Response to Reviewer 2

Reviewer comment: https://doi.org/10.5194/hess-2021-274-RC2

Reviewer comments are in black. Author responses are in blue.

The Penny et al. manuscript presents a variety of observations to investigate why the water balance for a Himalayan river has changed. They highlight the limits to a simple water balance calculation and instead include information on land-use, NDVI and water extent as markers for differences in vegetation or crop productivity (i.e., evapotranspiration) and groundwater contributions to surface water, respectively. They focus on the changes between a 15-year (1984-1999) and a 13-year (2000-2013) period, and approach the work as a 'method of multiple hypotheses', testing various hypotheses related to potential drivers of the change in streamflow. I found the manuscript relatively well-structured and I think that it can be a good addition to the literature and a good resource for water managers in the Jhelum river region. However, I do think the manuscript needs some work before it is ready for final publication.

Thank you for the careful consideration and constructive feedback, which we believe will strongly improve the manuscript. We respond to each of your comments below.

From a clarity perspective, I found that some words were not well defined ("secondary data", "drivers", "changing climate") and that the reader would benefit from describing these better. From a structural perspective, I found that the methods and results were not always matching up. How was the baseflow index calculated? How is a storm event characterized? Why is a basin-average correct value for precipitation preferred over a locally correct value? How were statistical analyses performed?

In the revised manuscript, we will clarify each of these terms upon their first use. The approach for calculating for baseflow was previously relegated to the SI, and we will port this to the manuscript. The remaining questions are raised in the minor and specific comments below, and we address them individually under each comment.

What was also clearly missing in the manuscript were uncertainty estimations on the data used to test the various hypotheses. The authors indicate an uncertainty in their initial water balance estimation of 15% (which by the way is more than increase in ET and almost as much as the decrease in streamflow of 117 mm?). An estimation of the uncertainty related to individual components of the water balance would be helpful here – or is it 15% of each component? Uncertainty estimations were given for none of the following analyses… (precipitation estimations from remotely sensed observations, evapotranspiration estimates using crop factors, NDVI estimations when only few (=< 5) images were available, and inferring catchment storage from surface water extent). I expect there to be considerable

uncertainty in each of these estimations, and expect that to be quantified and discussed in the text of the updated manuscript.

Thank you for this comment. Estimating uncertainty represents an important challenge in this study. Before describing how we address this comment, we wish to emphasize that the manuscript is oriented towards improving conceptual understanding of hydrological change. A key objective of the paper is to demonstrate the use of multiple hypotheses as an alternative approach for attribution in situations where data scarcity and uncertainty confounds the application of a model-based predictive inference approach. One important reason we included the water balance was to give a sense of the overall uncertainty with respect to water balance fluxes and therefore inform interpretation of the overall approach. That being said, we completely agree that observation uncertainties should not only be used to justify our approach, but also systematically incorporated into the approach itself.

With these clarifications in mind, we will improve our analysis of uncertainty in three ways, with an emphasis on providing a stronger statistical foundation for the evaluation of the hypotheses.

1. "Testing" each alternative hypothesis implies determining whether the signal is stronger than the noise. We will run statistical tests for hypotheses 1-5, and 8-9. In cases where we have confidence that errors are normally distributed, we will use a t-test and will apply a Wilcoxon signed-rank test otherwise. This will give greater confidence in our evaluation of each individual hypothesis. Hypotheses 6 and 7 deal with glacier and permafrost contributions to the change in streamflow. For these hypotheses we only have upper and lower bounds, and therefore instead of traditional hypothesis testing we will provide upper and lower estimates on the expected changes.
2. We will better describe uncertainty due to spatial variability of precipitation by showing a map of the differences in the two interpolation methods. We will also better describe how using the different regionalization approaches would affect our hypotheses. In particular, although the elevation-gradient interpolation yields a much larger estimate of annual precipitation, the results of the before-after comparison are consistent regardless of which dataset is used.
3. As an independent estimate of the uncertainty associated with each component of the water balance, we will compare our findings to what would be expected from a water balance closure. For example, we will compare our estimate of ET based on NDVI and temperature to its water balance estimate ($P-Q-\Delta S$) and describe how that might change our conclusions.

Ultimately, the benefit of the multiple hypotheses approach is that the findings from some hypotheses can corroborate or cast doubt on the findings from other hypotheses. The fact that the findings from many of the hypotheses lead to conclusions that are consistent gives credence to the findings that we present. With that said, we agree that a more careful

treatment of uncertainty will not only provide greater confidence in the approach but greater context for readers interested in the Upper Jhelum, and we are thankful to the reviewer for this suggestion.

Detailed comments:

Title: What do the authors mean with ‘secondary data’? – being specific would be helpful here to guide the potential reader to reading the article.

We will clarify that ‘secondary data’ refers to datasets collected by other sources -- in this case, mostly governmental agencies. The reason the emphasis on secondary data is important is because collecting primary hydrological data in remote (or transboundary) locations is often impractical, and therefore reliable approaches that build on publicly available data sources can support effective water management.

L28 Instead of pointing out that there are many watersheds where the hydrological drivers are unknown, which is not surprising given the amount of watersheds globally, highlighting prior studies that did identify and quantify drivers would be more helpful.

Thank you -- in response to this comment and comments by the other reviewer, we will condense this part of the introduction to focus on existing approaches and studies for attribution, and note how these studies might inform our attribution study.

L34 “Their associated drivers” which drivers are meant here?

We will clarify that by “associated drivers” we mean, broadly, the “causal drivers of hydrological change”.

L76 “finding”s please add the s

L78 remove “by”

Thanks for catching these grammatical errors and typos.

L124-129 consider removing this text since a standard manuscript format is followed, or adapt such that it contains information that is specific to this manuscript.

We will address these comments in the revised manuscript.

Fig 1 The panels 1d-g are square, but they are not square where they are indicated on map 1b. Did the coordinate system change? Then please indicate the new coordinates along insets d-g. If not, please make sure that the indication and maps match. Also, for me it

would have been helpful to have the line of the river drawn in panels d-g, and the river and catchment boundary are not shown in the legend (I assume light blue and black lines).

We will fix the projection in this figure so that the shapes of the subpanels are consistent and we will also include the main river channels and watershed boundary.

Fig 2 snow density is needed to calculate water content from area and depth, and should be included in this scheme as well as in the calculation.

We will clarify in the text that we do not calculate snow storage volumes, but rather focus on snow areal coverage. This is consistent with hypothesis 8, which states that "Reduced snow cover and earlier snowmelt generated an earlier peak in annual hydrograph." Although estimating snow storage volumes would be beneficial, we believe the benefits would be overshadowed by the various assumptions needed to estimate snow depth (ie, the need to spatially and temporally interpolate daily snowfall depths and daily temperatures, and model snowmelt and sublimation). As noted in response to another comment, we will present the dates of peak snow cover and streamflow.

L173-174 define "changing climate" and justify how a dataset that is shorter than 30 years can be representative of a change in climate.

We clarify that we mean by "changing climate" a shift in weather patterns -- that may or may not be associated with climate change.

L178 'a reduction in storm size' is not captured in the hypotheses yet – greater storm frequency does not always mean in a reduction in storm size. Also, there was no explanation on how 'storms' were characterized (precipitation magnitudes separated by an interval of how many hours at least?). Lastly, in many parts of the world a 'storm' can just be a period with a lot of wind, and does not necessarily imply rainfall. Please consider changing it to 'rainfall event' or 'precipitation event'.

Thanks for this comment. The reason we originally focused on storm frequency is due to the fact that remote sensing products are quite reliable for observing the occurrence of precipitation, but have high uncertainty when estimating the depths of storms (especially in mountainous regions). As stated in the manuscript, we use remotely sensed observations of daily precipitation because our gauge date is limited to monthly observations. To address this comment, therefore, we will convert frequency precipitation frequency to average event depth by combining the frequency observed from PERSIANN with the monthly total precipitation from gauges. We will also change the terminology of "storm event" to "precipitation event" and clarify that a precipitation event is any day in which precipitation was ≥ 1 mm.

L181 The authors hypothesize here about greater bare soil evaporation, but included no information on soil moisture (only groundwater), neither did they explain why this information is not considered.

We are unaware of any datasource that would provide reliable information about soil moisture before and after 2000. For this reason we can only hypothesize what could have changed. The intention of mentioning soil moisture in this passage is to describe how a change in storm frequency could affect the water balance (as has been observed in other watersheds). We will therefore clarify in the text that this is only a possible explanation that would relate changes in precipitation depths with the observed changes in streamflow. We do not explicitly evaluate a change in soil moisture.

L185-187 What does it mean for the water balance estimation that a Thiessen polygon results in better matching observations representing the elevation gradient in precipitation? What does this mean for the degree of reality with which the authors represent real-world processes? From the text here, it sounds like the water balance calculation might be quite wrong, and a precipitation distribution that the authors know is not correct has been selected to fit the uncertain water balance calculation. Please add more explanation/justification to show that the Thiessen polygons are still the best representation of precipitation, although they might not reflect the real-world precipitation patterns. A map showing where the two are different and how different they are could be a good start.

Unfortunately, a reliable estimate of net watershed precipitation within this basin would be quite difficult to achieve given the limited availability of data and strong elevation gradients. That the water balance closure is poor using the elevation-gradient suggests that the relationship between precipitation and elevation is nonlinear over the range of elevation in the basin (most likely the relationship flattens at higher elevations). With that said, the water balance closure provides some confidence in the Thiessen polygon interpolation -- it's otherwise impossible to generate an independent measure of the amount of bias in this interpolation. This is a common problem in hydrology and it's why, as we note, some researchers have argued to "leave the water balance open". Our approach relies on the fact that we use the same datasets to estimate precipitation before and after 2000, so that any consistent bias will be removed when calculating this change. Of course, there may be some nonstationary bias that appears or disappears between the two periods (before and after 2000) that we are unable to account for. We again note that our approach hedges against this possibility by using multiple hypothesis testing to corroborate findings and build our understanding of hydrological change. We will nevertheless make this limitation more explicit within the revised manuscript. We will also add the suggested map to show the differences between the two approaches to precipitation interpolation.

L227 how are season defined, and do these match up with crop life cycle? If not, how does that influence the analysis presented?

We use four 3-month seasons commonly defined for Kashmir (e.g., Khattak, 2011). Rice, wheat, and maize are the top three crops in the region. Although the seasons don't perfectly align with the growing season, Summer (June-August) is the primary growing season for paddy and maize. Both are typically planted in the spring and harvested in late summer (August) or early fall (September). Wheat is typically planted in October and harvested in June.

In the spring, the greatest change in NDVI is outside the valley (Figure 5a), which we associate with greater activity of native vegetation. The exception is along the southern portion of the valley, where we observe an increase in fruit orchards. In the summer, the greatest increase in NDVI is seen throughout the valley (Figure 5b), which consists primarily of traditional crops (e.g., paddy and maize) along with the orchards in the south.

We will clarify these details within the revised manuscript.

Khattak, M. S., Babel, M. S., & Sharif, M. (2011). Hydro-meteorological trends in the upper Indus River basin in Pakistan. Climate research, 46(2), 103-119.

L265 I am not convinced that hypothesis 8 or 9 are sufficiently supported with the data as presented now. For instance, quantifying by how much the snowmelt date and peak streamflow date have shifted would give better insight in their relation than just only stating "both were earlier, so the appear to be connected". Hypothesis 9 is not based on independent observations (as shown in Fig 2) but on an inference that because evapotranspiration estimates were higher less water will have recharged the groundwater so that now groundwater contributions are lower. Lastly, the use of 'baseflow' analysis in Figure 2 applicable to hypothesis 9 is confusing, because 'baseflow' analysis is a common term used for the analysis of water level data. In this case, there is some analysis of streamflow data (for which there are no methods - confusing?) but the bulk seems to be based on the extent of surface water, which to me is a less common of the term 'baseflow analysis'. Again, an estimation of the uncertainties that are involved are necessary here.

We will provide the dates of peak snowcover and streamflow before and after 2000, as suggested, and will clarify the limitations of our understanding of the relationship between peak snow cover and streamflow. As for hypothesis 9, we realize that we have not clearly articulated our approach to evaluating this hypothesis. We will first clarify that the hypothesis seeks to identify a reduction of groundwater in the valley. We do this primarily by assuming baseflow can be determined via a (non)linear reservoir model, which is commonly used to associate baseflow (Q_B) with groundwater storage (S) as:

$$Q_B = aS^b.$$

We don't apply or calibrate this model, but rather we use it conceptually to demonstrate that a reduction in baseflow is indicative of a loss of groundwater because there is a monotonic

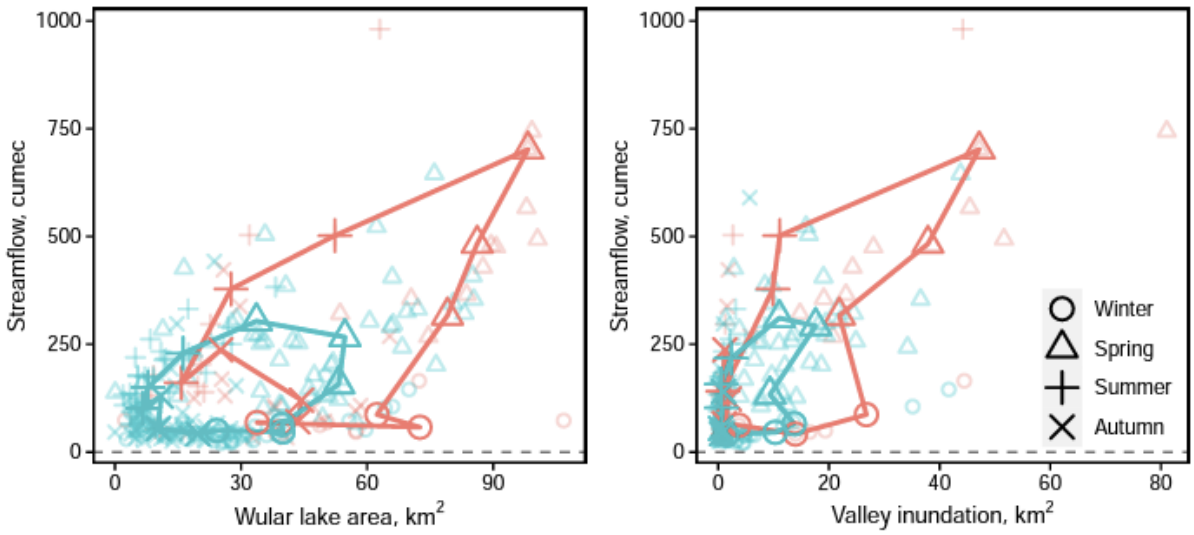
relationship between the two. By demonstrating that baseflow has declined over time, we infer that there has also been a loss of groundwater. As such the conclusions related to this hypothesis are necessarily circumstantial (we do not have the ability to observe a loss of groundwater). We take care as to how we describe our findings pertaining to this hypothesis. For instance, we state “This lends credence to Hypothesis 9, and we further discuss potential causes and implications of these opposing trends in the baseflow index with respect to saturated and unsaturated groundwater storage in the Discussion (Sect. 5)”. This approach is in agreement with the method, which favors holistic understanding through analyses of multiple hypotheses. However interesting, a complete analysis of groundwater is beyond the scope of this manuscript and would be difficult to implement.

L292 remove “by”

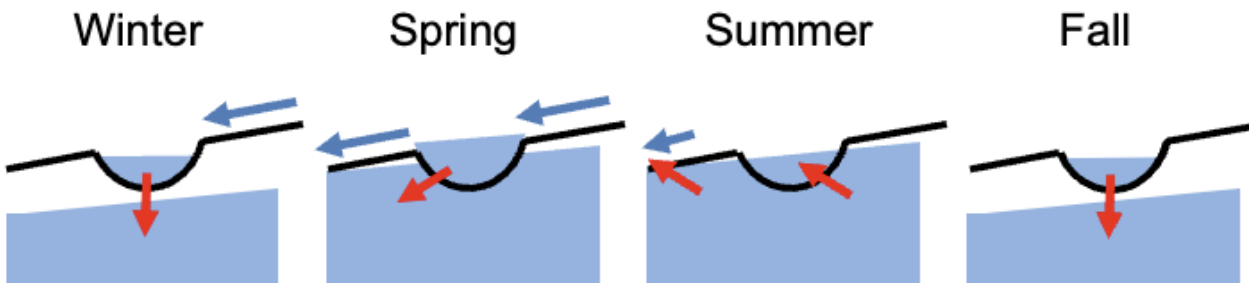
We will remove.

L310-317 This section is not clear to me. A drawing would be helpful.

The key idea of this paragraph is that the main river channels are connected to multiple large water bodies, notably Wular Lake and the wetlands in the valley. If, for instance, streamflow below the lak were controlled entirely by lake water levels, we would expect a 1-to-1 relationship between streamflow and the lake water area. Conversely, if streamflow were controlled by a combination of lake water levels and groundwater storage, we would expect the system to exhibit a property of memory based on groundwater storage. Because we expect the river to be coupled with groundwater, we hypothesized that there may be hysteresis in the relationship between water stored in surface water bodies and streamflow at the watershed outlet. We will clarify these points in the manuscript. We include here the original Figure S7 from the supplementary material, which shows this hysteresis:



We will also add the following figure as a subpanel to Figure S7. This figure shows how the observed hysteresis could be produced from the interacting relationships between streamflow, surface water bodies, groundwater, and runoff generation in the upland portion of the watershed. In particular, it could be explained by gaining condition in the summer and losing conditions in other seasons, although testing this follow-up hypothesis requires additional research.



L355 what is the threshold for a 'storm' if storms smaller than 1mm are included? Would the findings also hold if only storms > 5mm are included? (a common threshold)

Our threshold for daily precipitation events is 1 mm / day. The PERSIANN dataset includes data where precipitation is less than 1 mm on a particular day, and we therefore removed such days from our analysis of precipitation events. We will re-run the analysis using a threshold of 5 mm and include our findings in the SI.

L360 32 mm < 15% of 311 mm… please make clear to the reader that the change is smaller than the uncertainty.

Although this statement is true (and we will clarify it in the manuscript), context is required. Even though the 32 mm is less than the "missing" portion of the differential water balance, we have high confidence that there was an increase in ET. We will support this claim in the revised manuscript through a more careful treatment of uncertainty, in response to your second major comment above. In particular, we will compare the change in ET estimated from remote sensing datasets with change in ET estimated from the water balance residual.

L378 please quantify "a substantial level of uncertainty", and the implications that has for this analysis.

Because there are fewer observations, there are greater standard errors and uncertainty in estimating the mean. Although there is high uncertainty for these pixels, they represent a small portion of the watershed. We will clarify the effect of these pixels on the overall estimate of ET in the revised manuscript by showing the fractional areal coverage of these pixels in relation to the watershed.

Fig 6 I spent five minutes looking at this figure, but still don't fully understand what is shown. Moving the legend from panel a outside of the plot region would certainly help, but then still, I am not sure what the individual panels show and how the panels work together.

Thank you for pointing this out. The figure aims to show two pieces of information: (a) the land use categories in which ET increased the most (ie, on a per-unit-area basis), and (b) the land use categories that contributed the most to the change in ET (ie, averaged over the entire watershed). We will seek to address this confusion by:

- Moving the legends outside the plots
- Adding a y-axis title to the plot on the right
- Modifying the caption to: "Changes in land use and evapotranspiration. (a) Change in evapotranspiration within each combination of land use categories from 2001 and 2010. The size of the circle represents the fractional area covered by each pairwise grouping, and the shading of the circle indicates the average change in ET per unit area. For instance, the largest increase in ET occurred in pixels that changed from crops in 2001 to mosaic in 2010 (+80 mm), but this represented a small fraction of the watershed (<5%). (b) Net effect of each 2010 land use category on evapotranspiration, averaged over the entire watershed (i.e., ΔET * LU Area / Watershed Area). For instance, the pixels that changed from crops to mosaic reduced ET by 2 mm when averaged over the entire watershed. Overall, most of the increase in watershed ET (27 mm) occurred in regions where land cover remained consistent from 2001 to 2010 (outlined in black), compared with regions where land cover changed (5 mm)."

L428 how was the baseflow index calculated? What is the associated uncertainty?

Baseflow was calculated using a numerical filter (Nathan and McMahon, 1990) (we note that this information was in the SI and will be ported to the main text), and the baseflow index was calculated as (baseflow) / (total flow). There are two sources of uncertainty -- first, the uncertainty associated with using trimonthly observations of streamflow, and second, the uncertainty associated with the numerical filter parameter. In the revised SI, we will: (a) benchmark the trimonthly baseflow against baseflow from daily measurements, and (b) conduct a sensitivity analysis on the numerical filter by adjusting the filter parameter within commonly acceptable ranges.

Nathan, R. J. and McMahon, T. A. (1990): Evaluation of Automated Techniques for Baseflow and Recession Analysis, Water Resources Research, 26, 1465–1473, https://doi.org/10.1029/WR026i007p01465.

Fig 8 how was the significance of these analyses tested?

Thank you. We will clarify in the manuscript that the statistical significance in this analysis was determined as $p < 0.1$ for a nonzero trend from a linear least squares regression. All

statistically significant trends were also significant for $p < 0.05$ except for the Upstream trend in panel (a).

Fig 9 to me it's not clear what the different subsurface layers mean. Soil? Groundwater? Saprolite?

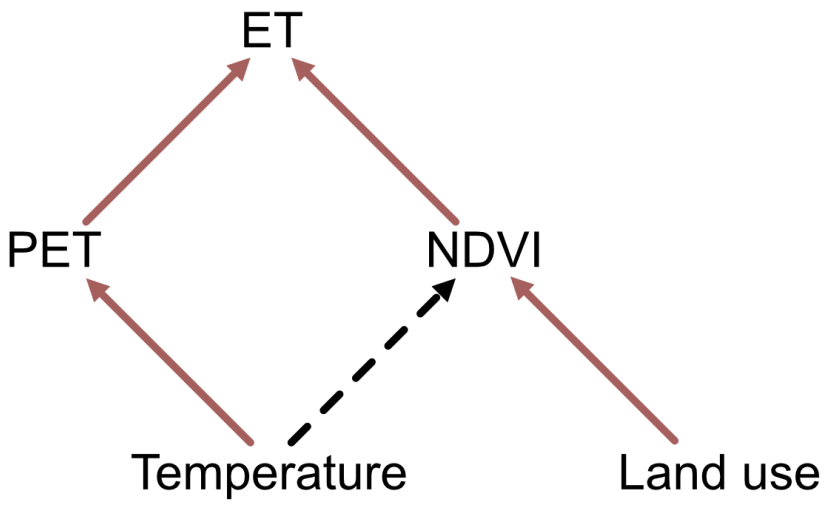
We will clarify in the revised manuscript that the soil layer in the figure represents unconsolidated deposits in the valley, which can be up to 1 km thick. Near the river channel, these are mostly alluvial deposits which overlay glacial deposits. Further up the slopes (but still in the valley) there are no alluvial deposits and the glacial deposits are uncovered.

L475 Doesn't the conclusion that NDVI was leading for ET follow from a model in which ET is calculated from NDVI? And, why did NDVI increase? The authors describe that warming temperatures were a reason for NDVI to increase, but said that temperature did not influence (evapo)transpiration directly? Precipitation amounts were much lower, and an increase in storms < 1mm don't bring much moisture to the soil either, that will be evaporated directly from the plant leaves. How can the climatic conditions be more favorable, while moisture and temperature apparently don't play a direct role?

We interpret this comment as highlighting the lack of clarity regarding our discussion of changing ET. We will therefore seek to clarify a couple of points. First, Temperature (T) can affect ET directly through the effect on potential evapotranspiration (PET). In the manuscript, we evaluate the effect that temperature has on ET via its effect on PET when we state: "Overall, of the 32 mm annual increase in ET that we detected, approximately 17% can be attributed directly to increasing air temperature." We will clarify that we are specifically referring to the effect through PET. Additionally, temperature can affect ET indirectly by changing growing conditions, vegetation structure and phenology, and therefore NDVI. In many places outside the valley where natural vegetation prevails (e.g., grassland / shrubs), there are large increases in NDVI and ET. This can be seen in Figure 6b by the yellow bar at the top -- land use remained the same but ET nevertheless increased considerably. While we note this relationship, we do not directly evaluate the effect of temperature on NDVI, as it is beyond the scope of this manuscript, and of course NDVI is also dependent on other variables.

Notably, NDVI also depends on land use. In the manuscript, we can make the broad assumption that any changes in NDVI can be attributed to land use change in pixels where land use actually changed. This would give us an upper bound on the effect of changing land use on ET, which we find to be 5 mm. This upper bound is nevertheless small relative to the estimated change in ET of 32 mm, allowing us to conclude that land use change is not a major driver of changing streamflow.

We will clarify the relationship among these variables by adding the following figure:



Caption: Temperature (T) can affect ET directly through the effect on potential evapotranspiration (PET), or indirectly by changing growing conditions, vegetation structure and phenology, and therefore NDVI. Land use also affects NDVI. We evaluate the direct effect of Temperature on ET using the ET model, and the effect of land use change on ET by assuming that any changes in NDVI where land use also changed are attributable only to the change in land use.

S1 pleas add 1:1 lines in the mass-mass plots

Thanks for the comment. We would like to note that double-mass plots should be linear, but not necessarily 1:1 (e.g., if one gauge reports more or less streamflow on average). To address the essence of your comment we will display linear trends on the double mass plots.

S2 what does the overestimation of precipitation say about the accuracy of your input data? What do the authors reckon is the importance of the spatial distribution of rain vs. the basin-average precipitation estimation. Spatial distribution might be very important, which is indicated by the small change in Q for various gauges. Commenting on this would be appreciated.

We don't have a way to validate the uncertainty on precipitation aside from using the water balance to see how far away we are from water balance closure. Because we utilize datasets that span both periods of analysis (before and after 2000), this issue would only affect our analysis if there is a new bias that appears (or disappears) between the periods. Of course, such a change could be plausible if, for instance, there is a change in the altitudinal precipitation gradient. We will therefore clarify in the manuscript that there may be changing biases that we cannot account for, that increase the uncertainty of our analysis (which we now quantify).

S6 what was assumed when zero or only few (defined as less than 5 in the text) images were available?

This comment refers to the number of summer landsat observations (for each pixel) prior to 2000 that were available to estimate NDVI. In some pixels, we had less than 5 cloud-free observations. We do not assume anything for these pixels, but rather note that they may bias the results. In the revised SI, we will provide a histogram of summer landsat observations to demonstrate that the net effect is likely small, because most pixels had a larger number of good observations.