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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

L45: ‘diagnostic’ tool may be better here than ‘predictive’ tool

We will make 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.

L72: You could pull this thought out of parenthesis.

We will make this change.

L75: ‘along’ river networks?

We will make this 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.

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).

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

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

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

We will make the requested change.

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.

L111: please clarify these are water years in North America

We will make 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.

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.

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.

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.

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

We will make 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.

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.

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.