This manuscript presents a large-scale analysis of seasonal and annual streamflow and stream-temperature trends across the United States. The extensive spatial coverage and analysis of multiple hydroclimatic and land-cover variables are valuable. However, I have fundamental concerns regarding whether the methodology addresses the purpose of the study and whether the resulting conclusions are supported by the analyses.

The principal motivation is that streamflow and stream temperature have generally been studied independently and that their behavior has not been broadly examined together. However, most of the manuscript still analyzes streamflow and stream temperature separately. Trends are calculated independently, relationships with potential drivers are examined separately, and the two variables are primarily compared based on coincident trend directions. The more direct paired analysis is largely limited to 218 sites in Figure 7. In my view, this does not fully deliver the joint streamflow–temperature analysis that motivates the paper.

I am also concerned by the repeated use of "attribution," "drivers," and related causal terminology based mainly on monotonic trends and bivariate Kendall correlations. This analysis identifies associations, but it does not constitute attribution. This distinction affects a substantial part of the Results, Discussion, and Conclusions.

Overall, while the continental-scale dataset is useful, the primary statistical methods are relatively conventional, and the aspects presented as the main advances including joint examination of streamflow and temperature and attribution of their drivers are not sufficiently developed. In its present form, I do not think the manuscript provides sufficient methodological or conceptual novelty to warrant publication in *HESS*.

Major comments

The Introduction identifies a clear knowledge gap. Streamflow and stream temperature have been widely studied independently, whereas their coupled behavior has received much less large-scale attention. I agree that this is an interesting and potentially important research question.

However, the methodology does not fully follow this motivation. Most analyses characterize trends in the two variables separately and subsequently compare their directions. Showing that stream temperature increases while streamflow decreases at a location is useful, but it is not the same as quantitatively characterizing their joint behavior, dependence, or co-occurrence.

There are established approaches that illustrate what could be gained from such an analysis. For example, Alizadeh et al. (2020) did not simply calculate separate trends in precipitation and temperature, they used an empirical copula to quantify the joint probability and return period of concurrent dry and hot conditions and then examined how these compound events changed over time. Similarly, Jalili Pirani and Najafi (2022) explicitly modeled dependence among multiple flood-generating variables and estimated joint and conditional probabilities using a vine-copula framework.

I am not suggesting that the authors necessarily need to adopt these exact methods. However, if the primary novelty is that streamflow and stream temperature are studied together, the analysis should quantify their joint behavior. For example, one could investigate quantities such as the probability of simultaneously experiencing unusually high stream temperature and low streamflow,

$$P(T_s > T_{90}, Q < Q_{10})$$

or the conditional probability

$$P(T_s > T_{90} \mid Q < Q_{10})$$

and determine whether these relationships are changing seasonally or temporally. Such an analysis would have direct ecological relevance and would more convincingly address the motivation presented in the Introduction.

Additionally, the attribution framework combines the significance of a temporal trend in a potential explanatory variable with the significance of its Kendall correlation with streamflow or stream temperature. These combinations are subsequently labeled as Limited, Moderate, or Robust attribution evidence. For example, a significant precipitation trend combined with a significant precipitation–streamflow correlation is interpreted as robust evidence that precipitation influences streamflow. I do not think this constitutes statistical attribution.

Precipitation, air temperature, streamflow, land cover, and other variables are mutually correlated and may share common temporal trends. A significant bivariate correlation cannot distinguish direct effects from confounding, indirect relationships, common trends, interactions, or lagged responses. Therefore, statements such as "increasing streamflow was attributed to..." or "precipitation was the primary driver..." are considerably stronger than the analysis supports (See methods on causality by Runge et al. (2019) or Zaerpour et al. (2025)).

This issue is also acknowledged by the authors themselves near the end of the manuscript, where they state that a more robust analysis, such as regression, would be needed to determine the effects of additional drivers and statistically attribute causation. This statement appears inconsistent with the attribution terminology used throughout the manuscript.

Zaerpour et al. (2025) provide one example of the distinction I am referring to. Rather than interpreting simple correlations as attribution, they used PCMCI+ causal discovery to condition on other variables and distinguish direct temporal relationships from indirect or potentially spurious associations. The point is not that the authors must use PCMCI+, but that attribution requires a framework capable of separating competing and interacting explanatory factors.

Moreover, the manuscript compares 1980–2020, 1990–2020, and 2000–2020 and interprets differences among these periods as recent shifts in some places. For example, Section 4.2 explicitly discusses a recent shift toward increasing streamflow.

The issue is not that Theil–Sen slopes from periods of different lengths cannot be compared descriptively. Rather, these are nested periods with very different record lengths and statistical power, and differences between their estimated slopes do not demonstrate that the underlying trend itself changed.

This concern is particularly important for stream temperature because the monitoring network changes substantially among periods. The manuscript reports only 52 sites with monthly stream-temperature trends for 1980–2020, compared with 110 for 1990–2020 and 364 for 2000–2020; annual-mean sample sizes are even smaller. Thus, differences in the percentage of warming or cooling sites may partly result from changes in the stations.

The authors themselves acknowledge the limited statistical power of the 21-year period, endpoint effects, and the possibility that low-frequency climate variability can appear as trends over particular periods.

At minimum, I suggest repeating the comparison using stations common to all periods, avoiding claims of shifts or reversals unless formally tested, and considering equal-length moving windows or change-point/segmented trend analysis if identifying temporal shifts is a central objective. Otherwise, the results should simply be described as differences among the estimated trends for the three analysis windows.

Also, I have an additional concern regarding the use of baseflow fraction as an explanatory driver of streamflow. The manuscript states that baseflow was obtained from the nonlinear separation model of Konrad (2022), and monthly and annual baseflow fractions were subsequently calculated. However, the Konrad model estimates the baseflow component from the streamflow time series itself; the associated data product also reports baseflow as a fraction of streamflow.

Consequently, streamflow and baseflow fraction are not independent quantities. An increase in total or quick flow can therefore cause BFF to decline even if absolute baseflow does not decline. A negative association between total streamflow and BFF may consequently arise partly from their mathematical construction (Anyway, the analysis of streamflow temperature and baseflow could also be an interesting study too).

Furthermore, I am not fully convinced by the interpretation of the Mann–Kendall likelihood categories. In particular, $p \leq 0.33$is treated as somewhat likely evidence and these observations are frequently subsequently described as warming, cooling, wetting, or drying trends. A p-value of 0.2–0.3 represents weak evidence against the null hypothesis and should not be interpreted as a probability that a trend exists. The terminology and interpretation should therefore be clarified, and the authors should distinguish weak evidence from statistically well-supported trends.

There is also a large multiple-testing problem given the number of stations, months, variables, periods, and correlations considered. Some treatment of false discoveries or field significance would strengthen the analysis.

Finally, the authors acknowledge that spatial correlation among neighboring and nested catchments is ignored. This is important because national percentages based on individual gauges can be disproportionately influenced by densely monitored regions and cannot necessarily be interpreted as percentages of independent U.S. catchments. At least a regional aggregation or spatial sensitivity analysis would be appropriate.

Overall, the manuscript assembles an extensive dataset and the large-scale seasonal analysis is potentially useful. Nevertheless, I do not think spatial scale alone provides sufficient novelty for HESS. Mann–Kendall/Theil–Sen trend analyses and correlations with hydroclimatic variables are well-established approaches. The manuscript's stronger potential novelty would be either (i) a genuine joint analysis of streamflow–stream-temperature behavior or (ii) a defensible attribution analysis of their controls, but neither is sufficiently realized by the current methodology.

In particular, the stated motivation of studying streamflow and stream temperature together is only partially addressed, while the attribution claims exceed what can be inferred from bivariate correlations. These are central rather than incremental concerns because they relate directly to the stated scientific contribution of the paper.

For these reasons, despite the considerable effort involved in compiling and analyzing the continental-scale datasets, I do not think the manuscript in its current form warrants publication in *Hydrology and Earth System Sciences*. Addressing these concerns would require substantial reconsideration of the analytical framework and, potentially, of the central framing and novelty of the study.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, *5*(11), eaau4996.This issue is also acknowledged by the authors themselves

Zaerpour, M., Hatami, S., Ballarin, A. S., Papalexiou, S. M., Pietroniro, A., & Nazemi, A. (2025). Agriculture's impact on water–energy balance varies across climates. *Proceedings of the National Academy of Sciences*, *122*(12), e2410521122.

Jalili Pirani, F., & Najafi, M. R. (2022). Multivariate analysis of compound flood hazard across Canada's Atlantic, Pacific and Great Lakes coastal areas. *Earth's Future*, *10*(8), e2022EF002655.

Alizadeh, M. R., Adamowski, J., Nikoo, M. R., AghaKouchak, A., Dennison, P., & Sadegh, M. (2020). A century of observations reveals increasing likelihood of continental-scale compound dry-hot extremes. *Science advances*, *6*(39), eaaz4571.