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.

The authors would like to thank the reviewer for their time and effort in providing useful feedback concerning our paper. In line with the reviewer's point, we will endeavor to make changes to our paper to better position our work with respect to prior research/literature. Also, we will reframe the novelty of tackling each individual shortcoming but rather clarify that it is the combination of the three shortcomings that we attempt to respond to in a pragmatic and systematic manner. We appreciate that the reviewer acknowledges this general aim as important and relevant.

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?

We will aim to provide in our revised paper a better review of where our study fits in with what has already been done. We will add references to these studies, which we build upon, earlier in the manuscript.

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) and highlight our adaptations to their approach in a clearer manner.

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.

Thank you for pointing out this issue. We had presented and used the binning method as many previous and current studies continue to use it. 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.

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

In our paper, we had already acknowledged that Zhang et al. (2017) uses seasonal data. However, we will add in an additional credit to their work in the revised paper.

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. We will work to better acknowledge the great work that has already been done by these authors and highlight our own addition to it by combining and expanding it.

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".

In this paper, we are building to the result that scaling rates need to be considered different between regions and across different seasons. It is precisely the methodology that we propose in this paper, where we suggest estimating different scaling rates for each month of the year across a region. Our method is

the act of computing scaling rates which rely on temporally independent data which have been normalized station-by-station and month-by-month. Furthermore, we propose only using data within a certain seasonal window, or up to a certain spatial distance, to fit the model. Doing this gives different results, with better model performance, than fitting the model to all of the data throughout the year. This approach of normalizing the data and only using data within a certain spatial and temporal window is central to our proposed methodology. The scaling rates in Figure 12 are the result of our proposed methodology, and if we were to compare model performance against the scaling rates from Figure 12, then we end up evaluating model performance against itself. That said, we will also include in our revised paper a benchmark model that relies on a scaling rate which is fit to all of the non-normalized data from the Upper Colorado River Basin (UCRB). We will, therefore, compare to 1) climatology, 2) a 7%/deg (theoretical benchmark), and 3) a benchmark scaling rate specific to the UCRB. That way, we can compare to an additional benchmark reference model which is specific to the UCRB, but it is not implementing the very approach which is central to our paper.

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 thank the reviewer for this comment. Our revised paper will include a new title that more accurately reflects our work.

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.

We are comparing against the theoretical Clausius-Clapeyron, which to our understanding does not change as a function of temporal scale (e.g., sub-daily to daily). However, as we stated above, we are going to provide another benchmark scaling rate for the UCRB using both hourly and daily data which has not been normalized.

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?

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. We will add that the normalization also helps us address different underlying storm types that are predominant in different times of the year.

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.

Indeed, we agree that the word "incorrect" might be too strong for what we try to address. We will adopt the wording to both make clear that predictability depends partly on the method used without attempting to judge on absolute correctness of one method versus another.

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?

Thank you for this point. We can improve our discussion concerning this point in the text.

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