~~Precipitation-temperature~~ Leveraging normalized data to improve point-scale estimates of precipitation-temperature scaling ~~: current challenges and proposed methodological strategies~~rates

Matthew Switanek[1], Jakob Abermann[1], Wolfgang Schöner[1], and Michael L. Anderson[2]

[1]Department of Geography and Regional Science, University of Graz, Graz, Austria
[2]California Department of Water Resources, Sacramento, California, United States

Correspondence: Matthew Switanek (matthew.switanek@uni-graz.at)

Abstract. Sub-daily to daily extreme precipitation intensities are expected to increase in a warming climate, consistent with the Clausius-Clapeyron (~~CC~~C-C) relationship, which predicts a ∼7% increase in atmospheric moisture-holding capacity per °C of warming. Many studies have benchmarked observed extreme precipitation–temperature (P–T) scaling rates against this theoretical value, finding that global averages align closely with ~~CC~~C-C, while regional and seasonal estimates often diverge substantially. Significant challenges remain, however, in accurately estimating and interpreting P–T scaling rates, particularly at point scales. In this study, we use observational station-based data from the Upper Colorado River Basin to ~~explore~~ illustrate these challenges and propose methodological improvements. Specifically, we compare multiple approaches, including those using raw (non-normalized) and normalized data, to estimate P–T scaling for hourly and daily extreme precipitation. Model performance is assessed using a cross-validation framework. Our results demonstrate that ~~normalizing data , independently for every station andeach calendar month,~~ when estimating P-T scaling rates using data pooled from multiple stations and/or months, it is essential to account for spatial and temporal climatological variability. ~~Without normalization, estimated scaling ratescan be inaccurate and misleading~~We find that using normalized data allows us to more effectively leverage pooled data, and thus improve our estimates of P-T scaling rates.

1 Introduction

Climate change is expected to increase the intensity of extreme precipitation events lasting from sub-daily to daily timescales ~~across many regions of the world (Lenderink and van Meijgaard, 2008; Lenderink et al., 2011; Ban et al., 2015; Tabari, 2020; Abatzoglou~~ (Lenderink and van Meijgaard, 2008; Lenderink et al., 2011; Ban et al., 2015; Tabari, 2020; Abatzoglou et al., 2022; Ali et al., 2022; Chia . This increase can primarily be attributed to the increased moisture-holding capacity of a warmer atmosphere (Alduchov and Eskridge, 1996; Allen and Ingram, 2002; Lenderink and van Meijgaard, 2010; Huang and Swain, 2022; Gu et al., 2023; Trenberth et al., 2003; Rahat et al., 2024). 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 ~~1.0~~1°C increase in temperature. ~~Extreme precipitation events depend on the~~ Due to the C-C relationship, as available moisture in the air column ~~, and hence, the intensities of~~

those extremes 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

.

Over the last couple of decades, a growing body of research has emerged concerning the estimation and application of precipitation-temperature (P-T) scaling rates (Lenderink and van Meijgaard, 2008; Berg et al., 2009; Lenderink and van Meijgaard, 2010; Lenderink et al., 2011; Ban et al., 2015; Prein et al., 2017; Fowler et al., 2021; Ali et al., 2022; Dollan et al., 2022; Marra et al., 2024). Many studies have used observations in an effort to better quantify P-T scaling rates (Jones et al., 2010; Lenderink et al., 2011; Utsumi et al., 2011; Ali et al., 2018; Wasko et al., 2018; Ali et al., 2021; Najibi and Steinschneider, 2023), while others have investigated P-T scaling rates using climate model data (Ban et al., 2015; Drobinski et al., 2018; Meresa et al., 2022; Donat et al., 2023; Jong et al., 2023; Martinez-Villalobos and Neelin, 2023; Chiappa et al., 2024; Estermann et al., 2025; Higgins et al., 2025). Past efforts to better quantify and/or estimate P-T scaling rates can be further separated by the choices of temporal and spatial extents. Some research has concentrated on daily extreme precipitation (Utsumi et al., 2011; Ali et al., 2018; Yin et al., 2021), while others have focused on hourly extreme precipitation (Lenderink et al., 2011; Prein et al., 2017; Ali et al., 2021)(Lenderink et al., 2011; Prein et al., 2017; Ali et al., 2021; Haslinger et al., 2025

. Likewise, the spatial extent of some prior work has been at the global scale (Ali et al., 2018; Tabari, 2020; Tian et al., 2023), while others have investigated scaling rates at the point or regional scale (Jones et al., 2010; Drobinski et al., 2018; Najibi et al., 2022; Martinez-Villalobos and Neelin, 2023).

Estimates of P-T scaling rates have primarily been obtained by conditioning extreme precipitation on either 2-meter air temperature (i.e., dry-bulb temperature) (Jones et al., 2010; Utsumi et al., 2011; Panthou et al., 2014; Wasko et al., 2015; Prein et al., 2017; Li et al., 2023; Marra et al., 2024) or 2-meter dew point temperature (Lenderink and van Meijgaard, 2010; Zhang et al., 2017; Wasko et al., 2018; Najibi and Steinschneider, 2023). For the most part, studies which have used dew point temperature have found greater consistency and more robust relationships than when using air temperature (Lenderink and van Meijgaard, 2010; Lenderink et al., 2011; Ali and Mishra, 2017; Wasko et al., 2018; Barbero et al., 2018). This can be attributed to the fact that dew point temperature also contains information concerning the available moisture in the atmosphere.

In addition to decisions concerning which data to use, prior reseach has also proposed a variety of methods to estimate P-T scaling rates. One Two of the most widely used approaches is are the binning method (Lenderink et al., 2011; Prein et al., 2017; Ali et al., 2

. The binning methodhas been applied, for example, to investigate how extreme daily precipitation changes as a function of (Lenderink et al., 2011; Prein et al., 2017; Ali et al., 2018; Drobinski et al., 2018; Fowler et al., 2021; Gu et al., 2023; Tian et al., 2023) and quantile regression (Wasko and Sharma, 2014; Ali et al., 2018, 2021; Gu et al., 2023; Marra et al., 2024). Visser et al. (2021) effectively pointed out the adverse impact that sample size can play in the binning method, whereby there are many more samples in the central and often-recorded temperature (or dew point temperature(Ali et al., 2018; Wasko et al., 2018). Often, the binning method is used with) bins and many fewer samples in the bins for the tails of the temperature distribution. Given this issue of varying sample sizes across temperature bins, a number of prior studies have advocated for using quantile regression (Wasko and Sharma, 2014; Molnar et al., 2015; Ali et al., 2018), which fits a function to data in a specified quantile,

such as the top 1% of precipiation. Implementation of either of these methods often relies on data pooled from ~~more than one station in the same region~~ multiple stations and across different times of the year (Utsumi et al., 2011; Drobinski et al., 2018). Pooling data in this manner can leverage an increased sample size in an effort to improve the robustness of the ~~estimates (Molnar et al., 2015; Ali et al., 2021). The binned averages, with or without pooling, provide estimates of how extreme precipitation scales as a function of the chosen temperature variable. It~~ estimated P-T scaling rates (Molnar et al., 2015; Ali et al., 2021).

When using pooled data with either the binning method or quantile regression, it is important to ~~recognize, however,~~ also recognize that climatological differences in both time and space can be present in the data ~~across both time and space~~ (e.g., California has more extreme daily precipitation in the winter, at lower dew point temperatures, than it does in the summer). ~~Due to~~ 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 ~~apparent scaling ratesif data normalization is not applied. Notably, Zhang et al. (2017) proposed a method which uses normalized data computed at the station-level across a three to four month season. Their method removes many of the climatological differences that are present in the data. However, even climatological differences across different months within a 3-month season can potentially have a large impact on the estimated scaling rates.~~ 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.

~~Another topic that has recieved significant attention is~~ Our work herein is built on the ~~so-called "hook" or peak structure which has been shown to be present in scaling rates at higher temperatures (Prein et al., 2017; Wang et al., 2017; Drobinski et al., 2018; Yin . This hook can be described as a shift from positive to negative scaling rates. A number of studies have found the "hook" pattern to be less prevalent when using dew point temperatures instead of air temperatures (Lenderink and van Meijgaard, 2010; Lenderink et al., 20 . However, some cases still show a "hook" pattern in the scaling rates which have conditioned extreme precipitation on dew point temperature (Lenderink and van Meijgaard, 2010; Lenderink et al., 2011; Panthou et al., 2014; Tian et al., 2023; Sokol et al., 2024) . Different physical mechanisms, such as limited moisture availability and atmospheric dynamics, have been proposed to explain the transition from positive to negative scaling (Berg et al., 2009; Jones et al., 2010; Utsumi et al., 2011; Molnar et al., 2015; Sun a . Boessenkool et al. (2017) have shown that sample size can also play a key role in the robustness of the estimates at higher temperatures.~~

~~Our aims in~~ foundation of prior work (Zhang et al., 2017; Wasko and Sharma, 2014; Molnar et al., 2015; Visser et al., 2021) . The aims of this paper are twofold. First, we ~~use the binning method to illustrate and~~ describe some common challenges or problems that ~~exist when estimating and interpreting~~ must be addressed in order to effectively estimate and interpret P-T scaling rates. Second, we suggest a methodology to resolve many of those problems~~which is an extension of the work by Zhang et al. (2017)~~. Lastly, we use ~~three different methods to produce estimates of~~ quantile regression to estimate P-T scaling rates ~~, and~~ under different modeling assumptions. Furthermore, we use these estimates to generate predictions of extreme hourly and daily precipitation through a cross-validated framework. The skill of these predictions are subsequently evaluated ~~against climatology and against assuming a theoretical Clausius-Clapeyron scaling rate. This allows us to quantify the added~~

value of certain methods over others with respect to our ability to predict changes to extreme precipitation as a function of changes in dew point temperature to determine which methods and/or assumptions provide the best performance.

2 Data

Hourly measurements of precipitation and dew point temperature for the Upper Colorado River Basin (UCRB) are obtained from the GH2D-MetNet dataset (Switanek, 2025), which is a quality controlled, global dataset of observed precipitation, temperature, and dew point temperature derived from the Global Historical Climatology Network - hourly (GHCN-hourly or GHCNH) dataset (Smith et al., 2011). Hourly data is used beginning at 00:00, January 1, 1951 and ending at 23:00, December 31, 2023. 2024. The dataset is relatively sparse until around the year 1999, when the density of the stations increases. The spatial distribution of these stations can be observed in Fig. 1a.

Daily measurements of precipitation are taken from the Global Historical Climatology Network - daily (GHCN-daily or GHCND) dataset (Menne et al., 2012). Many Most of these stations do not measure dew point temperature in-situ. Therefore, the ERA5 Reanalysis dataset (Hersbach et al., 2023) is used along with GHCN-daily to provide dew point temperatures at the GHCN-daily GHCND stations. For each GHCN-daily GHCND station, the nearest ERA5 grid cell is found, and the corresponding time series of dew point temperatures are used for that station. This procedure is repeated for all of the GHCN-daily GHCND stations in the UCRB. We use a common period of record for the hourly and daily data between January 1, 1951 through December 31, 2023. 2024. The distribution of the daily stations can be observed in Fig. 1b.

The hourly and daily datasets have been quality controlled to remove statistical outliers in both space (i.e., regional outliers) and time (i.e., temporal outliers). In Appendix A, we provide a comparison of the two quality-controlled datasets.

Throughout the paper, we use different data and/or indices. We begin by exploring the relationship between the raw (non-normalized) dew point temperature and raw precipitation data at the hourly resolution. Later in the paper, we rely on data at the station/month resolution, with one data point per station per month. For the hourly dataset, we find the maximum hourly precipitation amount at each station and for each month. For the duration of the paper, we refer to this value as Rx1hr. Also for the hourly dataset, we find for each station/month the concurrent dew point temperature which corresponds to the same time step as a given Rx1hr value. Additionally, we compute average monthly dew point temperature for each station/month. Likewise, for the daily dataset, we find the maximum daily precipitation amount at each station/month, referred to as Rx1day for the duration of the paper. Average monthly dew point temperatures are obtained from ERA5.

We additionally make use of normalized dew point temperature and precipitation data. Normalized anomalies of dew point temperature are computed as,

$$\mathbf{DPT}^{*}_{x,m,t} = \mathbf{DPT}_{x,m,t} - \overline{\mathbf{DPT}}_{x,m}, \tag{1}$$

where $\mathbf{DPT}_{x,m,t}$ is the average monthly dew point temperature (or concurrent dew point temperature) at station, x, month, m, and year, t, and $\overline{\mathbf{DPT}}_{x,m}$ is the mean dew point temperature over the calibration time period at station, x, and month, m.

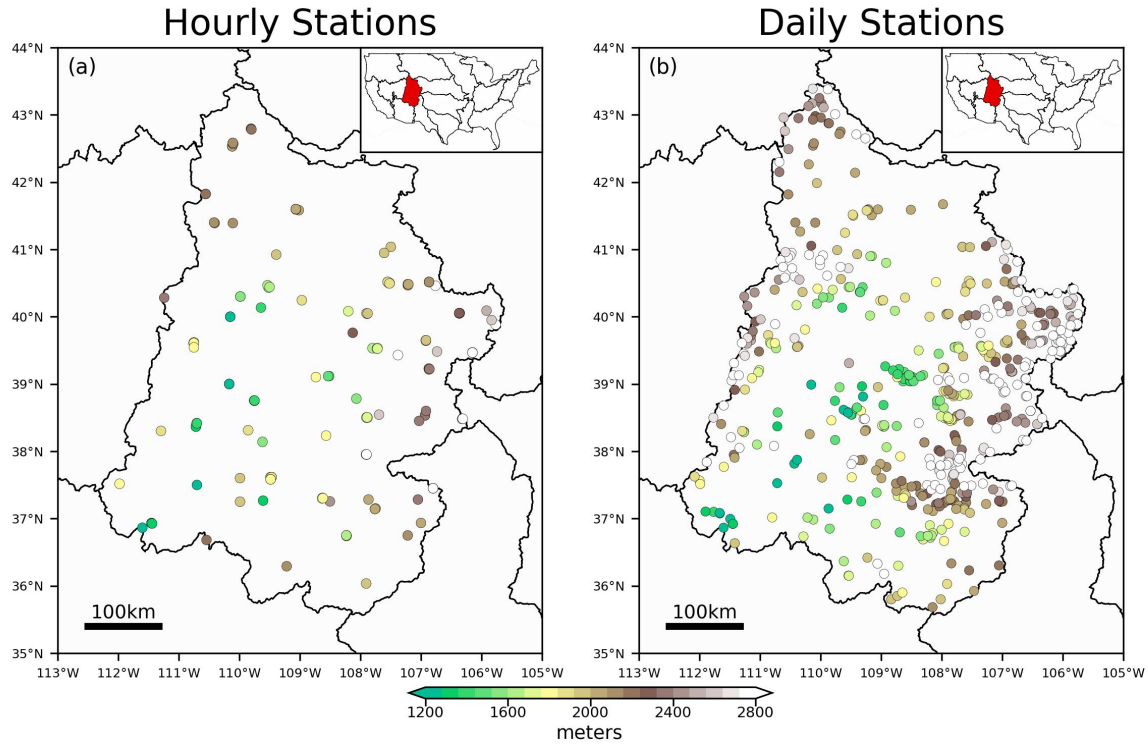

Figure 1. The hourly (a) and daily (b) distribution of stations across the Upper Colorado River Basin which are used in this study. The location of the study region is shown ~~, along with~~ as the inset of the subplots alongside other large-scale hydrological basins ~~in~~ across the contiguous United States~~, in the inset subplots in the upper right~~. The color of the stations corresponds to the station elevation.

Similarly, normalized anomalies of precipitation are computed as,

$$\mathbf{P}^{*}_{x,m,t} = \frac{\mathbf{P}_{x,m,t}}{\overline{\mathbf{P}}_{x,m}} \cdot 100 \ , \tag{2}$$

where $\mathbf{P}_{x,m,t}$ is either Rx1hr or Rx1day at station, x, month, m, and year, t, and $\overline{\mathbf{P}}_{x,m}$ is the mean of the respective precipitation (either Rx1hr or Rx1day) time series over the calibration time period at station, x, and month, m. 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.

2.1 Evaluation Metrics Used for Validation

We evaluate model performance ~~across a range of different cases~~ using the root mean squared error skill score ($\mathbf{SS}_{RMSE}$). The skill score, $\mathbf{SS}_{RMSE}$, is a function of the model and reference RMSE errors ($\mathbf{RMSE}_{MOD}$ and ~~$\mathbf{RMSE}_{REF}$~~ $\mathbf{RMSE}_{CLIM}$,

respectively). The $\mathbf{RMSE}_{MOD}$ is defined as,

$$\mathbf{RMSE}_{MOD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{mod,i} - y_{obs,i})^2} \quad . \tag{3}$$

where y_{obs} and y_{mod} are the observed and the modeled precipitation values, respectively. Likewise, ~~$\mathbf{RMSE}_{REF}$~~ $\mathbf{RMSE}_{CLIM}$, reflects the error associated with a ~~reference or baseline model .~~

~~$$\mathbf{RMSE}_{REF} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{ref,i} - y_{obs,i})^2} \quad .$$~~

~~where y_{ref} contains the reference precipitation values. We compare against two reference modeled values. First, we compare whether the model predictions are more skillful than~~ baseline model which always assumes climatology (i.e., always assuming 100% of normal, or equal to the climatological mean). ~~And second, we compare whether the model predictions are more skillful than if we had assumed a theoretical Clausius-Clapeyron (CC) relationship (i. e., using 7% per °C).~~

$$\mathbf{RMSE}_{CLIM} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{clim,i} - y_{obs,i})^2} \quad . \tag{4}$$

where y_{clim} contains the benchmark climatological precipitation values. The skill score can then be obtained as,

$$\mathbf{SS}_{RMSE} = 1 - \frac{\mathbf{RMSE}_{MOD}}{\mathbf{RMSE}_{REF}}\frac{\mathbf{RMSE}_{MOD}}{\mathbf{RMSE}_{CLIM}} \quad . \tag{5}$$

Skill scores of $\mathbf{SS}_{RMSE}$ above zero indicate that the model predictions are more skillful than ~~the reference predictions~~climatology, while scores below zero indicate that the model is performing worse than ~~the reference~~climatology.

3 Methods

3.1 Common ~~Methodological~~ Challenges ~~Pertaining to the Interpretation of~~ in Estimating P-T Scaling Rates

~~Ultimately, our goal~~ Our goal in this paper is to have a methodological approach that can more accurately estimate P-T scaling rates, and furthermore to use those estimates to make skillful predictions of extreme precipitation. We want to be able to estimate scaling rates with sufficient enough spatial and temporal resolution in order to say, for example, that a particular region (or station) in a given ~~month or season~~ time of year (or month) can expect some ~~specified~~ average percentage increase in extreme precipitation ~~, on average,~~ provided that the region experiences, for example, +1°C of warming in dew point temperature. The percentage increase per °C is our scaling rate and it will vary as a function of space (e.g., ~~some chosen region or basin~~between stations or between grid cells) and time (e.g., ~~month~~between months of the year).

Without carefully managing the underlying data, one can incorrectly estimate and interpret P-T scaling rates. There are ~~three~~ two primary concerns that must be addressed prior to using the data to estimate scaling rates:

1. ~~Using raw , or~~ Pooling raw (non-normalized~~, values~~) data of dew point temperatures and precipitation ~~rates~~ across multiple stations and/or months can lead to an inaccurate estimate of a scaling rate due to climatological differences that exist in both space (from station to station) and time (from month to month).

2. ~~Differences in sample sizes.~~

3. Data at hourly or daily resolutions cannot be assumed to be ~~temporally independent~~ statistically independent in time.

In Figure 2, we ~~plot an example of~~ illustrate the first of the aforementioned challenges. ~~This challenge relates to the conflation of different climatologies that can arise in space and time when using measured data that has not undergone any normalization.~~While we primarily implement a quantile regression methodology throughout our paper, here we apply a binning method in order to more clearly show the problems of pooling raw (non-normalized) data from across space (e.g., different stations) and time (e.g., months). In Fig. ~~2a, measured values of dew point temperature and precipitation rates from multiple stations, and across multiple months, are plotted together using a binning method. Figure~~ 2a~~plots~~ , we use all of the ~~pairings of hourly dew point temperature and precipitation rates using~~ hourly data from all of the stations ~~that fall~~ within the UCRB to look at pairs of dew point temperature and precipitation which are concurrent or collocated in time. The top ~~1.0% and top 0.1% of precipitation events are shown for~~ 1% of precipitation intensities are shown in blue for each dew point temperature ~~bins, where the bins are iterated for every 1°C with~~ bin (using a bin size of 2°C ~~(e.g., centering at 10°C and using values between 9°C~~ and ~~11~~incrementing by 2°~~C, then stepping to 11°C). The observed scaling rates for the top 0.1% and 1.0% are plotted as the solid and dashed-dotted red lines, respectively. The red lines are calculated using the average precipitation of the points that fall in the top 0.1% and 1.0% of each bin. Note, we found that using the median precipitation of the points (which is not shown), instead of the average precipitation, did not change the shape of the scaling rate curves.The binning method which uses the measured data, seen in Fig.2a, shows a "hook" pattern at higher~~ C). One problem that can arise when finding the extreme precipitation values (e.g., the top 1%), is that there can be certain months that are more extreme at the same dew point temperatures ~~, where the scaling rate transitions from positive to negative (Lenderink et al., 2011; Tian et al., 2023; Visser et al., 2021; Yin et al., 2021; Sokol et al., 2024). This "hook" pattern was also found using different bin sizes. The presence of this "hook" complicates the interpretation of the associated scaling rate, especially at these higher~~ than other months. Similarly, there can be some stations that are more extreme at the same time of year, and at the same dew point temperatures~~. It is unclear how exactly one should extrapolate new precipitation extremes given new high-valued dew point temperature extremes. The very presence of the "hook" or peak pattern, however, could simply be the result of different stationsand different months, all with different climatologies, being plotted together~~, than other stations. We can test if this is the case ~~here by removing climatological differences in the underlying data. To do this, we can compute z-scores~~ by implementing the binning approach at the station/month level. So, for each station we can use all of the hourly ~~time series, month by month and station by station. Hourly z-scores of~~ data for a given month (all of the Julys for example). Then, we find the top 1% of precipitation for each station/month at different dew point temperature ~~are computed as,~~

$$\mathbf{DPT}^{z}_{x,m,t} = \frac{\mathbf{DPT}_{x,m,t} - \overline{\mathbf{DPT}}_{x,m}}{\sigma_{\mathbf{DPT}_{x,m}}},$$

where $\mathbf{DPT}_{x,m,t}$ is the bins. This gives us extreme precipitation amounts that are specific to every station, every month of the year, and every dew point temperature at station x, month$m$, and hour t, and $\overline{\mathbf{DPT}}_{x,m}$ and $\sigma_{\mathbf{DPT}_{x,m}}$ are the mean and standard deviation of the dew point temperature time series at station x and month m, respectively. The maximum number of data points in the array, $\mathbf{DPT}_{x,m}$, for the month of July is 54bin. We can then plot the influence of climatological differences across time as seen in Fig. 2b. Using these top 1% of precipitation values that are derived at each station for each month, 312 ((24 hours)x (31 days) x (73 years)). Similarly, hourly z-scores of precipitation are computed as,

$$\mathbf{P}^z_{x,m,t} = \frac{\mathbf{P}_{x,m,t} - \overline{\mathbf{P}}_{x,m}}{\sigma_{\mathbf{P}_{x,m}}} \quad ,$$

where $\mathbf{P}_{x,m,t}$ is the precipitation rate at station x, month m, and hour t, and $\overline{\mathbf{P}}_{x,m}$ and $\sigma_{\mathbf{P}_{x,m}}$ are the mean and standard deviation of the precipitation time series at station x and month m, respectively. Figure Fig. 2b plots all of the standardized pairings (from Eqs. 4 and 5)of hourly illustrates how extreme precipitation differs between the month of May and the month of July. Within the same dew point temperature and precipitation rates using all of the stations that fall within the UCRB. The same data is used to produce Figs. 2a and 2b. However, bin, for example, which is centered at 5°C (between 4°C and 6°C), extreme precipitation is dramatically different between these two months. For the month of May, the data in Fig. 2b has undergone a transformation to remove climatological differences in the underlying data. The "hook" pattern from Fig. 2a is not present in Fig. 2b. We can clearly observe, in this case, that the "hook" structure in Fig. 2a can be attributed to climatological differences of the underlying data in space and time. After removing these variations in climatology, extreme standardized anomalies of precipitation are found to increase across the entire range of standardized anomalies of dew point temperature. Note, average precipitation that falls in the top 1% of this bin is 3.60 mm/hr (with 171 samples). In contrast, the z-scores of precipitation in Fig. 2b are quite large. This is due to fact that Eq. average precipitation in the top 1% of the same bin is 0.36 mm/hr for the month of July (with 252 samples). The average extreme precipitation in May at a dew point temperature of 5is computed using the entire time series, which contain many zeros. We also restricted the analysis to using only the data where there was positive precipitation, and we found the same result where precipitation anomalies are found to continue to increase °C is 10 times that of July (the distributions are statistically significantly different with a p<0.01). Next, let us focus on the influence of climatological differences across space. In Fig. 2c, the top 1% of precipitation values are plotted for the month of July, but the stations are split by elevation as being either below or above 1800 meters. Again, we can use a bin, such as the one centered at 11°C, to observe that even at the highest same dew point temperature anomalies.

(a) All of the pairings of measured hourly precipitation along with the corresponding dew point temperature for the stations in the UCRB are plotted in the background as the black scatter points. The top 0.1% and 1.0% of precipitation rates are plotted as the light blue and dark blue scatter points, respectively. The top 0.1% and 1.0% are for 2-degree bin windows with 1-degree increments. The solid and dashed-dotted red lines are the average precipitation rates for the top 0.1% and 1.0% of values that fall within each of the bins. (b) The same as (a), except that the data has been standardized (i.e., z-scores).

Figure 3 plots the number of points that fall within the top 0.1% of precipitation values for each of the dew point temperature bins in Fig. 2a. These different number of points corresponding to the bars in Fig. 3 are the various sample sizes used to compute the scaling rates via the binning method. One can clearly observe that the number of extreme precipitation events that

~~reside in the bin centered about 16~~and at the same time of year, certain stations exhibit more extreme precipitation than others. For stations above 1800 meters, the average precipitation that falls in the top 1% of this bin is 5.83 mm/hr (with 163 samples). In contrast, the average precipitation in the top 1% of the same bin is 1.49 mm/hr for the stations below 1800 meters (with 118 samples). For the month of July, the average extreme precipitation for stations above 1800 meters and with a dew point temperature of 11°C is ~~much less than the number of extreme precipitation events that fall within the bins between -5°C and 0°C. There are roughly 50 times the number of events at -3°C than at 16°C or 17°C.~~4 times that of stations below 1800 meters ($p<0.01$). These climatological differences in both time and space adversely influence the sampling frequencies, when applying a binning method or quantile regression using pooled raw (non-normalized) data. Some months will be sampled at higher rates, such as May as opposed to July. Similarly, some stations will be sampled at higher rates, such as stations above 1800 meters in elevation as opposed to stations below 1800 meters. This influence leads to inconsistent sampling, where some stations, at some times of year, might never be sampled from, while others are sampled at greater than our specified quantile (e.g., 1%). As a result, ~~the binning method is less robustly estimating what is considered "normal" for the top 0.1% of precipitation events for certain dew point temperatures than for others~~we would estimate a scaling rate using data that more heavily weights certain stations at certain times of year. By sampling the top 1% more heavily from stations above 1800 meters in May versus stations below 1800 meters in July, for example, we end up overfitting or underfitting the scaling rate to data at certain locations and at certain times of the year.

~~The number of samples, or the number of light blue data points, that fall in the top 0.1% for each of the dew point temperature bins from Figure 2a.~~

~~Due to the fact that both the dew point temperature and precipitation timeseries are positively autocorrelated~~Moving to the next issue, the data used in Figure 2 cannot be considered to be statistically independent in time. The data from one time step at a given station is not independent of the data at the next time step at the same station. Due to this lack of data independence, the effective sample sizes are actually much less than what is suggested in Fig. ~~3. Focusing on the precipitation time series, Figure 4a plots the autocorrelation of these time series, station by station,~~ 2. We can show the issue with statistical independence in two different ways using the hourly ~~data set. For each station~~precipitation data. First, we can ~~take a subset of the time series for which the current hour and the subsequent hour both contained data , and we can compute the~~ compute the temporal autocorrelation of the data for each station. We do this using two different lagged time steps, where we compute lagged-1 ~~autocorrelation. The~~ hour and lagged-4 hour autocorrelations. The lagged autocorrelations for all of the stations in the UCRB, which have been sorted, are are plotted as empirical cumulative distribution ~~function (CDF) of all of the different station autocorrelations from Fig. 4a is plotted~~ functions (CDFs) in Figure 3a as the thicker ~~black line in Fig. 4b. To determine the statistical significance of these autocorrelations, we additionally ran 100 randomized simulations. For each simulation, we randomly generated a shuffled realization of the precipitation data at each station, drawing upon~~ lines on the right. Next, we test whether these distributions are statistically significantly greater (or shifted to the right) from what we would expect by chance. We generate 1000 randomly generated time series for all stations. At each station, bootstrapping is used to randomly create a time series of data points drawing from the empirical distribution of ~~precipitation at that station, and computed the corresponding autocorrelation~~data points from that station. The CDFs of ~~these 100 simulations~~ the stations from these randomly

generated time series are seen as the ~~thin colored lines in Fig. 4b. The expected autocorrelation is zero through the null hypothesis, and we find that the observed autocorrelation of the hourly precipitation data set~~ thinner lines on the left side of Fig. 3a. Precipitation data which is separated by less than 4 hours cannot be considered to be statistically ~~significantly greater than this null hypothesis~~ independent of one another ($p<0.01$). ~~Figures 4c-4d more specifically investigate the conditional probabilities of the~~ In Fig. 3a, all of the hourly precipitation data is used, including zeros, and as a result it does not provide information specifically related to the independence of extreme precipitation events~~themselves. Using the hourly data at one stationas an example, we look at what the probability of having an extreme precipitation event~~. Therefore, we can next compute conditional probabilities for extreme precipitation values. Let us first focus on precipitation events which fall in the top 1%. For each station, we find all of the cases where if precipitation at hour, t, was in the top ~~0.1%for some specified~~ 1%, then what was the probability that hour, $t+1$, ~~given that the prior hour, t~~or hour, $t+4$, was also ~~an extreme event~~ in the top ~~0.1~~1%. The ~~spatial distribution of those probabilities are plotted for the hourly data set~~ thicker lines in Fig. ~~4c. The CDFs of the probabilities from Fig. 4c are plotted as the thick black line in Fig. 4d. Again, the probabilities are statistically significantly greater than what we would expect by chance~~ 3b show the CDFs of these two conditional probabilities across the stations. Again, we find the conditional probabilities of the extreme precipitation data to be greater than randomly generated data ($p<0.01$)~~, as seen by the thin colored lines. In fact, the average probability that an hourly value of precipitation at a particular station would be in the~~ . One can note that the lagged-1 conditional probability of the data is approximately 40 times what we would expect by chance (compare 0.4 as the approximate average of the thick blue line versus 0.01 as expected by chance). Fig. 3c shows the results using the top 0.1~~%, given that the previous hourly precipitation at that same station was also in the top 0.1%, is greater than 100 times more likely than if the data were temporally independent (averge probability of the black line in Figure 4d, which is 0.14, versus the event likelihood, which is 0.001). Figures 4e-4h show the same as Figs. 4a-4d, except now using the daily dataset. Note, that the colorbar ranges are different between Figs. 4e and 4a, and similarly between Figs. 4g and 4c. This indicates, as we would expect, that there is a higher degree of autocorrelation in the hourly temporal resolution than at the daily resolution. That said, the correlations in Fig. 4e and the probabilities in Fig. 4g, which use the daily data, are also found to be statistically significantly greater than the null hypothesis ($p<0.01$). For the daily data, it is more than 50 times as likely, than by chance alone, that a daily event at a particular station will be extreme given the prior day was also extreme. Adjacent hourly and daily values of precipitation, and dew point temperature for that matter, cannot be considered statistically independent. As a result, if one were to apply an approach such as the binning method, one cannot include all of the hourly or daily data points and treat them as statistically independent events.~~ % of precipitation, where again we find that extreme precipitation events cannot be considered independent of one another over time periods of less than 4 hours. As one continues to increase the duration of time between extreme precipitation events, beyond 4 hours for example, then those events begin to exhibit more statistical independence from one another.

~~Still another issue that requires consideration is the collocation in time of dew point temperatures at a given hour or day along with the maximum hourly or daily precipitation rates. Ideally, our goal~~ There are two other notable points to discuss here. First, when applying a binning method, one must additionally take care with respect to the various sample sizes across different bins. For this reason, we implement a type of quantile regression methodology for the duration of the paper. Second,

there is the issue of whether or not to use dew point temperature data which is concurrent in time with the Rx1hr or Rx1day precipitation value. Our primary goal in this paper is to devise ~~methods of finding "historical"~~ an effective method of estimating scaling rates where ~~we can then~~ they can be used to predict expected changes in extreme precipitation given changes in ~~our~~ dew point temperatures. ~~To that end, we can think of our maximum hourly precipitation as our dependent variable, and it depends on, or is conditioned on changes in~~ Estimating concurrent relationships between dew point temperature ~~. Given some "future" projected distribution of hourly dew point temperatures for a given month at a given station, it is not clear ahead of time at which hourly dew point temperature we will observe the maximum hourly precipitation rate. Figure 5a plots three empirical CDFs of hourly dew point temperatures for one randomly selected station for the month of July. The curves are for the same month, but using three different years. One can clearly observe in Fig. 5a that the distribution of~~ and extreme precipitation for a region in a given season is often done with knowledge of when exactly the extreme precipitation events have already taken place. In that case, the dew point temperature ~~from July 1996 is substantially warmer than the distribution from July 2020. However, the dew point temperature at which the maximum hourly precipitation occured is lower in the year 1996 than 2020. So, even though the mean of the distribution for the year 1996 is more than 5°C warmer than the mean of the distribution from 2020, the maximum hourly precipitation occured at a lower dew point temperature in 1996 than in 2020. One can also compare the~~ is found corresponding to the same hour of the extreme precipitation amount that fell in the top specified quantile, such as 1%. However, if we have a projected change in our hourly dew point temperature distribution ~~from 1996 to the year 2001. In this case, the dew point temperature distributions are nearly identical, but the dew point temperature at which the most extreme hourly precipitation occured is quite different. To restate the problem~~without prior knowledge of when the most extreme precipitation event happened, we cannot ~~know~~ say ahead of time at which ~~hourly~~ hour, or dew point temperature~~the most extreme precipitation rate will occur. More often than not~~, the most extreme precipitation ~~will occur in the upper end, or right-tail, of the dew point temperature distribution. This is not always the case, however, and there are even cases where the most extreme precipitation rate for the month occurs during the coldest recorded hourly~~ event will take place. It is possible that the most extreme hourly precipitation rate can be concurrent with any of the dew point temperature ~~for that month. Provided that we cannot know ahead of time at which value of our predictor will correspond to the most extreme precipitation , the predictor can more easily~~ values within the distribution from that station/month. Considering a distribution of 744 hourly values of dew point temperature at a given station in July, which of these hourly dew point temperatures do we use as our predictor of extreme precipitation? We cannot know exactly which hourly dew point temperature value corresponds with the extreme precipitation amount, and therefore our predictor can ultimately be thought of as ~~the entire distribution itself and whether or not the mean of that distribution has shifted to the left or right~~a shift in the mean of the distribution over the month. In fact, ~~we find that using the mean monthly dew point temperature even has the potential to improve the accuracy of the forecasts. Consider that we find for each station our maximum hourly precipitation values for each month in the time series. We also find, for each station, the dew point temperature at the hour where the maximum hourly precipitation for that month occured, in addition to the average monthly dew point temperatures for each month at each station. This gives us three multidimensional arrays: 1) the maximum hourly precipitation for each month and station, 2) the corresponding, or collocated, hourly dew point temperature corresponding with the time of maximum hourly~~

precipitation from the first array, 3) the average monthly dew point temperature for each month and station. These three arrays each have a size of (70,12,73), where 70 is the number of stations, 12 is the number of months in the calendar year, and 73 is the number of years in the data record. Using Eq. 4, we can compute standardized values of the second and third arrays containing different dew point temperatures, where the t, from Eq. 4, now corresponds to the year. Similarly, we can use Eq. 5 to compute standardized values of the first array containing precipitation amounts, where again the t, from Eq. 5, corresponds to the year. In Fig. 5b, the scatter points, along with the 2-d histograms, between the first and the second standardized arrays, are plotted for the month of July. We can compare Fig. 5b to Fig. 5c which uses the first and third standardized arrays also for the month of July. There is a stronger statistical relationship between the using monthly averages of dew point temperatures instead of concurrent dew point temperatures can even enhance the statistical relationship, or predictive power, between our predictor and our predictand. 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. Figure 4a plots the normalized concurrent dew point temperature anomalies along with normalized Rx1hr precipitation. Figure 4b plots the normalized monthly average dew point temperature and maximum hourly precipitationthan the collocated hourly dew point temperature. More generally, we find that during the summer months anomalies along with normalized Rx1hr precipitation. One can observe in the case of the UCRB, that there is a stronger statistical relationship between the average monthly (indicated by the difference in correlation coefficients between Fig. 4a and 4b) between the monthly average dew point temperature and the maximum hourly precipitation than when using the collocated Rx1hr than the concurrent hourly dew point temperatures. While in the winter, we find the correlation strength to be of a similar magnitude. It is for these reasonsoutlined here, that we have chosen to use the average monthly temperature. For these reasons, we use the monthly average dew point temperatures as our predictor of maximum hourly and maximum daily precipitationrates. Rx1hr and Rx1day precipitation.

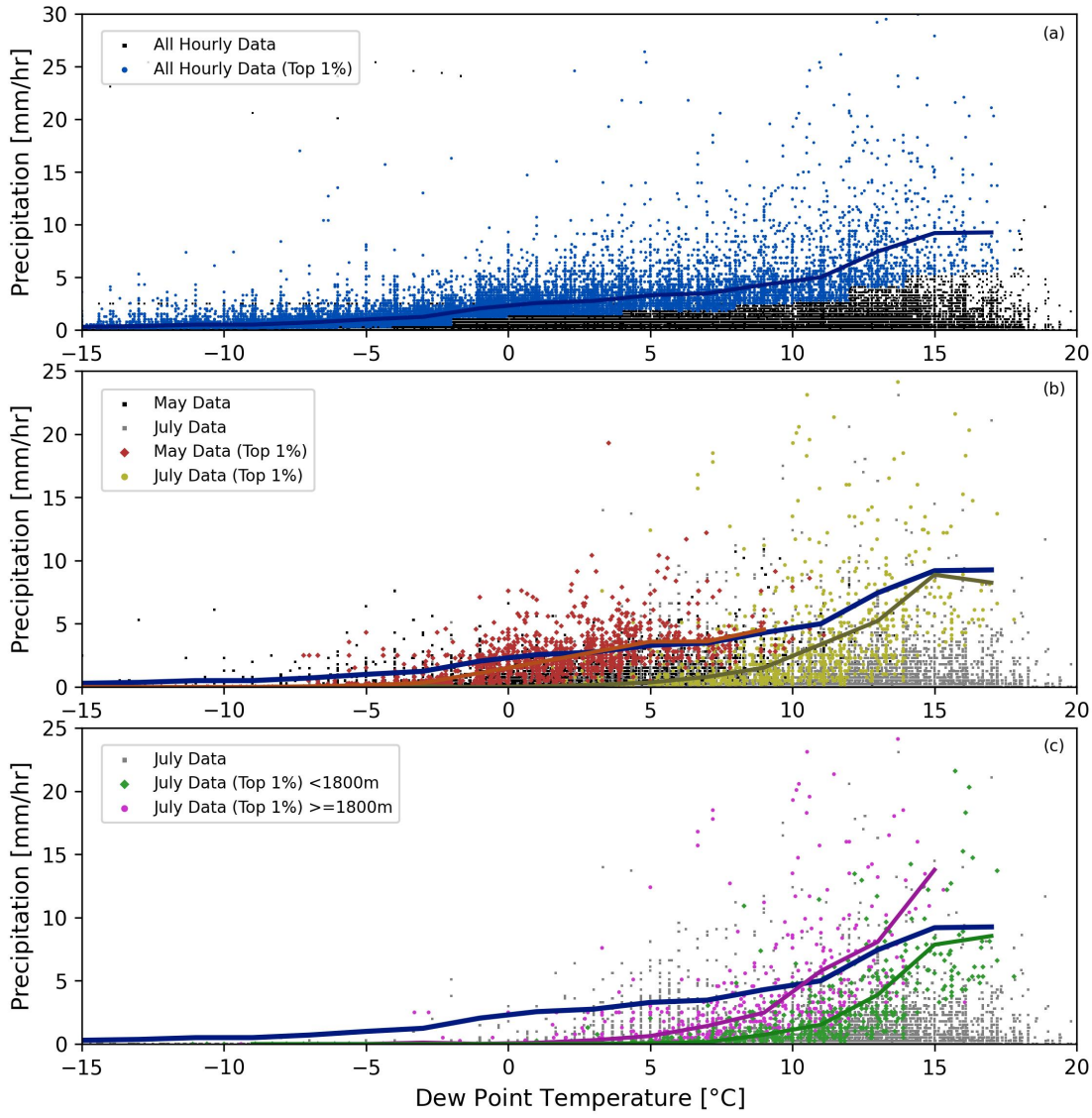

Figure 2. (a) All of the pairings of measured hourly precipitation along with the corresponding dew point temperature for the stations in the UCRB are plotted in the background as the black scatter points. The top 1% of precipitation rates are plotted as the blue scatter points. The average of the top 1% can be seen as the blue line. (b) and (c) use the top 1% of precipitation, but where the top 1% was found for each station and each month. (b) shows the extreme precipitation for May (red points and the average is plotted as the red line) and for July (yellow points and line), using all of the stations in the UCRB. The average of the top 1% from May and July can be contrasted to the blue line from (a). (c) shows the difference between extreme precipitation in the month of July (i.e., top 1%) using stations in the UCRB that are located below (green points and line) and above (magenta points and line) 1800 meters in elevation.

~~The top and bottom rows of subplots corresponds to hourly and daily data, respectively. (a) The lagged-1 autocorrelations of the hourly precipitation time series data are plotted for the stations in the UCRB. (b) The thick black line is the empirical cumulative distribution function (CDF) of all of the station lagged-1 autocorrelations from (a). The empirical CDFs of 100 randomly resampled realizations are the thin colored lines in (b). (c) The probability, at each station, of having an hourly precipitation rate fall in the top 0.1%, given that the previous hour was also in the top 0.1%. (d) The same as (b), except now computing the probabilities, instead of autocorrelations, given the randomized realizations. (e)-(h) These are the same as (a)-(d), except using daily precipitation data.~~

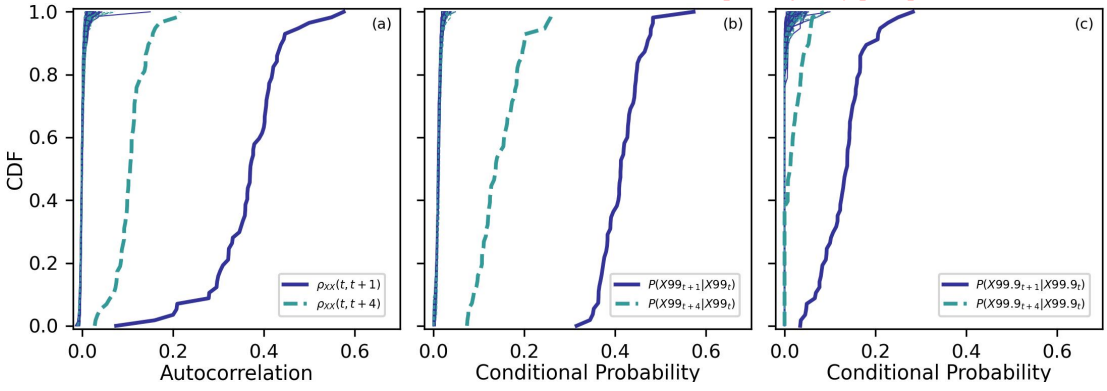

Figure 3. (a) Empirical cumulative distribution functions (CDFs) of the lagged-1 and lagged-4 hour autocorrelations from the stations in the UCRB are plotted as the blue solid and light blue dashed lines, respectively. CDFs from randomly generated time series are seen as the thin lines on the left. (b) Empirical CDFs of the lagged-1 and lagged-4 hour conditional probabilities are shown for extreme precipitation. For each station, the probability of an extreme precipitation (in this subplot, this is the top 1%) event is followed by another extreme precipitation event 1 or 4 hours later are plotted as the blue solid and light blue dashed lines, respectively. (c) same as (b) except using the top 0.1% instead of top 1% of precipitation.

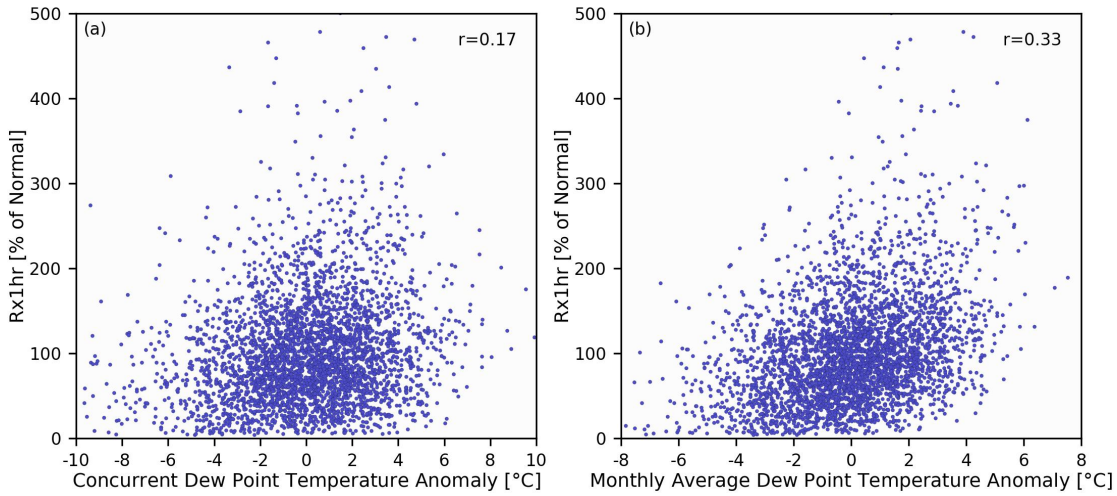

Figure 4. (a) ~~The distributions of~~ normalized concurrent dew point temperature anomalies are plotted ~~for the same station, but for three different Julys. The black line, brown dashed line, and gray line are the distributions for the month of July in the years 1996, 2001, and 2020, respectively. The dew point temperature values are shown on the x-axis, and the y-axis plots the non-exceedance probabilities. The dew point temperature corresponding to the hour at which the maximum hourly~~ along with normalized Rx1hr precipitation~~rate occured is enclosed by the hollow blue square, the circle, and blue diamond, respectively~~. (b) ~~Shows a 2-dimensional histogram between the standardized values of the collocated~~ normalized monthly average dew point temperature ~~and maximum hourly precipitation for the month of July (the green dots~~ anomalies are ~~the values used to create the histogram)~~plotted along with normalized Rx1hr precipitation. ~~(c) Shows a 2-D histogram~~ The correlation coefficients between the ~~standardized values of~~ data points in each subplot are seen in the ~~average monthly dew point temperatures and maximum hourly precipitation for the month of July~~upper-right corners.

3.2 ~~Using Normalized Data~~Proposed Approach for Estimating P-T Scaling Rates

In the ~~previous~~ prior section, we ~~documented three common challenges or~~ pointed out a number of problems that can ~~arise when estimating and interpretating~~ adversely influence our estimation of P-T scaling rates. ~~However, all three of these problems can largely be circumvented by implementing the following steps: Using the hourly dataset, find the maximum hourly precipitation rate for each month at each station . Similarly, using the daily dataset, find the maximum daily precipitation rate for each month at each station. Compute monthly averages of~~ 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~~for each month at each station. Normalize the datafrom the prior two steps by computing anomalies of the precipitation and dew point temperature time series. The normalized~~ . 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.

Our modeling approach uses a type of quantile regression, but not as it has traditionally been applied with respect to P-T scaling rates. Often, quantile regression is applied by using a chosen quantile which is found with respect to different dew point temperature ~~anomalies can now be computed as,~~

$$\mathbf{DPT}^*_{x,m,t} = \mathbf{DPT}_{x,m,t} - \overline{\mathbf{DPT}}_{x,m} \ ,$$

~~where $\mathbf{DPT}_{x,m,t}$ is average dew point temperature at station x, month m, and year t, and $\overline{\mathbf{DPT}}_{x,m}$ is the mean dew point temperature over the calibration time period at station x and month m. Similarly, the anomalies of precipitation can be computed as,~~

$$\mathbf{P}^*_{x,m,t} = \frac{\mathbf{P}_{x,m,t}}{\overline{\mathbf{P}}_{x,m}} \cdot 100 \ ,$$

bins. However, we have shown that this can be problematic when using pooled, non-normalized data which also lacks statistical independence. We propose instead to take the quantile with respect to time. That is essentially what we have with our Rx1hr and Rx1day values, where Rx1hr is approximately the top 0.14% (=1/720, with 720 hours in a month) and Rx1day is approximately the top 3.3% (=1/30, with 30 days in a month) for any given month and station. We can now fit a quantile regression model between monthly average dew point temperatures and Rx1hr (or Rx1day if we are using the daily dataset). The model fit minimizes the least-squares over the following exponential function:

$$y = a \cdot b^x \ , \tag{6}$$

where ~~$\mathbf{P}_{x,m,t}$ is either the maximum hourly or maximum daily precipitation rate at station x, month m,~~ x contains monthly-averaged dew point temperature anomalies, from the array $\mathbf{DPT}^*$ (or $\mathbf{DPT}$ if using non-normalized data), a is a multiplicative offset, b is interpreted as our scaling rate, and y would be either the normalized (or non-normalized) values from Rx1hr or Rx1day. The model fit is performed using the optimize function contained in Python's Scipy package. Consider a synthetic example where we were to fit the model to normalized data from Eqs. 1 and ~~year t, and $\overline{\mathbf{P}}_{x,m}$ is the mean of the respective precipitation~~ 2. Given that the model is fit to the normalized data, the value of a would approximately equal 100 (which is 100% of normal, this value ultimately depends on the distribution and the skewness of the data), and consider the value of b is found to be 1.08 (i.e., ~~hourly maximum, daily maximum) time series over the calibration time period at station x and month m. By normalizing the data with Eqs. 6~~ the scaling rate is 8% per °C). With this example, and given a +2°C average monthly dew point temperature anomaly, we would predict Rx1hr (or Rx1day) to be 116.6% of normal.

3.3 Predicting Extreme Precipitation

Predictions of Rx1hr and ~~7, we have effectively removed the three common challenges or problems that we discussed above. First, we compute anomalies of the data which removes climatological differences between stations in both space and time .~~

~~Second, we always use the same sample sizes (e. g., there are always 744 hours and 31 days in January, with the only exception being for the month of February, where there can be small differences in sample sizes given whether or not there is a leap year). And third, we observe the maximum hourly and maximum daily precipitation rates at the station-month scale to be statistically independent.~~ Rx1day are performed using leave-one-year-out cross validation over the period 1951-2024. For the ~~duration of the paper, any reference that we make concerning dataanomalies or normalized data will correspond to the values computed from Eqs~~daily dataset, which has sufficient data coverage throughout the period of record, we additionally perform a two-fold cross validation with the data being split into two equal time periods, with the first period being 1951-1987 and the second period being 1988-2024. We have four models that we use to make predictions and evaluate model performance. These are in increasing complexity, 1) always assuming 100% of normal, 2) fitting the exponential model from Eq. 6 to the normalized data and always assuming a C-C scaling rate of 7% per °C (i.e., this fits Eq. 6 to solve for a when b is fixed to a value of 1.07), 3) fitting the exponential model from Eq. 6 to the raw (non-normalized) data, and 4) fitting the exponential model from Eq. 6 ~~and 7.~~ to the normalized data. Through this procedure of implementing different modeling approaches with different data, we can determine which approach provides the best performance.

3.4 ~~Different Methodological Approaches to Estimating Scaling Rates~~

~~In this paper, we use three different methodological approaches to estimate P-T scaling rates. The first uses a binning method with the raw, measured datawhich has not undergone any transformation or normalization. The second uses a binning method again, but with the normalized data anomalies from Eqs.~~ We can use the fourth model (i.e., fitting Eq. 6 ~~and 7. For these first two methods, binned-averages are computed over 2° windows, or bin widths, and with increments of~~ to normalized data) from the prior paragraph to briefly outline our procedure to produce leave-one-year-out cross-validated predictions of Rx1hr. For each year, the average monthly dew point temperatures ($\mathbf{DPT}^*$ from Eq. 1~~°. As we have already prescreened the data for the precipitation extremes (i. e., maximum hourly and maximum daily precipitation), we compute binned-averages over all of the existing data points that fall within each bin. The third method fits an exponential curve to the same normalized data used in the second binning method. The model fit is performed using the optimize function contained in Python's Scipy package. This method minimizes the least-squares over the following function:~~

$$y = a \cdot e^{xb} \quad ,$$

~~where x contains monthly-averaged dew point temperature anomalies, from the array~~) and Rx1hr precipitation ($\mathbf{P}^*$ from Eq. 2) are computed using only the calibration years which exclude the year for which we are going to make model predictions of the anomalous Rx1hr. Likewise, the model is fit using Eq. 6 to the data from the calibration years, and model predictions are produced for the given validation year. For each validation year in turn, we additionally retain the observed precipitation anomalies computed using mean values of Rx1hr calculated from the current set of calibration years. These observed anomalies will be used in our model evaluation. Then, we step to the next year until we have covered the entire time period, 1951-2024. For the two-fold cross validation scenario, the average monthly dew point temperatures ($\mathbf{DPT}^*$ ~~, a is a multiplicative offset, and b is interpreted as our scaling rate~~from Eq. 1) and Rx1day precipitation ($\mathbf{P}^*$ from Eq. 2) are first computed using the

calibration years 1951-1987. The model is fit using Eq. 6 to the data from these calibration years, and predictions are produced for the validation period 1988-2024. Then, data from the years 1988-2024 is used for calibration, and the years 1951-1987 are predicted.

We begin by implementing the two binning methods across Each of the last three models (models 2-4 listed above), use a range of different cases of varying spatial and temporal extents over which to fit the model and make predictions. This is so that we can first show the added value of using the normalized data instead of the measured values. Model estimates of done to observe if, and to what extent, the pooling of data across different stations and/or months can improve the estimated P-T scaling rates are used to generate predictions of maximum daily and maximum hourly precipitation at the station-month level, where modeled values are produced for every month at every stationand the associated predictions. We use the two binning methods to evaluate the cross-validated model performance of the predictions given different combinations of spatial and temporal four spatial extents. The five spatial extents that we use spatial extents are: 1) using only the data from the station being predicted, 2) using stations within a 50 km radius of the station being predicted, 3) using stations within a 100 km radius of the station being predicted, and 4) using stations within a 200 km radius of the station being predicted, and 5) using all of the stations that fall within the entire UCRB. The region (in our case, the UCRB). Similarly, we use four temporal extentsthat we use . The temporal extents are: 1) the data solely from the month being predicted ,(i.e., 1-month window), 2) a 3-month window centered about the month being predicted, 3) a 5-month window centered about the month being predicted, and 4) using all 12 calendar months. Consider a synthetic example where we would like to produce modeled predictions of maximum daily precipitation for station "1" for the month of July over the last 30 years, 1994-2023. We begin by using only the data at this station from all of the Julys over the calibration period 1951-1993. This example corresponds to using the first spatial and first temporal extents. The maximum number of data points used to construct all of the binned values in this case would be 43 (corresponding to the number of data values in July, at one station, over the 43 years in the calibration period). The binned-averaged precipitation rate in this case is often only relying on only a few observations for each temperature bin. Consider another case where we apply the binning methods using stations within 100 km of station "1" and over a 3-month window corresponding to June-July-August. In this case, let us assume that there were 30 additional stations that fall within 100 km of station "1". Then, the maximum number of data points used to construct the binned values in this case would be 3999 (corresponds to (43 years) x (31 stations) x (3 months)). The same procedure is performed for all combination of cases using the measured and the

In Figure 5, we can observe the process of model fitting to non-normalized versus normalized data.

In Figure 6, we illustrate the three different methods using In these cases from Fig. 5, we use all of the 187 daily stations that fall within stations in the UCRB (i.e., spatial extent 5)using number 4), and a 3-month temporal window (i.e., temporal extent number 2). Figure 6a shows the binned estimates between the measured monthly-averaged The top row of subplots (Figs. 5a-d) show the relationship between either non-normalized, or normalized, average monthly dew point temperatures and Rx1hr precipitation. This is shown using the data from the winter (**DPT**) along with the measured maximum daily precipitation amounts (**P**) using all of the stations in our domain over the winter months December-February. Figure 6b shows the binned estimates between the normalized monthly-averaged dew point temperatures (**DPT***) and the normalized maximum daily precipitation amounts (**P***) for the same stations and months. Figure 6c shows an exponential fit to the same anomalous data

from Fig.6b.Figures 6d-6f show the same as Figs. 6a-6c, but now using the summer months of June-August. The blue and red lines are examples of the modeled P-T estimates used in this paper to produce predictions for the validation periods for these particular cases.

The three methods used in this paper are illustrated here. (a) Shows all of non-normalized values of monthly-averaged dew point temperature plotted against maximum daily precipitation rates. All of the daily stations in the UCRB are used, and the data is plotted for the winter months of December-January-February. (b) A binning method is applied again, but now using normalized data. (c) An exponential function is fit to the normalized data. (d)-(f) The same as (a)-(c), except now showing the summer months June-July-August.

3.4 Evaluating Model Performance

The number of observations of daily maximum precipitation is much greater than the number of hourly maximum observations. Therefore, we begin by evaluating the performance of the two different binning models in their ability to predict out-of-sample daily maximum precipitation. This is done using the various combinations of spatial and temporal extents as we discussed in the previous section. At this point , all model predictions are produced using the calibration period 1951-1993 and validated over the period 1994-2023. For example, for a given spatial and temporal extent, Dec-Jan-Feb) and summer (Jun-Jul-Aug) seasons. The bottom row shows the model bins are computed for one station-month using the data over the calibration period 1951-1993. Then, same as the top row, except using daily data (i.e., Rx1day). One can observe substantial differences in model fitting when we are controlling for climatological differences (using normalized data) and when we are not (using non-normalized data). A prime example of this is illustrated by Figs. 5e and 5f. The same data, with the same number of data points, is used in both of these cases. However, the binned values obtained from the calibration period, are used with the monthly-averaged dew point temperatures from the period 1994-2023 to produce values in Fig. 5f have been normalized using Eqs. 1 and 2. Without applying this normalization, we miss the fact that stations at higher elevations climatologically experience lower dew point temperatures while also experiencing climatologically greater Rx1day precipitation. We find that the elevation of the daily stations is correlated with the values of average Rx1day over the time period during the winter season with a correlation coefficient of 0.69, with the average Rx1day for stations centered about 3000 meters in elevation (between 2750-3250) being more than double that of stations centered about 1500 meters (between 1250-1750). Without controlling for climatological differences in space, which is the case in Fig. 5e, we end up conflating these climatological differences with the scaling rate. After normalization, we observe a positive scaling rate as seen in Fig. 5f. To produce the different sets of model predictions of Rx1hr and Rx1day, the procedure outlined in Fig. 5 is performed using a cross-validated framework, for the C-C, non-normalized, and normalized models. In addition, for each model case, we iterate through the four spatial and temporal extents. Model predictions which rely on the raw, non-normalized data initially provide predictions of maximum daily precipitation over the validation period. Model predictions are produced for each month at each station. The first binning method initially produces predictions as non-normalized values of Rx1hr precipitation (i.e., mm per hr). These modeled values are subsequently normalized using Eq. 7. 2 prior to model evaluation.

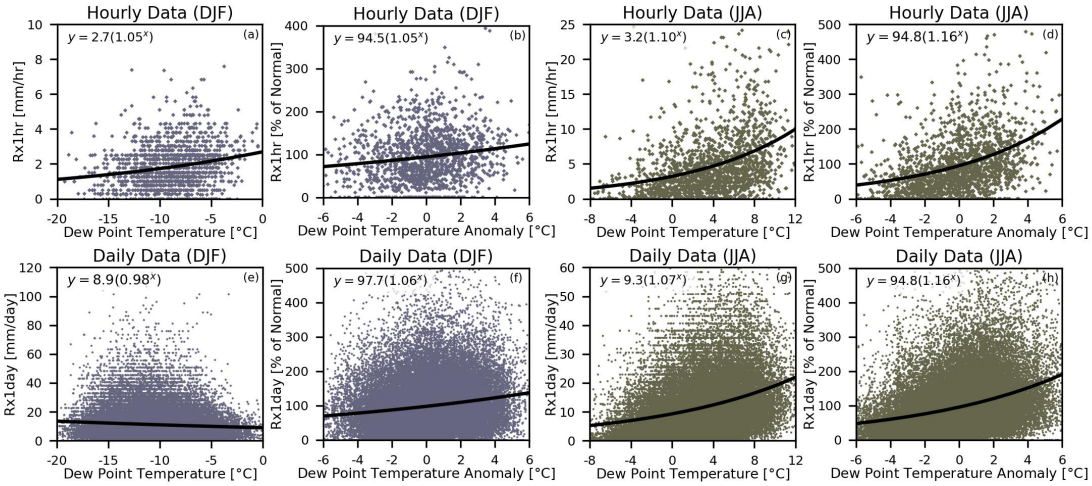

Figure 5. Comparing the model fit using non-normalized versus normalized data. The data points correspond to all of the stations in the UCRB and using a 3-month window corresponding to the winter and summer seasons. (a) plots all of non-normalized values of average monthly dew point temperature plotted against Rx1hr for the Dec-Jan-Feb winter season. The model fit using Eq. 6 is plotted as the black line. (b) is the same as (a) except using normalized data. (c) and (d) are the same as (a) and (b) except for the Jun-Jul-Aug summer months. (e-h) are the same as (a-d) except now using Rx1day instead of Rx1hr.

3.4 Evaluating Model Performance

The underlying empirical-statistical relationship between the normalized average monthly dew point temperature and normalized Rx1hr is plotted for all station/months in Figure 6a. In Fig. 6b, the individual leave-one-year-out cross-validated predictions of Rx1hr are plotted versus observed Rx1hr. The predicted Rx1hr in Fig. 6b are produced using the normalized data model with a spatial extent of 100 km and a 3-month temporal window. While individual predictions of ~~daily maximum~~ Rx1hr and Rx1day precipitation are produced, as depicted in Fig. 6b, the individual values themselves exhibit large fluctuations due to natural variability. This is due to the fact that we are solely conditioning changes in extreme precipitation on changes in dew point temperature. Given the large variability of the individual values themselves, we evaluate model performance as the average change in ~~extreme~~ Rx1hr or Rx1day precipitation given specified average monthly dew point temperature anomalies ranging between -3°C below normal to +3°C above normal. At each dew point temperature anomaly, we compute the mean extreme precipitation anomaly centered about that dew point temperauture anomaly. We use a ~~2~~1°~~window~~ C window, with 1°C increments, and find the modeled and observed extreme precipitation distributions corresponding to dew point temperature anomalies ~~of:~~ -3°C, -2°C, -1°C, 0°C, +1°C, +2°C, and +3°C. Model performance evaluates how well the predicted mean shifts in the normalized extreme precipitation align with ~~observations over the validation period.~~

~~(a) Normalized average monthly dew point temperatures, for all daily stations, against normalized maximum daily precipitation for the month of July. This is for the data in the validation period 1994-2023. The different dew point temperature anomalies where model performance is evaluated can be seen as the different colored dashed lines. The orange points located between~~

0°C and +2°C are used to construct the precipitation distribution corresponding to a +1°C dew point temperature anomaly. (b) Plots the distribution of the orange points highlighted in (a) as the thicker orange line. The mean shift of this example distribution is the thicker vertical orange dashed line. (a) and (b) are obtained from observed values in the validation period. (c) The modeled mean shifts are plotted against the observed mean shifts for the 12 months of the year and for the 6 dew point temperature anomalies. The modeled mean shifts were obtained using the binning model approach with normalized or anomalous data, and using all of the stations in the UCRB along with a 3-month window. One can see the observed mean shift corresponding to a +1°C anomaly for the month of July (orange points and distribution from (a) and (b)) is the value on the y-axis of larger orange diamond in (c). (d) The same as (c), except using CC to predict the mean shifts.

Figure 7 provides an illustration of how the model skill score is computed given the case of using the spatial extent of all of the stations within the UCRB (i.e., spatial extent number 5) and the 3-month temporal extent (i.e., temporal extent number 2). the observed mean shifts. In Fig. 7a, the observed monthly-averaged 6c, we can look look more specifically at the Rx1hr anomalies for each of these dew point temperature anomalies are plotted against the observed maximum daily precipitation anomalies. The values in Fig. 7a are shown using all of the stations within the UCRB for the month of July over the validation period 1994-2023. There are 4,974 data points in Fig. 7a, which corresponds to a 88% data coverage (88% = 4, 974/5,610, where 5,610 is equal to the 30 validation years for the month of July times the 187 stations that reside within the UCRB). The different empirical CDFs of normalized maximum daily precipitation, given the 6 using the summer months Jun-Jul-Aug. The orange points show, for example, the Rx1hr values that fall within the one degree window centered about a +1°C anomaly. The empirical CDFs corresponding to different dew point temperature anomalies, are plotted as the curves along with the means of those distributions, are shown in Fig. 7b. The mean shifts of the different distributions are plotted as the vertical dotted lines. This gives us 6 observed mean shifts computed over the validation period for the month of July. The same procedure is repeated for all 12 calendar months. Similarly, we can compute the model predicted mean shifts. Then6d. This provides us with seven mean observed Rx1hr precipitation anomalies (as % of normal) as a function of the seven dew point temperature anomalies anomalies for the 3-month window centered about July. We can similarly compute the seven mean predicted Rx1hr precipitation anomalies for this case (not shown). Similarly, we can use the modeled and observed mean shifts to compute how skillful the model predictions are (using Eqs. 1-3) with respect to climatology (i.e., 100% of normal), and with respect to a theoretical CC scaling rate of 7% per °C. In Fig. 7c, one can see all of the model predicted means plotted against the observed means. In Fig. 7c, there are 72 scatter points corresponding to the perform the same procedure for the 12 months times the 6 dew point temperature anomalies. The modeled means months of the year. By doing that, we have 84 predicted versus observed mean Rx1hr anomalies for the different dew point temperatures and for the different months of the year. These values are plotted in Fig. 7c is derived from the binning model which uses normalized data (e6e. In the case of this model (i.e., normalized data with 100 km spatial extent and a 3-month temporal window), the RMSE of the model is 6.4. The benchmark model RMSE, which always assumes climatology (always 100% of normal) is 21.9. g., see Figs. 6b and 6e). In Fig. 7d, one can see all of the model predicted means, which are derived from assuming a theoretical scaling rate of 7% per °C, plotted against the observed means. Eq. 1 is used to compute the model error for the binning model using normalized data for all of the points in Fig. 7c. And similarly, one can use Eq. 2 to compute the model error when assuming a theoretical scaling rate of

7% per °C (using the points from Fig. 7d) . Using Eq. 3, we can arrive at our RMSE skill score (5, we find our $\mathbf{SS}_{RMSE}$) for this case. The skill score ($\mathbf{SS}_{RMSE}$) is 0.12, with respect to the theoretical CC, for this case where the binning method is used with normalized data from all of the stations within the UCRB along with a 3-month time window. At the same time, the skill score ($\mathbf{SS}_{RMSE}$) of these modelpredictions with respect climatology is 0.51. Both of these skill scores are statistically significant (p<0.01). We provide a more thorough analysis of the statistical significance later in is 0.71. The same procedure is performed separately for Rx1hr and Rx1day using predictions from the paper. C-C model, the non-normalized model, and the normalized model, and where in each model case the four different spatial and temporal extents were implemented.

4 Results

4.1 Model Performanceof the Two Binning Models

Figure 8 shows model performances for the two different binning methods, with and without normalization. Figure 8a shows the RMSE skill scores of In Figure 7a, we show the range of model skills, $\mathbf{SS}_{RMSE}$, using leave-one-year-out cross validation for the 16 possible iterations of spatial-temporal extents using the C-C model, the binning modelwhich uses the raw, or non-normalized , data. In Fig. 8amodel, and the normalized model. Fig. 7a plots the skills of the cross-validated Rx1hr predictions for the three different models. In the case of Rx1hr, the skill is in reference to climatology. Each individual grid cell in Fig. 8a corresponds to the skill computed where the predictions are produced using different combinations of our spatial and temporal extents. Figure 8b shows the skill of the binning model which uses normalized data in reference to climatology. Figures 8c and 8d show the skill of the binning modelwhen using best obtained skill score is 0.45 for the C-C model, 0.66 for the non-normalized and normalized data, but now in reference to predictions generated using the CC scaling rate. The skill values of 0.12 model, and 0.51, corresponding to the example detailed 0.71 for the normalized model. The dotted black line in Fig. 7, can be observed in Figs. 8b and 8d at the intersection of "Basin" on the y-axis and "3-Month" on the 7a are the skill scores obtained from 1000 randomly generated predictions (or simulations), where each individual set of predictions can be analogously compared or contrasted to the points along the x-axis . The skills are seen to change as a function of the spatial and temporal extents. For the binning model which uses non-normalized data, the best model performance is achieved when data from both small spatial and temporal extents are used . This can be seen as the blue colors in the top-left corners of the Figs. 8a and 8c. In contrast, the optimal model performance of the binning model which uses the normalized data is obtained when using a temporal extent of approximatedly 3 months and a spatial extent equal to the entire UCRB. By using the non-normalized data itself, which does not correct for climatological differences in both time and space, in Fig. 6b, for example. These random predictions follow the same modeling framework used by the other models, especially as it comes to computing anomalous values through cross-validation, except we do not perform any actual model fitting. Instead, for each year and for one set of predictions, we randomly select a year from the calibration period to be used as the skill immediately begins to decrease as we use greater spatial and temporal extents. For example, the skill in estimating the maximum daily precipitation over the period 1994-2023 at one station for the month of July will decrease, on average, when data from June or August from the same stationis used andpredicted values. By randomly selecting all of the station/or data from July is used from stations within

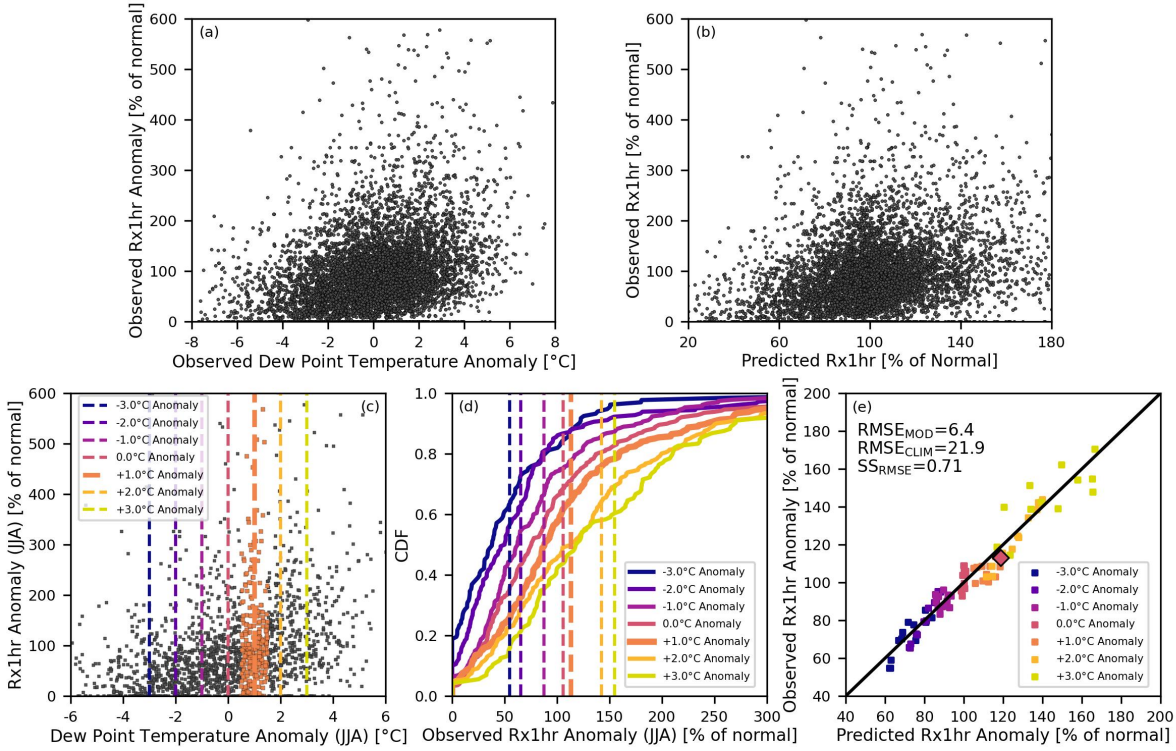

Figure 6. (a) plots the relationship between the normalized average monthly dew point temperature (our predictor) and normalized Rx1hr (our predictand) for all station/months. (b) plots the individual cross-validated predictions of Rx1hr versus observed Rx1hr using the normalized data model with a spatial extent of 100 km and a 3-month temporal window. In (a), our predictor corresponds to the x-axis, while in (b) our cross-validated predictions corresponds to the x-axis. The values along the y-axis of both (a) and (b) are the observed Rx1hr anomalies and they are the same in each subplot. (c) Dashed lines show the different dew point temperature anomalies over which the mean predicted versus observed Rx1hr anomalies are evaluated. This subplot shows the observed data for the 3-month window centered about July. The orange points show the distribution of Rx1hr along the y-axis corresponding to a +1°C dew point temperature anomaly. (d) The empirical CDFs, for the different distributions corresponding to different temperature anomalies, are plotted as the lines. The dashed lines show the mean shift of the observed distributions. The thicker orange dashed line shows along the x-axis the observed mean shift in Rx1hr (i.e., average of orange points from (d)) given the example case of a +1°C dew point temperature anomaly. (e) plots all of the seven predicted mean versus observed mean Rx1hr precipitation anomalies for the 12 months of the year, giving us 84 points. The larger orange diamond corresponds to the example case illustrated in (d) and (e).

580 ~~a 100 km radius (see Figs. 8a and 8c, the skill decreases as one traverses to the right and/or down from the top-left cases). However, we find that this same data can be leveraged to improve the skill of the binning model when data normalization is first applied. As shown in Figure 8, the best skills for the binning model using the measured, non-normalized, data is 0.49 and 0.08 with respect to climatology and CC, respectively. When using the normalized data with the binning model, the best model skill increases to 0.51 and 0.12, respectively. We can additionally compare these results for the same case (i.e., "Basin"~~

and "3-Month")against the skill of the exponential model which also uses the normalized months in a calibration year, we successfully preserve the spatio-temporal covariance of the observational data in our randomized simulations. We can then test how skillfully one can randomly predict Rx1hr (or Rx1day), when the randomized predictions exhibit the same spatio-temporal covariance structure as the underlying data. The best obtained skill of the exponential model for that case is 0.55 and 0.18, respectively. We find that the parametric exponential model provides a statistically significant improvement 1000 randomly generated predictions is 0.09. As a result, the best performing skill scores from all three models are found to be statistically significantly better than both climatology and what we would expect by chance ($p<0.01$)with respect to the non-parametric binning method.

Performance of the binning model when using non-normalized versus normalized data. The performance is measured using the RMSE skill score. (a) Skill scores of the binning model, which uses non-normalized data, are shown with respect to climatology. The skill scores are reflected by the colorbar in addition to values being printed on top of the colors. Skill is seen to change as a function of making predictions using data from varying spatial and temporal extents. (b) The same as (a), but for the binning model which uses normalized data. (c) and (d) The same as (a) and (b), except that the model performance is evaluated with respect to CC.

We can further highlight why it is that the skill of the binning model, which uses the measured or . If it is not already clear, we want to stress at this point, that all of the models, whether a randomized set of predictions, the C-C model, the non-normalized data, falls off so quickly as we further extend the temporal and spatial extents. For the UCRB, it is during the season of August-September-October that, on average, produce the most extreme precipitation events. One might decide that they want to find the maximum daily precipitation event that took place over that most extreme season (Aug-Sep-Oct) at a particular station, and they could also pair that maximum daily precipitation event with the monthly average dew point temperature for the month of that event. However, even among months over this 3-month season, the climatological differences across time can be quite large. Consider a case where we compute anomalies of dew point temperature and precipitation using Eqs. 6 and 7, but the anomalies are computed with respect to seasonal averages. So, for each station, we take the August and October time series of dew point temperatures and maximum daily precipitation rates, and compute the anomalies with respect to a August-September-October seasonal average for that station. Then, we proceed to do the same operation for all stations. In this way, we have removed any spatial climatological differences (because we have produced the anomalies station by station), but this approach preserves any temporal climatological differences that may exist in the data . We can plot these anomalies for the months of August and October in Figure 9a. The dew point temperature and the maximum daily precipitation anomalies computed using seasonal averages are plotted in Fig. 9a for all of the stations and all of the years in the data set.It is clear from Fig.9a that the dew point temperatures in the month of October are significantly cooler than those of August.They are, on average, 8.7model, or the normalized model, all have access to exactly the same data to produce the cross-validated predictions. The models only differ in how they make use the data (e.g., non-normalized versus normalized) and what assumptions are made (e.g., C-C always assumes 7% per °Ccooler. While at the same time, October has extreme daily precipitation amounts that are, on average, 7% greater than those of August. If we simply take the measured data values across a season such as August-September-October and subsequently compute anomalies on that data, then the climatological differences in time will

have a large impact. Due to these climatological differences between the months of August and October, for example, we would in this case underestimate the effective scaling rate. Molnar et al. (2015) and Visser et al. (2021) both showed a similar result across time at a single station. The gray dashed line in Fig. 9b, is). In Figs. 7b-d, we observe in what combinations of spatial and temporal extents each respective model sees its maximum skill score. We observe fairly uniform positive skill for the C-C model in Fig. 7b, with small variations present due to a changing sample size between different spatial and temporal extents. For the non-normalized model (Fig. 7c), its optimal skill is found to correspond to using the 50 km spatial extent along with the 1-month temporal window. One can clearly observe the skill decrease for the non-normalized model as the data is pooled from further away in both space and time. When pooling data from the exponential fit, or the estimated scaling rate, of all of the scatter points in the August-September-October season where the anomalies are computed using seasonal averages (these are the data points from entire UCRB region and using all of the data throughout the year (i.e., the bottom-right grid cell in Fig. 9a plus the data from September). We can then compare the gray dashed line in Fig. 9b to the exponential fit of each of the individual months in this 3-month season (also shown in Fig.9b).In Fig. 9b, we are highlighting the differences in the rate of change as a function of dew point temperature, and we therefore have all of the curves pass through 100% of normal at a 0.0°C dew point temperature anomaly. The estimated scaling rates of the three months, individually, can be seen to be very similar to one another and to what we estimate when fitting an exponential curve to all three months together using normalized data computed at the station-month level. The effective scaling rate is clearly greater than what is estimated if we had not accounted for climatological differences in time, where data values can vary substantially from month to month even at the same station. This shows how one would inaccurately estimate the scaling rate for this season simply by not accounting for the fact that October is typically cooler, but also exhibits modestly more extreme precipitation than August. Next, we can isolate the impact of climatological differences across space. To do this, we can take the measured values of the data across the stations for the single month of October. This is the same data as the blue points in Fig. 9a, though we have not yet computed any anomalies at this point. Next, we can calculate the mean over all of the stations in 7c), the UCRB and all of the years for this month of October. In Fig. 9c, we plot the scatter points for two different subregions of the UCRB, data east and west of 108° longitude, respectively. The data east of 108° longitude contains dew point temperatures 1.6°C cooler and maximum daily precipitation extremes that are 15% wetter, on average, than the data west of 108° longitude. Here, we have removed temporal climatological differences by using a single month, and we therefore isolate the climatological differences in space. The scaling rate which is fit to the scatter points of Fig. 9c is shown as the dashed gray line in Fig. 9d. Again, we observe the estimated scaling rate in gray would, in this case , underestimate the "true" scaling rate due to climatological differences in space. These examples show how important it is to remove climatological differences in both space and time. And it is because of these climatological differences, that the skill of the binning model which uses the measured data without normalization decreases so rapidly as skill is less than half of what it is in its optimal case, or 0.30 versus 0.66. This is primarily due to the fact that the spatial and temporal extent increases.

non-normalized model includes pooled data with very different climatologies in time and space. Another reason is that we are mixing different scaling rates across time and space, where each station can be considered to have a certain scaling rate for a given time of the year, such as July. For this second reason, we also see a decrease in skill for the normalized model when

pooling data from the largest spatial and temporal extents (see Fig. 7d). However, the normalized model is able to leverage data from larger spatial and temporal extents in order to improve the skill with respect to using non-normalized data. The maximum skill of the normalized model is 0.71 and corresponds to the grid cell bounded by the yellow box (this case was also illustrated in Figs. 6b and 6e). A t-test between the model errors (as depicted about the one-to-one line, for example, in Fig. 6e) that are associated with the optimal skill scores between the non-normalized and normalized models show the normalized model provides statistically significantly better skill than the non-normalized model ($p<0.01$).

4.2 ~~Evaluating the Exponential Model Performance~~

~~In Figure 8, we have already shown the split-period cross-validated performance of the two different binning methods. We found that~~ Figs. 7e-h show the same as Figs. 7a-d except evaluating the model performance in predicting Rx1day precipitation. Generally, the ~~binning method which uses normalized data produces more skillful predictions of maximum daily precipitation than the binning method which uses the measured values without normalization. Furthermore, we find that using the same normalized data, the parametric exponential model produces more skillful predictions than the binning method. We have shown this using a calibration period 1951-1993 and an independent validation period of 1994-2023. Here, in Figure 10, we now show the model performance for both maximum daily and maximum hourly precipitationrates. However, in Fig. 10 we use shorter validation periods. The reasons for doing this are two-fold. First, we need sufficient data in hourly dataset in order to both calibrate and validate the model. Second, we would like to illustrate the impact that different calibration and validation periods can have on model performance (the following paragraph provides further details regarding this issue). For the maximum daily precipitation, we now use a calibration period 1951-2003 and a validation period of 2004-2023. For the hourly dataset, the data can be quite sparse prior to the year 1999. Therefore, we make use of a shorter validation period to perform an independent cross-validation for the maximum hourly precipitation. We use the period 1951-2008 to calibrate the maximum hourly precipitation exponential model, and we validate over the period 2009-2023. In both of these cases of the daily and hourly data, we have a "sufficient"~~ results follow the same pattern with the optimal model skills for the three models being 0.63, 0.74, and 0.81, respectively. What is apparent now, is that the skill of the non-normalized model immediately begins to degrade as any amount of data ~~to both calibrate and validate the models. Following the results concerning the best skill of the exponential modeldiscussed above, the model is fit to the normalized dataover the calibration period using all of the stations within the UCRB and using a 3-month window. Using these calibration and validation periods, we now observe skill scores of 0.42 (~~pooling is performed. The non-normalized model provides the best results when the calibration data is used from only the station and month being predicted. The drop off in skill for the non-normalized model, which uses pooled data, is even more profound that what we observed for Rx1hr (Fig. 7c). Using data with a 100 km spatial extent and a 5-month temporal window, for example, gives worse skill than climatology (see the corresponding light red grid cell in Fig. 7g). The maximum skill of the normalized model is again bounded by the yellow box in Fig. 7h, and this skill is again found to be statistically significantly better than the optimal non-normalized skill score ($p<0.01$)~~and 0.35 ($p<0.01$) with respect to climatology in the model's performance of estimating the maximum daily and hourly precipitation rates, respectively. The model skill scores with respect to CC are 0.09 ($p<0.01$) for the maximum daily, and 0.12 ($p<0.01$) for the maximum hourly.~~.~~

The relationship between model predicted mean shifts versus observed mean shifts for the 12 months of the year and for the 6 dew point temperature anomalies. The mean shifts are obtained from predictions of the exponential model using normalized data for all stations within the basin and a 3-month window. The modeled versus observed mean shifts are shown for the maximum daily precipitation in (a) and the maximum hourly precipitation in (b).

Our results show that the skill of Figure 8 shows the same results as Figure 7, except we now predict Rx1day using two-fold cross validation. In this case of the model predictions of maximum daily precipitation are statistically significant for both of the chosen validation cases, 1994-2023 two-fold cross validation, the best obtained skill score is 0.68 for the C-C model, 0.66 for the non-normalized model, and 2004-2023. However, the skill is seen to be quite different between these two cases. The skill scores are 0.55 and 0.18 (versus climatology and CC, respectively)when validating over the last 30 years, and 0.42 and 0.09 when validating over the last 20 years. This leads us to another important issue that requires consideration when assessing the skillfulness of different methods. It is worth remembering that we are dealing with extreme events that are relatively rare to begin with, and as a result, natural variability can influence our apparent level of skill in two important ways. First, the estimation of what is considered a "normal" extreme precipitation event will vary from one calibration period to another. As the length of the time series in the calibration period increases, we improve our confidence in estimating what constitutes an "average" extreme event. Consider that we want to use a station with 20 years of measurements. If we split the time period in half for calibration and validation, then we are using only 10 measurements to compute what the "average" or mean extreme precipitation is for a given station and a given month. And the second way that natural variability can influence the apparent level of skill, is that the validation must also be performed over enough cases or data points in order to obtain an accurate picture of the "true" mean shift of precipitation extremes given the different dew point temperature anomalies. To largely circumvent these two issues, we would like to predict the entire extreme precipitation array through 0.76 for the normalized model. These values are quite similar to what we obtained and plotted Figs. 7e-h for the leave-one-year-out cross-validation. Before we do that, however, we must first determine whether the mean shifts in the observed precipitation extremes, as a function of dew point temperature anomalies, are systematically changing over time. That is to say, are the observed mean shifts in extreme daily precipitation for the 12 calendar months and the 6 dew point temperature anomalies systematically moving in one direction between the calibration and validationperiods? Can these mean shifts, or the scaling rates themselves, be considered stationary? Would we be making an invalid assumption that a model fit over some historic period can be applied to some "future" period?

Influence of natural variability in evaluating performance skill. Randomly chosen 20-year calibration and 20-year validation periods are repeated 100 times. The calibration and validation periods are chosen from only the first 53 years of data, corresponding to the years 1951-2003. For each dew point temperature anomaly, the scatter points are shown between the data which could be used in calibration (x-axes) and the data which could be used in validation (y-axes). The smaller points consist of the 12 months of the year for the 100 randomized realizations. The larger diamond scatter show the calculated mean shifts over the last 20 years (2004-2023) with respect to three different calibration periods: 1951-1983, 1961-1993, and 1971-2003.

~~The larger scatter points of Figures 11a-11f show the effect of the first point raised above. That point relates to~~ cross validation, especially seen in the fact that we can still leverage normalized data to improve the predictions of Rx1day precipitation. This result provides a better understanding as to how well the methodology can apply in the context of a changing climate. With the results from Fig. 7 alone, one might misinterpret that the skill we observe is entangled with the trends in the ~~fact that the predicted mean of~~ underlying data. However, the ~~extreme precipitation events can vary from one calibration period to another. The larger diamond scatter show the calculated mean shifts over the last 20 years (2004-2023) with respect to three different calibration periods: 1951-1983, 1961-1993, and 1971-2003. One can observe that the points move as the calibration period changes, which is illustrated by the fact that more than 12 diamond scatter points can be seen in each subplot. This is because what is considered "average" or "normal" changes as a function of the chosen calibration period. Next, we would like to observe if the scatter of these larger points differs systematically from randomly selected calibration and validation periods. If we observe a systematic change over time, then this would indicate that the scaling rates cannot be assumed to be stationary in time. To determine whether there is a systematic change in the observed mean shifts over time, we first generate 100 randomized 20-year calibration and 20-year validation periods chosen from only the first 53 years of data, corresponding to the years 1951-2003. For each scenario, we randomly choose two separate and independent 20-year periods, which are not necessarily sequential, from the years 1951-2003 to be used as calibration and validation. For each scenario, none of the years chosen for the calibration period are used in the validation period. We then plot the mean of the extreme precipitation anomalies for the calibration versus the validation periods for each of the 12 calendar months and~~ results from Fig. 8 show that is not the ~~6 dew point temperature anomalies. We repeat this procedure 100 times. We find that the cloud of these smaller scatter points generated exclusively from the data period 1951-2003, fully envelops the larger scatter points of the last 20 years. This tells us that much of the error that we observe between the scatter points and the one-to-one line in Figures 7c, 10a, and 10b can be attributed to natural variability. We do not find any discernible systematic change in the mean shifts of extreme precipitation over time. As a result, we can make a safe assumption that the P-T scaling rates are stationary with respect to time.~~ case, and that the method can successfully produce skillful predictions over multi-decadal lengths of time.

~~With this assumption, we can leverage the full length of the data sets by increasing both our calibration periods and the period over which we validate. To do this, we use leave-one-year-out cross-validation to make predictions of extreme precipitation and validate over the entire period 1951-2023. For each year, the mean dew point temperatures ($\overline{\mathbf{DPT}}$ from Eq. 6) and the mean extreme precipitation ($\overline{\mathbf{P}}$ from Eq. 7) are computed using only the calibration years which exclude the year for which we are going to make model predictions of the anomalous extreme precipitation. Likewise, the exponential model is fit to the data from the calibration years, and model predictions are produced for the given validation year. For each validation year in turn, we also retain the observed precipitation anomalies computed using mean precipitation calculated from the current set of calibration years. Then, we step to the next yearuntil we have covered the entire time period, 1951-2023. In Figures 12a and 12b~~

4.2 Estimated Scaling Rates

With regards to the models that we evaluated, our model which uses normalized data is found to provide the best cross-validated skill. For Rx1hr, the best model performance was using a 100 km spatial window and a 3-month temporal window. While for Rx1day, the best model performance was using a 50 km spatial window and a 3-month temporal window. We then use the normalized data with these respective model parameters governing the spatial and temporal extents to estimate the scaling rates for each station and each month of the year. The top row of panels in Figure 9a plots the scaling rates of Rx1hr at each station for each month of the year. In Figure 9b, we plot the ~~modeled versus the observed mean shifts in the maximum daily and hourly precipitation rates as a function of the 12 calendar months and the 6 dew point temperature anomalies. These mean distributional shifts are the average of the data anomalies which have been computed through the leave-one-year-out cross-validation. We observe that the skill of the predictions increases with respect to our split sample validation. This is because we use more data to calculate our means in Eqs. 6 and 7, and more data points on which to validate. Now, our skill scores with respect to climatology are 0.74 ($p<0.01$) and 0.71 ($p<0.01$) for predicting maximum daily and maximum hourly precipitation, respectively.Similarly, our skill scores are 0.39 ($p<0.01$)~~scaling rates of Rx1day at each station for each month of the year. In Figure 9c, a boxplot shows the distribution of scaling rates for both Rx1hr and ~~0.45 (p~~Rx1day for each month of the year. The median scaling rates for both Rx1hr and Rx1day, in the winter months, exhibit sub C-C (i.e., < ~~0.01) with respect to CC (see Figs. 12a and 12b). We can further disaggregate the skill, and plot it as a function of calendar month. In each month, for example, we use the predicted versus observed 6 dew point temperature anomalies to compute the corresponding skill scores. We use the same modeling framework to generate 100 randomized simulations to test the statistical significance of the skill as a function calendar month. For each of the simulations, randomly selected precipitation anomalies are chosen from the calibration period and used as the model predictions. Each set of the 100 randomized simulations of predicted precipitation anomalies are evaluated against climatology in order to establish statistical significance. The randomized predictions are also superimposed on top of the theoretical CC relationship, at the given dew point temperature anomalies, in order to evaluate the statistical significance of the skill scores with respect to the theoretical CC scaling rate. We show in Figs. 12c and 12d that the skills of our proposed methodological approach are statistically significantly ($p<0.05$) better than climatology for all calendar months for both the maximum daily and maximum hourly precipitation anomalies. At the same time, our proposed methodology performs statistically significantly better ($p<0.05$)than the theoretical CC scaling rate for nearly all months at both the daily and hourly resolutions. It is only for the maximum daily data, and for the months of January and March, that the exponential model fit does not statistically significantly outperform the theoretical CC scaling rate. What this tells us, is that the~~ 7% per °C). Additionally, the wintertime Rx1day is found to scale more strongly, on average, than Rx1hr. On the other hand, the median scaling rates in the summer months exhibit super C-C (> 7% per °C). In contrast to the winter, the summertime Rx1hr scales more strongly, on average, than Rx1day. Some stations in the summer months exhibit scaling rates that are approximately triple the C-C scaling ~~rates at this time is near CC, and that during instances such as these, we cannot necessarily provide model predictions that are an improvement over assuming the theoretical CC scaling~~ rate. The median P-T scaling rates for ~~maximum daily and maximum hourly precipitation are plotted in Figs. 12e and 12f~~ Rx1hr, as a function of ~~calendar month. The scaling rates for both the maximum daily and maximum hourly precipitation are observed to be higher during the summer months than they are for the winter months. We find that the spread between the summer and winter scaling~~

rates is greater for maximum hourly precipitation than it is for maximum daily precipitationmonth, range between 4.6% per °C for January to 17.0% per °C for June. While for Rx1day, the scaling rates range between 5.2% per °C in December to 12.9% per °C in June.

5 Conclusions

Using point-scale station data from the Upper Colorado River Basin, this study begins by illustrating some of the most common challenges in estimating P-T scaling rates. Our aim, herein, has not been to provide a comprehensive overview of every existing methodology related to P-T scaling rates, but rather to focus on some of the prevailing challenges confronting scaling rate estimation along with some proposed solutions. We find three that there are two primary challenges that require careful consideration prior to implementing any estimation of scaling rates. These are: 1) using pooled data across multiple stations and/or months, without normalization, can lead to inaccurately estimating the scaling rate due to climatological differences that exist in both space (from station to station) and time (from month to month), and 2) differences in sample sizes, and 3) the temporal the statistical independence of the data . Applying a binning method with non-normalized data which is pooled from multiple stations and/or months fails to address all three of these problems. The methodology proposed by Zhang et al. (2017) resolves many of these problems by using the normalized values of maximum precipitation accumulations and in time.

Drawing upon prior research (Wasko and Sharma, 2014; Zhang et al., 2017; Molnar et al., 2015; Visser et al., 2021), we propose a strategy which reduces the impact of these two problems. Our approach relies on using average monthly dew point temperatures which are found over a seasonal window. Their approach effectively deals with the second and as a predictor of Rx1hr and Rx1day precipitation. We then implement a regression model which fits an exponential function between these two variables for a multitude of cases. For each precipitation variable, such as Rx1hr, we make cross-valided predictions using three models and compare the model error with respect to if we had always assumed climatology. The three models are: 1) always assuming C-C scaling, 2) using the data without normalization, and 3) using the data with normalization. Additionally, we make the cross-validated predictions using a range of spatial and temporal extents. We find that both the non-normalized and normalized models are capable of providing better predictions of Rx1hr and Rx1day than either relying on climatology or the C-C scaling rate. Furthermmore, the third problems, while only partly addresses the first problem. In this paper, we have shown that even climatological differences across a season can lead to an inaccurate estimation of the effective scaling rates. In particular, we showed in the normalized data model is able to more effectively leverage pooled data when compared to the non-normalized data model. As a result, the best model performances are obtained through the normalized model for both Rx1hr and Rx1day. In the case of the UCRB that the average dew point temperatures in August are more than 8°C warmer than October, while at the same time the average extreme precipitation in August is slightly less than that of October. Without first normalizing the data at the station-month level (where anomalies are computed station by station and month by month), these climatological differences across time leads to an underestimation of the scaling rate in the UCRB for the August-September-October season.

~~Building on the work of Zhang et al. (2017), we advocate for estimating scaling rates~~ the UCRB, we find optimal performance using normalized data ~~, or data anomalies, at the station-month level. Using normalized data allows us to more effectively~~ along with a 3-month temporal window and approximately a 50 to 100 km spatial extent. Regarding the non-normalized model, it is clear that the best performance is achieved when using very narrow windows of spatial and temporal extents, such as using only data from the station itself for a given month (i.e., which uses all the the years for that station and that month). Given the relatively sparse hourly dataset, we found the non-normalized model is able to leverage data from ~~multiple stations and across multiple months. By using the normalized data, we circumvent all three of the problems we raise in the first part of our paper. Then, we use different methods to estimate P-T scaling rates over different calibration periods, with and without the normalized data . Next, we go a step further and use these estimates to make cross-validated predictions of extreme precipitation , at the station-month level, as a function of dew point temperature. The performance of the predicted mean shift in the distributions versus the observed shift is subsequently evaluated at different dew point temperature anomalies between -3°C and +3°C. Our results clearly show~~ adjacent stations up to 50 km away, but the performance degrades with attempts to pool data from further away in space or time. For the daily dataset, which has increased data coverage, the best performance is found without implementing any data pooling. After normalizing the data, however, we can more effectively pool temporally and spatially adjacent data, and as a result we can improve our estimates of our scaling rates and the associated predictions of precipitation extremes. 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 ~~the data prior to estimating any scaling ratesand producing predictions of extreme precipitation . The performance of a binning modelwhich uses raw, or non-normalized, data is seen to quickly degrade as stations are pooled from further than~~ 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. 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.

Using our best performing model, which uses normalized data pooled from 50 km ~~away and/or when adjacent months are used (e.g., using data in June and August to predict July). In contrast, the skill improves when using a binning model with normalized data which has been pooled across a basin/region and across~~ or 100 km (for Rx1hr and Rx1day, respectively) and a 3-month window~~. We find that the predictive skill is further enhanced when using a parametric exponential model fit instead of a binned model. The predictions from the exponential model are found to be statistically significantly more skillful than either climatology or assuming a theoretical CC scaling rate.~~

~~Lastly, we show the estimated scaling rates in the UCRB across the twelve calendar months for both the maximum hourly and maximum daily precipitation extremes~~, we finally show the scaling rates at each station for each month of the year. The scaling rates ~~in this region are observed to exhibit seasonality, with the scaling rates in the winter months being below the theoretical CC rate of 7% per °C and the scaling rates in the summer months being more than double CC. Additionally, we find the variability of the scaling rates between the summer and winter months to be greater for maximum hourly precipitation~~

~~than it is for the maximum daily precipitation.~~ associated with daily extreme precipiation, Rx1day, are found to exhibit less variability throughout the year than scaling rates associated with hourly extreme precipitation, Rx1hr. The median of the June scaling rates, for Rx1hr, is more than triple the median of the January scaling rates. Substantial variability is seen in the the scaling rates across both space and time, and as a result it is advisable that one estimates a unique P-T scaling rate at every given station (or grid cell if using gridded data) and for every given time of year (such as a month).

In this study, we have achieved skillful cross-validated predictions of extreme Rx1hr and Rx1day precipitation in the UCRB by ~~using normalized data in a parametric exponential model . With this knowledge in hand~~fitting an exponential model to normalized data. By applying this methodology to other regions of the world, we can ~~continue to improve our~~ gain a more detailed and comprehensive understanding of how P-T scaling rates vary across ~~different months in other regions. Furthermore, one can evaluate how skillfully those rates can be applied in a changing climate.~~ space and time. This information is essential in our efforts to improve our preparedness regarding extreme precipitation.

Data availability. Supporting data can be found at https://doi.org/10.6084/m9.figshare.29858954.v1 (Switanek et al., 2025).

Appendix A: Comparison of the Quality-Controlled Hourly and Daily Datasets

Both the hourly (GHCNH) and daily (GHCND) datasets used in this paper are taken from larger datasets, which contain global coverage. Here, we can provide some comparisons which draws upon data from these larger datasets for the UCRB, in addition to data from the broader contiguous United States (CONUS). In Figure A1, we show how well the datasets align with one another given the same period of temporal aggregation. Using the hourly dataset, we compute Rx1day (i.e., maximum aggregated precipitation over 24 hours, found for each station/month). Then, for each station in the hourly (GHCNH) dataset, we find the closest daily (GHCND) station. We can then compare the normalized quantities (by applying Eq. 2) over the 74 years and 12 months for all stations in the GHCNH dataset. In Fig. A1a, we plot the relationship between all of the normalized Rx1day values computed from the hourly dataset, and the Rx1day values from the closest daily stations. This is shown for CONUS and the UCRB. We can next look at the relationship between the daily data itself. We can plot in Fig. A1b the relationship between the normalized Rx1day at each GHCND station and its closest neighboring GHCND station. We find good agreement between the two quality-controlled datasets, with a correlation coefficient between the all of the values in Fig. A1a is 0.86, and a value of 0.78 for Fig. A1b. However, this decrease in correlation can be explained by the fact that the average distance between the stations is 4 km in Fig. A1a, while it is 8 km in Fig. A1b. Next, we compare in Fig. A1c the average monthly dew point temperature anomalies from the GHCNH dataset to the nearest grid cell of average monthly dew point temperature anomalies from ERA5 (which is used as our predictor of Rx1day). Again, we find good agreement between the two datasets with a correlation coefficient of 0.90.

Another way that we can see whether there are any unrealistic statistical outliers in the datasets is to plot all of the normalized dew point temperature anomalies versus normalized Rx1hr or Rx1day. We do this using all of the station-months in the UCRB

for the hourly (i.e., average monthly GHCNH dew point temperature anomlies versus Rx1hr precipitation anomalies) and daily (i.e., average monthly ERA5 dew point temperature anomlies versus Rx1day precipitation anomalies) datasets. We do not find the presence of any unrealistic outliers, in either the hourly or daily datasets, when considering the associated covariance between our two variables.

~~In (a), anomalies of dew point temperature are plotted against anomalies of maximum daily precipitation. For each station, anomalies are computed with respect to the August-September-October seasonal average for that station. (a) illustrates the climatological differences in time of the data. September is not shown in (a) in order to more clearly visualize the climatological differences between two differences months within a 3-month season. (b) Different scaling rates, which have been centered at 0°C along the x-axis and 100% or normal along the y-axis, are plotted. The scaling rate obtained from the data in (a), including the data from September, is plotted as the gray dashed line. The solid black line is the scaling rate using the same data over the same season, but where anomalies are computed at the station-month level. The red, green, and blue lines are the scaling rates from each month individually. (c) is similar to (a), but now isolating the impact of spatial climatological differences. For the month of October, anomalies are computed with respect to the all of the stations across the UCRB. (c) plots the anomalies east and west of 108° longitude. The scaling rate fit to the values in (c) is plotted as the gray dashed line in (d), while the scaling rate for the season centered about October, using anomalies computed at the station-month level, are plotted as the solid black line.~~

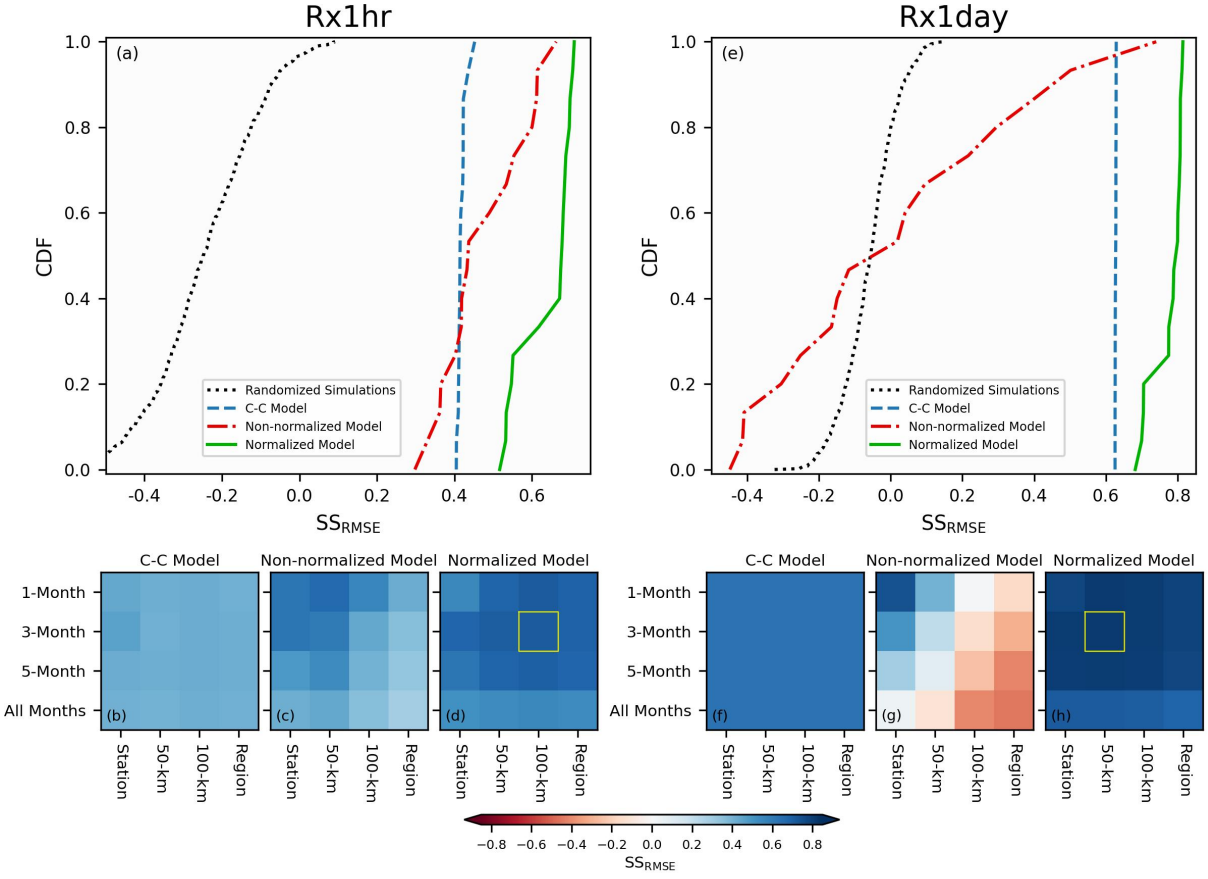

Figure 7. (a) plots the skills as empirical CDFs of the leave-one-year-out cross-validated predictions of Rx1hr for the three different models: C-C, non-normalized, and normalized. The CDFs are comprised of the skills from the different possible iterations of the spatial and temporal extents. Additionally, the CDF of 1000 randomized simulations is shown as the black dotted line. (b),(c), and (d) show the skill scores as a function of model (corresponding to subplot title), spatial extent (x-axis), and temporal extent (y-axis). The maximum skill achieved is highlighted with the bounded yellow box. (e-h) are the same as (a-d), except showing the skills in predicting Rx1day.

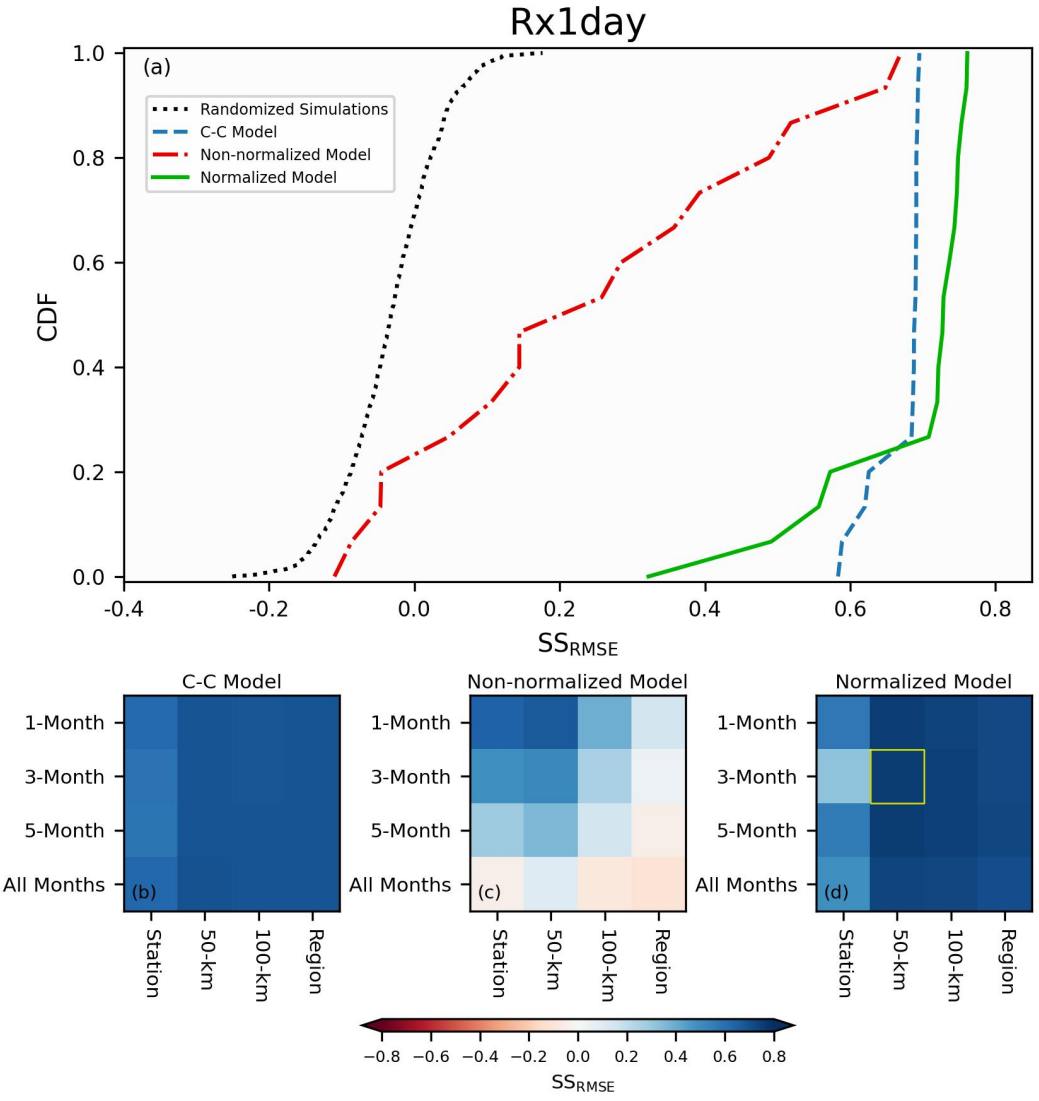

Figure 8. (a) ~~Leave-one-out~~ plots the skills as empirical CDFs of the two-fold cross-validated ~~modeled mean shifts in~~ predictions of Rx1day for the ~~maximum daily precipitation~~ three different models: C-C, non-normalized, and normalized. The CDFs are ~~plotted against~~ comprised of the ~~observed mean shifts for~~ skills from the ~~12 months~~ different possible iterations of the ~~year~~ spatial and temporal extents. Additionally, the ~~6 dew point temperature anomalies~~CDF of 1000 randomized simulations is shown as the black dotted line. (b)~~The same as~~ (ac), ~~except using maximum hourly precipitation.~~ and (cd) ~~The~~ show the skill ~~is plotted~~ scores as a function of ~~the time of year for the maximum daily precipitation predictions. The thick black and brown lines are the RMSE skill scores with respect to climatology and CC, respectively. The average skill scores of randomly generated forecasts are plotted as the thin dashed lines, with the 95% confidence interval represented by the shaded regions.~~ model (dcorresponding to subplot title)~~The same as (c), except using maximum hourly precipitation.~~ spatial extent (ex-axis), and temporal extent (~~f~~y-axis)~~Plot the scaling rates for the different months of the year for the maximum daily and maximum hourly precipitation, respectively~~. The ~~green bars are~~ maximum skill achieved is highlighted with the ~~scaling rates obtained by fitting the exponential function from Eq~~bounded yellow box.~~8 to the normalized data using all of the stations in the UCRB and using a 3-Month window. The 5% and 95% confidence intervals are obtained through bootstrapping.~~

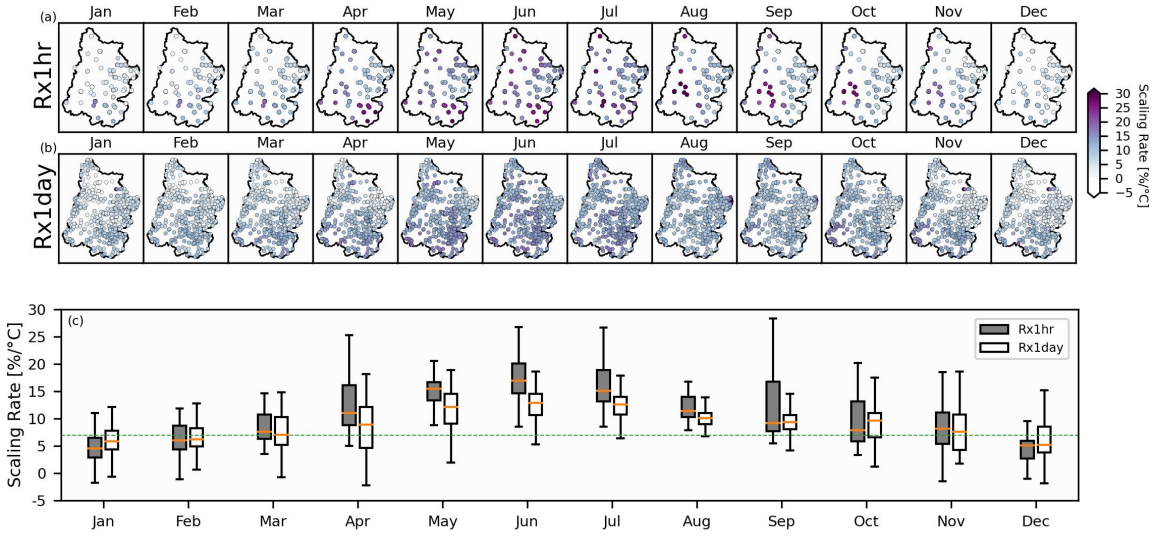

Figure 9. The top row of (a) plots the scaling rates of Rx1hr at each station for each month of the year. (b) same as (a) except for Rx1day. (c) uses a boxplot to show the distribution of the scaling rates as a function of Rx1hr or Rx1day along with the month of the year. In (c), each box, median, and whiskers are constructed using the scaling rates from all of the stations in the UCRB corresponding to that month and either Rx1hr or Rx1day. The gray boxplot for July, as an example, is constructed using the scaling rates from the stations in the "Jul" panel in subplot (a).

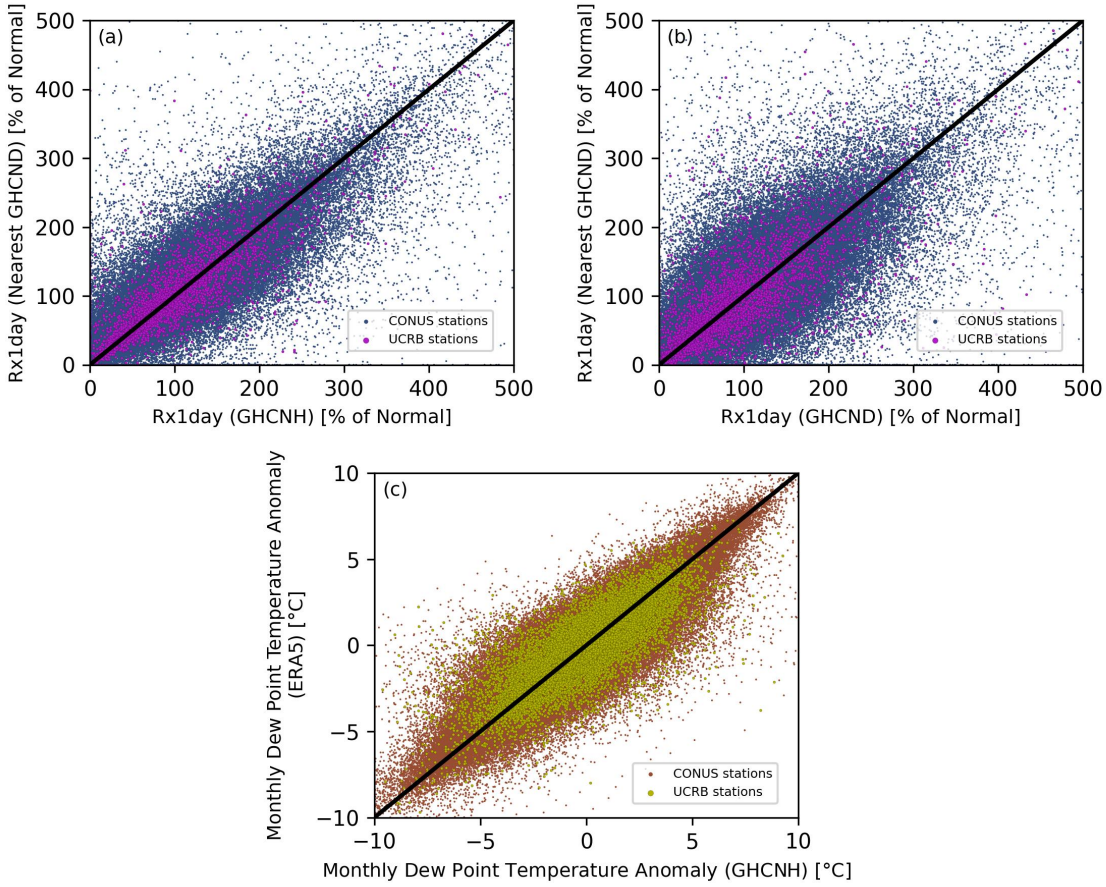

Figure A1. (a) Anomalies of Rx1day values are compared when Rx1day is computed from the hourly dataset (x-axis) or from the daily dataset (y-axis). This is shown for all station/months across CONUS and the UCRB. For each station in the hourly dataset, the data from the nearest station from the daily dataset is found and contrasted. (b) Plots the relationship between the daily data (GHCND) itself. For each GHCND station, the closest neighboring GHCND station is found and the anomalies of the two are compared. This is shown using all of the station/months across CONUS and the UCRB. (c) plots the average monthly dew point temperature anomalies from the GHCNH dataset against the nearest grid cell of average monthly dew point temperature anomalies from ERA5.

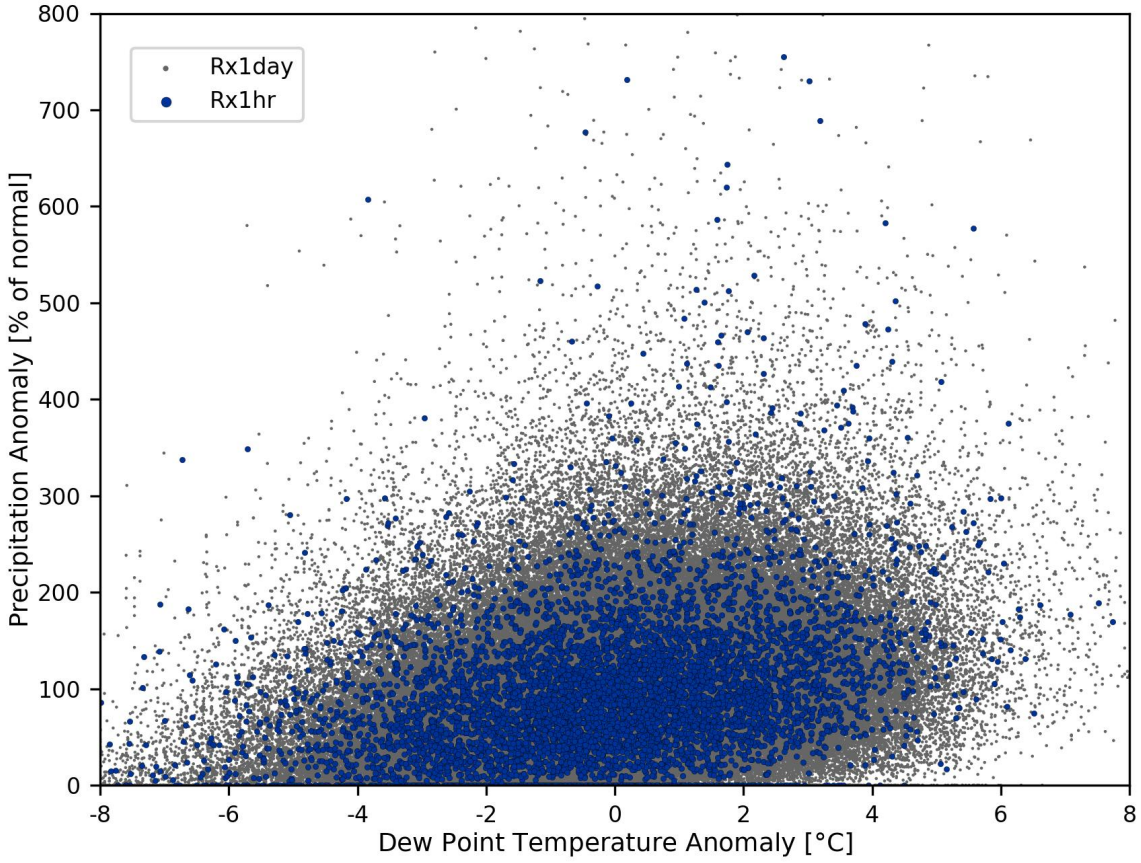

Figure A2. The small points show the normalized anomalies of dew point temperature versus the normalized anomalies of Rx1day using the UCRB stations from the daily dataset GHCND. The larger blue points show the normalized anomalies of dew point temperature versus the normalized anomalies of Rx1hr using the UCRB stations from the hourly dataset GHCNH.

Author contributions. The study was conceived by MS. All code, analysis, and figures were produced by MS with input from all coauthors. The paper was written by MS with assistance from all of the other coauthors.

Competing interests. The authors do not have any competing interests.

Acknowledgements. The authors want to thank the California Department of Water Resources (contract number: 4600015149) and the University of Graz for funding this research. The lead author would additionally like to thank Andreas Prein for some early discussions pertaining to this research.

References

Abatzoglou, J. T., Marshall, A. M., Lute, A. C., and Safeeq, M.: Precipitation dependence of temperature trends across the contiguous US, Geophysical Research Letters, 49, e2021GL095 414, https://doi.org/10.1029/2021GL095414, 2022.

Alduchov, O. A. and Eskridge, R. E.: Improved Magnus Form Approximation of Saturation Vapor Pressure, Journal of Applied Meteorology, 35(4), 601–609, 1996.

Ali, H. and Mishra, V.: Contrasting response of rainfall extremes to increase in surface air and dewpoint temperatures at urban locations in India, Scientific Reports, 7, 1228, https://doi.org/10.1038/s41598-017-01306-1, 2017.

Ali, H., Fowler, H. J., and Mishra, V.: Global observational evidence of strong linkage between dew point temperature and precipitation extremes, Geophysical Research Letters, 45(12), 12,320–12,330, https://doi.org/10.1029/2018GL080557, 2018.

Ali, H., Fowler, H. J., Lenderink, G., Lewis, E., and Pritchard, D.: Consistent large-scale response of hourly extreme precipitation to temperature variation over land, Geophysical Research Letters, 48, e2020GL090 317, https://doi.org/10.1029/2020GL090317, 2021.

Ali, H., Fowler, H. J., Pritchard, D., Lenderink, G., Blenkinsop, S., and Lewis, E.: Towards quantifying the uncertainty in estimating observed scaling rates, Geophysical Research Letters, 49, e2022GL099 138, https://doi.org/10.1029/2022GL099138, 2022.

Allen, M. R. and Ingram, W. J.: Constraints on future changes in climate and the hydrologic cycle, Nature, 419, 224–232, https://doi.org/10.1038/nature01092, 2002.

Ban, N., Schmidli, J., and Schär, C.: Heavy precipitation in a changing climate: Does short-term summer precipitation increase faster?, Geophysical Research Letters, 42, 1165–1172, https://doi.org/10.1002/2014GL062588, 2015.

Barbero, R., Westra, S., Lenderink, G., and Fowler, H. J.: Temperature-extreme precipitation scaling: a two-way causality?, International Journal of Climatology, 38, e1274–e1279, https://doi.org/10.1002/joc.5370, 2018.

Berg, P. and Haerter, J. O.: 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, 2013.

Berg, P., Haerter, J. O., Thejll, P., Piani, C., Hagemann, S., and Christensen, J. H.: Seasonal characteristics of the relationship between daily precipitation intensity and surface temperature, Journal of Geophysical Research: Atmospheres, 114, D18 102, https://doi.org/10.1029/2009JD012008, 2009.

Boessenkool, B., Bürger, G., and Heistermann, M.: Effects of sample size on estimation of rainfall extremes at high temperatures, Natural Hazards Earth System Science, 17(9), 1623–1629, https://doi.org/10.5194/nhess-17-1623-2017, 2017.

Chiappa, J., Parsons, D. B., Furtado, J. C., and Shapiro, A.: Short-duration extreme rainfall events in the central and eastern United States during the summer: 2003–2023 trends and variability, Geophysical Research Letters, 51, e2024GL110 424, https://doi.org/10.1029/2024GL110424, 2024.

Dollan, I. J., Maggioni, V., Johnston, J., de A. Coelho, G., and III, J. L. K.: Seasonal variability of future extreme precipitation and associated trends across the Contiguous U.S., Frontiers in Climate, 4, 954 892, https://doi.org/10.3389/fclim.2022.954892, 2022.

Donat, M. G., Delgado-Torres, C., Luca, P. D., Mahmood, R., Ortega, P., and Doblas-Reyes, F. J.: How credibly do CMIP6 simulations capture historical mean and extreme precipitation changes?, Geophysical Research Letters, 50, e2022GL102 466, https://doi.org/10.1029/2022GL102466, 2023.

Drobinski, P., Silva, N. D., Panthou, G., Bastin, S., Muller, C., Ahrens, B., Borga, M., Conte, D., Fosser, G., Giorgi, F., Güttler, I., Kotroni, V., Li, L., Morin, E., Önol, B., Quintana-Segui, P., Romera, R., and Torma, C. Z.: Scaling precipitation extremes with tem-

perature in the Mediterranean: past climate assessment and projection in anthropogenic scenarios, Climate Dynamics, 51, 1237–1257, https://doi.org/10.1007/s00382-016-3083-x, 2018.

Estermann, R., Rajczak, J., Velasquez, P., Lorenz, R., and Schär, C.: Projections of heavy precipitation characteristics over the Greater Alpine Region using a kilometer–scale climate model ensemble, Journal of Geophysical Research: Atmospheres, 130, e2024JD040 901, https://doi.org/10.1029/2024JD040901, 2025.

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., Lewis, E., Schaer, C., Sharma, A., Villarini, G., Wasko, C., and Zhang, X.: Anthropogenic intensification of short-duration rainfall extremes, Nature Reviews Earth and Environment, 2, 107–122, https://doi.org/10.1038/s43017-020-00128-6, 2021.

Gründemann, G. J., van de Giesen, N., Brunner, L., and van der Ent, R.: Rarest rainfall events will see the greatest relative increase in magnitude under future climate change, Communications Earth and Environment, 3, 235, https://doi.org/10.1038/s43247-022-00558-8, 2022.

Gu, L., Yin, J., Gentine, P., Wang, H.-M., Slater, L. J., Sullivan, S. C., Chen, J., Zscheischler, J., and Guo, S.: Large anomalies in future extreme precipitation sensitivity driven by atmospheric dynamics, Nature Communications, 14, 3197, https://doi.org/10.1038/s41467-023-39039-7, 2023.

Harp, R. D. and Horton, D. E.: Observed changes in daily precipitation intensity in the United States, Geophysical Research Letters, 49, e2022GL099 955, https://doi.org/10.1029/2022GL099955, 2022.

Haslinger, K., Breinl, K., Pavlin, L., Pistotnik, G., Bertola, M., Olefs, M., Greilinger, M., Schöner, W., and Blöschl, G.: Increasing hourly heavy rainfall in Austria reflected in flood changes, Nature, 639, 667–672, https://doi.org/10.1038/s41586-025-08647-2, 2025.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Sabater, J. M., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N.: ERA5 hourly data on single levels from 1940 to present, https://doi.org/10.24381/cds.adbb2d47, accessed on 10-09-2024, 2023.

Higgins, T. B., Subramanian, A. C., Watson, P. A. G., and Sparrow, S.: Changes to Atmospheric River Related Extremes Over the United States West Coast Under Anthropogenic Warming, Geophysical Research Letters, 52, e2024GL112 237, https://doi.org/10.1029/2024GL112237, 2025.

Huang, X. and Swain, D. L.: Climate change is increasing the risk of a California megaflood, Science Advances, 8(31), eabq0995, https://doi.org/10.1126/sciadv.abq099, 2022.

Jones, R. H., Westra, S., and Sharma, A.: Observed relationships between extreme sub-daily precipitation, surface temperature, and relative humidity, Geophysical Research Letters, 37, L22 805, https://doi.org/10.1029/2010GL045081, 2010.

Jong, B. T., Delworth, T. L., Cooke, W. F., Tseng, K. C., and Murakami, H.: Increases in extreme precipitation over the Northeast United States using high-resolution climate model simulations, npj Climate and Atmospheric Science, 6(1), 18, https://doi.org/10.1038/s41612-023-00347-w, 2023.

Lenderink, G. and van Meijgaard, E.: Increase in hourly precipitation extremes beyond expectations from temperature changes, Nature Geoscience, 1, 511–514, https://doi.org/10.1038/ngeo262, 2008.

Lenderink, G. and van Meijgaard, E.: Linking increases in hourly precipitation extremes to atmospheric temperature and moisture changes, Environmental Research Letters, 5, 025 208, https://doi.org/10.1088/1748-9326/5/2/025208, 2010.

Lenderink, G., Mok, H. Y., Lee, T. C., and van Oldenborgh, G. J.: Scaling and trends of hourly precipitation extremes in two different climate zones – Hong Kong and the Netherlands, Hydrology and Earth System Sciences, 15, 3033–3041, https://doi.org/10.5194/hess-15-3033-2011, 2011.

Li, X., Wang, T., Zhou, Z., Su, J., and Yang, D.: Seasonal characteristics and spatio-temporal variations of the extreme precipitation-air temperature relationship across China, Environmental Research Letters, 18, 054 022, https://doi.org/10.1088/1748-9326/acd01a, 2023.

Marra, F., Koukoula, M., Canale, A., and Peleg, N.: Predicting extreme sub-hourly precipitation intensification based on temperature shifts, Hydrology and Earth System Sciences, 28, 375–389, https://doi.org/10.5194/hess-28-375-2024, 2024.

Martinez-Villalobos, C. and Neelin, J. D.: Regionally high risk increase for precipitation extreme events under global warming, Scientific Reports, 13, 5579, https://doi.org/10.1038/s41598-023-32372-3, 2023.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G.: An Overview of the Global Historical Climatology Network-Daily Database, Journal of Atmospheric and Oceanic Technology, 29(7), 897–910, https://doi.org/10.1175/JTECH-D-11-00103.1, 2012.

Meresa, H., Tischbein, B., and Mekonnen, T.: Climate change impact on extreme precipitation and peak flood magnitude and frequency: observations from CMIP6 and hydrological models, Natural Hazards, 111, 2649–2679, https://doi.org/10.1007/s11069-021-05152-3, 2022.

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

Myhre, G., Alterskjær, K., Stjern, C. W., Hodnebrog, Marelle, L., Samset, B. H., Sillmann, J., Schaller, N., Fischer, E., Schulz, M., and Stohl, A.: Frequency of extreme precipitation increases extensively with event rareness under global warming, Scientific Reports, 9, 16 063, https://doi.org/10.1038/s41598-019-52277-4, 2019.

Najibi, N. and Steinschneider, S.: Extreme precipitation-temperature scaling in California: The role of atmospheric rivers, Geophysical Research Letters, 50, e2023GL104 606, https://doi.org/10.1029/2023GL104606, 2023.

Najibi, N., Mukhopadhyay, S., and Steinschneider, S.: Precipitation scaling with temperature in the Northeast US: Variations by weather regime, season, and precipitation intensity, Geophysical Research Letters, 49, e2021GL097 100, https://doi.org/10.1029/2021GL097100, 2022.

Panthou, G., Mailhot, A., Laurence, E., and Talbot, G.: Relationship between Surface Temperature and Extreme Rainfalls: A Multi-Time-Scale and Event-Based Analysis, Journal of Hydrometeorology, 15(5), 1999–2011, https://doi.org/10.1175/JHM-D-14-0020.1, 2014.

Prein, A. F., Rasmussen, M., Ikeda, K., Liu, C., Clark, M. P., and Holland, G. J.: The future intensification of hourly precipitation extremes, Nature Climate Change, 7, 48–52, https://doi.org/10.1038/NCLIMATE3168, 2017.

Rahat, S. H., Saki, S., Khaira, U., Biswas, N. K., Dollan, I. J., Wasti, A., Miura, Y., Bhuiyan, M. A. E., and Ray, P.: Bracing for impact: how shifting precipitation extremes may influence physical climate risks in an uncertain future, Scientific Reports, 14, 17 398, https://doi.org/10.1038/s41598-024-65618-9, 2024.

Smith, A., Lott, N., and Vose, R.: The integrated surface database: Recent developments and partnerships, Bulletin of the American Meteorological Society, 92(6), 704–708, https://doi.org/10.1175/2011BAMS3015.1, 2011.

Sokol, Z., Řezáčová, D., and Popová, J.: Change in the distribution of heavy 1 h precipitation due to temperature changes in measured values, model reanalyses and model simulations of future climate, Atmospheric Research, 304, 107 395, https://doi.org/10.1016/j.atmosres.2024.107395, 2024.

Sun, X. and Wang, G.: Causes for the Negative Scaling of Extreme Precipitation at High Temperatures, Journal of Climate, 35, 6119–6134, https://doi.org/10.1175/JCLI-D-22-0142.s1, 2022.

Switanek, M. B.: GH2D-MetNet: A global hourly-to-daily dataset of observed precipitation, temperature, and dew point temperature, (in preparation for Earth System Science Data), 2025.

Switanek, M. B., Abermann, J., Schöner, W., and Anderson, M. L.: Data for the paper, "Precipitation-temperature scaling: current challenges and proposed methodological strategies", figshare. Dataset. https://doi.org/10.6084/m9.figshare.29858954.v1, 2025.

Tabari, H.: Climate change impact on flood and extreme precipitation increases with water availability, Scientific Reports, 10, 13 768, https://doi.org/10.1038/s41598-020-70816-2, 2020.

Tian, B., Chen, H., Yin, J., Liao, Z., Li, N., and He, S.: Global scaling of precipitation extremes using near-surface air temperature and dew point temperature, Environ. Res. Lett., 18, 034 016, https://doi.org/10.1088/1748-9326/acb836, 2023.

Trenberth, K. E., Dai, A., Rasmussen, R. M., and Parsons, D. B.: The changing character of precipitation, Bulletin of the American Meteorological Society, 84(9), 1205–1218, https://doi.org/10.1175/BAMS-84-9-1205, 2003.

Utsumi, N., Seto, S., Kanae, S., Maeda, E. E., and Oki, T.: Does higher surface temperature intensify extreme precipitation?, Geophysical Research Letters, 38, L16 708, https://doi.org/10.1029/2011GL048426, 2011.

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.

Wang, G., Wang, D., Trenberth, K. E., Erfanian, A., Yu, M., Bosilovich, M. G., and Parr, D. T.: The peak structure and future changes of the relationships between extreme precipitation and temperature, Nature Clim. Change, 7, 268–274, https://doi.org/10.1038/nclimate3239, 2017.

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

Wasko, C., Sharma, A., and Johnson, F.: Does storm duration modulate the extreme precipitation-temperature scaling relationship?, Geophys. Res. Lett., 42, 8783–8790, https://doi.org/10.1002/2015GL066274, 2015.

Wasko, C., Lu, W. T., and Mehrotra, R.: Relationship of extreme precipitation, dry-bulb temperature, and dew point temperature across Australia, Environmental Research Letters, 13, 074 031, https://doi.org/10.1088/1748-9326/aad135, 2018.

Westra, S., Fowler, H. J., Evans, J. P., Alexander, L. V., Berg, P., Johnson, F., Kendon, E. J., Lenderink, G., and Roberts, N. M.: Future changes to the intensity and frequency of short-duration extreme rainfall, Rev. Geophys., 52, 522–555, https://doi.org/10.1002/2014RG000464, 2014.

Yin, J., Gentine, P., Zhou, S., Sullivan, S., Wang, R., Zhang, Y., and Guo, S.: Large increase in global storm runoff extremes driven by climate and anthropogenic changes, Nature Communications, 9, 4389, https://doi.org/10.1038/s41467-018-06765-2, 2018.

Yin, J., Guo, S., Gentine, P., He, S., Chen, J., and Liu, P.: Does the hook structure constrain future flood intensification under anthropogenic climate warming?, Water Resources Research, 57, e2020WR028 491, https://doi.org/10.1029/2020WR028491, 2021.

Zhang, X., Zwiers, F. W., Li, G., Wan, H., and Cannon, A. J.: Complexity in estimating past and future extreme short-duration rainfall, Nature Geoscience, 10, 255–259, https://doi.org/10.1038/NGEO2911, 2017.