Dear Reviewer 2

Thank you for your thoughtful feedback on our manuscript. We sincerely appreciate the time and effort that you have dedicated to the review process, and your support to this work. Please see below, in **blue**, the original review comments, and in **black**, our responses to the comments and how we will implement the changes in the revised version.

Araki et al. investigate hydrologic processes (signatures) and their drivers in 14,146 watersheds across the contiguous US. The study is a highly relevant contribution to an advanced understanding of hydrologic processes at the US continental scale. While the study is well designed and executed overall, the authors may want to address the following points before final publication.

Major comments:

In the results section 4.2, the short description of the corresponding region in the relevant subsections is useful. However, many parts of section 4.2 (and partly 4.3) include interpretations ("suggest…", "indicate…") and references that would rather fit into the discussion (e.g., section 5.1). A clear separation of results and discussion would make it easier for the reader to follow the main points presented.

Thank you for your valuable feedback. To allow Sections 4.2 and 4.3 to focus on the observed signature patterns, we will move interpretive content from these sections to the Discussion (Section 5.1). In particular, the East and South subsections (Sections 4.2.1 and 4.3.1) contained several interpretive passages, which will be relocated to Section 5.1 as follows:

"By mapping and categorizing the primary drivers of runoff processes, we can untangle which physical characteristics drive the hydrologic response in each region. In the East and South, soil, geology, and topography emerged as primary drivers, which consistent with regional hydrologic processes: in the Piedmont, wide and wet valley bottoms generate fast responses (Zimmer and Gannon, 2018); clay-rich soils along the Gulf Coast promote infiltration-excess overland flow (Miller, 1999; Fig. S6), producing mixed storage and water balance signatures despite deep bedrock (Fig. S5) and semi-consolidated sand aquifers; and Eastern Coastal Plain sandy soils, seasonal flooding, and wetland likely support a single dominant groundwater reservoir supplying baseflow (Fig. 3b; Holt and McMillan, 2025; Hupp, 2000). The machine learning approach is especially powerful for this purpose, as multiple landscape attributes often contribute simultaneously to the hydrologic response."

"In four of the six regions, soil texture, particularly silt or clay fraction, was identified as a recurring primary driver (Fig. 6b), though their roles differ by context. In the Northeast, silt dominates variable importance; silt is found in glacial till layer and supports high water storage and baseflow (Shanley et al., 2015) while facilitating subsurface stormflow under wet conditions (Detty and McGuire, 2010). In the South, despite silt being identified as a primary driver, clay is the dominant soil texture (Miller and White, 1998); in the Mississippi embayment, extensive confining units of clay and silt separate aquifers and control the groundwater flow (Renken, 1998; Clark et al., 2011). These two cases suggest that Shapley or permutation-based methods

may not fully separate correlated variables due to their treatment of joint variable distributions, high clay content may be implicitly captured through the absence of silt in regional analyses."

For the Region 2 (Midwest&Central) and Region 3 (West & Southwest) analysis in Sections 4.2 and 4.3, spatial patterns of some signatures and identified drivers vary with regional geography, and previous literature is integral to describing these patterns. We have retained language that makes clear these are observations rather than interpretations.

Lastly, to better reflect that Section 4.3 presents inferred drivers and their interpretation from the permutation analysis, we will rename it to "4.3 **Inferred** Climate and Landscape Drivers of Hydrologic Processes."

Minor comments:

Line 13: In general, "hydrologic behavior" always sounds quite humanized to me. Maybe change to "hydrologic functioning" or similar (here and at other relevant places in the manuscript).
Thank you for pointing this out. We will replace "behavior" with "function."

Line 21: I am not sure whether the study can be referred to as a "framework". The authors may want to rewrite this sentence and leave out the "framework" part.
Thank you for your suggestion. We will rewrite the sentence as *"Our approach of estimating dominant processes and their drivers facilitates extending process knowledge from research watersheds to the continental scale, assessing current hydrological understanding, and evaluating hydrological model structures."*

Lines 38-49: This paragraph is supposed to focus on hydrologic processes, but it describes more of the method or approaches ("likelihood of…", "statistics can help…"). It would be interesting to read more about the actual findings of the studies investigating these processes across the CONUS (even if not at a continental scale).
We will revise the sentence as follows to incorporate key findings from the papers:
"For example, Buchanan et al., 2018 assessed the likelihood of infiltration excess flow occurrence by comparing whether rainfall intensity exceeds saturated hydraulic conductivity, ***finding that saturation excess dominates across the contiguous U.S., while infiltration excess is regionally likely in the central U.S.*** *(Buchanan et al., 2018); Similarly, studies on baseflow indices have shown their* ***strong dependence on climatic and soil properties*** *(Beck et al., 2013; Xie et al., 2024), and Fang and Shen (2017) quantified the strength of runoff-storage connectivity through correlations between anomalies in streamflow gauge and satellite water storage observations****, highlighting large-scale interactions among groundwater table, soil thickness, topography, and snow****"*

Furthermore, we discuss these papers in the results and discussion section. For example, Buchanan et al. (2018), Beck et al. (2013), and Xie et al. (2024) are the main references used to compare and discuss the regional patterns of hydrologic processes identified in this study; and Barnhart et al. (2022) is used in Section 4.2.3 to explain snowpack contribution to sustaining baseflow.

Lines 93-94: Many large-sample studies are conducted across large climatic gradients, which may "override" landscape attributes and thus lead to comparably weak predictive power of the latter?

Yes, we agree and believe that it is one of the major reason reason, and it is currently implicitly stated as *"Regional analysis can mitigate climate influence and elucidate the contribution of non-climatic drivers, such as regional random forest models that revealed physiographic and anthropogenic controls on flow regimes (Almagro et al., 2024; Hammond et al., 2021). However, smaller regional sample sizes may limit prediction accuracy if datasets only provide tens of watersheds per region (Willard et al., 2024)."* However we agree that it was not clearly spelled out to the reader and we will add a sentence directly addressing this point before this passage, as follows:

*"**Additionally, large-sample studies across broad climatic gradients may be obscuring the influences of landscape attributes.** Regional analysis can mitigate this effect and elucidate the non-climatic drivers; for example, regional random forest models have revealed physiographic and anthropogenic controls on flow regimes (Almagro et al., 2024; Hammond et al., 2021). However, smaller regional sample sizes may limit prediction accuracy if datasets only provide tens of watersheds per region (Willard et al., 2024)."*

Line 99: Consider adding the study of do Nascimento et al. (2025), dealing with geological maps in large-sample studies.

Thank you for your suggestion, this study perfectly fits the context and we'll add the reference.

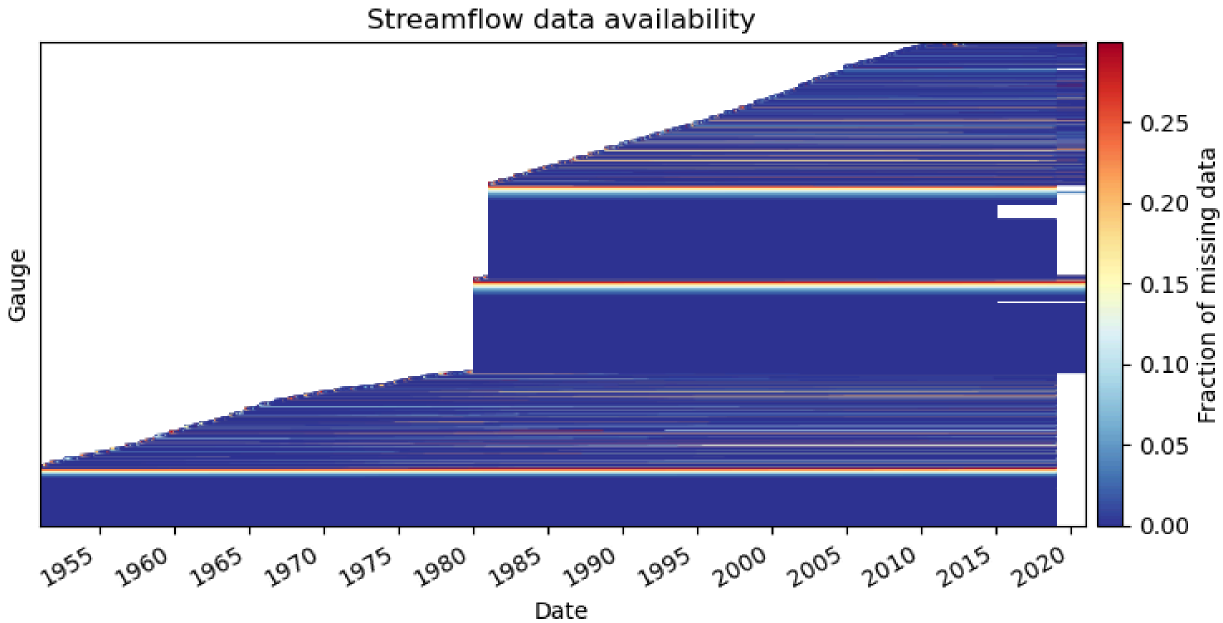
Lines 168-169: Any ideas on why the RF models yielded lower performance on all Caravan watersheds compared to other subset combinations? This may be something to add to the Text S2.

Thank you for your comment. We have several hypotheses: (a) signatures from GAGES-II + Caravan watersheds may be easier to predict, possibly because (a-1) streamflow data are of higher quality, leading to more accurate signature extraction, (a-2) signature-attribute relationship are stronger (e.g., these watersheds are less impacted by human influences). We will add a sentence in Text S2 to address this point.

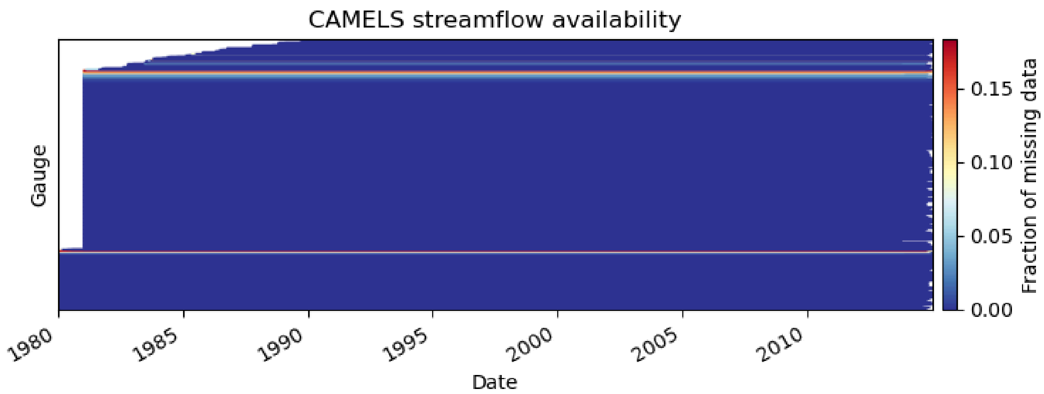
Line 172: Looking at the distribution of streamflow observation length in Caravan, a minimum of 5 years seems to be quite short. Why did the authors choose this length? A quick visualization of the observation length distribution of all data sources used (e.g., added to S1) would be helpful.

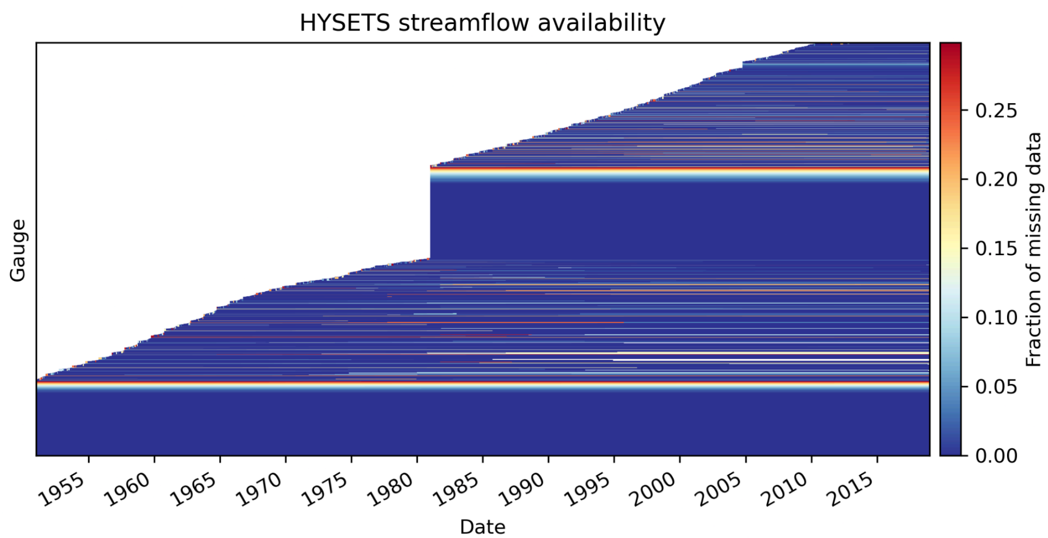
Thank you for your suggestion. We will add the plot displaying the length distribution of all data sources to S1. Given the large number of gauges, each gauge is represented as a horizontal bar spanning its available record period (from start to end date), colored by the fraction of missing data.

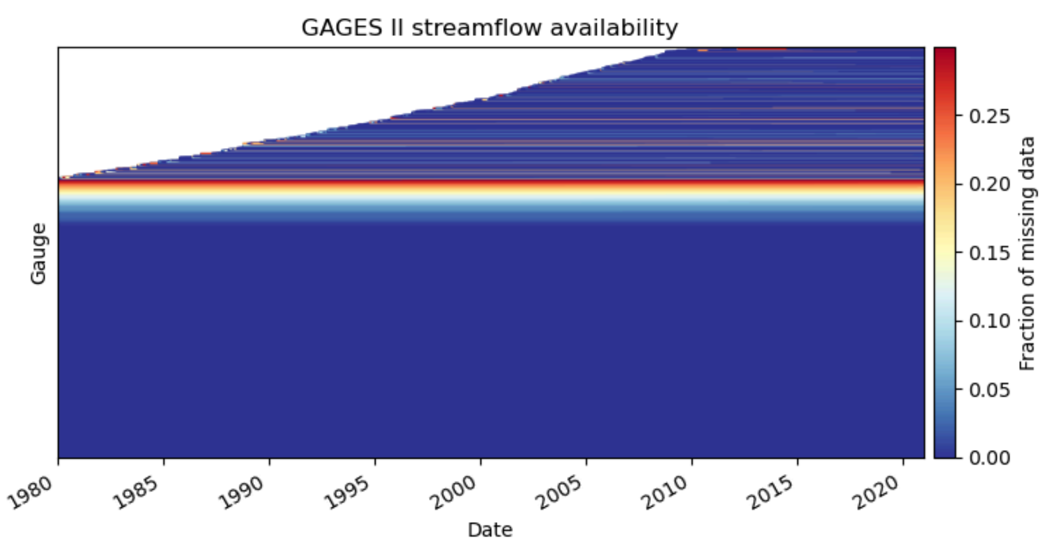
Figure. Record availability and data completeness for all gauge stations. Each horizontal bar represents the available record period of an individual gauge, spanning from its start to end date. Bar color indicates the fraction of missing data within the record period.



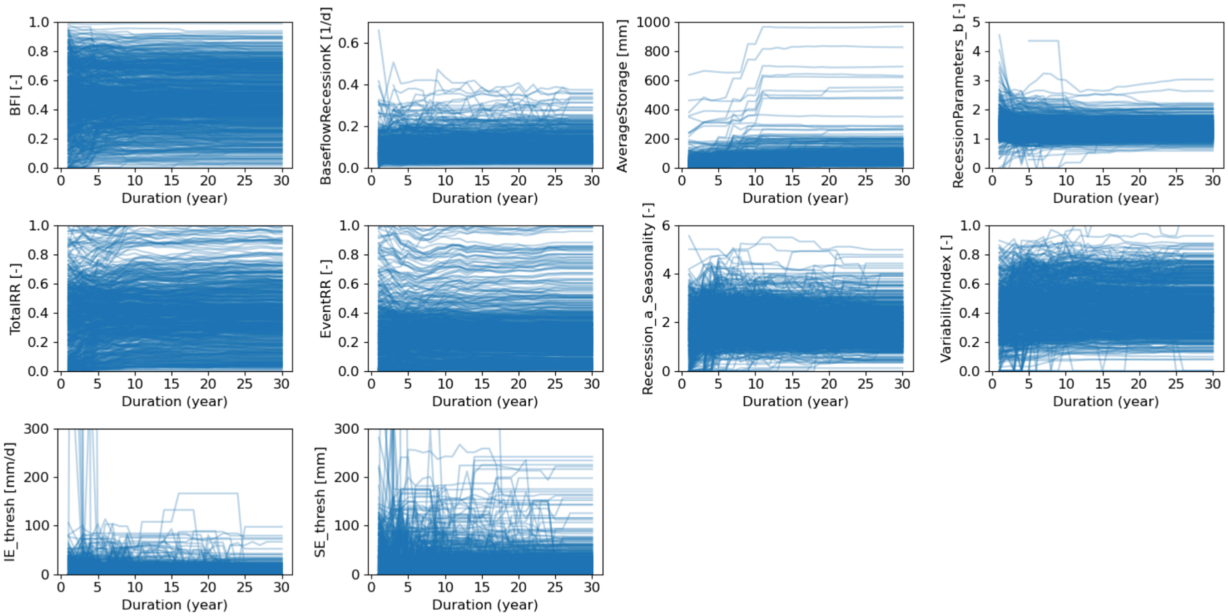
Here are the same plot from each sources for reference:







On the point of 5-years data availability threshold, studies suggest that temporal hydrologic variability is adequately captured with 3 to 5 years of data (Refsgaard and Storm, 1996; KlemeŠ, 1986; Merz et al., 2009). Further, we conducted preliminary experiments to confirm that the signatures stabilize with 5 years of observations. In the figure below, we calculated signatures with varying time lengths (the x-axis is the length of the record in years, starting from 1990; the y-axis is signature values; each line represents a CAMELS-US watershed). As demonstrated in the plot, the signature values tended to stabilize around 5 years.



Reference:

Refsgaard, J. C. and Storm, B.: Construction, calibration and validation of hydrological models. In Distributed hydrological modelling, in: Distributed Hydrological Modeling, edited by: Abbott, M. B. and Refsgaard, J. C., Kluwer Academic, Netherlands, 41–54, 1996.

Merz, R., Parajka, J., & Blöschl, G. (2009). Scale effects in conceptual hydrological modeling. Water Resources Research, 45(9). https://doi.org/10.1029/2009WR007872

KlemeŠ, V.: Operational testing of hydrological simulation models, Hydrol. Sci. J., 31, 13–24, https://doi.org/10.1080/02626668609491024, 1986.

Lines 168-179: The training sample is derived based on various (subjective?) quality criteria. How sensitive are the predicted hydrologic signatures to changes in the quality criteria? A short explanation in the discussion would be beneficial.

Thank you for your questions. The following are the rationales for the quality criteria, and we will add these to the revised manuscript.

- **Rationale for 5 years of continuous data**: Please see the previous response.

- **Rationale for 30 % missing data:** The signatures need at least 3 years of data to stabilize their values (see the previous response for reference); therefore, a minimum of 5 years of data × 30% ≈ 3 years.
- **Rationale for watersheds boundary area discrepancies 25% between GAGESII and Caravan:** We determined this threshold based on the distribution of area error values. Errors of <20% are possible due to differences in the watershed delineation tools or missing small tributaries (Ray, 2018). These areas are already used in streamflow (Q) normalization and landscape attribute subsetting through GAGES-II or Caravan, for which we do not have the ability to conduct a sensitivity analysis. Most of the signatures characterize the dynamics and relationships rather than the absolute values of the streamflow dataset. Signatures that use absolute values and could potentially be impacted by streamflow normalization by watershed areas are: *TotalRR* (uses the Q/P relationship), *EventRR* (uses the Q/P relationship), and *AverageStorage* (uses the Q = P − ET relationship).
- **Rationale for Snow-dominated watersheds with >20% snow fraction:** We determined this threshold based on the spatial distribution of snow-affected catchments, and the 20% threshold effectively identified snowy regions in the contiguous United States. A 30% threshold was used in a similar analysis (McMillan et al., 2022; Wu et al., 2021), and our threshold is lower (i.e., more strict).

Ray, L. K. (2018). Limitation of automatic watershed delineation tools in coastal region. Annals of GIS, 24(4), 261–274. https://doi.org/10.1080/19475683.2018.1526212

McMillan, H. K., Gnann, S. J., & Araki, R. (2022). Large scale evaluation of relationships between hydrologic signatures and processes. Water Resources Research, 58(6). https://doi.org/10.1029/2021wr031751

Wu, S., Zhao, J., Wang, H., & Sivapalan, M. (2021). Regional patterns and physical controls of streamflow generation across the conterminous United States. Water Resources Research, 57(6), e2020WR028086. https://doi.org/10.1029/2020wr028086

In the manuscript, we will describe it as:

"We filtered out watersheds from our signature calculations based on quality criteria for watershed area and snow used by previous studies, and on the timeseries length needed for signatures to stabilize. First, we removed watersheds from our analysis with uncertain topographic boundaries, showing high discrepancies (>25%) in the estimated drainage area between GAGES-II and Caravan datasets. Errors of <20% are possible due to differences in watershed delineation tools or missing small tributaries (Ray, 2018). Second, for overland flow signature analysis, we excluded snow-dominated watersheds (>20% snow fraction of total precipitation; a >30% criterion were used in McMillan et al., 2022 and Wu et al., 2021); this is because our overland flow signatures can be heavily influenced by periods with no flow response due to snow or frozen conditions. Third, we excluded watersheds with less than 5 years of streamflow observation record, and those with over 30% missing daily data over the period where streamflow was recorded (yielding at least three years of available data). Studies suggest

that temporal hydrologic variability is adequately captured with 3 to 5 years of data (Refsgaard and Storm, 1996; KlemeŠ, 1986; Merz et al., 2009).”

Furthermore, the uncertainty about watershed boundaries will be discussed in Section 5.3. Limitations and future work as:

Additionally, errors in watershed boundary delineation would affect signatures that use drainage area to normalize flow, such as runoff ratio (TotalRR, EventRR) and water balance (AverageStorage).

Lines 195-204: This rather long paragraph seems to be unnecessary, as Table 1 already provides a good overview.
Thank you for your suggestion. This paragraph was shortened as the following and combined with previous paragraph: *“The process hypotheses tested are described in Table 1 and cover six major hydrologic processes: baseflow, watershed storage, water balance, seasonal variability, overland flow dominance, and its type.”.*

Line 202: Please indicate the meaning of “small p-values” in this context.
Thank you for your comment. This “small” meant “p-values < 0.05”, but the phrase will be omitted in response to your comments on L195–204. The new L195–204 will briefly describe the processes tested, and guide the reader to Table 1, which explains the p-values as : *“p-values outside the range $0 \leq P\text{-value} \leq 0.05$ are deemed insignificant and clipped out. Within the range, the smaller P-value is, the more significant the threshold is.”*

Line 258: Nice figure!
Thanks for the compliment!

Line 273: Introductory sentence may be deleted.
Thank you, for your suggestion, deleted.

Line 317: Please avoid using “very” as a filling word. If statistically significant, consider using “significantly” instead.
Rephrased as *“no significant evidence of overland flow”.*

Lines 389-393: Clay and silt fraction are often collinear. Can the authors discuss mechanistic reasons (e.g., surface crusting, …) for the silt fraction (and less so clay) emerging as a dominant driver of signatures in the Northeast and South? Maybe add a few sentences to the discussion (lines 465-466).
Per text S1, clay and silt fractions are selected among three soil texture variables (clay/silt/sand fractions) as follows: *“There is a strong correlation ($|\rho|>0.8$) between sand fraction (snd_pc_sav) and clay fraction (cly_pc_sav), as well as between sand fraction (snd_pc_sav) and silt fraction (slt_pc_sav). Picked clay fraction (cly_pc_sav) and silt fraction (slt_pc_sav) as a representative of a soil texture, as information on clay and silt content constrains sand content and is more relevant to water retention capacity.”*

We will further discuss possible mechanisms why silt/faction may appear as stronger predictor. For the Northeast, we will add the following sentence: *"Silt is the most dominant soil texture in the Northeast (Miller and White, 1998), which consists of a glacial till layer, promoting high water storage and baseflow (Shanley et al., 2015) while facilitating subsurface stormflow under wet conditions (Detty and McGuire, 2010)."*

In the South, our manuscript currently states that *"Along the Gulf coast, either climate or soils and geology may dominate. In this region, although depth to bedrock is high (Fig. S5), and the area overlies semi-consolidated sand aquifers, the clay-rich soils are capable of generating infiltration excess flow (Miller, 1999; Fig. S6)."* We will add: *"In the Mississippi embayment, the extensive confining units consisting of clay and silt separate the aquifers and control the groundwater flow (Renken, 1998; Clark et al., 2011)."*

We will further discuss in the Discussion Section 5.1 as: *"In four of the six regions, soil texture, particularly silt or clay fraction, was identified as a recurring primary driver (Fig. 6b), though their roles differ by context. In the Northeast, silt dominates variable importance; silt is found in glacial till layer and supports high water storage and baseflow (Shanley et al., 2015) while facilitating subsurface stormflow under wet conditions (Detty and McGuire, 2010). In the South, despite silt being identified as a primary driver, clay is the dominant soil texture in many areas (Miller and White, 1998); in the Mississippi embayment, extensive confining units of clay and silt separate aquifers and control the groundwater flow (Renken, 1998; Clark et al., 2011). These two cases suggest that Shapley or permutation-based methods may not fully separate correlated variables due to their treatment of joint variable distributions, and high clay content may be implicitly captured through the absence of silt in regional analyses."*

Lines 426-434: In my view, this summary paragraph is not relevant to the discussion and may be omitted or moved to the last section of the manuscript.
Thank you for pointing this out. We will shorten the paragraph, move some sentences to the conclusions, and revise it so that it serves as the opening of the discussion section:

"This study creates comprehensive maps of hydrologic processes across the contiguous United States by using machine learning to analyze streamflow signatures and connecting these signatures to dominant watershed processes. The analysis from over 10,000 watersheds shows distinct regional patterns in estimated hydrologic processes and its potential drivers. These process maps provide novel information for selecting appropriate hydrologic models across large domains and help hydrologists anticipate how watersheds will respond to environmental changes such as altered climate or land use. In the following sections, we discuss how these maps provide new benchmarks (Section 5.1), inform hydrologic modelling (Section 5.2), and outline directions for future work (Section 5.3)."

Line 445: Up until this line, the section 5.1 is not about "new process understanding" but rather about the sample size itself. The authors may want to consider renaming this section.
Thank you for your suggestion. We will rename the Section 5.1 title to "New benchmark maps of process understanding over large domains" so that it reflects the purpose of discussing sample size in this paragraph.

Line 452: Consider adding the study of Stein et al. (2021) to this section, dealing with flood generation processes across the CONUS.
Thank you for the suggestion, we have integrated the paper into the passage:

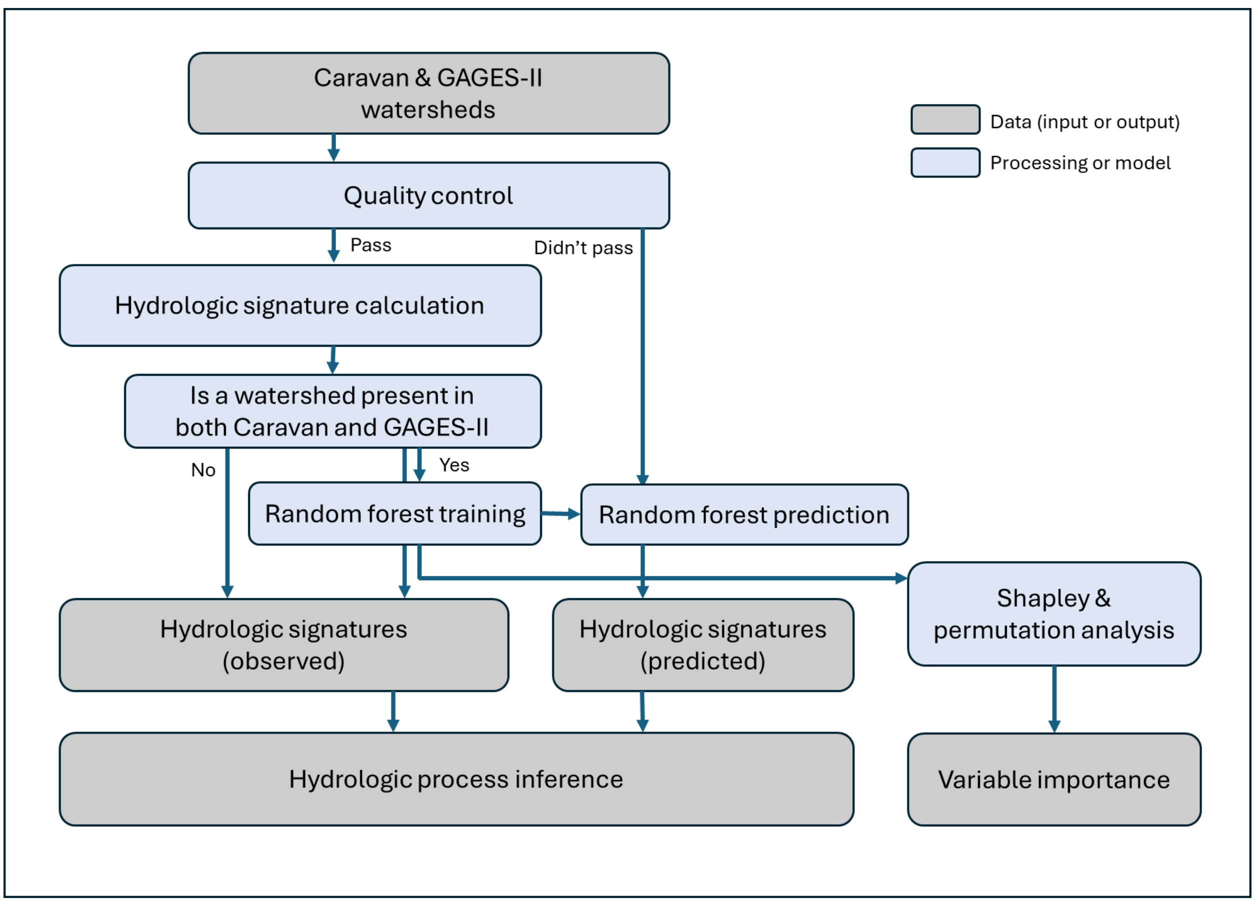
"This may reflect vegetation shifting the inferred overland flow mechanism toward saturation excess. Infiltration excess is inferred when overland flow is related to storm intensity rather than storm size. In arid and semi-arid catchments, vegetation can locally increase infiltration capacity and soil water retention, reducing the extent of infiltration excess overland flow (Stein et al., 2021). Additionally, where smaller storms are intercepted by canopies, signatures may incorrectly attribute the runoff to saturation excess rather than infiltration excess."

Line 538: Please renumber to "6 Conclusion".
Thanks for pointing it out, renumbered.

Technical points:

As the Method section comprises various steps, the authors may want to consider creating a simple flowchart containing the major working steps.

Thank you for your suggestion. We will add a flowchart to Text S1 as following:



References

do Nascimento, T. V., Rudlang, J., Gnann, S., Seibert, J., Hrachowitz, M., & Fenicia, F. (2025). How do geological map details influence the identification of geology-streamflow relationships in large-sample hydrology studies? Hydrol. Earth Syst. Sci., 29(24), 7173-7200. https://doi.org/10.5194/hess-29-7173-2025

Stein, L., Clark, M. P., Knoben, W. J., Pianosi, F., & Woods, R. A. (2021). How do climate and catchment attributes influence flood generating processes? A large‐sample study for 671 catchments across the contiguous USA. Water Resources Research, 57(4), e2020WR028300. https://doi.org/10.1029/2020WR028300