The reviewer's original comments are in black*, our responses are in blue italic fonts, and modifications made to the original manuscript in* blue plain text.

Referee #1

This is a review of "On the similarity of hillslope hydrologic function: a process-based approach" by Maina and colleagues submitted to the Hydrology & Earth System Science. This manuscript details hydrologic classification at the regional catchment scale focused on groundwater parameters gleaned from modeling, but also including a full range of related hydrologic and geomorphologic characteristics. The ability to classify hydrologic function across the landscape is important for large-scale hydrologic modeling, and in turn, for predictions of future water resources. To my knowledge, this work is novel particularly for the focus on including groundwater dynamics in the classification effort. This manuscript is within the scope of the journal and will be of value to the readership. Overall the writing is clear, and the figures are well drafted. My general and specific comments are provided below.

We thank the reviewer for their constructive comments and feedback. Based on the feedback, we have made significant changes to the manuscript. Please see below for our detailed responses to the comments. The reviewer's original comments are in black*, our responses are in blue italic fonts, and modifications made to the original manuscript in* blue plain text.

General Comments:
The manuscript may benefit from some improvement in organization. For example, the core purpose of the manuscript appears to focus on classification, however classification methods are not described or justified in the Methods section.

The revised manuscript has now a "methods" section with 4 subsections: modeling framework, hillslope delineation, hillslope clustering, and the methods we used to compare the different clustering approaches.

As far as I can tell, groundwater dynamics are only modeled and direct measurements are not used to calibrate or validate the model results. There appear to be some groundwater data in ESS-DIVE – would it be possible to use these to support the modeling results?
Williams K ; Carroll R ; Dong W ; Versteeg R ; Tokunaga T (2020): Water Level Data from Wells PLM1 and PLM6 for the East River Watershed, Colorado. Watershed Function SFA, ESS- DIVE repository. Dataset. doi:10.15485/1818367
Dafflon B ; Dwivedi D (2020): Groundwater level elevation and temperature at the Lower Montane in the East River Watershed, Colorado. Watershed Function SFA, ESS-DIVE repository. Dataset. doi:10.15485/1647040

These measurements were performed at a site called pumphouse (see the location of the measurements in blue in the Figure below). The distances between these measurements are a few meters smaller than the grid of ParFlow, as a result, most of these measurements (e.g., the wells PLM1, PLM7, etc.) fall into one grid of the numerical model. Because these ground measurements are located close to the river, water table depth fluctuations reflect the changes in river stages. Comparisons between measured and simulated river stages were shown in Foster and Maxwell, (2019).
The measured groundwater levels show two time periods: a decrease during baseflow conditions followed by an increase through the rest of the year as a result of snowmelt, as depicted in Figure 2 and discussed in section 2.3.
In Figure R1 we compare measured and simulated groundwater levels. Because the grid size of the numerical model is equal to 100 m and contains many measurements each of them with different values and timing, we averaged all the measurements that are within one ParFlow grid cell. The two other lines of the graph showed two measurements that are close to the center of the ParFlow grid cell. The differences

between measured and simulated values are likely due to the differences of scale and location between the measurements that provide value at a point location and ParFlow.
We have added this discussion to the revised manuscript, please refer to lines 269-270 and appendix A.

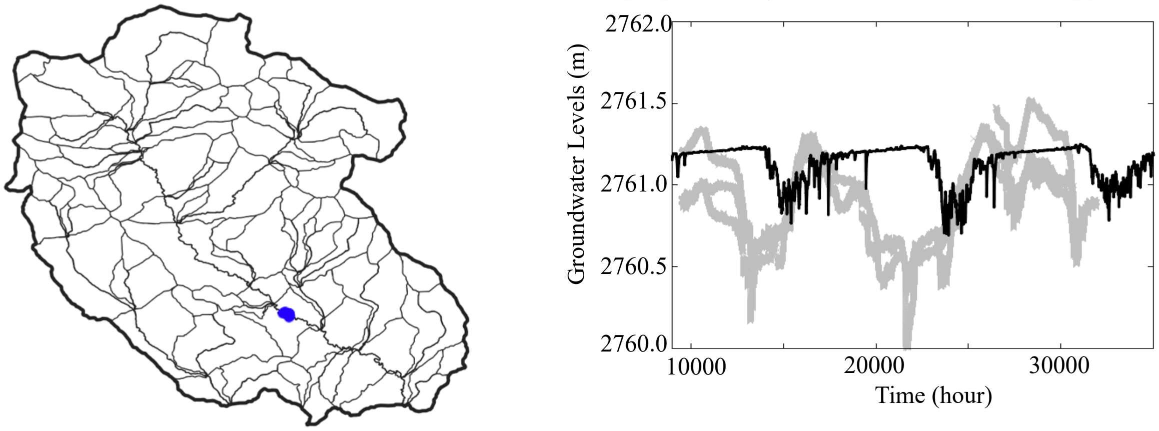

Figure R1: location of groundwater levels measurements and comparisons between measured (gray) and simulated (black) groundwater levels. The x axis indicates the number of hours since 10/01/2015.

I had difficulty interpreting the ‘distribution plots’ Figure 6-9. Perhaps it is because I am unaccustomed to interpreting this type of plot, or perhaps it is because the x-scale is squeezed to fit all of the lines on, but it could be more clear what type of patterns the reader should be looking for and how to draw specific conclusions from these graphs. Perhaps considering an alternative way to present this data would be useful, or maybe include an explicit explanation for how to interpret them for the reader.

We have changed the figure and added more description. Please see below.

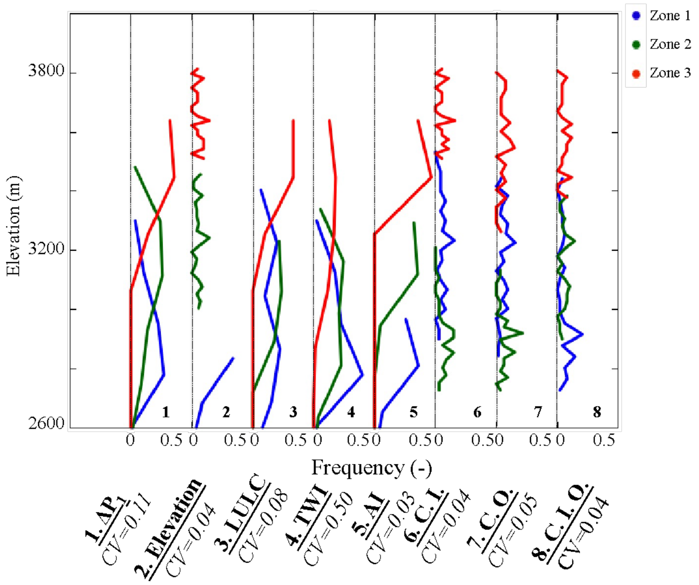

Figure 6: Frequency distributions of hillslope elevation. Clustering approaches are based on ΔP1, elevation, land cover (LULC), topographic wetness index (TWI), aridity index (AI), and machine-learning approaches (with inputs C.I., outputs C.O., and inputs and outputs C.I.O). Hillslope clustering approaches

are located across the x-axis. Note that we plotted the distributions of the 8 clustering approaches on the same graph, between each dotted line (frequency from 0 to 0.5) are plotted the frequency distributions of the three zones derived from the clustering.

The authors may wish to consider splitting section 3 up into separate 'results' and 'discussion' sections to improve organization and help guide the reader to where data are presented vs where they are contextualized. In general, the discussion content of this section builds primarily examines patterns within the results of this study and builds only limited links to past work – more complete referencing in the discussion may help to improve contextualization of this research within the broader body of scientific work.

We have now added a discussion section and moved the following sections to the discussion: Advantages of a similarity index based ΔP and Similarities in hydrologic responses to wet and dry conditions. The discussion section also contains comparisons and references to previous studies.

The 'summary and conclusions' section could be condensed by removing the summary and focusing on the core conclusions of this research.

We have removed the summary section.

Line Comments:
Line # | "quote from manuscript" , Comment .
69-71: Perhaps provide a reference as evidence that would call this assumption into question?

We have added the following reference: (McDonnell & Woods, 2004).

85: "300 mm" Does this mean 'within a single hillslope'? Where does this number come from?

We have clarified that this is within a single hillslope and added the following reference (Wainwright et al., 2022).

85-86: Where do the order-of-magnitude numbers come from?

They come from the following references: Hubbard et al, 2018 and Wainwright et al., 2022. We have added them to the revised manuscript.

98-100: "In this study...functional zonation." As I understand this, the authors are defining "functional zonation" as the seasonal change in groundwater levels. Later, on line 110, the authors appear to state that 'hydrologic function' is an equivalent term of 'functional zonation.' (my apologies if I have misunderstood this). If this is the case, why not just stick with the term as originally defined and be consistent, rather than introduce a synonym that may add confusion? Additionally, it is somewhat unclear how the 'integrated hydrodynamic response' can be effectively captured by simply the seasonal changes in groundwater level – this would seem to ignore any unsaturated zone dynamics that do not directly affect the water table such as storage, partitioning, plant water use, etc. I recognize the argument present on line 89 that groundwater is linked with unsaturated zone processes, however Maxwell and Condon found this on a continental scale using 1km model cell resolution, and it is not clear that the same relationship is robust at the hillslope scale. I am not suggesting that groundwater level is unimportant, but rather that it's unclear if it is truly appropriate as a proxy that 'integrates' the whole hydrologic "story" of a hillslope or watershed.

A hydrologic function of a watershed or a hillslope describes its ability to partition, store, and release water (Sivapalan, 2006; Wagener et al., 2007; Wainwright et al., 2022). We have clarified it in the following

sentence (lines 267-269): “These two key variables allow us to quantify water release (ΔP1) and recharge (ΔP2) within a hillslope, two key dynamics of the watershed hydrologic function (Sivapalan, 2006; Wagener et al., 2007, Wainwright et al., 2022).”
We proposed the use of a clustering based on the changes in groundwater to find hillslopes with similar hydrologic function because groundwater changes include the partition, storage, and release of water controlled by land surface processes, unsaturated zone and subsurface dynamics. Our work highlighted that groundwater is able to capture both unsaturated zones and land surface processes as shown in Figures 6-9. We have now clarified that our work is based on the assumption that groundwater dynamics can be a good proxy for identifying hillslopes with similar unsaturated zones and land surface processes and this assumption has been confirmed with our results. Although in some areas groundwater dynamics are somewhat disconnected from unsaturated and land surface processes in this watershed (zone 3), our clustering allows identifying these areas. By identifying hillslopes with similar groundwater dynamics, we were able to categorize hillslopes with similar snow dynamics, evapotranspiration, and atmospheric dynamics. These dynamics control the partition, storage, and release of water.

Figure 2: The y-axis tickmarks could be improved: WTD only 1m and 3m are labeled, and tick marks of apparently .3333 m are provided, which is a bit awkward as an uneven number. on the SWE exist, the 200mm and 800mm labels don’t appear to line up with any tick marks.

We thank the reviewer for catching this error. We have changed the labels of Figure. Please see below.

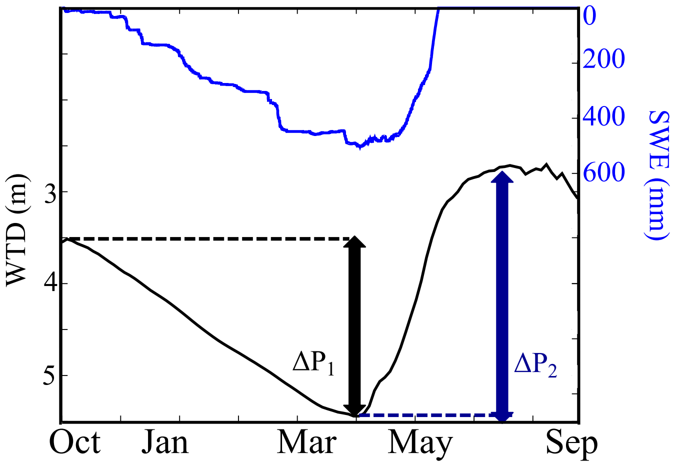

Figure 2: Temporal variations of water table depth (WTD) and SWE at an example hillslope. The location of the hillslope is shown in Figure 1.

242-246: I am unclear: is the WTD plotted in Fig 2 a measured value or a model output? from this sentence it appears to be a model output, but it is unclear.

We have clarified that the WTD is a model output not a measured value.

248: “beginning of snowmelt (i.e., May)”: based on Figure 2, it appears that snowmelt might begin in April or perhaps earlier?

We have now stated the exact beginning of snowmelt which is April.

249: perhaps this is nuanced, but the timing of events on Figure 2 is slightly different than noted in the text: apparently, the groundwater level begins to rise somewhat before the SWE begins to decrease substantially. Maybe the issue is just that the text is describing snowmelt, while the graphic is illustrating SWE (i.e., there may be substantial snowmelt occurring before SWE begins to decrease.

We have now clarified these timings and changed the sentence now to read (lines 245-250): "In this mountainous watershed, where the largest changes in WTD are mostly a result of snowmelt, WTD increases from the beginning of the WY (i.e., October) to the beginning of snowmelt (i.e., starting from April). As the snow starts to melt and precipitation starts to fall as rain instead of snow, WTD starts to rise. The shallowest WTD is June and July when the snow has completely melted and has had time to percolate through the unsaturated zone into the groundwater."

249: "peak discharge is mostly observed in June and July when the snow melts" Again, my apologies for being pedantic, but based on how I read the graph, it appears that SWE goes to zero by around May 13th, so somewhat before peak discharge.

*We have now clarified these timings and changed the sentence now to read (lines 245-250): "*In this mountainous watershed, where the largest changes in WTD are mostly a result of snowmelt, WTD increases from the beginning of the WY (i.e., October) to the beginning of snowmelt (i.e., starting from April). As the snow starts to melt and precipitation starts to fall as rain instead of snow, WTD starts to rise. The shallowest WTD is June and July when the snow has completely melted and has had time to percolate through the unsaturated zone into the groundwater."

253: deltaP1 & P2: are these parameters defined for the first time in this paper, or is there a reference that could be cited with a more precise definition? "This variable indicates the ability of the hillslope to release water" This seems vague: would it also be dependent on inputs, antecedent conditions, etc.? it appears to carry units of "meters" so it's unclear how it quantifies the ability of a hillslope to release water. Similar comment WRT delta P2 "contains information about the storage and the recharge capacity"

Yes, to the best of our knowledge, these parameters are defined for the first time in this paper. We have clarified these sentences now to read (lines 252-266):
The dynamics show two periods characterize the dynamics of the hillslope: from the initial conditions to the baseflow conditions when the hillslope is losing water, then from baseflow conditions to the peak of WTD when the hillslope is gaining water. To characterize these groundwater dynamics, we define two variables:

- ΔP_1 represents the change in WTD between the beginning of the water year and the deepest WTD during the baseflow conditions. This variable quantifies the amount of water released by the hillslope during the dry period at the beginning of the water year. It thus contains information about the amount of water that the hillslope typically releases/loses, mainly by ET and discharge, given its physical characteristics and climate dynamics.
- ΔP_2 represents the changes in WTD between the peak flow (i.e., the period with the shallowest WTD) and the baseflow conditions. ΔP_2 quantifies the amount of water gained in the hillslope by recharge, and thus contains information about the recharge ability of the hillslope given its physical characteristics and climate dynamics.

264-267: reference needed or more complete explanation?

We have added more clarifications (please refer to the previous response) since these parameters are defined for the first time in this paper and we do not have references.

268: “Figure 3 shows the classification” Is this really showing anything ‘classified’ – the caption seems to be more accurate “spatial distribution of average values. Perhaps I am misunderstanding and the text could just be clarified.

We have changed to “Figure 3 shows the spatial distribution of hillslope values”

280-281: “These two patterns are different from each other, and they are also different from the ones associated with the land surface processes...” It may be helpful to provide a brief characterization of how these patterns are different.

We have changed the sentence now to read (lines 316-323) “As expected, the hillslopes characterized by high SWE have high precipitation and low temperatures in contrast to the hillslopes with low SWE. However, ET shows a different pattern, because it depends on both water availability and ET demands, which depends on the type of land cover. The mid-elevation zone (i.e., zone 2) with a high coverage of forests has high ET. Hillslope with high ΔP_1 have a deep WTD on average, this is because the WTD increases significantly during baseflow conditions and reaches very large values as quantified by ΔP_1. Hillslopes with high ΔP_1 values generally correspond to hillslopes with high precipitation and low temperature and therefore high SWE values.”

283-285: “...complementary information, with areas with high ΔP1 having low WTD because the strong changes in groundwater levels, as quantified by ΔP1, lead to a deep WTD.” Suggest rewording in a more straightforward way to improve clarity. Also, it is not immediately clear to me why “strong changes in groundwater levels” should result in deep water table. Why could two systems not have the same mean value with different standard deviations?

We have changed these sentences now to read (lines 319-321) “Hillslopes with high ΔP_1 have deep WTDs on average because the WTD increases significantly during baseflow conditions and reaches very large values as quantified by ΔP_1.”

Figure 4: Overall this is a nice figure, but a possible suggestion to improve readability would be to color-code the bottom left boxes with a red-white-blue color scale scaled to the strength and sign of the correlation to make it easier to digest at a glance. Also, some type of lettering/numbering scheme may make it easier to draw the reader to the correct part of the figure when discussing the figure in the text.

We have color-coded the Figure and added letters and numbers to identify each graph.

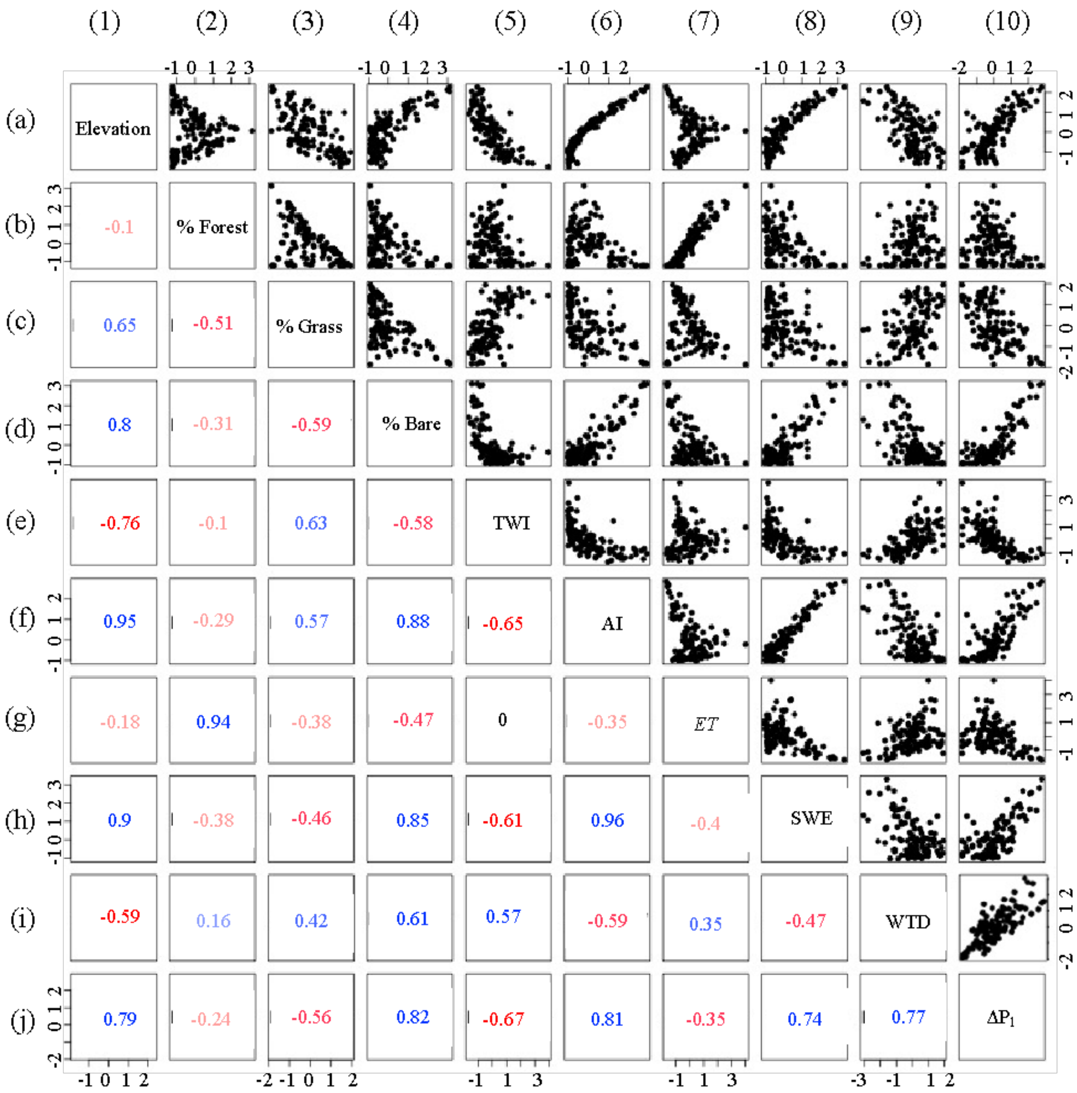

Figure 4: Pearson's correlations between the selected variables for hillslope clustering approaches: elevation, percent of the main land cover type (forest, grassland, and bare soil), topographic wetness index (TWI), aridity index (AI), evapotranspiration (ET), snow water equivalent (SWE), water table depth (WTD), and seasonal changes in groundwater ΔP1. Note that correlation coefficients are colored coded based on their values.

297-298: "...the two variables provide the same information." This is inconsistent with earlier in the text where ΔP1 and ΔP2 are defined (253-259) as containing different information.

We have removed this statement.

298-299: "TWI, AI, SWE, WTD, and ΔP1 are significantly correlated with elevation." What is the threshold for significance? The correlation between Elevation and % Bare is not included on this list (correlation coefficient = 0.8), while the correlation between Elevation and TWI is included (correlation coefficient = -0.76).

We have changed the sentence to avoid any confusion about the use of "significantly", see lines 333-335 and below:

"Results for ΔP2 are not shown because ΔP2 is strongly correlated to ΔP1. Bare soil, TWI, AI, SWE, and ΔP1 are strongly correlated (we define this as Pearson's correlation coefficient higher than 0.7) with elevation (Figure 4, column 1, lines d, e, f, h, and j)."

302-304: "A high correlation between the percent of forests and the elevation is found in the mid-elevation whereas grassland shows a high correlation in low and high elevations" I am unsure how to read the figure to interpret different correlations at different elevation ranges, as suggested by this text.

We have added ticks to the figure to make it consistent with the text, see below the new figure 4 in the previous answer.

304-305: The sentence feels repetitive, suggest rewording.

We have changed the sentence to (lines 339-341): "ET is positively correlated to the percent of forests (Pearson's correlation coefficient is higher than 0.9, Figure 4 2g)."

305-306: "ΔP1 is, in general, well correlated to all these variables" Does this suggest that a correlation of -0.24 and -0.35 indicates well-correlated variables? Suggest stating the metrics used for deciding if correlation is strong or not.

We clarified the metrics (correlation values we have used to determine a strong correlation). Now the sentences read (lines 333-349):
"Results for ΔP2 are not shown because ΔP2 is strongly correlated to ΔP1. Bare soil, TWI, AI, SWE, and ΔP1 are strongly correlated (we define this as Pearson's correlation coefficient higher than 0.7) with elevation (Figure 4, column 1, lines d, e, f, h, and j). In particular, elevation has a dominant control on AI and SWE with a Pearson's correlation coefficient higher than 0.9. We observe nonlinearity such that TWI increases in the lower elevation and that AI becomes constant at the lower elevation. High percentage of forests is only found in mid-elevation (Figure 4, 2a) whereas high percentage of grassland is well correlated to low elevations (Figure 4, 3a). ET is positively correlated to the percent of forests (Pearson's correlation coefficient is higher than 0.9, Figure 4 2g). ΔP1 has a Pearson's correlation coefficient higher than 0.7 for 6 out of 9 studied variables (elevation, percent of bare soil, TWI (correlation coefficient equal to 0.67), AI, SWE, and WTD, Figure 4, line j, columns 1, 4, 5, 6, 8 and 9); it, therefore, indicates that changes ΔP1 can reflect the changes of these variables. The two variables with low correlations with ΔP1 are ET and the percent of forests (Figure 4, line j, column 2 and 7). ET is related to groundwater dynamics in a nonlinear way (Condon et al., 2013; Ferguson & Maxwell, 2010; Rahman et al., 2016). As shown in these studies, regions with shallow WTDs have the highest ET fluxes and this flux typically decreases significantly with WTD. When WTD reaches a critical depth, the groundwater and the atmosphere disconnect and changes in WTD do not impact ET."

306-307: "...the selected variable contains valuable information about these variables." Suggest rewording to improve clarity.

We have removed this statement and changed the paragraph please refer to the previous answer.

312-314: "Regions with shallow WTDs have the highest ... changes in WTD do not impact ET." I am having trouble discerning the indicated relationships from Figure 4. Specifically, the exponential behavior and threshold are not clearly visible.

*These statements are related to the cited studies. We clarified the sentence which now states (lines 345-349) "*ET is related to groundwater dynamics in a nonlinear way (Condon et al., 2013; Ferguson & Maxwell, 2010; Rahman et al., 2016). As shown in these studies, regions with shallow WTDs have the highest ET fluxes and this flux typically decreases significantly with WTD. When WTD reaches a critical depth, the groundwater and the atmosphere disconnect and changes in WTD do not impact ET.*"*

325-316: "classifications...zones" I am slightly struggling with how 'classification' and 'zones' are being used here. What are identified as "zones" appear to me – as a reader – to be classes assigned to the underlying polygons. Perhaps it would be helpful to more explicitly define these terms.

We now refer to clustering as the method (ΔP, aridity index, etc.) we used to delineate the three zones. We have clarified it in the revised manuscript.

317-318: "...grouping was made based on the manual selection of natural grouping in the "probability density function." This is unclear. Perhaps the method could be elaborated on in the Methods section? The explanations between 318-366 are helpful, but they generally come across as arbitrary: for example, why are elevation cut-offs at 3000 and 3500 m used? Should this be based on some statistical property of the dataset? This applies to all categories except "clustering."

We defined these thresholds based on the distribution of the variables and by testing different thresholds. We have added it to the revised manuscript, please refer to the following sentence added to the revised manuscript (lines 352-354): "For the ΔP_1, elevation, TWI, and AI clustering approaches, we define the thresholds of each zone by analyzing the distributions of the hillslope values of these indices."

316-366: Seems like this could be in the "methods" section.

We have moved this paragraph to the methods section.

378-379: "...zones with the least variability..." It's unclear why this is 'an important metric that provides a degree of performance for the method's ability to delineate zones.' Also, this statement could benefit from a reference to support it.

We have removed this statement and explained in the new methods section the comparison procedure.

401-402: " the essence of that classification" Unclear what this means.

We have removed this statement.

402: "excellent index for identifying hillslopes with similar elevation" It's unclear why you would want to do this? why use these indirect observations when elevation is directly available?

We have clarified in the revised manuscript that identifying hillslopes with similar elevations is helpful because the hydroclimate depends on the elevation. Hillslope with similar elevations could be expected to have similar land surface processes. Because the elevation is directly available, one could cluster similar hillslopes based on this similarity in elevation.

408-411: I find it difficult to follow the logic here. Why is it desirable to 'distinguish zones of similar elevation?' How can similar results also indicate that they yielded the same results?

We have clarified in the revised manuscript that identifying hillslopes with similar elevations is helpful because the hydroclimate, which controls the changes in hydrological processes, depends on the elevation. Hillslope with similar elevations could be expected to have similar land surface processes and therefore similar subsurface processes.

412: "average percentage" – I'm not sure I follow: the table does not have any numbers expressed as percentages. I think they are possibly reported as fractional values, and just updating the numbers to percentage would make this clearer.

We have changed "average percentage" to "average hillslope fraction".

413-420: I am unsure how describing the contents of the table here is helping the reader to take away any particular conclusions. For example: "The selected classifications lead to similar conclusions, hillslopes associated with zone 1 have mainly grasses..." So in many (but not all) of the classifications, there is more than 50% grassland for zone 1... but what does this mean? what does this tell us about the classification or the hydrology? Furthermore, this statement is misleading because two of the classifications have grassland <40% for Zone 1. Similarly: "...zone 2 have mostly identical percentage of forest and grasses..." It's unclear what 'mostly identical' means since the numbers are not equal (i.e., they are not identical). The remainder of the paragraph describing Table 2 is similar – it's unclear how to interpret these results, or what they mean.

We have clarified the results presented in the table, please refer to the lines 407-414 and below:
"The land cover clustering indicates that grassland is the dominant land cover of zone 1, forests Zone 2, and bare soil zone 3. Only the machine learning clustering approaches using outputs lead to a similar conclusion whereas while the other clustering approaches capture the characteristics of zone 1 and 3, they do not identify a distinct forested zone 2. For the ΔP_1 clustering this could be attributed to the disconnection between groundwater dynamics and land surface processes that takes place in certain forested zones. Since clustering based on landscape characteristics and ΔP_1 do not identify such a distinct zone, it suggests that this zone may not be indicative of distinct hydrologic behavior."

429-430: "The classifications based on elevation and AI allows clearly distinguishing the hydroclimate associated with each zone" I am unable to interpret Figure 7 in such a way to understand how the information provided can 'clearly distinguish the hydroclimate.' I see the brief explanation provided in the following lines 430-432, however I still am unable to see how this information or interpretation is represented in the figure.

We have added the evidence that has helped us to make such a statement, the sentence is now (lines 420-421) "The AI clustering allows identifying hillslope with similar hydroclimate because it has low values of coefficients of variation."

440-441: "this type of classification mainly describes how a given hillslope release water based on its topographic structure" It is unclear what this means or how it is interpreted from the results presented.

To clarify the purpose of the TWI clustering approach we have changed the sentence to (lines 431-433) "TWI shows that clustering that includes only hydroclimate would miss important information on distinct hillslope hydrologic processes that strongly affect the response of the hillslope to meteorological forcing."

451-452: "A hillslope hydrologic function should aim to describe how a hillslope partitions, stores, retains, and releases water." Great – this is useful, however perhaps it could be presented in the Introduction to set up this concept for the manuscript. Also, it should be supported with references and specific definitions. What is the key parameter of interest for each of these process functions? timing? volumes? locations? all?

We have moved this statement to the introduction and methods section, added references and definitions as well as the most important aspects and/or parameters to quantify each function.

453: "...are simultaneously occurring..." occur simultaneously

Changed, thank you.

475-476: "As a result, the land cover based classification performs well at delineating hillslopes with similar ET rates (Figure 8b)" I'm just not sure how to interpret this from the figure.

This is because ET values mostly depend on the soil moisture and the land cover type. Soil moisture is also somewhat linked to the availability of water. We have clarified this statement in the revised manuscript now to read (lines 466-467) "The land cover clustering performs well at identifying hillslopes with similar ET because the latter strongly depends on the land cover (Figure 8b).".

478-479: "To some extent, the TWI and elevation classifications poorly distinguish hillslopes with similar ET." Why?

This is because TWI and elevation clustering approaches do not account for the processes that control ET which are soil moisture and the spatial distribution of land cover. We have clarified it in the revised manuscript, see below the new sentence (lines 469-471):
"The TWI and elevation clustering approaches do not separate hillslopes by ET because they do not account for varying land cover and soil properties that influence ET".

495: "regrouping" Unclear what it means for hillslopes to be 'regrouped' during classification.

We have changed the wording to clarify that the purpose of clustering (as now we refer to clustering instead of classification) is to identify hillslopes with similar characteristics.

496: "Because the TWI approach describes water transfer..." Based on the description of TWI provided on line 344, it is unclear why TWI would describe "water transfer" or what the definition of "water transfer" is.

We have clarified it by changing the sentence to (lines 483-484) "Because the TWI describes the characteristics that drive flow, it serves as a good indicator of soil saturation like the AI."
We have removed the term "water transfer" which could be misleading.

498-499: "The ΔP1 based classification has one of the lowest averages of CV..." This statement is misleading at best. Looking at the figure, the CV of ΔP1 is 0.12 – there are also two other classifications that achieve the same CV, one classification that achieves a *lower* CV, and the other three CV's equal 0.13, which appears to be only very slightly higher than 0.12. Perhaps this is "one of the lowest" however all CVs are very low and very similar, so it is unclear how 0.12 brings any significance to the argument (or that the low CV is due to the connection between GW and soil saturation as is claimed later in the sentence).

We have clarified it by changing the sentence to (lines 481-488): "As the land cover clustering adequately regroup hillslopes with similar ET, it also allows regrouping hillslopes with similar soil saturation. Because the TWI describes the characteristics that drive flow, it serves as a good indicator of soil saturation like the AI. Similar to the results above, the machine learning based clustering perform well. The ΔP_1 clustering has a low average CV due to the strong connection between the changes in WTD and soil saturation. It is only the elevation clustering that fails to identify hillslope with similar soil saturation, where the distributions of the three defined zones show overlap."

508: "Groundwater storage is mostly quantified in terms of WTD." Support with a reference?

*We have changed the sentence to (line 498) "*WTD is an important variable for determining groundwater storage.*"*
This is because groundwater storage is calculated using the specific storage, the porosity, and the WTD (Maxwell and Miller, 2005). Since the specific storage and the porosity are constant values, groundwater storages depend on WTD.

Maxwell, R. M., & Miller, N. L. (2005). Development of a Coupled Land Surface and Groundwater Model. Journal of Hydrometeorology, 6(3), 233–247. https://doi.org/10.1175/JHM422.1

512: "intermediary" Not sure this is the best word.

We have changed to (line 501): "Zone 2 exhibits a behavior that is in between those of Zone 1 and 3."

604-605: "...transcending the uniqueness of place inherent in traditional classifications..." Unclear, suggest rewording with more direct language.

We have changed the sentence to (lines 597-598): "ΔP1 is an important variable controlled by many hydrologic processes including land surface processes and hydroclimatic."

A. Parsekian

Referee #2

I think this is interested study that is worthy of publication. However, a lot of improvements have to be made for its current form. I agree with that ground water dynamics may be a good proxy of surface hydrological processes in some places. However, this may not be the truth in some other areas. So this proposed approach may have its limits. This have to be clarified in the introduction and discussion.

We thank the reviewer for their comments and feedback. Based on the feedback, we have made significant changes to the manuscript. Please see below for our detailed responses to the comments. The reviewer's original comments are in black*, our responses are in blue italic fonts, and modifications made to the original manuscript in* blue plain text.

We have clarified the limitations of using groundwater dynamics as they may not be suitable in other regions with perched systems for example. We have added the following statement to the introduction (lines 102-105) "Nevertheless, we acknowledge that groundwater dynamics in some regions such as arid areas could be disconnected to land surface processes and less dependent on many key physical features of the hillslope, which may impede the ability of the proposed clustering in these regions."

There are a lot of indices and methods used in this study were not presented in the Methodology section. I also feel that the methods section did not clearly present how the authors process the data and generate the results.

The revised manuscript has a "methods" section with the definition of all the indices as well as all the methods and data processing methods we used.

In addition, and especially, discussion of the findings of this study has to be strengthen. currently, the discussion is weak, maybe due to the reason that the results and discussion were combined. References are needed for the interpretations. Explanation of the results and comparions with other published studies have to be improved.

The revised manuscript has a discussion section in which we discussed and compared the findings of this study to previous works notably (Wainwright et al., 2022).

1. the abstract lacks quantified description

We have added some statistics to the abstract. We, specifically, added in lines 22-25:
"The ΔP clustering performs very well in identifying hillslopes with 6 out of the 9 characteristics studied. The variability among clusters as quantified by the coefficient of variation (0.2) is less in the ΔP and the machine learning approaches than in the others (>0.3 for TWI, elevation, and land cover)."

2. line 41-44, references are needed to support this statement.

We have added the following references: (McDonnell & Woods, 2004).

3. line 106, maybe give some examples of such models

We have named these models, the sentence is now in lines 96-102 "These models (e.g., HydroGeoSphere (Brunner and Simmons, 2012), ParFlow (Maxwell & Miller, 2005), Advanced Terrestrial Simulator, (Coon et al., 2016)) account for the two-way interactions between groundwater and land surface processes and can be constrained with ground observations and measurements at ultra-high resolutions through aerial or remote sensing (i.e., drones, planes, or satellites)".

4. line 138, maybe starting with a sentence to tell the reads the pupose or the reason of using ParFlow-CLM in this study

We have added the reason for using ParFlow-CLM in this study. Please refer to the sentence below added to lines 129-133: "We use the integrated hydrologic model, ParFlow, which has the advantages of simulating the water and energy balance from the bedrock to the lower atmosphere and therefore connect groundwater dynamics with land surface processes."

5. line 164, provide the examples of the application of ParFlow-CLM

We have now added examples of ParFlow-CLM applications in the following sentence in lines 172-175. "ParFlow-CLM has been used in many studies to understand the interactions between groundwater dynamics and lower atmosphere (Maina et al., 2022; Maina and Siirila-Woodburn, 2020) at different scales from the watershed (Foster and Maxwell, 2019; Maina et al., 2020) to the continental scale (Maxwell and Condon, 2016)."

6. line 239-240, what were those thresholds tested, specify

We have tested these thresholds, we have now clarified it in the revised manuscript by adding the following sentence in lines 352-354: "For the ΔP_1, elevation, TWI, and AI clustering approaches, we define the thresholds of each zone by analyzing the distributions of the hillslope values of these indices."

7. line 238-243, what were the thredholds of drainage area you finally used?

We used a threshold for drainage area equal to 810,000 m^2. We tested different thresholds to select this one, more details could be found in Wainwright et al (2022). We have added this information to the revised manuscript.

Wainwright, H. M., Uhlemann, S., Franklin, M., Falco, N., Bouskill, N. J., Newcomer, M. E., ... & Hubbard, S. S. (2022). Watershed zonation through hillslope clustering for tractably quantifying above- and below-ground watershed heterogeneity and functions. Hydrology and Earth System Sciences, 26(2), 429-444.

8. line 316, the clustering approaches have to be introduced in the methods section

We have now described the clustering approaches in the methods sections.

9. line 369-379, most of the part would be better to move to the methods section

We have moved this paragraph to the methods section

10. section 3.2.3, why were surface runoff not considered? I thinks it might be one of the most important hydrological processes.

We have not added surface runoff because not all hillslopes have a quantifiable surface runoff. For hillslopes where surface runoff is quantifiable, there is a connection between surface runoff and subsurface flow as a result, subsurface flow captures that behavior (see the comparisons between measured and simulated groundwater levels added to the Appendix A). In the hillslope without runoff located uphill, the changes in snow and evapotranspiration capture the land surface processes. If all the hillslopes had a quantifiable runoff we would have characterized the temporal variations of the runoff and use it as a metric to perform clustering.