Reviewer 2:

This paper proposes an integrated approach to improve our estimates of precipitation-dew point scaling rates. The idea relies on the seasonal normalisation of dew point (additive normalisation) and monthly maxima of precipitation at daily/hourly scales (multiplicative normalisation).

Overall, the manuscript collects a set of ideas from literature (although sometimes with incomplete referencing), and proposes an approach to integrate those ideas. For this reason, I think the title is slightly misleading, as it tends to overgeneralise the breadth of the contribution and oversell the novelty.

The authors would like to thank the reviewer for their time and effort in providing useful feedback concerning our paper. We will try to improve our referencing and clarify our contribution. Our revised paper will include a new title that more accurately reflects our work.

Before commenting, I'd like to say that I found the manuscript difficult to follow, so some of my comments may be associated with misunderstandings. I'll be happy to discuss them further.

I think the methodological idea has potential, but the implementation is rather convoluted and it hinges on some subjective choices not fully motivated in the text. Also, I am not convinced by the validation that, to my understanding, is based on mean deviations - while in precipitation-temperature scaling we are typically interested in extremes.

Thank you for the comment. We will work to improve the readability and the clarity of the revised paper. To be clear, we are using extremes throughout the paper. Please refer to our response, found below in this document, corresponding to your point at Line 308 in the original paper version.

Overall, the study has potential, but I think major work is required before it can be reconsidered for publication.

I provide below here my specific comments, in the order they appear in the manuscript.

- Line 7-8: here it is not yet clear what 'normalised' means. I suggest to briefly explain it
  We can add a brief description earlier in the paper. At this point in the abstract, we can rewrite the text starting at line 7 in the paper to read: "Specifically, we compare multiple approaches, including those using raw (non-normalized) and normalized data (using % of normal for precipitation and °C from normal for temperature or dew point temperature), to estimate P–T scaling for hourly and daily extreme precipitation."

- Line 20: is looks odd to have two significant digits for the 1 degree and not for the 7% (which is approximated)
  Thank you for this suggestion. We can adjust the number of significant digits to be consistent throughout the manuscript.

- Line 20: Extreme precipitation events depend on other variables, not only moisture in the column. I suggest to slightly rephrase.
  This is a valid point, and we will rephrase this sentence to include other variables which also influence extreme precipitation.

- Line 42: Technically, dew point is defined as “the temperature at which air saturates when cooled at constant pressure” (e.g., see wikipedia or any atmospheric sciences book). It follows that dew point over a given location contains information only on the available moisture, and not on temperature. In fact, knowing the dew point and the pressure, it is not possible to calculate the temperature. For this reason, also the sentence in lines 49-50 needs to be updated (“the chosen temperature variable” should be changed to something like “the chosen variable”).
  Thank you for this suggestion. We will update our definition of dew point temperature. Many prior studies which have focused on scaling rates have compared air temperatures and dew point temperatures. It is common practice to refer to dew point temperature as a temperature variable. However, we can more simply refer to the "chosen variable".

- Line 45: Marra & al 2024 do not use the binning method, please remove the reference. The reference instead can be relevant to the sentence at lines 65-67 and 67-68.
  Thank you for pointing this out to us. We will change the references accordingly.

- Line 54: data normalisation has not been defined. What do you mean by that? I expect to learn it later, but it should be explained earlier.
  Good point, we will provide a description in the text, that we then later define mathematically.

- Line 69: you seem to use the binning method. Indeed the introduction mentions this method, but does not mention there are some alternatives, namely the quantile regression, which is known to provide more robust estimates (if used properly) and to require much less subjective choices. This becomes more relevant given the fact that later you use an exponential model for the means and that the evaluation is done on mean

values. I believe quantile regressions could help you solve both these issues.

Thank you for this comment. 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 an additional reference or benchmark scaling rate estimate in the Upper Colorado River Basin.

- Line 75: the introduction fails to mention the literature that investigated the impact of process heterogeneity on the emerging scaling rates (e.g., Molnar et al 2015, which is cited but for other reasons, or De Silva & al 2025 Nat Geo). Given the focus on seasonality, this is a critical aspect that needs to be addressed. For example, is the normalisation handling the same problems? Is it only an approximation of what a classification would do in a more proper manner?

  We aim to provide one way to improve the accuracy of scaling rate estimates, and how they can skillfully provide predictions of extreme precipitation. We have used normalization to remove climatological differences found in the data. Classification will group data that are statistically similar, while we have transformed all of our data so that all of the data exhibit similar statistics by having a common frame of reference. We have already shown, and will aim to make more clear in our revised paper, that our approach is statistically significantly skillful. We will additionally make sure to add in these points and references in the Introduction. Perhaps in some future study, others can apply an adjusted version of our proposed methodology. However, that is beyond the scope of our work here.

- Line 108: “make skillful predictions of extreme precipitation”. It is not yet clear what you mean by “predictions”. What are you trying to predict? is it the scaling rate for a given place and month? Is it the precipitation magnitude associated with a given probability (percentile) at a given temperature? Needs to be clear right from here, otherwise it is difficult to follow.

  In our revised paper, we will try to be more clear about what exactly we are predicting. We will also change some of the terminology in our revised paper. Notably, we will use commonly defined index names to refer to our maximum hourly and daily precipitation values, which are obtained at each station and each month. These will be referred to as Rx1h and Rx1day. Consider that we are trying to predict daily maximum precipitation (Rx1day) given changes in monthly dew point temperature. Using cross-

validation, we predict, for each station and for each month, how much above or below normal the Rx1day would be given the dew point temperature anomaly at that station and in that month. For example, let's say that some station in July has a mean Rx1day across the calibration years of 15 mm, and the mean dew point temperature is 10°C. Through our model fit in the calibration period, let's say that Rx1day scales by 12%/deg. Now, if that station had a dew point temperature value of 11°C (i.e., 1°C above normal) in one year in the validation period, we would predict that the precipitation would be 12% greater than average (i.e., 112% of normal), or 16.8 mm. We perform the same procedure for all stations and months in the validation period. So, we are predicting how much the intensities of Rx1h and Rx1day are changing as a function of dew point temperature anomalies. We then evaluate these predictions versus observations, in the validation period, across different bins of dew point temperature anomalies. We evaluate model performance using deterministic values, corresponding to a linear exponential model fit. In our revised paper, we will also add some discussion about the utility and interpretation of deterministic versus probabilistic predictions.

- Fig. 2 and the related analyses. Technically, the hook structure could be created by lack of sufficient observations to properly estimate rare percentiles (Marra & al 2024). What the normalisation does is to remove the heterogeneity. The result is that the sample is homogeneous and its portion at high temperatures becomes more populated, allowing for a better estimate of the large percentiles. Therefore, the normalisation alone is not the “cure” to the hook structure, it also needs sufficient data sample. I suggest to better state this.
  This is an interesting comment. Indeed, increasing the sample size will have an impact. However, in the case of our data in the Upper Colorado River Basin, we did not change the data sample. In each case, using non-normalized versus normalized data, we took the average of the data within a bin only when at least 10 points fell in the top 0.1% or 1%. So, in the case of the UCRB, it is not the difference in sample sizes that explains the hook, but rather climatological differences in space and time. After normalizing the data, that alone was enough to remove the hook structure. We will restructure and restate this point in the revised version of our paper.

- Lines 198-205: this is a lot of text to say “precipitation is stochastic”.
  We will attempt to simplify the text here.

- Lines 209-210: this resembles some of the ideas behind Marra et al. 2024

Thank you for this suggestion. Instead of collocated temperature values, they use daily-averaged values of temperature. We aggregate the dew point temperature over a longer period of time (i.e., the month). Indeed, we agree that it is good advice to relate our results to their paper here.

- Line 236: why only the mean is included in the normalisation? Doesn't the variance also count? It should be stated something about the assumption behind this normalisation.

  There is not a precise, fixed definition of the terms normalization and standardization. The authors view standardization as the process of computing z-scores (because the data is being standardized by subtracting the mean and dividing by the standard deviation). We have interpreted and presented normalization as the process of adjusting values to a common scale, where the data can then be more easily compared and contrasted with a common frame of reference. Ultimately, we want to provide scaling rates as %/°C. Therefore, this informed us as to how we wanted to scale the data. We wanted the precipitation data to be "% of Normal" and dew point temperature data to be "°C from Normal". Then, after applying those normalizations, we can easily fit a model which provides us with a scaling rate as %/°C. This is not the case if we were dealing with z-scores, as those values are unitless, and would thus not provide easily interpretable scaling rate values such as % change of Rx1day precipitation per degree change in dew point temperature.

- Line 246-246: you repeatedly claim that the normalisation above "effectively timely removed the three common challenges". This needs to be shown, for example you can compare the distributions before and after for some example cases.

  We show this to be the case for the temporal independence of the data. The normalization of precipitation and temperature does put the data on similar scales for each respective variable. As stated above, we want to fit and estimate scaling rates as %/°C. To do this, our normalization procedure does not perfectly put all of the data to the same distribution. Some station could have greater skewness in its summer precipitation than its winter precipitation. Similarly, the variance of dew point temperatures at some station could be greater in the winter than for the summer. We don't perfectly collapse all of the data to one common distribution, but we have a common frame of reference that is then useful in providing scaling rates as %/°C. We can more carefully state that while we do not entirely remove these challenges, we do show how useful the normalized data is in improving our predictions of extreme events.

- Line 251-252: “any reference…” I suggest to include this part, in some way (perhaps referring to section 3.2), much earlier in the text.
  Good point, we will bring up earlier in the text what we mean by normalized data.

- Line 294: I don’t understand how this is possible. Perhaps the way daily and hourly maxima are defined is not sufficiently clear?
  As we state in the Data section 2, the hourly and daily data are different data sets. For the hourly dataset, there are not as many stations and the average period of record is shorter than with the daily dataset.

- Line 308 and Figures 6,7,8,11,12: if I understood correctly, the evaluation is done on the mean values of the bins, and not on the extremes. Is this correct? I don’t understand how is this useful for extremes, which are the target of P-T scaling applications. I think more reasoning should be provided here.
  All of our precipitation data used for all of those figures are extreme precipitation data. To clarify, we can provide a comparative example. Let's take the traditional binning method with data pooled from all months and stations within a region such as the Upper Colorado River Basin. Next, we could take bins of dew point temperature and find, for example, the hourly precipitation amounts in the top 0.1% in each bin. If one were then to take the average of these points that fall within the top 0.1%, we are still dealing with extreme events but we are finding the average extreme precipitation amount across each bin. In any distribution, if you were to take the top 1% or 0.1% of the values in that distribution, and then average those values, you are providing information about the extremes or the tail end of the distribution and not the mean of the entire distribution. Now, instead of using the binning method, we take the hourly maximum precipitation for each month and each station. We will try to improve our content in the paper revision as to why this is a good idea. Then, we perform our modeling and evaluation over these extreme precipitation values that fall in the top ~0.1% (for hourly data) and ~3.3% (for daily data).

- Fig 7 and several other results/validation: why are 2 degree C bins used? What is the sensitivity of the outcomes to this choice?
  Thank you for this comment. We plan to include a more thorough analysis of the statistical significance of our model and its sensitivity to evaluating the model performance when using different bin sizes.

- Fig 7 and 10: usually the reference value (observed in our case) is plotted in the x-axis to facilitate interpretation (model overestimation is above the 1:1 line, and underestimation below)
  While we appreciate the reviewers observation here, it is our preference to plot the results the way that they have been presented. We have clearly labeled the x- and y-axes, to help facilitate the reader making sense of the results.

- Line 460-461: I agree on this consideration, but I wonder how much statistically robust it can be considered. For example, the plots for -3 and +3 (Fig 11a,f) show quite many large dots at the boundaries of the distribution. Perhaps a statistical test can help on this. Perhaps the same Montecarlo used here can be used, if many more samples are generated?
  We respectfully disagree with the reviewers interpretation of this figure. The vast majority of the larger points fall well within the range of the smaller points. While a few of the larger points do fall near the edges of the cloud of smaller points, this is also to be expected with randomly generated data. We do not find the existence of more larger points that are near the boundaries than one would find with randomly generated data.

- Lines 492-494: I don't understand how this finding relates with the finding that normalised values allowed for putting different months together (e.g., fig 9). Isn't this result suggesting that we should not mix months even with normalisation? Please explain.

  Thanks for this comment, and we acknowledge that this point requires additional explanation. Our results from Figure 8 show that normalizing allows us to leverage data across space and time in a way that non-normalized data does not allow. What Figure 9 also shows is that the way we normalize matters. Even data across the same stations in the same season, for example, can be influenced enough by climatological differences to where we inaccurately estimate the underlying scaling rate (Figs 9a,9b). These results tell us that if we obtain maximum precipitation amounts at each station across a 3-month season, and then normalize those values, we encounter a problem with climatological differences across time. We showed how different the climatologies are between the months of August and October. Taking an extreme value across the August-October season will end up mixing statistically significantly different climatologies and will adversely affect the estimation of the scaling rate. However, if we normalize the values for each station and each month first (across all Augusts for a single station, for example), then this effectively allows us to

better leverage the data across months in a season, as depicted in Figure 8.