Development of a Long-Term (1884-2006) Serially Complete Dataset of U.S. Temperatures and Precipitation for Climate Services

Jinsheng You[1]
Kenneth G Hubbard
Martha Shulski
High Plains Regional Climate Center, School of Natural Resources, University of Nebraska, Lincoln, NE

Mark Svoboda,
Michael J. Hayes
National Drought Mitigation Center, School of Natural Resources, University of Nebraska, Lincoln, NE

2015

Journal of Service Climatology

http://www.journalofserviceclimatology.org/

Volume 8, Number 1

A Refereed Journal of the American Association of State Climatologists

[1] Corresponding author address: Jinsheng You, High Plains Regional Climate Center, University of Nebraska, Lincoln, NE 68583-0988. Email: jyou2@unl.edu

Abstract

Serially complete climate datasets with no missing data are necessary for a diverse group of users working in many economic sectors. In this article we describe the procedures used to create a Serially Complete Data set (SCD) for the U.S. We include the selection criterion applied to potential SCD stations, the various procedural steps and the details applied to each step. A few observations that were not previously digitized were obtained from observers official paper reports. The methods used to estimate missing data are the Spatial Regression Test and the Inverse Distance Weighting technique. Using the criterion for selecting stations we were able to include 2144 stations for the SCD that had at least 1 element (maximum/minimum temperature and/or precipitation) for a continuous period of at least 40 years. In addition, the quality control procedure assigned confidence intervals to all observations and many of the estimates. We continue to explore the options for estimating any missing data that remain after our 3 step approach and we look forward to changing the base data set form TD 3200 to GHCN.

Introduction

***1.1 Problem Statemen*t.** Thousands of scientists have obtained official climate data for use in their analyses only to face the problem of how to address missing data. Although missing data is a common problem there is no common solution. This has led to inconsistent datasets from one study to another and usually no statement of the confidence or uncertainty in regard to the observations and estimates therein.

1.2 Review of Previous Research. In the area of information visualization, missing data can cause visualization failure or misleading interpretations of data (Eaton et al., 2003).The effect of missing data, i.e. data gaps, in the calculation of applications such as monthly mean temperatures can result in errors that exhibit temporal and spatial patterns (Stooksbury et al., 1999). Areal average precipitation is generally derived by a weighting of available stations. From year to year the available stations may increase or decrease. As stations begin or cease operations, the areal averages may show a step change (inhomogeneity) in the averages. To address this problem, McRoberts and Nielsen-Gammon (2011) addressed monthly precipitation data and arrived at a homogeneous Serially Complete Dataset (SCD) for the climate divisions of the U.S. (1895-present). Another example of an SCD was in association with the development of a National Drought Atlas in 1994 (Werick et al.) The National Drought Atlas was first completed in 1994 as an outcome of the National Study of Water Management conducted during the period November 1989 to October 1993. It used monthly precipitation totals from 1119 sites in the National Climatic Data Center's Historical Climatology Network. There have been two decades of additional data collected since the most recent drought atlas.

Demands for daily data in a serially complete dataset (SCD) are increasing in the climate community and among federal agencies as the result of needs in water management, environmental systems, and natural resource modeling (Eischeid et al., 2000; Chen et al., 2006). Missing data have a significant impact on agricultural decision support systems e.g. an SCD is required to calculate drought indices such as the Standardized Precipitation Index (McKee et al., 1993), the Palmer Drought Severity Index (PDSI, Palmer, 1965), or the Self-Calibrating PDSI (Wells et al., 2004). When daily data are missing, the above indices are either not calculated or are calculated by excluding the data gap. Both alternatives produce inaccurate indices and potentially lead to incorrect climate-related decisions.

The type of estimation and how it is implemented does make a difference. Chen et al. (2006) in the development of an SCD provided estimated daily values for maximum temperature (Tmax), minimum temperature (Tmin), and precipitation (PRCP). However, the parameters used in the spatial regression test (SRT, Hubbard et al., 2005) were based on an annual time frame It was shown byYou et al. (2006) that SRT estimates based on an annual time frame are less accurate than estimates derived from a shorter time period because the regression results between nearby stations will change with time. Short term weather variation owing to weather events such as cold front passing or abrupt temperature changes are better reflected in shorter time periods but, become smoothed out over a longer time frame In addition, the algorithm of Chen et al. (2006) can be improved by using more than 5 stations as demonstrated in You et al. (2006).

An earlier development of an SCD was accomplished by Eischeid and his colleagues for the western United States (Eischeid et al., 2000). However, their dataset was limited to the western U.S. and is labor intensive for applications. Eischeid et al. (2000) preselected stations and discarded any with a total of 48 months or more of missing data over the period 1951 to 1991. A month was taken as missing when more than 14 days were missing. It was also necessary to classify the stations into different categories based on the time of observation (TOB), which can vary from year-to-year even for a single station and unfortunately supporting documentation for metadata regarding the TOB is often incomplete. In addition, the partition of stations based on the TOB dilutes the spatial density leading to further difficulties in providing estimates for quasi isolated stations in a particular TOB category.

The increasing involvement of the insurance industry in weather disasters and anomalies brings requests directly from the insurance providers or indirectly from the cooperating research agencies concerning the availability of an SCD for climate. With the increasing availability of weather data (Hubbard et al., 2004) and the recent development of drought monitoring tools (Wilhite et al. 2005), an improved drought atlas was proposed to the U.S. Department of Agriculture's Risk Management Agency in 2005 (Wilhite, 2007) and the work presented here was undertaken to provide the data for this atlas.

1.3 Goal. The goal of this manuscript is to describe the procedures used to develop an SCD for the new drought atlas. The drought atlas is one process by which the historical weather data are turned into information in order to meet the demands of a wide range of stakeholders. The details of how the new SCD was created are provided in the following section.

Methods and Data

2.1 Data. We began by retrieving the ACIS: maximum (Tmax) and minimum (Tmin) air temperature, and precipitation (PRCP) for the stations that met our criterion. Fig. 1 shows the 2144 long term stations that were selected. This ACIS data was a copy of the official data at NCDC prior to their adoption of GHCN as the official daily dataset (Menne et al., 2012). Official data in the paper archives, were obtained through NCDC WSSRD and HPRCC paper archives, and digitized to fill gaps in actual observations. The remaining missing data were filled using spatial estimation procedures as described below. The keyed data was ingested and combined with the data retrieved from ACIS to form the *Base Dataset.* This study provided data estimation of missing data for all stations from the conterminous U.S.A. during the period of record for each element. Only stations with 40 or more continuous years of operation for at least one variable within the period 1884 to 2006 were included. The data were taken by the NOAA Cooperative Observer Program (COOP). Since the number of operating stations varies from year to year, daily data were retrieved and processed on an annual basis.

The missing data estimations procedures were made on the basis of the official release of COOP data, after undergoing quality control (QC) at the National Climatic Data Center (NCDC). The data that would not be identical to the TD 3200 NCDC archives are the estimates made herein by using the spatial regression test (Hubbard et al., 2005; You and Hubbard, 2007) and the estimates from the inverse distance weighted methods as described in the next sections. Currently NCDC and the regional climate centers archive the Tmax and Tmin in degrees Fahrenheit and PRCP in inches. To be consistent with the official reporting of this data in

English units and consistent with the official dataset, we use degrees Fahrenheit for temperature and inches for precipitation in this paper.

2.2 Spatial regression test (SRT). Stations, within a circle of 250 kilometers radius, centered on the station of interest, are selected, for each element, and a linear regression performed, for a window of time (*l*),for each surrounding station paired with the station of interest (Hubbard and You, 2005). For each surrounding station and for all daily observations in the selected window, regression based estimates are formed. If the missing number of days (*m*) represents less than 50% if the window length, the spatial regression test (Hubbard et al., 2005) is used to estimate the missing value as described below

$$x_i = a_i + b_i y_i, \tag{1}$$

where y_i is the particular measurement (e.g. Tmax or Tmin) at the *i*th surrounding station, x_i is the regressed intermediate estimate for the station of interest based on the surrounding station y_i, and the parameters a_i and b_i of the linear regression function for the regression window of a specific length (*l*). The weighted estimate (x') is derived by utilizing the standard error of estimate (*s*) also termed the Root Mean Square Error (RMSE), in the weighting process. The estimates are obtained from the following equation (You and Hubbard, 2007):

$$x' = \sum_{i=1}^{N} x_i s_i^{-2} / \sum_{i=1}^{N} s_i^{-2} \ . \tag{2}$$

N is the number of stations within the circle that have R^2 greater than 0.5 and *N* is limited to the 15 nearest stations to limit computations. If $N < 2$, no SRT estimates are calculated in this study.

The weighted error of estimate (s') is calculated in accordance with You and Hubbard (2007) as shown in (3) below:

$$1/s'^{2} = N^{-1}\sum_{i=1}^{N} s_i^{-2} \ . \qquad (3)$$

To account for possible systematic time shifting of observations (this occurs when an observer consistently records his observation on the day before or after the actual date of observation) the surrounding station's data are each shifted by ± one day and the regression repeated. The shift (-1, 0, +1) that results in the lowest error of estimate is then used in (2) and (3). This process allows for the stations with different times of observations to be intermixed in the analysis without causing a systematic bias. The estimated confidence intervals are based on s' and we test whether or not the station value (x), when it exists, falls within the confidence intervals at f = 4 For normally distributed data we expect there to be no data beyond f=4 in (4) below. Thus when the following test is failed we assume the datum is an outlier and we write the outcome in a set aside table.

$$x' - fs' \leq x \leq x' + fs' \ . \qquad (4)$$

Our procedure merely tests the data and writes the details in a set aside table but, does not replace the data in the base *dataset*. Unlike distance weighting techniques, the SRT minimizes systematic differences between station data (the coefficients a_i and b_i remove the systematic bias). In our analysis we use an *f* value of 4.0 to limit the number of Type II flags. This is particularly important in the case of unique events such as cold fronts and hurricanes (You and Hubbard, 2006

2.3 Inverse distance weighting method. The inverse distance weighting (IDW) method (You, et al., 2008) was employed when data were insufficient to meet the SRT requirements. The IDW estimates are made based on the assumption that surrounding stations should receive more weight if they lie in closer proximity to the target station than other neighbors. This estimate is given by (5)

$$\hat{x} = \sum_{i=1}^{n} \left[y_i w_i \right] / \sum_{i=1}^{n} w_i \ , \qquad (5)$$

where $\hat{x}$ is the predicted variable, y_i is the particular measurement (e.g. Tmax, Tmin, or PRCP) at the *ith* surrounding station and the weighting function w_i is defined as the inverse of the distance between the target station and the *ith* surrounding station. The fifteen nearest stations are used (n=15), a number which has proven to provide satisfactory results (Hubbard and You, 2005). The RMSE only reduces by 0.1 F. when 30 stations are used instead of 15 and this reduction is below the measurement precision. Bias's that exist between the target station and the surrounding stations are not removed when IDW is used.

2.4 Step-wise Procedure Used to Develop this SCD. The following restrictions were imposed on the data estimation process, 1) Any data prior to the start or falling after the end of station operation, were not estimated, 2) For periods when an element was not observed (e.g. some times Tmax and Tmin were not measured and a station was collecting only precipitation), that element was not estimated, 3) Stations were not included if the length of data was less than 40 years or there were more than two months of continuous missing data. After the station selection procedure our list contained: 2144 stations with precipitation observations of which 1705 observed temperature.

To develop an SCD, three steps were adopted, as shown in Table 1. The data filling procedures were implemented on a yearly basis using all available observations for the individual year even those stations not in the SCD list. The data for the missing values were filled as described below.

2.5 Estimation of missing data. The estimation of missing data was accomplished in 3 steps as shown in Table 1. We began with Step 1 (see Table 1) and used the *Base Dataset* as input and the SRT to estimate missing data in independent windows (*l*=1-60, 61-120,…, 300-365). For the station being examined, the other stations within 250 km were potential estimators however if the R^2 for the comparison between data from the station and the potential estimator was less than 0.5 then that potential estimator station was not included in the estimation process. Additionally, if the days (*m*) with missing data represent more than ½ the window length *(l)*, the missing data is not estimated in this 1st step. The data successfully estimated were added to the *Base Dataset* to obtain an intermediary dataset that we describe as Dataset 1.

Dataset 1 then became the input dataset for Step 2. In this step the SRT method of estimation was again used, however in step 2 the window (*l*) was equal to 1-365 with the exception of leap years where *l was* 1-366. The estimates from Step 2 were then added to Dataset 1 to create another intermediate dataset, Dataset 2.

In Step 3 we used IDW to estimate missing data in Dataset 2. The IDW was performed on a daily basis using no more than 15 of the closet stations within a 250 km radius of the target station. There is no R^2 involved in IDW and if there are fewer than 2 surrounding stations the estimation is not performed. When the IDW estimates were added to the Dataset 2, the result is the final long term Serially Complete Dataset (SCD).

Both maximum and minimum daily temperature data were analyzed as shown in Table 1. When step 1 and step 2 were performed, the actual observations were subjected to the test in (4) and the statistical information for all values were placed in a set aside table for future reference. Because precipitation is not as strongly correlated from station to station, only Step 3 was used to estimate missing precipitation data.

Quality control procedures were also carried out in Step 1 and Step 2. Those values that failed the spatial regression test were placed in a set-aside table for optional use.

2.6 Organization of the SCD. When the data estimation processes were complete for all years from 1884 to 2006, the data were reorganized for the 2144 stations. The output consisted of the final SCD in two formats: the first was organized by station and the second by year.

3. Results

Table 2 shows a summary of the dataset with respect to the number of missing values and the roles of the different data estimation methods with respect to estimation of missing values. For example, for Tmax there was a possibility of 55,913,565 values at the 1705 stations. The fractions of data filled were 0.054, 0.530, 0.192, and 0.224 for key-in, Step 1, Step 2, and Step 3 respectively. A comparison of different window lengths for SRT on the RMSE is shown in

Figure 2. Consistent with the findings for single stations in Hubbard and You (2005) the SRT estimation using data for shorter windows gave better overall estimates than those with longer windows (i.e. the performance of the estimates increases with number of windows per year, up to 12 window. As an example we note that the proportion of stations meeting RMSE of 2 or less is about 0.7 for an annual window but about 0.75 for a 6 window per year case. About 97% of the estimates have an RMSE < 3 F for Step 1 while only 92% of the estimates have an RMSE < 3F for Step 2.

Of the total filled data, 5.2 and 5% data were keyed in from the paper archives for Tmax and Tmin, respectively (See Figure 3 and Table 2). Step 1 filled 53.0 and 51.5% of the total number for Tmax and Tmin, respectively. Step 2 estimated another 19.2% and 21.1% of missing data for Tmax and Tmin, respectively. Step 3 estimated 22.4% and 22.1% of the total filled data for Tmax and Tmin, respectively. The keyed data consists of only 2.8% of the filled precipitation values and the IDW estimated the other 97.2% of missing precipitation data.

Figure 4 plots the number of valid measurements, missing data, and the number of estimates obtained during each step for Tmax during the period 1884-2006. In addition to adding new stations, air temperature sensors were continuously added to existing stations that previously were PRCP only in the early 1960’s. Although all 2144 stations were in operation before 1988, some were precipitation only. In 2006 all these long-term stations were recording air temperature. Most of the missing data for Tmax and Tmin can be filled by IDW estimates; however, for this version we chose not to fill missing data when the air temperature sensors were not in operation.

The histogram of station operation for Tmax is shown in Figure 5. The total stations represented are 1705. The histogram for Tmin is similar.

4. Summary and Conclusions

The new SCD was produced for applications in the development of a national drought atlas with criteria for a long-term (at least 40 years) continuous (no data gaps longer than two months) dataset of Tmax, Tmin, and/or PRCP for a total of 2144 stations over the period 1884-2006. The distribution of data length of these stations was shown in Figure 5. The missing values in the original dataset retrieved from ACIS were filled with the data keyed from official paper records and the estimates using the SRT and IDW methods.

After producing the suite of estimates that we describe in Table 1, we still have a few missing values because the conditions for calculating our estimates were not met (e.g. there were no stations or only one station within 250 km). We will select one of the following options for providing these remaining estimates: direct substitution of the value from the nearest neighbor; an estimate from the SRT when only one neighboring station is within 250 km, estimation with R extended to 300 km; etc.

Additionally, where data can be estimated by the SRT procedure the estimates were assigned a confidence level. The observations in ACIS were the TD 3200 at the time and since the official dataset is now the GHCN, our future work will include revising the SCD on the basis of GHCN data. To our knowledge this SCD is the most extensive daily dataset available for users including decision makers, scientists, and insurance industries.

This is the first serially complete dataset where a statement of confidence can be associated with many of the estimates, i.e. SRT estimates. The RMS for maximum and minimum temperature is less than 1F in most cases and thus we are 95% confident that the value, if available, would lie between ±2F of the estimate. This dataset is available for interested parties and the directions for retrieving the dataset can be obtained by contacting Dr. Jinsheng You at jyou2@unl.edu Probabilities related to extreme rainfall for flooding and erosion potential can be derived along with indices to reflect impact on livestock production. The data is also of potential use in crop models, and in the assessment of severe heat, cold, and dryness. The dataset is offered as an option to distributing the official data to the users who need this level of spatial and temporal coverage but are not well positioned to spend time and resources on filling missing data gaps with acceptable estimates.

Analysis based on the long-term dataset will best reveal the regional and large scale climatic variability in the continental U.S., which affords an ideal dataset for the development of a new drought atlas and associated drought index calculations. Future data observations can be easily appended to this SCD with the procedures described herein.

5. Reference:

Bulley, H., D. Marx, J. Merchant, J. Holz and A. Holz. 2008. A Comparison of Nebraska Reservoir Classes Estimated from Watershed-Based Classification Models and Ecoregions, Journal of Environmental Informatics, **11**(2), 90-102.

Chen, Z., Goddard, S., Hubbard, K.G., Sorensen W.S. and You, J. (2006). A Serially Complete U.S. Dataset of Temperature and Precipitation for Decision Support Systems. Journal of Environmental Informatics **8** (2): 86-99.

Eaton, C., Plaisant, C. and Drizd, T. (2003). The challenge of missing and uncertain data, in *Proc. IEEE InfoVis Poster Compendium 2003*, IEEE Computer Society Press, pp. 40-41.

Eischeid, J.K., Baker, B.C., Karl, T.R. and Diaz, H.F. (1995). The quality control of long-term climatological data using objective data analysis. *J. Appl. Meteorol.*, **34**(12), 2787-2795.

Eischeid, J. K., Pasteris, P.A., Diaz, H.F., Plantico, M.S. and Lott, N.J. (2000). Creating a serially complete, national daily time series of temperature and precipitation for the western United States. *J. Appl. Meteorol.*, **39**(9), 1580-1591.

Hubbard, K.G. (2001). Multiple station quality control procedures, in *Automated Weather Stations for Applications in Agriculture and Water Resources Management*, World Meteorological Organization, AGM-3 WMO/TD, 1074.248.

Hubbard, K.G., Goddard, S., Sorensen, W.D., Wells, N. and Osugi, T.T. (2005). Performance of quality assurance procedures for an applied climate information system. *J. Atmos. Ocean. Technol.*, **22**(1), 105-112.

Hubbard, K.G. and You, J. (2005). Sensitivity analysis of quality assurance using spatial regression approach-A case study of the maximum minimum air temperature. *J. Atmos. Ocean. Technol.*, **22**(10), 1520-1530.

McRoberts, D. B., and J. W. Nielsen-Gammon, 2011: A new homogenized United States climate division precipitation data for analysis of climate variability and change. J. Appl. Meteor. Clim., **50**, 1187-1199, doi:10.1175/2010JAMC2626.1

Menne, M.J., I. Durre, R.S. Vose, B.E. Gleason, and T.T. Houston. 2012. An overview of the Global Historical Climatology Network daily database. J. Atmos. And Ocean. Tech. **29**::897-910. doi: http://dx.doi.org/10.1175/JTECH-D-11-00103.1

Stooksbury, D.E., Idso, C.D., and Hubbard, K.G. (1999). The effects of data gaps on the calculated monthly mean maximum and minimum temperatures in the continental United States, A spatial and temporal study. *J. Climatol.*, **12**: 1524-1533.

Werick, W. J., G. E. Willeke, N. B. Guttman, J. R. M. Hosking, J. R. Wallis (1994), National drought atlas developed, *Eos Trans. AGU*, **75(8):** 89.

Wilhite, D. (2007), Preparedness and Coping Strategies for Agricultural Drought Risk Management: Recent Progress and Trends. **In:** Managing Weather and Climate Risks in Agriculture. Springer Berlin Heidelberg. **2**: 21-38.

Wilhite, D., Svoboda, M. and Hayes, M. (2005). Monitoring Drought in the United States: Status and Trends. In: Boken, V.J., Cracknell, A.P. and Heathcote, R.L. (eds). Monitoring and Predicting Agricultural Drought: A Global Study. pp. 121-131. Oxford University Press. New York.

You, J. and Hubbard, K.G. (2006). Quality Control of Weather Data during Extreme Events. Journal of Atmospheric and Oceanic Technology **23**(2): 184–197.

You, J., and Hubbard, K.G. (2007). Relationship of Flagging Frequency to Confidence Intervals in the Statistical Regression Approach for Automated Quality Control of Tmax and Tmin. International Journal of Climatology. 27(9): 1257.

You, J., Hubbard, K.G. and Goddard, S. (2008). Comparison of Methods for Spatially Estimating Station Temperatures in a Quality Control System. International Journal of Climatology, 28: 777–787. doi: 10.1002/joc.1571

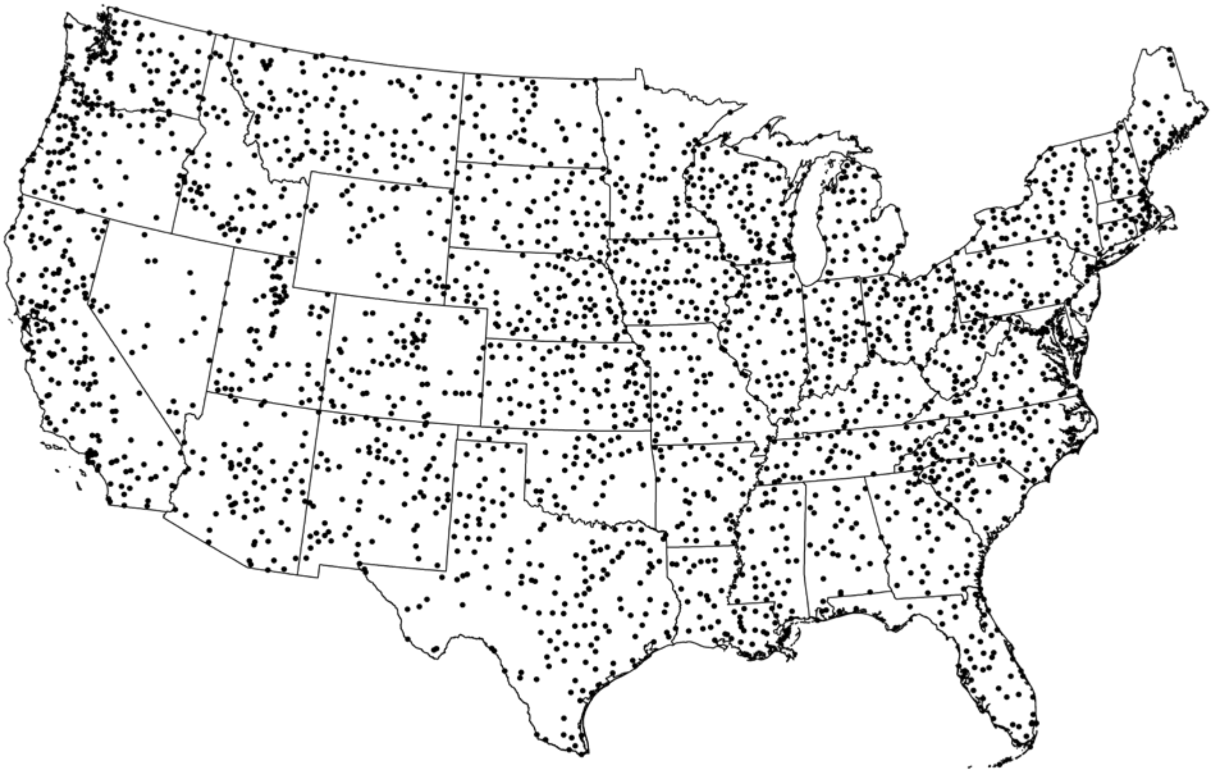

Figure 1. This map shows the spatial distribution of long term SCD stations (1884-2006)with period of record greater than 40 years.

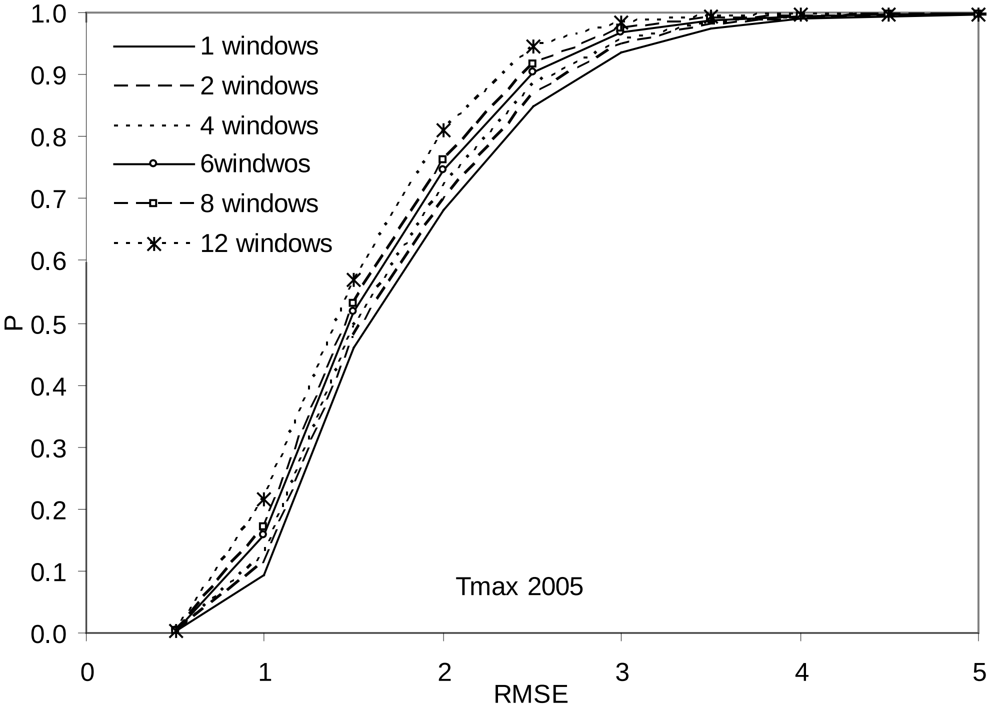

Figure 2 The proportion (P) of stations that have RMSE less than the value shown for multiple windows/year in 2005.

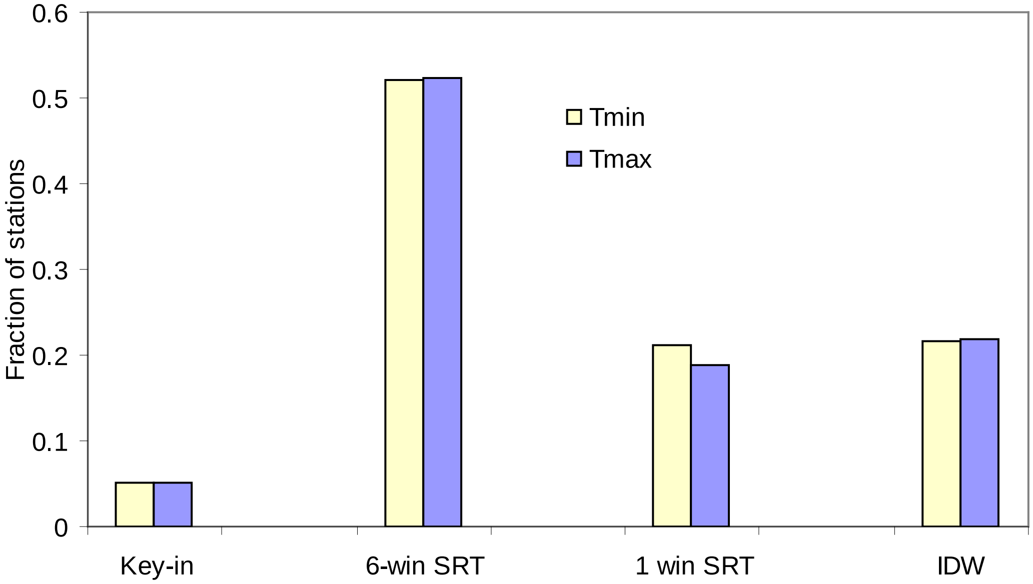

Figure 3. The fraction of filled ‘missing’ data using different sources and methods.

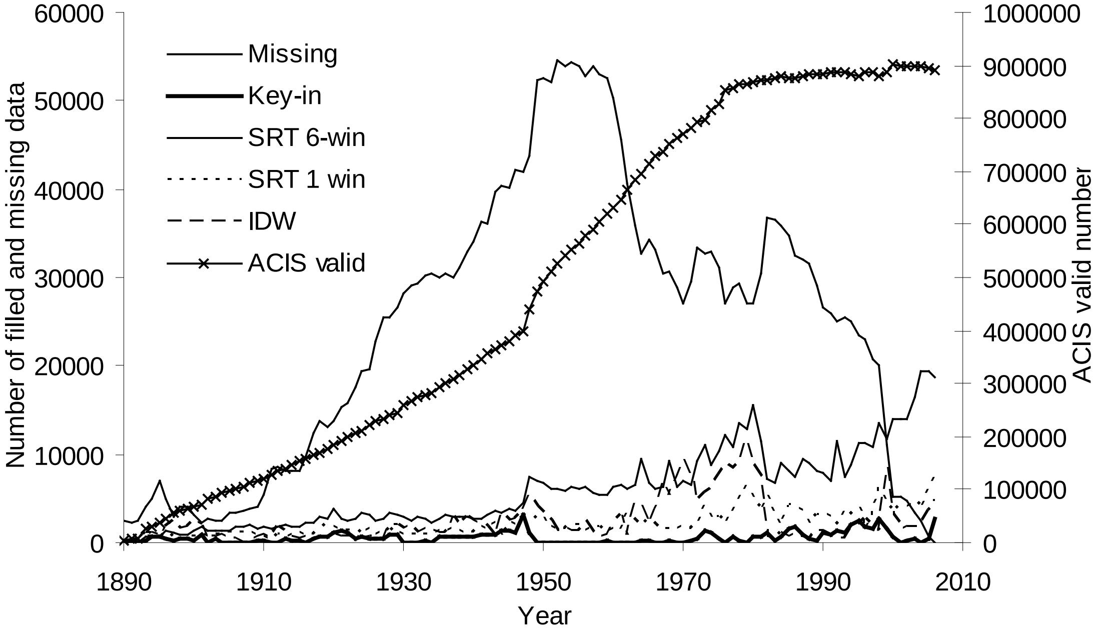

Figure 4. Numbers of valid data, missing values, and the number of estimates obtained from different steps through time. Step 1 involved 6 windows, Step 2 involved 1 window and IDW is implemented on each day independently.

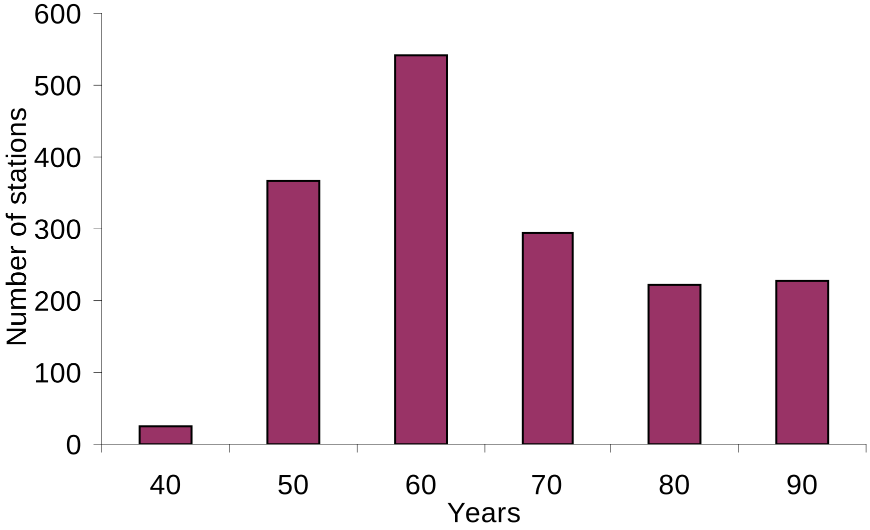

Figure 5. The distribution of data length for stations with maximum temperature in this SCD. The stations shown here together with the precipitation only stations totals 2144.

Table 1. The type of estimation and associated parameters are shown for each step of the data estimation. N is the number of station within 250 km up to 15 and m is the number of missing data values allowed at the candidate station. The window length is *l*.

Step	**Input Dataset**	**Estimator**	**Window l(in days)**	**Output Dataset**	**Up to 15 Stations within 250 km**	
1.	*Base Dataset*	SRT, *f*=4	60	Dataset 1	Best R^2 but ≥0.5	2≤N≤15 *m*<*l*/2
2.	Dataset 1	SRT, *f*=4	365(366 in leap years)	Dataset2	Best R^2 but ≥.5	2≤N≤15 m<*l*/2
3.	Dataset 2	IDW	1	Final SCD	No R^2 test	N≥2

Table 2. Statistics for the 2144 selected stations for the drought atlas project. Total Number (TN) is the potential number of days in the SCD or sum of all station days, TF=TN-ACIS-Missing Table and TF=Key-in + 6win SRT +1win SRT+IDW.

	Total Number (TN)	Missing	ACIS	Total Filled (TF)	Key-in	6win SRT	1win SRT	IDW
Tmax	55913565	2718144	52054393		61771	604527	218791	255939
	Ratio to TF			1141028	0.054	0.530	0.192	0.224
	Ratio to TN	0.049	0.930	0.020	0.001	0.010	0.004	0.005
Tmin	55913565	2715240	52036755		61124	597725	245555	257166
	Ratio to TF			1161570	0.053	0.515	0.211	0.221
	Ratio to TN	0.049	0.931	0.021	0.001	0.011	0.004	0.005
PRCP	55913565	1382	55026831		24820			860532
	Ratio to TF			885352	0.028			0.972
	Ratio to TN	2.47E-05	0.984	0.016	0.0004			0.015