The manuscript by Koven et al. highlights results from long-term projections of the CMIP6 models. It addresses an important issue of a legacy of carbon emissions in the Earth System dynamics. The period until 2300 is long enough to establish responses of ocean and land uptakes to stabilized CO2 concentrations. The paper is definitely important for the Earth Science community and should be published, but only after the authors address my concerns about three major weaknesses of the current manuscript.

1) First, the title promises surprises. Surprise is something one wouldn't expect, but here most of results are the same as one would expect from the MAGICC simulations or from simulations with ESMs until 2100. The last (very long) sentence in the abstract says about the possibility of surprises beyond the 21st century. Two reasons are: (i) the lack of agreement among the land models. I cannot count this as a surprise, it is a usual finding in the land model intercomparisons that land response to warming and CO2 increase is very different among models, see e.g. Arora et al. (2020); (ii) a recovery and overshoot of AMOC above the pre-industrial level in CESM2-WACCM in SSP5-3.4 simulation. This feature is indeed interesting enough and can be counted as surprise, but in that case, I miss important details. What are the physical mechanisms – is it a feature of sea-ice instability or deep convection change? If this feature is brought to the top message in the title, the authors should invest more time into analysis of what have happened in the ocean circulation and carbon cycle response. Preferably they show a geographical map of patterns of land and ocean carbon uptake, either averaged over the last couple of decades or integrated over the 23rd century. The AMOC overshoot is a rare event with potentially important implications for the carbon cycle. Independently of how plausible the AMOC overshoot is, it makes sense to investigate how land/ocean carbon uptakes are changing in response to the AMOC recovery. Up to know, discussion in the literature was mostly about slowdown of AMOC and its effect on climate and carbon cycle. How does it work when AMOC overshoots, what are implication for ecosystems on land and carbon uptake in the ocean? The authors need either to make it clear and justify what was unexpected, or rename the paper and update the abstract.
2) Second, there are limitations of concentration-driven runs in analysis of TCRE (Fig. 3). Emissions in these runs are not purely anthropogenic but ESM-inferred emissions. Using monotonously increasing CO2 scenarios is all right, but when concentrations start to decline, one can do wrong conclusions about temperature-cumulative emissions relationship if interactive carbon cycle is ignored. It is counter-intuitive: after cessation of fossil-fuel emissions in IAM scenarios, ocean naturally continues to take carbon, and ESM-inferring approach would count this ocean uptake as continuing anthropogenic emissions. Really confusing! This limitation must be discussed in depth for Fig. 3.
3) Third, about the paper conclusions. They are currently suboptimal and not that clear in terms of what new is found in these simulations. For example, a conclusion that land carbon uptake has many uncertainties – repeated several times - could be written without these simulations at all. This is very clear from ESM runs until 2100, and also can be seen in millennium-scale experiments, see eg Joos et al. (2013). Instead of these qualitatively vague conclusions on uncertainty, could you rather suggest how could we proceed in reducing uncertainty, what could be the main factors: response to droughts, CO2 fertilization, natural vegetation dynamics, land use? I also would suggest to focus on what is common among the models, and not only on differences. This would be helpful for carbon cycle emulators to be used for analysis of other future scenarios.

Minor comments

P2., l.44: the long lifetime of CO_2 in the atmosphere: I miss here citations of papers focused on this issue, eg Archer et al. (2009), Joos et al. (2013) – please refer to them in the introduction.

Section 2.1: ScenarioMIP protocol doesn't define how landuse forcing (including wood and crop harvest) and aerosol forcing are implemented after 2100. Land cover forcing in Hurtt et al. (2020) is provided only until 2100. How is it extended beyond 2100? Are these forcing extensions treated in the same way in all models? These details should be explained.

l.114-115. This statement belongs to the Results section, not scenarios (unless scenario forcings were averaged).

Section 2.2 Model descriptions here is extensive and inconsistent among models. It doesn't allow to capture at the first glance a difference between models. Arora et al. (2020) did a better job with their Table 2 summarizing all important details of participating models. Some details like on the permafrost carbon dynamics in CESM could be reported in addition to the new Table.
Also: some descriptions here refer to components which are most likely not used in the study (such as wetland emissions in CanESM2). Why do you need to mention them?

p. 15, l. 467-472 – very long sentence, cut it in two after the reference. Interesting is the overshoot behavior.

P.15 – last sentence is unclear, what exactly is meant by reference to paleoclimate? Does it mean that evaluation by paleo runs allows to avoid surprises or other way around?

P16., l. 508 what is meant by budgeting here?

l. 512 How would you propose to test carbon model dynamics on long timescale, against what evidence? Emitting ca. 5000 PgC in RCP8.5 goes beyond a scale of any available evidence for the last 30 million years, and the quality of data for deep paleo is really poor.

Fig. 2 a-b: dashed lines are hard to see. Why don't you use the same color code for models as on figure 3? It will make figure consistent. Labels on c-d are hard to see on a paper, this figure is not readable on a paper.

Fig. 2: I suggest to add a table with values of emissions, land and ocean net fluxes at the years 2100, 2200, and 2300 for every model and for ensemble-mean. This would help to see a difference between models and time slices.

Fig. 2 caption: using the term "biospheric fluxes" for net land and ocean fluxes is confusing. Ocean carbon uptake is mostly inorganic, unless the authors could disentangle changes in anthropogenic uptakes due to solubility and biology. I suggest consistently use land and ocean uptakes. Strictly speaking, landuse emissions are anthropogenic too, so one rather has to use "fossil-fuel" emission term.

Fig. 3. There is a principal inconsistency between C-driven and E-driven TCRE plots, as explained in the second major comment. This has implications for this plot, as C dynamics of MAGICC are different from the ones of ESMs. A negative betta feedback in E-driven run will likely stabilize the system much faster than in the C-driven simulation. Please discuss it.

Fig. 4-5. These Hovmöller diagrams are very useful, but they hide the fact that the zonal land distribution is inequal. This is especially valid for the latitudes to the south of 40°S. I suggest to cutoff these diagrams at 40°S or 50°S.
Also - ESMs have 2-D geographic distribution of carbon. You can add a figure of ensemble mean for land and ocean net fluxes at 2300 for two scenarios. It will be very useful to see such a plot corresponding to Fig. 2 c-d.

References

Archer, D., Eby, M., Brovkin, V., Ridgwell, A., Cao, L., Mikolajewicz, U., Caldeira, K., Matsumoto, K., Munhoven, G., Montenegro, A., Tokos, K., 2009. Atmospheric lifetime of fossil-fuel carbon dioxide, Annual Reviews of Earth and Planetary Sciences, 37, 117-134.

Joos, F., Roth, R., Fuglestvedt, J. S., Peters, G. P., Enting, I. G., von Bloh, W., Brovkin, V., Burke, E. J., Eby, M., Edwards, N. R., Friedrich, T., Frölicher, T. L., Halloran, P. R., Holden, P. B., Jones, C., Kleinen, T., Mackenzie, F. T., Matsumoto, K., Meinshausen, M., Plattner, G.-K., Reisinger, A., Segschneider, J., Shaffer, G., Steinacher, M., Strassmann, K., Tanaka, K., Timmermann, A., and Weaver, A. J.: Carbon dioxide and climate impulse response functions for the computation of greenhouse gas metrics: a multi-model analysis, Atmos. Chem. Phys., 13, 2793–2825, https://doi.org/10.5194/acp-13-2793-2013, 2013.