Reviewer 3:

The paper presents a relevant and interesting idea. Normalizing station-month data to reduce artefacts in precipitation-temperature scaling is sensible, and the exponential model fitted to normalized anomalies improves predictive skill in the Upper Colorado River Basin. However, several aspects of the analysis and interpretation need tightening before the conclusions can be considered robust.

The authors would like to thank the reviewer for their time and effort in providing useful feedback concerning our paper. We will work to improve our analysis and interpretation in our revised paper.

The main issue is the lack of discussion on data quality control. The paper doesn't explain how precipitation observations were checked or filtered. Since Ali et al. (2022) highlights errors from coarse measurement precision and faulty readings, this needs to be addressed directly in the data section, with a brief note on the checks used or potential uncertainties.

The hourly data was checked and quality controlled against the daily data set, where both have been aggregated to a monthly resolution. We will include a few additional lines outlining this process. We can also include an additional figure in an appendix that shows the how the scaling rates differ when using a precipitation resolution of 0.1mm versus 1.0mm.

The temperature binning method could also be improved. Fixed temperature intervals cause uneven sampling-cooler bins dominate while warm bins remain sparse. Using bins with roughly equal numbers of data pairs would produce more balanced estimates.

We agree completely that the mentioned points are common problems with the binning method. In our revised paper, we will shift to using quantile regression instead of the binning method to illustrate some challenges a researcher faces when estimating P-T scaling rates. Additionally, quantile regression will be used to provide a reference or benchmark scaling rate estimate in the Upper Colorado River Basin.

Normalization is useful for removing spatial and seasonal effects, but it can hide genuine long-term trends. Subtracting historical station-month means risks erasing real climate signals in dew point or rainfall. The assumption of stationarity and the leave-one-year-out validation don't fully test for this. If the data are non-stationary, the resulting scaling estimates may be biased.

Thank you for this point. We use Figure 11 to show that we find no evidence of non-stationarity or systematic changes in scaling rates over time. However, we can add a note of caution about interpreting the results, given that the scaling rates may indeed be non-stationary.

From a statistical standpoint, a hierarchical model would be more robust than treating all stations equally or independently. It would allow shared information across stations while preserving local variations. Alternatively, quantile regression or a generalized additive model could capture nonlinear relationships without relying on arbitrary bins, and would better describe high-end percentiles.

Thank you for this comment. To reiterate our point from above, we will shift our content in the revised paper to implement quantile regression instead of the binning method.

The fitted exponential coefficient also needs clearer interpretation. The slope parameter b is treated as "% per °C," but the correct expression is (exp b - 1). Using b directly can slightly misstate the scaling rate.

This is a good point. We will change the exponential function to more explicitly have it to be interpreted as "% per °C". This can be done with a model taking the form $y = ax^b$, where b are the monthly dew point temperature values, a is a multiplicative offset, and x provides information concerning the scaling rate. For example, assume a=1, b=2 and x=1.08, then $y = 1*(1.08)^2 = 1.166$. The scaling rate is simply (x-1), and now the values actually scale at 8% per degree. In this example with a +2 degree anomaly, that would be 1.08 per degree or 1.08*1.08, for 2 degrees of warming, which is equal to 1.166.

Equation 7 divides precipitation by its mean for each station-month, but many of these means are very small or zero, inflating anomalies and adding noise. Although this is mentioned briefly, its effect isn't explored or corrected.

The reviewer is correct that this can lead to larger values. To deal with this, we did not use stations which had a mean extreme precipitation amount less than or equal to 0.1 mm. So, we are always dividing by a mean that is greater than 0.1 mm. After normalizing the precipitation data, we did not find the presence of any abnormally large or unrealistic values. We will add a statement regarding this point in the revised paper.

Using monthly-mean dew point as a predictor helps correlation but weakens the physical link to rainfall extremes, which depend on short-term moisture and dynamics such as CAPE or large-scale ascent. Higher-frequency predictors would strengthen the physical interpretation.

Our aim to to provide a robust estimate of how extreme precipitation would be expected to change in a warming world. Of course, the more variables that one uses to predict extreme precipitation, the better those forecasts should be. However, we are not attempting to perform numerical weather prediction, but rather we want to provide a zoomed-out view of how something like 1 degree of further warming in a region would translate into expected average changes to

extreme precipitation for a given season. Further predictors and complexity can always be added, but that is beyond the scope of this paper.

The assumption that station-month maxima are statistically independent isn’t fully demonstrated. Dependence across years or from climate modes like ENSO could still exist. Block-bootstrap or similar resampling methods would provide more realistic uncertainty estimates.

Thank you for this point. We will provide a better analysis of the statistical dependence of the data which we use in our modeling approach. However, we generate randomized predictions that exhibit the same temporal and spatial autocorrelation as the data itself. Therefore, we have already included the underlying data's statistical dependence in our model evaluation. We will include a more thorough analysis of the model performance in the revised paper.