Author's response for 'A clustering approach to reduce computational expense in land surface models: a case study using JULES vn5.9'

Elizabeth Cooper

We thank both reviewers for their comments and suggestions, which will help to improve the manuscript. We list the reviewers' comments below in black, with our responses in blue. We have copied the detailed comments from reviewer one into this document with line numbers corresponding to the original document, and we have numbered our responses to those comments for ease of cross reference. We give proposed additions to the manuscript in italics.

Reviewer #1

This paper presents a case study where the JULES land model is run with a vector-based spatial configuration, instead of the more typical grid-based setup. The main benefit of such an approach is that it can lead to computational efficiency, if grid cells with similar hydroclimatic behaviour can be grouped into a single computational unit. As the authors already acknowledge, this approach is not new (in addition to the already existing mention of Swenson et al., 2019, see also for example Gharari et al., 2020).

Thank you for pointing us to the work in Gharari et al., 2020, we will add discussion of this study to the introduction (see response 4 to detailed comment from line 37).

The main novelties in this paper are application to the JULES model, as well as a comparison with observations of three soil moisture observation stations. I believe this is within scope of EGU journals, though HESS may be more appropriate than GMD because the tools the authors use, as well as the general concept, already exist.

We agree that the novelty in our paper relates to use of clustering techniques with the JULES land surface model and suggest that the paper fits into the development and technical paper category for GMD as a paper '... relating to technical aspects of running models' ; we note that we provide evaluation against observations and other model output as per this guidance (https://www.geoscientific-model-development.net/about/manuscript_types.html#item2). We conceived of this work as a first proof-of-concept technical step towards a more thorough analysis of the effect of clustering on the internal dynamics of processes within the model. We suggest that this future work would be more suitable for publication in HESS.

I do think that the paper needs to be clarified in multiple places (see comments in the uploaded .pdf). Briefly, the paper currently assumes more background knowledge from the reader (both on terminology, as well as on the JULES model) than I think is appropriate. Additionally, methods need to be clarified in multiple places, with a particular focus on explaining to the reader why various choices are appropriate. I particularly want to highlight the choice to run JULES at a daily time step.

We will provide clarification as suggested by adding a subsection to section 2 to introduce the JULES model, and describe in more detail how it was run here. In particular we will highlight that we in fact used a 30 min timestep in the model, making use of the CHESS daily meteorological drivers with the JULES internal disaggregator, which converts daily drivers to subdaily timesteps. (See response 7 to detailed comment from line 65).

I further think that the presented analysis on clustering outcomes and similarity between vector-based and gridded setups (section 3.1.1) would benefit from extra analysis. Currently the reader is only presented with two sets of plots for a single time step (out of a 3-year run), and only the most minimal statistics for (I believe) a spatial comparison only.

We presented snapshots of soil moisture for a single day in section 3.1.1 (figs 3 and 4) to compare the spatial patterns in soil moisture for the different model configurations in a purely qualitative way. We then compare the absolute values against time series of observations at three points in the domain for a more quantitative analysis in section 3.1.2. Our motivation here was that we have no direct observations of the spatial soil moisture patterns and we felt that there was limited value in comparing model against model when we do have some observations in the domain. Additionally, the gridded output has 1km resolution, whereas all the clustered outputs are mapped back to 50m resolution; this makes direct, quantitative comparisons difficult over the domain. We will make this clearer in the text, change from one daily snapshot to several monthly mean snapshots, and swap the order of sections 3.1.1 and 3.2.2 as the reviewer suggests.

More in-depth analysis is needed to support the statements that:
- Land use and soil type are the most important covariants,
- 1000 vector-based clusters produce sufficiently similar results to a more high-resolution gridded setup

We were careful to state that our results only imply that elevation and soil properties are important (line 131). However, to increase generality we would like to replace the text from line 129 with the following:

We note that only the simulations which include soil properties as a covariate are able to reproduce the same large-scale soil moisture patterns as in the gridded simulation, suggesting that soil properties are important for correct modelling of soil moisture in this particular domain.

Finally, the main conclusion of the paper is that similar JULES performance can be obtained by using a clustering-based spatial discretization scheme as one can get with a traditional gridded setup. First, the paper focus heavily on aggregated efficiency metrics (KGE, NSE, MAE, etc) to compare the gridded and cluster-based setups. While these aggregated scores are indeed similar between both setups, the time series plot in the paper make it very clear that these similar scores are obtained as the result of very different internal model dynamics. I believe the conclusions need to be more nuanced to reflect this.

We will make the conclusions more nuanced to reflect the fact that, although the metrics for the river flow simulations are similar, the flows themselves look quite different for the gridded and 1000LRU simulations.

Second and related, I believe that more investigation of these internal dynamics would strengthen the paper. This can involve more detailed analysis of model states and fluxes, as well as comparison to additional external data sources (such as ET) to determine if either of these two setups (gridded or cluster-based) is closer to reality - this would add a new dimension to the paper, in the sense that we would then better understand whether reducing the computational demand of running JULES also comes with a trade-off in model realism.

We agree that an investigation of the internal model dynamics would be really interesting. However, such an analysis is beyond the scope of this paper, which is intended to show that clustering approaches can be applied using the JULES model to give satisfactory soil moisture and river flow simulations in a test domain. We already compare model outputs to two sets of observations; we agree that comparison with ET would certainly be interesting, but we do not have such a dataset covering the time period for our simulations. We note that these experiments were performed without any calibration or tuning of parameters, which means that we cannot gain insight into model dynamics by comparison of optimal parameter sets as is in Gharari et al, (2020).

Third (and this is more of a suggestion), the application domain of this test case seems somewhat small to me in both time and space. I believe the paper would be strengthened if the model domain would be larger and/or the simulation times were longer.

We agree, and plan to extend this work to larger domain sizes in future work.

Please see the uploaded pdf for further comments.

We have copied each of the comments from the pdf provided and show them here in black with line numbers corresponding to the original manuscript. We have numbered our responses for ease of cross referencing, and give suggested additions to the text in italics.

Line 20: These acronyms should probably be defined before being used.

1. We will add definitions here: River Flow Model (RFM) and Total Runoff Integrating Pathways (TRIP)

Line 21: I believe the interpretation of this statement will depend heavily on the subfield one works most often in. Adding an example of what the authors consider a dedicated hydrological model, as well as how a specific process is modeled compared to an LSM might help. Perhaps a one-line comparison of snowmelt implementations or something similar is enough to clarify the point.

2. As the reviewer suggests in a subsequent comment, we will clarify which processes we are considering land surface processes, and which hydrological for the purposes of this study. See response number 5 for more details.

Line 21: Unclear if this refers to "dedicated hydrological models" or "LSMs".

3. We will clarify this by starting the sentence with 'Hydrological models....'

Line 37: Another relevant reference to discuss here is Gharari et al. (2020), where the VIC model (traditionally run as a gridded setup) is converted to an HRU-based configuration. The focus on soil moisture is still a novelty of the authors' manuscript.

Gharari, S., Clark, M. P., Mizukami, N., Knoben, W. J. M., Wong, J. S., and Pietroniro, A.: Flexible vector-based spatial configurations in land models, Hydrol. Earth Syst. Sci., 24, 5953–5971, https://doi.org/10.5194/hess-24-5953-2020, 2020

4. Thank you for pointing the authors to this study - we will add a short discussion of this work here:

Another example of this is Gharari et al. (2020), in which the Variable Infiltration Capacity model (VIC) is used in various vector-based configurations to show that more computational units do not necessarily provide better results, and that optimised parameter values do not necessarily translate across scales.

Line 42: This introduction puts quite a bit of emphasis on drawing a line between hydrologic processes and models on the one hand, and land surface processes and models on the other. Such a separation may have exist traditionally, but the lines between these communities are becoming increasingly blurred (see e.g. Archfield et al., 2015). I think this introduction would benefit from the addition of a paragraph where the authors clearly outline which processes they consider to be hydrologic, and which processes they classify as land surface. This will be especially helpful to readers at earlier career stages, who may not have the necessary experience to intuitively follow the authors' division between hydrology and land surface processes. Archfield, S. A., et al. (2015),Accelerating advances in continental domain hydrologic modeling,WaterResour. Res.,51, 10078–10091,doi:10.1002/2015WR017498.

5. We will add the following paragraph to make clear the division we are using for this study:

Land surface models are also used for hydrological predictions (e.g., Lewis and Dadson, 2021; Lewis et al., 2018; Martínez-de la Torre et al., 2019), and land surface and hydrological modelling communities are increasingly becoming less distinct, as discussed in Archfield et al. (2015). Despite the blurring between land surface and hydrological models, for the purposes of this study we make a distinction between land surface and hydrological processes. We consider land surface processes in JULES to include infiltration of precipitation into the soil, vertical transport of water through the soil column, evaporation and transpiration from soil and vegetative land cover. LSM outputs considered here are then soil moisture stores, and surface and subsurface water runoff. These land surface runoffs can then be used as inputs for a hydrological routing model, and JULES has a number of in-built options for routing surface and subsurface runoff outputs into river flow predictions, including the River Flow Model RFM (Bell et al., 2007; Dadson et al., 2011) and Total Runoff Integrating Pathways TRIP (Oki et al., 1999).

Line 42: After reading the full paper it's really not clear to me why we would need a new acronym here. We already have Grouped Response Units and Hydrologic Response Units, and the associated confusion about what either is and how (in)comparable these are, and I don't think the community would benefit from throwing LRUs into the mix. I would strongly suggest to simply stick with the more common terminology.

6. We don't suggest this as a new term for the community - just for this paper. We will note that GRU would also be an appropriate term here and add the following:

We note that we could equally well use the terminology 'Grouped response Units' or GRUs here. But we use LRUs throughout this paper to reflect the act that we have only used clusters for the land surface part of modelling and revert to the standard gridded approach for the hydrological simulations.

Line 65: This section should include a brief description of JULES. Apart from a brief description of the model, an explanation for the choice of Jules v5.9 while Jules vn7.0 is available would need to be added too.

7. We will add a JULES subsection to the Methods. We used vn5.9 rather than more recent versions just because updates happen relatively quickly compared to the timescale of some projects.

2.1 JULES land surface model

JULES is a sophisticated land surface model, used to simulate energy and water fluxes (Best et al., 2011), carbon fluxes and vegetation dynamics (Clark et. al., 2011). In this study we are primarily concerned with soil moisture and water fluxes; JULES uses the Darcy–Richards equation to model soil hydraulic processes. JULES provides estimates of soil moisture at various depths and we use the standard setup here, which comprises four layers with depths 0 to 10 cm, 10 to 35 cm, 35 to 100 cm, and 100 to 300 cm.

JULES requires meteorological driving data in standalone mode, and in this paper we have used the CHESS meteorological dataset (Robinson, 2020). The CHESS drivers have a spatial resolution of 1km and are daily averages, but are specifically designed for running JULES with its internal disaggregator switched on; this converts daily driving data to a subdaily timestep. Here we ran JULES with a timestep of 30 minutes for numerical stability, and output soil moisture, surface and subsurface runoffs as daily averages. JULES has two options for surface runoff generation, and we used the topmodel option in this work. JULES also requires (static) soil, landcover and topmodel parameters (often called ancillary files/data). For these we used data on the same grid as the CHESS meteorology, as used in Matrinez-de la Torre(2018); Blyth et al. (2019).

Line 73: Is the discrepancy between station representativeness (radius 200m) and LRU size somehow accounted for? How are situations dealt with where multiple stations are in a single LRU, or an LRU has no stations in it? Does JULES directly simulate soil moisture in the units these stations operate in or is some form of conversion needed?

8. We don't take account of the footprint discrepancy. Instead we treat the grid cell (50m or 1km) closest to each COSMOS-UK station location as our best estimate of the soil moisture there. The COSMOS-UK measurement footprint size and depth are both variable, depending on soil moisture, with the instrument being more sensitive to the soil moisture conditions closest to its position. We account for the variable depth effect using an effective observation depth from the COSMOS dataset using the approach from https://doi.org/10.5194/hess-25-2445-2021. We do not account for the variable footprint size; this will introduce representativity error and we will make this clearer in the paper by adding the following:

The COSMOS-UK soil moisture measurements correspond to variable depths and footprint extents, with the instrument measuring over a larger (smaller) footprint and deeper (shallower) into the soil under drier (wetter) conditions. When comparing COSMOS-UK observations with model output we make use of the effective observation depth provided in the COSMOS-UK dataset (Stanley et al. 2023) and calculate the average JULES soil moisture corresponding to the same depth; we do this using the depth-weighting technique outlined in Cooper et. al (2021a) and also used in Pinnington et. al (2021). We do not account for the variable footprint size or the measurement and this will be a source of representativity error. However, field scale COSMOS-UK observations are likely to be more representative of a 50m or 1km model grid cell average than a point scale soil moisture sensor. For more details about COSMOS-UK see Cooper et al. (2021c); Stanley et al.(2023).

Line 82: Why is tiling not appropriate, and why is using a dominant cover a better approach?

9. We will clarify our choice to use dominant land cover here by amending the text here to:

The CHESS landcover dataset assumes a tiled vegetation approach to account for sub km scale vegetation heterogeneity but the clustering algorithm requires one landcover type per 50m grid cell. We therefore assigned each 50m cell to be the dominant landcover type of the 1km CHESS cell which it is located in. An advantage of clustering with a finer scale underlying grid is that if 50m land cover data were available then we could use that directly here, rather than relying on a tiling approach at 1km grid scale.

Line 86: A table might be useful to communicate exactly which variables from each dataset are used for the clustering. "Meteorological driving data" and "JULES ancillary data" are not specific enough for an audience not familiar with JULES.

10. We will add more details about JULES and our JULES configuration in section 2.1 - see response 7 for proposed text. We also coloured the inputs in the schematic to show which datasets are used.

Figure 1: What units is the DTM in?

11. The units are m/10, as given by the iHDTM. We will add this information to the figure legend.

Line 88: It's my understanding that JULES is typically ran at a sub-daily timestep, and a brief look at the documentation seems to confirm this (e.g.: "In order for the numerics to remain stable, it is recommended to run JULES with a model timestep of 1 hour or shorter. If the data timestep is longer than the model timestep, interpolation is required. How interpolation is performed for a particular variable depends on whether the variable is an instantaneous state or a flux."). Why does this experiment use a daily timestep instead? Do the results of this experiment have much practical relevance if they derive from a model setup that is unlike what's used in practice?

12. We will clarify in section 2.1 that although we used the CHESS met data, we ran the model at a subdaily (30 min) timestep using the internal disaggregator. See response 7 for proposed text.

Line 92: Is there a justification for using these attributes and not others? For most of these I can see some relation with land surface processes, but it's unclear to me what the explanatory power of longitude would be, or why something like aspect isn't included.

13. We chose these as relating to land surface processes and datasets we had easy access to. Others could certainly be included, and we will make this point.

Line 92: Can these numbers be justified? The jump from 20 to 1000 seems large

14. We chose these numbers fairly arbitrarily to give some first qualitative understanding of the resulting cluster shapes.

Line 97: Does JULES use some sort of internal routing within a given model element (before the runoff from the model element is fed into the routing routine)? Given the domain size and the small number LRUs in most cases, I would imagine that simple transit times of water within the LRU would have a large impact on the accuracy of streamflow simulations, unless this travel time is somehow accounted for.

15. We remapped surface and subsurface JULES water fluxes back to the 50m grid in physical space before performing traditional gridded river routing. We will make this clearer in the text, amending line 108 to

- we map the JULES output values (soil moisture surface runoff and subsurface runoff) back onto the 50m grid using mapping files created by the clustering algorithm

Line 109: I think adding a section that discusses just the LRU outcomes is welcome, so that the reader has some way to understand how JULES is actually trying to simulate the domain with 1/10/20/1000 LRUs.

16. We note that the panels in figure 4 illustrate what the clusters look like for the various numbers of LRUs that we consider. We will swap round figures 3 and 4 as the reviewer suggests in a previous comment, and make that point more explicitly with the following text at the start of section 3.1.1:

The results presented in section 3.1.1 are not intended to provide a quantitative analysis of modelled soil moisture, but give an insight into the ways in which different clustering choices can influence the broad spatial characteristics of the clusters and resulting soil moisture patterns. In the first set of simulations, we used a constant number of clusters (N=10) and varied our choice of clustering covariates. We chose 10 LRUs for this so that the spatial patterns are relatively easy to pick out by eye, rather than any expectation that this will be a sufficient number of LRUs to match either gridded simulations or observations.

Line 112: Where can the reader find what these fixed covariates are?

17. We will state these here and add legends to each panel of updated figs 3 and 4 to make clear which simulation each set of results corresponds to

Line 113: $kg \cdot m^{-1}$ might be clearer.

18. We will change this throughout the manuscript

Line 118: It is unclear to me why 50m grids are mentioned in this sentence. The 1000 LRU setup consists of 1000 (irregularly shaped) LRUs, right? Why is the grid size of the data that went into creating the LRUs important?

19. We will make it clearer in the text that the JULES outputs are mapped back onto a 50m resolution grid before plotting and comparing to observed data. See also response number 15.

Line 121: I don't believe the one set of plots and one set of statistics about a single simulation day are sufficient to support this claim. How do these comparisons vary in time? What are the maximum deviations between the 1km-gridded setup and each of the LRU setups? (Much) more analysis is needed here.

20. See response 16 to related comment - we will add some results from drier conditions and use monthly means rather than daily snaphots. We will also clarify that the aim here is not to make a quantitative comparison, just to show that the clustering choices impact on the spatial features of the clusters and the resulting soil moisture. See response 16 for proposed additional text.

Figure 3: The text on these figures is quite small and vague.

21. We will increase the resolution and add captions on the revised figures 3 and 4 to make the clustering choices clearer.

Line 122: Why N=10? The previous results seems to show a substantial loss in simulation accuracy with 10 LRUs compared to the gridded setup. Why is N=10 then a helpful number of LRUs to use here?

22. We wanted to qualitatively demonstrate that the choice of clustering covariates produces different spatial patterns of clusters and hence soil moisture. This is much more obvious for 10 LRUs than a larger number. We make this clearer in the text as per response 16.

Line 130: The caption says figure 4f is based on land cover and soil type, not elevation.

23. Thanks for spotting this typo. We would like to change the lines here to:

We note that only the simulations which include soil properties as a covariate are able to reproduce the same large-scale soil moisture patterns as in the gridded simulation, suggesting that soil properties are important for correct modelling of soil moisture in this particular domain.

Line 131: 1. Upon reading this paragraph I think I understand its purpose better than at the start. If the intent here is to find out which covariates are helpful, it might make more sense to have these results as the first section under 3. Results.

24.1 Yes - we agree, and will switch the order of figs 3 and 4.

Line 131: 2. I do think that more systematic testing of all combinations of covariates, or justifying why these specific combinations of covariates make sense to test, is needed.

24.2 This would indeed be interesting, but the scope of this paper was really just to see whether we could implement this technique with JULES and get sensible results compared to various observation sets. Future work would involve a more systematic testing of combinations of covariates as the reviewer suggests, as well as an analysis of how different model processes are affected.

Line 131: 3. That said, the "similarities between panels a and f in figure 4" (line 129-131) are hampered by the use of N=10 clusters. This analysis would make more sense with a higher number of LRUs, because that limits the differences between the LRU outcomes and gridded outcomes, and therefore allows a cleaner attribution of any differences in the figure panels to difference in covariates. I would suggest to re-do this with N=1000.

24.3 We would like to change the lines mentioned here as detailed in response 23.

Line 131: 4. Also here I believe that showing plots for a single day is insufficient to support the claims made. More analysis is needed.

24.4 We will provide more aggregated snapshots here, with monthly mean results for wet and dry conditions. We will also amend the text to clarify that we are only aiming to show the broad spatial clustering patterns. See response 16 for details.

Line 131: 5. Finally, this analysis is currently somewhat left hanging, because both the previous results (how many LRUs are needed) and the next section (simulation accuracy) use a setup based on all covariates (and not just on land cover and soil type).

24.5 We will reword those lines to say that our results suggest that soil properties need to be included in this domain - we agree that our results do not justify any more substantial claims about which covariates would be best to use. (See response 23 for more details)

Line 145: Specifically, that the simulations are better than the annual mean

25. Yes - we will state this here.

Line 164: Can the authors speculate why the bias is so much better in the LRU1000 SHEEP run?

26. It may be that the soil properties are not well specified in the HWSD at some or all of the sites. This work assumes that the soil textures are accurate.

Line 168: It would be nice to see those plotted in the same figure that shows where the COSMOS stations are

27. We will add gauge locations to figure 1.

Figure 5: These figures need a unit on the y-axis

28. Thanks for spotting this - we will add y-axis labels.

Figure 5: I assume this is the total KGE score. If so, I expect that calling this "KGE" instead of "KGM" will be more intuitive to readers.

29. Thanks for spotting this inconsistency, we will change the subplot title here.

Figure 6: Units on the y-axes would be helpful

30. Thank you for spotting this - we will add y-axis labels

Figure 6: Given that the text compares these simulations against KGE = -0.41, it seems odd to cap these axes at KGE = 0.

31. We didn't see much value in adding the extra y-axis span here but will mention it in the text.

Line 178: I don't see those regridded simulations in Fig. 7. Is that intentional?

32. We did not include the regridded hydrographs but will include them in an updated manuscript.

Line 202: Judging from Figure 8, it seems that most, if not all, of this is a consequence of the higher resolution routing grid: the boxplots of gridded and regridded simulations (red and green) seem to mostly overlap, especially for overall KGE and alpha and beta metrics. The one thing that may reasonably be claimed is that the LRU approach seems to slightly improve the correlation component (presumably by reducing the excessive flashiness seen in Fig 7). It seems like straightforward test to just run a JULES simulations at a 50m grid and either confirm or disprove the highlighted sentence, even if a 50m resolution would be too computationally expensive to use as the resolution for typical use. If this is deemed to expensive, than disaggregating the 1km gridded results to 50m and routing those may be an acceptable substitute too.

33. Our 'regridded' simulations were designed to remove the effect of the 50m resolution river routing as the reviewer suggests - i.e. we first ran the 50m 1000LRU runoff outputs at 50m river resolution, and then aggregated the 50m, 1000LRU clustered runoff outputs to 1km and routed them at 1km. We will make this clearer in the text.

Line 224: I think this conclusion is incomplete without mentioning the more detailed insights available in the results. While the aggregated scores are comparable, the time series plots show that using LRUs instead of gridded approaches also means trading a certain set of errors for a completely different set:

- In the soil moisture plots it can be seen that there is no consistency in the bias of the LRU runs compared to the gridded setup. By switching to LRUs bias might remain the same, become better or worse.

- In the streamflow plots, a model configuration that produce much too flashy hydrographs is replaced by a model config that has too little variability.

In other words, by switching to a more computationally efficient JULES setup that uses LRUs, the aggregated metrics might remain similar but the model's internal dynamics become substantially different. Prior experience with (gridded-)JULES internal functioning may no longer apply to understanding of the LRU-based setup, nor may previous assessments of JULES internal functioning apply to these new LRU-based configurations. I believe this needs to be made explicit in these conclusions.

34. We agree that we could make these points more explicitly in the conclusions but note that we did not so any parameter tuning or calibration in these experiments, which means we are unable to provide a more detailed analysis of differences in model dynamics (as in e.g. Gharari et. al, 2020) . We propose to add the following text:

We note that although the aggregated metrics for river flows were similar for the gridded and 1000LRU regridded river flow metrics, the hydrographs were very different, suggesting different model dynamics. A detailed analysis of how JULES land surface and river routing dynamics vary across scales is beyond the scope of this paper but will be vital in understanding how to best employ techniques such as the clustering used here.

Line 233: This link requires login credentials, which according to the website require an institutional sponsor. This section needs to be updated with instructions on how to obtain access to the code, or an explanation about why this code base cannot be made open.

35. We will add this

Line 234: Also requires institutional access

36. We will add this information

Reviewer #2

This paper presents a case study of the JULES land model while adapting the clustering approach to update the traditional grid approach. This effort is shown to improve the computational efficiency of the JULES model on the premise that the accuracy is similar to the two approaches. However, the HydroBlocks model and the cluster approach were clearly presented in Chancy et al. (2016) as the authors cited, the novelty of the JULES vn5.9 is supposed to be more stressed in Methods, or this work is more like an application of the HydroBlocks tool, which is a significant effort but doesn't fall in the GMD's Development and technical paper scope very well. I also think that the manuscript needs to be more structured to present a logical and complete work.

The novelty of our study is the use of JULES and we propose to add a subsection in section 2 with more details on the JULES model (see response 7 to reviewer 1's pdf comments for details). We suggest that the paper fits into the development and technical paper category for GMD as a paper '... relating to technical aspects of running models' ; we note that we provide evaluation against observations and other model output as per this guidance (https://www.geoscientific-model-development.net/about/manuscript_types.html#item2)

Below is a list of my detailed comments:

1. Detailed information on each set of simulations of this study is suggested (e.g. in section 3.1.1), which might include the value of the covariates, the variable N, and the grid resolution. For this, a table might be useful. The comparisons under different conditions will be clearer and more reliable, or the reader will be confused.

We will add captions to each panel in updates to figures 3 and 4 to make it clearer which simulation each set of results corresponds to.

2. The authors mentioned “This indicates that a ten fold reduction in JULES compute expense can yield comparable results to the 1km gridded approach.” (L119-120) and “… the JULES regridded approach gives similar overall KGE results to the original gridded approach, while still benefiting from a ten fold reduction in compute resource.” (L204-205). What does the compute resource refer to in L205?

We have ten times fewer JULES runs to do if we use 1000LRUS, as we need only one per LRU (1000) rather than one per grid cell (10, 450). We will make this clearer in the text.

3. It is suggested to clarify the distribution characteristics of the soil moisture data because the KGE metric was built up based on the normal distribution.

The KGE metric is commonly used for river flow time series metrics; we are also aware of several cases in which the KGE metric is used for soil moisture time series – e.g. Vergopolan, N., Chaney, N.W., Pan, M. *et al.* SMAP-HydroBlocks, a 30-m satellite-based soil moisture dataset for the conterminous US. *Sci Data* **8**, 264 (2021). https://doi.org/10.1038/s41597-021-01050-2, and Yong Chen, Huiling Yuan, Yize Yang, Ruochen Sun,Sub-daily soil moisture estimate using dynamic Bayesian model averaging,Journal of Hydrology, 2020, https://doi.org/10.1016/j.jhydrol.2020.125445.

4. The locations of the 26 gauges the authors mentioned in L168 are suggested to be mapped.

We will add the gauge locations to figure 1.

5. I do think that Introduction is more focused on the technique while lacking a view of the further influence of the tool and the meaning of the work to regulators and decision makers. The sentence in L52-54 seems like the conclusion of this study and is not recommended in this part. Moreover, the novelty of the JULES vn5.9 should be stressed in Methods.

This proof-of-concept study is not aimed at regulators and decision makers. Our paper is more of a case study to demonstrate that clustering techniques applied to JULES can give satisfactory results in terms of soil moisture and river flow simulation. This is likely of interest in the land surface and hydrology modelling communities. We will add a JULES description section as part of the methods to make the model capabilities clearer (see response 7 to reviewer’ 1 s detailed pdf comments for details).

6. The format of units should be uniform (italic or not). The units of the y-axis in most figures should be presented. The prat of the title in the Fig. 6 should be checked. The set 100LRU (L120) is not mentioned and there might be a writing mistake.

We will make sure all units are consistent in presentation, and add y-axis labels where missing. Thank you for spotting a typo in line 120, we will correct 100LRU to 1000LRU.