Review of "Diagnosing hydrological limitations of a Land Surface Model" by Le Vine et al, for HESSD

This study examines the extent to which a land surface model (JULES) offers a useful starting point for so-called "hyper-resolution" modelling. The model is applied at 1km resolution over a chalk basin in southern England. A carefully considered "diagnostic approach" is used to guide successive improvements to the model, using a variety of data sources and additional model components.

Overall this is a good manuscript and deserves to be published after minor revisions. I think the methodology and results presented are good, and my comments in later sections below are largely suggestions as to how to improve clarity at a few points in the manuscript. However I will start by considering some more general points that I would like to see addressed.

Would an intermediate configuration suffice for some applications? Table 3 shows that most of the chosen metrics were relatively stable over many configurations, and in particular those to do with runoff or flow (i.e. not soil moisture) are stable starting at JULES+WG+PDM (if anything they deteriorate slightly after that). I accept the line followed by the authors and the further insight that is gained in the CHALK, GW and GWadj stages, but I'm left wondering whether for some applications WG+PDM would be sufficient. Fig.6 shows good agreement between modelled and observed hydrographs, but by then I was wondering what the modelled hydrographs would look like for some of the earlier configurations. Configuring and running a distributed groundwater model is a non-trivial exercise and, in the context of global hyper-resolution modelling, likely requires data (e.g. characterising subsurface properties) that are not available for all catchments. At the outset my assumption might have been that a chalk catchment is one place where a good representation of groundwater processes would be important, yet the results suggest that, at least in this catchment and for some metrics, the groundwater model doesn't add that much – which is probably good news if the aspiration is global or even regional modelling. On the other hand, perhaps the large depth to the chalk water table simplifies things, in that only a very weak coupling was required between the land surface and groundwater models, and other environments with shallower water tables would be trickier. Given that the study is essentially (and in part) looking at whether land surface models are a reasonable starting point for hyper-resolution modelling, it would be good to see some sort of conclusion on this point, even if it is necessarily couched in caveats about this being a single study, etc. – perhaps the conclusions could cover this. At any rate I would like to see some mention and discussion of these issues in the manuscript, even if the ultimate answer comes down to the "uniqueness of place" idea that the authors mention in their conclusions.

In a similar vein, the catchment studied (the Kennet) is clearly very data rich and many sources of data were used by the authors. Clearly most locations, even within Europe never mind globally, will have fewer data. From memory, the data requirements of hyper-resolution modelling were considered in the Wood et al paper and subsequent discussions. I would like to see some mention of data requirements and limitations where there are fewer data – even if it's just one sentence.

The "diagnostic approach" seems to work quite well and in particular Table 3 shows that later configurations generally do not lose the advantages of earlier configurations, which the authors suggest is due to the physical basis of the model and the modifications. Or is there an element of good luck and/or selective reporting of experiments? Many modellers are familiar with the experience of finding that later changes improve some aspects of model performance but worsen others, and I'm not convinced that "physically-based reasoning" (my phrase) is sufficient to guarantee that a diagnostic approach will result in such clean, monotonic trend towards improvements (which is not something the authors claim either). Any comments?

I'm not convinced that "Standard JULES" is an appropriate configuration (assuming I've understood it correctly) – see comment below.

Specific points

Abstract, line 5 "A diagnostic approach to model evaluation" – could be "evaluation and improvement", given that it seems to me that part of the process is thinking about how process representation can be used to make improvements.

P7545 L5 Assumption of 1-D vertical flow – I wonder if this is partly a reflection of the historical use of LSMs with large gridbox sizes (e.g. 100km), at which scale vertical flow is likely dominant.

P7545 L23 Note that the LSMs in the Boone et al. study (Rhone-Agg, or something similar) were NOT coupled to an atmospheric model; they were driven offline by prescribed meteorology.

P7546 L7 "to maintain an overall water balance" – I didn't really follow this phrase initially, although I now realise that "overall water balance" is the phrase used in the cited Yilmaz et al. paper. My first understanding was that you were meaning the model just had to conserve water, which is a zeroth order requirement! I think a phrase such as "provide reasonable estimates of individual components of the water balance" (although clearly not perfect!) conveys the idea better, at least for me.

P7547 Somewhere about here it would be appropriate to mention that you use v2.2 of JULES. Otherwise that information only comes in Table 3 I think. Also, you say JULES "typically" uses a 3m depth of soil, but don't clarify that that's what you used, nor the number (and thicknesses) of layers. Even more importantly, I think this might be a good place to clarify how runoff can be generated by JULES, in particular that infiltration excess runoff is a possibility (which becomes apparent in later discussion about rainfall rates). Also – how do you specify the initial state and consider "spin up"?

P7548 "base flow index" – is this a sufficiently generic term that it requires no further information or citation?

P7549 L18 Unless I've missed something, the AWS data are not used here and don't need to be mentioned.

P7550 Section 3 Think about adding a bit more text in the introduction here, to clarify that you're about to list and go through the details of several experiments. Also mention the configuration names that I think otherwise first appear in Table 3, e.g. clarify that 3.1 describes JULES+WG.

P7550 Sec3.1 Somewhere (maybe here, or in model description) I think you need to clarify that the default setup is to use the daily meteorological data with no further temporal disaggregation. At present this was not very clear to me and only came out when I was trying to understand later sections.

In fact I have reservations about whether your "Standard JULES" configuration is useful. Land surface models such as JULES are designed to be driven by sub-daily meteorological data. The most obvious effect that I can think of is in the parameterisation of stomatal conductance/photosynthesis, which responds non-linearly to shortwave radiation and saturates at high levels of radiation. Forcing a model with daily average radiation might be akin to perpetual twilight, and at any rate the response

of the system to this average forcing is possibly rather different to the average response to time-varying forcing. To me JULES+WG should be the baseline parameterisation. “Standard JULES” is essentially a poorly designed configuration. So at best the improvement to JULES+WG quantifies the effect of not setting the model up correctly to start with and is therefore of limited value.

P7550 L21 Rephrase as at present you effectively have “sub-daily precipitation depends on mean daily temperature”. I think I follow your meaning, but it’s a bit confusing. I think the type of precipitation depends on the daily T, but that is then fixed within any one day. I also note a high threshold temperature for convective rainfall: 27 degC. I assume this means there is almost no convective precip in this catchment! I think you also need to explain the significance of these different types of precip, in particular the hydrological significance (in the model) of the distinction between convective and large scale precip. For large-scale applications this is typically that convective precip is assumed to cover only part of a gridbox, whereas “large-scale” precip covers all or more of the gridbox. What is assumed in your model setup?

P7551 Sec3.2 It would be interesting to know the range of value for the b parameter that you found (possibly even via a map?). Were they very variable between locations? If not, how much performance is lost if a single value is used across the catchment (as might be required in many catchments for which fewer data are available)? On L20 the word “range” can be removed. Also perhaps you can make it clearer that a series of spatially-distributed runs of JULES were used to evaluate drainage:total runoff, then the best parameter field used for JULES+WG+PDM (explanation or a map of the range of b values might help to clarify this in readers’ minds).

P7552 Sec3.3 Can you make clearer how you implemented the dual curve representation in JULES? I guess you essentially have two curves, with a breakpoint separating when each is used. This is clear in Fig.3, less so in the text at this point. Perhaps we could have Fig.3 at this point, not later?

P7554 Sec3.5 Initially I was unclear how ZOOMQ3D was to be used. In particular, note that the 2-D hillslope model and ZOOM are used rather differently – the former (essentially) to check assumptions used in JULES, the latter is coupled to JULES. It would be good to make this as clear as possible.

P7556 L14 See earlier comment re clarification of runoff generation mechanisms in JULES. I think this is the first time I found myself thinking “infiltration excess runoff” – which I assume is what you’re referring to.

P7557 Sec4.2 See earlier comment re need to know layer thicknesses. Fig.2 suggests far fewer model layers than obs layers (many of the model curves look identical at different depths).

P7558 Sec4.3 I’m not sure I follow why the lower boundary condition is considered under “temporal redistribution”. Similarly the insights from the 2-D hillslope model (e.g. negligible lateral fluxes in the unsaturated zone) are not obviously “temporal redistribution” (although they are related). Can you provide insight into your thinking and classification?

I’m not convinced we need Figure 4! The first panel is trivial. The second shows that the gradients at 6m and 5.5m are often similar, which doesn’t strike me as particularly surprising.

It’s unclear what was finally used in JULES, mainly because I’m not 100% sure what soil depth was used (apologies if I missed that). You used a persistent gradient at the bottom of a 3m column? And

the better relationship in the 2-D model at 6m suggests that a deeper column should be used with JULES? Please clarify.

P7559 final paragraph has confusing references to various timesteps and models. I think the results all refer to the 2-D model, but this should be made clearer. Similarly, daily values (L28) are used just because this is the ZOOM timestep length, not because ZOOM itself was used here.

Tables and Figures

Table 3 – Two columns appear as "RBias_Q". I assume one should be "RBias_SR". Footnote 2 - likely remove final "s" from "For a model configurations".

Fig.2 We don't need so many panels. Try to select a few representative or illustrative depths, then fewer panels will allow us to see more detail in each.

Fig.4 Clarify that second panel shows gradient at 5.5m.

Minor points

There are a few American spellings, e.g. "modeling", "meters", whereas I assume HESS uses UK English (e.g. modelling).

P7544 L13 "Whereas, CLM…" – Rephrase, e.g. "In contrast", "However".

P7545 L3 "i.e." – I think this should be "e.g.", as it's only one of several possible examples.

P7546 L15 "only...physically meaningful way" – I'm not sure "physically meaningful" means much here! How about just "currently available way"?

P7547 L13 "global circulation" – I'd remove that, as I know the Unified Model can be used in limited-area configurations too (i.e. not global).

P7547 L27 Missing full stop.

P7556 L10 "When this is done" – Rephrase. At present it sounds a bit like "constant temporal disaggregation is dine using the weather generator", when in fact this is exactly what the generator avoids. L15 Remove comma.

P7564 L23 I haven't studied the references, but I did notice a missing o in "Viterbo" in the first line.