Figure 2.19. Trends in the number of EHE-days during the 1970-2010 period for different EHE types and intra-seasonal focus. All three types of EHEs are given, with three seasonal focuses and the symbology is the same as figure 2.4.

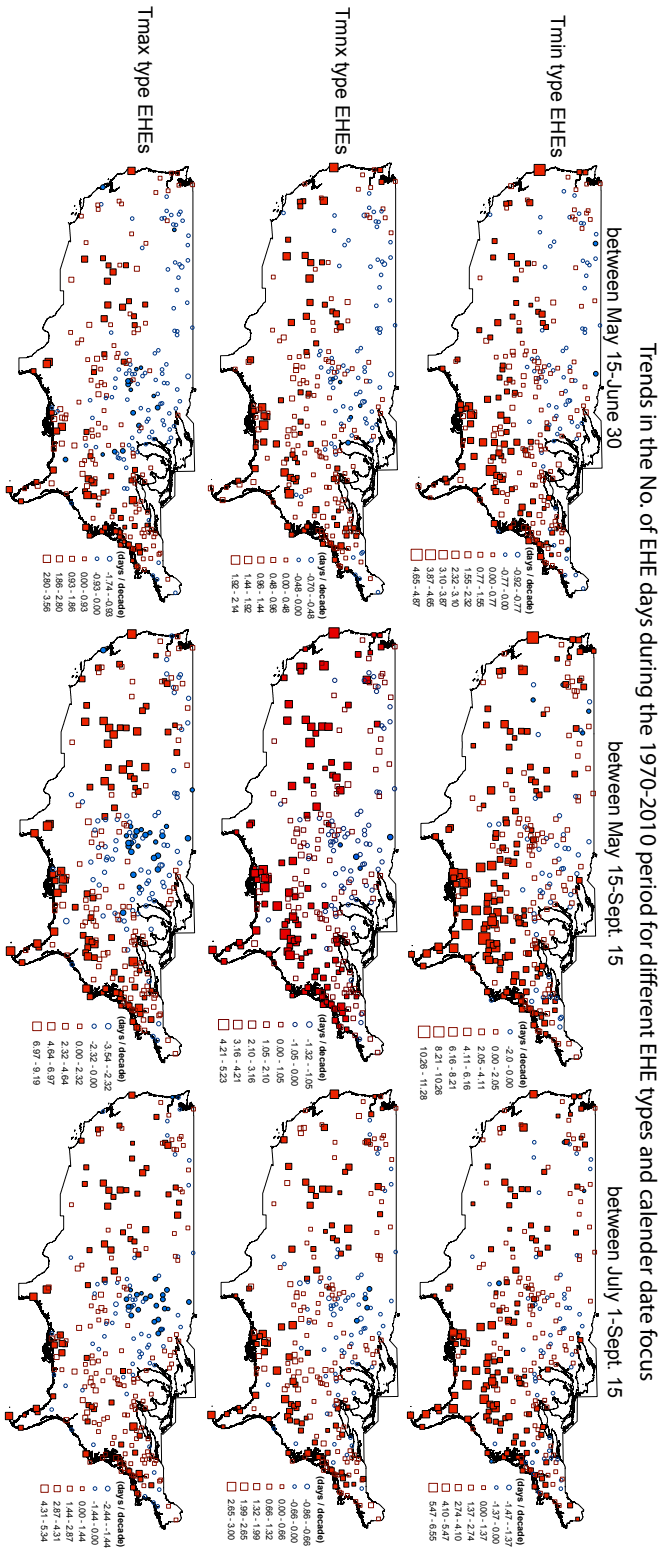


Figure 2.20. Trends in the number of EHE-days during the 1930-2010 period for different EHE types and intra-seasonal focus. All three types of EHEs are given, with three seasonal focuses and the symbology is the same as figure 2.4.

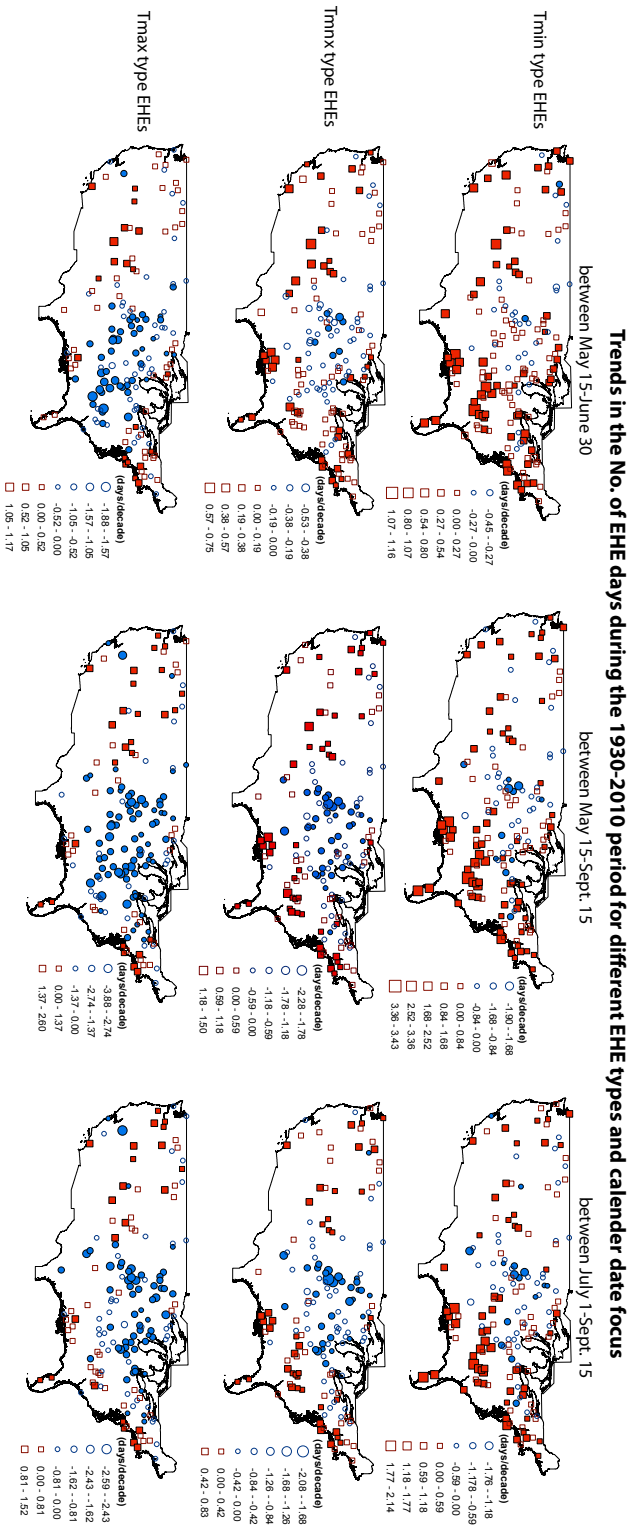


Figure 2.21. Trends in the number of Tmin EHE days and summer average daily minimum temperatures. All three time periods are provided and the symbology follows that of figure 2.4.

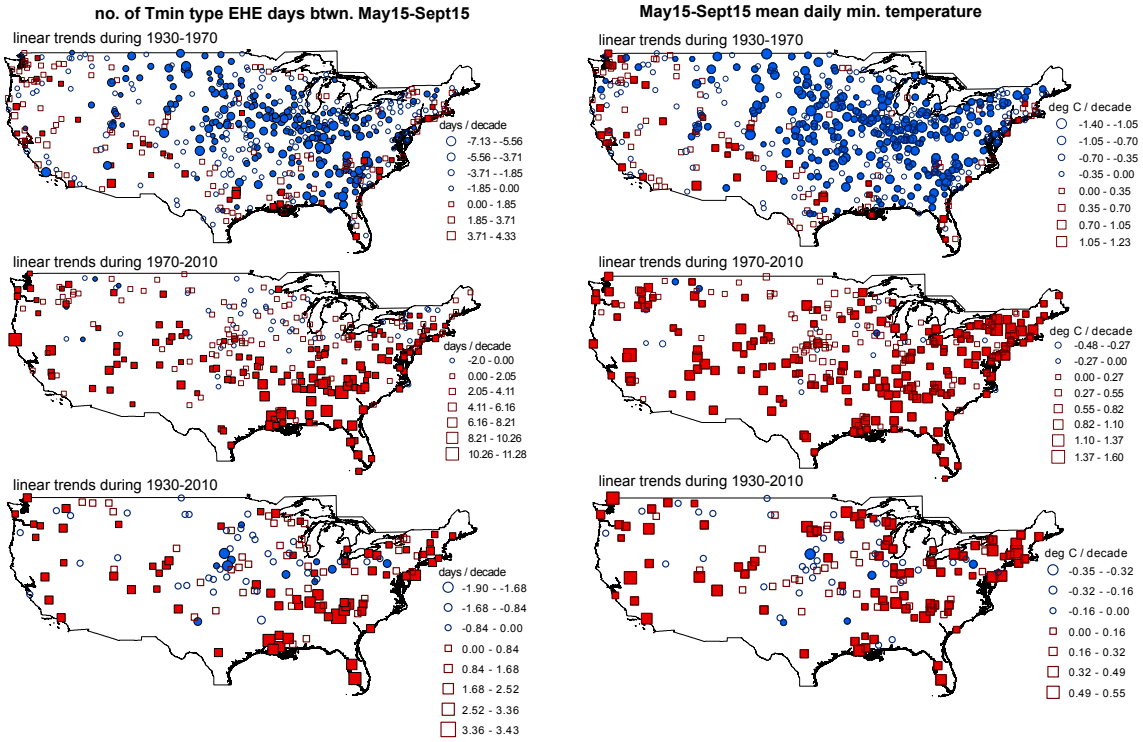


Figure 2.22. Trends in the number of Tmax EHE days and summer average daily maximum temperatures. All three time periods are provided and the symbology follows that of figure 2.4.

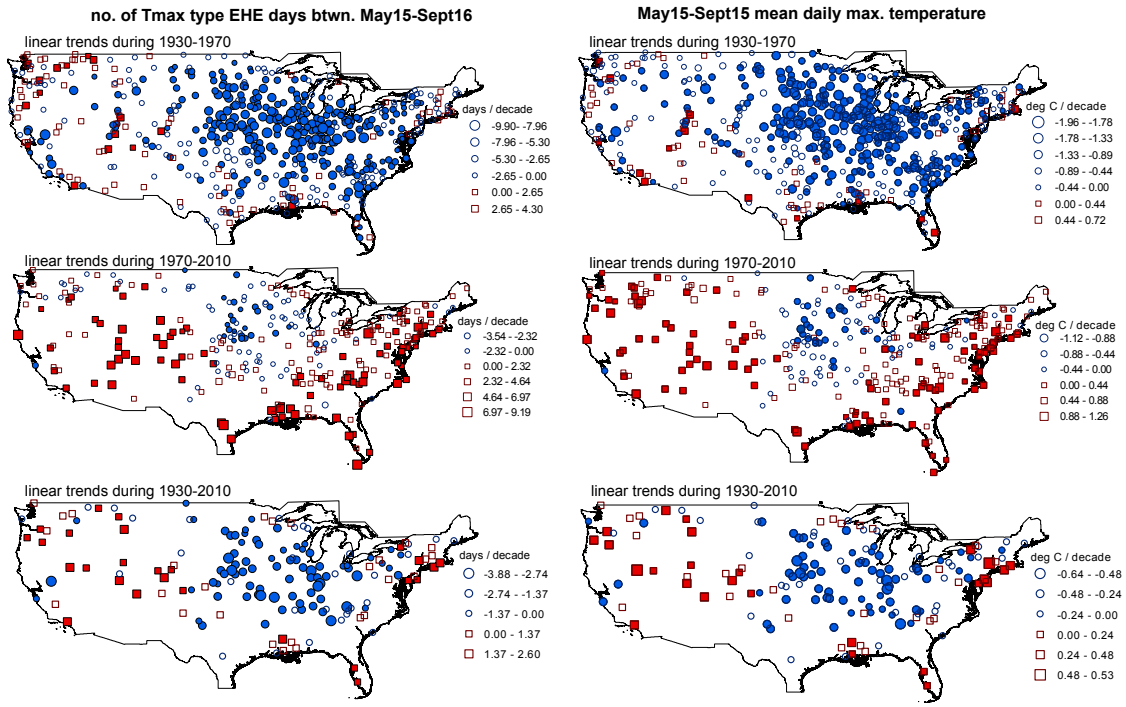


Figure 2.23. Trends in the number of Tmnx EHE days and summer average daily mean temperatures. All three time periods are provided and the symbology follows that of figure 2.4.

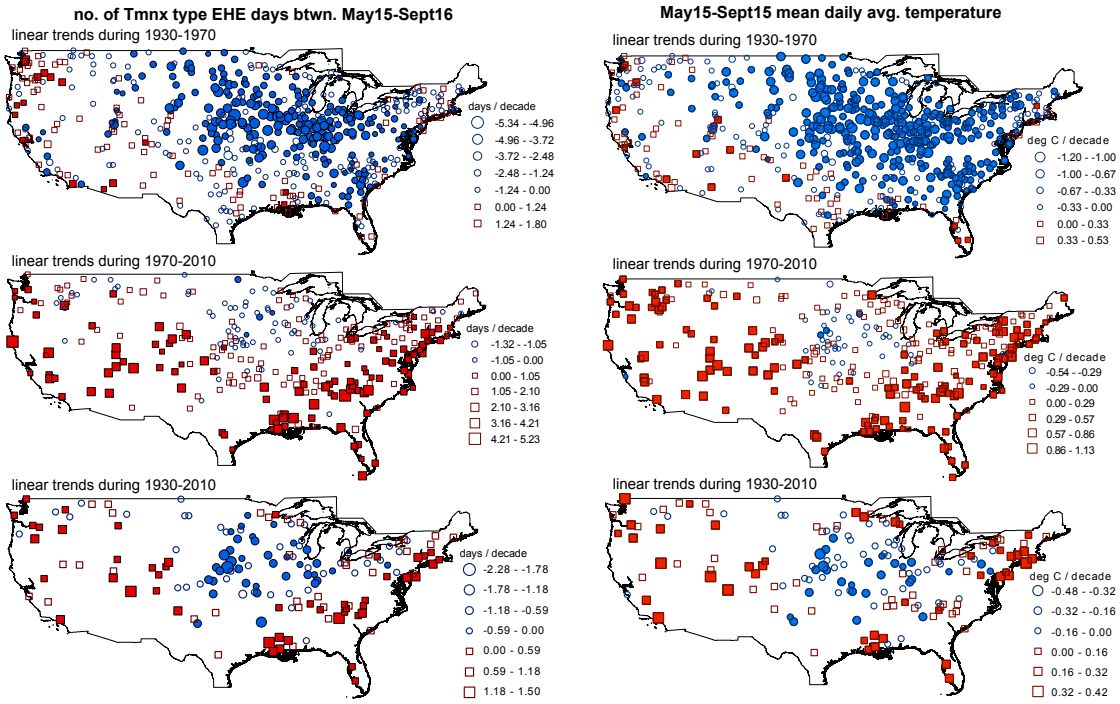


Figure 2.24. Trends of the EHE days over the CONUS over the 1950-2006 period. Trends shown at the 161 stations that span the 1930-2010 period. All three EHE types are shown and the symbology follows that of figure 2.4.

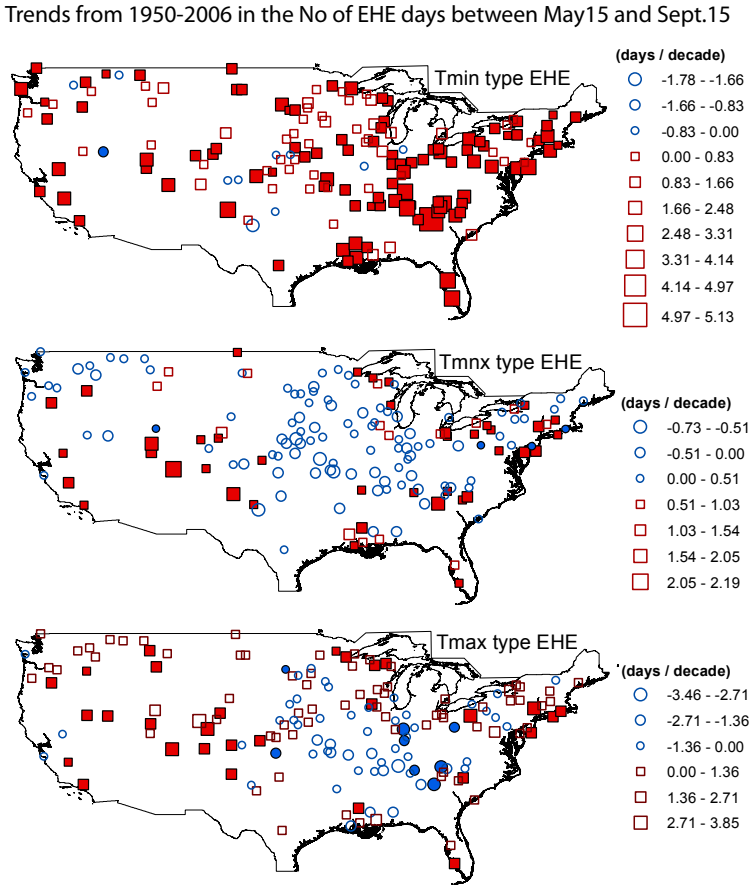


Figure 2.25. Trends of the EHE days over the CONUS over the 1960-1996 period. Trends shown at the 161 stations that span the 1930-2010 period. All three EHE types are shown and the symbology follows that of figure 2.4.

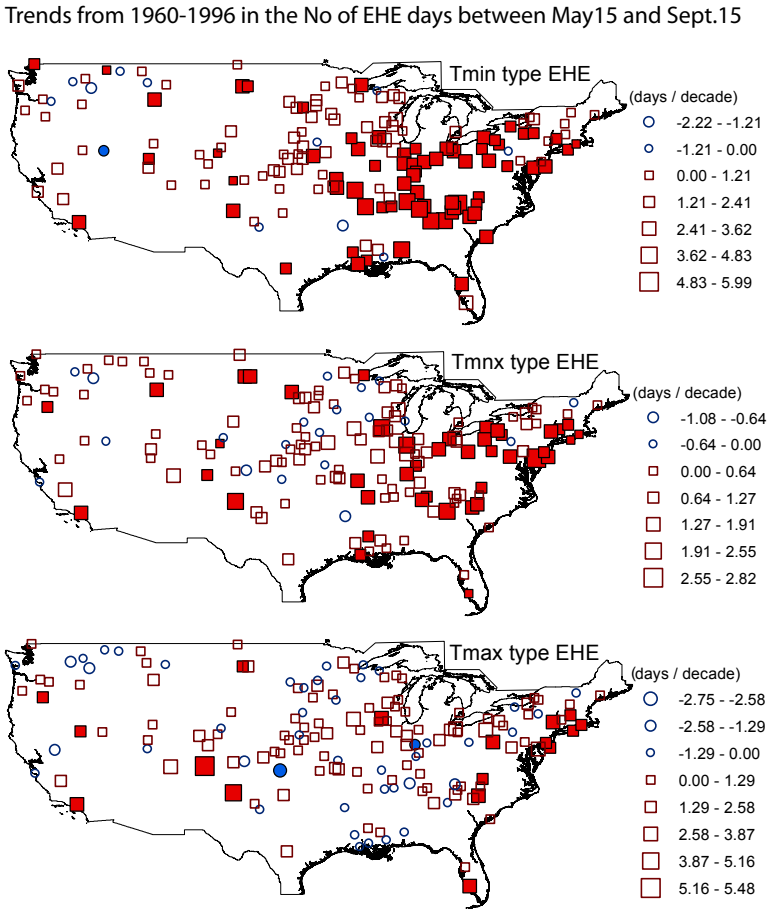


Figure 2.26. Trends of the EHE days over the CONUS over the 1950-2004 period. Trends shown at the 161 stations that span the 1930-2010 period. All three EHE types are shown and the symbology follows that of figure 2.4.

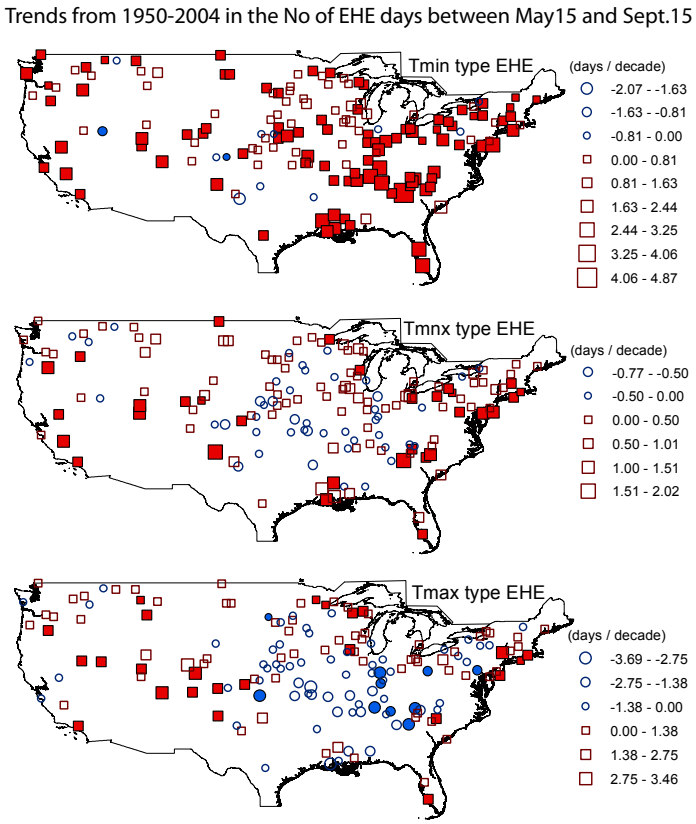
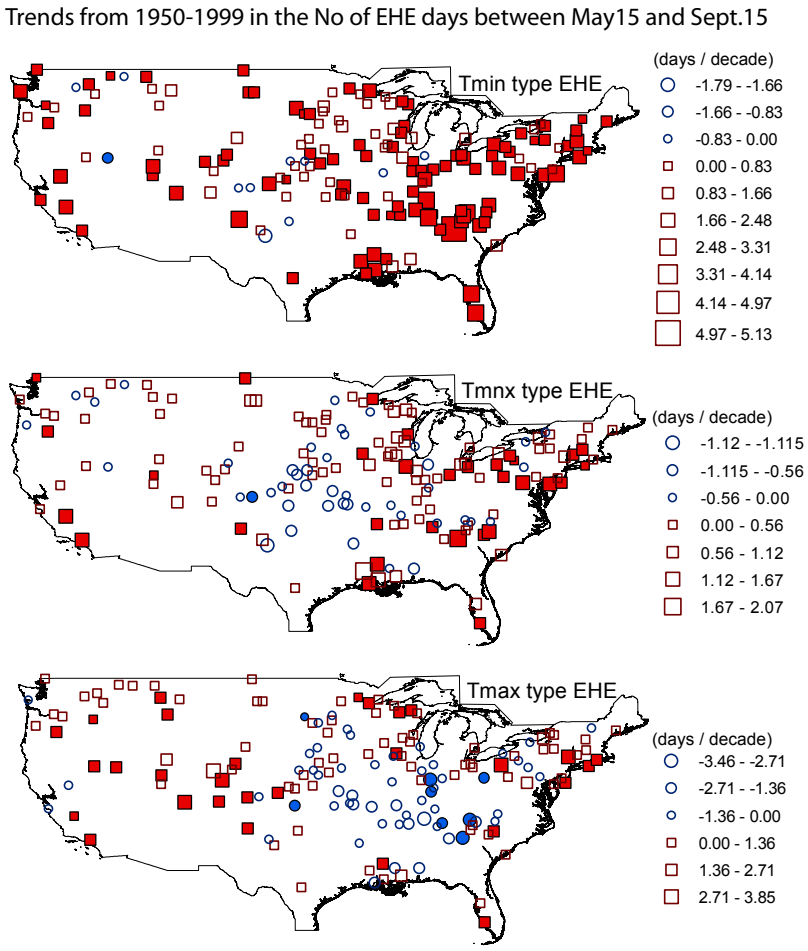


Figure 2.27. Trends of the EHE days over the CONUS over the 1950-2006 period. Trends shown at the 161 stations that span the 1930-1020 period. All three EHE types are shown and the symbology follows that of figure 2.4.



2.9. Tables

Table 2.1. Typical values of EHE characteristics. Listed are the sample means (standard deviations) of the of stations temporal means. Each EHE characterization metric, both periods and all three EHE types are listed.

1930-1970 (n=541)			
EHE characteristic	**Tmin EHE**	**Tmax EHE**	**Tmnx EHE**
No. of EHEs per year	2.1 (0.3)	2.5 (0.4)	1.0 (0.4)
No. of EHE days	7.5 (1.7)	11.0 (2.3)	3.7 (1.7)
No. of early EHE days	2.5 (1.4)	3.9 (2.7)	1.3 (1.0)
Mean EHE duration	2.7 (0.4)	3.5 (0.6)	1.9 (0.7)
Mean EHE intensity	10.0 (1.6)	12.9 (2.1)	15.3 (5.7)
Sum of EHE intensities	27.6 (6.6)	41.9 (9.7)	30.9 (15.3)
1970-2010 (n=295)			
EHE characteristic	**Tmin EHE**	**Tmax EHE**	**Tmnx EHE**
No. of EHEs per year	2.3 (0.3)	2.5 (0.4)	1.0 (0.3)
No. of EHE days	8.1 (1.7)	9.8 (1.8)	3.3 (1.1)
No. of early EHE days	3.1 (1.7)	3.6 (2.0)	1.4 (0.9)
Mean EHE duration	3.0 (0.4)	3.2 (0.5)	1.9 (0.5)
Mean EHE intensity	10.6 (1.4)	11.5 (1.7)	15.0 (4.0)
Sum of EHE intensities	29.6 (6.1)	36.0 (6.7)	28.3 (8.9)

Table 2.2. CONUS spatial average of the decadal trends in EHE characteristics. Values are arranged by EHE characteristic and type of EHE. In each cell, listed from left to right, are the values for 1930-1970, 1970-2010 and the 1930-2010 period.

EHE characteristics	**Tmin EHE**	**Tmax EHE**	**Tmnx EHE**
Number of EHEs	-0.24, 0.43, 0.10	-0.38, 0.28, -0.05	-0.19, 0.23, 0.02
Number of EHE days	-1.02, 2.13, 0.51	-2.23, 1.56, -0.42	-0.84, 0.96, 0.06
Mean EHE duration	-0.13, 0.34, 0.12	-0.30, 0.20, -0.08	-0.23, 0.29, 0.04
Mean EHE intensity	-0.63, 1.36, 0.38	-1.57, 0.99, -0.41	-2.37, 2.50, 0.24
Sum EHE intensity	-4.33, 8.09, 1.71	-10.22, 6.46, -2.02	-8.07, 7.91, 0.14

Table 2.3. Regional average decadal trends in the number of EHE days per summer. The rows correspond to the north western region (NW), south western region (SW), north central region (NC), south central region (SC), north eastern region (NE) and south eastern region (SE); and the columns correspond to the three EHE types.

1930-1970

region(stations)	Tmin	Tmax	Tmnx
NW (99)	-0.84	-0.82	-0.44
SW (58)	-0.04	-1.03	-0.26
NC (129)	-1.9	-4.39	-1.92
SC (106)	-1.08	-3.3	-1.17
NE (60)	-0.56	-1.61	-0.48
SE (89)	-1.75	-3.26	-1.12

1970-2010

region(stations)	Tmin	Tmax	Tmnx
NW (50)	1.27	1.15	0.66
SW (25)	2.4	2.82	1.6
NC (70)	0.34	-1.11	-0.22
SC (53)	2.91	1.59	1.16
NE (49)	1.08	1.52	0.71
SE (48)	3.65	2.27	1.36

1930-2010

region(stations)	Tmin	Tmax	Tmnx
NW (27)	0.33	0.38	0.16
SW (14)	0.53	0.18	0.31
NC (44)	-0.06	-1.48	-0.47
SC (29)	0.58	-1.24	-0.09
NE (25)	0.47	-0.16	0.13
SE (22)	1.23	-0.79	0.31

Table 2.4. Results of the Student's t-tests of mean between regions. Two sided, two sample Students t-tests were employed, and here 1 indicates significantly different trends and a 0 indicates the means were unable to be distinguished at the 99% confidence level.

1930-1970

Tmin type EHE trends

region(stations)	NW	SW	NC	SC	NE	SE
NW (99)	0	X	X	X	X	X
SW (58)	0	0	X	X	X	X
NC (129)	1	1	0	X	X	X
SC (106)	0	1	1	0	X	X
NE (60)	0	0	1	0	0	X
SE (89)	1	1	0	0	1	0

Tmax type EHE trends

	NW	SW	NC	SC	NE	SE
NW	0	X	X	X	X	X
SW	0	0	X	X	X	X
NC	1	1	0	X	X	X
SC	1	1	1	0	X	X
NE	0	0	1	1	0	X
SE	1	1	1	0	1	0

Tmnx type EHE trends

	NW	SW	NC	SC	NE	SE
NW	0	X	X	X	X	X
SW	0	0	X	X	X	X
NC	1	1	0	X	X	X
SC	1	1	1	0	X	X
NE	0	0	1	1	0	X
SE	1	1	1	0	1	0

1970-2010

Tmin type EHE trends

region(stations)	NW	SW	NC	SC	NE	SE
NW (50)	0	X	X	X	X	X
SW (25)	0	0	X	X	X	X
NC (70)	1	1	0	X	X	X
SC (53)	1	0	1	0	X	X
NE (49)	0	1	1	1	0	X
SE (48)	1	0	1	0	1	0

Tmax type EHE trends

	NW	SW	NC	SC	NE	SE
NW	0	X	X	X	X	X
SW	1	0	X	X	X	X
NC	1	1	0	X	X	X
SC	0	0	1	0	X	X
NE	0	1	1	0	0	X
SE	1	0	1	0	0	0

Tmnx type EHE trends

	NW	SW	NC	SC	NE	SE
NW	0	X	X	X	X	X
SW	1	0	X	X	X	X
NC	1	1	0	X	X	X
SC	0	0	1	0	X	X
NE	0	1	1	1	0	X
SE	1	0	1	0	1	0

1930-2010

Tmin type EHE trends

region(stations)	NW	SW	NC	SC	NE	SE
NW (27)	0	X	X	X	X	X
SW (14)	0	0	X	X	X	X
NC (44)	1	1	0	X	X	X
SC (29)	0	0	1	0	X	X
NE (25)	0	0	1	0	0	X
SE (22)	1	0	1	0	1	0

Tmax type EHE trends

	NW	SW	NC	SC	NE	SE
NW	0	X	X	X	X	X
SW	0	0	X	X	X	X
NC	1	1	0	X	X	X
SC	1	1	0	0	X	X
NE	0	0	1	1	0	X
SE	1	0	0	0	0	0

Tmnx type EHE trends

	NW	SW	NC	SC	NE	SE
NW	0	X	X	X	X	X
SW	0	0	X	X	X	X
NC	1	1	0	X	X	X
SC	0	0	0	0	X	X
NE	0	0	1	0	0	X
SE	0	0	1	0	0	0

Table 2.5. Pearson's correlation coefficients and Student's t-test restuls between trends of different EHE types. Displayed are three values, representing the three time periods (1930-1970, 1970-2010, 1930-2010) for each metric and between all three different EHE types. Regular font values indicate the means are statistically not equal at the 0.10 significance level, and bold font values signify a failure to prove they are not equal at that significance level.

EHE characteristic	**Tmin/Tmax**	**Tmin/Tmnx**	**Tmax/Tmnx**
No. of EHEs per year	0.36, 0.38, 0.35	**0.75**, 0.70, 0.79	0.69, **0.78**, 0.69
No. of EHE days	0.42, 0.42, 0.32	**0.78**, 0.76, 0.82	0.76, **0.78**, 0.69
No. of early EHE days	0.27, 0.36, 0.19	0.67, 0.62, 0.72	0.48, **0.69**, 0.55
Mean EHE duration	0.21, 0.36, 0.34	0.72**, 0.72**, 0.82	0.59, 0.67, 0.69
Mean EHE intensity	0.26, 0.32, 0.35	0.72, 0.69, 0.82	0.65, 0.67, 0.71
Sum EHE intensity	0.44, 0.39, 0.34	0.77, **0.73**, 0.82	0.79, **0.78**, 0.70

Table 2.6. The CONUS spatial average decadal trends in the number of EHE days differentiated calendar date, or intra-seasonal, focus. Three values are listed per time period and EHE type: between 15 May and 30 June, between 1 July and 15 September and between 15 May and 15 September, respectively. Shaded boxes indicate the sample mean trend from the 15 May to 30 June values is statistically different from the 38% of the 15 May to 15 September value it should be.

	Tmin EHE type	**Tmax EHE type**	**Tmnx EHE type**
1930-1970	-0.07, -0.87, -1.02	-0.41, -1.61, -2.23	-0.07, -0.74, -0.84
1970-2010	0.71, 1.35, 2.13	0.51, 1.03, 1.56	0.37, 0.58, 0.96
1930-2010	0.25, 0.25, 0.51	-0.10, -0.25, -0.42	0.07, -0.01, 0.06

Table 2.7. Spatial correlation coefficients of trends in the number EHE days between different intra-seasonal focus. The first value given, per EHE type and time period, is the coefficient between early season and whole season. The second value is between early and late season.

	Tmin EHE	**Tmax EHE**	**Tmnx EHE**
1930-1970	0.82, 0.63	0.89, 0.67	0.77, 0.55
1970-2010	0.87, 0.70	0.88, 0.69	0.86, 0.60
1930-2010	0.90, 0.78	0.87, 0.57	0.87, 0.72

Table 2.8. Characterization and comparison of sign and significance of trends in EHEs and seasonal average temperatures. The top four rows in the table provide summer average temperature trends and the bottom four the trends in EHE-days; columns corresponding to different extreme heat manifestations. The first value listed for each entry is the percent within stations of the sample with negative trends, the second is the percent of stations within the sample with positive trends. The values in parenthesis behind those values are the percentage of stations within the sample with corresponding significant trends.

	Daily min.	**Daily max.**	**Daily mean**
1930-1970	79(49), 21(7)	87(58), 13(3)	85(60), 15(3)
1970-2010	6(1), 94(64)	32(10), 68(32)	16(1), 84(45)
1930-2010	25(3), 75(55)	67(43), 33(16)	47(16), 53(26)
	Tmin EHE-days	**Tmax EHE-days**	**Tmnx EHE-days**
1930-1970	72(34), 28(6)	82(47), 18(2)	78(41), 22(3)
1970-2010	21(2), 79(41)	39(8), 64(24)	27(2), 73(36)
1930-2010	34(7), 66(39)	68(45), 32(14)	49(22), 51(26)

Table 2.9. Comparison of the average EHE and summer average Pearson's correlation coefficients (through time) at each station. From left to right are listed the sample-average correlation coefficients for the 1930-1970, 1970-2010 period and 1930-2010 periods and then the average (of those three periods sample-averages) standard deviation. These values are arranged vertically by seasonal average and horizontally by EHE type.

	Tmin type EHE	**Tmax type EHE**	**Tmnx type EHE**
Daily min.	0.68, 0.68, 0.68, 0.08	0.43, 0.42, 0.39, 0.17	0.50, 0.51, 0.50, 0.13
Daily max.	0.48, 0.50, 0.47, 0.17	0.80, 0.78, 0.80, 0.07	0.58, 0.58, 0.58, 0.21
Daily mean	0.63, 0.65, 0.64, 0.11	0.73, 0.70, 0.71, 0.09	0.62, 0.62, 0.62, 0.11

Table 2.10. Comparison of Pearson correlation coefficients between summer average temperature trends and the trends in the number of EHE days. The leftmost column indicates daily minimum (Tmin) temperature, daily maximum (Tmax) temperature and daily mean (Tavg) temperature trends. The topmost column represents EHE types. Each cells three values from left to right represent the 1930-1970 period, the 1970-2010 period and the 1930-2010 period, respectively.

	Tmin type EHE	**Tmax type EHE**	**Tmnx type EHE**
Daily min.	0.84, 0.68, 0.75	0.38, 0.27, 0.23	0.69, 0.55, 0.63
Daily max.	0.39, 0.27, 0.32	0.90, 0.84, 0.93	0.69, 0.64, 0.66
Daily mean	0.70, 0.55, 0.62	0.81, 0.78, 0.81	0.83, 0.77, 0.81

CHAPTER 3. A CHARACTERIZATION OF THE ACCURACY OF HIGH-RESOLUTION GRIDDED OBSERVATIONAL CLIMATE DATASETS FOR TREND ANALYSIS AND DOWNSCALING

Full citation of corresponding manuscript

Oswald E. M. and R. B. Rood: A characterization of the accuracy of high-resolution gridded observational climate datasets for trend analysis and downscaling. *Submitted to the International Journal of Climatology.*

Abstract

Gridded observational climate datasets with high spatial resolution and daily temporal resolution are currently some of the only tools publicly available for many areas of climate related research (e.g. frost days, extreme heat events). This study compared the representation of temperature and extreme heat climate indices of three such datasets with a more trusted climate dataset. That dataset is modern, well trusted and homogenized by a well-evaluated method; and was subsequently temporally downscaled using the daily version of the dataset. The high-resolution gridded observational climate datasets being evaluated were originally constructed to force hydrological models and did not address issues regarding non-climatic biases, but are now used in other fields including atmospheric science and global climate model output downscaling. These datasets were compared at the locations of the reference network monitoring sites and over the durations of the datasets being evaluated.

The results exhibited statistically significant differences in temporal averages and linear trends, for all three datasets. Explicit examinations at the continental, regional and small scales indicated that while the small scale dominates the spatial variability of the differences, the differences were also statistically significant at the regional and continental scales. Maps of the results confirmed such conclusions. While it is well known that erroneous discontinuities arise from non-climatic biases (e.g. urbanization, time of observation bias, instrument changes), proxies for them had weak relationships with the differences. Results indicated the differences were closest related to homogenization adjustments made after the “time of observation bias” adjustments. Weak relationships with proxies for discontinuities and homogenization adjustments suggested the gridding process convolves the discontinuities from multiple surrounding stations to create especially discontinuous time series at all grid points.

It was concluded that the next generation of high-resolution gridded climate observational dataset must have an amount of homogenization applied to the underlying data network. Future studies focusing on how to homogenize a spatially dense underlying network are needed. Acknowledgement of the uncertainty within downscaling products that use these datasets is essential moving forward.

3.1. Introduction

Currently, gridded datasets available for evaluation of daily-scale climate at high spatial resolutions are not in full agreement with more trusted climate datasets provided by the climate community. The lack of agreement stems from a larger disagreement about how to deal with inhomogeneities in climate data. Inhomogeneities arise from using data with discontinuities in the time series caused by non-climatic biases, and moreover additional inhomogeneities can arise from the gridding process. Unfortunately because the differences between non-homogenized high-resolution climate datasets and climate community standard datasets are so rare, it is unclear if and how non-homogenized high-resolution gridded datasets can be confidently used.

High-resolution climate datasets originated because small-scale process modeling over large spatial domains (e.g. hydrological modeling) requires high-resolution serially complete gridded datasets for meteorological inputs (Christensen et al. 2004; El-Sadek et al. 2011). These datasets have subsequently evolved to become some of the more commonly used tools to assess climate signals at high spatial resolution. Regional and small-scale climate signals are particularly important to decision makers, as they often make decisions on those spatial scales. Climate datasets that enable quantification of sub-monthly events (frost days, heat events, flooding) are important because those events are often very impactful. Furthermore, these gridded datasets are used not only for studying historical records, but also for downscaling climate model output forecasts.

The climate community approaches their datasets with a guarded methodology, as results are occasionally contested (e.g. Pielke Sr. et al. 2007; Fall et al. 2011). The gridded datasets developed by the applications communities address their dataset needs from a standpoint that estimates of the variables are better than no estimates. In general, these very high spatial resolution datasets are possible due to relaxed requirements for station inclusion and infilling (of missing values), as well as from use of highly resolved secondary information (e.g. elevation) to enhance spatial interpolation performance. They also employ statistical models to estimate a suite of secondary variables (e.g. humidity, incoming radiation).

Disconnects between these high-resolution datasets and the more trusted climate community datasets have been previously acknowledged but poorly quantified. Early evaluations of similar gridded climate datasets (e.g. Moberg and Alexandersson 1997) concluded these datasets averaged over large spatial domains would be accurate due to the cancelling out of conflicting inhomogeneities. A 2005 study by Hamlet and Lettenmaier, established a method to effectively adjust the grid cell values in the Maurer et al. (2002) dataset to compensate for the discontinuities the underlying Co-Op data contained. Subsequently numerous studies (Hamlet et al. 2005; Bonfils et al. 2008; Lobell et al. 2008) have mentioned the Maurer et al. (2002) dataset was less than optimal for use and decided to either use other datasets or use the Hamlet and Lettenmaier (2005) method to construct similar, but homogenized, gridded products.

Yet to date no manuscripts by the high-resolution datasets creators have been located, which discuss the topic of homogenization of the datasets. Unpublished masters-thesis work by Scully (2010) compared monthly and annual daily maximum and minimum temperatures between the PRISM (Daly et al. 2008) and DAYMET (Thornton et al. 1997) datasets. They found no trends and no seasonality in the absolute error or bias, and the averaged absolute error and bias were on the order of a hundredth to a tenth of a degree Celsius, but there was concerns regarding the dataset used to evaluate the gridded datasets. Guentchev et al. (2010) tested the PRISM dataset and the dataset describe by Maurer et al. (2002) for inhomogeneities in the precipitation time series over the Colorado River basin, and found both to contain them but the Maurer et al. (2002) dataset contained more. Hasenauer et al. (2003) aimed to validate the concept of the DAYMET dataset over Austria for 1960-1999, but again the dataset was evaluated against a questionable time series.

The purpose of this study was to evaluate three popular and highly resolved gridded climate datasets. Those datasets were described by Maurer et al. (2002), Daly et al. (2008)/Di Luzio et al. (2008) and Thornton et al. (1997) and are referred to as the Maurer, PRISM/DiLuzio and DAYMET datasets. The evaluation took place both at the daily scale via extreme heat indices and at the seasonal level via conventional temperatures. It focused on the summer season, which is important

because the effects of discontinuities are often a function of season and the summer season is important to the heat-health discussion. We are unaware of a similar evaluation of these datasets.

The beginning of this evaluation examines whether significant temporal mean differences existed. The next part addresses whether these datasets reproduced the temporal trends of the reference climate dataset. Then subsequent focus was on what spatial scales could reproduce the trends of the reference dataset. Lastly, attention was given to linking the differences to urbanization, instrument changes, time of observation changes and different homogenization adjustments made at the stations.

3.3. Background and datasets

3.3.1 Climate datasets

The Maurer dataset spatial domain covers the CONUS plus parts of Canada and Mexico. It has a spatial resolution of roughly 12 km, daily data and is serially complete over the 1949-2010 period. Its widespread use has risen due to the dataset being used to downscale global climate model output (Maurer et al. 2010; Hayhoe et al. 2010) through the bias-correction and spatial disaggregation downscaling technique (Wood et al. 2004). It has also been used to assess climate signals, for example summer nighttime temperature trends in the California's Central Valley (Bonfils et al. 2007), trends in annual maximum temperatures in Florida (Waylen et al. 2012) and the U.S. spatio-temporal patterns in surface temperature caused by the El-Niño/Southern Oscillation (Zhang et al. 2012). The dataset is based on the Co-Op network observations (as described in section 1.3). There is no mention of disqualifying any stations due to quality concerns and the stated average density of stations in the underlying network implies that all Co-Op stations were used. There is no mention of homogenization or addressing non-climatic biases.

The DAYMET dataset spatial domain covers the United States, Mexico, and Canada. It spans only 1980-2008 but is at a higher resolution (1 km). The published and unpublished literature regarding the DAYMET dataset could be larger. The only

known published literature regarding temperature is a proof of concept paper (Thornton et al. 1997), which provides analysis of a 400000 km^2 area in the north west CONUS during 1989. The temperature values in the DAYMET dataset are widely used in several application communities such as those modeling past and current fire hazard and risk (Keane et al. 2010), modeling productivity of forests (Turner et al. 2011; Littell et al. 2010), modeling biogeochemical cycling rates (Hartman et al. 2011; Pan et al. 2009), mapping past and future corn pest risk (Diffenbaugh et al. 2008) and modeling the transmission risk of human diseases (Konrad et al. 2011; Wimberly et al. 2008). The DAYMET dataset literature states that it is based on observations from the Co-Op network and several hundred from the SNOTEL network to create its grids. There were no mentions of disregarding stations due to quality concerns, or any steps taken to homogenize the time series.

The final dataset evaluated was a daily version of the PRISM dataset. The PRISM dataset has a 4 km resolution from 1895-1997 covering the CONUS. Scientists of various backgrounds, including climate, created the PRISM dataset. The PRISM dataset has relatively sophisticated consideration of geographical features and quality control (Daly et al. 2008). For example, it takes distance to coast into consideration (for temperature) when creating its grids. The PRISM dataset is frequently used for its precipitation fields but studies also use it for its temperatures fields such as the projected impact the 21st century climate changes will have on tree growth (Williams et al. 2010), river basin crop yields (Srinivasan et al. 2010) and riverine nitrogen flows (Schaefer et al. 2009). The NWS Co-Op station observations provide the temperature data, and there was no mention of not using stations due to quality. There was no mention of homogenization efforts. The temporally downscaled version of the PRISM dataset described in a study by Di Luzio et al. (2005) spans 1960-2001. A similar method to that described by Hamlet and Lettenmaier (2005) was employed for this downscaling. That downscaled dataset is the dataset evaluated and is referred to as the PRISM/DiLuzio dataset.

Within the climate community a number of observational datasets (of surface climate) have been designed for determining trends, oscillations and the behavior of temperature at multiple spatial and temporal scales. There is, and has been,

continuous effort to develop and improve the creation of the independent/optimal and systematic processes of homogenization, quality control, and subsequent updating that each dataset has in current operation. These datasets are often funded by major agencies (e.g. NASA) and are widely accepted to be of high quality and are made publicly available; thus informally they are considered to be standards or "trustworthy".

These climate-community observational dataset standards include the National Climatic Data Center's (NCDC) Global Historical Climatology Network (GHCN)/United States Historical Climatology Network (USHCN), (Lawrimore et al. 2011; Menne et al. 2009), the Hadley Centre and University of East Angila's Climate Research Unit (CRU) datasets (Mitchell and Jones 2005), the National Aeronautics and Space Administration's Goddard Institute for Space Studies (NASA-GISS) dataset (Hansen et al. 2010). These datasets have monthly resolution and are either ungridded (GHCN, USHCN) or low resolution (CRU and GISS are 0.5° and 2.0°, respectively). These datasets generally agree with one another (Rohde et al. 2013; Hansen et al. 2010). The dataset for this study used was the downscaled form of the USHCNv2.0-monthly dataset (Menne et al. 2009), as described in section 2.2.1. The USHCNv2.0-monthly dataset is appropriate to evaluate trends, but needed to be temporally downscaled to the daily scale (described in section 2.2.1). We again want to acknowledge that an empirical method of arriving at daily data was used, but chose this route since there were no homogenized daily climate datasets available (that homogenized properly at the daily scale).

The continental United States (CONUS) has a robust meteorological observing network called the Cooperative Observer (Co-Op) Network Program (McCarthy 2007) with roughly 7600 stations reporting daily minimum and maximum temperatures. Observations from this network encompass the temperature information in the Maurer, DAYMET, PRISM/DiLuzio and USHCN datasets. The USHCN network is a subset of 1218 Co-Op stations with longer records. The Maurer dataset states that it is based on the Co-Op stations and does not acknowledge removing any stations. Available information regarding the PRISM/DiLuzio dataset recognized the Co-Op network as its source of temperature observations but also did

not mention removing stations due to quality concerns. The DAYMET dataset literature does not mention any needs of removing stations for any reasons, and lists the Co-Op network as its source of temperature observations.

3.3.2 Supplementary datasets

A gridded population dataset available at NCDC (Owens and Gallo 2000) was used to generate a proxy for urbanization. The dataset spans 1930-2000 at the decadal temporal resolution with a spatial resolution of 1 km. This dataset has previously been used to diagnose observing stations as either rural or urban (Hausfather et al. 2013) from the surrounding areas.

Information regarding the instrument type at each station was used as a proxy for discontinuities stemming from instrumentation changes. The list of USHCN sites with Maximum Minimum Temperature Sensors (MMTS) and Cotton Region Shelter liquid in glass thermometers (CRS) used in the Menne et al. (2010) study was used for establishing the majority of stations. Stations not included in that set were diagnosed using the NCDC's Historical Observing Metadata Repository. The equipment listed as of summer 2010 was extracted and each station was subsequently categorized as MMTS, CRS, hygrothermometer or other.

3.3.3. Background

The causes of non-climatic temporal discontinuities in climate records are understood and there exist methods of, to some degree, correcting most of them (i.e. homogenization as in section 1.4.1). The pairwise comparison method is a popular method of homogenization, and fully described elsewhere (Menne and Williams Jr. 2009). This method uses the time series of difference between a station and its highly correlated neighbors to detect the time and magnitude of discontinuities. This method was used to homogenize the USHCNv2.0-monthly dataset (Menne et al. 2009) and has been defended as robust and objective (e.g. Menne et al. 2010).

Discontinuities stemming from changes in what time of day the temperatures were observed at, referred to as the "time of observation bias" are well documented

(Baker 1975; Karl et al. 1986; DeGaetano 1999; Belcher and DeGaetano 2005) and are corrected for in the USHCNv2.0-monthly dataset (Menne et al. 2009; Vose et al. 2003). Likewise, discontinuities stemming from development/urbanization of the local scale surroundings affecting the temperatures at a station are documented (Karl et al. 1988; Gallo et al. 1996; Menne et al. 2009). The pairwise comparison method the USHCNv2.0-monthly employs corrects those biases (Hausfather et al. 2013). Similarly, changes in microscale surroundings and station relocations also create erroneous discontinuities. The pairwise comparison method corrects these errors as well. Instrument changes at a station also create discontinuities and should be corrected by the pairwise comparison method. Metadata indicating when the non-climatic biases occurred at each station were not necessary. The pairwise comparison method does not require documentation to detect points in time where non-climatic biases caused shifts in the time series.

3.4. Methods

3.4.1 Comparison description

The comparison occurred during the summertime (i.e. May 15 through September 15) using five metrics. Three of which were extreme heat indices; specifically summertime percentile exceedence and the sum of participating extreme heat event (EHE) days per summer, referred to as percentile exceedence counts (PEC) and EHE-days. The 90^{th} percentile was chosen as the threshold because it equates to the widely used “warm night” and “hot day” climate indices. An EHE was defined as a period that began when two or more consecutive daily maximum and minimum percentile exceedences occurred and ended when either the running mean percentile for the daily maximum or minimum ceased to exceed the threshold (following the methodology in section 2.2.2). The other two comparison metrics were summer-mean daily maximum and minimum temperatures. The comparisons were through the differences in those metrics, and those differences were calculated by subtracting the value of the metric in the USHCN dataset from that in the gridded

dataset, for each location of comparison and year of available data. These differences were referred to as residuals.

These comparisons between the gridded datasets and the downscaled USHCN dataset occurred at the locations of selected USHCN stations, referred to as comparison-locations. Bilinear interpolation from the nearest grid points provided the time series from the gridded datasets at comparison-locations. Since the USHCN is a subset of the Co-Op network and the gridded datasets use all of the Co-Op network stations it was assumed that the same Co-Op station within the USHCN network should predominantly influence the time series at the comparison-locations within the gridded datasets. The difference is the time series of the Co-Op stations at the comparison-locations were homogenized within the USHCN network but not in the underlying data within the gridded datasets. The high spatial resolution of these grids should minimize both the influences of the interpolation techniques used to estimate time series at the gridpoints and the bilinear interpolation used to estimate the time series at the comparison-locations.

Appropriate sets of USHCN stations to be used for comparison-locations were chosen based on meeting several requirements (nearly identical to those in section 2.2.1). The sets were unique to each dataset being evaluated because each gridded dataset spanned a different time period. The evaluation of each dataset matched the time span of that dataset: 1949-2010 for the Maurer dataset, 1960-2001 for the PRISM/DiLuzio dataset and 1980-2008 for the DAYMET dataset. The first set of requirements were having at least 80% of the years available, 70% available in both the first and second half of the period and having available data both the first or second-to-first and last or second-to-last years of each period. A year's data was only considered available if all five summer months (i.e. May, June, July, August and September) were present in both the daily maximums and minimums. This left 325 USHCN stations in the Maurer evaluation set, 408 stations in the PRISM/DiLuzio set and 272 stations in the DAYMET set.

The next requirement was that no station could have the worst site-quality rating in the surfacestations.org project. Consequently, stations with compromised sittings were not included in the set of comparison-locations. This left 315 stations in

the Maurer evaluation set, 395 stations in the PRISM/DiLuzio set and 267 stations in the DAYMET set.

The next requirement ensured the climatology period would have a robust amount of data at each station. The Maurer and PRISM/DiLuzio climatology periods were 1971-2000, but because the DAYMET dataset didn't have any data in the 1970s, the climatology period was 1980-2010. The requirements were that at least 135 (out of 150) months were available, and that 27 (out of 30) years were available for each month. This left 272 stations in the Maurer set, 360 stations in the PRISM/DiLuzio set and 248 stations in the DAYMET set.

The last two requirements ensured the monthly data in the USHCNv2.0-monthly dataset were robustly original, rather than estimated by the USHCNv2.0-monthly infilling algorithms. Specifically, the first requirement was that 90% of the available data in the time period was original for each station. This left 245 stations in the Maurer set, 301 stations in the PRISM/DiLuzio set and 196 stations in the DAYMET set. The second requirement ensured the climate base period specifically, had original data. Thus 90% of the data during the respective 30-year climate base period were required to be original. This left 202 stations in the Maurer set, 247 stations in the PRISM/DiLuzio set and 156 stations in the DAYMET set. The bulk of the comparisons focused on climate indices of extreme heat. These indices were specifically summertime percentile exceedence and the sum of participating extreme heat event (EHE) days per summer, and again are referred to as percentile exceedence counts (PEC) and EHE-days. Summertime was defined as May 15 through September 15. The 90^{th} percentile was chosen as the threshold because it equates to the widely used "warm night" and "hot day" climate indices. An EHE was defined as a period that began when two or more consecutive daily maximum and minimum percentile exceedences occurred and ended when either the running mean percentile for the daily maximum or minimum ceases to exceed the threshold (following the methodology in section 2.2.2).

3.4.2. Percentiles, biases and residual trends

The time series of temperatures, from both gridded and USHCN station datasets at each comparison-location, were turned into time series of percentiles. The percentiles at each comparison-location were determined relative to the past values at that comparison-location, on similar calendar dates within the climate base period. This climate base period was built following the suggestions in the Zhang et al. (2005) study evaluating percentiles and climate base periods. For example, one consideration was the method of subsampling used to select the values within the climate base period for calculating each percentile. This analysis used a window size of 15 consecutive dates centered on the calendar date (e.g. the sample for June 15th consisted of data from June 7th to 22nd) (as described in section 2.2.2; Figure 2.1).

The climate base sample that the percentiles were established from, for both the USHCN and the gridded datasets, was built using only the USHCN data. This was done because the DAYMET dataset only had 29 years worth of data and because the USHCN data was not afflicted by any erroneous discontinuities. Furthermore, holding the climate base sample constant allows for clearer interpretation of the results. This required the times series from the gridded dataset be brought into line with their USHCN counterparts. The mean difference was quantified over all (summertime) calendar dates and all years of available overlapping data, separately at each comparison-location and for each daily temperature extreme. These values were subsequently used to force the means of the gridded time series to equate to those of the USHCN time series (i.e. subtracted the quantified mean difference value from each individual value within the gridded time series). These temporal mean temperature differences comprise two of the comparison metrics, as will be described in this section.

The percentiles were calculated by first calculating the empirical cumulative distribution function (Kaplan and Meier 1958), which assigned a percentile (y_o) to each temperature value (x_o) from the climate base period sample. Subsequently, bilinear interpolation was used to find the value (y_i) of the aforementioned percentile function (y_o) at the target temperature value point (x_i) in the climate base period temperature function (x_o). If the target temperature value (x_i) was larger (smaller)

than any value in the climate base period sample (x_o) then it (y_i) was assigned a 1.0 (0.0).

Of the five comparison metrics, only the summer mean daily maximum and minimum temperatures evaluated the differences in the temporal average, and were referred to as biases. The biases, as described previously, were the mean residuals (i.e. differences) quantified over all summertime calendar dates in all years of available overlapping data. They were tested for statistical significance using Student's t-tests. Significance was at the 90% confidence (or alpha = 0.10) level. Conversely, the trends were evaluated via the three heat indices by evaluating the linear trends in the residuals, referred to as the residual trends. Residual trend magnitudes were estimated through the ordinary least squares method and the statistical significance was indicated by non-inclusion of the zero value for the 90% confidence intervals of that estimated trend. Trends calculated using non-parametric methods were evaluated (for assurance), but were decided against because of uncertainty in handling missing values. Those results, however, confirmed the trends presented.

Because both the USHCN and gridded datasets use the same data from the Co-Op network, the USHCN is not in the conventional position to evaluate the absolute accuracy of the values in the gridded datasets (e.g. the USCRN network observations (Heim Jr. 2001), which are in pristine locations with multiple simultaneous observations). However since the noteable differences in the datasets lie in station quality requirements, homogenization and the subsequent grid creation, if the biases and temporal residuals were statistically significant and can be attributed to the effects of the non-climatic discontinuities then a reasonable conclusion was the lack of homogenization generated error(s) in the data.

3.4.3. Continental averages

For all five comparison metrics, the continental mean residuals each year were calculated to display time series of the residuals for all three datasets. The continental averages of the residual linear trends were also provided. Determination

of the CONUS spatial averages required creating a gridded product from the ungridded distribution of values; this was done for both residual trends and biases using Esri ArcGIS software (Esri Corporation, Redlands, CA). The common “inverse distance weighting” method (exponent of distance weighting of two) estimated the values at each grid point using the underlying comparison-location values. For each grid point, values from the ten nearest comparison-locations were used to determine the value. To ensure sensitivity to map-making methods with the spatial distributions of the evaluation sets was not substantial, grids were also constructed via the “Ordinary Kriging” (OK) interpolation method (Holdaway 1996). Kriging is a geostatistical estimator used to infer values at unobserved locations based on values at known locations, and a popular interpolation method.

Only grids with centroids within the continental United States were used in the averaging process. When calculating the continental spatial mean each grid cell was weighted by the cosine of its latitude. This was done because the spatial area of northern gids was smaller than that of southern grids, due to the arrangement of the grid. The averaging was done using a bootstrapping procedure that created 7500 means with sample replacement and subsequently the median mean was determined along with the 90% confidence intervals.

3.4.4. Spatial variability

Evaluation of the spatial variability was partially accomplished via maps of the comparison-locations. Maps were created showing either the residual trend magnitudes, or biases, at each comparison-location and their statistical significance. This allowed visual analysis of spatial patterns and offered support to the results of tests in spatial variability at different scales.

3.4.5. Regional and small-scale variability

Objective testing of the agreement in residual trends at the regional scale prompted splitting the country into regions: north west, south west, north central, south central, north east and south east. The country was delineated using the 100-

degree longitude line, the 39.72-degree latitude line and either the 83 or 87-degree longitude line (in the north and south, respectively). To avoid influence on the comparison-location distributions within each region, the grid points (see section 3.5.1) within each region were used instead of the comparison-locations.

Next the spatial autocorrelations of the residuals trends were examined, using Esri ArcGIS software. The distances between each comparison-location were estimated with the simple spherical law of cosines formula. Using the distances between comparison-locations and the residual trends at each one, the "Moran's I test" (Moran 1950) estimated the Moran's index statistic, which indicates where on a spectrum (-1 to +1) the pattern was from perfectly dispersed (i.e. spatial anti-autocorrelated) to perfectly clustered (i.e. spatial autocorrelated). It also provides P-values for inference of the statistical significance of each index value estimated. Lack of meaningful autocorrelations was taken as indicative of large amounts of small-scale variability relative to the regional-scale variability.

3.4.6. Relationships with non-climatic bias proxies

In an attempt to discuss the physical drivers behind the differences these high-resolution gridded climate datasets had with the USHCN dataset; residual trends were assessed for relationships with external information variables inherent to those comparison-locations. Pearson's correlation coefficients were calculated over the full set of comparison-locations evaluating each dataset (i.e. estimated the spatial correlation). These correlations were with proxies for known causes for non-climatic bias type discontinues.

Before correlations with the proxy for urbanization were calculated from the population dataset, sensitivity of the correlations to the sampling radius (of the population 1km grids) surrounding each comparison-location was breifly examined. After preliminary examination of the correlation coefficients over a wide range of radii it was decided to move forward using the radius corresponding to the largest coefficient between the nearest 5 to 65 grid cells (~1-4km radius). This was separately done with each dataset and heat index. Subsequently the ordinary least

squares estimated linear trend in population over that radius from 1950-2000, 1960-2000, 1980-2000 for the Maurer, PRISM/DiLuzio and DAYMET datasets (respectively), provided the population change values. The sets of these values, for each dataset, represent the urbanization proxy.

Proxies for different amounts of homogenization applied to the USHCN dataset were quantified and subsequently correlated with the residual trends in the heat indices for each dataset being evaluated. This was feasible (within the scope of this work) because the USHCNv2.0-monthly dataset is publicly available in three versions: adjusted for time of observation bias only (TOB adjusted), fully adjusted (fully adjusted) and completely unadjusted (raw). After aggregation of USHCNv2.0-monthly values into summer mean temperatures, three time series were calculated by subtracting: TOB adjusted values from raw values (proxy for TOB adjustments), fully adjusted values from raw values (proxy for all adjustments) and fully adjusted values from TOB adjusted values (adjustments post TOB adjustments). This was done separately for each daily temperature extreme; subsequently the linear trends were estimated for each of these resulting six time series. These linear trends were the proxies for three levels of homogenization.

As a proxy for changes in instrument type; the residual trends were grouped by current instrument type (MMTS, CRS or hygrothermometer) and the group means compared with one another. This was done in residual trends across the three heat indices and datasets. The comparison was done by the median of 7500 means estimated via bootstrapping, and significant differences were determined by non-overlapping 90% confidence intervals.

3.4.7. Interpretation of gridded extreme heat index trends

It was potentially a beneficial exercise to produce and examine grids of trends in extreme heat indices from a gridded dataset, post the conclusions about the uncertainties within these datasets. The Maurer dataset was chosen and the time period was 1970-2010 to match the chapter 2 analysis. The methodology for processing these grids (e.g. percentiles, EHE definition) was like that described in

chapter 2, except here the Maurer dataset observations were used for its climate base period. The only EHE characteristics quantified here were the number of EHEs and the mean duration of those events each summer. All three EHE types (as described in section 2.2.2.) were quantified and thus six maps were made. Statistical significance was also determined at each gridpoint and indicated on the map.

3.5. Results

3.5.1. Biases and continental averages

Nearly all comparison-locations (89-99%) had statistically significant biases (Table 3.1). Most comparison-locations showed positive bias (i.e. gridded time series were warmer than USHCN station time series) of the daily maximum, and the daily minimum bias was less consistent but was generally negative. Continental averages of the biases (Table 3.2) supported those results including the (spatial) statistical significance of the biases. While statistically significant, these values were two to three orders of magnitude smaller than the summertime mean temperatures they were evaluating.

The time series of the annual continental mean residuals (Figure 3.1) displayed the residuals varied through time from short to long time-scales, and appeared to vary independently of the other datasets as well as from the other daily temperature extreme. For example, the time series seemed to (temporally) vary from inter-annual to long-term trends and the daily maximum summer-average temperatures and PEC residuals varied dissimilarly (Pearson's temporal correlation coefficient for the PRISM/Diluzio residuals was 0.35). The jump in the PRISM/DiLuzio biases in Daily minimum PEC in the late 1990s was also notable.

Regarding residual trends, the popular sign and percent of significant residual trends showed 25-75% of the comparison-locations were significant depending on heat index and dataset (Table 3.1). The continental mean results (Table 3.2) indicated the trends were commonly (5 out of 6) the opposite sign of the corresponding biases (summer mean daily maximum and daily maximum PEC), and results from table 3.1 did not disagree with that (4 out of 6). Bootstrapping of the

continental averages indicated all were statistically significant. Continental mean biases and residual trends both varied by heat index and dataset. The EHE-days index consistently showed characteristics similar to both the two PEC trends. The magnitudes of the trends in EHE-days within the USHCN dataset and gridded datasets were also provided, and show the trends to be from double to an order of magnitude larger than the residual trends. Statistically significant distinctions were not seen between methods of map making (Table 3.3).

The continental mean residual trends varied by dataset; which was anticipated in small part due to different spatial coverages of comparison-locations, but primarily because of time period differences. It was unexpected that the biases and residual trends were statistical significant at the continental mean level. Both the biases and residual trends at the continental level were notably smaller than the temperatures and index trends they evaluated, but the biases were much smaller than their counterparts. The residual trends were of the opposite sign of the mean biases, which signifies that further backwards in time the residuals were larger. This is typical of differences between homogenized and non-homogenized time series (Peterson and Vose 1997).

3.5.2. Spatial Variability

The spatial patterns of biases (Figure 3.2) were spatially 'noisy' at small scales, supported the table 3.1 and 3.2 results, and did not resemble the linear trends in the summer mean temperatures during this period (not shown). Particularly visible was the abundance of comparison-locations over 1 °C (theoretically 32% of comparison-locations exist outside +/- 1 standard deviation) including 1 comparison-location in the north east region with biases over 4 °C. Thus biases at the small scale were of a more similar magnitude to the continental mean summertime temperatures.

Evaluation of PEC residual trend maps of each comparison-location's magnitude and statistical significance (Figure 3.3) indicated spatial patterns appeared generally anti-correlated with the patterns in the biases (Figure 3.2). This

was confirmed by (spatial) correlation coefficients between biases and residual trends ranging from -0.13 to -0.57 (not shown). The spatial patterns were different depending on daily temperature extreme, but how different, varied with dataset. For example, the spatial patterns of Tmin and Tmax PEC looked dissimilar in the Maurer dataset, but similar in the PRISM/DiLuzio dataset. All spatial patterns were dominated by small-scale variability, however there did exist some distinguishable regional coherency. For instance, the south west region of the Tmin PEC residual trends in the Maurer dataset appeared to have positive residual trends (i.e. larger trends in the gridded dataset). The mapped results also confirmed that the abundances of comparison-locations with statistically significant residual trends were not sensitive to daily temperature extreme but did vary with dataset being evaluated (~45-74%; Table 3.1).

Maps of the EHE-day residual trends (Figure 3.4) also exhibited the dominance of small-scale variability. Comparisons with the PEC residual trend maps (Figure 3.3) suggested the spatial distributions of EHE-days had similar spatial patterns, but were neither strictly an average of the two daily temperature extreme PECs nor identical to one particular daily temperature extreme PEC. It appeared that there were fewer instances of statistically significant residual trends in the EHE-day index than either of the PEC indices. Comparisons were provided with maps of EHE-day trends in the gridded datasets and in the USHCN station data (Figure 3.5), which illuminate how much more (less) small-scale (large-scale) variability exists in the trend residuals (trends) of the heat indices we were comparing.

Anti-correlation of the spatial patterns between biases and residual trends was indicative that the residuals generally shrank forward in time, as homogenization typically induces. Contrary to previous conclusions that suggested they should average out at larger scales, results indicated regional variability and continental bias existed. As illustrated in the time series, the differences in residuals between datasets and daily temperature extremes were indicative of the impacts the methods of interpolation and spatial resolution have.

3.5.3. Regional and small-scale variability

Further investigation of the biases was omitted due to the relatively smaller size of the biases compared to what they evaluated. The results of grouping the residual trends by geographical region indicated regional-scale variability existed in the trend residuals (Table 3.4). Across all three datasets and heat indices, significant disparities existed between the different regional mean residual trends. The ratio of dissimilar to similar regions varied from roughly 3 to 5.5, depending on dataset. The sensitivity to map-making method again was not statistically significant (Table 3.5). Features seen in the maps were confirmed here, such as the south west region daily minimum PEC residual trends being positively biased.

The results of the evaluation of the spatial patters from an autocorrelation standpoint indicated dominance of small-scale variability over regional clustering. The Moran's I index values for the residual trends were very small (Table 3.6), indicating the variability within regions was large compared to the variability between regions. Comparing the residual trends to the trends in the heat indices themselves (not shown) displayed the Moran's I index values were an order of magnitude smaller in the residual trends. That implied the small-scale variability was relatively larger than the regional variability in the residual trends than it was in the trends in the heat indices.

3.5.4. Relationships with proxies

The population trends surrounding each comparison-location (i.e. the urbanization proxy) had the largest correlation coefficients at 3.61km, 3.00/1.12km and 1.80km for the Maurer, PRISM/Diluzio and DAYMET datasets, respectively (Table 3.7). Radii were consistent between heat indices for two out of the three datasets, but were not consistent across datasets. Correlation coefficients ranged from 0.03-0.17 and less than half were statistically significant.

Comparisons of the residual trend means for each instrument type (Table 3.8) displayed that about a quarter of the groupings were statistically dissimilar. The

DAYMET dataset had the most instances of dissimilar groupings, and the PRISM/DiLuzio dataset had no groupings deemed to be statistically dissimilar. The heat index with the most dissimilar groupings was EHE-days, and the grouping that was most dissimilar was the “MMTS compared to CRS” grouping.

The correlations with the levels of homogenization proxies (proxy for TOB adjustments, adjustments post TOB adjustments and all adjustments) showed coefficients (Table 3.9) ranged from 0.03-0.84. The proxy for TOB adjustments exhibited substantially smaller coefficients (0.03-0.37) than the coefficients for the proxy for adjustments made post TOB adjustments (0.16-0.82). The proxy for all adjustments had the largest coefficients (0.18-0.84). The Maurer dataset typically had larger coefficients than the other datasets did. The coefficients corresponding to the adjustments made post TOB adjustments proxy had larger coefficients with the daily maximums than minimums, but both the proxy for adjustments post TOB adjustments and the proxy for all adjustments showed larger coefficients with the daily minimums.

The proxy for all adjustments being larger than the other two homogenization proxies was anticipated, but the coefficients for the proxy for all adjustments was lower than expected. The small spatial correlation coefficients might have been due in part to poor performance of the proxies, however the correlations were very low and had substantial variability. The variability across datasets with the population dataset sampling radii was expected however, as these datasets all have different spatial resolutions and interpolation methods.

3.5.5. Construction of gridded extreme heat index trends

The grids of an extreme heat index (number of EHE days per summer) produced by the high-resolution datasets for post conclusions discussion, were available (Figures 3.6-3.11) for visual analysis. These resulting grids will be evaluated and interpreted in the discussion section.

3.6. Conclusions and discussion

3.6.1 Conclusions regarding residuals

Both the summer mean temperature residuals (i.e. biases) of the datasets being evaluated were statistically significant. The percentages of comparison-locations with significant biases and the statistically significant continental averages cemented such conclusions. Albeit, the biases were substantially smaller than the summer mean temperatures being evaluated and thus for some applications the biases may be irrelevant. Regardless this was meaningful because these datasets are used for downscaling global climate model output. Issues at the temporal mean level within these high-resolution gridded climate datasets likely impact that downscaled output, regardless of downscaling method. Thus while such downscaling spatially resolves model output it likely introduces uncertainty, which is regularly not recognized and/or discussed.

Our conclusions diverge from the notion that the inaccuracies from discontinuities cancel one another out over large scales. Results in the Scully (2010) work related to large-scale averages were mixed: the daily maximum temperature bias in the DAYMET dataset was trivial, but those in the daily minimums and both daily temperature extremes in the PRISM dataset were not. There were similarities (similar magnitude of PRISM bias values to our PRISM/Diluzio bias) and dissimilarities (0.18 vs. 0.00 °C daily maximum bias) between the continental mean biases found in our study and those Scully (2010) indicated in the annual temperature biases. However any comparison is difficult because the Scully (2010) investigation focused on annual temperatures and did not fix the discontinuities in the time series used to evaluate the gridded datasets. Hasenauer et al. (2003) indicated essentially zero biases over a smaller region (Austria) for the DAYMET dataset, but a similar lack of homogenization and summer season focus mades comparisons difficult.

The residual trends of the heat indices were statistically significant across spatial scales, and largest at small scales. The percentage of comparison-locations with significant trends and the significance of the continental mean residual trends

again prompted such conclusions at the large scale. Regional and autocorrelation tests drove the small to regional-scale conclusions, and maps supported those conclusions as well. While more comparable than the biases, the residual trends still were sometimes an order of magnitude smaller than the trends they evaluated. Results also suggested residual trends to be more comparable to what they evaluated at smaller scales. These conclusions imply trends of features of any spatial scale should be considered to have uncertainty attached to them related to the lack of homogenization. This was meaningful because it has not been quantified previously and thus contributes to the informed use of the dataset for trend analysis. Particularly uncertainty should be acknowledged in small-scale features of trends. Also important was the existence of errors at the continental level because of previous conclusions that large-scale features would be free of error. Disagreement with the Hasenauer et al. (2003) conclusions of no residual trends on the spatial mean was reasonable due to lack of homogenization of the reference time series.

The residuals were highly variable. They varied temporally and across time-scales, which was seen in the time series. The biases varied spatially, as did the residual trends. Both were explicitly confirmed to be significant at the continental scale and the residual trends also confirmed at the regional and small scales. Maps indicated that the biases likely vary as well at the regional and near certainly at the small-scale, but variability at these scales was not actually tested. The residuals acted differently depending on comparison metric and dataset, as was also seen in the time series. These conclusions were noteworthy because they imply that easy solutions/corrections to the datasets are unlikely and homogenizing the underlying data may be the only solution. Due to paucity of evaluations of these datasets there aren't other studies to compare these conclusions to.

Both biases and residual trends were primarily caused by non-climatic discontinuities. These conclusions were based on the way the residuals generally shrank forward in time and the correlation coefficients between the residual trends and the proxy for all the homogenization adjustments. Our disagreements with the Scully (2010) and Hasenauer et al.(2003) studies also support such conclusions as they evaluated the gridded datasets against non-homogenized datasets. Such

conclusions were meaningful because they imply that proper homogenization of the underlying data can potentially substantially improve the quality of these datasets. The improvements the Hamlet and Lettenmaier (2005) method bring to these datasets are a testament to how non-climate discontinuities are a major problem in these datasets.

The residual trends were not strongly correlated with proxies for non-climatic biases or homogenization adjustments at each comparison-location. The results of correlations between residual trends and the proxies (homogenization and urbanization), and the lack of distinctions seen in the residual trends between comparison-locations of different station instrument type, prompted such conclusions. These results were important, as will be discussed below, because they provide insight into how these differences arise.

Results suggest that the adjustments correcting discontinuities unrelated to the TOB were responsible for the majority of the residual trends. This implies the TOB-related adjustments were a relatively small part of this homogenization process. Causes of non-climatic biases adjusted post TOB-adjustment includes: urbanization, microclimate influences, instrument changes and station relocations. However, weak relationships were found between residual trends and proxies for urbanization and instrument changes, which suggests microclimate changes and station relocations play meaningful roles.

Furthermore it can be argued the spatial distribution of station relocations and microclimate influences would be relatively dominated by small-scale variability. Biases due to instrument changes are generally a function of the climate they monitor in, and thus should have regional variability (Quale et al. 1991). Similarly, the influence of urbanization is also a function of the climate being monitored in (Grimmond et al. 2010) and furthermore land use changes during the 1950-2000 period were geographically clustered (Brown et al. 2005). Biases that arise from TOB discontinuities are a function of the levels of temperature persistence (Belcher and DeGaetano 2003), which is also geographically controlled (e.g. coastal areas). Conversely, station relocations and microclimate influences have no geographical/climate controls and thus the errors they instill would have relatively

large small-scale variability.

Another source of the residuals was the gridding process and comparing grids to point sources. During the gridding process discontinuities can arise from changes in the set of comparison-locations used to determine the grid points, via missing data or stations going in and out of use during the dataset duration. Discontinuities also arise at a comparison-location due to the aliasing of discontinuities in surrounding stations into that estimated time series. The time series at each grid point was constructed using an interpolation method (methods are not consistent between datasets) that employs input of numerous surrounding stations' time series. Since the four surrounding grid points were interpolated to generate the time series at the comparison-locations, that time series was a function of numerous stations surrounding the comparison-location. Effectively merging all time series discontinuities from several surrounding stations into the time series at every grid point. This error source potentially explains the lack of stronger correlations with proxies of non-climatic biases and homogenization adjustments in this study. These errors are a source of uncertainty rarely discussed. Thus, while ideally the time series at each USHCN location could be reconstructed by interpolating the four surrounding grid points, it makes more sense that such a reconstruction would be dissimilar.

3.6.2. Discussion regarding gridded extreme heat index trends

As explained in section 3.4.7 an interpretation of the trends of an extreme heat index in a gridded dataset is provided below. The trends in Tmin EHE duration and frequency (as described in section 2.2.2 and 2.2.3) within the Maurer dataset are provided in figure 3.6 and figure 3.7. While it was clear in both grids that small-scale variability/features exist, they should be only lightly trusted here. This study demonstrated the trends in a similar extreme heat index were inaccurate (on the order of days per decade), and the bulk of those errors operate at the small-scale. Thus it would be less than robust to have much confidence in these small-scale features. However there exists regional variability in these figures as well. The

majority of the south showed increase (excluding southern California, Nevada and an area from Kansas through north Texas). There also exists a region in the north central US that displayed slight decreased EHE activity, but it was less consistent (e.g. the frequency slightly increased).

Since this study demonstrated that there existed errors operating at the regional scale, there was an amount of skepticism, but a region of some amount of increase in the south could still be a confident conclusion. However, there was less confidence in the existence of a region of decrease in the north – partly due to a weak signal but also due to smaller spatial extent. On the continental scale, there seems to be a small dominance of increasing trends in both EHE frequency and duration. Since this study's results indicate an amount of uncertainty at the continental scale, such a conclusion might be in relatively low confidence

Both maps of Tmax type EHE (as described in section 2.2.2.) trends also show small-scale variability in the trends of both EHE characteristics (Figures 3.8, 3.9). Again, while some of these features can be interesting limited confidence should be given for features of small scale in this gridded dataset.

Conversely, in both maps there was an increase in a region centered on Utah or Colorado and a region of decrease in the north central region. Other regions of increase and decrease exist but were not coherent across frequency and duration. The trends were convincing in both regions, and so mild confidence could be had in an increase/decrease in these regions. On the continental scale, neither sign was particularly strong, perhaps decreasing trends dominated. Regardless the existence of uncertainty at the continental scale might eliminate what little confidence in that decrease one can acquire from a visual analysis of these maps.

The maps of the Tmnx type EHE (as described in section 2.2.2.) trends display a relatively large amount of regional scale variability in addition to the typical amount of small-scale variability (Figures 3.10-3.11). Again, most all small-scale features should be considered with caution due to the uncertainty at that scale in the Maurer dataset.

The north central region displayed decreasing frequency and duration trends, and the south west and south east regions had frequency and duration trends of increase. It was hard to know whether the south central region actually separated the regions of increase because it was a small area (of generally mixed sign trends), and thus uncertainty was relatively large. Similarly, the confidence in the decreasing region (north central) was less than the regions of increase because it was smaller in areal extent than the increasing region(s). On the continental scale the grids only suggest a very small positive bias and thus confidence in a positive continental-scale trend was very small.

3.6.3 Discussion regarding the Hamlet and Lettenmaier (2005) correction method

The work in both this chapter, and chapter 2, lent itself to thinking about the role of homogenization, scales and gridding in observational datasets. Specifically in chapter 2 the temporal downscaling of the monthly data was not dissimilar from the Hamlet and Lettenmaier (2005) (HL) method of adjusting unhomogenized Co-Op derived datasets in order to better align with more trusted climate datasets. Moreover, this chapter compared gridded Co-Op data to points for comparison with a well-homogenized climate dataset, which was similar to comparing grids of homogenized and non-homogenized climate data (as is done in the HL method).

The HL method has not been explicitly evaluated in the literature; the only known published evaluation was by said authors and its impact on the ability to model normalized stream flow. Its theoretical workings have not been discussed in the literatures either. However there are numerous studies that use the method of correction. Thus thoughts concerning the method, its limitations and how it can be improved are briefly given here in hopes to inform the authors of the next generation of gridded products.

The HL method consists of six simple steps. First a monthly, homogenized dataset – typically the USHCN (1218 stations) – is gridded to $1/8^{th}$ degree spatial resolution (adjusting also for elevation at each point). Then, post infilling of individual days, a (much larger) subset of the daily data from the Co-Op network is gridded to a

matching 1/8th degree resolution (also adjusting for elevation). Thus while the gridded day-to-day variability operates at a high spatial resolution (because of the dense data source distribution of the Co-Op subset), the gridded USHCN monthly averages are effectively at a much lower resolution (e.g. that of the USHCN network). The Co-Op daily grid is subsequently aggregated to produce monthly 1/8th degree resolution grids. Then both the USHCN and Co-Op monthly grids have a three-monthly-running mean applied to each grid point, and subsequently the differences between the USHCN and Co-Op values for each month-grid point are quantified. Those differences are used to adjust the Co-Op monthly grids to match the USHCN grid points. Then the day-to-day anomalies (with respect to the monthly average) are calculated based on the daily Co-Op grids. Subsequently those anomalies are used in conjunction with the adjusted Co-Op monthly grids to make daily grids loosely equal to the USHCN dataset. Full details are available in the Hamlet and Lettenmaier (2005) study.

Subsequent mathematical combination of these two grids results in the daily scale variability and monthly scale variability not operating at the same spatial resolution. The problem is clear when such a dataset is averaged to the monthly scale and the spatial variability drops down to the much lower resolution, a physically not realistic scenario. Furthermore, in an attempt to incorporate a portion of the high spatially varying information, discontinuities are less than optimally handled. Specifically, by forcing the three-month-running mean grids to equate instead of the raw grids to equate, abrupt discontinuities are adjusted over three-month periods instead of instantaneously. Thus discontinuities will still occur but are gradually, and erroneously, 'corrected' over three months; this could have a significant impact on the month before and month after a discontinuity.

A way to move forward on both fronts is herein proposed. It method starts with the full 1218 USHCNv2-monthly stations that are well homogenized as explained in the Menne et al. (2009) study. Iteratively, a pairwise homogenization algorithm is used with those stations to homogenize the time series of nearby Co-Op station not previously included in the USHCN network of 1218, and then subsequently add those newly homogenized stations to the group of USHCN stations

(herein referred to the “homogenized stations sample”). This can continue until the desired increase in the number of stations, and thus station density, is accomplished. Next this set of homogenized stations is turned into a grid using the regression-decision making process that the PRISM dataset uses to interpolate the data down to the 1/8th degree resolution. This allows climatologically smart (i.e. with respect to relationships between physical geography and climate) interpolation of the station data to the high-resolution grids. Then the combination of these monthly-homogenized grids and the daily Co-Op grids should be done, similarly to the HL method but without the temporal averaging. Thus, the disconnect in spatial variability (between monthly and daily scales) will be smaller since the effective resolution of the monthly-homogenized grids will be higher. Moreover the advantages in the PRISM method of interpolation would be retained. Lastly, temporal discontinuities will not be erroneously corrected since the daily data is adjusted to equal the raw, as opposed to temporally averaged, monthly data grids.

3.9. Acknowledgements

This research was sponsored by NOAA Climate Program Office grant NA10OAR4310213, through the Great Lakes Regional Integrated Sciences and Assessments Center (/GLISA/). The Great Lakes Adaptation Assessment for Cities (GLAA-C) also financially helped this work happen, which is supported by the University of Michigan Graham Environmental Sustainability Institute and the Kresge Foundation. I would also like to thank Drs. Brian Belcher and Jeff Andresen for their helpful conversations, as well as the comments by my thesis committee members Drs. Frank Marsik, Marie. S. O'Neill and Allison Steiner.

3.10. References

Baker, D. G., 1975: Effect of observation time on mean temperature estimation. *Journal of Applied Meteorology* **14**, 471-476. doi: 10.1175/1520-0450(1975)014<0471:EOOTOM>2.0.CO;2.

Belcher, B. N. and A. T. DeGaetano, 2003: A method of operational detection of daily observation-time changes. *Journal of Applied Meteorology* **42**, 1823-1836. DOI: 10.1175/1520-0450(2003)042<1823:AMFODO>2.0.CO;2.

Belcher, B. N. and A. T. DeGaetano, 2005: A method to inter time of observation at US Cooperative Observer Network station using model analyses. *International Journal of Climatology* **25**, 1237-1251. doi: 10.1002/joc.1183.

Bonfils, C., P. B. Duffy, B. D. Santer, T. M. L. Wigley, D. P. Lobell, T. J. Phillips, C. Doutriaux, 2008: Identification of external influences on temperatures in California. *Climatic Change* **87**: S43-S55. doi: 10.1007/s10584-007-9374-9.

Brown, D. G., K. M. Johnson, T. R. Loveland, D. M. Theobald: 2005. Rural land-use trends in the conterminous United States, 1950-2000. *Ecological Application*s **15**: 1851-1863. doi: 10.1890/03-5220.

Christensen, N. S., A. W. Wood, N. Voisin, D. P. Lettenmaier, R. N. Palmer, 2004: Effects of climate change on the hydrology and water resources of the Colorado River Basin. *Climatic Change* **62**: 337-363. doi: 10.1023/B:CLIM.0000013684.13621.1f.

Daly, C., M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, P. P. Paseris, 2008: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology*, **28:** 2031-2064.

DeGaetano, A. T., 1999: A method to infer observation time based on day-to-day temperature variations. *Journal of Climate* **12**, 3443-3456. doi: 0.1175/1520-0442(1999)012<3443:AMTIOT>2.0.CO;2.

Diffenbaugh, N. S., C. H. Krupke, M. A. White, C. E. Alexander, 2008: Global warming presents new challenges for maize pest management. *Environmental Research Letters* **3**: 044007. doi: 10.1088/1748-9326/3/4/044007.

Di Luzio, M., G. L. Johnson, C. Daly, J. K. Eischeid, J. G. Arnold, 2008: Constructing retrospective gridded daily precipitation and temperature datasets fo the conterminous United States. *Journal of Applied Meteorology and Climatology* **47**: 475-497. doi: 10.1175/2007JAMC1356.1.

El-Sadek, A., M. Bleiweiss, M. Shukla, S. Guldan, A. Fernald, 2011: Alternative climate data sources for distributed hydrological modeling on a daily time step. *Hydrological Processes* **25**, 1542-1557. doi: 10.1003/hyp.7917.

Fall S., A. Watts, J. Nielsen-Gammon, E. Jones, D. Niyogi, J. R. Christy, R. A. Pielke Sr., 2011: Analysis of the impacts of station exposure on the U.S. Historical Climatology Network temperatures and temperature trends. *Journal of Geophysical Research* **116**: D14120. doi:10.1029/2010JD015146.

Guentchev, G., J. J. Barsugli, J. Eischeid, 2010: Homogeneity of gridded precipitation datasets for the Colorado River basin. *Journal of Applied Meteorology and Climatology* **49**: 2415-2424. doi: 10.1175/2010JAMC2484.1.

Gallo, K. P., D. R. Easterling, T. C. Peterson, 1996: The influence of land use/land cover on climatological values of the diurnal temperature range. *Journal of Climate* **9**, 2941-2944. doi: 10.1175/1520-0442(1996)009<2941:TIOLUC>2.0.CO;2.

Grimmond, C. S. B., M. Roth, T. R. Oke, Y. C. Au, M. Best, R. Betts, G. Carmichael, H. Cleugh, W. Dabberdt, R. Emmanuel, E. Freitas, K. Fortuniak, S. Hanna, P. Klein, L. S. Kalkstein, C. H. Liu, A. Nickson, D. Pearlmutter, D. Sailor, J. Voogt, 2010: Climate and more sustainable cities: Climate information for improved planning and management of cities (producers/capabilities perspective). *Procedia Environmental Sciences* **1**: 247-274. doi: 10.1016/j.proenv.2010.09.016.

Hamlet, A. F., D. P. Lettenmaier, 2005: Production of temporally consistent gridded precipitation and temperature datasets for the continental United States. *Journal of Hydrometeorology* **6**: 330-336. doi: 10.1175/JHM420.1.

Hamlet, A. F., P. W. Mote, M. P. Clark, D. P. Lettenmaier, 2005: Effects of temperature and precipitation variability in snowpack trends in the Western United States. *Journal of Climate* **18**: 4545-4561. doi: 10.1175/JCLI3538.1.

Hansen, J., R. Ruedy, M. Sato, K. Lo, 2010: Global surface temperature change. *Review of Geophysics* **48**: RG4004. doi: 10.1029/2010RG000345.

Hartman, M. D., E. R. Merchant, J. W. Parton, M. P. Guttmann, S. M. Lutz, S. A. Williams, 2011: Impact of historical land-use changes on greenhouse gas exchange in the U.S. Great Plains, 1883-2003. *Ecological Application* **21**: 1105-1119.

Hasenauer H., K. Merganicova, R. Petritsch, S. A. Pietsch, P. E. Thornton, 2003: Validating daily climate interpolations over complex terrain in Austria. *Agricultural and Forest Meteorology* **119**: 87-107. doi: 10.1016/S0168-1923(03)00114-X.

Hausfather, Z., M. J. Menne, C. N. Williams Jr., T. Masters, R. Broberg, D. Jones, 2013: Quantifying the effect of urbanization on U.S. Historical Climatology Network temperature records. *Journal of Geophysical Research: Atmospheres* **118**, 481-494. doi: 10.1029/2012JD019509.

Hayhoe, K., J. VanDorn, I. I. T. Croley, N. Schlegal, D. Wuebbles, 2010: Regional climate change projections for Chicago and the U.S. Great Lakes. *Journal of Great Lakes Research* **36**: 7-21. doi: 10.1016/j.jglr.2010.03.012.

Heim Jr., R. R., 2001: New network to monitor climate change. *EOS, Transactions, American Geophysical Union* **82**: pp. 143. DOI:10.1029/EO082i012p00143.

Holdaway, M. R., 1996: Spatial modeling and interpolation of monthly temperature using kriging. *Climate Research*, **6**: 215-225.

Karl, T. R., C. N. Williams Jr., P. J. Young, W. M. Wendland, 1986: A model to estimate the time of observation bias associated with monthly mean maximum, minimum, and mean temperature for the United States. *Journal of Climatology and Applied Meteorology* **25**, 145-160. doi: 10.1175/1520-0450(1986)025<0145:AMTETT>2.0.CO;2.

Karl, T. R., H. F. Diaz, G. Kukla, 1988: Urbanization: Its detection and effect in the United States climate record. *Journal of Climate*, **1**, 1099-1123. doi: 10.1175/1520-0442(1988)001<1099:UIDAEI>2.0.CO;2.

Kaplan, E. L. and P. Meier, 1958: Nonparametric estimation for incomplete observations. *Journal of the American Statistical Association*. **53**: 457-481.

Konrad, S. K., S. N. Miller, W. K. Reeves, 2011: A spatially explicit degree-day model of Rift Valley fever transmission risk in the continental United States. *Geojournal* **76**: 257-266. doi: 10.1007/s10708-010-9338-x.

Keane, R. E., S. A. Drury, E. C. Karau, P. F. Hessburg, K. M. Reynolds, 2010: A method for mapping fire hazard and risk across multiple scales and it's application in fire management. *Ecological Modeling* **221**: 2-18. doi: 10.1016/j.ecolmodel.2008.10.022.

Lawrimore, J. H., M. J. Menne, B. E. Gleason, C. N. Williams Jr., D. B. Wuertz, R. S. Vose, J. Rennie, 2011: An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3. *Journal of Geophysical Research* **116**: D19121, doi: 10.1029/2011JD016187.

Littell, J. S., E. E. Oneil, D. McKenzie, J. A. Hicke, J. A. Lutz, R. A. Norheim, M. M. Elsner, 2010: Forest ecosystems, disturbance, and climatic change in Washington State, U.S.A. *Climatic Change* **102**: 129-158. doi: 10.1007/s10584-010-9858-x.

Lobell, D. B., C. J. Bonfils, L. M. Kueppers, M. A. Snyder, 2008: Irrigation cooling effect on temperature and heat index extremes. *Geophysical Research Letters* **35**: L09705. doi: 10.1029/2008GL034145.

Maurer, E. P., A. W. Wood, J. C. Adam, D. P. Lettenmaier, 2002: A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States. *Journal of Climate* **15**: 3237-3251. doi: 10.1175/1520-0442(2002)015<3237:ALTHBD>2.0.CO;2.

Maurer, E. P., H. G. Hidalgo, T. Das, M. D. Dettinger, D. R. Cayan, 2010: The utility of daily large-scale climate data in the assessment of climate change impacts on daily streamflow in California. *Hydrology and Earth System Sciences* **14**, 1125-1138. doi: 10.5194/hess-14-1125-2010.

McCarthy, D., 2007: Cooperative station observations. National weather service manual 10-1315, pp. 122. (Available at http://www.docstoc.com/docs/977267/Cooperative-Station-Observations)

Menne, M. J., C. N. Williams Jr., R. S. Vose, 2009: The United States Historical Climatology Network Monly Temperature Data – Version 2. *Bullentin of the American Meteorology Society* **90**, 993-1007, doi:10.1175/2008BAMS2613.1.

Menne, M. J. and C. N. Williams Jr., 2009: Homogenization of temperature series via pairwise comparisons. *Journal of climate* **22**: 1700-1717. doi: 10.1175/2008JCLI2263.1.

Menne, M. J., C. N. Williams Jr., M. A. Palecki, 2010: On the reliability of the U.S. surface temperature record. *Journal of Geophysical Research* **115**: D11108. doi: 10.10.29/2009JD013094.

Mitchell, T. D. and P. D Jones, 2005: An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *International*

Journal of Climatology **25**: 693-712. doi: 10.1002/joc.1181.

Moberg, A., H. Alexandersson, 1997: Homogenization of Swedish temperature data. Part II: Homogenized gridded air temperature compared with a subset of global gridded air temperature since 1861. *International Journal of Climatology* **17**: 35-54. doi: 10.1002/(SICI)1097-0088(199701)17:1<35::AID-JOC104>3.0.CO;2-F.

Moran, P. A. P., 1950: Notes on continuous stochastic phenomena. *Biometrika* **37**: 17-23. doi: 10.2307/2332142.

Owens, T. W. and K. P. Gallo, 2000: Updated population metadata for the United States Historical Climatology Network States. *Journal of Climate* **13**: 4028-4033. doi: 10.1175/1520-0450(1986)025<0145:AMTETT>2.0.CO;2.

Oswald, E. M. and R. B. Rood, 2013: A trend analysis of extreme heat events in the continental U.S. since 1930. Submitted (accepted with revisions to the *Journal of Applied Meteorology and Climatology* on March 1, 2013.

Pan, Y., R. Birdsey, J. Hom, K. McCullough, 2009: Separating effects of changes in atmospheric composition, climate and land-use on carbon sequestration of U.S. Mid-Atlantic temperate forests. *Forest Ecology and Management* **259**: 151-164. doi: 10.1016/j.foreco.2009.09.049.

Peterson, T. C. and R. S. Vose, 1997: An overview of the global Historical Climatology Network temperature database. Bulletin of the American Meteorological Society, **78**, 2837-2849. doi: 10.1175/1520-0477(1997)078<2837:AOOTGH>2.0.CO;2.

Pielke, Sr. R. A., J. Nielsen-Gammon, C. Davey, J. Angel, O. Bliss, N. Doesken, M. Cai, S. Fall, D. Niyogi, K. Gallo, R. Hale, K. G. Hubbard, X. Lin, H. Li, S. Raman, 2007: Documentation of uncertainties and biases associated with surface temperature measurement sites for climate change assessment. *Bulletin of the American Meteorological Society* **88**: 913-928. doi: 10.1175/BAMS-88-6-913.

Quale, G. R., D. R. Easterling, T. R. Karl, P. Y. Hughes. 1991: Effects of recent thermometer changes in the cooperative station network. *Bulletin of the American Meteorological Society*, **72**, 1718-1723. doi: 10.1175/1520-0477(1991)072<1718:EORTCI>2.0.CO;2.

Rohde, R., R. Muller, R. Jacobsen, S. Perimutter, A. Rosenfeld, J. Wurtele, J. Curry, C. Wickham, S. Mosher, 2013: Berkeley Earth temperature averaging process. *Geoinformatics and Geostatistics: An overview* **1**:1. doi: 10.4172/gigs.1000103.

Schaefer, S. C., J. T. Hollibaugh, M. Alber, 2009: Watershed nitrogen input and riverine export on the west coast of the U.S. *Biogeochemistry* **93**, 219-233. doi: 10.1007/s10533-009-9299-7.

Scully, R. A., 2010: Intercomparison of PRISM and DAYMET temperature interpolation from 1980-2003. M.S. thesis. Watershed Sciences, Utah State University. 63 pp.

Srinivasan, R., X. Zhang, J. Arnold, 2010: SWAT ungauged: hydrological budget and crop yield predictions in the upper Mississippi river basin. *Transactions of the ASABE*, **35**, 1533-1546. (Available at http://ddr.nal.usda.gov/handle/10113/46714)

Thornton PE, Running SW, White MA. 1997. Generating surfaces of daily meteorological variables over large regions of complex terrain. *Journal of Hydrology* **190**: 214-251. DOI: 10.1016/S0022-1694(96)03128-9.

Turner, J., S. R. Colwell, G. J. Marshall, T. A. Lachlan-Cope, A. M Carleton, P. D. Jones, V. Lagun, P. A. Reid, S. Iagovkina, 2004: The SCAR READER project: toward a high-quality database of mean Antarctic meteorological observations. *Journal of Climate* **17**: 2890-2898. doi: 10.1175/1520-0442(2004)017<2890:TSRPTA>2.0.CO;2.

Vose, R. S., C. N. Williams Jr., T. C. Peterson, T. R. Karl, D. R. Easterling, 2003: An evaluation of the time of observation bias adjustment in the U.S. Historical Climatology Network. *Geophysical Research Letters* **30**, 2046. doi: 10.1029/2003GL018111.

Waylen, P., D. Keellings, Y. Qiu, 2012: Climate and health in Florida: Changes in risks of annual maximum temperatures in the second half of the twentieth century. *Applied Geography* **33**: 73-81. doi: 10.1016/j.apgeog.2011.06.007.

Williams, A. P., J. Michaelsen, S. W. Leavitt, 2010: Using tree rings to predict the response of tree growth to climate change in the continental United States during the twenty-first century. *Earth Interactions* **14**, 1-20. doi: 10.1175/2010EI362.1.

Wimberly, M. C., A. D. Baer, M. J. Yabsley. 2008: Enhanced spatial models for predicting the geographic distributions of tick-borne pathogens. *International Journal of Health Geographics* **7**: 15, doi: 10.1186/1476-072X-7-15.

Wood, A. W., L. R. Leung, V. Sridhar, D. P. Lettenmaier, 2004: Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic Change* **62**: 189-216. doi: 10.1023/B:CLIM.0000013685.99609.9e.

Zhang, X., G. Hegerl, F. W. Zwiers, J. Kenyon, 2005: Avoiding inhomogeneity in percentile-based indices of temperature extremes. *Journal of Climate* **18**: 1641-1651. doi: 10.1175/JCLI3366.1.

Zhang, Y., Y. Qian, V. Dulière, E. P. Salathé Jr., L. R. Leung, 2012: ENSO anomalies over the Western United States: present and future patterns in regional climate simulation. *Climatic Change* **110**: 315-346. doi: 10.1007/s1058-011-0088-7.

3.11. Figures

Figure 3.1. Time series of the continental average residuals of the metrics being evaluated. Specifically summer average temperatures, percentile exceedence counts and the number of extreme heat event days. All three datasets span their evaluation period.

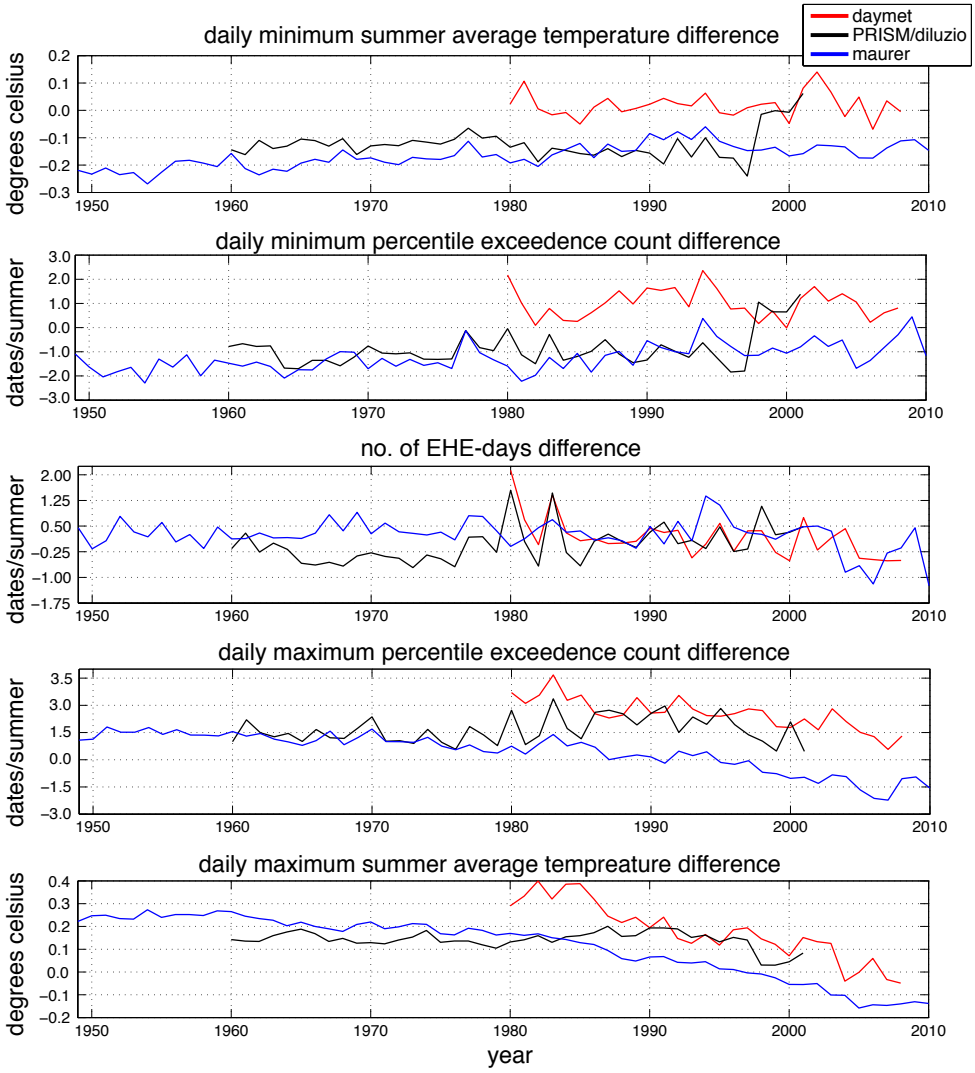


Figure 3.2. Spatial distribution of summer average temperature biases for all datasets and both daily extremes. Symbols increase by standard deviations away from the zero values, and thus the groupings are different for each map. Symbols are shaded/filled in if the biases are statistically significant.

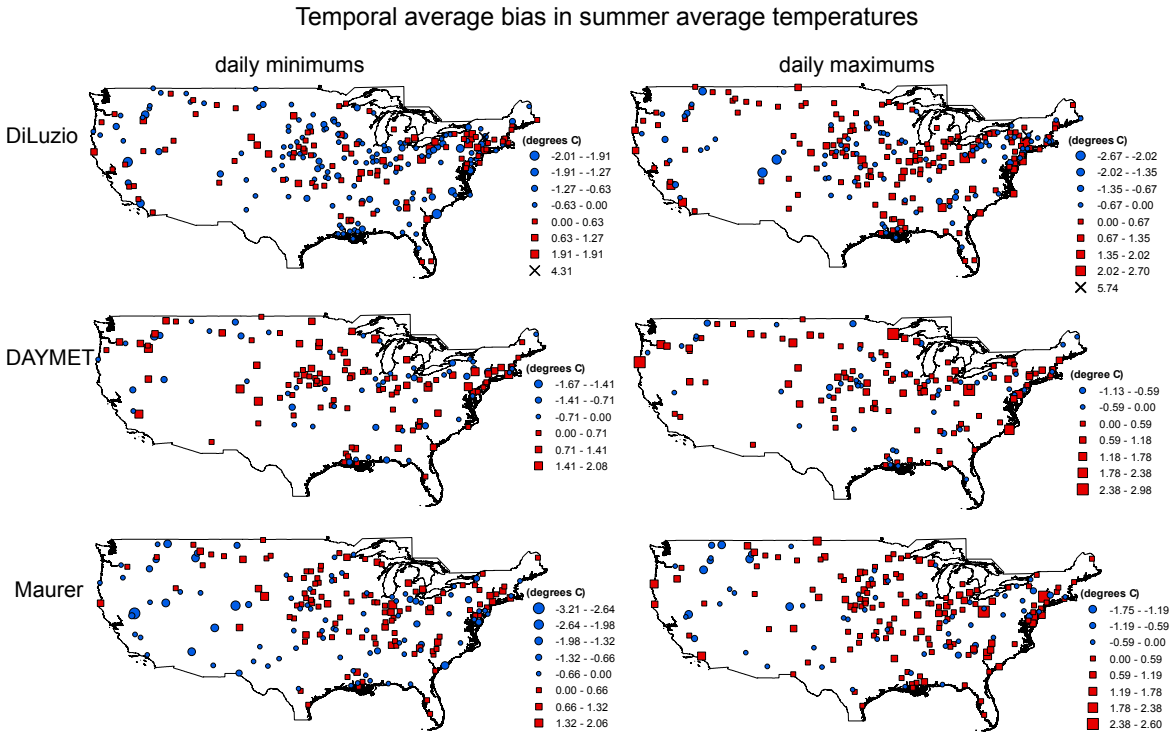


Figure 3.3. Spatial distribution of residual trends in 90th percentile summer exceedence counts for all datasets and both daily extremes. Symbology follows figure 3.2 except for trends.

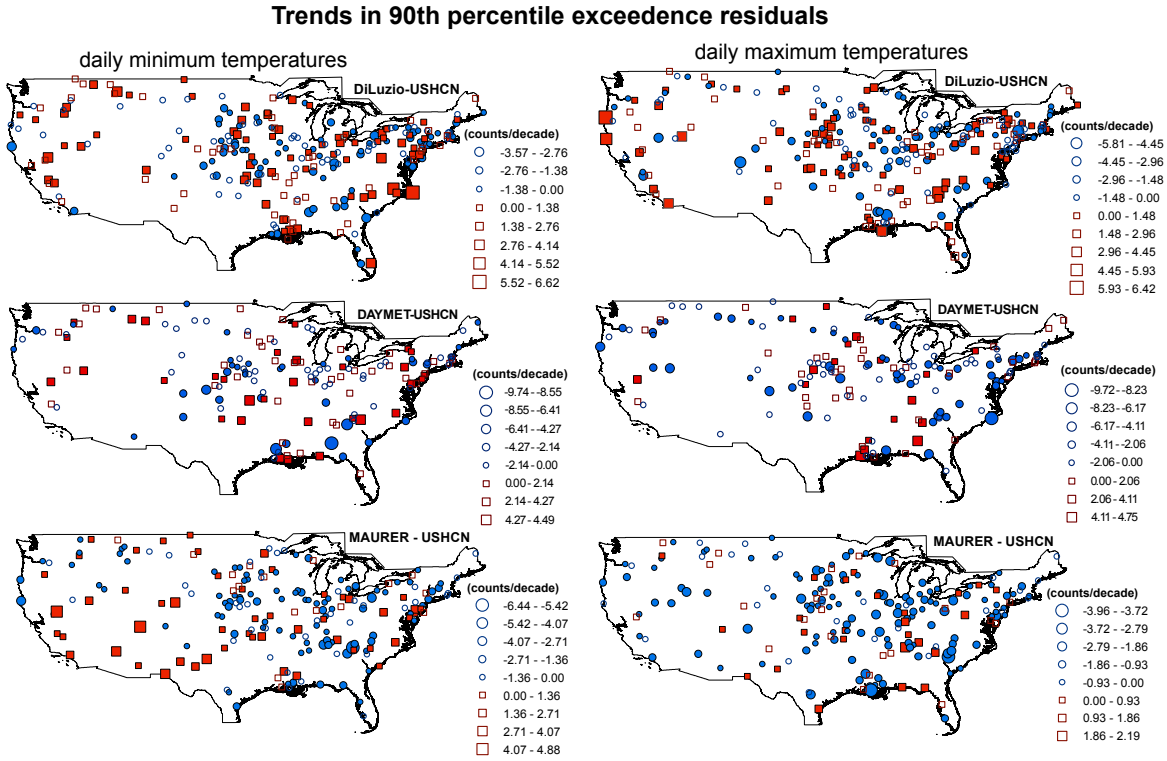


Figure 3.4. The spatial distribution of residual trends in extreme heat event days per summer for all datasets. Symbology follows figure 3.3.

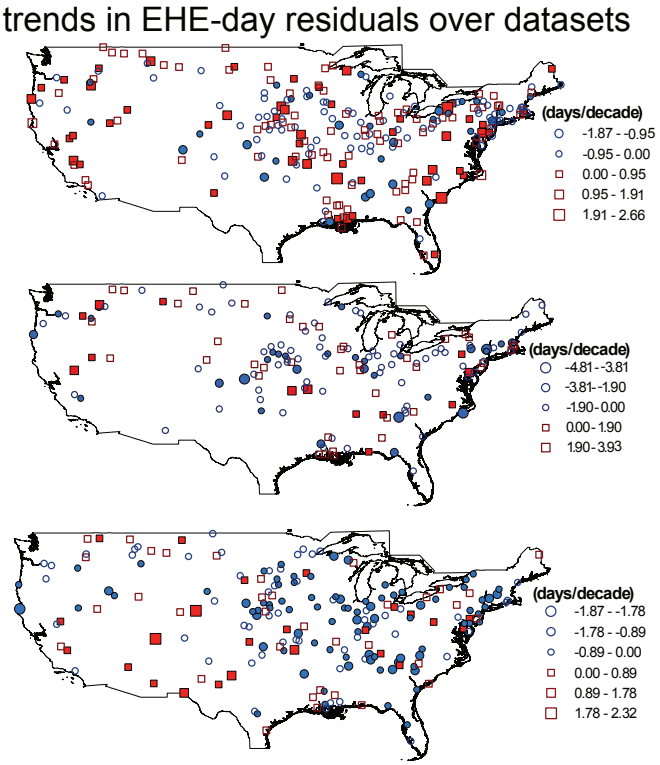


Figure 3.5. Comparison of trends and trend residuals in EHE-days. The left column shows the trends in the gridded dataset, the right column shows the trends in the USHCN station data and the middle column is the trends in the residuals. For each dataset, maps in all three columns have symbols based on the standard deviation of the dataset being evaluated. Statistical significance indicated by symbol infilling.

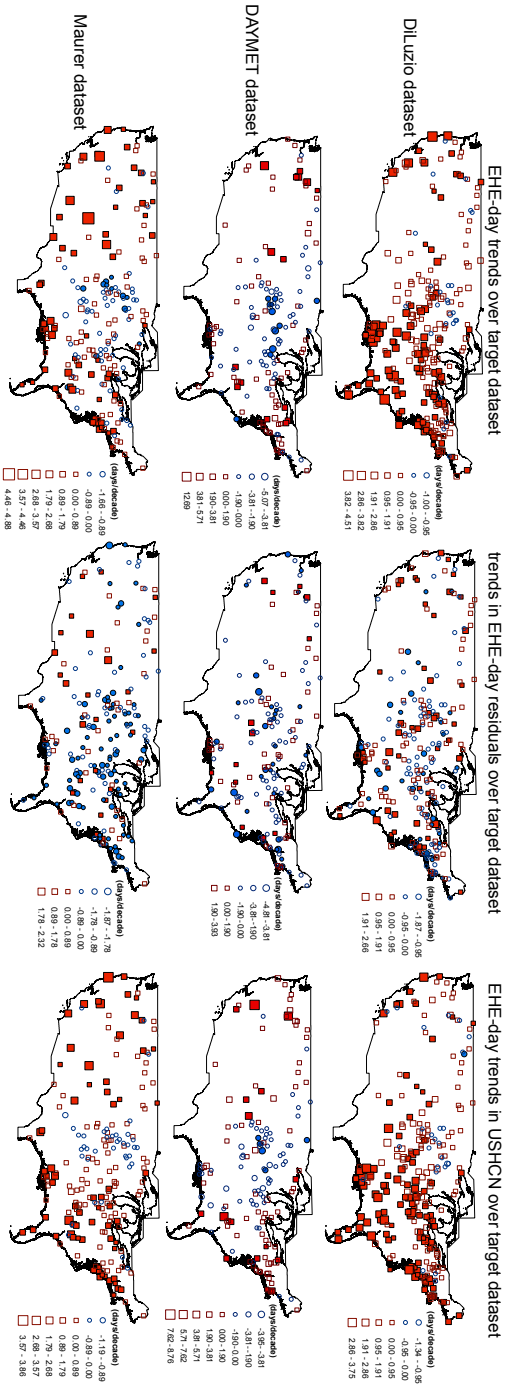


Figure 3.6. Decadal trends in the number of Tmin extreme heat event per summer during the 1970-2010 time period in the Maurer dataset. Plus signs within a grid cell signify statistically significant trends. Groupings are by color and signify standard deviations away from the zero value.

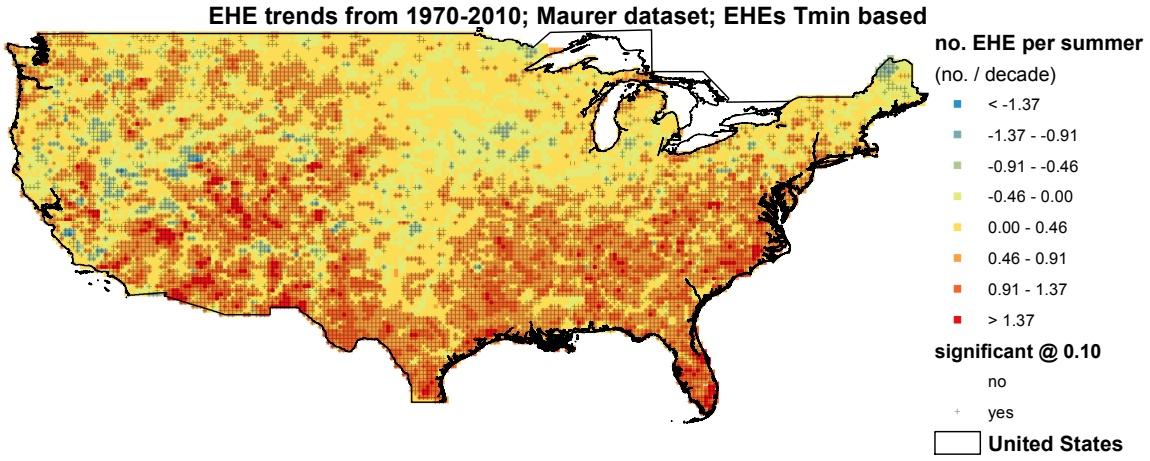


Figure 3.7. Decadal trends in the annual mean duration (in days) of Tmin extreme heat events during the 1970-2010 time period in the Maurer dataset. The symbology regarding color groupings and statistical significance is the same as figure 3.6.