Response to the comments by Reviewer 2

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.

We thank the reviewer for helpful and detailed comments on how to improve clarity of the manuscript. The comments are addressed below in blue. There is no reference to the exact pages/line numbers as those are not known before the typesetting.

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 nontrivial 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.

To reflect the reviewer’s point, the following has been added to the conclusions:

“For some applications, the intermediate model configurations might be sufficient. For example, while JULES+WG+PDM configuration cannot provide flow/groundwater level hydrographs as it lacks surface and subsurface water routing, the configuration can still be used to represent the water balance over an area. This is useful for regions where no groundwater model and/or detailed geology data are available. It is to be noted that the findings are catchment specific and result from a weak surface-groundwater coupling, and as such cannot be readily generalised to other environments with shallower water tables."

The modeled hydrographs for the configurations that do not have the water routing component are (as one might expect) pulses of effective rainfall (i.e. excess rainfall), and do not meaningfully represent a temporal water redistribution at the catchment outlet. Hence, JULES+WG+PDM configuration is only representative of the water balance in the catchment.

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 conclusions section has already contained the data requirements (see below), one sentence on the 'fewer data' limitations has been added to that (in italic).

"Diverse sources of information were used to guide the model assessment and include remotely sensed data (topography, land use), spatially extrapolated point data (soils, weather conditions), point measurements (soil moisture and matric potential, flow, groundwater level), regionalised hydrological information (BFIHOST), and states/fluxes extracted from an auxiliary physics based hillslope model [Ireson and Butler, 2013]. *Fewer data might result in a less detailed representation of the water cycle depending on the specifics the hydrological system being investigated."*

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.

Some of the model weaknesses (and hence targets for improvement) had been identified at the research proposal stage, i.e. the chalk parameterisation, lower boundary condition and lack of groundwater/surface water routing (see the corresponding proposal abstract on http://gotw.nerc.ac.uk/list_full.asp?pcode=NE%2FI006621%2F1). And some, i.e. temporal disaggregation of weather variables and lack of surface runoff, were realised at the initial stage of model usage (in this case they're weaknesses in how the model handles the relatively low resolution - temporal or spatial – of the available data). The model change sequence has been arranged in a logical way, following the strategy of Yilmaz et al, 2008, to reflect the different levels of detail represented by the model. As indicated in the conclusions (and Section 4), other changes might have been implemented in the model as the cause-effect mechanism is sometimes not that clear.

For example, a localized higher intensity rainfall is a potential alternative to the spatially non-uniform soil properties (that are represented by the PDM add-on) that allows producing an additional surface runoff from each model grid.

The comments on the "standard JULES" are addressed below as they arise.

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.

Modified to "evaluation and improvement".

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.

The use of LSMs to provide the lower boundary condition for climate models might have indeed affected the assumption.

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.

The reference to the atmospheric model – LSM coupling is removed.

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.

An explanation has been added (in italic):

"…to maintain an overall water balance *(i.e. water partitioning between different water cycle components)…*"

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"?

The following has been added at the end of the section describing JULES:

"JULES is able to generate an infiltration-excess (when PDM is used, or when rainfall intensity exceeds the near-surface infiltration rate) as well as saturation-excess (when TOPMODEL is used, or through the upward movement of water from saturated soil layers) surface runoff.

The study uses and implements modifications to JULES version 2.2, termed the 'standard JULES'. The standard setup is used with a 3 m depth of soil, four soil layers: 0.1, 0.25, 0.65, and 2 m deep, starting from the surface. The model is spun-up over three years, repeating the weather inputs for the first year of available data three times (one of the model ' warming-up' options provided), and initialising soils with saturated conditions."

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

A short explanation has been added for clarity: "*…i.e. the proportion of total flow as base flow…*".

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

As CHESS data is daily, and a finer temporal resolution meteorological data from a local AWS is used to calibrate Ksat for chalk. This is to account for any soil moisture sensitivities to the actual rainfall timing and intensities (those are only approximated by the weather generator). And the following sentence has been added to the last paragraph of Section 2:

"The sub-daily weather data are used to account for any soil moisture sensitivity to the rainfall timing and intensity."

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.

The introduction to Section 3 already refers to Table 2 that contains all the details about the modifications, sources of information and the configuration names. To clarify what modifications are considered below the introduction has been amended (in italic):

"The hydrological process representation in JULES is assessed with respect to the four primary functions of a hydrological system [Yilmaz et al, 2008] 1) overall water balance, 2) vertical redistribution, 3) temporal redistribution, and 4) horizontal spatial redistribution. Table 2 lists the assessment metrics for each of the four functions, the examined model assumptions/simplifications, the implemented model modifications, and the information sources used to inform the model modifications. Each of these information sources is described in the following sub-sections. *The implemented model modifications considered below consist of a sub-daily weather generator, representation of sub-grid scale heterogeneity, dual Brooks and Corey curve representation of chalk hydraulic properties, change of the lower boundary condition and coupling to a groundwater model.*"

Further, references to the resulting configuration names have been added at the end of each sub-section in Section 3 that introduces a new model component/adjustment, e.g.

"The model configuration that includes the weather generator is referred to as JULES+WG (see Table 2)."

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 timevarying 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.

To clarify how JULES version 2.2 extrapolates the meteorological drivers, the following has been added to the model description given in Section 2 (in italic):

"The model runs at a sub-daily time step, using meteorological drivers of rainfall, incoming radiation, temperature, humidity and wind speed. *When meteorological data have coarser temporal resolution than required by the model, the standard model (version 2.2) disaggregates the data as constant values.*"

The 'Standard JULES' might indeed seem to be a poorly designed configuration, nevertheless, there is nothing in version 2.2. to warn about/stop a user from supplying daily meteorological data as a model input; and the meteorological data are disaggregated as constant values, regardless of the data temporal resolution. Further, sub-daily forcing data are not always available (e.g. for model inter-comparison), and a weather generator will be required. The lead author has observed a number of cases when young researchers were oblivious of the input data temporal resolution effects that lead to incorrect representation of the processes. Hence, it is perceived important to include the 'Standard JULES' as a default configuration.

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?

There has been a typo for the value of convective rainfall temperature threshold – it should be 20C, not 27C.The significance of the different types of precipitation is the event duration – 2 hrs for convective and 5 hrs for large scale precipitation; the precipitation is assumed to be uniform over a corresponding model grid. This is how the manuscript describes the rainfall disaggregation (added/modified text is in italic):

"This differentiation is based on the mean daily temperature. Precipitation is defined as: snow if the temperature is below 0 C; convective if the temperature is above 20 C; and large scale rainfall, otherwise. It is set to start at a random time during a day and to continue for a specified number of hours over *the entire corresponding model grid*: two hours for a convective storm, and five hours for large scale precipitation."

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).

The following explanation has been added to Section 4.1:

"The parameter b of the PDM model is selected based on regionalised information from BFIHOST (Section 3.2) and ranges, mainly, between 0 and 0.4, except for two grids where b is set as 0.7 and 1. Further, the parameter b is assigned 0 value over approximately 60% of the catchment for the locations with permeable (chalky) top-soils."

If the parameter b is assigned a single most frequent value of 0, then it is equivalent to switching PDM model off. And as Table 3 shows, it results in a negligible surface runoff amount (configuration JULES+WG).

The word 'range' is needed to justify the selected interval for b values. A word 'considered' is added to clarify the meaning:

"Due to the high computational requirement of JULES, only 21 regularly spaced values between 0 and 2 are considered. The *considered* parameter b range is found to provide suitable drainage-to-total-runoff ratios for the catchment soils and meteorological conditions."

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?

Figure is a part of the results and as such belongs to the Results section, not to the Method section that describes how to obtain such a dual-curve representation. The following has been added to Section 3.3 to clarify the dual curve representation (in italic):

"Due to the two distinct flow domains in chalk - matrix and fractures, two *intersecting* Brooks and Corey curves are employed when fitting a chalk soil moisture retention curve. *The effective soil moisture at the curves' intersection is estimated using available observations*."

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.

Section 3.4 (on the detailed physics-based model of a chalk hillslope) explicitly states that the 2D model will be used to check the assumptions:

"Fluxes and states of the chalk hillslope model for the period 1970–2000 are examined to assess the following two assumptions underlying the JULES hydrology: (a) there is no hydrological interaction between neighboring vertical soil columns, and (b) a unit gradient flow is a satisfactory approximation of the lower boundary condition at the 3 m base of the soil column on a hill-slope location with a typically deep unsaturated zone. Further, the hillslope model is used to evaluate the nature of coupling between the unsaturated zone and groundwater, as well as the nature of water transport in the deep unsaturated zone located between the base of the JULES soil column and the water table."

To make the distinction even clearer the section describing ZOOMQ3D has been amended with the following:

"The model configuration JULES+WG+PDM+CHALK is coupled to ZOOMQ3D based on the findings from the detailed 2D model (Section 3.4) and is referred to as JULES+WG+PDM+CHALK+GW (Table 2). When groundwater model parameters are adjusted to examine the sensitivity of model response, the configuration is referred to as JULES+WG+PDM+CHALK+GWadj."

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.

A description has been added to Section 2.1 (JULES outline) (see the response to the earlier comment).

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).

The setup is now clarified at the beginning of Section 4.2 as:

"Both JULES+WG+PDM and JULES+WG+PDM+CHALK configurations use 4.5 m long soil columns with 0.1 m thick soil layers to facilitate the comparison with the observed soil moisture."

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?

> The lower boundary condition is a key to the LSM-groundwater model coupling. A groundwater model provides the subsurface water routing, and hence the temporal water redistribution at the outlet.

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 might be trivial, however the standard JULES still uses a unit gradient lower boundary condition (the left-hand side panel of the figure), while the figure shows that it is inaccurate. Moreover, the figure illustrates the level of approximation offered by the proposed persistent boundary condition.

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.

> To clarify the following has been added to Section 4.3 (in italic):
>
> “Based on the above findings, JULES+WG+PDM+CHALK *is used with a 6 m deep soil column and 0.1 m thick soil layers, and* is coupled via a weak two-way coupling to the groundwater model ZOOMQ3D implemented through the lower boundary condition (persistent gradient).”

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.

> The following clarification has been added (in italic):
>
> “Lastly, to draw a connection between the modelled potential recharge at 6 m depth and the modelled actual recharge at the water table, temporally averaged vertical fluxes *extracted from the 2D hillslope model* are considered for 6-hourly (the model step), daily, weekly and 30-day periods.”

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”.

> Corrected.

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.

> The number of panels has been reduced.

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

Clarified in the figure caption.

Minor points

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

Corrected.

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

Changed to 'In contrast'.

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

Corrected.

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

Modified as suggested.

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).

Modified as suggested.

P7547 L27 Missing full stop.

Corrected.

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.

Changed to "*When weather variables are temporally disaggregated* using the weather generator (WG)..."

L15 Remove comma.

Corrected.

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

Corrected.