Answer to the comments by anonymous reviewer#1

We thank the reviewer for their constructive comments on our work. For a fruitful open discussion and to clarify some points we have drafted this early reply to the reviewer. A more detailed reply will follow the revised version identifying point by point changes to the manuscript based on reviewer's comments/suggestions. For convenience, the reviewer's comments are given in blue.

Gharari et al present an application of the VIC model, using Grouped Response Units to define computational units, rather than grids. It is acknowledged that this concept was already presented in 1993. I do think it is justified to re-introduce older concepts if these can serve the science of today, however, then the re-introduction should also deal with some of the challenges of today, and this is currently not the case.

As the reviewer rightly mentioned, this manuscript does not intend to introduce the concept of GRU, instead it tries to used it as a base for implementation of the VIC model which has been used worldwide and hopefully draw the attention for wider land model community to use the concept of vector-based setup based on GRUs.

Solving the "challenges of today" (which the reviewer describes in more depth in their later comments, and which we respond to in more depth later in this document) is not the main goal of this paper. In this manuscript, we try to point out the technical and scientific advantages of using a vector-based setup based on the concept of GRU. To do so, here we reflect on one of often not very well explored challenges in land modeling community, that is trade-off between accuracy of the land models' spatial representation and their performance. We hope our paper sheds some light on the ongoing discussion. The GRU concept was very helpful in this respect as we could change the resolution of forcing without really affecting the parameter values at the GRU level as there is no upscaling to the grid resolution. We think that is major advantage that we highlighted in this manuscript.

Firstly, the reader has to do quite some searching to fully capture the concept of GRU's, and its comparison to HRU's. Only when the investigated cases are presented it be- comes clear what a GRU exactly is and the choices it encompasses when defining GRU's. This seems to be the result of an overall quite weak structure in the manuscript; the introduction does not clearly present the aim or goal, probably because the structural test (case 3, presented in the intro in line 103-110) seems to be completely out of context. In the same fashion, 3.1.3 is not well embedded. Furthermore, sections are not logically structured, e.g. subsection 3.3 only consists of 2 sentences while some subsubsections are longer, and parameter are presented well after the calibration is discussed and the cases are introduced. I suggest restructuring the manuscript, clearly introducing the concepts with simple examples, and omitting parts that do not fit the aim or goal of the study.

We agree with the reviewer that the introduction structure can be improved. We agree that the third experiment is somewhat out of the scope of this paper (although experiment three is related to the parameter/process uncertainty; will be mentioned in the following). We will revise this manuscript, clarify the concept of the GRU earlier in the manuscript and remove the experiment three (or perhaps move it to an appendix and refer to it shortly).

One of the key questions in defining the spatial discretization of models is of course the calibration. Whereas the GRU's conceptually might make sense compared to grid cells, it introduces new questions on how to calibrate the parameters, and this is not well explained in the text. Does each GRU receive its own set of parameters? And is this then related in any way to the underlying data? As the authors rightly suggest, parameter ranges can be adapted based on soil type or land use, but it seems this was not done by the authors. Not surprisingly, the results demonstrate some of the already known flaws from calibrating on discharge outlet; the everlasting problem of equifinality and overparameterization. If the authors believe the GRU concept is valuable to reintroduce (and I can see it has potential), this value should be demonstrated in a more sophisticated calibration. If the same calibration is done as for usual grid-models, of course we know we can achieve good model performance because there are enough buttons to push, but what do we learn from it compared to a grid-based model and what does it add? 1000 evaluations in the calibration procedure seems rather limited given the dimensions of the problem; this is understandable from a computational point of view, but also a chance to demonstrate why GRU's make more sense than grids within these bounds, by making use of the opportunities that GRU's offer in comparison to grids.

We thank the reviewer for this comment. We fully agree with the reviewer's comments on the parameters' values and estimation. Everything boils down to how GRUs are parameterized. We will better explain our approach to GRU parametrization during our revisions. We will explain how the parameters for each GRU are derived and how this relates to available data. Just a brief explanation here:

1- The soil layers get the same set of parameters for bulk density, saturated hydraulic conductivity (no difference between the vertical soil layers).
2- The conceptual soil parameters such as binf are unified across the scale (similar to most of VIC application).
3- The soil depth that conceptually define the storage of the system are defined based on land cover. The forested areas have deeper soil (or root zone) to allow for larger storage and transpiration. Just to mention that is this is an advantage of the vector-based implementation.
4- K_{slow} is similar for the entire system (or a gauges) as it can be inferred/calibrated only from the recession analysis.

Of course, more intuitive and sophisticated parameter allocation can be explored but the above-mentioned parameter selection is aligned with what is often done for calibrating the VIC model. This is purposefully not to make the regionalisation so complex that the manuscript deviates from its own message (which is vector-based implementation and accuracy-trade-off implementation).

I know my colleagues who work with MESH model sometime do this distinction between parameter values of various GRUs in their applications/scientific explorations, for example, different soil with different land cover have different parameters. I personally do not move to

that direction for few reasons: (1) the parameters of the spatially largest GRU will be the most sensitive ones when calibrating against the observed streamflow (or polishing of smaller GRUs that have very small contribution, <1%, may be needed), (2) expansion of parameter for calibration that we don't know how to tied actually will unnecessarily add to the dimension of the problem (no information tangible construct them). There is ongoing effort to relate the parameters to physical characteristics but each of the decisions in itself is an assumption and cannot be inferred directly from the data (for example Mizukami et al., 2017 Table-3). We totally agree with the reviewer on "this value should be demonstrated in a more sophisticated calibration" but at the same time we have not much data for the sophisticated calibration especially the entire subsurface flow movement.
We should also emphasize that part of the motivation for GRU is computational efficiency with respect to optimization, for example in the MESH model. The underlying assumption is that grouping units from a parametrization perspective, since we expect them to behave in a physical similar way, allows us to characterize the variability with respect to the forcing and of course the subsequent hydrological response, while maintaining the degrees of freedom for parameter estimation reasonable. Our accuracy-performance trade-off is aligned with this mentality also.

Reflecting on reviewer comment, there might be two benefits of the GRU implementation:

1- Technical aspect; which is the ease of parameter allocation to a GRU (as each GRU has a specific land cover and soil type), or better implementation of regionalization rules if applicable. Easier coupling with vector-based routing.
2- The scientific values of implementing the models in a GRU approach. That is a grand challenge and an ongoing development. For example, how to effectively parameterize the model simulation at GRUs. I personally think part of the reason that the GRU was overlooked or not implemented widely for land models, as the reviewer mentioned with sophisticated calibration, is the lack of data and proper understanding of how parameters behave at the scale of modeling. One of the scientific applications we had here is the trad-off between accuracy of spatial representation and model performance.

Perhaps, I agree that the current manuscript has more emphasis on the first than the latter but at the same time the GRU implementation can be vehicle to explore the more scientific questions. For sake of simplicity and to emphasis on the advantages of the GRU implementation, we tried to put less emphasis on the sophisticated parameterization. As computational hydrology team at University of Saskatchewan we are moving to face this grand challenge in future.

The modeling set up, calibration and parameter perturbation for calibration are based on the objective of the modeling. Our objective here is efficiently-accuracy trade-off for land model and hydrograph simulation.

Reflecting on the computational costs of the setups:

1- 1000 simulations are selected as an arbitrary value we chose based on the computational infrastructure (and time) we had available for the Case4-4km which has 6000 computational units (the most computationally expensive setup).

2- For context, 6000 computational units will be equivalent of approximately 1/3 of the CONUS domain with standard gridded simulation at 0.25 by 0.25 degree lat/lon.
3- Consequently, running 1000 simulations for the Case4-4km takes 5 days on 50 CPUs of ComputeCanada infrastructure (assuming nothing goes wrong with the job), or approximately 8 months on a single CPU.
4- We have tried the impact of longer calibration runs for simpler cases (up to couple of thousands of simulations) but did not find a noticeable increase in NS scores.
5- Also, our choice for hourly forcing data increases the storage space needed by almost a factor 40 compared to daily inputs that is used for VIC-4 and earlier.

An example: The results from Figure 4 are criticized in the text as: "The result indicates the two parameters that are often fixed or a priori allocated based on look up tables can exhibit significant uncertainty and non-identifiability". The Brooks-Corey coefficient is from such a high conceptual level that it might be challenging to find good values in lookup tables, but Ksat might be able to be estimated. The lookup tables can then provide an indication for a search range for the parameter and decrease the equifinality issues with these two parameters.

The soil data that need to inform this choice are themselves so uncertain that they may not guarantee more "realistic k_{sat} ranges". Also sub-resolution heterogeneity is not account for, nor is k_{sat} very relevant if the dominant flowpath is macropores (which might be very well the case given the significant elevation differences in the region of study). Therefore, we didn't try to a priori limit the k_{sat} calibration ranges too much. This uncertainty in k_{sat} also has high implication for the future regionalization that might be built partly on k_{sat}.
We also wanted to indicate that parameters may be more uncertain that what is suggested in look up tables. It is often the case the land modeling community "kill" the potential uncertainty in parameters and processes either by assigning (hardcoded) parameter values from look up tables or using calibration techniques that yield single best solutions (Mendoza et al., 2014).

Shortly, I can see why GRU's might have added value in land-surface modeling. How- ever, the re-introduction of this concept in this manuscript might not make a very good case to convince people of this fact, given that calibration is one of the main challenges and the potential for GRU's in this context is not well explored.

We thank the reviewer for his/her constructive suggestions. We try to improve the manuscript flow. Meanwhile, we would like emphasis that the focus of this manuscript is not to come up with the parameterization scheme for the model but instead to provide an alternative representation of spatial data that can be from benefit for land modeling community. We demonstrated that how an existing model, such as VIC, can easily be implemented in a vector-based framework. Moreover, the vector-based implementation is not tied to any calibration strategy. A modeller may use the vector-based set up for a land model while avoid any automatic calibration.

Other suggestions:

In section 2.3, it remains unclear why structural changes to the model were made. Some of the most sensitive parameters of the model (Ds, Dm) have been replaced by a linear reservoir coefficient. Furthermore, the description focusses on VIC4 while VIC5 was explored. Why is that?

It is a very good point for discussion.

1- VIC has a baseflow formulation that has 6 parameters, 4 for the baseflow formulation and 2 for the physical specification of the depth and porosity of the lower layer. These 6 parameters are impossible to be inferred from the recession analysis of a hydrograph.
2- The common structure of the land models does not allow for recession analysis based on master recession curve. The soil layers act like a cascade of reservoirs. The water movement from the second layer of VIC can already be damped enough that it may not need even need a slow baseflow reservoir (Gharari et al., 2019).
3- One simple solution to that is to basically seize the micropore water movement to the baseflow layer, and only allow macropore water movement based on fraction of surface runoff for example (experiment three).
4- The recession coefficient of the Bow River at Banff is in scale of 0.01 day^{-1} or 100 days. This is similar to what we get from the automatic calibration. K_{slow} is one identifiable and sensitive parameter given the regime of the Bow River.

Going back to the "sophisticated calibration". Here we are somehow calibrating the model structure based on the general hydrological knowledge that water movement is mostly preferential (macropore) in areas with high elevation differences. It can be assumed that the preferential flow pathways are activity passing the water in comparison with micropore water movement in the region of study that has such a huge elevation difference. Subsequently any inference or assumption on K_{sat} based on an ill posed model structure may not be a valid assumption. This if one of the points the manuscript tries to make.

We mentioned VIC-4 to emphasis why VIC was used so widely. We will remove the explanation on VIC-4.

It is not explained how the parameters in Table 2 were selected for calibration. It is for instance remarkable that no snow parameters, such as snow roughness, are included in the calibration – is this because GRU's focus on soil and land use? Furthermore, it is not clarified to which soil layer Eexp and Ksat refer, or is this kept constant over both soil layers?

We have indeed chosen to only calibrate those parameters that relate to aspects of our GRU configuration (and more importantly forcing resolution for accuracy-efficiently trade-off). We will clarify this in the text. We also set the K_{sat} and E_{exp} similar for both layers. We will clarify that as well.

Minor for tables and figures:

Table 1 the unit of forcing resolution is missing (degree) Figure 3 the a,b,c labels are missing, the legend is not readable. Figure 4 not sure if this is very informative. More interesting to see a boxplot of every parameter to demonstrate the wide range. Figure 5 Caption says "deviation" but you demonstrate NSE compared to benchmark run, and not the deviation in NSE.

We will fix the forcing resolution in Table-1.

We will fix Figure-3.

We will replace figure 4 with a more informative possibly a box plot.

Please not that this is deviation from synthetic case and not standard deviation. We will rephrase this part.

We thank the reviewer for the constructive comments, and we hope to enrich our manuscript by addressing the reviewer's comment sufficiently and successfully.

With kind regards,

Shervan Gharari, on behalf of the co-authors

References:

Gharari, S., Clark, M., Mizukami, N., Wong, J.S., Pietroniro, A. and Wheater, H., 2019. Improving the representation of subsurface water movement in land models. *Journal of Hydrometeorology*, (2019).

Mendoza, P.A., Clark, M.P., Barlage, M., Rajagopalan, B., Samaniego, L., Abramowitz, G. and Gupta, H., 2015. Are we unnecessarily constraining the agility of complex process-based models?. *Water Resources Research*, *51*(1), pp.716-728.

Mizukami, N., Clark, M.P., Newman, A.J., Wood, A.W., Gutmann, E.D., Nijssen, B., Rakovec, O. and Samaniego, L., 2017. Towards seamless large-domain parameter estimation for hydrologic models. *Water Resources Research*, *53*(9), pp.8020-8040.