Reviewer 1 Comments

In their study, McGill et al. characterized the thermal sensitivity of streams in two watersheds of the Pacific Northwest of the United States which describes how changes in stream temperature track changes in air temperature. They characterized thermal sensitivity using the conventional method looking at the slope between air and stream temperatures. They also used a novel approach using time-varying coefficients to capture how thermal sensitivity varies through the year – this is a truly interesting contribution to the field to assess in a continuous way the seasonality of thermal sensitivity. McGill et al. then performed a clustering analysis on the annual average time series of thermal sensitivities and used classification and regression trees to identify drivers of thermal sensitivity.
Overall, the manuscript is well written and beautifully illustrated. Methods are well described and sound. While results per se are mainly of regional interest, their use of time-varying coefficients offers a methodological contribution of interest to the journal. With a few revisions, this would make a quality contribution to HESS. The main points to address are the following:

We thank the reviewer for their interest and positive assessment of the topic and methodological approach. We also appreciate their thoughtful critiques, which we address below.

1) Clearly present hypotheses/predictions underlying the study
The manuscript identified three broad research questions (lines 82-85) and most of the results section then goes on to describe observed patterns found posteriori. Using a descriptive research approach is perfectly sound but I believe the study would be more informative if it used an explanatory research approach where the goal is to understand underlying causal mechanisms. In fact, at numerous places in the manuscript, authors talked of expected results: “we expected thermal sensitivity to increase with river size” (line 418), “we expected land cover characteristics such as open water and forest cover to be important predictors” (line 426) or “expectations of a negative relationship between thermal sensitivity and groundwater influence” (line 362). While not presented this way, it appears authors had hypotheses/predictions underlying their work.

I believe framing the manuscript to more clearly present hypotheses/predictions would make conclusions of broader interest to the stream temperature community in comparison to the current presentation of results which can be difficult to interpret without regional knowledge. For example, presenting results for the Chiwawa, White and Little Wenatchee rivers (lines 270), the tributaries to the mainstem and Raging River (line 259) or the Chumstick Creek (line 415) is factually correct but bears little meaning to someone unfamiliar with the study region.

There is a large body of work examining drivers of air and water temperature correlations, therefore we had numerous hypothesized drivers based on first principles and previous literature. The background work and these hypothesized drivers informed our decision about the suite of potential predictors to include. The drivers are often highly correlated, and we therefore attempted to summarize the structure of predicted drivers and their impacts on thermal sensitivity in Table 3. We chose to present the summary metric component as an exploratory analysis for a variety of reasons. First, exploratory research provides a flexible framework for investigating complex and multifaceted topics, enabling the generation of novel ideas and hypotheses. Overreliance on hypothesis testing can pose dangers to the research process, including an overemphasis on statistical significance and p-hacking, which compromises the integrity and reproducibility of research findings (See Special Issue in The American Statistician 2019 Volume 73, Statistical Inference in the 21st Century: A World Beyond $p < 0.05$; Amrhein et al. 2019; Wasserstein & Lazar 2016). Importantly, the structure of our data lends itself more to an exploratory analysis than testing of a suite of individual hypotheses. Our study utilized a series of spatially distributed sites across the basin, and the configuration of these sites was designed to capture the range and variability of air and water temperature across the basin but not to test hypotheses about specific, causal mechanisms of thermal sensitivity. For example, ideally, if we wanted to test the impact of watershed slope on thermal sensitivity, we would have a series of more-or-less identical sites where only watershed slope varied between them to isolate slope as a driver. As variables across our basin are highly correlated, and our sample size only moderate, it would be difficult to parse apart the impact of specific drivers. We therefore believe that it is best not to frame our work in an explicit hypothesis testing framework for this manuscript.

However, as both reviewers brought up the same point, we clearly did not emphasize our statistical decision-making framework enough in our manuscript and will work to clarify it throughout. In particular, we will modify the methods paragraph on L127-133 to 1) explicitly state that our summary metric analysis was exploratory in nature to better understand patterns to set up future hypothesis testing, 2) ensure readers understand that relationships between thermal sensitivity and basin properties shown in Table 3 are hypotheses based on first principles that we lay out but do not explicitly test, 3) remove linear fits from Table 3 and instead include loess curves to aid the reader in visualization and avoid implying a regression was run, and 4) modify our phrasing of "summary metrics" results section accordingly.

Additionally, we agree that the discussion contains substantial regional knowledge that the average reader may not be familiar with. We believe that the details are important to include as they're often critical for understanding processes within a given basin and useful for local resource managers and practitioners. To provide context for readers, we will add a map with subbasin names, lakes, dams, and elevation as a supplementary figure. We will also modify lines where specific places are called out (e.g., Lines 259, 266, 270, etc.) so important elements about the basin are described, with the name of the location in parentheses to limit the necessity of regional knowledge to understand results.

Along the same lines, authors performed three distinct cluster analyses (air temperature, water temperature, thermal sensitivity) and their goal was unclear until I reached the discussion and understood that they wished to show that thermal sensitivity clusters offers additional information to what we find when studying solely air/stream temperatures. If that was one of the goals of the study, I suggest it be clearly stated and communicated as a take-home message.

We will amend the manuscript to include this. We will modify L83-84 from "What are the characteristic regimes of air-water temperature correlations and how do they cluster on the landscape?" to "What are the characteristic regimes of air-water temperature correlations, how do they cluster on the landscape, and how do they differ from clusters of only air and water temperature?" to highlight that this was a goal of our study. Furthermore, we will add a sentence in L181 (Agglomerative Hierarchical Clustering in the Methods section) reiterating that we ran three distinct clustering analyses.

2) Conclusions need to be better supported by analyses and results.

A few statements in the results are not sufficiently supported by analyses. Moreover, the results section often lacks precision and statements are often little quantified. For example, the abstract states that thermal sensitivity regimes "differed in both timing and magnitude of sensitivity" and while Figures 4-5 offer a nice illustration of regimes, no formal analysis clearly compared the timing/magnitude of clusters. There are a few other examples in the results section:

We will modify Table 4, which currently includes measures of the cluster-specific thermal sensitivity range, mean, and stability, to also include timing for maximum and minimum thermal sensitivity.

The manuscript states that "thermal sensitivity estimates were not entirely consistent" (line 230) although it is not clear what consistent refers to and if it was quantified. Similar wording regarding a "consistent seasonal signal" (line 243) should be revised.

We will modify our use of the word consistent to identical.

The manuscript states that "only SWE displayed a relationship with thermal sensitivity" (line 233) while no formal analysis was done – a visual assessment is not sufficient to determine the presence/absence of a relationship. At minimum, R^2 values should be presented in Figure 3 to assess the strength of relationships. I also question here the presence of the "wedge-shaped pattern" (line 234) for SWE in Figure 3 which does not stand out clearly and may be simply due to fewer data points for large SWE values. Further analysis is required to assess the strength of this relationship.

Our "thermal sensitivity metrics" section was exploratory in nature due to the structure of our data, which we will more clearly reiterate in the manuscript (see the above response for further details).

Similarly, I question some of the relationships between variables that are discussed in Table 3. It is not clear to me that we can see a "consistent negative relationship between thermal sensitivity, distance upstream and MWE" (line 236). Figures in Table 3 do not present the correlation coefficient but my visual assessment is that it is likely close to 0 for MWE. The results section also points towards a "weakly positive and parabolic" (line 238) relationship between hydraulic conductivity and thermal sensitivity in the Snoqualmie basin, yet a linear regression is plotted in the figure in Table 3. Overall, many different relationships between thermal sensitivity and environmental variables appear to be weak and should be confirmed using statistical analyses.

We thank the reviewer for pointing out that including linear fits to the data in Table 3 suggests that regressions were run; we will remove the lines from this table and instead add loess curves, in addition to explicitly stating that no regressions were run and our summary metric was exploratory in nature. Given that we don't necessarily expect linear relationships, we are hesitant to include correlation coefficients, particularly when adding a nonlinear smoothed line to aid visualization.

Furthermore, we agree that patterns are often weak and inconsistent, and explicitly state "Overall, weak and inconsistent patterns emerge in summer between thermal sensitivity and landscape and climate variables (Figure 3; Table 3)" on Lines 232-234. Inherent covariation in river basins can hinder statistical efforts to identify mechanistic links between landscape gradients and features of aquatic ecosystems (Lucero et al. 2011); variables may have a small impact that went undetected due to noisy observations or limited variability within our study region (Lines 428-440 discuss this in the manuscript). We nevertheless thought it important to include the summary metric analysis in our results, as these covariates are assumed to be important controls on thermal sensitivity, and we aimed to clearly set up a framework in which future studies could conduct more targeted analyses.

Last, the paragraph from lines 256-275 should be more precise and quantify some statements such as "somewhat high mean thermal sensitivities" (line 263), "overall high thermal sensitivity and low variability" (line 267), "cluster 3 had the greatest variability through time" (line 271)

We will add more specific text to this section, in particular referencing a modified Table 4 to include exact values for cluster-specific thermal sensitivity range, mean, stability, and timing for maximum and minimum thermal sensitivity.

3) Better consideration of interannual variability in thermal sensitivity

The cluster analysis of thermal sensitivity relies on an annual average time series of thermal sensitivities. I suggest that the manuscript better lay out the implications of having sites with fewer years of data. Did this have a strong influence on the clustering? For example, was the clustering similar if performed using a single and most common year of data available? Section 4.5 in the discussion does a very good job of discussing limitations in general terms but adding a more formal analysis would be more convincing.

The reviewer brings up a good point that we don’t consider interannual variability explicitly in our clustering analysis. Ideally, we would be able to run our clustering algorithm for each year individually to assess how clusters differ across specific years. However, the issue that arises is that the set of sites with continuous data can be quite different between years, limiting our potential to compare across years. Therefore, to assess cluster sensitivity to interannual variability, we will use a “leave-one-out” approach similar to the stability analysis outlined in Lines 199-205 (assessing cluster stability when leaving out sites). We will leave one year out when calculating average annual time series, and subsequently run the clustering analysis on the annual average time series for N-1 years of data. We will then compare cluster similarity to our results reported for all years of data. This analysis will be completed for each year we have data for. Results from the analysis will allow us to assess if any specific years have a particularly strong influence on clustering results (i.e., clustering results differ substantially when data from 20XX is removed). Results will be reported in detail in the supplementary material, and implications of the sensitivity analysis will be discussed in the “caveats and limitations section” of the discussion. This assessment will be a first step towards more comprehensively assessing interannual variability. Results can also be compared to a large body of work assessing interannual variability of water temperature, particularly in the Snoqualmie River (Steel et al. 2019).

MINOR POINTS

line 52: Define thermal memory as it is not a widely accepted concept.

We will make the requested change.

line 179: what is the dissimilarity matrix d^c xy ?

This is notation used to reference the spatially weighted dissimilarity matrix, whereas dxy references the Canberra distance matrix. We will clarify this in the methods.

line 246: Although presented in Supplementary Material (Table S2), I suggest adding a sentence to give an idea of the variability in the number of clusters according to the method used.

We will make the requested change.

line 249: Without regional knowledge, it is not clear from figures that “air and water temperature correspond closely with elevational gradients”.

We will add elevation shading to the basins in a new supplementary figure map to clarify this point.

line 302: thermal sensitivities varied substantially between sites? I suggest being more explicit as to what is being compared here.

We will modify this sentence to “Fall thermal sensitivities were relatively homogeneous, with 90% of values falling between 0.47 and 0.70, whereas spring and summer thermal sensitivities exhibited a broader range of values, with 90% of values falling between 0.30 and 0.84 in spring and 0.25 and 0.78 in summer.”

line 306: non-redundant aspects relative to what? I suggest being more explicit as to what is being compared here.

We will modify this sentence to “Thermal sensitivity regimes reflect non-redundant aspects of river dynamics relative to air and water temperature alone.”

line 323: This statement is a bit strong and little supported by results. For example, static thermal sensitivity (e.g. Table 2) may in fact align well with clusters defined using the time-varying approach, something the manuscript did not look into.

The way thermal sensitivity is typically measured, it is often conceptualized as a single, stationary value, rather than an average of multiple estimates. We believe that this is an important distinction; recognizing that a parameter shifts over time and using the average is fundamentally different from assuming a parameter is static through time. Our point here was that recognizing variability in this parameter is important, and we will work to clarify this in the manuscript.

line 334: To what does the buffering refer to?

Buffering refers to the process wherein snowmelt-influenced streams have lower thermal sensitivity (i.e., buffering against climate variability). This is due to a direct input of cold water and a corresponding increase in flow rates and depths which mitigates the impact of surface heat exchanges by increasing thermal inertia (van Vliet et al. 2011; Siegel et al. 2022). We will work to clarify this in the manuscript.

line 335: A comma is missing after “summer”

We will make the requested change.

line 361: Do “summary metric regression” refer to Table 2?

Yes, the summary metrics refer to Table 2, however, this is a mistake in wording on our part. We will amend the sentence to state “… results from the summary metric exploratory analysis were mixed…”.

line 435: Are there large dams in the two studied basins? If so, it should be clearly stated as this could explain why certain environmental variables had little influence.

There is a dam and reservoir on a major tributary to the Snoqualmie River, the Tolt River. Several small dams exist on tributaries to the Wenatchee River, and a large lake (Lake Wenatchee) sits at the junction of the White and Chiwawa Rivers. We will include all basin names, lakes, and dams on a map in the supplementary material and reference their potential to influence results in the manuscript.

line 457: What were the bandwidth and averaging periods used? I couldn’t find this information anywhere in the methodology.

We thank the reviewer for pointing out this omission. We will include the bandwidth used in the methods section.

Citations

Amrhein, V., Greenland, S., McShane, B. 2019. Scientists rise up against statistical significance. Nature, 567: 305-307, DOI: https://doi.org/10.1038/d41586-019-00857-9.

Wasserstein, R.L. & Lazar, N.A. 2016. The ASA statement on p-values: context, processes, and purpose. The American Statistician, 70(2): 129-133, DOI: https://doi.org/10.1080/00031305.2016.1154108.

Steel, E.A., Marsha, A., Fullerton, A.H., Olden, J.D., Larkin, N.K., Lee, S.Y., Ferguson, A. 2018. Thermal landscapes in a changing climate: biological implications of water temperature patterns in an extreme year. Canadian Journal of Fisheries and Aquatic Sciences, 76(10): 1740-1756, DOI: https://doi.org/10.1139/cjfas-2018-0244.

Lucero, Y., Steel, E.A., Burnett, K.M., Christiansen, K. 2011. Untangling Human Development and Natural Gradients: Implications of Underlying Correlation Structure for Linking Landscapes and Riverine Ecosystems. River Systems, 19(3): 207–24, DOI: https://doi.org/10.1127/1868-5749/2011/019-0024.

van Vliet, M.T.H., Ludwig, F., Zwolsman, J.J.G., Weedon, G.P., Kabat, P. 2011. Global river temperatures and sensitivity to atmospheric warming and changes in river flow. Water Resources Research, 47:W02544, DOI: https://doi.org/10.1029/2010WR009198.

Siegel, J.E., Fullerton, A.H., Jordan, C.E. 2022. Accounting for snowpack and time-varying lags in statistical models of stream temperature. Journal of Hydrology X, 17: 100136, DOI: https://doi.org/10.1016/j.hydroa.2022.100136