Reviewer 1 comments

I'm generally satisfied with the responses provided by authors.

Major comment #1: Clearly present hypotheses/predictions underlying the study
Authors now state upfront that they are performing exploratory analyses and made substantial changes to Table 3 to make this clear. Authors also added a map in Supplementary Material to help understand results with regional knowledge.

Major comment #2 : Conclusions need to be better supported by analyses and results.
Authors improved in numerous places the description of results which are now described with more precision. Removing the linear fit in Table 3 also solved some of my issues with the presentation of results.

Major comment #3: Better consideration of interannual variability in thermal sensitivity
Authors did a really good job of addressing interannual variability with additional analyses presented in Appendix A.

We are grateful that the reviewer is satisfied with our revisions.

Reviewer 3 comments

The manuscript by McGill et al. examines spatiotemporal variability of stream thermal sensitivities for two watersheds in Washington, USA. They collected water and air temperature data from 73 sites distributed across the Snoqualmie and Wenatchee basins. Most loggers ran for seven years. The data were used in statistical models to estimate thermal sensitivities (the slope coefficient of air temperature in a linear regression model). Seasonal models were applied, as well as time-varying coefficient models. Clustering analysis was performed to group sites that shared similar thermal sensitivities. These clusters were then used to explore how thermal sensitivity varied with climate and landscape variables. They argue that thermal sensitivity showed strongest relationships with elevation, snow water equivalent, and variables representing groundwater influence. Some variables which were expected to be related to thermal sensitivity, such as percent riparian forest cover, did not vary in a systematic way. Some key conclusions made by the authors are: (1) it is essential to acknowledge the non-stationarity of the relationship between air and water temperature, (2) snow and geological characteristics shape the relationships between air and water temperatures at the study site, and (3) classifying rivers based on thermal sensitivity is a powerful tool when planning for global change.

This is my first review of this manuscript (I did not review the original submission). Overall, the manuscript is generally well written and covers a topic suitable for HESS. My general feeling is that there is considerable amount of unexplained variability in the thermal sensitivity estimates. Reporting these sorts of noisy findings can be useful, but some of the key conclusions seem more inconclusive than how they are stated. I share a couple key comments, followed by some specific feedback. I reviewed the first round of reviewer comments and replies, and I echo some of those comments and feel as though the current version could still be improved with regards to structure and results/discussion.

1) Improving structure and flow of the manuscript

As the other reviewers pointed out, the presentation of the study would be much improved if structured around some key hypotheses. The authors replied that this is an exploratory study and that they want to avoid the use of null hypothesis significance testing. I am sympathetic to those concerns (I'm glad that a p-value is nowhere in sight), but the authors could focus on framing the study around scientific (distinct from statistical) hypotheses informed by the extensive literature on spatiotemporal variability of river temperature. Even exploratory studies will have hypotheses driving the direction of the exploration. The content for doing this is already more or less in the manuscript; however, the organization is challenging to follow. For example, some hypothesized drivers are listed in the hybrid Table 3, but this table isn't referenced until page 10. In addition, there is considerable geological context provided in the discussion that could be introduced earlier so that it doesn't feel so unexpected. I suggest using a paragraph or two at the end of the introduction to better frame the study and expected outcomes from the analyses.

We attempted to clarify the structure of the manuscript, and the driving hypotheses of our study, in several ways. First, we substantially restructured the introduction. We removed a paragraph detailing differences between statistical and process-based approaches, and we now devote the second paragraph of the introduction to laying out the hypothesized impact of climate, landscape, and hydrogeologic features we examine within our paper. Table 3 (now Table 1) is referenced in the second paragraph of the introduction to better set up our scientific objectives. Additionally, we have included an introduction to the geology of the basins in the methods section, L101-106, to familiarize readers with the geologic context of the basins much earlier.

2) Challenges in linking results to underlying controls

Much of the discussion tries to link the patterns in thermal sensitivities to underlying process controls. This is difficult since the study is exploratory, focuses on correlations, and uses statistical abstracts that can be a few steps removed from the actual observational data. For example, although the coefficient estimate associated with the air temperature term in the regression models gives you an idea of how water temperature co-varies with air temperature, you lose some information about the thermal regime at that site. Since there isn't a systematic relationship between air temperature and thermal sensitivity (Figure 3a), comparing thermal sensitivities can't tell you whether a particular stream is colder or warmer than another stream during the summer (for example). However, this can be important information for diagnosing key controls on stream thermal regimes (e.g., we might expect a colder stream, in summer, to have more groundwater influence, for example). This challenge is further compounded in this study because the thermal sensitivities are then used within clustering analyses and regression trees. The authors are familiar with these sites, and which streams have colder vs warmer or stable vs dynamic thermal regimes (or what the dominant geology of the site is), but as a reader, I find it difficult to follow these connections. Understanding these connections is crucial for interpreting how the results support the conclusions of this study. I challenge the authors to rethink how this information is presented. I provide some suggestions below, but it might be helpful to show more of the actual stream temperature time series for the individual sites (maybe in the supporting information) and referring back to those data when making interpretations.

We agree that coupling thermal sensitivity with water temperature can be a useful tool to diagnose controls on stream thermal regimes, however, this effort was outside the scope of our study. The goals of this specific study were threefold: determine the spatial and temporal distribution of commonly used air-water temperature metrics across each basin, quantify the representative thermal-sensitivity regimes and determine how clusters of similar sites differ from clusters based solely on air and water temperature, and determine the landscape or climate factors that best predict thermal sensitivity cluster membership. Future work could certainly include a more thorough examination of water temperature regimes in conjunction with thermal sensitivity. Many studies exist in the Snoqualmie and Wenatchee basins looking at facets of water temperature across the year, and our thermal sensitivity insights could be coupled with empirical observations of stream temperature (Steel et al. 2016) or process-based simulations (Cristea and Burges 2010, Yan et al. 2021) as an avenue of future research. Additionally, previous studies have used bivariate clustering of the slope (thermal sensitivity) and intercept of time varying regressions (Li et al. 2016).

We do agree that air and water temperature can provide context for interpreting thermal sensitivities, and all air and water temperature data, and thermal sensitivity estimates are available to view in detail within the associated RShiny application: https://lmcgill.shinyapps.io/TimeVarying_AWC/. Users can select a basin and site from a drop-down menu to simultaneously visualize daily air and water temperature and estimates for thermal sensitivity. We encourage the reviewer to explore this feature and now reference it in the text on L186, in addition to the Data Availability statement, in the event that other readers share similar concerns. Furthermore, annual average time series for water and air temperature for every site are visualized in Figures 4 and 5.

Partly related to the above, in my own experience I have found that there can be considerable uncertainty in the estimate of the slope coefficient/thermal sensitivity. What kind of uncertainties were associated with these estimates for this study? Could these be added as uncertainty intervals to the figures?

We agree that as empirical thermal sensitivity is a statistical relationship between two time series, there are many ways to calculate its value! We would argue that every modeling framework inherently requires a series of decisions that introduce uncertainty, ranging from data selection to parameter choices, which will collectively shape the model's structure and outcomes. We attempted to address these issues through several avenues within the manuscript. First, we showcase two alternative methods in our manuscript by comparing thermal sensitivities calculated through both standard linear regression (e.g., summary metrics) and through varying coefficient linear models. Second, we attempted to clearly articulate our reasoning behind data selection in L495-513 and the uncertainties of each modeling approach in L514-532. For example, we acknowledge that selection of the bandwidth parameter for time-varying coefficient models and the averaging period (e.g., monthly, seasonal, annual) for summary metrics will impact the final calculation of thermal sensitivity. In our RShiny application we allow users to explore various bandwidth parameters to examine how shifting this value up or down impacts time varying thermal sensitivity estimates. We additionally included leave-one-out-cross-validation analyses to assess the impacts of both interannual variability and the sensitivity of relative importance estimates within the CART analysis. These sensitivity analyses are referenced within the text and results can be found in the Supplementary Material (Appendix A, Figure S7). Lastly, the uncertainties for the summary metrics can be

conceptualized by the R^2 value, which is an indicator of how well water temperature can be approximated by air temperature. Summaries of this parameter are shown in Table 2.

Finally, there does not appear to be any assessment of the performance of the CART model. Many of the key conclusions rely on the results of this modelling; therefore, showing the overall performance of these models seems important. These CART models have a tendency to overfit and can be sensitive to individual data points. I recommend including some sort of evaluation of the models (e.g., leave-one-out cross validation).

We have included a leave-one-out cross validation approach for the CART modeling to determine how individual points influence estimates of relative importance within our model framework (L245-248, Figure S7). In our analysis, we show that although individual points can clearly impact relative importance estimates, when all points are considered simultaneously, estimated relative importance estimates generally line up with median values from the LOOCV analysis (Figure S7). This suggests that the CART analysis consistently identifies certain variables as more influential in making predictions, and results are relatively robust to individual data points. We have also included text within Section 4.5 of the discussion that further discusses difficulties and opportunities in collecting and analyzing data on dynamic stream networks.

Specific comments:

L54-56: I would perhaps qualify this as '... is often the most important...' since there are conditions when solar radiation is a secondary driver of river temperature (e.g., winter periods for well-shaded reaches - see Leach et al. (2023) and maybe references within for some examples).

We completed the requested change.

L58-59: I'm not sure the Webb and Zhang (1999) or Mohseni and Stefan (1999) are the best references to support the statement that runoff composition and groundwater inflow are important influences on river temperature. The former focused on essentially point-scale heat budgets with an emphasis on energy exchanges at the air-water interface and the latter looked at air-water temperature relationships. A better reference might be Cadbury et al. (2008).

We have included the suggested reference.

L77: Typo.

We fixed the typo.

L85-86: The second objective is awkwardly worded. It seems to state whether clusters of air-water temperature correlations differ from clusters based on air and water temperature. Before reading the rest of the manuscript, this objective seemed to me to be asking the same thing. Consider rephrasing for clarity.

We have amended objective two to state “What are the representative thermal sensitivity regimes, how do they cluster on the landscape, and how do these clusters differ from clusters based on air and water temperature individually?” We hope that this clarifies the objective for readers.

L98-109: This paragraph would benefit from some specifics. For example, provide mean January and July air temperatures and give some idea of precipitation amounts. 'Wenatchee receives a greater proportion of winter precipitation as snow' - how much greater? Figure 2 provides some context, but include some summary statistics within the text, as readers unfamiliar with this region will have little context for these general statements.

We moved long-term average annual temperature and precipitation information to this section and provide additional details about individual years of data used in this analysis in L252-259.

L116: Was air temperature also logged hourly?

Yes. We clarified this point in the manuscript.

L164: This question may not make sense, as I'm not familiar with TVCMs: What window size (in days) corresponds with a bandwidth of 0.2?

The window size corresponds with 20% of the annual data, or around 73 days, with higher weight given to data closer to the point of interest.

L220-221: How were clusters with mean Jaccard coefficients between 0.5 and 0.75 treated?

We clarified this point in the manuscript. New text states “Clusters with a coefficient larger than 0.75 were considered stable, clusters with a coefficient between 0.5 and 0.75 indicate that the cluster is measuring a pattern in the data but exact site assignment may be doubtful, and clusters with a mean Jaccard coefficient of less than 0.5 were considered unstable and may not reflect a true pattern in the data (Maheu et al. 2016, Savoy et al. 2019).”

L240-242: I see you include these long-term air temperature and precipitation values here. As I noted above, I suggest moving some of these long-term values up to the study site description. Also, what do you mean by 'long-term'? Figure S1 seems to suggest 1901-2000, but this is not clear. Also, are these DayMet output? Weather station data (if so, which stations)? Please clarify where these values come from.

The long-term data mentioned in the Supplementary Material is from the NOAA National Centers for Environmental Information climate divisional time series, which has been clarified in the text. We also included the specific years included in the long-term average in the main body of the text.

L257: I thought the data only focused on total SWE, but this statement suggests a relationship with 'snowmelt events'. It's not clear when and how the analysis focused on snowmelt events. Or are the authors assuming a single snowmelt event occurring in the spring? Is this reasonable to assume? My

guess is that, given the region, these watersheds are located within a transient snow zone and snowpacks can form and melt multiple times per winter, but maybe that's an incorrect assumption?

The SWE variable used in our analysis was calculated seasonally as the difference in SWE at the start of the season and the end of the season (L151-153). We have changed the notation to ΔSWE to clarify this in our manuscript in Figure 3 and L269-270.

L254-262: I was waiting to see if there was any explanation of how these landscape variables were calculated/estimated. There is a reference to Hill et al. 2016 in Table 1, but that citation is not in reference list. I would guess that mean slope and elevation were derived from a DEM, but I have no idea where a hydraulic conductivity estimate would come from. Is this estimate for the channel bed? Surficial geology of the upslope area?

We have amended L134-141 to include more detail about covariate calculation and corrected the citation list to include Hill et al. 2016. The new text reads "Watersheds for each site were delineated and covariates describing the watersheds were derived from commonly available geostatistical products (Table 2). Covariates were divided into four broad categories: basin topography (watershed area, mean watershed elevation, average stream slope, and distance upstream), land use (percent watershed forest, riparian forest, and lake area), climate (average temperature, precipitation, and percent precipitation falling as snow), and hydrogeologic (baseflow index, hydraulic conductivity, and soil depth to bedrock). Temperature, precipitation, and percent precipitation as snow were obtained from DAYMET Daily Surface Weather data (Thornton et al. 2020) and all other landscape covariates were obtained from the Stream-Catchment (StreamCat) Database (Hill et al. 2016)." StreamCat documentation describes the development and processing of all metrics, including hydraulic conductivity and baseflow index values. For example, the hydraulic conductivity is calculated from mean lithological hydraulic conductivity (micrometers per second) content in surface or near surface geology, which we now state in Table 1. We do not currently include processing details in our manuscript, as they are easily accessible in the StreamCat database, but we would be happy to include further details in Table 2 if the editor wishes to see the change.

L282-301: Can the number of sites within each cluster be included in the text? It's done for a few clusters but including all of them would limit the need to reference back to the table.

We completed the requested change.

L302: What is meant by 'hydrogeology' here?

A previous reviewer noted that baseflow index was a hydrologic property, whereas hydraulic conductivity and soil depth to bedrock described geologic aspects of the watersheds. Hydrogeologic simply indicates variables that describe, either directly or indirectly, the distribution and movement of groundwater through soil and bedrock. We hope the changes to L134-141 clarify this point.

L317-319: I don't understand this statement, especially '... reflect aspects of river dynamics not redundant with water and air temperature.' But aren't air temperature and climate related? Also, do the results of this

study support this statement? It seems like most of the landscape variables (I assume some of these are what the authors mean by 'geology') have very weak correlations with thermal sensitivity. Even the CART analysis seems to suggest minimal explanatory power of these variables.

This statement is simply meant to indicate that thermal sensitivity reflects unique properties of river thermal regimes that are not captured by water or air temperature alone. We have clarified this statement, and the new text reads “We find that underlying geology and climate are important controls on thermal sensitivity across two Pacific Northwest river basins, and thermal sensitivities reflect aspects of river dynamics not redundant with water and air temperature.”

L335: Perhaps include Kelleher et al. (2021) here. Although focused on river temperature trends, not thermal sensitivity, they make a similar key point that seasonal trends can differ from annual or just summer patterns.

We completed the requested change.

L356-357: How are the processes controlling river temperatures 'more diverse' in spring/summer than in fall/winter? I would argue all the same energy exchange processes are occurring (radiative and turbulent exchanges, advection, etc.), it is just the relative magnitudes that differ seasonally.

We have amended this sentence to use the above wording, specifically, “The greater variability of responses in spring and summer indicates that the relative magnitude of energy exchange processes controlling river temperatures are more diverse than in fall or winter”.

L372: This is the first mention of glacial influence in these watersheds. How much glacial coverage is there? Which sites had upstream glaciers? Why wasn't glacial coverage included as a landscape variable?

We utilized the 2019 National Land Cover Database for the percent of the upstream watershed classified as ice/snow land cover when drafting this statement. Values are generally small within our basins, and range from 0-0.7% in the Snoqualmie basin, and 0-3.2% in the Wenatchee basin. As true glacial input is minimal within our basins, we have amended this statement to state “This is likely due to snowmelt inputs within these catchments, and points to the importance of high elevation, late-summer snowpack melt as a significant source of summer baseflow and control on water temperatures during the months of greatest heating within these watersheds.”

L387: This seems to be the first time that 'geologic controls' is clarified to mean baseflow index, hydraulic conductivity and soil depth. Although this may seem obvious to some readers, I think this should be clearly stated earlier in the manuscript. Baseflow index can be influenced by factors other than groundwater (e.g., persistent, high-elevation snowpacks, glaciers, or flow regulation - especially downstream of a dam/lake, which seems to be the case for some of these sites). In addition, there are no details on where these hydraulic conductivity and soil depth estimates come from and what they represent.

Soil depth indicates the mean depth (cm) to bedrock of soils within the watershed and hydraulic conductivity is calculated from mean lithological hydraulic conductivity (micrometers per second) content in surface or near surface geology. L133-140 now includes more details on covariate selection and Table 1 includes a more detailed description of hydraulic conductivity. We assume that soil depth to bedrock and baseflow index are familiar enough to readers of HESS that we do not need to include a description, although we would be happy to include further details in Table 2 if the editor wishes to see the change.

L393: Are 'groundwater metrics' clearly important? Some of the variables that could be associated with groundwater influence often have relative variable importance values of less than 10% - that doesn't seem very important to me. Also, there is no performance evaluation of the CART model.

We have changed this wording to “hydrogeologic” to better reflect that we are referring to baseflow index, hydraulic conductivity, and soil depth in this sentence. Additionally, see our revised leave-one-out-cross-validation analysis for greater detail on the evaluation of the CART model.

L401-403: It is difficult to follow the logic here. The authors highlight that the relationships between thermal sensitivities and groundwater metrics were mixed (and in some cases they were counter-intuitive). They note uncertainty in using these metrics to capture groundwater influence, especially in mountain headwater streams. They then conclude that thermal sensitivity is a promising indicator of groundwater influence. I don't see how the results of this study support this statement.

We believe that the use of the term “a promising indicator” does imply that more work on the topic needs to be completed. We have amended L418-422 to state “The ability to use thermal sensitivity as an empirical measure of groundwater influence, therefore, shows great promise for understanding catchment processes and informing management and restoration actions at ecologically relevant scales (Snyder et al. 2015). Although our approach moves us closer to a mechanistic understanding of the relationship between thermal sensitivity and groundwater, mixed results from our analyses emphasize the need for additional targeted studies” to clarify our thinking.

L410-411: Looking at Figure 6, I can't tell that soil depth, hydraulic conductivity and baseflow index are high in streams that overlay the lower portion of the watershed. Can these be shown in a more clear and convincing way?

We have amended this line to specify sites from Clusters 1 and 4, which will provide spatial context for readers.

L404-432: A lot of geological context is suddenly provided in this section. Have the authors considered putting some of this context within the study area description? Also, are there maps to show where the measurements sites are relative to these geological features?

We amended the methods to provide geologic context earlier in the manuscript, in L101-106.

L471: Did the authors explore the sites that were located downstream of reservoirs and lakes? Could that explain some of the spatial variability observed in this study? A number of studies have highlighted that reservoirs and lakes can have a strong influence on downstream thermal regimes.

We explored this option through the inclusion of percent upstream lake area as a potential covariate.

Figure 1: Why is there a dashed line for thermal sensitivity = 0.5?

The dashed line at 0.5 is just included as a reference for easier visualization across graphs. We now state this in the figure legend.

Figure 2: Where were the SWE and precipitation data collected? How representative are these values for the entire watersheds?

We have amended the Figure 2 legend to state "Average annual discharge, SWE, and total precipitation for the outlets of the Snoqualmie and Wenatchee basins across the sampling time frame (black dashed lines) and interannual variability across the seven water years included in this analysis (gray lines). Discharge gage locations can be found in Figure 1A and 1B, and SWE and precipitation data is from DAYMET Daily Surface Weather data for the upstream watershed of each discharge gage (Thornton et al. 2020)." Discharge gages are already present in Figures 1A and 1B.

Figure 3: Please label the subplots with (A), (B), and (C), as indicated in the caption. Also, it would be interesting to see thermal sensitivity plotted against mean summer stream temperature.

We labeled the subplots with A, B, and C, and thank the reviewer for pointing out this omission. We do not show the relationship between thermal sensitivity and mean water temperature here, in order to keep the figure limited to climate covariates. However, we hope that the RShiny application and Figures 4 and 5 are illustrative of the relationship between thermal sensitivity and water temperature within our basin.

Figure 4 and 5: Can the number of sites within each cluster be shown on these figures (e.g., change the facet labels to show: 'Cluster 1 (n = XX)').

We have completed the requested change.

Table 1: What is the Hill et al. 2016 data source? It is not listed in the reference list.

We have corrected this oversight and thank the reviewer for pointing it out.

Table 2: How were the data grouped to compute these metrics? This is not clear to me. Are these simply summaries of daily mean air and water temperatures grouped by site, season and year? Or are these the means of the inter-annual thermal sensitivities estimated by the time-varying coefficient models?

We have amended the Table 2 legend to state “Air water correlation average summary metrics by basin and season. Averages are calculated as the mean value of summary metrics at all sites across each basin and season.”

Figure S1: Please show precipitation anomaly in SI units.

We have completed the requested change.

References

Cadbury, S. L., Hannah, D. M., Milner, A. M., Pearson, C. P., & Brown, L. E. (2008). Stream temperature dynamics within a New Zealand glacierized river basin. River Research and Applications, 24(1), 68-89.

Kelleher, C. A., Golden, H. E., & Archfield, S. A. (2021). Monthly river temperature trends across the US confound annual changes. Environmental Research Letters, 16(10), 104006.

Leach, J. A., Kelleher, C., Kurylyk, B. L., Moore, R. D., & Neilson, B. T. (2023). A primer on stream temperature processes. Wiley Interdisciplinary Reviews: Water, e1643.

References

Cristea, N. C., and S. J. Burges. 2010. An assessment of the current and future thermal regimes of three streams located in the Wenatchee River basin, Washington State: some implications for regional river basin systems. Climatic Change 102:493–520.

Li, H., X. Deng, C. A. Dolloff, and E. P. Smith. 2016. Bivariate functional data clustering: grouping streams based on a varying coefficient model of the stream water and air temperature relationship. Environmetrics 27:15–26.

Steel, A. E., C. Sowder, and E. E. Peterson. 2016. Spatial and Temporal Variation of Water Temperature Regimes on the Snoqualmie River Network. Journal of the American Water Resources Association 52:769–787.

Yan, H., N. Sun, A. Fullerton, and M. Baerwalde. 2021. Greater vulnerability of snowmelt-fed river thermal regimes to a warming climate. Environmental Research Letters 16:054006.