Original reviewer comments are in blue.

Following HESS review policy, we initially replied to each of the reviewer's comments, but did not prepare a revised manuscript. These comments are in black.

After the HESS editor decision to revise and resubmit, we prepared a revised version of our manuscript. Specific changes to our manuscript are tracked in red.

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.

We changed the title of section 2.2. to “Exploratory analysis of air-water correlation summary metrics” and modified L133-141 to state “A large body of literature examines landscape-level drivers of air and water temperature correlations within rivers. We therefore summarized hypothesized drivers of thermal sensitivity based on previous literature and their covarying landscape variables within our basins. We then conducted an exploratory analysis of the relationship between landscape covariates and thermal sensitivity to better understand patterns in our data and set up future hypothesis testing. Due to the correlated nature of our dataset, no formal statistical tests were conducted. We plotted summer thermal sensitivity metrics against hypothesized drivers, including mean watershed elevation (MWE), watershed slope, distance upstream, percent riparian forest cover, and substrate hydraulic conductivity. Loess curves were plotted to aid in data visualization, and correlation coefficients between thermal sensitivity and each landscape covariate were used to quantify the strength of the linear relationship. Covariate descriptions and sources are found in Table 1.”

We also removed linear fits from the plots in Table 3 and now explicitly state that loess curves were included to aid in visualization in the caption to Table 3, and report correlation coefficients for each plot to describe the linear relationship between thermal sensitivity and landscape covariates. We amended L258-259 to explicitly state that “For landscape variables, correlation coefficients were overall small (|ρ|

< 0.3), indicating weak to non-existent linear relationships between landscape covariates and observed thermal sensitivity."

We have also included a supplementary figure with subbasin names, lakes, dams, and elevation to provide greater context for interested readers (Figure S4 and S5).

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.

We modified L87 accordingly and reiterated our intention to compare clustering of thermal sensitivity, air, and water temperature in Section 2.3, L150-153 and L198-200.

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.

Table 4 formerly reported the mean, maximum, and minimum thermal sensitivity for all individual stations within a cluster. We modified this slightly to report the cluster-average mean, maximum, and minimum to improve clarity and facilitate comparison across different clusters more easily. We also included the timing of the cluster-average minimum and maximum values and cluster-averages for air and water clusters in addition to thermal sensitivity.

We included several quantitative metrics from Table 4 to describe cluster differences in Section 3.2, L269-301. A few examples from this paragraph are described below.

1) L275-276 we state, "For example, within both basins seasonal water temperatures were synchronized, with the cluster minimum and maximum water temperatures occurring within a day of each other (Table 4)."

2) L288-289 we state, "Cluster 2 was characterized by a mean thermal sensitivity of 0.52 and the highest annual variability, with a cluster-average range of 0.45".

3) L294-296 we state, "Clusters 1, 4, and 5 demonstrated similar seasonal patterns in thermal sensitivities, with minimum values occurring in late Spring (water days 216, 207, 214) and maximum values occurring in late summer (water days 324, 331, 330)."

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.

We have made the requested change. Furthermore, the figure this sentence describes is now in a separate appendix (Appendix A) devoted to a description of interannual variability within our dataset.

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).

See above for a more detailed description of completed changes. The phrase was edited to state "only SWE appeared to have a linear relationship with thermal sensitivity."

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.

We removed linear fits from the plots in Table 3 and now explicitly state that loess curves were included to aid in visualization in the caption to Table 3, and report correlation coefficients for each plot to describe the linear relationship between thermal sensitivity and landscape covariates. We amended L258-259 to explicitly state that "For landscape variables, correlation coefficients were overall small ($|\rho| < 0.3$), indicating weak to non-existent linear relationships between landscape covariates and observed thermal sensitivity."

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.

We have modified this section to include more specific descriptions of cluster properties, particularly drawing from cluster-average values listed in Table 4. A few examples from this paragraph are described below, but please see the revised manuscript to view all changes.

1) L275-276 we state, "For example, within both basins seasonal water temperatures were synchronized, with the cluster minimum and maximum water temperatures occurring within a day of each other (Table 4)."

2) L288-289 we state, "Cluster 2 was characterized by a mean thermal sensitivity of 0.52 and the highest annual variability, with a cluster-average range of 0.45".

3) L294-296 we state, "Clusters 1, 4, and 5 demonstrated similar seasonal patterns in thermal sensitivities, with minimum values occurring in late Spring (water days 216, 207, 214) and maximum values occurring in late summer (water days 324, 331, 330)."

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).

We have completed a sensitivity analysis to assess if the removal of data from a specific water year had a particularly strong influence on clustering results. Results indicated that, although cluster agreement between our reported results and the reduced dataset was not perfect, our analysis using average annual data generally captured broad patterns within our dataset. More detail on this new analysis is included in the revised Appendix A of our manuscript.

MINOR POINTS

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

We will make the requested change.

We have changed “thermal memory” to “annual hysteresis” as we believe this is a more widely accepted term.

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.

We have added the phrase “$d_{xy}^{c}$is the spatially weighted dissimilarity matrix” to L182 in our manuscript.

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.

We have added the range of clusters suggested by cluster validity indices to L269-272.

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.

We have added supplementary figures S4 and S5 showing elevation gradients for each basin.

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.”

We modified L248-253 in the results 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” to better support this statement in the discussion.

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.”

We modified 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.

We have modified L313-321 to state “Thermal sensitivity varies throughout the year and reflects hydrologic conditions at a given time and place within a watershed; therefore, it should not be conceptualized as a static value. Although summary metrics of thermal sensitivity, such as average values over the summer, can still prove useful and informative, it is essential to acknowledge the non-stationarity of the relationship between air and water temperature to obtain a more accurate understanding of how river temperature responds to changing conditions.”

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.

We have modified L366-368 to state “Importantly, snowmelt buffering, the process wherein snowmelt-influenced streams have lower thermal sensitivity due to a direct input of cold water and a corresponding increase in flow rates and water depths (van Vliet et al. 2011, Siegel et al. 2022), diminishes throughout the summer.”

line 335: A comma is missing after “summer”

We will make the requested change.

We completed 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…”.

We completed the stated change.

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.

We have included supplementary figures (Figure S4 and S5) with the location of the single reservoir and large lake within our basin.

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.

We have included the bandwidth utilized (0..2) in the methods on L164.

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

Reviewer 2 Comments

Overall, I really like this study, from the conceptual development, to the data collection, to much of the analysis (especially continuous time series of stream thermal sensitivity), and discussion. I think there is great transferrable value of interest to HESS readership. I have some criticisms of the way the sensitivity metric data are visualized and discussed in Figs 1 and 3, but I really like the metric time series analysis is shown in Fig 4 and 5.

We thank the reviewer for their positive assessment of our manuscript. We also appreciate their thoughtful critiques, which we address below.

It would be nice to show representative streamflow from those basins over the same time periods to help assess how thermal sensitivity may be driven by the volume of water in the channel at any one time (determines channel water thermal inertia to changes in net heat flux). Low stream discharge volume may be a primary driver of increased thermal sensitivity at many sites in late summer, though I do not see discharge included in any of your quantitative analysis of controlling parameters (though baseflow index is derived from stream discharge, and is included here in a general way).

We agree that streamflow likely impacts thermal sensitivity, particularly in the dry summer months when discharge is lowest, temperatures highest, and features such as groundwater seeps may show up clearly. Discharge was not included in our analysis due to the lack of spatially and temporally resolved streamflow data across the basins. There are relatively few USGS and locally maintained discharge gauges in the Snoqualmie and Wenatchee basins, and most gauges do not directly correspond to our temperature sites. Watershed area is likely the best proxy for average annual discharge, with baseflow index loosely corresponding to specific discharge in summer. We agree that representative time series of discharge would be useful for readers and will include average discharge at the outlet of each basin as a panel on Figure 1. The location of these outlet gauges is already shown on the maps.

We have included annual time series of discharge, SWE, and precipitation for the outlet of the Snoqualmie and Wenatchee basins in a new figure, Figure 2.

As mentioned by Reviewer 1, given the 'expectations' listed in Table 3 it would be nice to frame the study as hypothesis driven/testing, which would not be a major change to what you have now. Below I list some more major and minor points that could be considered during the revision process.

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.

We changed the title of section 2.2. to “Exploratory analysis of air-water correlation summary metrics” and modified L133-141 to state “A large body of literature examines landscape-level drivers of air and water temperature correlations within rivers. We therefore summarized hypothesized drivers of thermal sensitivity based on previous literature and their covarying landscape variables within our basins. We then conducted an exploratory analysis of the relationship between landscape covariates and thermal sensitivity to better understand patterns in our data and set up future hypothesis testing. Due to the correlated nature of our dataset, no formal statistical tests were conducted. We plotted summer thermal sensitivity metrics against hypothesized drivers, including mean watershed elevation (MWE), watershed slope, distance upstream, percent riparian forest cover, and substrate hydraulic conductivity. Loess curves were plotted to aid in data visualization, and correlation coefficients between thermal sensitivity and each landscape covariate were used to quantify the strength of the linear relationship. Covariate descriptions and sources are found in Table 1.”

We also removed linear fits from the plots in Table 3 and now explicitly state that loess curves were included to aid in visualization in the caption to Table 3, and report correlation coefficients for each plot to describe the linear relationship between thermal sensitivity and landscape covariates. We amended L258-259 to explicitly state that “For landscape variables, correlation coefficients were overall small ($|\rho| < 0.3$), indicating weak to non-existent linear relationships between landscape covariates and observed thermal sensitivity.”

1. L15: ‘…it is critical to both understand the underlying processes causing stream warming and identify the streams most and least sensitive to environmental change.’ Measurement of air-water temperature relations across the landscape provides an efficient way to address this important topic. However, it is a localized measurement that may not reflect general behavior across the stream system as other related studies have shown, especially when there is strong variability in groundwater discharge (eg Z. Johnson et al papers). This point is discussed somewhat in the body text, but still could be made more clear throughout. Local stream channel heat exchange process can dominate the local air-water temp sensitivity metrics, which speaks to collecting spatially distributed datasets, as you nicely did for this study.

We agree with the Reviewer’s point that air-water temperature measurements can be localized in space and time, and believe our manuscript highlights this fact throughout. We will emphasize the fact that local stream channel heat exchange processes such as groundwater inflow can be a dominant control on thermal sensitivity in certain situations.

2. Although stream thermal sensitivity is quantified relative to changes in air temperature, air temperature warming may not always be the primary driver of stream temperature warming. Sensible heat fluxes are often dwarfed by solar and latent heat fluxes along the stream corridor. L39 acknowledges this important point. However, climate warming as typically described is primarily driven by the impacts on the global long wave radiation budget by accumulation of greenhouse gasses, not changes in solar short wave radiation input. The point that air temperature itself may not be the primary driver of stream temperature change at the seasonal timescale should be more clear, throughout. For example there is this statement on L122: 'The slope of this relationship, the thermal sensitivity, indicates how sensitive a given stream's water temperature is to changes in air temperature.' I am not sure that is true, more that air and stream temperature are sensitive to solar radiation in more or less coupled ways. This is kind of a nuanced point, but I have interacted with several people who interpret these type of metrics as air temperature often being the primary driver of stream temperature, presumably through sensible heat exchange.

The reviewer brings up an excellent point that air and water temperatures are correlated primarily due to a similar response to solar radiation, not because air temperature drives water temperature. This is a point we want to emphasize to readers, and we will amend L122 to more accurately reflect this and attempt to make it clear throughout the manuscript. We thank the reviewer for the suggested wording.

We have modified L125-128 to state "The slope of this relationship, the thermal sensitivity, indicates the average difference in water temperature when comparing time periods with a one-degree difference in air temperature. For example, a thermal sensitivity of 0.5 would indicate that, based on historical data, when air temperature at a site differs by 1°C, water temperature differs on average by 0.5°C (Leach and Moore 2019)". This new phrasing avoids implying that air temperature controls water temperature.

3. L41 and elsewhere: Addition of water to the stream channel impacts thermal inertia and stream temperature sensitivity, even if that water is of the same temperature as the channel. How are these patterns impacted by variable stream discharge at locations over time and along the stream network continuum? For example, clusters 2,3, and 4 show substantial increases in thermal sensitivity in late summer during presumably the lowest flows.

We agree that high thermal sensitivity in summer is likely mediated by low discharge, as in both the Snoqualmie and Wenatchee basins discharge is lowest in late summer. We will emphasize this in the manuscript by adding discharge time series at the outflow of each basin to Figure 1 and stating that low summer discharge values likely contribute to increased thermal sensitives in late summer in L328-341 of the discussion.

Figure 2 was added to the manuscript to illustrate basin-wide discharge regimes.

We have modified L505-508 to state "For many of our study sites, thermal sensitives were highest in late summer during the hottest, lowest flow portion of the year. Previous studies have found that the impact of fluctuations in discharge generally increases during dry, warm periods, when rivers have a lower thermal capacity and are more sensitive to atmospheric warming (van Vliet et al. 2013)."

4. I found the 'Identification of environmental drivers in thermal sensitivity' section most questionable given the relatively small sample size and lack of representation across varied types of watersheds. Also, hydrologic attributes downstream in a network are inherently influenced by physical attributes upgradient in the network, and your spatial sampling spans upstream to downstream. I think that statements such as: 'Annual patterns in thermal sensitivity are largely controlled by underlying geology and climate across two Pacific Northwest river basins' are too definitive given the sparse nature of the datasets across a range of geologic and climatic variables.

It may be that stream network position is more important that some of the apparent shifts in the tested physical variables.

We will amend this sentence to say “Underlying geology and climate are important controls on annual patterns in thermal sensitivity across two Pacific Northwest river basins”, which more accurately reflects the results of our CART analysis. We include both upstream distance and watershed area in our examined covariates for the clustering analysis, both of which had middling-to-low importance.

We amended L317-319 to state “Underlying geology and climate are important controls on annual patterns in thermal sensitivity across two Pacific Northwest river basins.” Additionally, we intentionally limited our conclusions to the Snoqualmie and Wenatchee basins, as we do not feel that we sufficiently sampled across a broad enough range of geology and climate variables to draw general conclusions.

5. The air-water temp sensitivity metrics in Fig 1 are somewhat difficult to interpret, as data are plotted seasonally over years for individual sites all by elevation. Given some sites appear at quite similar elevation, its not possible to disentangle changes by site and changes by elevation, and which sites are upstream/downstream of each other. I do not have any great advice with how to deal with this, however. Different colors for all sites would be overwhelming. Apparent trends in thermal sensitivity with elevation in some seasons may be somewhat of an artifact of plotting both watershed datasets together. Taken alone, seasonal datasets from either watershed would not seem to show an increasing trend with elevation. Given the inherent hydrogeological and climate differences between the two study watersheds I am not sure it is appropriate to depict and analysis the season metrics together.

We acknowledge that it can be difficult to show all aspects of the data in a single plot; it was not our intent to show interannual differences or upstream-downstream effects with this figure, but rather to visualize general patterns within and across river basins. Comparing across basins can be a powerful tool and is a common practice in hydrologic sciences, and our inclusion of differing colors for the basins was designed to acknowledge that basic-specific differences exist beyond the parameter (elevation) shown.

6. There are numerous places in the paper where a statistical test is inferred but it is not clear if a statistical test (along with p-value) was performed. For example: L233 ‘Overall, weak and inconsistent patterns emerge in summer between thermal sensitivity and landscape and climate variables’. While ‘patterns’ does not indicate a test, ‘weak’ does. Also, L230 ‘Thermal sensitivities for sites with consistent data coverage tended to covary,..’. Covariance is a statistical test and should be associated with a significance level. My biggest problem is with the fourth column of Table 4, where linear fits are shown to the datasets without significance levels being directly indicated. I am pretty sure that many of those fits are not significant, and therefore should certainly not be shown. Plotting the best fit lines tends to influence the reader’s perception of trends, and if they are not statistically significant, they do now exist according to those significance metrics (eg p value levels). Labeling the column ‘observed relationship’ indicates all linear fits shown are significant and I see that as highly problematic.

See the above comment for a more detailed response to the themes addressed in this comment. In short, 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 and that no statistical tests were performed, 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 avoid implying a regression was run, and 4) modify our phrasing of “summary metrics” results section accordingly.

We changed the title of section 2.2. to "Exploratory analysis of air-water correlation summary metrics" and modified L133-141 to state "A large body of literature examines landscape-level drivers of air and water temperature correlations within rivers. We therefore summarized hypothesized drivers of thermal sensitivity based on previous literature and their covarying landscape variables within our basins. We then conducted an exploratory analysis of the relationship between landscape covariates and thermal sensitivity to better understand patterns in our data and set up future hypothesis testing. Due to the correlated nature of our dataset, no formal statistical tests were conducted. We plotted summer thermal sensitivity metrics against hypothesized drivers, including mean watershed elevation (MWE), watershed slope, distance upstream, percent riparian forest cover, and substrate hydraulic conductivity. Loess curves were plotted to aid in data visualization, and correlation coefficients between thermal sensitivity and each landscape covariate were used to quantify the strength of the linear relationship. Covariate descriptions and sources are found in Table 1."

We also removed linear fits from the plots in Table 3 and now explicitly state that loess curves were included to aid in visualization in the caption to Table 3, and report correlation coefficients for each plot to describe the linear relationship between thermal sensitivity and landscape covariates. We amended L258-259 to explicitly state that "For landscape variables, correlation coefficients were overall small ($|\rho| < 0.3$), indicating weak to non-existent linear relationships between landscape covariates and observed thermal sensitivity."

7. As mentioned above, plotting data from the two study watersheds together to assess apparent changes in the sensitivity metrics across elevation and other physical variables may be problematic given the inherent differences in settings. Essentially all of the apparent patterns shown in Fig 1 and 3 would not exist if either watershed dataset was plotted alone.

Comparing across basins can be a powerful tool and is a common practice in hydrologic sciences, and our inclusion of differing colors for the basins was designed to acknowledge that basic-specific differences exist beyond the parameter (elevation) shown.

8. I am not sure I universally agree with this statement that leads the Discussion: 'Thermal sensitivity varies throughout the year and reflects hydrologic conditions at a given time and place within a watershed; therefore, it should not be treated as a static value.' Just because a parameter may show variability over time, does not mean the average value is not meaningful in assessing differences between sites. Daily temperature is one example, or anything else that varies diel or seasonally. I do agree there can be great value in inspecting short term to seasonal variation in air-water temp sensitivity metrics, but that is not a requirement of all studies to be useful.

We agree with the reviewer that summary metrics can be useful and informative! However, 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 (even if a mean value is eventually used), and we will work to clarify this in the manuscript.

We edited the initial paragraph of the discussion to state "Thermal sensitivity varies throughout the year and reflects hydrologic conditions at a given time and place within a watershed; therefore, it should not be conceptualized as a static value. Although summary metrics of thermal sensitivity, such as average values over the summer, can still prove useful and informative, it is essential to acknowledge the non-stationarity of the relationship between air and water temperature for a more accurate understanding of how river temperature responds to changing conditions."

9. It is typical to not assess air-water temp relations when stream temperature falls below some threshold close to freezing, as described by Ben Letcher's work and others. Was a cutoff value used here (eg 0.5 or 1 deg C?) It does not appear so for some of the winter datasets, which may not make sense conceptually. Stream and air temperature must decouple as the water starts to freeze, though perhaps these streams do not freeze (or come close)?

We did not use a cutoff value, and fully expect streams to decouple when air temperatures drop below freezing. The only stations where freezing occurs are high-elevation sites within the Wenatchee Basin. We will acknowledge this in the manuscript.

In L348-353 we state "Observed low thermal sensitivities in winter are likely due to the non-linear relationship between air and stream temperature at cold temperatures when air temperatures can dip below the water temperature-freezing limit (Mohseni et al. 1998, 1999). Air temperature covaries strongly with elevation in Pacific Northwest basins, and sites that are high in the watershed will experience a greater number of sub-freezing days, and therefore greater decoupling between air and water temperatures."

10. What do you think may drive the super low thermal sensitivities observed at some sites (eg less than 0.01?) That would seem to be possible mismatch of air and water temp data or a spring run creek totally dominated by groundwater near to the discharge source.

Numerous potential reasons for very low thermal sensitivities exist. As stated above, periods of time when air temperatures fall below freezing could cause a complete decoupling of air and water temperatures. Intense snowmelt over the spring season could result in decoupling if high temperatures melt snowpack, reducing water temperatures. Additionally, as the reviewer suggests, small tributaries dominated by groundwater could also decouple air and water temperatures.

Minor comments

L37: This statement could use a range of supporting citations

We will make the requested change.

We have made the requested change and included two citations to support this statement.

L41: addition of water to the stream channel impacts thermal inertia and stream temperature sensitivity, even if that water is of the same temperature as the channel.

We will include this point in the manuscript.

We have modified this sentence to state "Stream temperature is also influenced by discharge through changes to thermal inertia and residence time (Meier et al. 2003) and runoff composition where snowmelt, surface runoff, or groundwater inflow entering the stream have different temperature signatures than the stream itself (Webb and Zhang 1997, Mohseni and Stefan 1999)."

L45: 'diagnostic' tool may be better here than 'predictive' tool

We will make the requested change.

We have made the requested change.

L65: what do you mean here by ‘insensitive data’? Do you mean difficulty in collecting appropriate data to calibrate/validate heat budget models or something else?

Here we are referring to data necessary to parameterize a physically based hydrologic model, such as land use and soil parameters, surface flow characteristics and input data of rainfall, evapotranspiration, and stream flow. These data generally need to be spatially distributed and may be unavailable for certain basins or regions. We will modify the sentence to include examples of necessary data.

We have modified L38-41 to state: Issues exist with process-based modelling, including intensive data and computational needs (e.g., spatially distributed land use and soil characteristics, meteorological and discharge data, etc.), limited ability to generalize across basins, and difficulty representing groundwater and subsurface flow paths (Safeeq et al. 2014).

L72: You could pull this thought out of parenthesis.

We will make this change.

We have made the requested change.

L75: ‘along’ river networks?

We will make this change.

We have made the requested change.

L78: It is not clear here whether you are referring specifically to statistical cluster analysis or more qualitatively to spatial groupings of streams that show similar response across the landscape

In this sentence, we were referring generally to spatial groupings of similar streams. We will modify the word “clusters” to “groupings” to avoid confusion with our formal analysis.

We switched the wording from “clusters” to “groupings”.

L82: mention generally where the two experimental basins are regionally

We will add a sentence stating that the basins are located within the Pacific Northwest (western United States).

We modified L83 to explicitly state “two Pacific Northwest river basins”.

L83: it is not clear what you mean here by ‘characteristic regimes’

We will modify the phrasing from “characteristic” to “typical or representative” regimes.

We switched the wording from “characteristic” to “representative” regimes.

L85: perhaps add ‘(decreased thermal sensitivity)’ after ‘decoupling between air and water temperature’ for clarity

We will make the requested change.

We modified this sentence to state “What are the landscape or climate factors that best predict cluster membership?”

L107: Can you clarify the subscripts for number of loggers in each basin, and also list what specific Tidbit model(s) was used?

We will make the requested change. We used HOBO TidbiT v2 (UTBI-001) water temperature data loggers, which we will include in the manuscript.

We modified L113 and L116 to state the logger models used: HOBO TidbiT v2 (UTBI-001) for water temperature and HOBO Pendant (UA-002-64) for air temperature. We also modified subscripts on L110 to explicitly state $N_{Snoqualmie}$ and $N_{Wenatchee}$ to improve clarity.

L111: please clarify these are water years in North America

We will make the requested change.

We have made the requested change.

L117: Solar shields were also used for the Tidbit loggers deployed in the water?

Yes, solar shields were fashioned to house both water and air temperature loggers.

L141: drop ‘original’

We will make the requested change.

We have made the requested change.

L141: when you say ‘continuous’ metric what is the realized timestep of the output? Is it calculated by season or over entire datasets?

The varying coefficient linear model utilized mean daily air and water temperature for the entire time series.

We have modified L147 to state “…we employed a varying-coefficient linear model to obtain continuous, daily estimates of thermal sensitivity”.

L162 and elsewhere in this section: It would be helpful to have topical sentences explaining plainly why these various calculations were done before diving into the nuts and bolts of how they were done.

This is a good point, thank you. We will make the requested changes.

We have included topical sentences for each of our methods paragraphs.

L199: Can you better explain ‘the stability of clusters’ concept? Again, these methods subsections tend to dive right into the details of the calculations without a clear explanation up top of why the calculations were performed. The ‘why’ can be gleaned, but may not be clear for readers from varied scientific backgrounds.

We will make the requested change.

We have modified L216-218 to state “To determine whether clusters assignment were stable, or preserved under a perturbed dataset similar to the original and therefore likely reflective of real differences, we conducted a bootstrapping approach where sites were sampled with replacement and then AHC was performed on the resampled data using the fpc R package (Hennig 2020).” The underlying premise behind analyzing stability is that a good clustering of the data will be reproduced over an ensemble of perturbed datasets that are nearly identical to the original data.

L220: you may want to reminder what years you are talking about.

We will make the requested change.

We have made the requested change.

L230: Are you assessing covariance by eye or statistically?

We assessed covariance informally initially, however, in our updated interannual sensitivity analysis (see above response to Reviewer 1) we will add a statistical measure of interannual covariance.

The subsection 3.2 title may be better posed not as a question

We will make the requested change.

We have changed the subsection title to “Patterns of clustering for water temperatures, air temperatures, and thermal sensitivities”.

Table 1. Its probably OK, but a little odd to list Baseflow Index as a geologic variable, given the importance of groundwater levels in addition to geologic materials.

We will change the wording from “geologic” to “hydrogeologic” to clarify this.

We changed the wording from “geologic” to “hydrogeologic”.

Citations

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.

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.