Reviewer 1:

Let me start by stating the manuscript is well written and I believe worthy of publication. However, I am concerned that the novelty has been misrepresented. The main issue here is the authors provide 3 shortcomings they have overcome building on Zhang et al (2017). I do not agree with this statement as I believe each individual shortcoming has been addressed in previous literature. Here the novelty lies in combining the approaches from three different manuscripts. I believe this combining of methods is a worthy contribution and an important one.

We would like to thank the reviewer for their comments. They have greatly assisted us in improving our paper. In light of of all of the great comments from the three reviewers, we have heavily revised our paper.

We appreciate that the reviewer acknowledges this general aim as important and relevant. In line with the reviewer's point, we have made changes to our paper to better position our work with respect to prior research/literature. Please refer more specifically to lines 46-65 in the revised version of the paper.

I have two major comments which I hope the authors can address.
Major comment 1.
I strongly believe that data for P-T scaling should not be pooled without standardising – and agree with the authors, but this was demonstrated in Visser et al (2022) and Molnar et al (2015). The authors have cited these papers in their manuscript and to their own admission at line 379 they state their work bears a strong resemblance to Visser et al (2022) and Molnar et al (2015) so why not make this point in the introduction that they build on these authors work?

Thank you again for this point. We have more clearly referenced how our work builds off some of these prior studies. Please refer to lines 57-65 in the revised paper.

Further, Figure 2: A standardisation was already proposed in Visser et al (2022) and I quote in their introduction “We introduce standardized pooling…”. This should be acknowledged here.

Thank you for this point. We will better acknowledge the standardization that was performed in Visser et al. (2021). In their paper, they do standardize by subtracting the mean and dividing by the standard deviation. They do this for each station. However, they do not apply the method month by month. As a result, they control for some climatological differences across space, but not in

time. Additionally, we apply a different approach to the normalization, whereby we produce percent of normal (or average) precipitation values and degrees from normal (or average) dew point temperature. However, we agree that we need to better acknowledge, for example, the work done by Visser et al. (2022). We now have beginning at line 61:
"Using pooled data, then, without accounting for these climatological differences, one runs the risk of inaccurately estimating the effective scaling rates. In light of this problem, some have advised using normalized or standardized data (Zhang et al., 2017; Visser et al., 2021)."

Figure 3 and 4: The issue that bins at the extremes have less events, and some binning techniques don't consider independence were both points made in Wasko and Sharma (2014) and hence quantile regression using independent events was proposed. This point also relates to Line 406. This should be acknowledged here.

For the reasons that you have clearly outlined, our revised version of the paper now focuses almost exclusively on using quantile regression instead of the binning method. One exception is for the content found in new Figure 2. This is when we point out and discuss the problems that can stem from mixing different climatologies when pooling raw (or non-normalized) data.

Figure 5: The use of a monthly (or seasonal) temperature was proposed by Zhang et al (2017). This should be acknowledged here.

Thank you for this comment. Please refer to lines 208-210, where we now write:
"Both Zhang et al. (2017) and Marra et al. (2024) have previously explored using predictor data which is not concurrent. Zhang et al. (2017) used only wet-days to compute an average seasonal dew point temperature, while Marra et al. (2024) used average daily temperatures."

In sum, while the justification of the proposed methodology presented in this manuscript is much more elaborate than previous manuscripts (and hence I am a proponent of it being published) the framing needs to change I believe to duly pay respect to the previous research. The method proposed here is more a combination of methods proposed by Visser et al (2022), Wasko et al (2014), and Zhang et al (2017) and the introduction and conclusion should be restructured accordingly.

Thank you for this assessment, and for helping us to better place our work. We have worked to better acknowledge the great research that has already been done by these authors and highlight our own addition to it by combining and

expanding it. Please refer to our revised Introduction and Conclusion where we better acknowledge that our work is built on the foundation of these prior studies. In particular, we begin our last paragraph of our Introduction, at line 65, with this sentence:
"Our work herein is built on the foundation of prior work (Zhang et al., 2017; Wasko and Sharma, 2014; Molnar et al., 2015; Visser et al., 2021)."

Major comment 2.
It is odd that the authors choose 7% per degree as their truth when calculating the skill, when by their own admission in Figure 12 the scaling is not aligned with CC? In some way this should addressed, with at least more focus on the actual scaling rates. The reason is – these are empirical relationships, without a "truth".

This is a great comment, and it is actually central to what we are trying to show. We have put in a great deal of effort in order to streamline our results, and to better illustrate and communicate this particular information that the reviewer is looking for here. In our prior version of this paper, we stepped through different validation frameworks. We acknowledge that our prior approach was perhaps a little confusing and convoluted. Therefore, we have streamlined our validation approach, where we now apply leave-one-year-out cross-validataion for both Rx1hr and Rx1day, and additionally we perform a two-fold cross validation for Rx1day. This is done by using four different modeling approaches: 1) always assuming climatology (100% of normal), 2) always assuming a C-C scaling rate of 7% per degree C, 3) fitting an exponential function to non-normalized data, and 4) fitting an exponential function to normalized data. We additionally have a fifth model, which is producing randomized predictions, and it is used to provide statistical significance of the skill scores. In your comment, you ask for more focus on the "actual" scaling rates. We put the word "actual" in quotes, because the empirical-statistical scaling rates vary as a function of the data used (non-normalized versus normalized) and the amount of data pooling applied. The data pooling relates to how far spatially and temporally we pool data from. We agree that there is no real "truth". However, different approaches under different conditions, will produce better estimates of scaling rates, and thus better predictions. When using non-normalized data with minimal data pooling (e.g., using yearly data from a single station and for a single month), we do obtain similar scaling rates, compared to using normalized data. However, using normalized data that is pooled does provide an improvement in model performance. Using the daily dataset, we clearly show in Figs. 5e and 5f (in the new version of the paper) an example of when the scaling rates can differ dramatically depending on whether or not we use normalized data to fit our model.

Minor comments:
Title: The title suggests a review and noting that some of the current challenges have been resolved the title could be amended.

We have a new title which we believe more accurately reflects our work. It is now:
"Leveraging normalized data to improve point-scale estimates of precipitation-temperature scaling rates".

Line 1-2: Does sub-daily rainfall scale at 7%? There is now much review/meta-analysis work showing it is likely higher? e.g. Fowler et al (2021); Wasko et al (2024). The IPCC reports also point to higher than 7% scaling for sub-daily rainfall.

The scaling rate will ultimately depend on the location of a given station and the time of year. We find the median scaling rates of hourly extremes (i.e., Rx1hr), for the stations in the UCRB, to be less than 7% in the winter and substantially higher than 7% in the summer (see Figure 9). Because it is not clear ahead of time, what scaling rates we can/should expect, we have additionally provided a comparison to scaling rates assuming C-C. However, we also compare to a method which uses the exact same data, but it has not been normalized. That approach, which uses the non-normalized data from the UCRB is not constrained by any upper or lower limit of the scaling rates.

Line 60 onwards: The point of pooling resulting in “incorrect” scaling was well made in Molnar et al (2015) and has been made in papers by Berg and Haerter – making the point that a lot of this has to do with different storm types, but this was never mentioned here?

Thank you for this comment. One can view the use of normalization, station-by-station and month-by-month, as a proxy for different storm types. Molnar et al. (2015) stated, "the scaling rates for all events are systematically higher than those of the individual lightning and no-lightning subsets because of the mixing of stratiform events at low temperatures and convective events at high temperatures." They are pointing to the fact that different seasons experience different types of storms. We agree that this is a reason why the raw, non-normalized data should not be mixed. Please refer to line 222, where we now write,:

"By normalizing the data, we provide a more homogeneous framework for dealing with different types of precipitation (Berg and Haerter, 2013; Molnar et al., 2015)."

Line 113: “incorrect” is a strong word when we don’t know the truth, scaling’s are correlations and they’re all true in some way regardless of the method.

Thank you for this suggestion. We have removed strong wording such as "incorrect" throughout the paper.

Figure 8 nicely presents that the pooling of standardized data works, but Figure 8d also shows that monthly data can be safely pooled after standardization and the performance is similar (Column 1 vs Column 4) – could this point me made in the text?

We now include more discussion about how the data can be effectively pooled, or not, given whether the data is normalized or not. Please refer to the content regarding Figures 7 and 8 in the revised paper.

References:

Berg, P., Haerter, J.O., 2013. Unexpected increase in precipitation intensity with temperature — A result of mixing of precipitation types? Atmospheric Research 119, 56–61. https://doi.org/10.1016/j.atmosres.2011.05.012

Fowler, H.J., Lenderink, G., Prein, A.F., Westra, S., Allan, R.P., Ban, N., Barbero, R., Berg, P., Blenkinsop, S., Do, H.X., Guerreiro, S., Haerter, J.O., Kendon, E.J., Lewis, E., Schaer, C., Sharma, A., Villarini, G., Wasko, C., Zhang, X., 2021. Anthropogenic intensification of short-duration rainfall extremes. Nature Reviews Earth & Environment 2, 107–122. https://doi.org/10.1038/s43017-020-00128-6

Molnar, P., Fatichi, S., Gaál, L., Szolgay, J., Burlando, P., 2015. Storm type effects on super Clausius–Clapeyron scaling of intense rainstorm properties with air temperature. Hydrology and Earth System Sciences 19, 1753–1766. https://doi.org/10.5194/hess-19-1753-2015

Visser, J.B., Wasko, C., Sharma, A., Nathan, R., 2021. Eliminating the “hook” in Precipitation-Temperature Scaling. Journal of Climate 34, 9535–9549. https://doi.org/10.1175/JCLI-D-21-0292.1

Wasko, C., Sharma, A., 2014. Quantile regression for investigating scaling of extreme precipitation with temperature. Water Resources Research 50, 3608–3614. https://doi.org/10.1002/2013WR015194

Wasko, C., Westra, S., Nathan, R., Pepler, A., Raupach, T.H., Dowdy, A., Johnson, F., Ho, M., McInnes, K.L., Jakob, D., Evans, J., Villarini, G., Fowler, H.J., 2024. A systematic review of climate change science relevant to Australian design flood estimation. Hydrology and Earth System Sciences 28, 1251–1285. https://doi.org/10.5194/hess-28-1251-2024

Zhang, X., Zwiers, F.W., Li, G., Wan, H., Cannon, A.J., 2017. Complexity in estimating past and future extreme short-duration rainfall. Nature Geoscience 10, 255–259. https://doi.org/10.1038/ngeo2911

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.

We would like to thank the reviewer for their comments. They have greatly assisted us in improving our paper. In light of of all of the great comments from the three reviewers, we have heavily revised our paper.

We have improved our referencing and worked to clarify our contribution. We also have a new title which we believe more accurately reflects our work. It is now:
"Leveraging normalized data to improve point-scale estimates of precipitation-temperature scaling rates".

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.

We have worked to improve the readability and clarity of our revised paper. In an effort to be more clear that we have always been dealing with extreme precipitation throughout our work, we introduce the terms: Rx1hr (the maximum hourly precipitation rate found for each station and each month) and Rx1day (the maximum daily precipitation rate found for each station and each month). Using different modeling approaches, we fit scaling rates between average monthly dew point temperatures and these Rx1hr or Rx1day values. We subsequently make cross-validated predictions, and the predictions are evaluated to determine which approach is the most skillful.

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 introduce this term earlier in the paper, but we more clearly define it in the Data section of the paper. This is in line with how other authors have presented material such as this. Visser et al. (2021), who introduces the term standardized in their abstract and in the Introduction, but they do not explain or define it until their "Data and methods" section of the paper.

Visser, J. B., Wasko, C., Sharma, A., and Nathan, R.: Eliminating the “Hook” in Precipitation–Temperature Scaling, Journal of Climate, 34(23), 9535–9549, https://doi.org/10.1175/JCLI-D-21-0292.1, 2021.

- Line 20: is looks odd to have two significant digits for the 1 degree and not for the 7% (which is approximated)

Thank you again for this suggestion. We have changed this.

- Line 20: Extreme precipitation events depend on other variables, not only moisture in the column. I suggest to slightly rephrase.

We have rewritten this. Please refer to the content starting at line 18, where we now write:
"The Clausius-Clapeyron (C-C) relationship defines the theoretical rate at which the moisture-holding capacity of the atmosphere scales with temperature. It states that there is approximately a 7% increase in moisture-holding capacity of the atmosphere for every 1◦C increase in temperature. Due to the C-C relationship, as available moisture in the air column increases with warmer temperatures, intensities of extreme precipitation rates are also expected to increase (Panthou et al., 2014; Westra et al., 2014; Wasko et al., 2015; Myhre et al., 2019; Fowler et al., 2021; Gründemann et al., 2022; Harp and Horton, 2022; Donat et al., 2023)."

- 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”).

We now more simply refer to either temperature or dew point temperature throughout the paper. To be clear, though, dew point temperature measurements do contain some information about temperature. It is true that with dew point temperature alone, you cannot know exactly what your corresponding temperature is. However, we do know that with a given measured dew point temperature value, the corresponding temperature value will be equal to or greater than the measured dew point temperature value.

- 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. We have removed the reference from the noted location.

- 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.

After reading again the literature that we cite in our paper, we have decided to define this term in the Data section of the paper. This is in line with how other authors has presented similar methods. Please refer to our comment above.

- 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.

This is a great suggestion. Our revised version of the paper now focuses almost exclusively on using quantile regression instead of the binning method. One exception is for the content found in the new Figure 2. This is when we point out and discuss the problems that can stem from mixing different climatologies when pooling raw (or non-normalized) data.

- 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?

In our paper, we investigate whether or not we can improve our estimates of our scaling rates, and the associated predictions of Rx1hr or Rx1day, if we pool normalized data. We find that normalized data, pooled from up to 50 or 100 km away and using a 3-month window, does provide more skillful predictions than using non-normalized data. We have focused on presenting the data and results in such a way as to highlight the impact of normalization. Normalizing the data is one way to handle the heterogeneity in the data. Classification can be another. We additionally handle heterogeneity in the scaling rates themselves, by implementing spatial and temporal windows listed above (e.g., ~50km and 3-month). We discuss other authors advocating for standardized or normalized data starting at line 57:
"When using pooled data with either the binning method or quantile regression, it is important to also recognize that climatological differences in both time and space can be present in the data (e.g., California has more extreme daily precipitation in the winter, at lower dew point temperatures, than it does in the summer). Molnar et al. (2015) clearly showed this impact, by fitting a regression model to larger sample of pooled data, and then comparing to regression fits which separate the same data by whether there was lightning or not. Using pooled data, then, without accounting for these climatological differences, one runs the risk of inaccurately estimating the effective scaling rates. In light of this problem, some have advised using normalized or standardized data (Zhang et al., 2017; Visser et al., 2021). An additional control for seasonality was proposed by Zhang et al. (2017), where they used normalized data over the summer season."

We also have the following statement written at line 222:
"By normalizing the data, we provide a more homogeneous framework for dealing with different types of precipitation (Berg and Haerter, 2013; Molnar et al., 2015)."

- 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.

We have worked to more clearly state what is being predicted and how we evaluate that skill. As we have already said above, we changed 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 Rx1hr 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 across different bins of dew point temperature anomalies.

- 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.

We have decided to remove the discussion of the hook pattern in our revised paper. We found that the content related to the hook pattern was detracting from our primary message. Our primary message being: We can leverage normalized data, in a way that we cannot with non-normalized data, to improve P-T scaling rate estimates and the associated predictions. Put another way, using non-normalized data does not allow one to reliably pool data from multiple stations and from multiple months of the year.

- Lines 198-205: this is a lot of text to say “precipitation is stochastic”.

Thank you for this comment. We have removed this content from the revised paper.

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

Thank you for this suggestion. Instead of concurrent (or 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. Please refer to line 208, where we write:
"Both Zhang et al. (2017) and Marra et al. (2024) have previously explored using predictor data which is not concurrent. Zhang et al. (2017) used only wet-days to compute an average seasonal dew point temperature, while Marra et al. (2024) used average daily temperatures."

- 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.

Our revised paper clearly defines how we apply the normalization. As we have already stated, we want to have the normalization in such a way so that we can readily fit a scaling rate which gives us %/°C. This is what we have done by computing normalized values of Rx1hr or Rx1day to be, % of normal, and by computing normalized values of dew point temperature as, degrees C from normal. The reviewer is correct that the variance of the data can also matter. That is part of the reason why, even when using the normalized data, we find optimal data pooling at spatial and temporal extents that are not all that large (i.e., ~50 to 100 km and 3-month). See Figures 7d and 7h.

- 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.

Thank you for this comment. We have softened our language regarding "effectively removing" the challenges. Please refer to the content starting at line 217, where we write:
"In the prior section, we pointed out a number of problems that can adversely influence our estimation of P-T scaling rates. For our modeling approach, we use data at the station/month resolution (one value per station per month). This gives us data that exhibits a greater degree of statistical independence than data at the hourly or daily resolution. Next, as we illustrated in Figure 2, raw or non-normalized can lead to over- or undersampling certain stations and/or months. This is due to the fact that there are climatological differences in time and space, and some stations in some months will on average have more extreme precipitation events than other station/months, even at the same dew point temperature. By normalizing the data, we provide a more homogeneous

framework for dealing with different types of precipitation (Berg and Haerter, 2013; Molnar et al., 2015). It is therefore advisable to normalize the data, following Eqs. 1 and 2, prior to estimating any P-T scaling rates."

- 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.

We have restructured the paper, and this specific content has been removed.

- Line 294: I don’t understand how this is possible. Perhaps the way daily and hourly maxima are defined is not sufficiently clear?

Please refer to our statement above and to the revised paper. We have attempted to provide more clarity and distinction between the hourly and daily datasets by using the terms Rx1hr and Rx1day.

- 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.

Thank you for this comment. We have put in a great deal of effort to make it more clear that we are dealing with extreme precipitation throughout the revised version of the paper. Notably, we now use the terminology, Rx1hr and Rx1day, to indicate hourly and daily extreme precipitation.

- Fig 7 and several other results/validation: why are 2 degree C bins used? What is the sensitivity of the outcomes to this choice?

Given the advice of the three reviewers, we have removed using the binning method to produce any of the results in the paper. It is only used to discuss the problems associated with pooling non-normalized data, illustrated in Figure 2. We are not producing any predictions using the binning method. Using different bin sizes in Figure 2 does not change the fact that extremes at certain stations, in certain months, and in a given dew point temperature bin, can be very different from other stations in either the same or different months. Please refer to the paragraph beginning at line 133.

- 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 have removed this plot from our revised paper. We now focus on model performance using leave-one-year-out cross validation for Rx1hr and Rx1day, in addition to a two-fold cross validation for Rx1day. Every model that we compare has access to precisely the same data and number of data points. The validation procedure is identical. As a result, we can observe what kinds of information can be most beneficial in making cross-validated predictions. We find that by normalizing the data, we can more effectively pool the data to improve predictions. We now include an improved view of what kinds of model skills can be achieved through 1000 randomized predictions, where each set of randomized predictions also preserves the spatio-temporal covariance structure of the observational Rx1hr or Rx1day data (see Figures 7 and 8).

- 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.

Thank you for this comment. We have worked to more clearly explain the circumstances in which we can more effectively use pooled data. The results that you point to in this comment can now be found in Figures 7 and 8 of the revised paper. We can take a look at the RMSE skill score of the models which used non-normalized and normalized data (for example, see Figs. 7c, 7d, 7g, and 7h). In Figures 7c and 7g, we show the skill scores of the model which uses non-normalized data to predict Rx1hr and Rx1day, respectively. As one traverses rightward and downward from the upper-left grid cell in these figures, the spatial and temporal extents over which the data is pooled increases. We can see that for the non-normalized data, the skill decreases as we increase the spatial

and temporal extents over which we are pooling data. The same cannot be said for the normalized data (seen in Figs. 7d and 7h), which sees improved skill with data pooling up to approximately 100 km and using a 3-month window. It is true that the skill of the normalized model cannot effectively pool data from all months in the year. Though, that makes sense because we are also showing in this paper that the scaling rates in January at a given station can vary dramatically with respect to July, for example. Therefore, it would not be wise to include either non-normalized or normalized data from January in order to estimate July scaling rates.

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.

We would like to thank the reviewer for their comments. They have greatly assisted us in improving our paper. In light of of all of the great comments from the three reviewers, we have heavily revised our 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.

We now include an Appendix A (see line 429) which provides an analysis of the quality controlled data.
Additionally, we have added a few lines in the conclusions raising the point about the measurement precision. We have now written at line 413:
"Another issue worth mentioning is the potential impact of the measurement precision of the precipitation data (Ali et al., 2022). In this study, we used data at a measurement precision of 0.1 mm. However, we did additionally try rounding our precipitation data to the nearest 1.0 mm prior to normalization, and this was not found to noticeably impact our results."

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.

Thank you for this comment. In light of your comments, and the comments of the other reviewers, we have removed using the binning method to produce any of the results in the paper. We are not producing any predictions using the binning method. It is only used to discuss some problems associated with pooling non-normalized data, illustrated in Figure 2.

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.

We should first be clear that every model that we compare has access to precisely the same data and number of data points, and our validation procedure is identical. As a result, we can observe what kinds of information can be most beneficial in making cross-validated predictions. We find that by normalizing the data, we can more effectively pool the data to improve predictions. It is true that we make an assumption that the scaling rates are stationary over the period of record. We have now written at line 409:
"We should note that our methodology has assumed that the scaling rates are stationary over our period of record, which appears to be a good assumption if we compare the results of leave-one-year-out and two-fold cross validation cases for Rx1day. Our focus, here in this paper, has been to illustrate the value of normalizing data in an effort to more accurately estimate P-T scaling rates. However, future research can focus on investigating whether or not scaling rates can be considered stationary."

Our goal in this paper has been to illustrate the effectiveness of leveraging normalized data in a way that the non-normalized data cannot be leveraged. With the normalized data, we can safely pool data up to ~100km and with a 3-month window. We believe the question of stationarity is not necessarily whether the data is stationary, but whether or not the scaling rates themselves can be considered stationary? We expect the data to be non-stationary. We expect that over long enough periods of time, that the shifts in mean dew point temperatures will lead to shifts in extreme precipitation, following the effective scaling rates that we are attempting to estimate. That said, we do additionally use a two-fold cross validation using the Rx1day data (see Figure 8). We find very similar results between the two validation cases, which indicates that trends from the data do not appear to bias the results of the leave-one-year-out validation case.

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.

We now implement a quantile regression throughout our methods. The primary aim of our paper is to provide a method that shows the improvement in scaling rate estimation when we use normalized data versus non-normalized data. We successfully demonstrated this result with robust and skillful predictions.

Additionally, we now include an improved view of what kinds of model skills can be achieved through 1000 randomized predictions, where each set of randomized predictions also preserves the covariance structure of the observational Rx1hr or Rx1day data (see Figures 7 and 8). While outside of the scope of our current study, future research can compare the effectiveness of other modeling approaches to what we have proposed here.

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 great point. We have changed the exponential function to more explicitly have it to be interpreted as “% per °C”. This can be done with a model taking the form $y = ab^x$, where x are the monthly dew point temperature values, a is a multiplicative offset, and b provides information concerning the scaling rate. See Equation 6 from the revised paper.

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.

As we stated earlier, we now include some analysis of the quality-controlled data. We have added the content starting at line l03:
"We additionally only computed the normalized Rx1hr or Rx1day if the mean at station, x, and month, m, is greater than 1.0 mm. This helps us to avoid infinite or unrealistic values in the normalized data."
Again, we did not find the presence of any abnormally large or unrealistic values.

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.

We are trying to provide a better understanding of which data can be best leveraged to effectively estimate P-T scaling rates. Additionally, we advocate for estimating scaling rates for each station and each month of the year. Our entire framework is build around "P-T", which is looking at the empirical-statistical relationship that a single predictor, "T" (such as dew point temperature), has on extreme precipitation, "P". One should exercise caution when interpreting any one single predicted versus observed value. That falls more under the umbrella of numerical weather prediction, and obviously depends on many different variables. That is why we have advocated for the validation framework that we

do, which evaluates the performance over many samples. In climate change studies, it is common to look at the isolated influence that one variable could have on another. How might a warmer climate impact snow depth? How might warmer sea surface temperatures in the "main development region (MDR)" impact hurricane intensity? Of course, a given value of seasonal snow depth at a particular location, or whether or not a specific hurricane makes landfall depends on many factors. However, we want to isolate the averaged response of extreme precipitation to increasing temperatures or dew point temperatures, and this is the approach we have taken with P-T scaling.

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.

In our revised paper, it is true that the Rx1hr and Rx1day arrays are not completely statistically independent. However, as we discuss in the results section related to Figures 7 and 8, we generated 1000 random predictions, and we compared their skill scores to what we observed with our modeling framework. Importantly, the randomized predictions also preserve the spatio-temporal covariance structure of the observational Rx1hr or Rx1day data (see Figures 7 and 8). Those skill scores of the randomized predictions do not achieve anything close to what we find with the other modeling approaches outlined in the paper. Not shown in our results, is that we additionally tried randomly selecting five-year segments of data from the calibration period as predictions. We did not find those results to be any different that the black dotted lines in Figures 7 and 8. As a result, we found that longer-term climate modes such as ENSO do not appear to influence our results.