ~~such as CLASSIC, LPJ-GUESS, LPJwsl, LPX-Bern, OCN, ORCHIDEE, SDVGM~~. All models have separate C3 and C4 PFTs, except VISIT ~~and OCN (al~~though in OCN these were defined as tropical/temperate grasses). ~~Some models do not have a separate bare soil such as LPJwsl, LPX-Bern (except for an 'Urban Bare' class), VISIT, and YIBs.~~"

L317 Which ‘certain analyses’?

Thank you for pointing out this was unclear. We have changed this to: “For ~~certain~~ **analyses exploring whether mean PFT cover explains how well models capture GPP IAV**, we split the non-woody vegetation group further to compare grasses/herbs versus crops and C3 versus C4 types”

L319 Presumably the peat graminoid would be zero in your study region anyway? I’m not sure I would recommend to just drop PFTs that are a bit annoying based on their non-traditional definition if they might have a strong contribution to your variable of interest

This is a fair point. We checked and indeed there are zero peat graminoids in the study area. But we have now included peat graminoids with C3 grasses in our scripts. We have removed this sentence in the main manuscript to shorten the manuscript and because the information is clear in Table S2.

L321-322 Did you test this?

Yes we did confirm this with an analysis of the changes between each month. We have revised this sentence so it now reads:

“However, **testing monthly differences for** ISBA-CTRIP **and** ~~does not model spatial distribution of vegetation dynamically and therefore we expect to have no monthly variability in fractional coverage of its PFTs. Although~~ **LPX-Bern** within each year ~~models spatial distribution of vegetation dynamically, from visually inspecting its monthly data using Panoply we did not identify any monthly~~ **revealed there were no** changes in the **monthly** PFT ~~distributions~~ **fCover** (only annual changes were identified). ”

L323-324 I don’t think ‘visually inspecting with Panoply’ is a robust method

Yes we agree and we did confirm this with an analysis of the changes between each month. Please see our response to the previous comment (L321-322) for changes we made to the text to reflect this.

L328 I find the soilcrust PFT argument a bit random. Would this not rather be an extra layer of the soil column?

We agree that it would be part of the soil column and therefore not relevant for the PFT section. So on reflection this is more of a discussion point. We have moved it to the discussion section 4.4.

L328 You are also missing a reference here

Please see our response to the L328 comment just above. We have added a reference for this sentence about biocrusts in the discussion.

L330 Why was nearest neighbour interpolation used here? Isn't this more commonly used for categorical datasets?

Please see our response to L376 for why nearest neighbour regridding was used here.

L334-347 I think this could be shortened, e.g. it is mentioned multiple times that it is an upscaled dataset. Once should be enough to get that point across :)

Thank you, we agree. We have removed this sentence and have merged parts of it into the final paragraph of the introduction where we first introduce DryFlux.

L356-358 Is it really necessary to give that much detail on different errors?

True. We have removed these sentences but referred to remote sensing based fCover map uncertainties in discussion Section 4.2.

L361 How did you regrid your dataset?

We used spatial averaging for all fine grids within each 0.5°x0.5° pixels for regridding. These details – including the reasoning for using this regridding method for this dataset – have been added to this section of the methods. Please see our response to L376 for further details.

L370-371 Again, is it really necessary to provide detail such as the exact errors for this dataset? The methods are already very long

We agree. Please see our response above for L356-358.

L376 I would suggest, instead of mentioning it separately for each dataset, to have a summary statement on how the regridding is done in this section. Presumably all regridding has been done with the same method?

We did not apply the same method to all regridding because different data types needed different processing for regridding; therefore, we have included the method for regridding in all relevant sections. We selected approaches based on the characteristics of each dataset to minimize bias. Data that are not in the geographic co-ordinate system we geotransformed. For the fine resolution 30 arch second (there are 3,600 pixels in a 0.5 degree pixel area) aridity index and much finer fractional cover data (e.g., RAP), we took a spatial average of all pixels within the course pixel. This approach reduces high-frequency noise and preserves the mean signal at the coarser scale. Spatial averaging is better suited for aggregating high-resolution data to coarser grids without introducing artificial gradients. For coarser-resolution fractional cover model outputs and masks, we used nearest-neighbor interpolation when mapping to a finer grid. This avoids smoothing and preserves values of land-cover fractions from course to fine resolution and preserve categorical masks, which could otherwise be distorted by interpolation methods like bilinear averaging. For coarser resolution GPP flux and burned area model outputs we applied a normalized conservative regridding approach to ensure that the total quantity is preserved during the regridding process, which is critical for total flux processing. We have not added the reasons why we used separate regridding methods to the document to save space, but we have added the methods used to each individual section per the comments above.

L382 Already mentioned earlier

Thank you for pointing this out. We have removed this sentence as we describe the use of aridity index to create a dryland mask in Section 2.1. Instead, we have just said we applied the dryland mask to all model and data products.

L383-384 I hesitate to suggest adding yet another paragraph in the methods, but i) why does the aridity dataset not deserve its own paragraph like the other datasets, and ii) why was the aridity index not derived from the climate forcing used to drive the TRENDY models? Was it used to train DryFlux?

i) We have included additional details about the aridity index dataset where we first mention it in Section 2.1 (please see our response to L250 above). Essentially it has its own paragraph there.

ii) The aridity index provided by Trabucco and Zomer (2022) was already calculated based on an in-depth method described in their paper. In part it is based on a model for calculating evapotranspiration (ET), which itself would be different from each of the ET models in the TRENDY DGVMs. Therefore it would not be simple to calculate the aridity index from the same climate forcing dataset as the models and each model would give a different estimate based on their ET model. The aridity index dataset we used is also the one that is used in other dryland studies that we have cited such as Fawcett et al., (2022) and Wang et al. (2022) so for consistency with those past studies we used the same aridity index dataset here.

DryFlux v1.0 used an older version of the aridity index dataset that was available at that time developed by "*Sorensen, L. (2007). A spatial analysis approach to the global delineation of dryland areas of relevance to the CBD Programme of Work on Dry and Subhumid Lands. UNEP-WCMC, Cambridge."* As we have updated aridity index data developed by Trabucco and Zomer (2022) (and for reasons described in our response to L250) we used the updated data in this study.

L388 Why 1970-2000? Wouldn't it make sense to match the study time period, i.e. 2001-2016?

We used the mean aridity index data from Trabucco and Zomer (2022), which provided the mean over the 1970-2000 period. While we see the point that it would make sense to match the study time period, we used this dataset for the reasons provided in our response to the previous comment (L383-384). This information about the aridity index dataset (mean over the 1970-2000 time period as well as the method) is now included in Section 2.1 where we fully describe the aridity index dataset. We also moved the information about where the aridity index dataset was downloaded from to the Data and Code Availability statement. Please see our response to comment L250 for further details.

L393 I would drop ‘That means [...] for any model’, this is already clear from the description

Ok thank you. We have removed this sentence.

L395 In this section I would suggest to have different subheadings for each of the statistical methods that were applied

Thank you for this suggestion. We have added in subheadings that match the results sections.

L398 It is not clear here how anomalies were derived / whether some normalisation method was applied and if so, which one

We agree this was unclear. We have replaced the text originally in parentheses with a new sentence: “Annual anomalies were calculated for each of the 15 TRENDY models and the DryFlux product by dividing the spatially aggregated annual GPP 2001 to 2015 time series by its mean.”. This is what we meant by “mean-normalized”.

L399-400 Was the data detrended for the pixel-wise comparison of time series with a Pearson correlation?

We agree this is unclear here. This correlation refers to the correlation between the spatially averaged annual GPP (per square meter) over the whole study area for the TRENDY ensemble mean and DryFlux, not a pixel-wise comparison. However, the data were not de-trended. We have now detrended the annual GPP time series for the models and DryFlux before calculating the Pearson correlation. The results have not changed much (see new Figure 2).

L423-425 Is this referring to the coefficient of determination? In the results it was not obvious this was considered for the analysis of the slope, at least not in all of the analyses

We have switched from using both the coefficient of determination to only using standard deviations in other model variables and model annual GPP. Please see our

response to comment L555-558 below and our responses to the general comments above for further details. We have also completely revised Section 2.5 to account for this and other changes. Please see our response to L442 below.

L442 I don't understand what 'typical uncertainty' means, and I'm not sure how including the 10% threshold helps grouping your data according to PFTs

Thanks for pointing this out. We were assuming any pixels that have a fractional cover below 10% are not reliable because of the uncertainty in the dryland fractional cover datasets. However, as 10% was an assumption based on a separate dryland fractional cover validation paper we read we decided not to use this threshold and have therefore removed this section of text in the document. Therefore, we updated previous Figure 9 (now Figure 4) with including any pixels that contain the cover types. We only excluded grid cells with 0% cover. All other pixels that have any fractional cover of the PFTs examined are included. We have completely updated the text in Section 2.5 but have not pasted it here as this is already a long document. Note in the new version of Figure 4 (inserted below) we also changed the circle size from the number of pixels to proportional area.

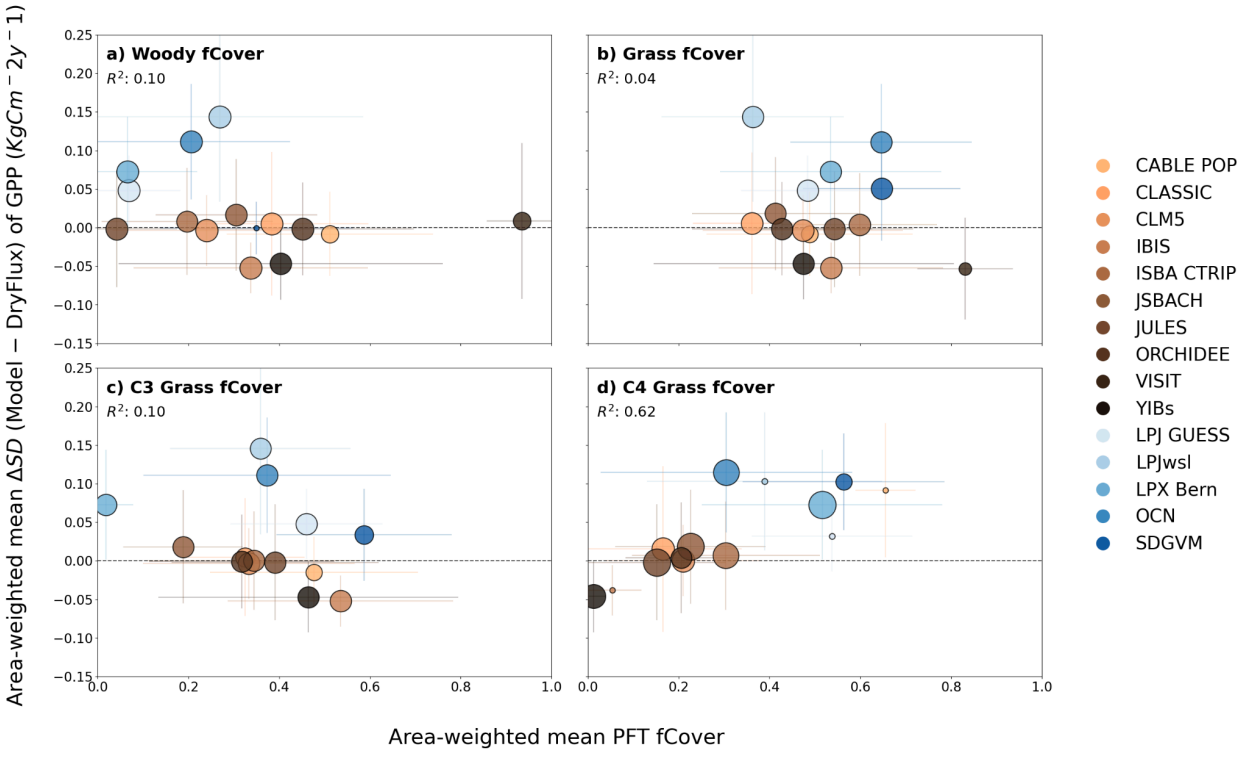


Figure 4: Scatter plots showing the area-weighted mean fractional cover (fCover) across the study area for a) woody, b) grass, c) C3 grass and d) C4 grass cover greater than 0 versus the area-weighted mean of differences in standard deviation in annual GPP between the models and DryFlux (models–DryFlux) (equivalent to the MBE in Table 2). Brown points represent models that generally capture or underestimate spatial patterns in DryFlux GPP IAV (~zero or negative differences in standard deviations), while blue points represent models that tend to overestimate DryFlux GPP IAV (positive differences in standard deviations). For each scatter plot the point size is proportional to the area of the grid cell, the horizontal error bar represents 1 s.d. spread from the mean fCover and the vertical error bar represents 1 s.d. spread from the area-weighted difference in standard deviations in annual GPP (models – DryFlux) across all grid points. The dashed horizontal line shows mean differences in annual GPP standard deviations of 0.0.

L435-445 I found this description a bit convoluted and hard to follow so my comment might reflect my lacking understanding - but would it not make sense to use the PFT-level GPP output to partition directly which PFT or PFT group contributes most to IAV in GPP?

That would be ideal but actually only 10 of the models output GPP per PFT and of these 10 models only 7 output C3 and C4 grasses separately, so we could not do a full analysis with all models. However, we could do an analysis of GPP for each of the major groupings (woody, grass, C3 and C4) for these 7 models to determine if variability in C4 grass GPP is considerably higher than GPP for other PFTs. This would be an exploratory analysis but for the dynamic vegetation enabled models we still wouldn't be able to discern whether the variability is because C4 photosynthesis itself is more variable or whether C4 grass fractional cover is varying more year-to-year. We haven't added such an analysis to the manuscript yet, partly because we are trying to shorten the manuscript and partly because we need to think this through more carefully, but if the reviewer thinks this would be a valuable addition we will explore this analysis further for the subset of models we have.

L456 Typo - 'tDryFlux'

Thank you. Corrected.

L462/Figure 2: I think it would be interesting to also plot the standard deviation or CV as an extra panel here

This is a great suggestion, thank you.. We plotted SD as additional subplot c) in figure 2. Thank you for suggesting this. We have updated the text in Section 3.1 to refer to this extra figure. The new text reads:

"~~Figure 2 shows the time series (2001-2016) of annual GPP summed over the entire study area.~~ Although the TRENDY ensemble mean total GPP **time series** (thick grey curve**s** in Fig. 2) is lower than the DryFlux GPP (thick black curve**s** in Fig. 2), the sign of the annual anomalies in TRENDY ensemble mean GPP match those of DryFlux well (Figs. 2a and b; R: = 0.96). However, individual TRENDY models simulate both higher and lower annual total GPP (Fig. 2a) **and standard deviation in total annual** GPP **(as a measure of GPP** IAV) (Fig**s**. 2b **and c**) compared to ŧDryFlux~~(Fig. 2a)~~. Most models **approximate DryFlux annual GPP IAV well or tend to** underestimate the~~both DryFlux mean GPP and~~ IAV (brown curves**/bars** in Figs. 2**b**~~a~~ and **c**~~b~~; **% difference in total annual GPP standard deviation (model-DryFlux) of +18% to -69% in Fig. 2c**). A~~while a~~ select~~ed~~ few models (OCN, SDGVM and the LPJ family of models) exhibit much higher annual GPP variability compared to DryFlux ~~tend to overestimate both the mean~~ **(and overestimate the DryFlux annual GPP magnitude)** ~~and the IAV~~ (blue curves**/bars** in Figs. 2a**–c** ~~and b~~; **>60% difference in total annual GPP standard deviation between model and DryFlux)**."

L464 I would suggest to rewrite to 'The ensemble mean'

Thank you for this suggestion, we have changed the text.

L467-469 See my general comments - I am not convinced that the slope of the linear aggression alone is very robust.

We agree. We have removed the slope analyses in this manuscript following the reviewer's general comment above (please see our response to that). We switched from analysis the spatial distribution of the slope of the linear regression for each grid cell in Figs. 3c and d to look at the differences in standard deviation, and calculating metrics to summarise those spatial patterns (see our response to L507 below, including the new Figure 3). In the original Section 3.4 (now Section 3.2) we relate mean PFT fCover of the major PFT types with mean GPP IAV. Originally we used the spatial mean slope across the study area as the metric of mean GPP IAV, but we have now replaced that with area-weighted mean of the differences in standard deviation.

L469 I don't find it obvious that the standard deviation is generally lower in Figure 2, perhaps consider my suggestion to include this metric as an extra panel

We agree. We have added the extra sub-figure and edited the text accordingly. Please see our response to L462/Figure 2 above.

This also reflects the slight change in emphasis with revised Figure 2 and our updates to Figure 3 (please see our response to major comments above and our response to L507 below). We now highlight that the models we previously said underestimate the model tend to either approximate DryFlux well or underestimate over a large part of the study area (Fig. 3a and c top two rows), while the other models (bottom row in Figs. a and c) clearly overestimate DryFlux over most of the study area. This is reflected in this new subplot (Fig. 2c) suggested here showing the standard deviations in total annual GPP across the study area for each model and DryFlux.

L481 Why is the data not shown - might make a nice addition in form of a multi-panel plot in Figure 1

We agree. We have added those plots to Figure 1. Please see our response to L256-269. We have changed the "data not shown" here to "Fig. 1c".

L507 Would it not make sense to show the differences in standard deviation between models and DryFlux? I think it would also be interesting to also include the ensemble mean, also in panel c. In panel c, the colorbar is hard to read and it is hard to see where values are negative (is this a frequent occurrence? And if so what does a negative slope between the observation and simulation of the same variable mean?) It would also be nice to have a high resolution figure here, there is a lot of granular detail in this figure that is hard to see in the low resolution

This was a great suggestion, thank you. As mentioned in our response to the major comments we have chosen to replace the original Figs c and d with maps and KDE plots of the differences in the standard deviations in annual GPP across the study area to show spatial patterns of where models are under- or over-estimating DryFlux annual variability (or capturing DryFlux annual variability well). We have kept the actual standard deviations in annual GPP for both the models and DryFlux (and their KDE plots) in Figs. 3a and b as it is helpful just to see the "real" patterns of variability as opposed to the differences. Here is the new Figure 3:

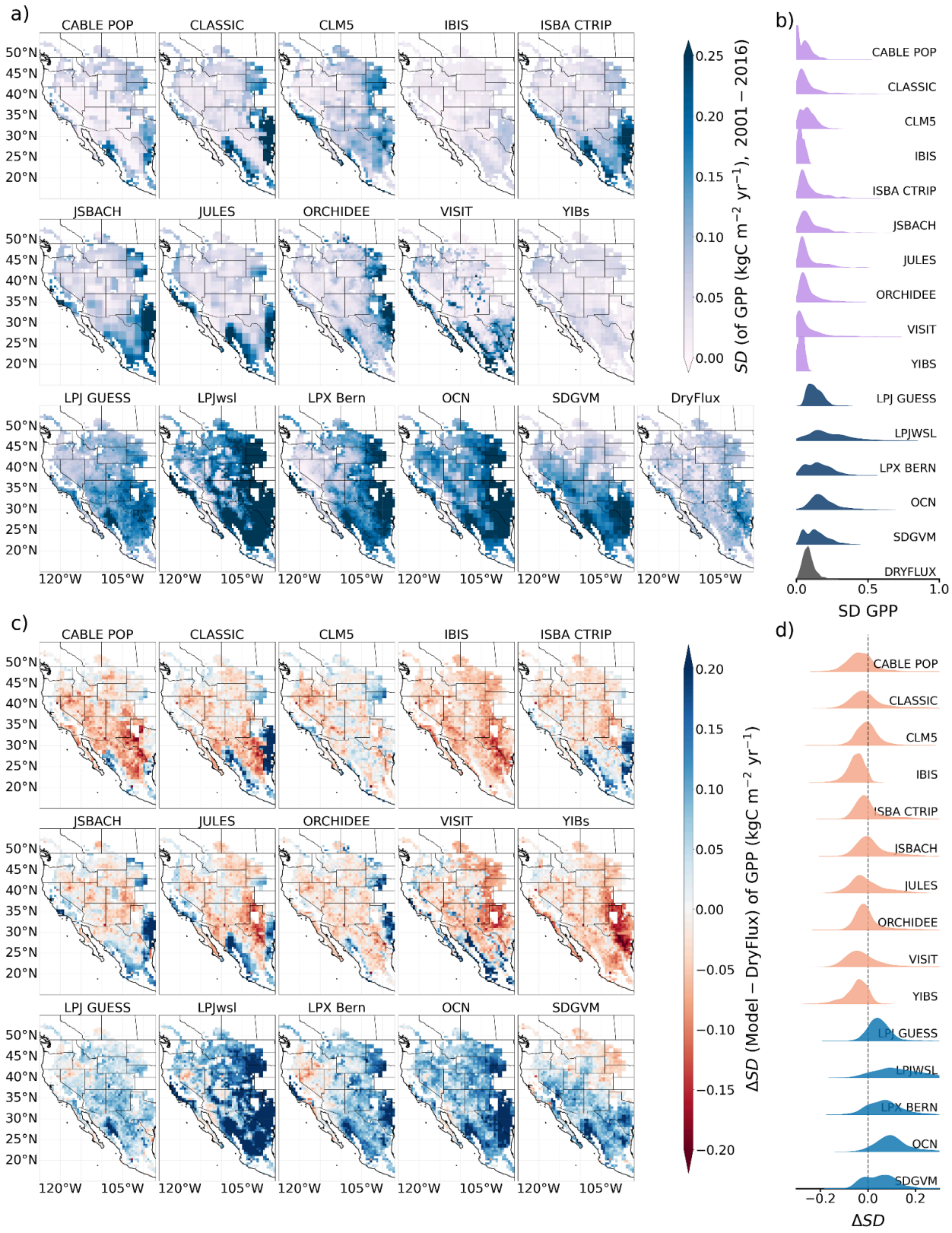

a)
CABLE POP
CLASSIC
CLM5
IBIS
ISBA CTRIP
JSBACH
JULES
ORCHIDEE
VISIT
YIBs
LPJ GUESS
LPJwsl
LPX Bern
OCN
SDGVM
DryFlux
50°N
45°N
40°N
35°N
30°N
25°N
20°N
120°W
105°W
0.25
0.20
0.15
0.10
0.05
0.00
SD (of GPP (kgC m−2 yr−1), 2001 − 2016)
b)
CABLE POP
CLASSIC
CLM5
IBIS
ISBA CTRIP
JSBACH
JULES
ORCHIDEE
VISIT
YIBS
LPJ GUESS
LPJWSL
LPX BERN
OCN
SDGVM
DRYFLUX
0.0
0.5
1.0
SD GPP
c)
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
ΔSD (Model − DryFlux) of GPP (kgC m−2 yr−1)
d)
−0.2
0.0
0.2
ΔSD

Figure 3: Spatial distribution across the western North American study area for: a) standard deviation in annual GPP (2001-2016) (as a measure of GPP IAV) for all TRENDY v11 models used in this study and the DryFlux data; b) a ridgeline plot of the Kernel Density Estimates (KDE) representing the area weighted spatial distribution of standard deviation in annual GPP for each of the TRENDY models and DryFlux; c) and d) same as a) and b) but for the differences in standard deviation in annual GPP between model and DryFlux (model – DryFlux). The vertical dashed line in the (d)shows a difference in annual GPP standard deviation of 0.0. The top two rows of (a) and (c) and the light purple and red KDE plots in (b) and (d) correspond to models that tend towards similar or lower standard deviation values than DryFlux. The bottom row of maps in (a) and (c) and the dark and lighter blue KDE plots in (b) and (d) correspond to models that tend to have higher standard deviations in annual GPP with a broader distribution of difference values.

We have removed the maps and KDE plots of the slope from the main text (original Figs. 3c and d) but we are considering including these in the supplementary along with scatter plots showing the slope vs R2 as described in the response to major comments above.

To support these new standard deviation difference figures (and in response to comments L552 and L558 below) we have calculated the mean of standard deviation differences (mean bias error) and the RMSE to provide an “effect size” of the difference in spatial patterns between the models and DryFlux and documented these values in a new Table 2. We have updated all text in Section 3.1 to refer to these new subfigures and new Table 2. We feel that providing a tracked change version of these paragraphs in Section 3.1 would be confusing. Instead, the following are the new sections of text that describe Figure 3 and Table 2:

“Figure 2 shows the time series (2001-2016) of annual GPP summed over the entire study area. Although the TRENDY ensemble mean total GPP (thick grey curves in Fig. 2) is lower than the DryFlux GPP (thick black curves in Fig. 2) the sign of the annual anomalies in TRENDY ensemble mean GPP match those of DryFlux well (Figs. 2a and b; R: = 0.96). However, individual TRENDY models simulate both higher and lower annual total GPP (Fig. 2a) and standard deviation in total annual GPP (as a measure of GPP IAV) (Figs. 2b and c) compared to DryFlux. Most models approximate DryFlux annual GPP IAV well or tend to underestimate the IAV brown curves/bars in Figs. 2b and c; % difference in total annual GPP standard deviation (model-DryFlux) of +18% to –69% in Fig. 2c). A select few models (OCN, SDGVM and the LPJ family of models) exhibit much higher annual GPP variability compared to DryFlux (and overestimate the DryFlux annual GPP magnitude) (blue curves/bars in Figs. 2a–c; >60% difference in total annual GPP standard deviation between model and DryFlux).

As seen in the total annual GPP time series (Fig. 2), most models generally have similar or slightly lower standard deviations in annual GPP (therefore similar or lower GPP IAV; top two rows of Fig. 3a) compared to DryFlux across much of the study area (except in the eastern and southeastern region – see below). This corresponds to mostly negative differences in standard

deviations of annual GPP between the model and DryFlux (Fig. 3c). In contrast, several models (LPJ-GUESS, LPJwsl, LPX-Bern, OCN, and SDGVM) generally have higher standard deviation in annual GPP (and positive differences in standard deviation between the models and DryFlux) (bottom row of Figs. 3a and c). LPJwsl and OCN in particular have larger regions of much higher standard deviation in annual GPP compared to DryFlux in the east and southeast of the study region (Figs. 3a and c). While LPJ-GUESS, LPX-Bern and SDGVM generally overestimate annual GPP variability, they display lower annual GPP variability compared to DryFlux in the northeastern or western regions (Fig. 3c).

The higher standard deviation of annual GPP in the east and southeastern part of the study area seen in the DryFlux product is captured well by most models except IBIS, VISIT and YIBs (Fig. 3a), although the magnitude of the standard deviation is higher than DryFlux for many models (Fig. 3c). This region corresponds to an area with high annual rainfall variability (Fig. 1c). We note that high values of standard deviation in annual GPP in the SE region do not necessarily correspond to high mean annual GPP values (Fig. S2). The KDE ridgeline plots for the differences in standard deviation in annual GPP (models – DryFlux) corroborate these spatial patterns: models in the top two rows of Figs. 3a and c either have KDE modes centered around zero difference in standard deviation or are skewed towards negative differences in standard deviation, albeit with long positive tails (Fig. 3d). In contrast, the KDE plots for the LPJ models, OCN and SDGVM (bottom row in Figs. 3a and c) exhibit much broader KDE distributions of positive differences in standard deviation in annual GPP between the models and DryFlux (Fig. 3d). The differences between model and DryFlux spatial patterns described here are captured by the mean bias error (MBE) and root mean squared error RMSE of the standard deviations in annual GPP (Table 2), which correspond to the area-weighted spatial mean of the differences in standard deviation (or the middle of the KDE plots) and the standard deviation in KDE distributions. Several models (CLASSIC, CLM5, ISBA-CTRIP, JSBACH and ORCHIDEE) have MBE values less than ±0.01 with RMSEs of <0.1.LPJwsl, LPX-Bern, SDGVM and OCN all have a much higher mean difference in annual GPP standard deviation (MBE) as well as positively skewed and broader KDE distributions (higher RMSE in standard deviation of annual GPP) (Figs. 3d and Table 1).”

Finally, we agree the Figure 3 quality was poor in the original manuscript. We have improved the quality and resolution of the figure for the resubmission.

L525-532 I find the conclusion a bit speculative here, and this is also reflected in the wording (e.g. L531 ‘possibly with the exception of SDGVM’). Given the LPJ model family is singled out here, it might as well be any other process this model family does differently compared to the other models. How can you conclude that your hypothesis

around dynamic vegetation is partially right, when you basically have only two model types in the dynamic vegetation which show different responses?

We completely agree that this section was too speculative. We have also reordered the manuscript so we discuss the relationship mean PFT type and fCover to annual GPP variability mismatches with DryFlux, so this section is now out of place. It also repeats many of the conclusions we come to later in Section 3.3. Therefore we removed this section and this decision also helped us to shorten the manuscript.

L552 It is not obvious to me how LPJ-GUESS captures the spatial patterns of GPP IAV better than any other model? Did you run any statistical test to support this claim?

We agree that in the original manuscript we did highlight LPJ-GUESS as the best model based on spatial patterns, which was wrong because actually many of the models that we originally described as underestimating DryFlux GPP IAV actually capture the standard deviations in annual GPP well – both in terms of the total annual GPP timeseries in Figure 2 and the spatial patterns and their area-weighted distributions shown in Figure 3. This is more obvious with the updates to Figures 2 and 3 suggested above (please see responses to L462/Figure 2, L469 and L507 above). Therefore, we have revised the description of which models best represent spatial patterns in Section 3.1 (see revised text in response to L507). As also described in our response to L507, we have also calculated metrics to support these claims, and these are documented in a new Table 2. This is not a statistical test of significance, but serves as to show the "effect size" (i.e. the magnitude of the differences between the models' and DryFlux's GPP IAV). We are not considering whether these differences are statistically significant for a couple of reasons:

1. We think effect sizes are more important to discuss here rather than significant differences. Significant differences would be more important if we were attempting to identify the "best" model(s). And while we acknowledge that we depicted LPJ GUESS as the "best" model in the original version of the manuscript, we did not in fact intend to identify the "best" model in this sense. Instead, we were using that to describe differences between models when we came to the section that this comment refers to, which focuses on dynamic vegetation enabled models. Our new text at the end of Section 3.1 does not refer to the best model.
2. Performing a test of statistical significance to identify which spatial pattern best matches that of DryFlux can be done in multiple ways, and all are relatively

complex due to the need to account for spatial autocorrelation. Given the manuscript is long, we do not want to inflate the methods and the results text by including this analysis.

Together with the revised text at the end of Section 3.1, we hope these new metrics (MBE and RMSE described in our response to L507) that describe the mean and spread in differences in standard deviation across the study area are sufficient to address this suggestion.

Accordingly, we have edited the text at this point in Section 3.3 to remove the part that refers to LPJ GUESS as the best model in terms of spatial patterns.

L553-554 Please reference the appropriate figure here (referring to the R2 values)

Good point, thank you. We have removed this reference because we have just cited the correct Figure 5 in the previous sentence.

L555-558 Would a non-linear fit maybe produce a better fit? Also for LPJ-GUESS (?) Where is the value 0.05 coming from?

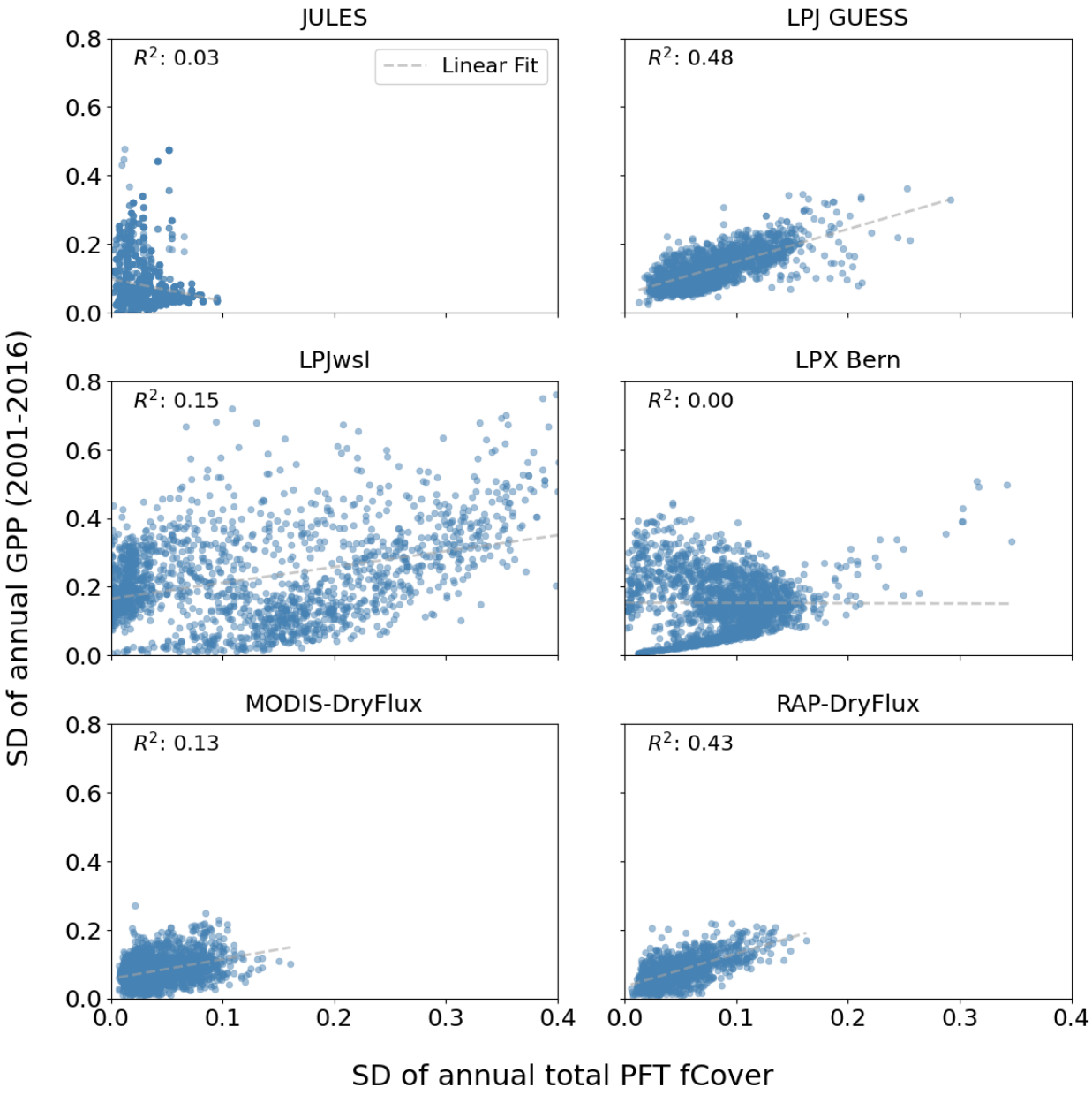
We have changed to using the standard deviation in annual GPP – rather than the coefficient of variation – as the coefficient of variation also has issues as a metric of variability (given it is a ratio to the mean). We also have decided to stick specifically to the use of standard deviations as our metric of IAV to simplify and shorten the text. Therefore, Figure 5 has been updated (see below) and is now Figure 6 as we have changed the order to show the maps of standard deviation in PFT fCover first. The relationships between standard deviations in PFT fCover are different from the coefficient of variation. In general the correspondence between standard deviation in annual total PFT fCover and standard deviation in annual GPP are much weaker (lower R2 values) and we no longer have the non-linear concave relationship mentioned here. In fact, we have deleted this entire first section of text and replaced it with a simpler version of the 2nd paragraph in Section 3.3 that describes both Figure 5 (original Figure 6) and the new scatterplot (now Figure 6) inserted below. Note that Figure 6 now includes a comparison with the two reference datasets as well. We provide the new Figure 6 with all these changes and the new section of text describing Figure 5 and 6 in response to comment L561 below.

L561 It would be useful to be consistent with the variable naming, the x-axis shows fCover? Defined as 100% - bare soil? Why were remotely sensed datasets not included in this analysis?

Good point, thank you. We added in “fCover” to the x-axis labels and edited the caption for consistency. We have also added the following sentence into the figure 5 caption: “The annual total PFT fCover is calculated as the sum of all PFT fCover minus fCover of bare soil or bare ground.”

We have now included comparisons between the DryFlux GPP and the two remotely sensed datasets into this figure (see below). Based on the changes described in our response to L555-558 and the additional reference dataset subplots added into this figure, we have a new, shortened section of text that describes both Figure 5 (original Figure 6) and Figure 6 (original Figure 5). This new text is an edited version of the 2nd paragraph of Section 3.3 in the original manuscript. However, we provide the revised version below for the sake of clarity.

“Dynamic vegetation enabled models exhibited much higher standard deviations in annual total PFT fCover compared to the reference datasets (Figs. 5 and 6). The standard deviation in annual total PFT fCover in the dynamic vegetation enabled models is mostly from the herbaceous PFTs , which is also true in two independent reference vegetation fCover datasets – MODIS VCF and RAP (last two columns in Fig. 5).The spatial patterns of standard deviation in annual PFT fCover predicted by JULES match both the MODIS VCF and RAP products well (Fig. 5). However, these spatial patterns do not correspond to the spatial patterns seen in JULES standard deviation of annual GPP (Figs. 3a; low R2 value of 0.03 in Fig. 6) and the c. All variants of the LPJ model have higher standard deviation in annual total PFT and grass fCover compared to MODIS VCF and RAP and the spatial patterns are not consistent with the reference datasets (Fig. 5). The LPJ models appear to be overestimating standard deviation in annual PFT fCover more in the arid central and southwestern part of the study area. These regions typically have sparse dwarf shrubs and higher bare soil cover compared to models (see RAP in Figs. S1a and b). LPJwsl also has a higher standard deviation in annual woody plant fCover in a north central and eastern region of the study area (and high mean woody plant fCover – Fig. S1b) that is not consistent with the two independent data products (Fig. 6). The positive bias in annual PFT fCover standard deviation in the LPJ models could help explain why those models also overestimate standard deviation in annual GPP (Fig. 3). However, the spatial patterns in standard deviation in annual PFT fCover (Fig. 6) do not match completely the spatial patterns in standard deviation in LPJ annual GPP (Fig. 3a; R2 values < 0.5 in Fig. 6). The exception is LPJ GUESS (cf.Fig. 5 with Fig. 3a). LPJ GUESS has an R2 of 0.48 between standard deviation in annual total PFT fCover and standard deviation in annual GPP, which matches well with the R2 between the RAP and DryFlux products.”



L593 Abrupt transition into new paragraph

Agreed. We have actually changed the start of this paragraph to compare the models to the two reference datasets (see our response to L619) and therefore have deleted the original sentence here. The opening of this paragraph now reads:

"We expected that models with fire enabled models would have higher grass cover, higher fCover variability, and higher GPP IAV. However, percent fire burned area IAV is low in the study area – both in the models and in the two satellite-derived reference datasets GFED5 and FireCCI 51 (Fig. 7). The two reference datasets exhibit closely aligned spatial patterns of burned-area variability; however, none of the models reproduce the dominant spatial gradients evident in the data (Fig. 7)."

L597 Again, why are those maps not shown (in the supplement?)

We initially did not do this just because we felt we were overloading the manuscript with plots. We have actually shortened this section considerably to focus on the key message that burned area variability is low and does not correspond well with either PFT fCover variability or GPP variability. Therefore we have removed the reference to mean burned area maps here but we have added the mean burned area to Supplementary Figure S5 and refer to it in the discussion Section 4.3 (please see our response to L807 below). We plan to follow up on this study with a more extensive comparison of fire burned area mean, IAV and trends in this region using ISIMIP simulations (see major comments above).

L599 I think this comparison is hard to make visually at least because all maps in Figure 7 have the same colorbar but very different magnitudes. I'm not saying it's not right to plot the map like this but could it be that some models have a similar pattern which just get lost in the value range?

We agree, in the new version of this figure 7 that includes reference satellite-derived burned area datasets (see our response to the following comment L619), we have allowed all models to have a different colorbar range. Otherwise, it is impossible to see any similarity in the spatial patterns between the models and between the models and the data.

L619 Why isn't there a reference dataset for remotely sensed burned area in this figure?

We agree there should have been and we have now added two satellite-derived burned area datasets as references: GFED5 and FireCCI51. We have also included two new subsections in the methods section 2.3.2 to describe these reference satellite derived

burned area datasets. Please see our response to L593 above for the new text that compares the models to these two reference datasets.

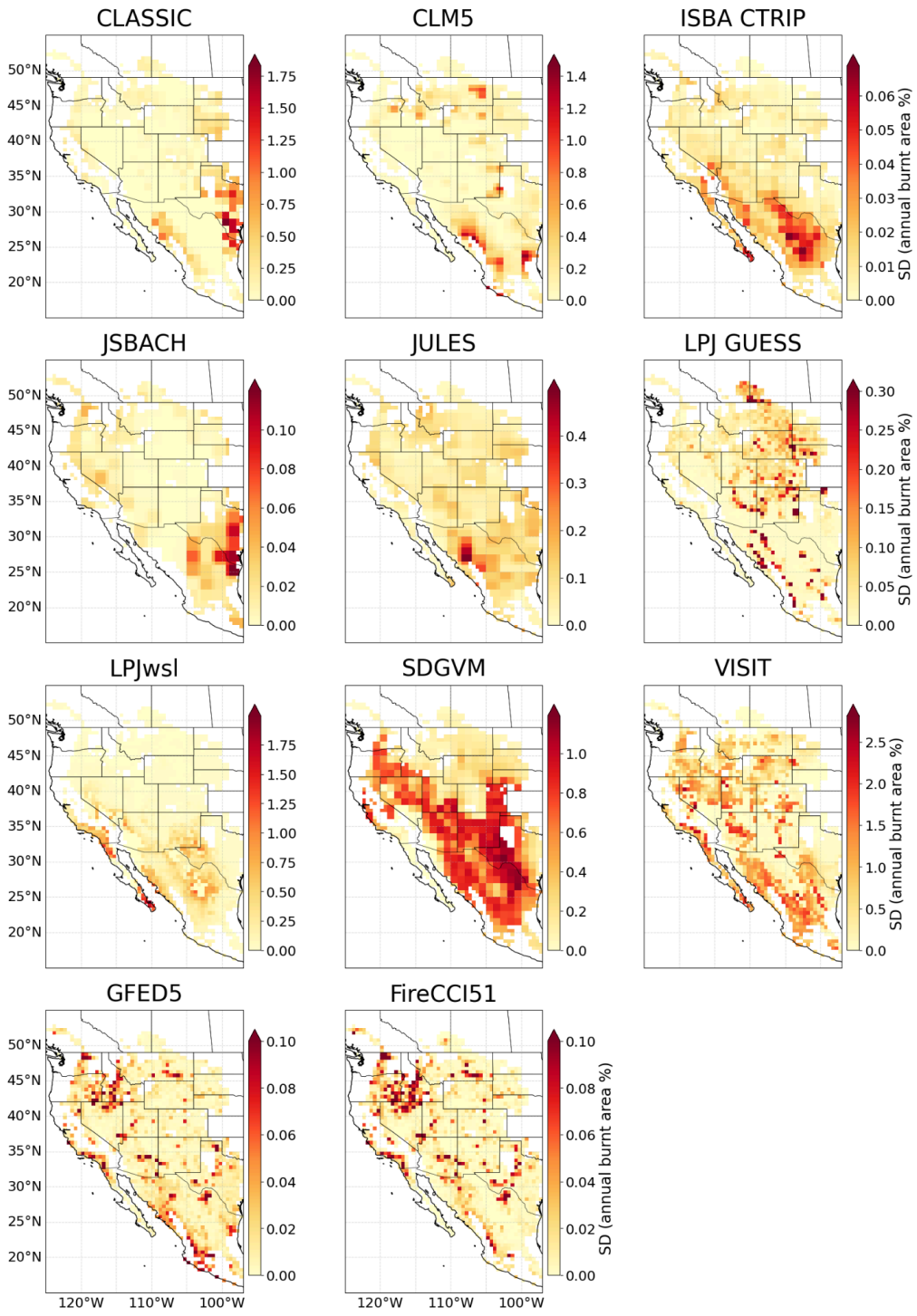


Figure 7: Spatial distribution across the western North American study area of the standard deviation in annual burnt area (%) for each grid cell for the 2001 to 2016 period for fire enabled TRENDY models, GFED5, and FireCCI51.

L623 Typo ‘showing for all the’ - please remove ‘for’. Again, would it not be interesting to also include what the comparison of the remotely sensed datasets would look like to get a sense for the ‘ideal’ relationship in the models?

Thank you, we have edited this Fig. 8 caption (see below) and we have added in the scatterplots between standard deviation in GFED and FireCCI burnt area and standard deviation in DryFlux GPP (see figure inserted below). Also see our responses to L619 above and L801 and L832-835 below for further details.

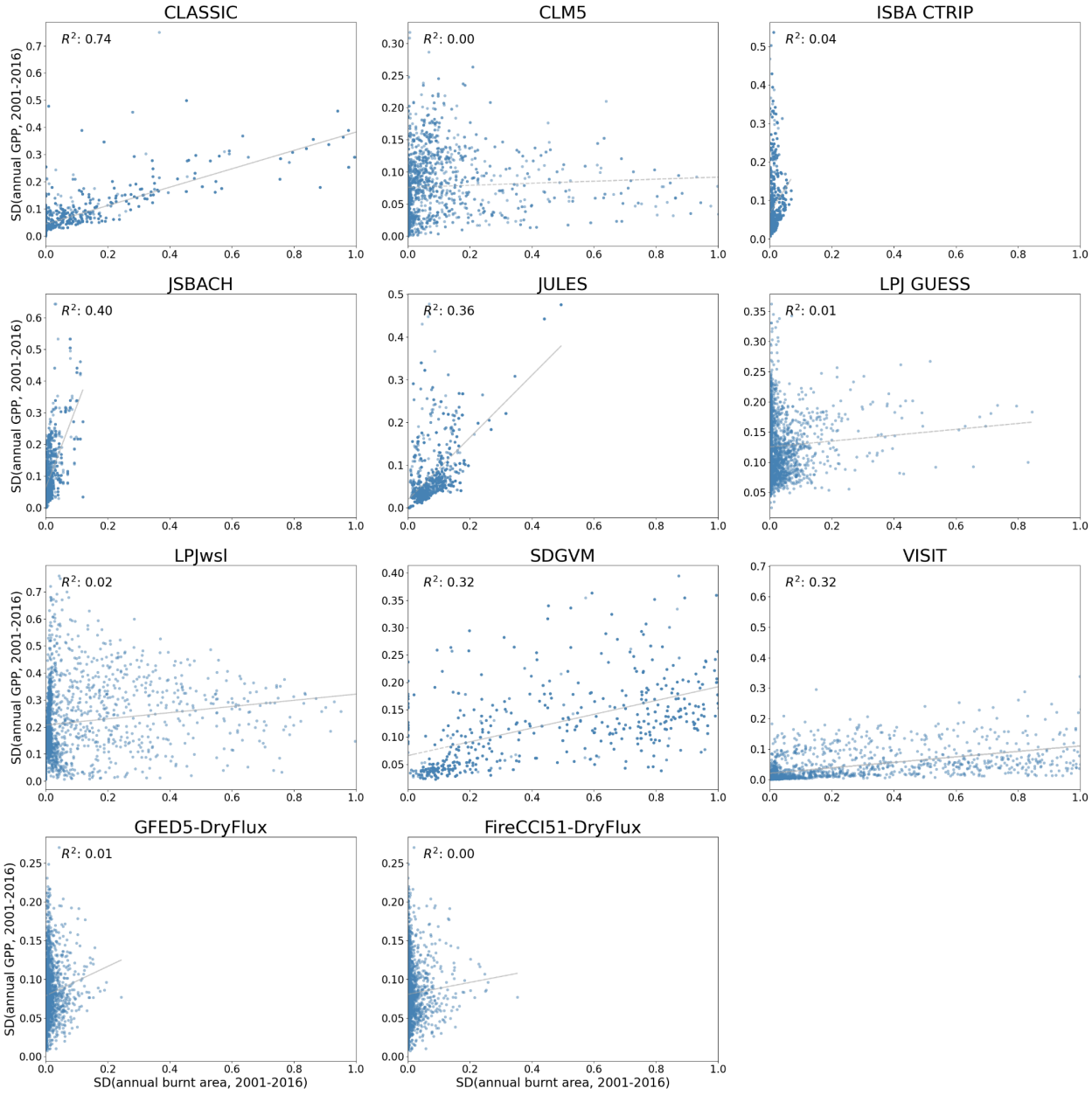


Figure 8: Scatterplots showing for each ~~all the~~ fire enabled TRENDY model and the two reference burned area datasets the pairwise grid cell comparisons of the standard deviation in of the total annual percent burnt area versus the standard deviation in annual GPP calculated over the 2001 to 2016 time series for each grid cell.

L692-713 This section is very thoughtful and covers a lot of potential shortcomings, but 1 page is quite long for something that can be boiled down to 'there is nothing wrong with the DryFlux dataset'. I assume this is also not meant to be the key result of this study so I would suggest to shorten that section and move it towards the end of the discussion where potential shortcomings in the study set up might be discussed (although clearly this dataset was not a shortcoming in this study)

We have shortened this section to remove the final sentences referring to the Pervin et al. study that has yet to be submitted following RC2's comment L708. We want to keep the rest of the text here though as we feel readers would wonder why we are not discussing uncertainties in DryFlux if we only mentioned it briefly.

L703 Typo, '..'

Thanks again for catching the Typo. Corrected.

L725-742 This section kind of lumps together different aspects of PFT distribution, and is a very long description of defining different bioclimatic limits which can be summarized to the fact that all models discussed here use a temperature threshold to define where C3 and C4 grasses are allowed to grow

We agree, we have removed most of this text from L726 to L739 in the original manuscript. We then merged the following paragraph with this one.

L755 Here it would be interesting to know (briefly) how this new distribution map was derived

Absolutely, here is the revised sentence: "A new approach **that leverages the contrasting seasonal peaks in C3 (winter) and C4 (summer) grass growth observed in remotely sensed vegetation indices** ~~utilizing the differences in timing of peak growth in C3 versus C4 grasses~~ has been ~~used~~ **applied** to map C3 and C4 grass distributions across Australia (Xie et al., 2022)."

L757 NEON hasn't been defined

Thanks for pointing this out. We added the definition of NEON.

L760-761 Then why wasn't this tested in this study?

Well, the short and honest answer to this is that we were only made aware of this new approach/study fairly late in the process of preparing this manuscript. But there are other reasons: we didn't run any of the TRENDY models used here. We are analysing their output (in collaboration with the modelers who performed the simulations). We could re-run one of the TRENDY models to test something like this, but that would have a) required an additional voluntary effort from one or some of the TRENDY model partners, and b) added additional analyses just for one or a few models and we didn't want to add additional analyses to an already bulky paper. Furthermore, while Luo et al. (2024) do provide netcdf files with their new C4 map, it is not trivial to replace the current C4 distribution in any model with this new map. If the new map suggests C4 should be in a location that C4 is not present in the model's original map, then we simple replace what was in the model's original map with C4. However, issues arise when the model's original map had C4 grass in a location, but the new map from Luo et al. (2024) does not. What do we put in the void created if the new map C4 grass is not actually there? Do we assume it's still grass, but a C3 grass? These kinds of subjective decisions cause issues in models all the time and are partly to blame for why prescribing vegetation distributions is so tricky in these models. However, we appreciate that the sentence, as is, implies what we suggested is a simple thing to do. We have replaced this sentence with the following broader sentence suggesting future studies look into the impact of alternative C4 distributions, but without going into the nuances detailed in this response for the sake of brevity and the need to shorten the manuscript: "Future studies should explore what impact alternative methods for deriving C4 distributions have on predictions of dryland vegetation and carbon cycle dynamics."

L788 Wasn't one conclusion that the (non-)inclusion of dynamic vegetation and/or fire could not be *really* be related to GPP IAV??

Partly true. It is hard to pinpoint exactly what is playing a role in inaccurate GPP IAV simulations but we think lack of dynamic vegetation may help to explain part of why some models underestimate GPP. We have revised this sentence so it now reads:

"**T**~~As expected, t~~he lack of dynamic vegetation **may**~~and fire~~ help~~s~~ to explain why some models have low GPP IAV compared to DryFlux. "

L792 Again, speculative wording ‘slight exception’

We have removed this part in parentheses as it was incorrect.

L801 But the ‘reality’ wasn’t shown in this analysis

We agree this was a big gap in the analysis as discussed in our response to the general comments above. We have added two independent reference datasets (GFED5 and FireCCI) to Figures 7 and 8 (see our responses to L619 and L623). This new analysis refutes this statement at L801 in the original manuscript. Instead, the comparison with two independent satellite-derived burned area datasets shows models tend to have a similar range in magnitude of variability in fire burned area, albeit that the spatial patterns in burned area variability do not match the reference datasets for any model. Therefore, we removed this part of the original sentence and the following sentence in the revised manuscript and edited the sentence so it now reads:

“All the dynamic vegetation enabled DGVMs included in this study also included a representation of fire; however, most of these models have low standard deviation in annual percentage burnt area, **with magnitudes (but not spatial patterns) matching reference datasets** (Fig. 7). ~~which does not represent reality~~.”

Otherwise the discussion points here remain the same – that fire is not causing high grass dominance in these models (or in reality).

L799 I wonder whether it would be better to refer to the models in TRENDY as terrestrial biosphere models (TBMs), because, as stated in the paper, not all TRENDY models simulate vegetation dynamically, which the name DGVM (dynamic global vegetation model) implies

We agree with the reviewer here. However, in all TRENDY papers they refer to the models as DGVMs (see for example Sitch et al. (2024), which is the most updated reference describing the TRENDY MIP) and we didn’t want to go against that standard practice. We also think that highlighting in this study that many of the “DGVMs” are not, in fact, DGVMs, and also highlighting that the ones that are, do not work so well, may contribute to the wider discussion of “what *is* a DGVM?” and “how should dynamic vegetation be modeled”. But this is a minor consideration. We are happy to change it though if the reviewer or editor still think it would be a good idea.

L807 This statement is not really supported by the analysis as ‘only’ the IAV is studied. Dominance of a PFT is also linked to long term averages, and if a region experiences on average high fire activity (even with low IAV), I’d expect grass dominance for this case too

Right, the reviewer is correct here. However, we also knew that the average burnt area was not that high either, we just did not show that data originally. We have added in “mean annual percent burned area” to this sentence and added a mean burned area plot to Supplementary Figure 5.

L808-809 ‘based on our knowledge of the study area’ is not an appropriate reference

Fair point. This part has been removed and the sentence revised to: “In the S3 simulations pasture covers much of these areas (as specified by LUH2). ~~(likely too much based on our knowledge of the study area)., but~~ However, LPJwsl does not consider a pasture category (Table S2), and JULES counts pasture as a natural grass in TRENDY simulations (S. Sitch, Pers. Comm.). While LPJ-GUESS and LPX-Bern do have high pasture fCover in the S3 simulations, they also predict high grass cover in ~~the even the LPJ~~ S2 simulations that do not include pasture land cover class ~~predict high grass cover~~ (data not shown). ”

L809 Why is this analysis not included?

It wasn’t included because we felt the paper was already too cumbersome. As we are trying to shorten the paper, we still have not shown the pasture fCover from S3 and S2 from the LPJ-GUESS or LPX-Bern simulations. We have provided the link to the files in the Data Availability statement so anyone can easily go and find those files and verify this information very quickly.

L828 This is nothing you can do about in this experiment set-up, but I would say it is inherently difficult to derive a meaningful relationship between fire and other ecosystem processes in any model using the GlobFIRM model which has been shown to neither capture mean or interannual variability in burned area anywhere in the world. The strong focus on fire does make me wonder whether it would have been more useful to make use of the ISIMIP simulations, with a similar experiment protocol compared to TRENDY AND an additional sensitivity experiment where fire is switched off.

We agree that using the ISIMIP simulations with fire switched on and off would be much more appropriate for this analysis. As we mentioned in our response in the major comment above those simulations were not available when we started this analysis. But we have added a sentence to have revised the sentence in Section 4.3 that discussed doing factorial simulations with and without fire (lines 839-840 in the original manuscript) to explicitly mention the ISIMIP simulations. THis sentence now reads:

"The role that fire plays in GPP variability in dryland ecosystems could be tested by comparing~~running factorial simulations with and without the fire module enabled. This is possible with~~ **ISIMIP3a historical simulations with and without fire enabled**. Such an analysis~~This~~ should be a priority ~~in the near term for~~ **to help** ~~further~~ diagnos**e**~~ing~~ the causes of **discrepancies in model** dryland **burned area and** C flux**es** ~~model-data discrepancies~~."

We are strongly considering doing such an analysis in the near future.

L832-835 This is purely a model result. Is this supported by observations? It might as well be an artefact where in those fire models burned area and GPP are too tightly

Originally we only wanted to explore what may be happening in the model. And actually this figure showed us that fire does not seem to be as strong a driver of GPP variability as we were expecting. However, we agree that without an independent reference dataset it is hard to untangle whether what is going on in the model is realistic or not, and therefore it is useful to examine the relationships seen in the independent data products, even though they have their own uncertainties that could confuse the relationships we "observe". Therefore, we have added scatterplots of standard deviation in GFED and FireCCI burnt area versus standard deviation in DryFlux GPP to Figure 8. Please see our response to L623 above. Indeed the reviewer is correct and we find that there is no relationship between either of the two satellite burned area products' IAV and DryFlux GPP IAV. Some models also show no strong relationship, but as we discussed in these lines in the original manuscripts, some models do show more of a clear relationship (with R2 values >0.3). We have revised the discussion in this section (first part of this paragraph) so it now reads:

"For many of the models that simulate higher standard deviation of annual percentage burnt area ~~(CLASSIC, SDGVM and VISIT)~~ there was a reasonable relationship with GPP variability (R2 values > 0.3; Fig. 8), **although most of these models (with the exception of SDGVM) still generally underestimated GPP IAV compared to**

DryFlux (Figs. 2 and 3). However, the independent datasets show no strong relationship between fire burned area variability and GPP variability (Fig. 8), indicating that models showing a reasonable relationship between burned area and GPP IAV may impose an overly strong coupling between fire and GPP. ~~This finding suggests that fire may play an important role in improving simulations of GPP IAV if it was better represented.~~ ”

L836 Not clear - does different fire parametrisations refer to different fire model ‘groups’, e.g. grouping all TBMs using GlobFIRM into one sub-group?

Yes, exactly. We have amended this sentence as follows to improve clarity: “Future evaluations of modeled burnt area should consider grouping models according to the different fire parameterisations **(e.g., grouping all GlobFIRM models together)** to try to disentangle the level of complexity needed”

L837 Typo - two full stops

Thank you, we have corrected this.

L840 As I mentioned above, these simulations actually already exist in a different MIP (ISIMIP)

Thank you. We agree. Please see our response to L828 above.

L847 Is the Walter hypothesis in the majority of TBMs or why is it highlighted here? What does it entail? If this detail is not relevant this mention should be removed

No, the hypothesis is not included in models per se. We have removed a mention of the Walter hypothesis.

L898-920 In section 4.5, the only reference to other literature is Bogucki et al. in prep. I noticed that this section was also written by L. Bogucki (according to the author contribution statement). I understand that it is tempting to rely on ongoing work but given this work is not even submitted yet, I would suggest the authors make an effort in supporting and validating their statements with published studies.

The unpublished paper in Section 4.5 – cited as “Bogucki et al. (in prep)” in the original submission – has since been submitted, undergone a round of revisions, and has just been accepted in *Environmental Research: Climate*, so we will include the reference here. We believe it will be published soon and we will update volume and page number in the final version. No other changes made to this section.

Bogucki, L., Feldman, A., Moore, D.J.P., Amaral, C., Wang, L., Green, J., Babst, F., Ojima, D., Pervin, R., Reed, S., Smith, W.K., Zhang, W. and MacBean, N. (2025) Tropical and subtropical drylands dominate variability in global net terrestrial carbon flux irrespective of different global ecosystem classifications and geographic scales. Accepted in *Environmental Research: Climate*.