The Land Surface, Snow and Soil moisture Model Intercomparison Program (LS3MIP):

aims, set-up and expected outcome (original title)

Reply to reviewers

Bart van den Hurk et al, 13 July 2016

We thank the reviewers, editor and representatives from the various organisational units of the CMIP panel for their constructive comments on the LS3MIP documentation paper. Although the LS3MIP protocol as described in this paper can be regarded as a reference document guiding the implementation of the experiment, many questions on technical and scientific aspects still arise as the experiment is being set up and critically evaluated. We therefore welcome the editors suggestion to provide a version number to the experimental description, as future redesigns or reconsiderations may be likely.

Reply to the editor

We have changed the title of the manuscript into "LS3MIP (v1.0) contribution to CMIP6: The Land Surface, Snow and Soil moisture Model Intercomparison Program – aims, set-up and expected outcome"

Reply to reviewer Paul Dirmeyer

- *Objectives section: An obvious "omission" is anything to do directly with vegetation or the carbon cycle, which will probably stir up questions in the minds of readers. The authors should declare the territory of this MIP up front, presaging Fig 3, that the focus is on the "GEWEXy" bits, especially the water cycle. State explicitly that there are other MIPs (e.g., LUMIP) that are concerned with the vegetation aspect of the land surface (can say "...as described later..." as this does get addressed eventually with Figs 2 and 3 on p.6).*

  Good point, also commented on by other reviewers. In the "objectives" section we added a paragraph explaining the link to LUMIP: "While vegetation, carbon cycle, soil moisture, snow, surface energy balance and land-atmosphere interation are all intimately coupled in the real world, LS3MIP focuses – necessarily – on the physical subdomain in this complex system. Interactions with vegetation and carbon cycle are included in the analyses wherever this is possible without loosing this essential focus. In the complementary experiments Land Use MIP (LUMIP; see Lawrence et al. submitted) and C4MIP (Jones et al, 2016) vegetation, the terrestrial carbon cycle and land management are the central topics of analysis. LS3MIP and LUMIP share some model experiments

and analyses (see below) to allow addressing the complex
interactions at the land surface and yet remain able to
focus on well-posed hypotheses and research approaches."

- *Highly relevant work on snow-climate coupling should be cited (Xu and Dirmeyer 2011) especially regarding the albedo versus delayed hydrologic effects (Xu and Dirmeyer 2013), the latter also at line 191.*

  Sorry we missed these references. We've added a sentence in the introduction section "Temporal dynamics of the snow-atmosphere coupling during various phases of snow depletion (Xu and Dirmeyer 2011, 2012) are crucial for a proper representation of the timing and atmospheric response to snow melt." and added a reference near line 191

- *Para lines 176-192: This is rather redundant with the first paragraph of the introduction. Probably needs to be in only one place. Same para: Is WCRP the only "customer" for this project? Seems that the potential audience is much broader - should be stated here. Along with the next paragraph, gives short shrift to the broader impacts of L3MIP.*

  Apart from removing some redundant references to WCRP, we've added a sentence describing the potential revenues of LS3MIP: "LS3MIP is geared to extend and consolidate available data, models and theories to support human awareness and resilience to highly variable environmental conditions in a large ensemble of sectoral domains, including disaster risk reduction, food security, public safety, nature conservation and societal wellbeing."

- *L297: Which NCEP reanalysis? There are 3 separate unique NCEP reanalyses.*

  This can be found in the documentation of the CRU-NCEP data set and is of minor relevance for this paper

- *Para lines 427-439: Really need to avoid prescribing surface soil layer moisture as well, because it can cause highly unrealistic Bowen ratios where net radiation is high, (cf. GLACE2 experiment "S"; Koster et al. 2006)*

  Yes, we are aware of this, and have added a reference to this notion. As we made clear in the manuscript, a standardization of this approach is difficult, and should be carefully tested by the modelling groups.

- *L510-12: What is the protocol for ensemble construction? Are there suggestions of a prioritized list of preferred approaches like there were for GLACE-1 and -2?*

  Good point. We've added the phrase "The procedure to initialize the land surface states in the ensemble members is left to the participant, but should allow to generate sufficient

spread that can be considered representative for the
climate system under study. Koster et al. (2006) proposed
a preference hierarchy of methods depending on the
availability of initialization fields, and LS3MIP will
follow this proposal."

- *Tier 2 experiments in LFMIP: The mean AOGCM climatology of SST will certainly differ from that of an AMIP run based on observed SST, introducing two differences between the experiments, not one. What are the implications?*

  The implications will be substantial, but also a systematic biases in SSTs is an inherent part of the analysis of the role of SSTs on land-atmosphere coupling.

- *L699-702: It seems agricultural areas in general should be a focus.*

  Indeed, added as such

- Other minor corrections and suggestions for citations: changed as suggested

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.

Reply to review by Ron Stouffer

- *In the Introduction, this paper needs to clearly state what is its focus and what is found in the other strongly related GMD CMIP6 papers. The split between the physical climate and the carbon MIPs needs to be made much clearer and early in the paper.*

  This comment is also made by Paul Dirmeyer and Gab Abramowitz. We've added a paragraph on the LS3MIP focus and links to LUMIP and C4MIP in particular (see comments above)

- *Page 11, Lines 389-402 – You may want to note here that these runs will be performed sometime in the future after the ESM data is available in the CMIP6 archive. This could be a year or 2 or more in the future.*

  Pointed out in a comment

- *Page 12, line 405 – Is there an interaction between LFMIP and FAFMIP? It seems there should be and it should be noted in this section.*

  We did make a reference to FAFMIP in this section, but plans for coordinated analyses have not yet been discussed with the FAFMIP panel: "This set-up is similar to the Flux Anomaly Forced MIP (FAFMIP, Gregory et al. 2016), where the role of ocean-atmosphere interaction at climate time scales is diagnosed by idealised surface perturbation experiments.". However, a stronger link with OMIP is included in the new version, to cross-reference offline generated freshwater fluxes: "Interactions with the Ocean MIP (OMIP; Griffies et al. 2016) are arranged by the use of terrestrial freshwater fluxes produced in the LMIP simulations as a boundary condition for the forced ocean-

```
only simulations in OMIP, in addition to the forcing
provided by (Dai and Trenberth 2002)."
```

- *Page 14, line 523 – "A perfect boundary condition" – several studies have shown that prescribed SSTs are less than perfect since it breaks the atmosphere-ocean coupling and feedbacks. This issue distorts the variability in models forced by SSTs relative predicted SSTs. I assume the land surface will have even larger issues since it has much smaller heat capacity. Reword.*

  Also Paul Dirmeyer made a comment on the notion that prescribed SSTs are not necessarily perfect. We've rephrased it as "pseudo-observed boundary condition" experiment.

- *In the table "direction" should be changed to "Positive direction". Just to be very clear.*

  See reply above

Reply to Anonymous Referee #3

We thank the reviewer for the overall positive assessment of the experimental set-up and its description, and his/her recommendation to have the paper published in GMD.

- *Data volume estimates for the requested ESM model output are currently missing and it is recommended to add the information for instance in table 1. This is easy to compute if the cost of 1-year of output (mandatory/extra) is made available. The information can be very helpful to plan storage of the output and runs throughput.*

  Although this is a valuable comment, we cross-checked a few other CMIP6 papers in GMD, and none of them provide these estimates. I understand the CMIP coordination panel is preparing a paper describing the planned data exchange and storage, and I would expect that document to act as a reference for resource planning by the modelling groups.

- *Links to other projects such as the PRIMAVERA-H2020 https://www.primaverah2020.eu or CRESCENDO-H2020 is also worth mentioning.*

  These projects are well known to us. We did make cross-references to a number of earlier experiments, but chose to confine ourselves to those experiments that have a direct relation with the LS3MIP protocol and analyses. We are aware that projects like CRESCENDO (and also others) will be used to carry out the simulations and analyses mentioned in LS3MIP (and other MIPs)

- *There is no mention to the reproducibility of the results and whether the data repository will facilitate for instance re-run the Land experiment series with another model at a later stage.*

  We don't have a lot of experience with reproducibility, but generally outcomes are pretty sensitive to computer platforms, initialization, subtle configuration settings

etc that make direct reproducibility limited. Later participation to the experiment by other modelling groups is encouraged and facilitated by the infrastructure. A comment on this is added in the "Data Availability" section: "This infrastructure makes it possible to carry out the experiments in a distributed matter, and to allow later participation of additional modelling groups."