Answers to Reviewer 1

Thank you very much for this detailed comment and the deep review of our analyses and manuscript. We highly appreciate you taking the time to write such comprehensive feedback and suggesting various ways to improve the quality of the manuscript. Underneath we split your comment by topic and addressed each part separately. The respective parts of the reviewers comment are marked bold, our responses italic.

Reviewer 1 Answers:
First, it is unclear why the authors chose the PCMCI algorithm in particular. As the authors discuss, there exist numerous techniques for analyzing causal relationships between X and Y in multivariate time-evolving contexts, but there is little justification for why this particular algorithm is best-positioned to address questions about the snowmelt-streamflow relationship vis-a-vis other simpler, more transparent methods. As part of algorithm selection, I would expect the testing of multiple methods and extensive validation that demonstrate their applicability to the system in question (see, for example, Docquier et al., 2024).

The choice of PCMCI+ was motivated by the multivariate structure of the problem: the method is specifically designed to handle many simultaneous actors while controlling for indirect links and common drivers, which simpler bivariate methods cannot adequately address. Docquier et al. (2024), the study cited by the reviewer, systematically benchmarked PCMCI against an alternative causal method (Liang-Kleeman information flow (LKIF)) across models of increasing complexity and explicitly concluded that "PCMCI is best with a larger number of variables." We consider this an independent confirmation that our choice is appropriate for the setting at hand.

We will revise the manuscript to make the methodological choices more transparent. In our view, a formal multi-method comparison goes beyond the scope of this large-scale exploratory study and we consider it a valuable direction for follow-up work. We note that both the PCMCI+ algorithm and the causal effect estimation step are flexible with respect to parameterization, and our choices were guided by systematic development testing. For PCMCI+, temporal resolutions from weekly to biweekly aggregation were explored, and tau_max = 12 biweekly steps (equivalent to a 6-month lag window) was selected as appropriate for capturing the seasonal snowmelt–discharge relationship. For the causal effect estimation, a seasonally stratified, time-of-year-appropriate intervention value per biweekly step was found to be more appropriate than a global average.

The authors write:

To confirm the basic functionality of the method, we carry out a set of sanity checks. We introduce modifications to selected variables (snow or rain) and examine whether the detected effects (on river discharge) changed in the expected qualitative manner. These checks serve as sanity tests rather than a sensitivity analysis." (ll. 196-199)

I would argue that these are not simple sanity checks, but absolutely necessary validation exercises that should be formalized and a fundamental part of the manuscript. Not all methods are applicable to all systems, and when applying a method to a novel system, as is done here, one needs to rigorously demonstrate that it is appropriate and identify the contexts in which it does and does not perform well.

We acknowledge the concern of the reviewer that what we designated as "sanity checks" should be formalized. We will include these tests as an additional figure with the relevant explanation in the revised version of the manuscript.

There are a few reason to believe that the chosen analytical framework may not, in fact, be particularly appropriate. First and foremost is the stationarity requirement for the PCMCI algorithm. As the authors note (ll. 150-155) and reference extensively from the literature—and indeed what motivates the study in large part—snowpack, precipitation, and streamflow all exhibit a high degree of nonstationarity in many regions of the world. Despite acknowledging this, there is no detrending of the variables prior to analysis—the rolling mean applied does not address the non-stationarity issue—which means that there is a high likelihood of identifying spurious causal relationships in strongly-trending variables.

We agree that stationarity is a genuine concern and that the variables in question exhibit trends in many regions. However, our analytical framework addresses this in two ways.

First, the inclusion of a time index as an explicit variable in the causal graph relaxes the assumption of stationarity and means that we only assume stationarity from year to year rather than across the full 43-year record, which directly addresses seasonal non-stationarity without removing it from the data. Second, the rolling window analysis further restricts each estimation to a shorter temporal period (20 years), reducing the influence of long-term trends on causal effect estimates.

We deliberately chose not to detrend the variables prior to analysis because detrending risks biasing causal effect estimates in non-linear environments by removing variance that may be causally relevant. Our approach instead reduces the effective sample size and thereby increases variance of the estimates. This trade-off is inherent to any analytical method applied to trending hydroclimatic time series, and we consider our approach appropriate for our study.

Additionally, the authors note, when discussing the identification of the effects of rain versus snowmelt, that the former is more identifiable because "…rain is nearly detectable all year round. In contrast, snowmelt signal for only occurs in spring and early summer, and is thus detectable in less data points" (ll. 594-595). If the goal of the study is to analyze the snowmelt-streamflow relationship, it seems troubling that the strong seasonality of snow accumulation and melt, which occurs during only a few months of the year, is a complicating factor for causal inference. One would desire a causal inference method that leverages that information, rather than being complicated by it.

We acknowledge that this sentence in the limitation section was not well formulated and did not clearly convey what we intended. The sentence refers to the nature of rain and snowmelt as signals rather than to a shortcoming of the causal inference framework itself. Snowmelt occurs only during a limited period of the year and, at our biweekly resolution, yields approximately 4–8 data points per year per basin. Rainfall, by contrast, occurs throughout the year and yields up to 26 data points per year, providing a considerably larger sample for PCMCI+ to leverage. This difference in effective sample size influences the statistical power to detect causal links, which is why the snowmelt signal is more difficult to pick up than the rainfall signal. As this limitation is due to the physical nature of the variables, any causal method that is informed by the data would face this issue.

We note that PCMCI+ allows causal links to be specified manually when physically justified, which would partially address this limitation. In this study, we made use of physical background knowledge to constrain the directionality of potential links, restricting the search space to physically plausible causal structures. However, we deliberately chose not to insist on the

presence of any specific link. This reflects the exploratory objective of the study: prescribing the existence of a snowmelt→discharge link would undermine the evaluation of the framework's capacity to detect it from data. We further note that where PCMCI+ does not detect a link for a given basin and period, the causal effect estimation will return zero. This may in part reflect the limited effective sample size of the snowmelt signal, as discussed above, rather than a genuine absence of a causal relationship. We will clarify this reasoning in the revised manuscript.

Even assuming that the PCMCI is a technically appropriate causal inference algorithm, the results it produces are quite opaque and difficult to interpret in terms of what relationships it is identifying and over what time-scales. For instance, all else equal, enhanced wintertime snowmelt will increase cold-season streamflow, but reduce warm-season streamflow since a smaller snowpack will persist into those months. As such, there is a short-run positive relationship and a longer-run negative relationship. It is not at all clear how these competing effects are combined into the single "causal effect" the authors report, nor why that is a particularly meaningful measure. Hydrologists and water resource managers typically care about streamflow volume and temporal distribution. An analysis that placed the results in those terms would be much more readily interpretable and useful.

The causal effect reported in this study represents the sum of path coefficients across all lags, and thus reflects the net effect of a unit change in snowmelt on annual streamflow integrated over the full lag structure. This definition captures both the short-run positive and the longer-run negative effects described by the reviewer, combining them into a single interpretable quantity: the overall change in yearly streamflow volume attributable to snowmelt variability.

We acknowledge that this was not sufficiently explained in the original manuscript and that greater transparency regarding the causal effect definition will benefit readers from a hydrological background. We will add a dedicated methodological clarification in the revised manuscript, including an illustrative example for a single basin that will walk through the causal graph, the lag structure, and the resulting effect estimate. We further acknowledge that the temporal distribution of streamflow, which is of direct relevance to hydrologists and water resource managers, is not explicitly addressed in this study. We consider this an important direction for future work, and note that the present analysis is intended as a first indication of the large-scale causal role of snowmelt in driving discharge variability.

Figure 10, for example, is quite inscrutable. How should we interpret the ratio? Based on my reading of the text and figure caption, it would seem that values >1 indicate that the causal effect or a perturbation in snowmelt/rain is greater than annual mean discharge? Given that each trigger only accounts for a subset of streamflow, this seems to defy reasonable expectation. It is once again also unclear how the changing nature of the relationships over the course of the seasonal cycle is collapsed into a single annual value. Additionally, throughout the analysis, it seems that the values presented correspond to calendar years, rather than the water years (October-September or less commonly September-August) snow hydrologists typically use, as they better correspond with the seasonal hydrologic cycle in the Northern Hemisphere mid- and high-latitudes.

Regarding the normalization and interpretation of the causal effect values in Figure 10, we address this point in detail in response to a related comment below. We agree that values >1 are hard to interpret, and thank the reviewer for pointing this out. We believe these values were due to computational noise from using an intervention outside the data support. In the comment below, we describe a revised intervention design comparing $E[Y|do(X = \mu)]$ to $E[Y|do(X = \mu - \sigma)]$. This revision directly resolves the interpretability issue because, as both interventions remain within the support of the observed data, values exceeding 1 are physically implausible under the new framework, rendering additional normalization unnecessary.

Regarding the use of calendar years versus water years: the annual values shown in the analysis are averages of rolling window estimates across all 20-year windows that cover a given year. Each window contains 520 biweekly time steps. The difference between a calendar year and a hydrological year boundary corresponds to a shift of approximately 6 biweekly steps at each end of the window, affecting roughly 2% of the data included in any given estimate. Its influence on the overall results can therefore be considered negligible.

My last concern about the overarching methodological framework is about the use of the random forest model for "causal effect" estimation. Random forests are fundamentally predictive models that learn predictive associations between variables, not causal models. [...] The manuscript should either clarify the assumptions under which the causal interpretation is valid or revise the interpretation accordingly.

We acknowledge that the terminology used in the manuscript may have caused confusion regarding the role of the random forest model in our framework. However, the random forest model is not used as a causal model but as a non-parametric parametrization of the conditional expectation E[Y|X, Z], where Z denotes the adjustment set. The causal effect is identified via the adjustment formula E[Y|do(X)] = \sum_Z E[Y|X,Z]E[Z], a standard approach in causal inference. The random forest is used solely to estimate this conditional expectation from finite data, and the perturbation experiment computes the contrast $E[Y|do(X = x_1)] - E[Y|do(X = x_0)]$ under this identified quantity, not a model sensitivity measure. We will add a dedicated figure to the revised manuscript showing the causal graph and the explicit adjustment formula.

Additionally, in the perturbation experiments, the motivation behind the different intervention experiments is not apparent. [...] the 'zero' interventions are both unphysical and likely well beyond the support of the data [...]. The authors also note that they intervene the calendar week index variable (ll. 244-245). This intervention makes little sense; it is a necessary statistical control to account for seasonality.

We agree that the zero intervention is likely outside the support of the data, particularly under a non-parametric estimator, and that values exceeding 1 in Figure 10 are consistent with this issue. We will replace the zero intervention with a statistically motivated intervention comparing $E[Y|do(X = \mu)]$ to $E[Y|do(X = \mu - \sigma)]$, where μ and $\mu - \sigma$ represent typical and reduced snowmelt conditions, respectively. This contrast reflects a meaningful scenario consistent with projected reductions in snowpack under climate change, while remaining within the support of the observed data. Both interventions are defined relative to the within-window distribution, ensuring comparability across basins and time periods.

Regarding the calendar week index: this variable is fully exogenous by construction, as no other variable in the system influences it. Including it as a condition in the analysis is a technical requirement of the algorithm's implementation for handling seasonal non-stationarity, not a causal intervention in the conventional sense. Its inclusion does not introduce a problematic causal interpretation. We will clarify this in the revised manuscript.

On the whole, the claim that the findings "highlight the power of causal inference over conventional statistical measures in enhancing the analysis of large-scale snow hydrological regimes by adding depth to existing approaches" (ll. 15-17) does not pass muster, as there is no comparison to conventional statistical measures and the added value of the casual inference technique chosen is not made readily apparent.

We acknowledge that the phrasing in ll. 15-17 overstates the comparative advantage of causal inference, as no formal comparison to conventional statistical methods was conducted in this study. The study is designed as an exploratory methodological contribution, as stated in Section 4.2. We propose to revise the sentence to: "These findings demonstrate the potential of causal inference as a complement to conventional statistical measures in the analysis of large-scale snow hydrological regimes.

My second major concern is about some apparent statistical errors in the application of the rolling window and calculation of trends. The authors apply a 20-year rolling window—it is unclear whether it is a centered window, or left- or right-justified, but I am assuming centered—to 43 years of data, then average the effects by year to produce a 43-year time series. This means that the edge years are estimated with a much smaller sample size than the central years, which has the potential to distort the results. The common practice when applying moving averages is to only retain years fully covered by the window to avoid such inhomogeneities. Additionally, testing the sensitivity of results to windows of different widths, which will have different tradeoffs between signal identification and the length of time series able to be analyzed, would be a valuable robustness check.

The rolling window method introduces very strong temporal autocorrelation, as adjacent years now share 19/20 of their data. This serves to dramatically reduce the variability of the time series, artificially inflating the Mann-Kendall test statistic and overstating statistical significance.

We agree that the rolling window approach results in reduced sample sizes at the temporal edges of the time series. This is explicitly discussed in Section 3.5, where a Wilcoxon signed-rank test confirms that a significant portion of dispersion originates from these edges. This is a primary reason we present trends rather than absolute values. Regarding window width: we tested multiple window sizes during development and selected 20 years as a balance between providing sufficient data for reliable causal effect estimation within PCMCI+ and limiting the degree of temporal autocorrelation introduced by overlapping windows. A 30-year window, for instance, would result in 29/30 overlapping years between adjacent windows. Regarding trend significance: we acknowledge that the rolling window introduces strong temporal autocorrelation, which can inflate the Mann-Kendall test statistic. We propose to address this by replacing the standard Mann-Kendall test with the autocorrelation-corrected version of Hamed and Rao (1998), and will update the significance statements accordingly.

The authors rely on a single observational or reanalysis dataset for each variable in their model. There are, however, large numbers of data products for each, and particularly for hydroclimatic variables such as SWE and rainfall, the observational uncertainty is quite large and different datasets may not agree with one another on magnitude, variability, or even the direction of long-term trends of the variable of interest (Gottlieb and Mankin, 2024; Mortimer et al., 2020; Mudryk et al., 2025). The authors claim in multiple places that no other independent validation data exist, but ensuring that the findings are consistent regardless of the particular data product chosen is itself an important validation. Additionally, there are multiple independent streamflow datasets used in large-sample hydrology that are based on actual streamflow measurements, rather than a reanalysis like GloFAS, e.g., the data compiled by Han et al., (2024) or the GRDC-Caravan dataset (Färber et al., 2025).

The choice of GloFAS as the streamflow dataset was motivated by its consistent spatial and temporal coverage across all 119 FAO basins over the full 1980–2022 period. Station-based datasets such as GRDC-Caravan or the compilation by Han et al. (2024), while valuable, tend to have uneven spatial coverage at the hemispheric scale, which may limit their applicability as a direct substitute or comparison across the full study domain.

We note that our claim of no independent validation data refers specifically to the causal effect estimates themselves, not to the individual input variables. To our knowledge, no observational benchmark exists against which hemispheric-scale snowmelt-streamflow causal effects can be directly validated, which is precisely what motivates the exploratory framing of this study. We agree that consistency across data products is a valuable form of validation and consider a multi-product sensitivity analysis an important direction for future work, but regard it as beyond the scope of the current study given the computational demands of the pipeline.

Relatedly, given that one of the key variables in the study is snowmelt, it is unclear from the Data and Methods sections what SWE data are actually used. Figure 3 suggests ERA5, Section 2.1 mentions SnowCCI and both ERA5 and ERA5-Land (which are very different data products when it comes to snowpack), and Table 1 mentions SnowCCI gap-filled with ERA5, and also ERA5-Land. Consistency and clarity here, including details on the gap-filling procedure are necessary.

We acknowledge that references to SnowCCI and ERA5 in various parts of the manuscript are remnants of earlier iterations and do not reflect the data actually used. ERA5-Land was the sole SWE dataset used throughout the analysis, and we will correct all inconsistent references in the revised manuscript, including Figure 3 and Table 1.

The choice of ERA5-Land was motivated by its performance in independent benchmarking studies. Mudryk et al. (2025), the SnowPEx+ intercomparison cited by the reviewer, identified ERA5-Land as the top-performing SWE product across a comprehensive range of evaluation metrics, supporting its use as the primary dataset for a hemispheric-scale study. We will add this justification explicitly to the Data section of the revised manuscript.

l. 34-35: I do not believe either of these studies assesses changes in the interannual variability of SWE over time.

We agree that neither of the cited studies directly assesses trends in interannual SWE variability over time, and we acknowledge that this claim was not adequately supported by the cited literature. We will remove this statement from the revised manuscript.

Figure 1 and the associated text are a nice overview, but not particularly germane to the study, which is focused more narrowly on streamflow.

Thank you very much for this remark. We would like to keep the figure as an introduction to frame the importance of snowmelt in various systems.

l. 53: I am assuming this is referring to the ratio of total annual snowmelt to total precipitation. Throughout the manuscript, SWE and snowmelt are interchanged quite a bit. Most of the analysis, it seems, focuses on snowmelt (as measured by negative changes in SWE), so that term should be used to avoid confusion.

We acknowledge the confusing use of snowmelt and SWE and we will revise the manuscript to make sure that both terms are used adequately.

l. 64: I would argue Han et al. (2024) is a study that examines the relationship between snow and river discharge on a hemispheric scale.

We thank the reviewer for pointing out this omission. Han et al. (2024) is indeed a hemispheric-scale study linking snow dynamics to streamflow seasonality and we will add this reference in the revised manuscript accordingly.

ll. 260-261: The sentence "The normalized Theil-Sen trends of all variables in the analyzed basins were computed using the basin mean and the maximum absolute value for each variable" is quite confusing. How are both the mean and maximum used in normalization?

The sentence was poorly worded and referred to two separate steps that were conflated. The Theil-Sen trends were computed on causal effects that had already been normalised by mean river discharge per basin and period, ensuring comparability across basins. For visualisation purposes, the resulting slopes were additionally normalised by the maximum absolute slope across all basins, mapping the most extreme basin to ±1. We will rewrite this sentence clearly in the revised manuscript.

Figure 5: The time scale on which these trends are being calculated is unclear. Presumably annual values over the 1980-2022 period?

The trends are indeed computed from annual values over the 1980–2022 period. We will update the caption accordingly in the revised manuscript.

Section 3.4: Some of the results for each basin regarding the mechanisms driving different changes in the causal influences of snowmelt and rain on discharge feel overinterpreted and unsupported by quantitative results.

Section 4.1: Much of the discussion around specific basins feels redundant with section 3.4 and could be tightened considerably.

We thank the reviewer for both comments and the suggestions regarding the basin analyses. We will restructure both sections in the revised manuscript: Section 3.4 will be revised to present the causal effect trends descriptively, reporting only what the data directly show without mechanistic interpretation. The mechanistic discussion and contextualization with existing literature will be consolidated in Section 4.1, where interpretations that go beyond the direct results will be explicitly framed as hypotheses.
This restructuring will reduce redundancy and more clearly separate results from interpretation.

Figure 10: Significant at what confidence level? Also, see Major Comment 2 about data smoothing and significance testing.

Significance is assessed at the 95% confidence level ($p < 0.05$) using the Mann-Kendall test. As described in our response to Major Comment 2, we will replace this with the Hamed-Rao corrected Mann-Kendall test in the revised manuscript, and will update the figure caption accordingly.

Section 3.5: I am having a difficult time understanding what is being tested here and why, and how it represents a validation of the method.

We acknowledge that the purpose of Section 3.5 was not sufficiently explained. As stated at the beginning of the section, no independent validation data exist for the causal effect estimates themselves. Section 3.5 therefore does not validate the absolute values of the causal effects, but assesses the internal consistency of the method: specifically, whether the estimated effects are stable within individual basins over time or dominated by noise. A low coefficient of variation after trend correction indicates that the residual variability around the estimated trend is small relative to the mean, which we interpret as evidence that the method produces consistent rather than erratic results. We will revise the section introduction in the revised manuscript to make this distinction and the purpose of the analysis explicit.

We hope that our responses could address all points raised and that our proposed revisions are in line with the intentions of the reviewer. We thank you again for your time and the comprehensive review.

Answers to Reviewers 2

We thank the Reviewers for the thorough and constructive review of our manuscript. The detailed feedback and concrete suggestions have helped us identify several opportunities to improve clarity and presentation. In the following, we address each point individually, structured according to the topics raised in the review. The respective parts of the reviewers comment are marked bold, our responses italic.

The authors list all their data sources starting in section 2.1. However, the authors do not address the different spatial resolutions of each dataset. The authors do partially address this matter in section 2.3.3 with the MERRA-2 data. Highlighting these differences might help make the point in Section 2.3 more clear why the proportional approach is clearly needed. Additionally, having a last access date will be helpful.

We will add a column specifying the spatial resolution of each dataset to Table 1 and include last access dates for all data sources. However we wanted to outline why it was not included in the first version. The spatial aggregation of all variables to the basin scale, which encompasses areas of up to several million km^2, differences in native resolution play a minor role in the presented analyses.

Table 1: For Snow, it's stated that SnowCCI v3.1 was used alongside ERA 5 Land. In the remarks column, it says ERA 5 was used to fill data gaps. Then on line 125 it says "ECWMF reanalysis land (ERA 5 Land/ERA 5). Are the authors using ERA5 and ERA 5 land interchangeably? Aren't they two different products? Also, since it states "PMW and in-situ", I'm guessing the SnowCCI dataset is a mixture of remotely sensed and in situ measurements of SWE. This is not clear in the manuscript. Also, how is the gap-filling conducted?

The references to SnowCCI in Table 1, Section 2.1, and Figure 3 are remnants of a previous version of the manuscript. The analysis was conducted exclusively using ERA5-Land SWE data. We apologize for the confusion and will correct all affected sections to consistently refer to ERA5-Land as the sole SWE data source, removing all references to SnowCCI and the associated gap-filling procedure.

Line 136: What is the justification to choose variables from ERA5 and ERA5 Land? Why didn't the authors stick with just ERA5 or ERA5 land for each variable?

Where possible, we used ERA5-Land as it is specifically designed for land surface applications. The sole exception is rain, which was derived from ERA5 because the precipitation type variable (used to distinguish liquid from solid precipitation) is only available in ERA5 and not in ERA5-Land. We will clarify this in Section 2.2.

Section 2.3.2, the authors state that snow is subtracted from the ERA5 total precipitation to generate a liquid precipitation data. However, it is unclear what is being subtracted. Is it snowfall or the SWE data?

As described in Section 2.3.2, liquid precipitation was derived by applying the precipitation type variable from ERA5, which classifies precipitation as frozen or liquid on an hourly basis. Only the liquid fraction was retained, and the resulting dataset was subsequently averaged to biweekly resolution. No SWE data were subtracted. We will revise the wording in Section 2.3.2 to make this distinction clearer.

Section 2.4: The authors computed a range of statistical measures (mean, median, variance, and standard deviation) and applied a Theil-Sen estimator and Mann-Kendall test. However, it is unclear what they are taking a mean or median of. Is it a yearly average for each year in the 43 year period between 1980 and 2022? Do the authors then do the Theil-Sen and Mann-Kendall test? In addition, in Section 3.1 the authors state normalized Theil-Sen trends were computed using both the mean and the maximum absolute value, but do not state the latter in section 2.4.

We acknowledge that these calculations were confusingly stated in the manuscript. The statistical measures (mean, median, variance, standard deviation) were computed over the full biweekly time series per basin and variable across the entire 1980–2022 period. The Theil-Sen slopes were computed on causal effects already normalised by mean river discharge per basin and period, ensuring comparability across basins. For visualisation purposes, the resulting slopes were additionally normalised by the maximum absolute slope across all basins, mapping the most extreme basin to ±1. We will clarify this two-step procedure explicitly in Section 2.4 of the revised manuscript.

Line 189: The authors switch from using PCMCI to using PCMCI+ without giving a description as to what PCMCI+ is and how it is different from PCMCI. I'm guessing that PCMCI+ is a specific type of algorithm. Also, if the authors are using a software package to run their PCMCI+, it would be nice to know which package they used and see the software package cited in the manuscript.

PCMCI+ is an extension of PCMCI that additionally allows for the detection of contemporaneous causal links, i.e. causal relationships occurring within the same time step, in addition to lagged relationships. This distinction is relevant for our setup as some hydrological interactions may occur at the biweekly timescale. The analysis was conducted using the Tigramite Python package (Runge, J., et al., Tigramite, https://github.com/jakobrunge/tigramite), which implements both PCMCI and PCMCI+. We will add a clarifying sentence to Section 2.5.1 and cite the software package accordingly.

Section 2.5: On line 196-197, the authors mention the use of sanity checks through modifications. What are these modifications? If these modifications are highlighted in a different section, this should be clearly highlighted in the text here. Also, have the authors considered doing the analysis on surrogate data that randomizes the time series but retains seasonal trends?

The modifications used in the sanity checks consist of two mirrored scenarios applied to the input time series: in the first, a compounding scaling factor is applied to artificially increase the snowmelt signal while simultaneously decreasing the rainfall signal, and in the second, the opposite scaling is applied. The causal effects detected by the pipeline changed in the expected qualitative direction in both cases, confirming the basic functionality of the framework. As noted in our response to Reviewer 1, we will formalize these tests as an additional figure with a dedicated explanation in the revised manuscript, and will add a cross-reference to that section here.

Section 2.5.1. In line 156, the authors claimed their approach to avoid seasonality and shifts of seasonal patterns were described in more detail. I am still somewhat confused as to how this was considered outside of the calendar week being an index variable. Did this actually remove the non-stationarity problem? Aren't there statistical tests to prove a time series is stationary?

Our framework addresses non-stationarity without modifying the data in two ways: (1) we address seasonal non-stationarity through the inclusion of a calendar week index as an explicit variable in the causal graph, which relaxes the stationarity assumption to year-to-year rather than across the full record, and (2) we address long-term trends through the rolling window approach, which restricts each estimation to a 20-year period where the trend is small enough to be ignored. We confirmed this using Theil-Sen and Mann-Kendall trend analysis across all 20-year moving windows: even for variables with a significant long-term trend over the full 43-year record, the vast majority of individual 20-year windows showed no statistically significant trend, and even where significant trends were detected, they were confined to specific sub-periods rather than present throughout the record. Since we address non-stationarity within the method instead of modifying the data, most formal stationarity tests are not useful or needed in this context. We will discuss this point more explicitly in the revised manuscript and welcome further guidance from the reviewer on how stationarity could be more rigorously addressed within the constraints of the current framework.

Section 2.5.2: I am also confused at how the random forest estimator was applied. Figure 4 goes into slightly more detail saying the RF estimator is fit using input data using the optimal adjustment set identified from the causal graphs, but it is still unclear what is being predicted or estimated.

The random forest is used as a non-parametric estimator of the conditional expectation E[Y|X, Z], where Y is river discharge, X is the trigger variable (snowmelt or rain), and Z is the adjustment set identified from the causal graph. The random forest is trained to predict river discharge given the trigger variable and all adjustment variables simultaneously. The causal effect is then computed as the contrast E[Y|do(X = μ)] - E[Y|do(X = μ - σ)] by evaluating the fitted model (E[Y|do(X)] = \sum_Z E[Y|X,Z]E[Z]) at two different values of X. We will add a dedicated figure to the revised manuscript showing the causal graph, the adjustment set, and the intervention design, which will make this procedure explicit.

Section 3: I wonder if section 3 can be improved by presenting the results in a slightly different order. Since this is a paper about causal statistics, I think presenting the causal results (Sect. 3.2) first would highlight its importance to the analysis. This can be followed by the results of Sect. 3.1 as explainers for the causal results.

We considered reordering the sections but we would like to to retain the current structure, in which the descriptive variable trends (Section 3.1) precede the causal results (Section 3.2). We believe this order benefits the reader by providing the necessary hydroclimatic context before the causal effects are introduced and interpreted. Presenting the causal results first would require the reader to interpret them without this contextual foundation.

In addition, my understanding is that casual graphs are an important part of PCMCI, especially for possible feature selection. Why didn't the authors show the casual graphs or include them in supplementary materials for discussion for some of the exemplary basins in section 3.4? I think showing these causal graphs and linking them back to the literature can help answer whether these causal graphs can produce reasonable results and help differentiate sections 3.4 and 4.1.

Given the complexity of our setup (8 variables, 12 time lags), the resulting causal graphs contain a large number of potential links and are difficult to present in a publication-quality figure. We therefore decided not to include the full causal graphs in the manuscript or supplementary materials. However, we will add a dedicated figure to the revised manuscript showing the causal graph structure, the adjustment set, and the explicit intervention design for an illustrative basin, which will address the reviewer's concern regarding transparency and interpretability.

The authors seem to have included maps with temporal trends of the snowmelt-to-discharge and rain-to-discharge causal effects, so we can see how they are changing over time, but they don't show maps with the actual magnitude of the causal effects (e.g., how does the strength of the causal effect in the Arctic compare to the western U.S., for instance)? They do this for select basins, normalized to river discharge (e.g. Figure 10), and present the results in tables in Appendix B, but it would be helpful to see it in a map.

As discussed in Section 4.2, absolute causal effect values are dependent on the method's detection capacity within each basin and are therefore not directly comparable across basins. For this reason we deliberately refrained from presenting absolute values as a map, as we are concerned that such a visualization would invite cross-basin comparisons that the data do not support. We believe the current combination of trend maps, exemplary basin time series, and tabulated values in Appendix B provides the most appropriate representation of the results given this constraint.

For Figure 10, couldn't the trend in river discharge result in increasing or decreasing trends, even if the strength of the causal effect remains constant over time? If so, then why normalize by discharge, which could make the plots harder to interpret? Or do the authors normalize by a single mean discharge value calculated from the entire times series?

We acknowledge that normalizing by a declining discharge trend could in principle introduce an artefact in the ratio. However, as described in our response to the related comment above, we will replace the zero intervention with a physically motivated intervention comparing E[Y\|do(X = μ)] to E[Y\|do(X = μ - σ)]. This revision renders the normalization question largely moot, as the resulting contrast is expressed directly in discharge units, remains within the data support, and does not produce values exceeding 1 under physical conditions.

Figure A1: Why did they specify that evapotranspiration has no causal link with river discharge? It seems to me that an increase in ET could definitely decrease streamflow.

The evapotranspiration variable used in this study intentionally excludes several components of total land evaporation. Open-water evaporation was excluded as it operates directly at the river surface rather than through terrestrial pathways. Snow sublimation was excluded as it would contradict the snowmelt definition used in this study. Within this definition, evapotranspiration influences river discharge indirectly through soil moisture depletion, which is explicitly included as a mediating variable in the causal graph. A direct link from transpiration to river discharge was therefore not specified. We will clarify the variable definition explicitly in the Methods section of the revised manuscript.

Paragraph staring on Line 52 ("A brief overview of the role…") and Figure 2 feels very out of place where it is placed. I don't think it hurts the introduction if the paragraph and figure are removed from the introduction.

We would like to retain Figure 2 and the associated paragraph as it serves a dual purpose: it provides context for the role of snow across the analyzed basins and functions as the primary study area overview, as no separate study area figure is included in the manuscript. We will however revise the paragraph and its placement within the introduction to improve the flow of the section as suggested.

Lines 124 to 131: These sentences can use refinement and reordering to make it more clear and justify why the time period was chosen and how the 199 basins were selected.

Line 136: The "(e.g. evapotranspiration datasets)" does not clarify the authors previous sentence on line 135.

Line 136: Manuscripts says, "For rain and temperature we use the ERA 5 and ERA 5 Land data respectively". However, on Fig. 3, Temperature is listed as ERA5. Also, why didn't the authors choose precipitation for ERA 5 Land?

We will revise and reorder lines 124–131 in the revised manuscript to more clearly present the basin selection criteria and the justification for the study period. The 119 basins were selected based on a minimum of 30 years of snow coverage over 1980–2022, and the study period was

chosen based on the availability of the soil moisture dataset, which relies on passive microwave satellite data available from 1980 onward. We will make this reasoning more explicit.

Regarding line 135: we will remove the example in parentheses as it does not add clarity to the preceding sentence.

Regarding line 136: we acknowledge the inconsistency. ERA5-Land was used for temperature, and ERA5 was used for precipitation, as ERA5-Land does not provide a precipitation type variable, which was required to separate liquid from solid precipitation. The figure caption will be corrected accordingly in the revised manuscript.

Figure 10: What does the black dashed line represent in the graphs? I don't think it was addressed what these plotted in the manuscript or the caption. Additionally, x-axis labelled have a typo: "Riverdischarge".

The black dashed line represents the Theil-Sen trend line fitted to the annual causal effect values. We will add this information to the figure caption in the revised manuscript and correct the typo in the x-axis label.

Line 474: What is the author referring to with "(5)"?

The "(5)" is a reference to Figure 5, which shows the declining trends in soil moisture and baseflow flux in the Colorado basin. This was incorrectly formatted and will be corrected to "Figure 5" in the revised manuscript.

All other comments will be addressed as the reviewers suggested.

We hope that our responses could answer all open questions and that our proposed revisions meet the intentions of the reviewers. We thank you again for your time and the thorough review.

Changes made according to the comments

Global changes (entire document)

- Terminology “PCMCI” consistently replaced with “PCMCI+” wherever the PCMCI+ algorithm specifically is meant (Abstract, Introduction, Methods, Discussion). Triggered by RC2's question on line 189 (“authors switch from PCMCI to PCMCI+ without giving a description… which software package”).
- Significance test for trend calculation changed throughout from Mann-Kendall to the Hamed-Rao test (autocorrelation-corrected). Direct response to RC1 Major Comment 2 (rolling window inflates Mann-Kendall significance).
- Intervention design fundamentally changed: from a zero intervention to the contrast do(X=μ) vs. do(X=μ−σ). Responds to RC1 Major Comment 1.3 (zero interventions unphysical/out of data support) and to RC2's Figure 10 question (“couldn't the trend in river discharge result in increasing/decreasing trends… why normalize by discharge.
- New decomposition of the causal effect into “anomaly magnitude” (σ) and “coupling strength” (β = CE/σ). Not a literal request from either reviewer; introduced by the authors to address the interpretability concerns behind RC1 Major Comment 1.2.
- As a consequence of the changed intervention, essentially all quantitative results changed in Results, Discussion, and Conclusion.
- SWE data source unified to ERA5-Land, all SnowCCI references removed Responds to RC1 Major Comment 3 (SnowCCI/ERA5/ERA5-Land inconsistency across Fig. 3, Section 2.1, Table 1) and to RC2's separate Table 1 question about the same inconsistency.
- Two new figures added: Figure 4 (sanity check) and Figure 6 (causal effect estimation illustration), see the relevant sections below for attribution.

Section-by-section changes

Abstract

- Closing sentence toned down (removed the unsupported claim of “highlighting the power of causal inference over conventional statistical measures”) — matches RC1 Major Comment 1's closing paragraph (“the claim… does not pass muster, as there is no comparison to conventional statistical measures”).
- Altered statements regarding the results (adjusted to the new variability intervention; see global changes)

1.1 Snow as a river flow driver

- Unsupported statement on interannual SWE variability removed — matches RC1 minor comment on l. 34–35.
- Figure 2 and corresponding references: “SWE-to-precipitation ratio” → “snowmelt-to-precipitation ratio” — matches RC1 minor comment on l. 53 (SWE/snowmelt used interchangeably; snowmelt term should be used for clarity).
- Lines 60ff: Han et al. (2024) added as a hemispheric snow–streamflow study; research-gap phrasing sharpened — matches RC1 minor comment on l. 64 (“I would argue Han et al. (2024) is a study that examines the relationship between snow and river discharge on a hemispheric scale”).

1.2 Background on Causal Inference in Environmental Science

- PCMCI+ explicitly defined as an extension of PCMCI, with the Runge et al. (2020) citation and the distinction between the two direct response to RC2's line 189 question.

- Description of the included variables clarified (snowmelt rather than SWE as the causal-analysis input). Response to both RC1 and RC2 comments regarding clarity in the use of SWE/snowmelt.
- New paragraph justifying PCMCI+ over alternative approach. In response to Major Comment 1.1.

2.1 Studied Basins and Period

- SnowCCI removed as a data source, basin selection based solely on ERA5-Land. RC1 Major Comment 3 and RC2's Table 1 question.
- Wording of the selection criterion sharpened ("at least 30 of these 43 years"). Response to RC2 comment regarding basin selection.

Justification for the 1980–2022 period added (l. 143) and the selected basins (l.138). Matches RC2's comment on preprint lines 124–131 ("sentences can use refinement… justify why the time period was chosen and how the 119 basins were selected").

2.2 Data

- Added justification on snow product selection (l. 150)
- ERA5 vs. ERA5-Land variable choice now explicitly justified (l. 153). Matches RC2's question on preprint line 136 ("What is the justification to choose variables from ERA5 and ERA5 Land?").
- Evapotranspiration exclusions (open-water evaporation, snow sublimation) justified (l. 159). Matches RC2's Figure A1 question ("Why did they specify that evapotranspiration has no causal link with river discharge?").
- Table 1, SWE/Snowmelt row corrected (SnowCCI removed, source unified). Response to RC1 Major Comment 3 and RC2's Table 1 question.

2.3.1 SWE / Snowmelt & 2.3.2 Precipitation / Rain

- 2.3.1: clarified that the snowmelt conversion applies only to the causal analysis, SWE used directly for the trend analysis. General clarification of the SWE/snowmelt distinction raised in RC1's l. 53 comment.
- 2.3.2: rain-calculation description corrected (no longer "subtracted from precipitation", now derived from the ERA5 precipitation-type variable). Matches RC2's repeated question on Section 2.3.2 ("it is unclear what is being subtracted").
- Table 1: Added column with spatial resolution. Response to RC2.

2.4 Statistical Metrics and Trend Calculation

- L. 201 ff Paragraph added to clarify why SWE was used for these analyses SWE.
- Two-step normalization of the Theil-Sen slopes explicitly described (l. 205). Matches both RC1's minor comment on preprint ll. 260–261 ("how are both the mean and maximum used in normalization?") and RC2's Section 2.4 question, which raises the identical point independently.

2.5 Causal Inference (overview)

- Non-stationarity paragraph restructured and expanded (l.226ff). Responds to RC1 Major Comment 1.1 (stationarity requirement of PCMCI+, no detrending) and to RC2's Section 2.5.1 question (“Did this actually remove the non-stationarity problem? Aren't there statistical tests to prove a time series is stationary?”).
- “sanity checks” clarified to “plausibility checks using artificially modified input data” (l. 236ff)— response to RC1 (Major Comment 1.1) and RC2 separately asks “What are these modifications?” (Section 2.5 question).
- New Figure 4 (plausibility check, Danube) added. Answers RC1's request to formalize the sanity checks as “necessary validation exercises” (Major Comment 1.1) and RC2s point (“What are these modifications?”).
- Figure 5 caption updated to the new intervention notation — consequential of the intervention-design change above.

2.5.1 Causal Discovery

- PCMCI+ vs. PCMCI distinction and Tigramite software citation added. Matches RC2's preprint line 189 question (now l. 245f)
- Calendar-week variable clarified as a technical device rather than a substantive causal driver. Matches RC2's Section 2.5.1 question (now l. 251ff).

2.5.2 Causal Effect Estimation

- L. 280ff Random forest reframed explicitly as a non-parametric estimator of $E[Y|X,Z]$, not a causal model; adjustment formula introduced. Matches RC1 Major Comment 1.3 (“random forests are fundamentally predictive models… not causal models”) and RC2's Section 2.5.2 question (“unclear what is being predicted or estimated”).
- Intervention design changed to $do(X=\mu)$ vs. $do(X=\mu-\sigma)$. see global changes above (RC1 1.3 and RC2's Figure 10 question).
- New σ (anomaly magnitude) / β (coupling strength) decomposition introduced. Authors' own addition, not a literal request (L. 305ff).
- New Figure 6 (causal graph, adjustment set, intervention design for an example basin) added Matches RC2's request to show causal graphs/adjustment sets (“why didn't the authors show the causal graphs… help differentiate sections 3.4 and 4.1”) and its Section 2.5.2 question about Figure 4 [old numbering] being unclear.
- L. 315. Window-size justification (20 vs. 30 years) expanded. Matches RC1 Major Comment 2 (“testing the sensitivity of results to windows of different widths… would be a valuable robustness check”).
- Changed from Mann-Kendall to Hamed-Rao. RC1 Major Comment 2.

3.1 Variable Trends

- Figure 7 (formerly figure 5) caption now states the period explicitly (1980–2022). Matches RC1 minor comment (“the time scale on which these trends are being calculated is unclear”).

3.2 North Hemisphere Causal Trends

- All basin counts and significance figures updated. Consequential of the intervention-design change.
- New anomaly-magnitude/coupling-strength paragraph added. Addresses the opacity concern behind RC1 Major Comment 1.2, not a literal request (l. 360ff).
- Figures 9 and 10 redesigned into three panels with Hamed-Rao significance. Consequential of RC1 Major Comment 2 and 1.2.

3.3 Using Causal Effects as Indicators of Water Stress

- Basin counts updated. Consequential of the intervention-design change.

3.4 Exemplary Basins Analysis (all six basins)

- All six basin subsections rewritten with explicit quantitative anomaly/coupling language. Matches RC1 minor comment on Section 3.4.
- Speculative explanations removed and replaced with data-supported argumentation based on the anomaly/coupling decomposition. (Response to RC1)

3.5 Consistency and Fluctuations of Trends

- Introductory paragraph rewritten to better define the section's purpose. Matches RC1 minor comment on Section 3.5 (“I am having a difficult time understanding what is being tested here and why, and how it represents a validation of the method”).

4.1 Contextualization of Results

- Opening sentence sharpened (“validate the causal effect trends”). Ties to RC1's overall validation critique (Major Comment 1.1).
- Added subsections for better overview and readability
- Basin paragraphs rewritten around the anomaly/coupling decomposition. Consequential of the intervention-design change (RC1 1.2/1.3, RC2's Figure 10 comment).
- Section restructured around the anomaly/coupling decomposition, differentiating it from the basin-level detail in Section 3.4 and addressing the redundancy noted between the two sections. (RC1)

4.2 Methodological Discussion

- Rain vs. snowmelt sample-size limitation quantified (26 vs. 4–8 biweekly points/year). Addresses RC1 Major Comment 1.1's seasonality concern, not a literal request (l. 704ff).
- Adjusted intervention discussion part (l. 679ff)
- “PCMCI is calculated per basin” → “PCMCI+ is run per basin”. Terminology, consequential of RC2's line 189 question (l. 728).

5 Conclusion

- Bullet-list numbers updated throughout. Consequential of the intervention-design change.
- Bullet on the six basins rewritten around the coupling-strength pattern. Authors' own addition.
- Final bullet sharpened (anomaly magnitude vs. coupling as the driver). Authors' own addition.

Appendix A — Hydrological System for Causal Discovery

- New explanatory paragraph on the evapotranspiration–discharge link added — matches RC2's Figure A1 question.
- Figure A1 caption expanded accordingly.
- Figures B1-B4 adjusted and extended to the new results and Continent column removed according to RC2 comment.

Appendix B — Full Results

- Updated Figures according to new results
- Added two new columns to display coupling strength and anomaly size
- Added short introductory sentence to match structure with Appendix A.