Response to reviews of manuscript:

Validation of Terrestrial Biogeochemistry in CMIP6 Earth System Models: A Review

We thank each reviewer for their constructive comments. We have responded to each comment below. The reviewer comments are presented in *red italicized text*, while author responses are in regular font, and the relevant changes to the manuscript are in blue.

Referee 1 Comments:

This article presents a review of the current practices for validating the terrestrial biogeochemistry in CMIP6 Earth system models. The authors use the literature to show that the terrestrial biogeochemistry is a major source of uncertainty in future climate projections, and that this uncertainty can be linked to model structure. They study how 11 modeling groups participating in the CMIP6 exercise validated the terrestrial biogeochemical cycles in their land surface models and their fully coupled Earth System Models. They analyze the different validations presented by the modeling groups in terms of number of variables validated, quantity represented (for instance GPP), spatial and temporal scales, reference dataset and statistical metrics used. They also present two community methods designed to validate land surface models (ILAMBv2.1) and Earth system models (ESMValTool2.0). They present a critique of the validation approaches and suggest ways forward: mainly developing a standard protocol for validation which could be based on the existing communal software packages ILAMBv2.1 and ESMValTool2.0.

The article is very well written, very clear and very detailed. The authors read in detail the articles describing the validation of the CMIP6 models. They give a very detailed and informative description of the techniques used by each modelling group in Annex A.

I have a few detailed comments (see below) and only two general comments:

- *I agree that the use of communal validation software packages would be highly beneficial to the community. However, these should be used to help understand the behavior of the models and help improve them. But I would be reluctant to see them used as tools to select models allowed to participate in certain exercises. If this was the case it would contradict Lovenduski and Bonan's recommendation to improve process understanding and observation accuracy instead of reducing model spread (line 378). The authors don't suggest to use these validation softwares as selection tools but it might be implied by their remark about GCP on line 217.*

  We agree with and thank the referee for bringing this to our attention. We have adapted the manuscript as follows:

  Lines 212-218: "The validation of particular variables by different participants occasionally employed the same datasets, though in many cases inconsistent reference datasets were used for the same variable, and the spatial and temporal dimension of

validations was often distinct. This contrasts with other works employing multiple models such as the Global Carbon Project (Friedlingstein et al., 2020; 2019; Le Quéré et al., 2018) which provide ~~stringent quality~~ explicit validation criteria ~~for model inclusion~~, such as simulating recent historical net land-atmosphere carbon flux within a particular range and being within the 90% confidence interval of specified observations. The stringency of such criteria must be carefully chosen to acknowledge the role of observational uncertainty as well as uncertainty stemming from potential model tuning to forcing datasets. The use of different validation approaches impedes the comparison of performance across models, however it also provides a diverse collection of example methods."

- *The authors don't mention at all the links between the terrestrial biogeochemical cycles and the hydrological cycle. They mention that the modelers should evaluate the response of the biogeochemical cycles to temperature or the effect of nutrient limitations but never mention the effect of moisture conditions. I think the evaluation of the coupling between water and biogeochemical cycles at the land surface is as much of a concern.*

We thank the referee for bringing this important point to our attention. While validation of the physical components of land surface models and ESMs is outside of the focus of this review, we agree the moisture conditions are an important driving variable for terrestrial biogeochemistry. We have adapted the first and fourth sections of the manuscript as follows:

Lines 22-25: "The future capacity of the terrestrial biosphere to sequester CO2 emissions is uncertain due to non-linear feedbacks such as CO2 fertilization, growing season extension in cold-limited regions, enhanced heterotrophic respiration, and potentially other~~s~~ feedbacks, as well as environmental and physiological constraints such as moisture availability, nutrient limitations, and stomatal closure (Fleischer et al., 2019; Green et al., 2019; Xu et al., 2016; Wieder et al., 2015)."

Lines 245-250: "Another approach to validation which combines high-level variables and re-parameterization efforts is the assessment of functional relationships or emergent constraints, such as the relationship between GPP or turnover times and temperature, moisture, growing season length, and nutrient stoichiometry (Danabasoglu et al., 2020; Swart et al., 2019; Anav et al., 2015; McGroddy et al., 2004). Physically interpretable emergent constraints can aid in identifying model components which are particularly influential to climate projections (Eyring et al., 2019), such as the temperature control on carbon turnover in the top metre of soil in cold climates (Koven et al., 2017), GPP responses to soil moisture availability (Green et al., 2019), or regional carbon-climate feedbacks (Yoshikawa et al., 2008)."

Lines 388-391: "Field experimental data provide unique insight as to the functional responses of vegetation to elevated CO2 concentration (Goll et al., 2017), temperature change (Richardson et al., 2018), moisture availability (Williams et al., 2019; Hovenden

and Newton, 2018), and nutrient limitations (Fleischer et al., 2019), outside the current context of observations."

Lines 405-410: "Additionally, the universal inclusion of often overlooked processes such as moisture limitation, nitrogen and phosphorus cycles, dynamic vegetation, prognostic leaf phenology, and natural disturbance regimes should be a priority focus for participants in developing diagnostic models as these processes are highly influential on terrestrial biogeochemistry and physics (Eyring et al., 2020; Fleisher et al., 2019; Piao et al., 2019; Wieder et al., 2015; Achard et al., 2014; Richardson et al., 2013; Heimann and Reichstein, 2008; Tucker et al., 1986), and their omission contributes to widespread bias (Green et al., 2019; Anav et al., 2015). While outside the focus of this review, equal attention should be applied to the physical components of terrestrial biogeochemical cycles, including explicit representation of permafrost and riverine carbon transport dynamics. In fact, a study including four CMIP5 ESMs found that soil moisture variability prompted variability in terrestrial NBP on the order of gigatonnes, with non-linear responses to both moisture scarcity and excess (Green et al., 2019). Further, many of the merits and limitations of the validation approaches discussed herein apply to the validation of these physical components as well."

Lines 417-419: "The strategic situation of nitrogen, ~~and~~ phosphorus, and soil moisture monitoring which coincides with current Fluxnet sites (Jung et al., 2020) could provide high fidelity insight as to nutrient and environmental limitations on GPP, coherent turnover time assessments, and broadly applicable functional relationships to facilitate upscaling."

Detailed comments:

L94 I find the figure a bit difficult to understand. An example would help like GPP being the variable corresponding to the left-most bar (if I understand correctly).

To improve the clarity of Figure 2, we have clarified the figure caption as follows.

"Frequency of a given variable being validated ~~by~~ across participants (treating ESMs and LSMs separately). Most variables were validated only once across participants (leftmost x-axis), while GPP was validated by 11 participants (rightmost bar)."

Table 6: the authors choose to separate vegetation carbon and vegetation biomass although these two type of data are very much related. I guess this is due to a choice made by iLAMB and ESMValTool. But Saatchi et al for instance, although being a tropical biomass dataset also gives data in terms of aboveground and belowground vegetation carbon mass. I'd suggest adding a comment specifying that these 2 quantities are not independent.

We have modified the caption of Table 6 as follows: "Table 6: Select observation-based reference dataset sources for ESMValToolv2.0 (Eyring et al., 2020) and ILAMBv2.1 (Collier et al., 2018), including Net Biome Production (NBP), Leaf Area Index (LAI), Land Cover (LC), Gross Primary Production (GPP), 205 Net Ecosystem Exchange (NEE), Soil Carbon (SC),

Vegetation Carbon (VC), Ecosystem Carbon Turnover (ECT), Vegetation Biomass (VB), and Burned Area (BA). Note that vegetation carbon is dependent upon vegetation biomass. ”

L215: The problem with the quality criteria of the Global Carbon Project is that they neglect the uncertainties related to the atmospheric forcing. As shown by Lawrence et al, 2019 (and also shown in Figure 3 in this paper) these may have a pretty strong impact. GCP switched from the CRU-NCEP forcing to the GSWP3 forcing in 2019, and model results were very different. Modeling teams had to retune their models to fit the criteria. That doesn't mean criteria shouldn't be imposed, it just shows their limit.

To acknowledge the limitations of imposing quality criteria, we have modified the manuscript as shown above in response to the first general comment and below as well:

Lines 212-218: “The validation of particular variables by different participants occasionally employed the same datasets, though in many cases inconsistent reference datasets were used for the same variable, and the spatial and temporal dimension of validations was often distinct. This contrasts with other works employing multiple models such as the Global Carbon Project (Friedlingstein et al., 2020; 2019; Le Quéré et al., 2018) which provide ~~stringent quality~~ explicit validation criteria ~~for model inclusion~~, such as simulating recent historical net land-atmosphere carbon flux within a particular range and being within the 90% confidence interval of specified observations. The stringency of such criteria must be carefully chosen to acknowledge the role of observational uncertainty as well as uncertainty stemming from potential model tuning to forcing datasets. The use of different validation approaches impedes the comparison of performance across models, however it also provides a diverse collection of example methods.”

L 249: I would argue that the soil moisture control on all the variables of the biogeochemical cycles is as crucial in a climate change perspective as temperature, if not more because more uncertain (for instance the future of peatland/wetland in the high latitudes)

We have modified the manuscript as follows (shown above as well):

Lines 245-250: “Another approach to validation which combines high-level variables and re-parameterization efforts is the assessment of functional relationships or emergent constraints, such as the relationship between GPP or turnover times and temperature, moisture, growing season length, and nutrient stoichiometry (Danabasoglu et al., 2020; Swart et al., 2019; Anav et al., 2015; McGroddy et al., 2004). Physically interpretable emergent constraints can aid in identifying model components which are particularly influential to climate projections (Eyring et al., 2019), such as the temperature control on carbon turnover in the top metre of soil in cold climates (Koven et al., 2017), GPP responses to soil moisture availability (Green et al., 2019), or regional carbon-climate feedbacks (Yoshikawa et al., 2008).”

L320-324: I totally agree that site-level evaluation is very important to really understand how a model behaves. This is partly because at site-level, there is usually much more information available: the type of vegetation, the type of soil, the presence of an aquifer, land-use practices (irrigation, multi-cropping) etc. The meteorological forcing is also much more precise. In a

global simulation, this type of information comes from global dataset (for instance soil texture, depth to bedrock, vegetation type) or are calculated. And this adds a huge part of uncertainty not related to the model structure but to model dataset. I think it might be interesting for the readers to get that insight

We thank the reviewer for their comment. We have added the following to the manuscript to highlight the value of site-level evaluations:

Lines 318-326: "Despite these caveats, such global-scale data products provide a critical resource to the CMIP community in conducting model validation (Collier et al., 2018), and the relatively common use of Jung et al. (2011) for validations by CMIP6 participants coincidentally reduces the influence of observational contradiction (Xie et al., 2020; Anav et al., 2015). Site-level GPP evaluation with observations from the tropics by Delire et al. (2020) and Vuichard et al. (2019) demonstrates a strategic approach to addressing the representation bias in GPP validations. Site-level evaluations often benefit from a wealth of available information including spatially consistent meteorological forcing, and avoid the influence of spatial extrapolation error. While Jung et al. (2011) do not provide uncertainty measures, several forms of uncertainty are explicitly presented for the Fluxnet2015 dataset by Pastorello et al. (2020). Therefore the utility of Fluxnet GPP data products could be improved with standardized use by participants in junction with other independent data products, select site-level evaluations, explicit uncertainty quantifications, and improved ecological representation in underlying site-level data."

L419: I probably miss something here but I don't see the link between N and P monitoring and turnover time assessment

We have addressed the confusion regarding the potential strategic link between N, P, and turnover time monitoring as follows:

Lines 417-422: "The strategic situation of nitrogen, ~~and~~ phosphorus, and soil moisture monitoring which coincides with current Fluxnet sites (Jung et al., 2020) could provide high fidelity insight as to nutrient and environmental limitations on GPP, coherent turnover time assessments, and broadly applicable functional relationships to facilitate upscaling. The co-situation of multiple observational monitoring objectives at Fluxnet sites would enhance the utility of each site-level dataset and alleviate errors due to spatiotemporal inconsistencies between datasets in both performing evaluations and developing large scale data products. Following increased collaboration between empirical and modelling communities to strategically expand observations, and their inclusion in a comprehensive evaluation software, the CMIP-designated use of such software would standardize, conserve, and augment validation efforts."

Technical comment

L23: I am not a native speaker but I find this sentence strange: "growing season extension in cold-limited regions, enhanced heterotrophic respiration, and potentially others, as well as environmental"

To improve the readability of this sentence, we have modified the manuscript as follows:

Lines 22-25: “The future capacity of the terrestrial biosphere to sequester CO2 emissions is uncertain due to non-linear feedbacks such as CO2 fertilization, growing season extension in cold-limited regions, enhanced heterotrophic respiration, and potentially others feedbacks, as well as environmental and physiological constraints such as moisture availability, nutrient limitations, and stomatal closure (Fleischer et al., 2019; Green et al., 2019; Xu et al., 2016; Wieder et al., 2015).”

L378: I believe it should be “precedence” instead of “president”

This has been fixed.

Referee 2 Comments:

This paper does an excellent job of reviewing the land biogeochemistry subcomponent evaluation approaches conducted by different modeling groups for 16 CMIP6-generation land surface models, both in a coupled and uncoupled context. This is useful since finding all the individual land model evaluation papers can be challenging. While Table 1, reprinted from Arora et al. (2020), provides such references, it lists only 11 models, not the 16 models identified in Figure 1. This paper would better serve the community with a table similar to Table 1 that includes all the models reviewed in the paper.

While the scope of this paper is to review only the land biogeochemistry subcomponents of land surface models, describing differences in hydrology and radiation/energy subcomponents and summarizing the assessments of them would be beneficial. Even if the authors do not wish to expand the scope to include other interacting model subcomponents, I recommend adding more key model configuration characteristics to an updated Table 1 to provide additional context to the summary. For example, the table could identify which models employ a river transport scheme, which models have depth-resolved soil carbon, which have an explicit permafrost representation, which used a prognostic dynamic vegetation scheme, etc.

The authors discuss the use of community-developed model evaluation and benchmarking packages, ESMValTool and ILAMB, which are increasingly being adopted as a means of standardizing model evaluation metrics and observationally constrained reference datasets. The wide variety of variables, datasets, and metrics employed in assessing model performance by different modeling groups makes direct comparison across models challenging.

Section 4.1, Variable Choice, begins with, "A comprehensive validation of a process-based model should include all simulated interacting variables for which a reliable empirical reference is available." However, the discussion that follows includes only biogeochemistry variables and there is little acknowledgement of the important interactions with water and energy variables. It may be useful to add a sentence or two that explicitly mentions the interdependence of and need to co-evaluate carbon, water, and energy cycles.

We agree that the validation of hydrology, radiation, and energy subcomponents is equally important to review, however we feel that such an effort would be better served with its own paper led by authors who are experts in these topics. We have adapted the manuscript to acknowledge the importance of physical model component validations in response to a similar comment by Reviewer 1 (see the above response to the second general comment as well). We have added the following to the manuscript to convey this to readers:

Lines 22-25: “The future capacity of the terrestrial biosphere to sequester CO2 emissions is uncertain due to non-linear feedbacks such as CO2 fertilization, growing season extension in cold-limited regions, enhanced heterotrophic respiration, and potentially other~~s~~ feedbacks, as well as environmental and physiological constraints such as moisture availability, nutrient limitations, and stomatal closure (Fleischer et al., 2019; Green et al., 2019; Xu et al., 2016; Wieder et al., 2015).”

Lines 245-250: “Another approach to validation which combines high-level variables and re-parameterization efforts is the assessment of functional relationships or emergent constraints, such as the relationship between GPP or turnover times and temperature, moisture, growing season length, and nutrient stoichiometry (Danabasoglu et al., 2020; Swart et al., 2019; Anav et al., 2015; McGroddy et al., 2004). Physically interpretable emergent constraints can aid in identifying model components which are particularly influential to climate projections (Eyring et al., 2019), such as the temperature control on carbon turnover in the top metre of soil in cold climates (Koven et al., 2017), GPP responses to soil moisture availability (Green et al., 2019), or regional carbon-climate feedbacks (Yoshikawa et al., 2008).”

Lines 388-391: “Field experimental data provide unique insight as to the functional responses of vegetation to elevated CO2 concentration (Goll et al., 2017), temperature change (Richardson et al., 2018), moisture availability (Williams et al., 2019; Hovenden and Newton, 2018), and nutrient limitations (Fleischer et al., 2019), outside the current context of observations.”

Lines 405-410: “Additionally, the universal inclusion of often overlooked processes such as moisture limitation, nitrogen and phosphorus cycles, dynamic vegetation, prognostic leaf phenology, and natural disturbance regimes should be a priority focus for participants in developing diagnostic models as these processes are highly influential on terrestrial biogeochemistry and physics (Eyring et al., 2020; Fleisher et al., 2019; Piao et al., 2019; Wieder et al., 2015; Achard et al., 2014; Richardson et al., 2013; Heimann and Reichstein, 2008; Tucker et al., 1986), and their omission contributes to widespread bias (Green et al., 2019; Anav et al., 2015). While outside the focus of this review, equal attention should be applied to the physical components of terrestrial biogeochemical cycles, including explicit representation of permafrost and riverine carbon transport dynamics. In fact, a study including four CMIP5 ESMs found that soil moisture variability prompted variability in terrestrial NBP on the order of gigatonnes, with non-linear responses to both moisture scarcity and excess (Green et al., 2019). Further, many of the merits and limitations of the validation approaches discussed herein apply to the validation of these physical components as well.”

Lines 417-419: “The strategic situation of nitrogen, ~~and~~ phosphorus, and soil moisture monitoring which coincides with current Fluxnet sites (Jung et al., 2020) could provide high fidelity insight as to nutrient and environmental limitations on GPP, coherent turnover time assessments, and broadly applicable functional relationships to facilitate upscaling.”

We have also clarified the scope of the manuscript as follows:

Lines 58-63: “Here we focus on validations of the stocks and biological fluxes of fully coupled ESMs and associated LSM releases from 2017 onwards with explicit terrestrial biogeochemical cycle representation contributed by CMIP6 participating modelling groups (hereafter participants; Table 1; Arora et al., 2020). Validations are analyzed in terms of variables included, spatiotemporal scales, reference datasets, and metrics of performance. Section 2 compares the methods of historical terrestrial biogeochemical cycle validation used by participants, Section 3 summarizes the methods used in community analyses of CMIP5 era models, and a critique of these methods. A future outlook is presented in Section 4.”

We have updated Table 1 as shown below. All 16 models included in Figure 1 are provided in the table, though the recently updated land surface model component (where available) is referenced on the same row as each ESM to preserve space and avoid redundancy (each ESM has identical model configurations (ie. dynamic vegetation, prognostic LAI, etc.) as its associated LSM). We have provided more information for readers in Table 1, though did not include model features which are not consistently included in descriptions of land surface model components for the included models (ie. riverine carbon transport). We felt adding these additional model features to Table 1 would confuse readers if we did not mention these features throughout the manuscript. In addition, further model configuration details are provided in Table 2 of Arora et al. (2020).

"Table 1: Modelling group contributions to C^{4}MIP of CMIP6 from Arora et al. (2020).

Modelling Group	**ESM**	**Land Surface Model Biogeochemistry Component**	**Explicit N Cycle**	**Dynamic Vegetation**	**Prognostic LAI**	**Prognostic Leaf Phenology**	**Reference(s)**
CSIRO	ACCESS-ESM1.5	CABLE2.4	Yes	No	Yes	No	Ziehn et al., 2020
BCC	BCC-CSM2-MR	BCC-AVIM2	No	No	Yes	Yes (for deciduous)	Wu et al., 2019; Li et al., 2019
CCCma	CanESM5	CLASS-CTEM	No	No	Yes	Yes	Swart et al., 2019
CESM	CESM2	CLM5	Yes	No	Yes	Yes	Danabasoglu et al., 2020; Lawrence et al., 2019
CNRM	CNRM-ESM2-1	ISBA-CTRIP	No	No	Yes	Yes (from leaf carbon balance)	Séférian et al., 2019; Delire et al., 2020
GFDL	GFDL-ESM4	LM4.1	No	Yes	-	-	Dunne et al., 2020
IPSL	IPSL-CM6A-LR	ORCHIDEE, version 2.0	No	No	Yes	Yes	Boucher et al., 2020; Vuichard et al., 2019
JAMSTEC	MIROC-ES2L	VISIT-e	Yes	No	Yes	Yes	Hajima et al., 2020
MPI	MPI-ESM1.2-LR	JSBACH3.2	Yes	Yes	Yes	Yes	Mauritsen et al., 2019; Goll et al., 2017
NCC	NorESM2-LM	CLM5	Yes	No	Yes	Yes	Seland et al., 2020
UK	UKESM1-0-LL	JULES-ES-1.0	Yes	Yes	Yes	Yes	Sellar et al., 2019

"

The authors do not describe the difference between metrics useful for evaluating offline models versus those that can be used for fully coupled models. Fully coupled models should exhibit the same statistical variability over decadal time scales as indicated by observational data, but the observed timing of ENSO and other drivers of climate variability are not reproduced in fully coupled models. Thus, metrics that assess biases or RMSE of time series model output should not be used when evaluating fully coupled model output. This discussion of applicable approaches should likely be included in Section 4.3.

To acknowledge the need for model specific evaluation metrics, we have altered the manuscript as follows:

Lines 327-374:"**4.3 Statistical Metrics and Validation Approaches**
Several participants relied primarily on residual-based metrics such as bias (simulated-observed) for validation of terrestrial biogeochemical cycle model components. On a spatial basis bias can identify significant regional over- or under-estimations of a given variable. However, the attribution of model error from global maps of bias can be ambiguous, as the displayed bias is the combined result of different forms of uncertainty, including model structural representations, unforced variability, and spatial disagreement (Deser et al., 2020; Lovenduski and Bonan, 2017; Koch et al., 2016). Such residual-based metrics may not indicate how well the model would perform in simulating future conditions beyond the current contextual envelope of observations (Gulden et al., 2008), and neglect the contribution of uncertainty from observations. These limitations are considerable in the context of ESMs and LSMs as tools for predicting terrestrial biogeochemical function. A more contextualized bias assessment is the Wilcoxon test as applied by Swart et al. (2019) to filter insignificant bias. In a LSM evaluation, Orth et al. (2017) provides an observationally robust bias assessment by subtracting mean seasonal cycles from each grid cell and correlating the resulting anomalies between observation-based datasets and model output. In addition, RMSE normalized by the mean or standard deviation of the observed quantity, NRMSE, contextualizes the difference between simulated and observed variable quantities in terms of the magnitude or inherent variability of the variable of interest (Swart et al., 2019; Fan et al., 2018), which is advantageous for variables such as GPP with large interannual variability. Caution is warranted however in the evaluation of fully coupled model output due to the inability of fully coupled models to reproduce the timing of internal climate variability phenomena such as El Niño-Southern Oscillation (ENSO) and volcanic eruptions (Flato et al., 2013). While the magnitude of observed and simulated internal climate variability may be statistically consistent, bias, RMSE, and NRMSE assessments of fully coupled model output should encompass decadal or longer periods to address the influence of temporal mismatches in simulated internal climate variability relative to observational records. Alternatively, as offline simulations can be directly forced with historical observation data, the output of offline simulations can be validated on a finer temporal scale.

Beyond these, a variety of targeted model skill metrics have been published for process-based modelling which provide detailed assessments of different forms of model uncertainty (Collier et al., 2018; Orth et al., 2017; Eyring et al., 2016b; Koch et al., 2016; Law et al., 2015; Kumar et al., 2012; Taylor, 2001; Kobayashi and Salam, 2000). Mean squared deviation, the sum of squared bias, squared difference between standard deviations, and lack of correlation weighted by standard deviations, presented by Kobayashi and Salam, (2000), was used by Vuichard et al.

(2019). This metric is readily applicable to the objective validation and improvement of mechanistic models, as its dissection allows for the accurate attribution of different sources of model errors. Additionally, a Taylor diagram (Fig. 4, Taylor, 2001) conveys several dimensions of model error and allows for the concise simultaneous display of variables and models and was utilized in the evaluation of BCC-AVIM2 (Li et al., 2019), and NORESM2 (Seland et al., 2020), as well as several LSMs and ESMs by Anav et al. (2015) and is incorporated into ILAMBv2.1 (Collier et al., 2018). The Taylor diagram was designed for simultaneous performance comparison of several simulated variables and serves as a concise and informative validation tool.

The validation process of terrestrial biogeochemical cycles and dissection of model uncertainty may also be enhanced through offline simulations or models with intermediate complexity as these allow for a greater replication of simulations with different initializations, forcing datasets, and model configurations, due to their computational affordability (Bonan et al., 2019; Umair et al., 2018; Orth et al., 2017). Offline simulations also reduce the potential for incidental compounding error from coupling components, though this leads to an underestimation in uncertainty for equivalent fully coupled simulations. Replicate simulations with different initial conditions allow for the attribution of uncertainty from unforced variability, such as performed by Danabasoglu et al. (2020), which accounted for half of the inter-model spread in key variables previously (Deser et al., 2020; Eyring et al., 2019). In addition, replicate simulations with different forcing datasets can indicate the role of forcing uncertainty (Wei et al., 2018), which Lawrence ~~and Bonan~~ et al. (2019) found to be significant. Further, sensitivity analyses or perturbed parameter analyses involving replicated simulations with one or more variables fixed as performed by Hajima et al. (2020) and Lawrence et al. (2019) illuminate structural uncertainty. The use of well-established statistical and model performance metrics in addition to strategic simulations facilitates a detailed analysis of model uncertainty."

"

The citation to Lawrence and Bonan et al. (2019) in line 371 should likely be Bonan et al. (2019).

We have fixed this error (see above). The intended citation was Lawrence et al. (2019), who demonstrated the important role of forcing uncertainty through the use of three different forcing datasets.

On line 378, "president" should be "presedence".

We have fixed this error.

On line 426, "in junction" likely should be "in conjunction".

We have fixed this error.

Developing a standard validation protocol for model intercomparison activities within CMIP would be useful, and it has been done to some extent for CMIP6 historical land and ocean model performance in comparison with corresponding CMIP5 models in the IPCC AR6 Working Group I

report in Figure 5.22, currently accessible at https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_Chapter_05.pdf#page=214

We thank the reviewer for bringing this to our attention. It is great to see some work has already been done to address the need for a standard validation protocol in AR6.

***Citation**: https://doi.org/10.5194/gmd-2021-150-RC2*

References

Green, J. K., Seneviratne, S. I., Berg, A. M., Findell, K. L., Hagemann, S., Lawrence, D. M., and Gentine, P.: Large influence of soil moisture on long-term terrestrial carbon uptake, Nature, 565, 476-479, https://doi.org/10.1038/s41586-018-0848-x, 2019.

Hovenden, M., and Newton, P.: Plant responses to CO2 are a question of time, Science, 360, 263-264, https://doi.org/10.1126/science.aat2481, 2018.

Williams, K. E., Harper, A. B., Huntingford, C., Mercado, L. M., Mathison, C. T., Falloon, P. D., Cox, P. M., and Kim, J.: How can the First ISLSCP Field Experiment contribute to present-day efforts to evaluate water stress in JULESv5.0?, Geosci. Model Dev., 12, 3207–3240, https://doi.org/10.5194/gmd-12-3207-2019, 2019.