Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-72-RC2, 2016

Geoscientific Model Development Discussions

Interactive comment on "The Land Surface, Snow and Soil moisture Model Intercomparison Program (LS3MIP): aims, set-up and expected outcome" *by* Bart van den Hurk et al.

G. Abramowitz (Referee)

gabsun@gmail.com

Received and published: 2 May 2016

Overall this is a well written document that describes an extensive set of experiments aimed at understanding the interaction between land and atmosphere as part of CMIP6. It wasn't particularly clear to me whether I should simply be commenting on the clarity of the description of the experimental protocol (i.e. that the program was already decided) or additionally giving feedback on the experimental protocol itself. There is a mixture of both below. To ensure my review was particularly annoying, I have cited several papers I have been involved with - apologies.

Broader comments / issues:

1. 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.

2. 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.

One of the stated objectives of this work is to "diagnose systematic biases and process-level 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?

3. 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.

Particulars:

L101-2: this use of parentheses seems a little unorthodox - perhaps write "sub-seasonal and seasonal"?

L108: remove right parenthesis after Zampieri et al. 2012

L169: "experimend"

L178: add an 'is' to "and [is] directly related"

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.

L322/323: maybe worth noting that SWdown / LWdown are not shown? L373: might be clearer to say "GSWP3 forcing" here.

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!

L470-472: an excellent idea!

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

L629: should be "several time scales"

L644: perhaps use effects 'on' calculated land temperature, rather than "to calculated land temperature"

L736-737: "optimise ... parameterizations or its forcing" should maybe be "or their forcing"?

References:

Best, M.J., G. Abramowitz, H.R. Johnson, A.J. Pitman, G. Balsamo, A. Boone,

M.Cuntz, B. Decharme, P.A. Dirmeyer, J. Dong, M. Ek, Z. Guo, V. Haverd, B.J.J van den Hurk, G.S. Nearing, B. Pak, C. Peters-Lidard, J.A. Santanello Jr, L. Stevens, N. Vuichard, 2015: The plumbing of land surface models: benchmarking model performance, Journal of Hydrometeorology, 16, 1425-1442.

Evans, J., F. Ji, G. Abramowitz and M. Ekstrom, 2013: Optimally choosing small ensemble members to produce robust climate simulations, Environmental Research Letters, 8, 044050, doi:10.1088/1748-9326/8/4/044050.

Haughton, N., G. Abramowitz, A.J. Pitman, D. Or, M. J. Best, H.R. Johnson, G. Balsamo, A. Boone, M.Cuntz, B. Decharme, P.A. Dirmeyer, J. Dong, M. Ek, Z. Guo, V. Haverd, B.J.J van den Hurk, G.S. Nearing, B. Pak, J.A. Santanello Jr, L. Stevens, N. Vuichard, The plumbing of land surface models: is poor performance a result of methodology or data quality? Journal of Hydrometeorology, in press.

Lenhard, J., E. Winsberg, 2010: Holism, entrenchment, and the future of climate model pluralism. Studies in History and Philosophy of Modern Physics, 41, 253-262.

Interactive comment on Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-72, 2016.