

Leveraging normalized data to improve point-scale estimates of precipitation-temperature scaling rates

Matthew Switanek¹, Jakob Abermann¹, Wolfgang Schöner¹, and Michael L. Anderson²

¹Department of Geography and Regional Science, University of Graz, Graz, Austria

²California 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 (C-C) relationship, which predicts a $\sim 7\%$ increase in atmospheric moisture-holding capacity per $^{\circ}\text{C}$ of warming. Many studies have benchmarked observed extreme precipitation-temperature (P-T) scaling rates against this theoretical value, finding that globally averaged scaling rates of daily extreme precipitation ~~is~~ are largely consistent with C-C, while hourly extremes have been observed to scale at rates greater than C-C. 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 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 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. 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 (Lenderink and van Meijgaard, 2008; Lenderink et al., 2011; Ban et al., 2015; Tabari, 2020; Abatzoglou et al., 2022; Ali et al., 2022; Chiappa et al., 2024; Marra et al., 2024; Haslinger et al., 2025). 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°C increase in temperature. Due to the C-C relationship, as available moisture in the air column increases with warmer temperatures, intensities of extreme precipitation rates are also expected to increase (Panthou et al., 2014; Westra et al., 2014; Wasko et al., 2015; Myhre et al., 2019; Fowler et al., 2021; Gründemann et al., 2022; Harp and Horton, 2022; Donat et al., 2023).

25 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, 30 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~~ researchers have 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; Haslinger et al., 2025). Likewise, ~~the spatial extent of some prior work has been~~ some scholars have investigated scaling rates at the global scale (Ali et al., 2018; Tabari, 2020; Tian et al., 2023), while others have ~~investigated scaling rates at the~~ targeted point or regional ~~scale~~ scales (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 40 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 45 to the fact that dew point temperature contains direct information on the available moisture in the atmosphere.

In addition to decisions concerning which data to use, prior research has also proposed a variety of methods to estimate P-T scaling rates. Two of the most widely used approaches are the binning method (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 50 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) 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 precipitation. Implementation of either of these methods often relies on data pooled 55 from 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 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 also recognize that climatological differences in both time and space can be present in the data (e.g., California has more extreme daily precipitation in the winter, at lower dew point temperatures, than it does in the summer). Molnar et al. (2015)

60 clearly showed this impact, by fitting a regression model to a sample of pooled data, and then comparing to regression fits which separate the same data by whether there was lightning or not. They found that the scaling rate of the raw, pooled data was nearly double that of the samples which were split by the occurrence lightning or no-lightning. Using pooled data, then, without accounting for these climatological differences, one runs the risk of inaccurately estimating the effective scaling rates. In light of this problem, some have advised using normalized or standardized data (Zhang et al., 2017; Visser et al., 2021).
65 An additional control for seasonality was proposed by Zhang et al. (2017), where they used normalized data over the summer season.

Our work herein is built on the foundation of prior work (Zhang et al., 2017; Wasko and Sharma, 2014; Molnar et al., 2015; Visser et al., 2021). The aims of this paper are ~~twofold. First, we describe as follows. We begin by describing~~ some common challenges or problems that must be addressed in order to effectively estimate and interpret P-T scaling rates. ~~Second~~Next,
70 we suggest a methodology to resolve many of those problems. ~~Lastly, we use exponential regression. Then, an exponential regression model is used~~ to estimate P-T scaling rates under different modeling assumptions. ~~Furthermore~~Lastly, we use these ~~estimates~~estimated scaling rates to generate predictions of extreme hourly and daily precipitation through a cross-validated framework. The skill of these predictions ~~are~~is subsequently evaluated to determine which methods and/or assumptions provide the best performance.

75 **2 Data**

Hourly measurements of precipitation and dew point temperature for the Upper Colorado River Basin (UCRB) are obtained 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, 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
80 Fig. 1a.

Daily measurements of precipitation are taken from the Global Historical Climatology Network - daily (GHCN-daily or GHCND) dataset (Menne et al., 2012). Most of these stations do not measure dew point temperature in-situ. Therefore, the ERA5 Reanalysis dataset (Hersbach et al., 2023) is used to provide dew point temperatures at the GHCND stations. For each GHCND station, the nearest ERA5 grid cell is found, and the corresponding time series of dew point temperatures are used for
85 that station. This procedure is repeated for all of the 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, 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~~our Supplemental Material, we provide a comparison of the two quality-
90 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

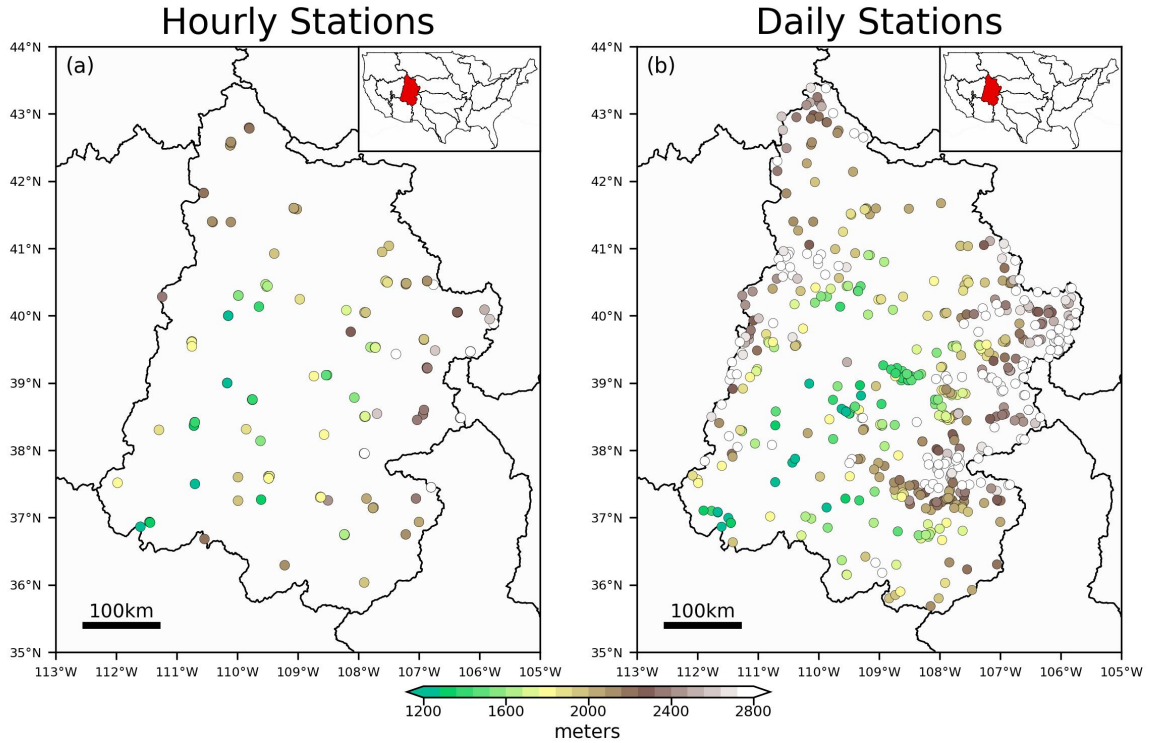


Figure 1. The hourly (a) and daily (b) distribution of stations across the Upper Colorado River Basin which are used in this study. The location of the study region is shown as the inset of the subplots alongside other large-scale hydrological basins across the contiguous United States. The color of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

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. We also 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,

$$\text{DPT}_{x,m,t}^* = \text{DPT}_{x,m,t} - \overline{\text{DPT}}_{x,m}, \quad (1)$$

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

Similarly, normalized anomalies of precipitation are computed as,

$$105 \quad \mathbf{P}_{x,m,t}^* = \frac{\mathbf{P}_{x,m,t}}{\overline{\mathbf{P}}_{x,m}} \cdot 100 \quad , \quad (2)$$

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 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.

110 2.1 Evaluation Metrics Used for Validation

We evaluate model performance 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}_{CLIM} , respectively). The \mathbf{RMSE}_{MOD} is defined as,

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

115 where y_{obs} and y_{mod} are the observed and the modeled precipitation values, respectively. Likewise, \mathbf{RMSE}_{CLIM} , reflects the error associated with a baseline model which always assumes climatology (i.e., always assuming 100% of normal, or equal to the climatological mean).

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

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

$$120 \quad \mathbf{SS}_{RMSE} = 1 - \frac{\mathbf{RMSE}_{MOD}}{\mathbf{RMSE}_{CLIM}} \quad . \quad (5)$$

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

3 Methods

3.1 Common Challenges in Estimating P-T Scaling Rates

125 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 time of year (or month) can expect some average percentage increase in extreme precipitation 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
130 a function of space (e.g., between stations or between grid cells) and time (e.g., between months of the year).

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

1. Pooling raw (non-normalized) data of dew point temperatures and precipitation across multiple stations and/or months
135 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. Data at hourly or daily resolutions cannot be assumed to be statistically independent in time.

In Figure 2, we illustrate the first of the aforementioned challenges. 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, we use all of the
140 hourly data from all of the stations within the UCRB to look at pairs of dew point temperature and precipitation which are
concurrent or collocated in time. The top 1% of precipitation intensities are shown in blue for each dew point temperature
bin (using a bin size of 2°C and incrementing by 2°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 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
145 temperatures, than other stations. We can test if this is the case by implementing the binning approach at the station/month
level. So, for each station we can use all of the hourly 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 bins. This gives us extreme precipitation
amounts that are specific to every station, every month of the year, and every dew point temperature bin. 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
150 derived at each station for each month, Fig. 2b illustrates how extreme precipitation differs between the month of May and
the month of July. Within the same dew point temperature 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 average precipitation that
falls in the top 1% of this bin is 3.60 mm/hr (with 171 samples). In contrast, the 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
155 temperature of 5°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 same dew point temperature and at the same time of year,
certain stations exhibit more extreme precipitation than others. For stations above 1800 meters, the average precipitation that
160 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 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, 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.

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. 2. We can show the issue with statistical independence in two different ways using the hourly precipitation data. First, we can compute the temporal autocorrelation of the data for each station. We do this using two different lagged time steps, where we compute lagged-1 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 functions (CDFs) in Figure 3a as the thicker 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 data points from that station. The CDFs of the stations from these randomly generated time series are seen as the 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 independent of one another ($p < 0.01$). 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. 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 1%, then what was the probability that hour, $t + 1$, or hour, $t + 4$, was also in the top 1%. The thicker lines in Fig. 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 ($p < 0.01$). 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% 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. When it comes to conditional probabilities, we find for data in the UCRB, that the data on average approaches statistical independence for events separated by approximately 12 hours for the top 0.1% and approximately 24 hours for the top 1%.

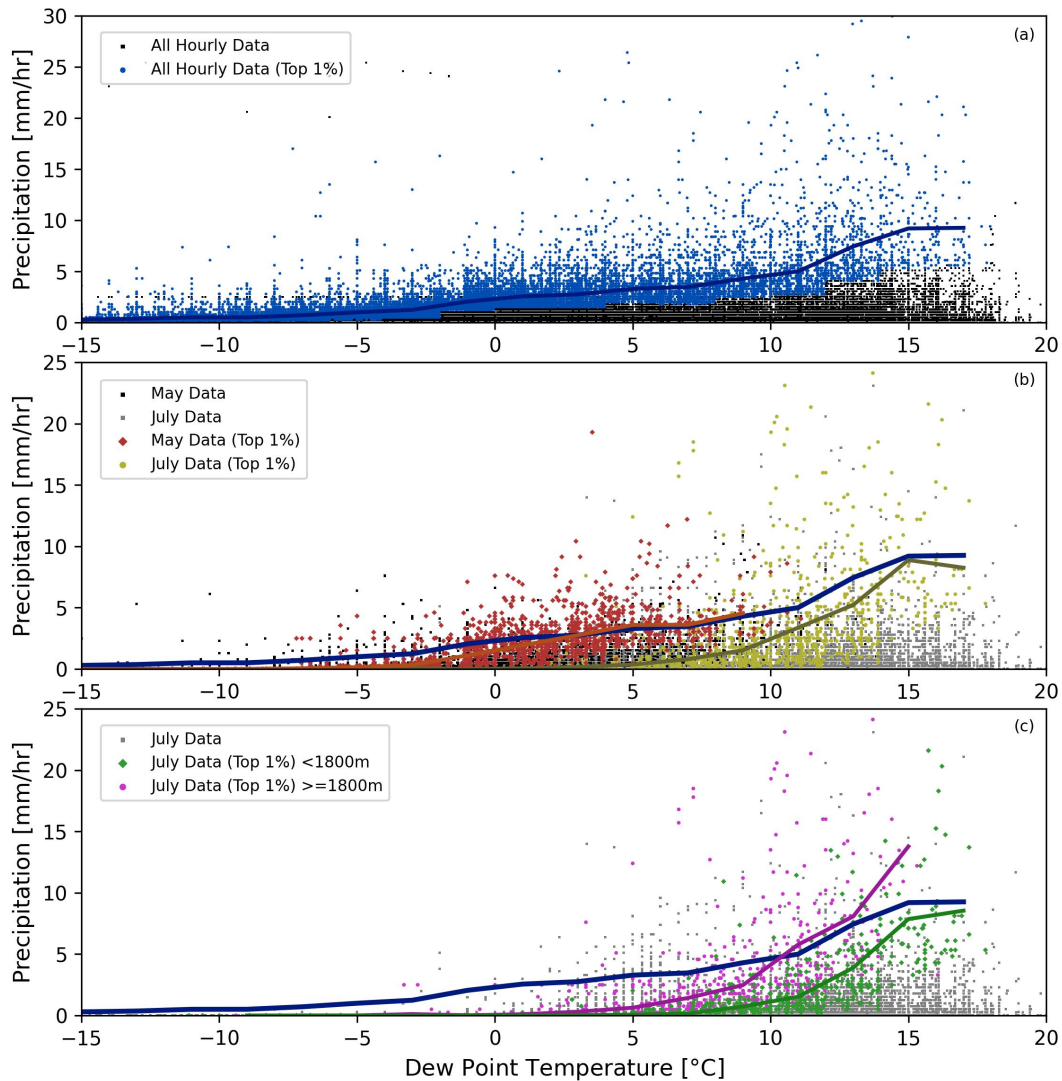


Figure 2. (a) All of the pairings of measured hourly precipitation along with the corresponding dew point temperature for the stations in the UCRB are plotted in the background as the black scatter points. The top 1% of precipitation rates are plotted as the blue scatter points. The blue line is produced through the binning method, whereby the average of the top 1% of precipitation rates are computed for each dew point temperature bin. (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.

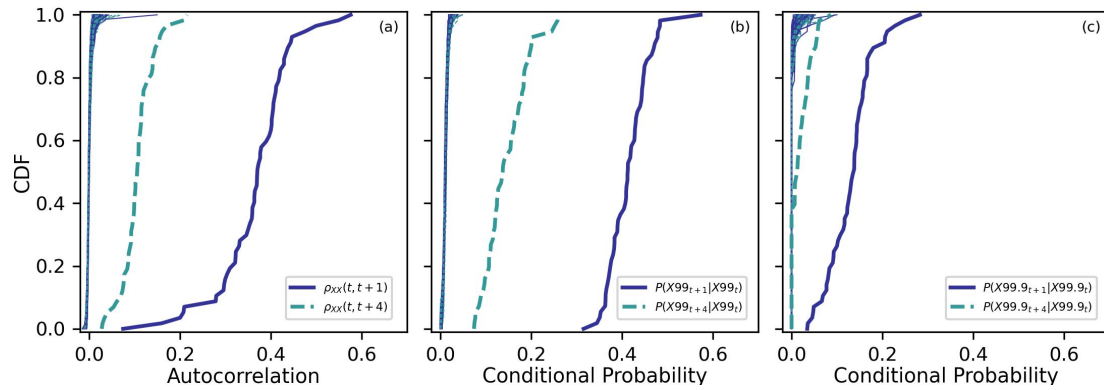


Figure 3. (a) Empirical cumulative distribution functions (CDFs) of the lagged-1 and lagged-4 hour autocorrelations from the stations in the UCRB are plotted as the blue solid and light blue dashed lines, respectively. CDFs from randomly generated time series are seen as the thin lines on the left. (b) 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.

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 methodology which fits an exponential regression methodology using model to the](#) monthly precipitation maxima Rx1hr and Rx1day. 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 an effective method of estimating scaling rates where they can be used to predict expected changes in extreme precipitation given changes in dew point temperatures. Estimating concurrent relationships between dew point temperature 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 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 without prior knowledge of when the most extreme precipitation event happened, we cannot say ahead of time at which hour, or dew point temperature, the most extreme precipitation event will take place. It is possible that the most extreme hourly precipitation rate can be concurrent with any of the dew point temperature 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 a shift in the mean of the distribution over the month. In fact, 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 [with extreme precipitation](#).

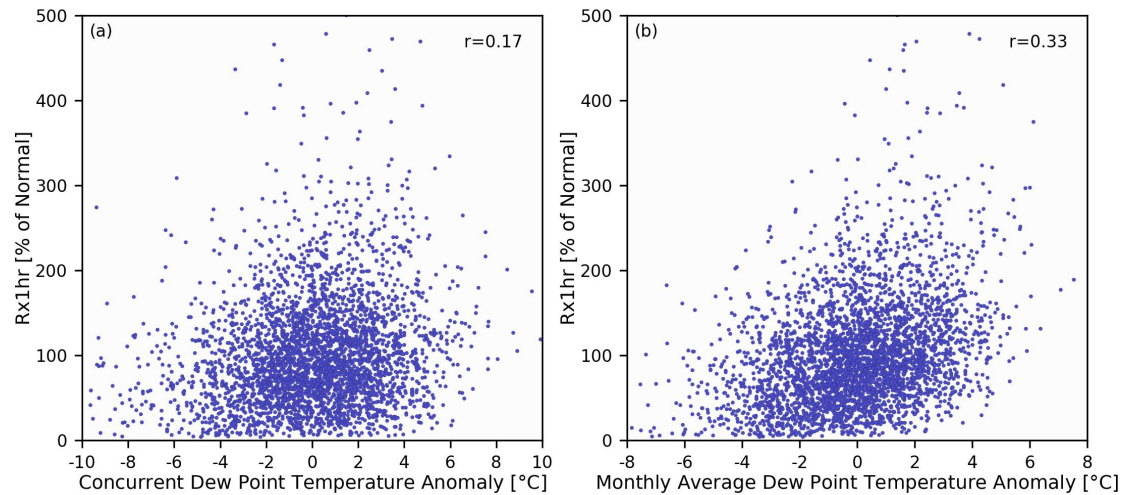


Figure 4. (a) normalized concurrent dew point temperature anomalies are plotted along with normalized Rx1hr precipitation. (b) normalized monthly average dew point temperature anomalies are plotted along with normalized Rx1hr precipitation. The correlation coefficients between the data points in each subplot are seen in the upper-right corners.

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 anomalies along with normalized Rx1hr precipitation. One can observe in the case of the UCRB, that there is a stronger statistical relationship (indicated by the difference in correlation coefficients between Fig. 4a and 4b) between the monthly average dew point temperature and Rx1hr than the concurrent hourly dew point temperature. For these reasons, we use the monthly average dew point temperatures as our predictor of Rx1hr and Rx1day precipitation.

3.2 Proposed Approach for Estimating P-T Scaling Rates

In the prior section, we pointed out a number of problems that can adversely influence our estimation of P-T scaling rates. For our modeling approach, we use data at the ~~station/month-resolution~~ (resolution of 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~~ stations and/or months, even at the same dew point temperature. By normalizing the data, we provide a more homogeneous framework for dealing with different types of precipitation (Berg and Haerter, 2013; Molnar et al., 2015). It is therefore advisable to normalize the data, following Eqs. 1 and 2, prior to estimating any P-T scaling rates.

Our approach uses an exponential regression model. Previously, we have shown some of the problems that can arise when using pooled, non-normalized data and which also lacks statistical independence. We therefore fit an exponential 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:

$$y = a \cdot b^x \quad , \quad (6)$$

where x contains monthly-averaged dew point temperature anomalies, from the array **DPT*** (or **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.

3.3 Predicting Extreme Precipitation

Consider a synthetic example where we were to fit the model to normalized data from Eqs. 1 and 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., 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.

Predictions of Rx1hr and Rx1day are performed using leave-one-year-out cross validation over the period 1951-2024. For the 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 to the normalized data. Through this procedure of implementing different modeling approaches with different data, we can determine which approach provides the best performance. With ~~regards~~regard to the fourth model, we also investigated fitting the exponential model to standardized anomalies (i.e., z-scores, or standard deviations from the mean). Using standardized data in the model produced very similar results to what we found using the normalized data (not shown). However, our proposed approach of using normalized data benefits from having less complexity and provides more easily interpretable scaling rates expressed as % per °C.

We can use the fourth model (i.e., fitting Eq. 6 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 (**DPT*** from Eq. 1) and Rx1hr precipitation (**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 (DPT^* from Eq. 1) and Rx1day precipitation (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.

Each of the last three models (models 2-4 listed above), use a range of spatial and temporal extents over which to fit the model and make predictions. This is 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 and the associated predictions. We use four spatial extents ~~-.The spatial extents are~~ where we pool data from: 1) ~~using only the data from the~~ only 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~~ all of the stations that fall within the entire region (in our case, the UCRB). Similarly, we use four temporal extents ~~-.The temporal extents are~~ where we pool data from: 1) ~~the data solely from~~ solely 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.

In Figure 5, we can observe the process of model fitting to non-normalized versus normalized data. In these cases from Fig. 5, we use all of the stations in the UCRB (i.e., spatial extent number 4), and a 3-month temporal window (i.e., temporal extent number 2). 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 (Dec-Jan-Feb) and summer (Jun-Jul-Aug) seasons. The bottom row shows the 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 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 as non-normalized values of Rx1hr precipitation (i.e., mm per hr). These modeled values are subsequently normalized using Eq. 2 prior to model evaluation.

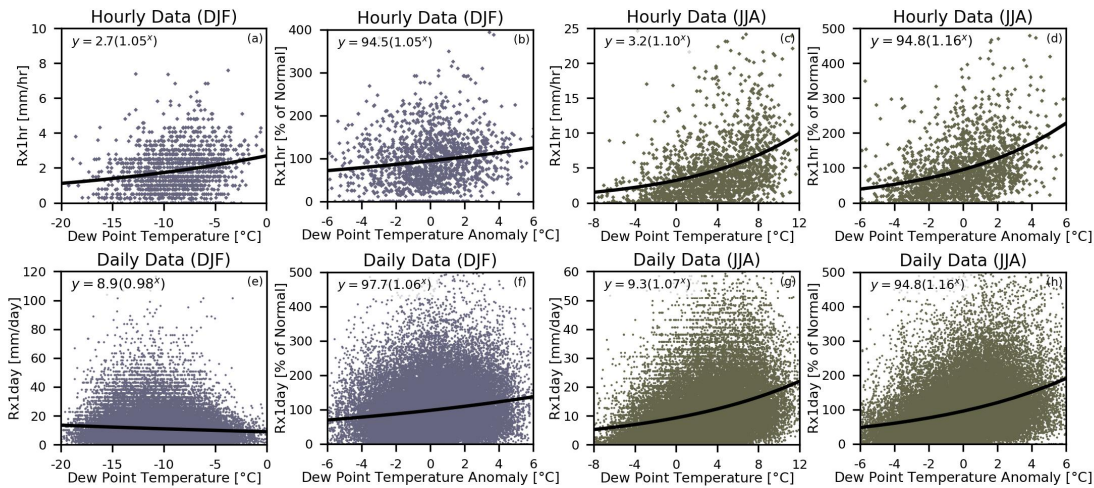


Figure 5. Comparing the model fit using non-normalized versus normalized data. The data points correspond to all of the stations in the UCRB and using a 3-month window corresponding to the winter and summer seasons. (a) plots all of non-normalized values of average monthly dew point temperature plotted 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.

300 4 Results

4.1 Evaluating Model Performance

The underlying empirical-statistical relationship between the normalized average monthly dew point temperature and normalized Rx1hr is plotted for all ~~station/stations and all~~ 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
305 normalized data model with a spatial extent of 100 km and a 3-month temporal window. While individual predictions of 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 Rx1hr or Rx1day precipitation given specified average monthly dew point temperature anomalies ranging
310 between -3°C below normal to $+3^{\circ}\text{C}$ above normal. At each dew point temperature anomaly, we compute the mean extreme precipitation anomaly centered about that dew point ~~temperauture-temperature~~ anomaly. We use a 1°C window, with 1°C increments, and find the modeled and observed extreme precipitation distributions corresponding to dew point temperature anomalies -3°C , -2°C , -1°C , 0°C , $+1^{\circ}\text{C}$, $+2^{\circ}\text{C}$, and $+3^{\circ}\text{C}$. Model performance evaluates how well the predicted mean shifts in the normalized extreme precipitation align with the observed mean shifts. In Fig. 6c, we can look ~~look~~ more specifically at the
315 Rx1hr anomalies for each of these dew point temperature anomalies 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, along with the means of those distributions, are shown in Fig. 6d. 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~~ We then perform the same procedure for ~~the all~~ 12 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. 6e. 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. Using Eq. 5, we find our SS_{RMSE} is 0.71. The same procedure is performed separately for Rx1hr and Rx1day using predictions from the 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.

4.2 Model Performance

In Figure 7a, we show the range of model skills, SS_{RMSE} , using leave-one-year-out cross validation for the 16 possible iterations of spatial-temporal extents using the C-C model, the non-normalized model, 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 best obtained skill score is 0.45 for the C-C model, 0.66 for the non-normalized model, and 0.71 for the normalized model. The dotted black line in Fig. 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 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 predicted values. ~~By randomly selecting all of the station/months in a calibration year~~ Through this process, 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 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 ($p < 0.01$). 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 model, 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 °C). 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

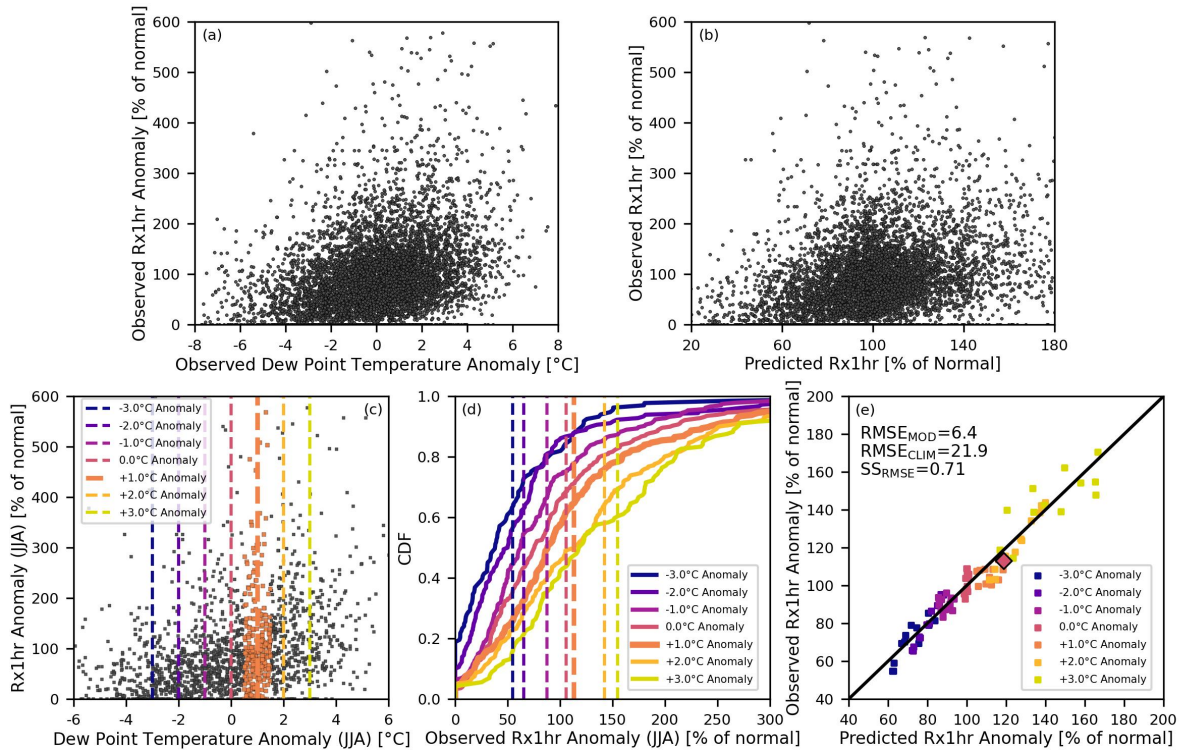


Figure 6. (a) plots the relationship between the normalized average monthly dew point temperature (our predictor) and normalized Rx1hr (our predictand) for all [station/stations and all](#) months. (b) plots the individual cross-validated predictions of Rx1hr versus observed Rx1hr using the normalized 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).

350 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 entire UCRB region and using all of the data throughout the year (i.e., the bottom-right grid cell in Fig. 7c), the 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 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

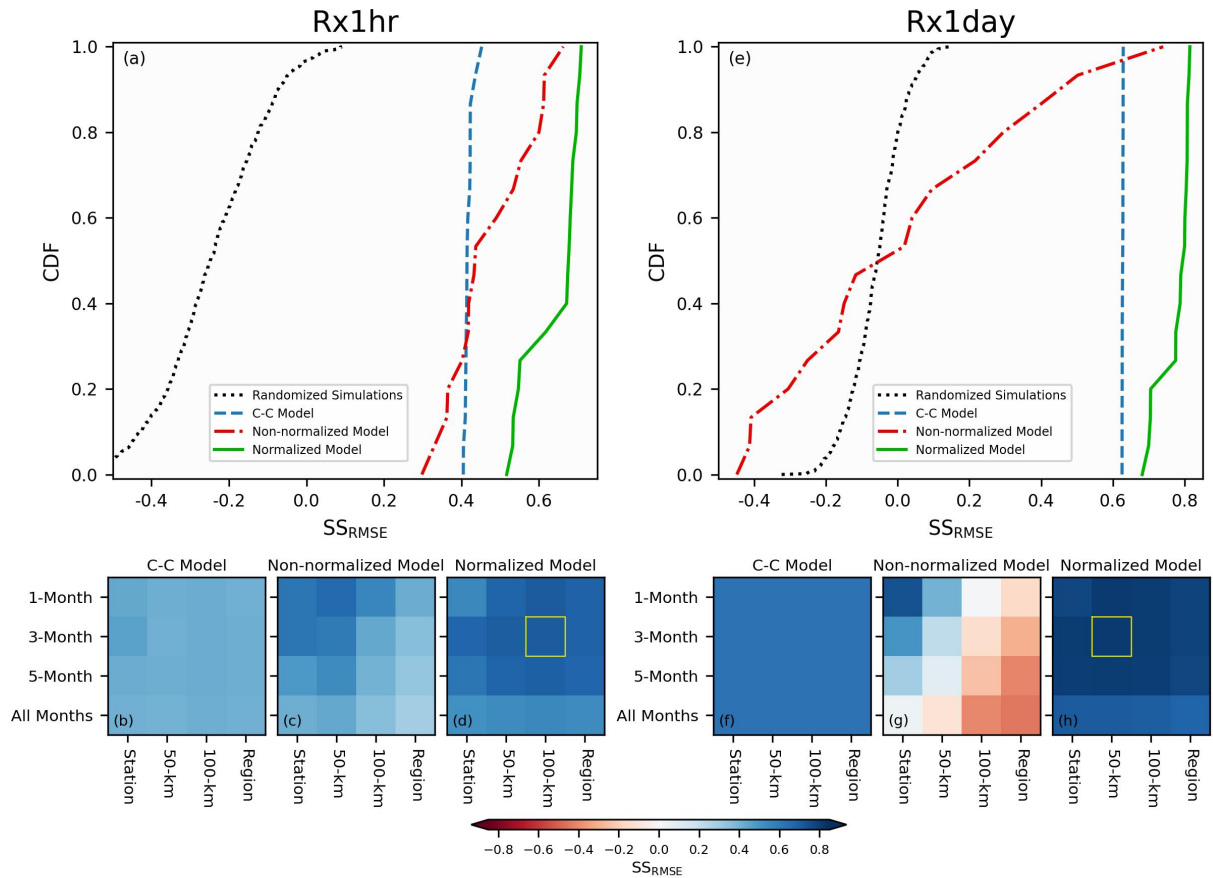


Figure 7. (a) plots the skills as empirical CDFs of the leave-one-year-out cross-validated predictions of Rx1hr for the three different models: C-C, non-normalized, and normalized. The CDFs are comprised of the skills from the different possible iterations of the spatial and temporal extents. Additionally, 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.

355 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
 360 (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 ($p < 0.01$). Figs. 7e-h show the same as Figs. 7a-d except evaluating the model performance in predicting

Rx1day precipitation. Generally, the 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
365 degrade as any amount of data 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 than 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). Please refer again to Figs. 5e-5h, to observe the impact that controlling for climatological differences in space and time
370 can have on model fitting. 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 ($p < 0.01$).

Figure 8 shows the same results as Figure 7, except we now predict Rx1day using two-fold cross validation. In this case of the 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 0.76 for the normalized model. These values are quite similar to what we obtained and plotted Figs. 7e-h for the leave-one-
375 year-out 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 underlying data. However, the results from Fig. 8 show that is not the case, and that the method can successfully produce skillful predictions over multi-decadal lengths of time.

380 4.3 Estimated Scaling Rates

With ~~regards~~ regard 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
385 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 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 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., $< 7\%$ per $^{\circ}\text{C}$). Additionally, the wintertime Rx1day is found to scale more strongly, on average, than Rx1hr. On the other hand, the median scaling rates in the
390 summer months exhibit super C-C ($> 7\%$ per $^{\circ}\text{C}$). 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 rate. The median P-T scaling rates for Rx1hr, as a function of month, range between 4.6% per $^{\circ}\text{C}$ for January ~~to~~ and 17.0% per $^{\circ}\text{C}$ for June. While for Rx1day, the scaling rates range between 5.2% per $^{\circ}\text{C}$ in December to 12.9% per $^{\circ}\text{C}$ in June.

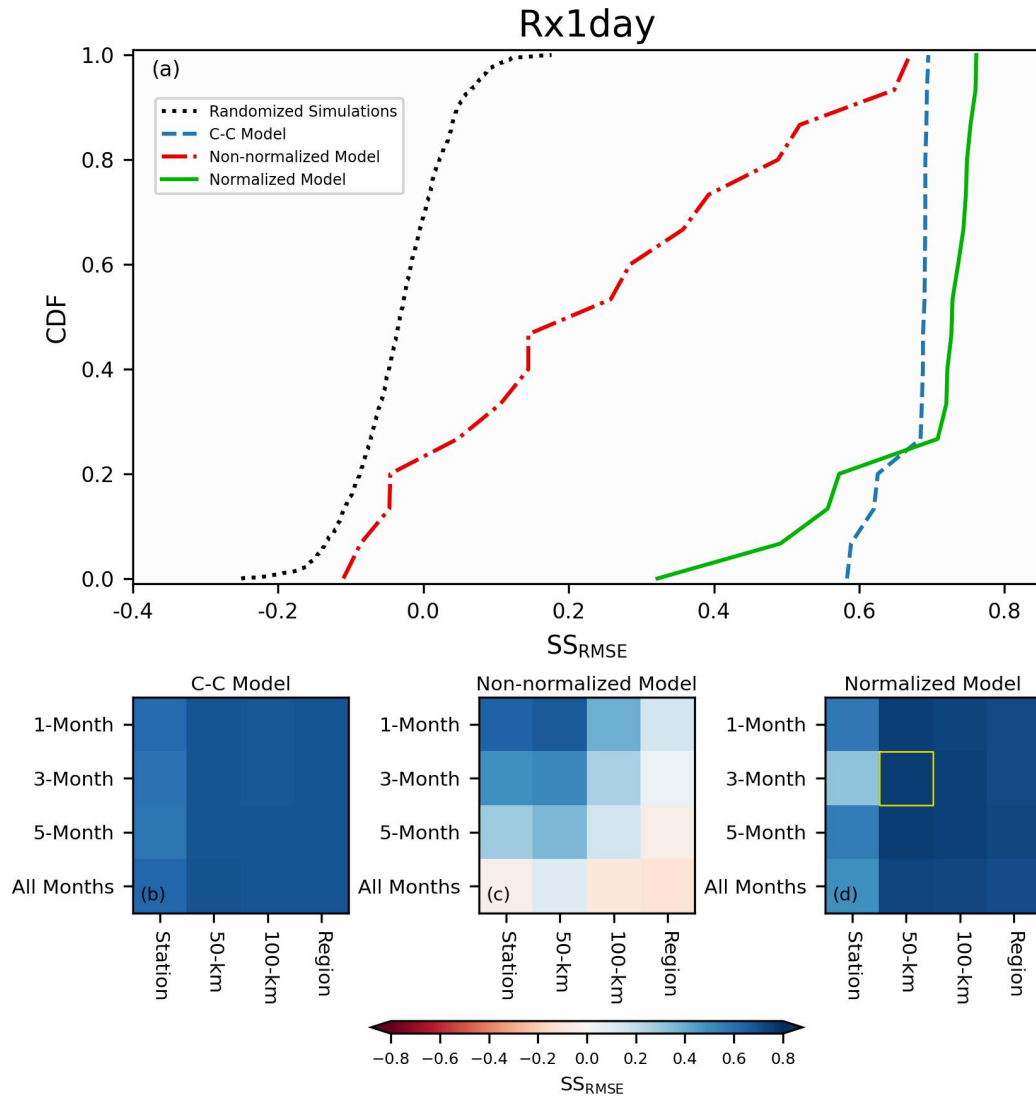


Figure 8. (a) plots the skills as empirical CDFs of the two-fold cross-validated predictions of Rx1day for the three different models: C-C, non-normalized, and normalized. The CDFs are comprised of the skills from the different possible iterations of the spatial and temporal extents. Additionally, the 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.

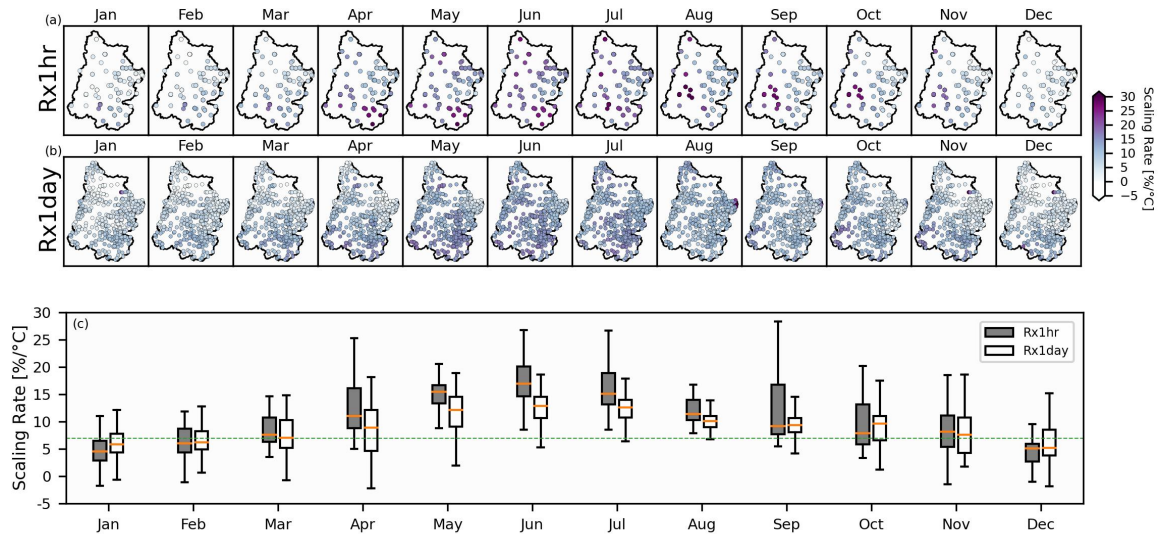


Figure 9. The top row of (a) plots the scaling rates of Rx1hr at each station for each month of the year. (b) same as (a) except for Rx1day. (c) uses a boxplot to show the distribution of the scaling rates as a function of Rx1hr or Rx1day along with the month of the year. In (c), each box, median, and 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).

5 Conclusions

395 Using point-scale station data from the Upper Colorado River Basin, this study begins by illustrating some common challenges
 in estimating P-T scaling rates. Our aim, herein, has not been to provide a comprehensive overview of every existing method-
 ology 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 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
 400 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) the statistical independence of the data 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 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,
 405 we make ~~cross-validated~~ cross-validated 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
410 Rx1hr and Rx1day than either relying on climatology or the C-C scaling rate. ~~Furthermore~~Furthermore, 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, we find
optimal performance using normalized data 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
415 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 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 of the non-normalized
model is found without implementing any data pooling. After normalizing the data, however, we can more effectively pool
420 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,~~paper's focus has been to illustrate the value of normalizing
data in an effort to more accurately estimate P-T scaling rates. However, future research can focus on investigating whether
425 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. Lastly, while our proposed method is capable of modeling either positive and negative scaling
rates, it is not capable of handling cases where the normalized data experiences a transition between positive to negative scaling,
430 or vice-versa.

Using our best performing model, which uses normalized data pooled from 50 km or 100 km (for Rx1hr and Rx1day,
respectively) and a 3-month window, we finally show the scaling rates at each station for each month of the year. The scaling
rates associated with daily extreme ~~preecipitation~~precipitation, 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
435 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 fitting an exponential model to normalized data. By applying this methodology to other regions of the world, we can gain
440 a more detailed and comprehensive understanding of how P-T scaling rates vary across space and time. This information is
essential in our efforts to improve our preparedness regarding extreme precipitation.

6 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 anomalies versus Rx1hr precipitation anomalies) and daily (i.e., average monthly ERA5 dew point temperature anomalies versus Rx1day precipitation anomalies) datasets. This is shown in Figure A2. 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.

(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.

475 ~~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.

480 *Acknowledgements.* The authors ~~want to thank the~~ acknowledge the financial support of 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, e2021GL095414, <https://doi.org/10.1029/2021GL095414>, 2022.
- 485 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.
- 490 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, e2020GL090317, <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, e2022GL099138, <https://doi.org/10.1029/2022GL099138>, 2022.
- 495 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.
- 500 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.
- 505 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, e2024GL110424, <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.
- 510 Donat, M. G., Delgado-Torres, C., Luca, P. D., Mahmood, R., Ortega, P., and Doblus-Reyes, F. J.: How credibly do CMIP6 simulations capture historical mean and extreme precipitation changes?, *Geophysical Research Letters*, 50, e2022GL102466, <https://doi.org/10.1029/2022GL102466>, 2023.
- Drobinski, P., Silva, N. D., Panthou, G., Bastin, S., Muller, C., Ahrens, B., Borga, M., Conte, D., Fossier, G., Giorgi, F., Güttler, I., Kotroni, V., Li, L., Morin, E., Öno, B., Quintana-Segui, P., Romera, R., and Torma, C. Z.: Scaling precipitation extremes with temperature 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, e2024JD040901, <https://doi.org/10.1029/2024JD040901>, 2025.
- 520
- 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.
- 525
- 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.
- 530
- Harp, R. D. and Horton, D. E.: Observed changes in daily precipitation intensity in the United States, *Geophysical Research Letters*, 49, e2022GL099955, <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.
- 535
- 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, e2024GL112237, <https://doi.org/10.1029/2024GL112237>, 2025.
- 540
- 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, L22805, <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.
- 545
- 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, 025208, <https://doi.org/10.1088/1748-9326/5/2/025208>, 2010.
- 550
- 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, 054022, <https://doi.org/10.1088/1748-9326/acd01a>, 2023.
- 555

- Marra, F., Koukoulou, 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.
- 560 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.
- 565 Myhre, G., Alterskjær, K., Stjern, C. W., Hodnebrog, M., 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, 16063, <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, e2023GL104606, <https://doi.org/10.1029/2023GL104606>, 2023.
- 570 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, e2021GL097100, <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.
- 575 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, 17398, <https://doi.org/10.1038/s41598-024-65618-9>, 2024.
- 580 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.
- 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.
- 585 Tabari, H.: Climate change impact on flood and extreme precipitation increases with water availability, *Scientific Reports*, 10, 13768, <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, 034016, <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.
- 590 Utsumi, N., Seto, S., Kanae, S., Maeda, E. E., and Oki, T.: Does higher surface temperature intensify extreme precipitation?, *Geophysical Research Letters*, 38, L16708, <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.
- 595 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
600 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., Guo, S., Gentine, P., He, S., Chen, J., and Liu, P.: Does the hook structure constrain future flood intensification under anthropogenic
605 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.