Reply to reviewer Gab Abramowitz

- *For this to work as a stand-alone paper, I feel like a little more contextualisation of the divisions between CMIP6 related projects might make sense. Why, for example, is there so little carbon cycle discussion (noting it's not circled in Figure 2) in an experiment that is ostensibly about all things land surface ("LS3MIP fills a major gap by considering systematic land biases and land feedbacks")? The carbon cycle is clearly relevant for a water resources discussion when CO2 is rapidly increasing. If there is a clear science rationale for the dividing line between another CMIP6 project (say, C4MIP) that investigates the land component of the carbon cycle it really should spelled out in detail here. C4MIP is only mentioned in passing and isn't shown on the diagram of LandMIPs (Figure 3). I would have thought it evident that the carbon cycle affects the water cycle, and that its effect is not limited to "impacts of snow and soil moisture processes . . . on terrestrial carbon exchanges" (L219-220). Alternatively, if there are historical institutional and/or political reasons for such a division I think that needs to be laid bare in a journal article describing the rationale for a science program.*

  This point was also made by Paul Dirmeyer. We've stressed the complementarity of LS3MIP with LUMIP and C4MIP, as indicated in the new text (see reply to Paul Dirmeyer above).

- *As a description of what LS3MIP participants will produce and why, this document is clear in its motivation and detail, and is well thought out. What's less clear, to me at least, is how we can meaningfully evaluate the model output that this experiment will produce. I understand that analysis of CMIP data is not coordinated in the way that the production of simulations is, but nevertheless the production protocol significantly affects what can or cannot be investigated.*

  Indeed, the manuscript primarily focuses on the experimental protocol, and gives examples of analyses and important research questions that can be addressed with these experiments. As such it does not describe so much the dynamics of the research network that is active in the planning, execution and analysis of LS3MIP. I've added a paragraph on this in the "time line/participation" section: "The organisational structure of LS3MIP consistently relies on active participation of modelling groups. Coordination structures are put in place for the collection and dissemination of data and model results (Eyring et al. 2015), and for the organisation of meetings and seminars (by the core team members of LS3MIP, first five authors of this manuscript). Different from earlier experiments such as GSWP2 and GLACE1/2, no central "analysis group" is put in place that is responsible for the analyses as proposed in this manuscript. The execution and publication of analyses is considered to be a community effort of participating researchers , under coordination of the core LS3MIP team members, for instance in order to avoid duplication of efforts and coordinate the production of scientific papers.".

- *One of the stated objectives of this work is to "diagnose systematic biases and processlevel deficiencies in the land modules of current Earth System Models". This requires an ability to 'ground-truth' a sufficient subset of model states and fluxes, at high temporal scales, to be able to categorically identify and quantify the fidelity of process representation. At this point in time, as I understand it, we don't come close to having this kind of observational data collection at gridded scales (despite the many products described on p17/18). While this experiment (laudably) uses multiple gridded driving data sets in Land-Hist2, this very real uncertainty, together with the significant disagreement amongst the multiple historical gridded evapotranspiration products that are available (as an example), means that we are usually unable to categorically describe the cause of differences between a model simulation and evaluation products. This problem is even tougher in the coupled environment. Essentially I don't think we can use this approach for model diagnosis, unless model problems are extreme. It is essentially a confirmation holism problem, well described in the broader climate modelling context by Lenhard and Winsberg (2010). It is clearly also problematic when we try to "quantify the associated uncertainties" with the land surface in climate projections – another stated objective. How do the authors propose we get around this issue?*

  This is an interesting and well posed issue: the complexity of the true climate system will not allow a comprehensive analysis of all its relevant interactions and dynamics given the limited ability of models and observations to capture these. Personally I am not a believer of "reducing uncertainty" as a key role of climate (model) research, but am convinced that within the limits of "understandability" valuable statements on plausibility of processes or events to occur can be derived from well designed model experiments. It goes too far to devote an extensive discussion on this issue in this manuscript, but we included a reference to Lenhard and Winsberg in the discussion section: "`Within the limits to which complex models such as ESMs can be evaluated with currently available observational evidence (see e.g. the interesting philosophical discussion on climate model evaluation by Lenhard and Winsberg; 2010), it will lead to enhanced understanding of the contribution of land surface treatment to overall climate model performance…`"

- *A partial antidote to the problem outlined in (2). Despite the glaring scale mismatch for model application, using a broad collection of site-based data sets to thoroughly understand the fidelity of process representation might well help regional and global scale applications. Lines 354-356 indicate that some forcing from single sites will be included in LS3MIP, but there is very little detail, presumably because the authors felt this spatial scale was not especially relevant for global scale simulations. My feeling is that if we really want to diagnose process level deficiencies and provide the means to quantify uncertainty, this really needs to be the starting point, since it's the only context in which we can meaningfully understand the uncertainties in both the forcing and evaluation data. This is not to suggest that model biases / errors at this* *scale necessarily translate directly to larger scales, of course. The results in Best et al (2015) and Haughton et al (2016) illustrate the power of the constraint that observational data provide at these scales. Do the authors have any reason to believe, if we had "true" gridded forcing and evaluation data at global scales, that the benchmarking results from these papers would not still be evident at gridded scales? If there is any doubt, I think a comprehensive set of site-based experiments would be very useful as part of LS3MIP, at least as its objectives currently stand. Again, I'm not sure of the extent to which the experimental protocol is already fixed, but if not, this may be a useful addition.*

Allthough we do agree with this notion, the exact point the reviewer wants to make is not clear. The experimental design is not particularly geared towards either local or global evaluation, but indeed analyses of larger scale interactions have a stronger emphasis than process evaluation at the local scale. However, also analysis using in situ observations must be put in the broader context in order to gain insight and inspiration for model development, and the "holistic view" described by Lenhard and Winsberg similarly applies to in situ data. We felt there is not a very clear message from this statement that we could include in the revised version of the manuscript.

- *L319 / Figure 5: are these the PLUMBER sites from Best et al 2015? If so, a simple reference gives readers enough information to get a lot more from this figure.*

  Indeed, it was the PALS data set that was used here. We've indicated that in the figure caption.

- *L393-394: How is the choice to "represent the ensemble spread efficiently and reliably" going to be made? Evans et al (2013)? Global temperature trend? Could be controversial!*

  We are aware of the controversy but have not yet made a decision on how this choice will be made. The reference to Evans et al is added for inspiration.

- *L501-509: this seems a little vague - are periods for extremes analysis part of LS3MIP or not? If so, which periods, why?*

  At this point in time it is very difficult to be more specific: early results should give inspiration to zooming in on particular episodes.

- Other minor text suggestions and citations have been included as suggested.