INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **30**: 601–611 (2010)
Published online 9 April 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/joc.1917

RMetS
Royal Meteorological Society

An error analysis for the hybrid gridding of Texas daily precipitation data

Andreas J. Rupp, Barbara A. Bailey, Samuel S.P. Shen,* Christine K. Lee and B. Scott Strachan

Department of Mathematics and Statistics, San Diego State University, San Diego, CA 92182, USA

ABSTRACT: This paper reports the error analysis results for the gridded daily precipitation data over the state of Texas of the United States from January 1, 1901 to December 31, 2000. The Global Daily Climatology Network dataset is used for both the data gridding and error analysis. The station data have been interpolated onto a $0.2^\circ \times 0.2^\circ$ grid which starts at the base point (25°50′N, 106°38′W). The data gridding approach is a hybrid method, which is a blend of two simple methods: inverse distance weighting and nearest station assignment. Our gridding results are compared with those obtained by other gridding methods. The cross-validation method is used for the error analysis. Our error analysis of the interpolated products includes not only the conventional errors, such as the mean bias error, but also the probabilistic distribution of the relative errors of precipitation frequency and the spatial distribution of a major Texas historical storm. The following results have been found: (1) a simple arithmetic average of station data usually overestimates Texas' average precipitation by 2.4 mm per day, (2) the relative error of the precipitation frequency follows a lognormal distribution, and (3) the hybrid gridding data do not have obvious bias and can reasonably display storm-covered areas in Texas. The gridded data and error results are useful for the validation of climate models, calibration of satellite borne remote sensing devices, and numerous agricultural and hydrological applications. The statistical methods of our analysis and some of our results are applicable to other regions of the world. Copyright © 2009 Royal Meteorological Society

KEY WORDS daily precipitation; error analysis; data gridding; hybrid method

Received 22 February 2008; Revised 4 March 2009; Accepted 8 March 2009

1. Introduction

Gridded daily precipitation data have numerous applications, ranging from climate model validation and soil quality modelling to drought modelling and climate change studies such as the assessment of agroclimatic changes of the Province of Alberta, Canada (Shen *et al*., 2005). Many gridded datasets have been produced by using various kinds of methods for different purposes. A few examples are the daily gridded dataset with a coarse resolution for the South America data (Liebmann and Allured, 2006), global daily dataset (Jolly *et al*., 2005; Chen *et al*., 2008), Alberta daily dataset (Shen *et al*., 2001), hourly gridded dataset for the contiguous United States (Higgins *et al*., 2000), daily US dataset (Thornton *et al*., 1997), and the global daily dataset based on both *in situ* and remote sensing data (Huffman *et al*., 2001). With the existing methods and datasets, there is a need for detailed error analysis and method dissection from different perspectives. Ensor and Robeson (2008) examined the errors of the US Midwest daily precipitation data gridded by a modified Cressman method and considered precipitation frequency, annual maximum of daily precipitation, and return values of extremes with a given time period. Our current study examines the Texas daily precipitation data gridded by a hybrid method that blends the inverse distance weighting (IDW) method with the nearest station assignment (NSA) method, and considers probability density function (pdf) of the precipitation frequency, the spatial distribution of a major storm, and the bias of regional precipitation totals due to improper gridding or averaging methods.

As reviewed in Shen *et al*. (2001) and Jolly *et al*. (2005), due to large spatial and temporal variances of daily precipitation data, a single conventional mathematical method of interpolation and extrapolation appears not to be able to effectively grid the daily station data onto a grid. These conventional methods include the spectral method of orthogonal polynomials (Dunkl and Xu, 2001), kriging (Gandin, 1963; Cressie, 1993), spline fitting (Wahba, 1990), multi-variate regression (Johnson and Wichern, 1992), empirical orthogonal function (EOF) reconstruction (Smith *et al*., 1996, 2005; Shen *et al*., 2004), and the Cressman smoothing method commonly used in meteorology via IDW with a spatial length scale. Many daily precipitation data applications, such as drought monitoring, validation of climate models, and ground truth calibration of satellite borne remote sensing devicestropical rainfall measuring mission (TRMM)

* Correspondence to: Samuel S.P. Shen, Department of Mathematics and Statistics, San Diego State University, San Diego, CA 92182, USA.
E-mail: shen@math.sdsu.edu

(Simpson *et al*., 1988; North *et al*., 1994), require several facets of information. These include precipitation frequency, intensity, spatial and temporal variations, as well as the total precipitation amount over a region in both daily and longer time scales. While the classical methods may provide adequate total precipitation amount in a longer time scale, like a month (Xie and Arkin, 1997; Chen *et al*., 2002), its direct application to daily data can be problematic (Robeson and Ensor, 2006; Xie *et al*., 2007). Robeson and Ensor pointed out that a hybrid of multi-methods appears to be the sensible way to resolve this precipitation interpolation problem. For example, a hybrid of two methods may be arranged in such a way that one method determines whether a grid point is dry (i.e. zero precipitation when the measurable daily precipitation is less than the threshold of 0.2 mm) or wet (i.e. non-zero daily precipitation greater than 0.2 mm), while another method determines the amount of precipitation for the day at the grid point. Here, 0.2 mm is a commonly used threshold of minimum measurable precipitation (Stensvand and Eikemo, 2005). Some hybrid methods have already been developed. Shen *et al*. (2001) developed a hybrid method based on a regression against the data of the nearest station. Thornton *et al*. (1997) also developed a method that can also be considered as a hybrid method. It first calculates the dry-wet indicator by interpolating a binary indicator of dry or wet at surrounding stations (wet = 1 and dry = 0) by using a truncated Gaussian filter. If the interpolated indicator is greater than a threshold value (equal to 0.52 in their paper), the grid point is considered wet. One then proceeds to find the precipitation amount for the grid point also by using the Gaussian filter but with a regression correction with respect to an elevation factor. Jolly *et al*. (2005) compared the ordinary kriging method and the IDW method to the truncated Gaussian filtering method and found that the mean absolute error (MAE) and mean bias error (MBE) on the spatially averaged total are comparable for all the three methods, although the errors from the simple IDW are slightly smaller. This agrees with the literature review of Shen *et al*. (2001) and their numerical experiments. Namely, the simple IDW method is likely the best approach to render good results for precipitation amount, when the more detailed dynamic relationships with elevation, wind, and other factors are not well parameterized.

Our Texas data gridding uses the hybrid method of Shen *et al*. (2001), which is a hybrid of two simple methods: IDW and NSA (Griffith, 2002; Rupp, 2007). Our error analysis of the interpolated products includes not only the conventional root mean square error (RMSE), MAE, and MBE but also the probabilistic distribution of the relative errors of precipitation frequency and the spatial distribution of a major historical storm. The MBE is a very basic check of the quality of a gridding product. A reasonable product necessarily has its MBE approximately equal to zero. The RMSE and MAE measure the error variance, which is usually an increasing function of the spatial variance of precipitation and is also related to precipitation frequency and temporal extremes. Although we have calculated the RMSE, MAE, and MBE as were done in Shen *et al*. (2001) and Jolly *et al*. (2005), the focus of this paper is on precipitation frequency error characteristics and the reduction of the overestimate bias of the simple spatial average. This overestimate bias resulted from the simple average of station data for Texas' spatially averaged precipitation can be 10% higher than a more accurate method. Using the cross-validation method, we have found that the relative error of the precipitation days in the hybrid interpolation is log-normally distributed and that the gridded data can help remove the 10% overestimate bias. Our assessment of the bias due to an improper averaging or gridding method can be applied to any region where a clear uneven spatial distribution of precipitation and gauges exists. Our pdf method of assessing the errors of precipitation frequency provides an approach to understand the source and range of errors and is helpful to the development of more advanced hybrid methods for reducing the errors, which has a procedure similar to Ensor and Robeson (2008) assessment of precipitation frequency errors by using three categories: light, moderate, and heavy precipitations.

The contents of this paper are arranged as follows. Section 2 describes the data and the method for our analysis. Section 3 describes the results of both gridding and the error analysis. Section 4 contains conclusions and discussion.

2. Data and interpolation method

2.1. Data

The daily station data are from the Global Daily Climatology Network (GDCN), Version 1.0, from the United States National Climatic Data Center (Gleason, 2002). The GDCN dataset includes daily precipitation data from 32 857 stations on the entire globe from March 1, 1840 to November 30, 2001. The data were collected by different countries and organizations and have gone through various data quality assurance reviews, such as bounds and climatological outlier checks. Thus, we apply no further quality control procedures to the data, although some problems can still exist in the data quality (Durre *et al*., 2008).

The daily precipitation's spatial length scale is around 40–60 km in the Texas and Oklahoma areas (Graves *et al*., 1993; North and Nakamoto, 1989). We want our gridded data to catch the mesoscale weather events and atmospheric motions so that the gridded data can be used for drought and flood monitoring, weather forecasting, and weather simulations. We choose a $0.2^\circ \times 0.2^\circ$ grid, starting at 25°50′N, 106°38′W, which is the southwest corner of the latitude–longitude box (25°50′N–36°36′N, 106°38′W–93°38′W) that covers Texas (Figure 1). All of the 1555 stations inside this box that had at least one record are naturally included for our interpolation. Among these 1555 stations, 975 stations are within Texas.

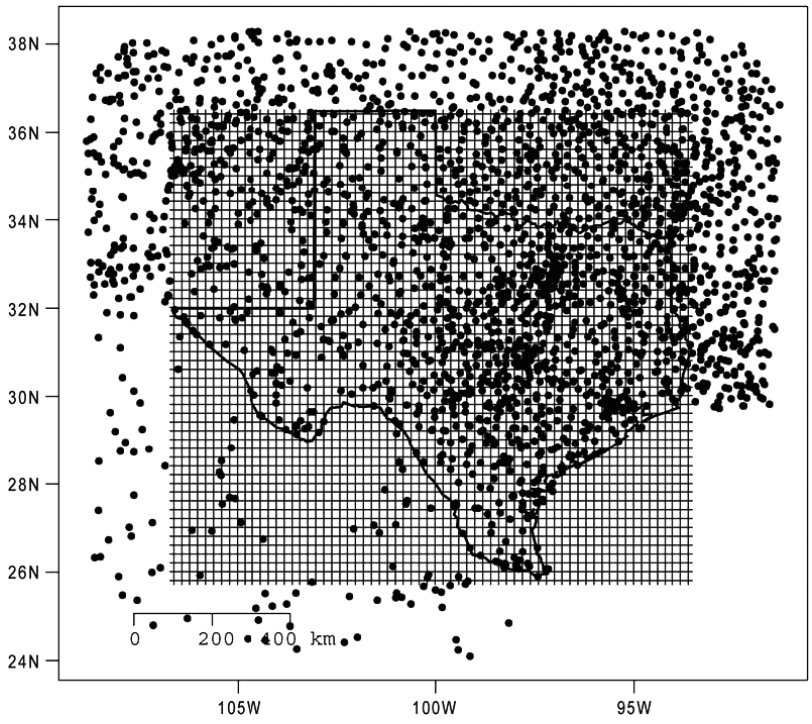

Figure 1. Locations of the 2401 stations used for data interpolation for Texas.

Since the stations outside of Texas but nearby the Texas boundary can be highly relevant to the precipitation measurement in the border regions of Texas, the stations within 200 km from the latitude–longitude box (25°50′N–36°36′N, 106°38′W–93°38′W) are included in our interpolation. The total number of stations in this extended region is 2401. Each of those stations had at least one record in the entire interpolation period. Figure 1 shows the locations of all the stations used in our interpolation. The figure indicates that the stations are unevenly distributed with low station density over western Texas, a dry area with an annual precipitation less than 500 mm, and high station density over eastern Texas, which has a high precipitation frequency and an annual precipitation as high as 2000 mm due to the highly moist air from the Gulf of Texas. Eastern Texas also has a greater population and more agriculture. It is thus natural for eastern Texas to have a more dense observational network for precipitation than western Texas.

Of course, not all the 2401 stations had data throughout the entire period from January 1, 1901 to December 31, 2000. Most stations did not start until after World War II, and many stations did not last for long. The number of stations that had data on a given day is shown in Figure 2, which is uneven in time, with about 50 stations in 1901 and nearly 1000 stations since World War II. The number of stations increased slowly from about 50 in 1901 to about 350 before World War II. An abrupt increase of the number of stations occurred right after World War II, and the number became close to 1000 in 1946. Thus, the quality of precipitation observation improved dramatically after World War II. Due to the availability of the station data, we consider our gridded data to be reliable after 1946. When using the gridded data before and during World War II, one needs to carefully consider the interpolation error and the application standards.

2.2. Methodology of data gridding

The gridding procedure used here is adopted from the hybrid method of Shen *et al*. (2001). This hybrid method employs the IDW method to calculate the monthly total of a grid point from the daily station data and uses the NSA method to downscale the monthly total into daily precipitation. The detailed procedures are given below.

The daily data of the eight nearest stations within 60 km from a grid point $\vec{g}_j$ are interpolated to the grid point $\vec{g}_j$ by the IDW formula:

$$\hat{G}_j(t) = \left(\sum_{i=1}^{M_j(t)} \frac{1}{d_{ij}} \right)^{-1} \sum_{i=1}^{M_j(t)} \frac{S_i(t)}{d_{ij}}, \qquad (1)$$

where $S_i(t)$ is the station data at station $\vec{s}_i$ and on day t, d_{ij} is the great circle distance between the grid point $\vec{g}_j$ and the station location $\vec{s}_i$, and $M_j(t)$ is the number of stations used for interpolation for the grid point $\vec{g}_j$ and day t. The stations $\vec{s}_i$ are sorted in an ascending order of the distance sequence d_{ij} for $i = 1, 2, \cdots, M_j(t)$. For station-sparse areas, there may be less than eight stations within 60 km from the grid point $\vec{g}_j$ on a given day t. For an extremely station-sparse area, it can happen that there is no station within 60 km from the grid point $\vec{g}_j$ on a day t, then the station nearest to the grid point is taken, and $M_j(t) = 1$. In that case, the precipitation value from the nearest station is assigned to the grid point. This is the case for many grid points during the earlier period of the 20th century. For example, on

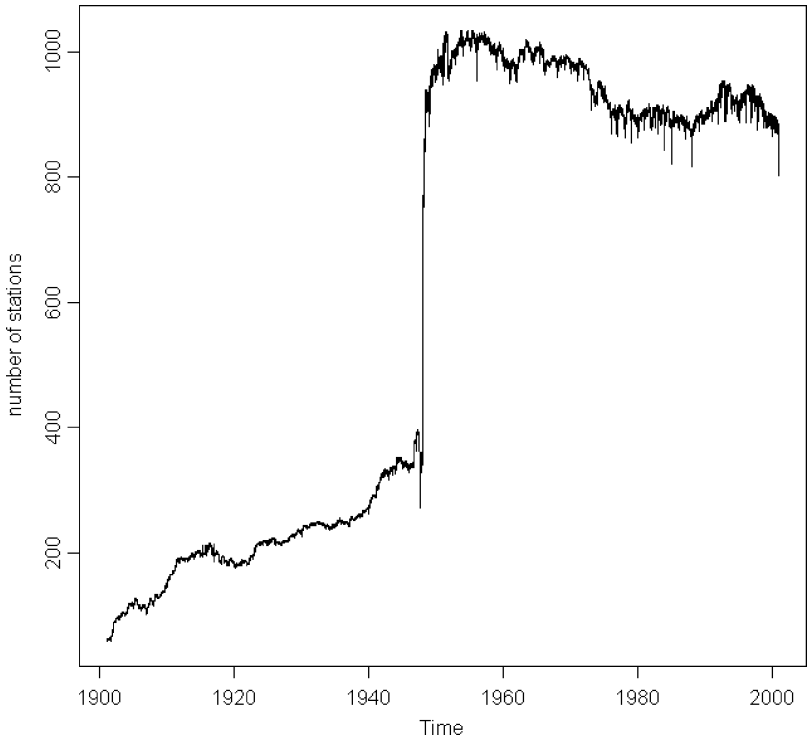

Figure 2. The history of the number of stations inside Texas shown in Figure 1.

July 7, 1901, Texas had only 46 stations. The station nearest to the grid point (29°38′N, 103°38′W) is located at 30°8′N, 102°23′W and the distance between the station and the grid point is 133 km. Many grid points have their distances to the nearest stations greater than 60 km. As the distance between a grid point and the nearest station is large, the interpolation accuracy level will be reduced. Therefore, the value range of $M_j(t)$ is between one and eight. The 60-km length scale is adopted as the upper boundary of the spatial length scale of daily precipitation (Huff and Shipp, 1969; Graves *et al*., 1993).

If the grid point $\vec{g}_j$ is on a station, then $d_{ij} = 0$. We naturally assign

$$\hat{G}_j(t) = S_1(t) \tag{2}$$

The monthly total precipitation at the grid point $\vec{g}_j$ is calculated from the daily values

$$\hat{G}_{M,j}(m) = \sum_{t=1}^{M_m} \hat{G}_j(t) \tag{3}$$

where M_m is the number of days for the month m, and is equal to 31 for January, March, May, July, August, October, and December, 30 for April, June, September, and November, and 28 for February, but 29 for a leap year's February.

Apparently, the NSA is a special case of the IDW when the number $M_i(t)$ of stations in Equation (1) used for interpolation is limited to one. We will use the NSA to interpolate the dry-wet index, which is equal to one if it has precipitation greater than 0.2 mm, and zero otherwise, where 0.2 mm is a commonly used threshold of a rainy day and is the resolution of the traditional standard cylinder rain gauge, developed about 100 years ago.

The hybrid method of Shen *et al*. (2001) is to downscale the monthly total precipitation $\hat{G}_{M,j}(m)$ on the grid to the daily scale by using the nearest station as the dry-wet indicator. The mathematical expression is:

$$R_j(t) = \frac{\hat{G}_{M,j}(m)}{S_{M,\text{nearest}}(m)} \times S_{\text{nearest}}(t) \tag{4}$$

where $S_{M,\text{nearest}}(m) = \sum_{t=1}^{M_m} S_{\text{nearest}}(t)$ is the monthly total precipitation of the station nearest to the grid point. The nearest station for a grid point may not be the same station since the nearest station in some days of the month may not have data, and hence the nearest station will be the next nearest station that does have data. As discussed in Shen *et al*. (2001), Equation (4) conserves the water mass since the sum of the two sides of the equation throughout all the days of the month is equal to $\hat{G}_{M,j}(m)$, which is the monthly total precipitation at the grid point obtained from the IDW interpolation method, i.e.

$$\begin{aligned}\sum_{t=1}^{M_m} R_j(t) &= \frac{\hat{G}_{M,j}(m)}{S_{M,\text{nearest}}(m)} \times \sum_{t=1}^{M_m} S_{\text{nearest}}(t) \\ &= \frac{\hat{G}_{M,j}(m)}{S_{M,\text{nearest}}(m)} \times S_{M,\text{nearest}}(m) \\ &= \hat{G}_{M,j}(m)\end{aligned} \tag{5}$$

Thus, the main function of Equation (4) is to distribute the monthly total precipitation $\hat{G}_{M,j}(m)$ obtained by the IDW method to daily precipitations.

3. Results

3.1. Spatial distribution of October 18, 1998 storm

One of the heaviest rainfalls in the Texas history occurred in Texas on October 17 and 18, 1998. The storm resulted in a wide area of flooding with an estimated property damage of $750 million and loss of 31 human lives. The rain started in the early morning of October 17, 1998 northwest of San Antonio (29°32′N, 98°28′W), extended to eastern Texas, and hit Louisiana on October 18. The 2-day storm had a peak precipitation south of San Marcos (29°53′N, 97°56′W).

The three panels of Figure 3 display the spatial distribution of the October 18, 1998 storm precipitation based on the gridded data from the hybrid, IDW, and NSA methods. Figure 3(a) shows the hybrid result and demonstrates that the storm was spread out over a large area with a high peak value (351 mm) in the grid box centered at 30°2′N, 97°50′W in the western part of the San Marcos area. In this region a large spatial gradient of precipitation exists, but cannot be reproduced by the NSA method, which results in strong discontinuities. The IDW method (Figure 3(b)) cannot recover the peak values around the San Marcos area due to its oversmoothing. The IDW interpolation would result in a maximum precipitation of only 255 mm over the grid box centered at 29°38′N, 98°2′W. The spatial gradient of the IDW gridded data is too small. In contrast, the NSA results may be unrealistically spatially discontinuous. The NSA interpolation would result in a high peak value of 466 mm over the grid box centered at 29°38′N, 98°14′W. This particular 466 mm peak value is from the New Braunfels station (ID 42500416276) (29°44′N, 98°7′W). As aforementioned, due to the missing data of this station on October 17, the real peak value might be smaller than 466 mm. Nonetheless, this does not conclude that the maximum precipitation over Texas is definitely less than 466 mm on October 18. Because of the gauge overflow problem, it might have happened that the data from the overflowed gauges were eliminated in the data quality control process. Therefore, it is still subject to further investigation to find out the exact peak values of the October 18, 1998 storm over Texas. Radar data and mesocale modelling data might be helpful in solving this problem, if station data cannot be satisfactorily recovered. Before these results become available, the hybrid results appear most credible.

Besides the overflow problem with gauges, other error sources exist for the gauge data, such as some suspicious missing data. Not recording the gauge level on time for a manual rain gauge may lead to a rain accumulation for more than 24 h. For example, New Braunfels station (ID 42500416276) (29°44′N, 98°7′W) reported 466 mm rain with a data flag E (here 'E' means estimate) on October 18, 1998, but had missing data on October 17, 1998. It might have happened that the station did not record the data on October 17 because of the stormy weather. Thus, 466 mm record on October 18 might be due to an accumulation from October 17. Therefore, when an estimated precipitation is flagged in the data, it is better to utilize the multiple neighborhood stations to revise the data without dramatically changing the spatial distribution properties. The hybrid method is appropriate for this purpose.

We also compared our extremes with the extreme of Chen *et al.* (2008) $0.5^\circ \times 0.5^\circ$ daily dataset, which is 187 mm occurred at 30°00′N, 96°30′W. This small extreme value may be a consequence of many factors: the grid is too coarse, the station dataset includes fewer stations, and the spatial distribution of precipitation is oversmoothed in the interpolation process.

Although a major storm may have a large spatial gradient, the transitional zone still exists from the zone of the peak precipitation to that of weak or zero precipitation. For the Texas October 18, 1998 storm, the notable steep transitional zone is along the parallel of 29°53′N, particularly near San Marcos. Thus, the IDW method alone would have yielded a too smooth field and hence underestimated the extreme precipitation. The NSA would have yielded a highly discontinuous field if stations are sparse and the precipitation field has a large spatial variance. Therefore, Figure 3 supports the idea of the development of a hybrid of the IDW and NSA methods. The hybrid method can display a storm field more objectively compared to other available methods, since it can retain extreme precipitation and spatial locality, and at the same time can maintain some degree of spatial smoothness.

Now, we examine the covered area of a heavy storm. Different gridding methods may yield quite different amounts of total precipitation and different values of the heavy precipitation area. If we use 50 mm as the threshold for heavy storms, the heavy storm coverage area A_{50} over Texas on October 18, 1998 was 157 000 km^2 based on the hybrid data, 152 000 km^2 based on the IDW data, and 146 000 km^2 based on the NSA data. The total precipitation R_{50} over the heavy storm region is calculated through numerical volume integration throughout the grid boxes,

$$R_{50} = \int_{A_{50}} R(\vec{r})\mathrm{d}\Omega \approx \sum_{j=1}^{N} R_j A_j, \qquad (6)$$

where R_j is the gridded precipitation data from Equation (4) in the case of hybrid interpolation, A_j is the area of the jth grid box, and N is the total number of grid points within the heavy storm region A_{50}. The calculated R_{50} value is 16.5 billion m^3 from the hybrid data, 15.0 billion m^3 from the IDW data, and 17.0 billion m^3 from the NSA data. The above analysis indicates that R_{50} value of 16.5 billion m^3 from the hybrid method is a relatively more accurate result than the IDW and NSA results. Accurate assessment of the heavy storm areas and total precipitation in a specific period of time is very important for flood risk management and engineering designs for many hydrological projects.

3.2. Errors of precipitation frequency

We analyze the error of precipitation frequency to support the claim that the hybrid method interpolates not only

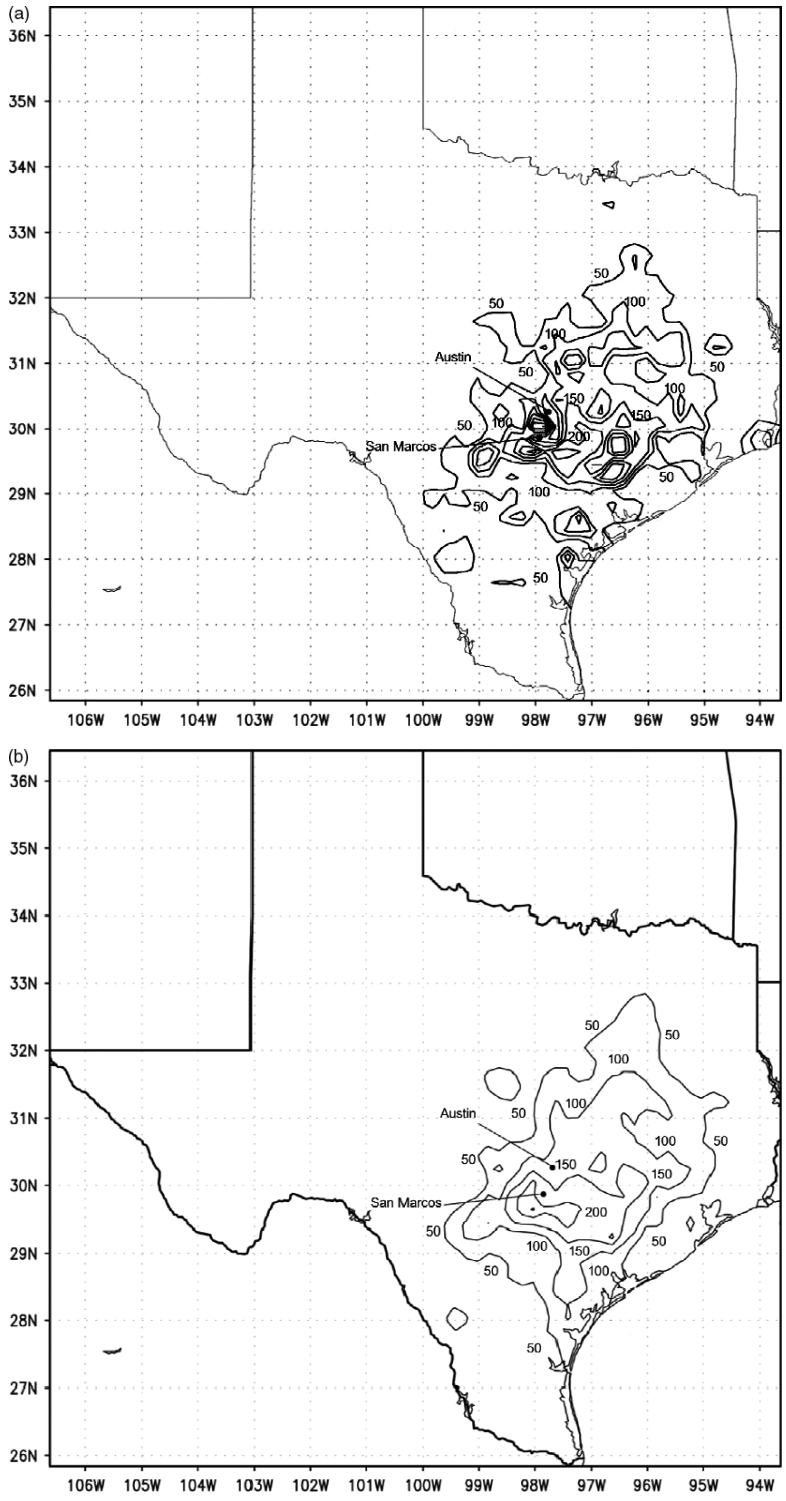

Figure 3a. Spatial distribution of the storm precipitation on October 18, 1998 according to the interpolated data by (a) the hybrid method, (b) the IDW method, and (c) the NSA method (Units: mm).

the precipitation amount with reasonable accuracy but also the number of days with precipitation. For a given day, a grid point is defined rainy if the grid point has a precipitation of 0.2 mm or greater. The relative percentage error can then be defined as:

$$e_{\text{rel}} = \frac{N_i - \hat{N}_i}{N_i} \tag{7}$$

where N_i denotes the true number of days with precipitation at the station location i and $\hat{N}_i$ is the number of estimated days with precipitation at this location. We use the cross-validation method to analyze this relative error. The procedures are (1) to withhold a long-term station, (2) to interpolate the data from the remaining stations to this station, and (3) to compare the differences between the true data (i.e. the withheld data) and the interpolated

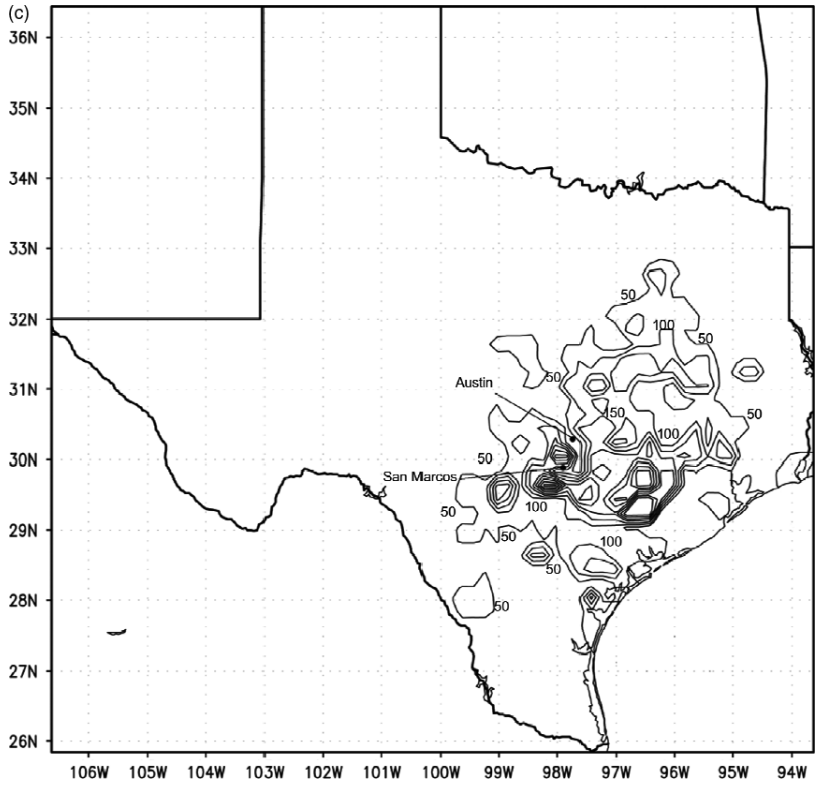

Figure 3b. (*Continued*).

data. Two sets of long-term stations are considered here. They are the stations whose data record length is greater or equal to 50 and 80%, respectively, of the interpolation period of January 1, 1901 to December 31, 2000. Figure 4 shows the histogram of the relative errors and the fit of a lognormal distribution. The spatial distribution of these two sets of stations is shown in Figure 5(a) and (b). The IDW method's overestimation of precipitation frequency is obvious because of the mean value around −60%. The hybrid method's median and expected values are around zero. The lognormal distribution has been used for modelling daily precipitation (Kedem *et al*., 1990). The lognormal distribution is used here to fit the distribution of relative errors after the errors are shifted to left by 2.0, i.e. error plus 2, to make the lognormal variable non-negative. Figure 4(a) and (b) indicates that the lognormal is a reasonably good fit for the network of long-term stations when using the hybrid method for interpolation. The skewness of this lognormal distribution is small and the mean is close to zero. As a matter of fact, the distribution is not too far away from being normal. The skewness is apparently much more severe for the IDW error results. We also fitted the histogram with the lognormal distribution, and the lower panels of Figure 4 indicate that the lognormal is not a good fit to the precipitation frequency error of the IDW method.

Of course, the relative errors from the hybrid and NSA methods should have identical results.

With the lognormal pdf, we can calculate the expected value and variance of the percentage error of the precipitation frequency when using the hybrid method or NSA method for interpolating the daily precipitation. The results provide a rough idea of the quality of the network on the precipitation frequency. The histogram and the fitted distribution are also helpful to examine the error properties of the precipitation frequency of other interpolation methods, such as Gaussian smoothing, thin-plate spline fitting, and kriging, as well as help to determine if a particular interpolation method is sound.

To test the robustness of our results, we have considered the following seasonal and regional dependence. Two seasons are considered: the warm season of May–October and the cool season of November–April. Two regions are considered: western Texas and eastern Texas partitioned by longitude 99 °W. With three spatial regions (entire Texas, eastern Texas, and western Texas) and three temporal periods (entire year, warm season, and cool season), nine cases exist altogether. For instance, the warm season of eastern Texas is a case, whose histogram is shown in Figure 4(b). Although the data points are not as many as shown in Figure 4(a) for the entire year and entire Texas, the histogram still supports a lognormal distribution of the precipitation errors of the hybrid method. The overestimate problem of precipitation frequency in the IDW method seems to be so serious that the method alone should not be used to interpolate the daily precipitation data. The other seven cases have similar conclusions and their figures are not shown in this paper.

3.3. Overestimate bias of the Texas precipitation by simple averaging

Here we describe the error results of the overestimate of total precipitation over Texas due to simple averaging.

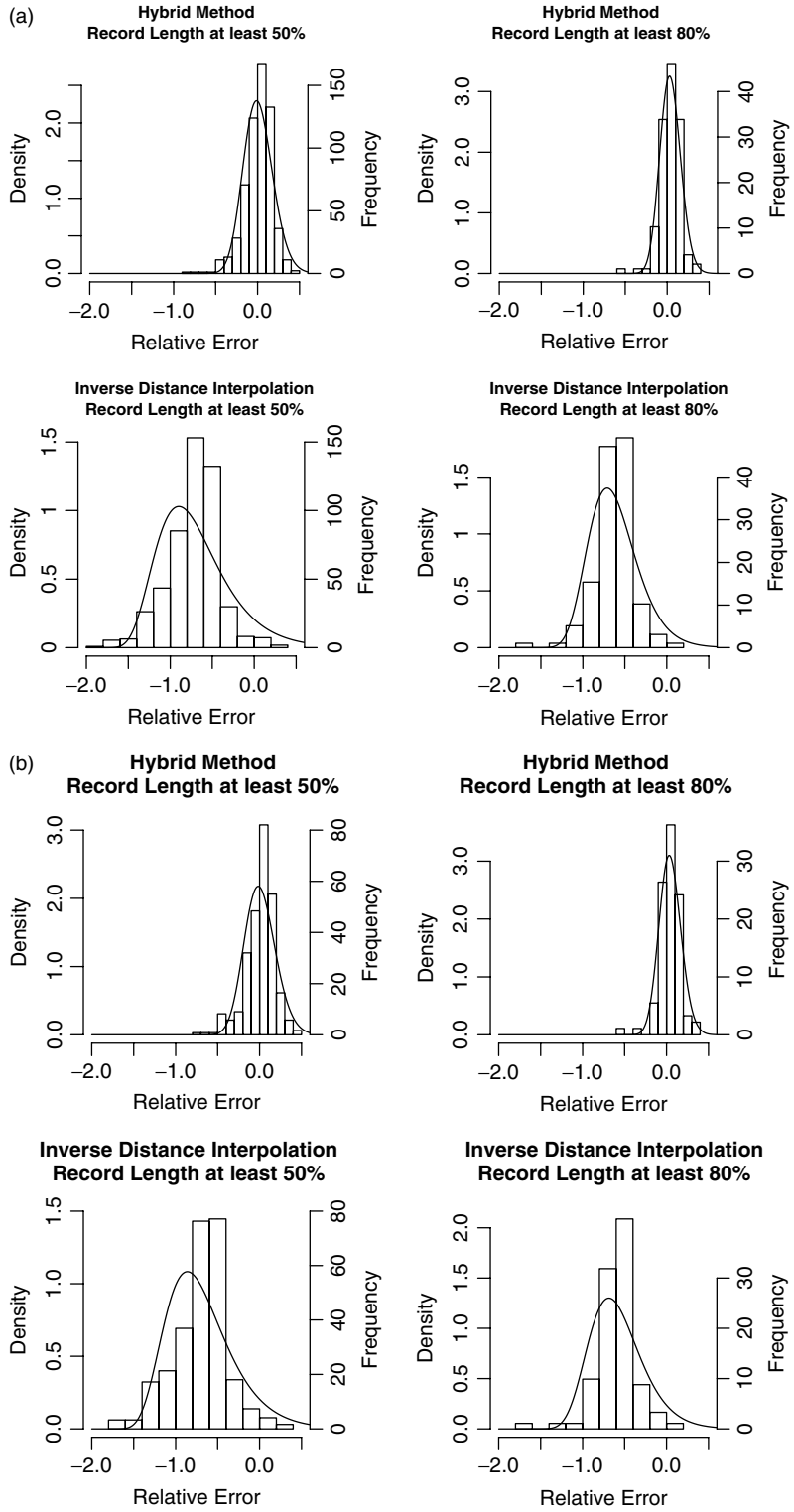

Figure 4. Histogram and lognormal fit of relative errors of precipitation frequency: (a) for the entire Texas and for the entire year and (b) for eastern Texas and for the warm season from May to October.

Texas has a very non-uniform precipitation structure. The precipitation amount increases dramatically from western to eastern Texas. From the western border to the centre of Texas, the annual precipitation is between 0 and 750 mm, whereas the annual precipitation between central and eastern Texas ranges between 750 and 2000 mm. Fewer weather stations exist in western Texas, and the station density is higher in eastern Texas because of higher population density and more complex precipitation events. If the arithmetic average of all the station data is used to calculate the average precipitation of Texas, i.e. the simple uniform weight average of all the station data, an overestimate will be obtained because of the spatial variation of station densities. The average precipitation of Texas

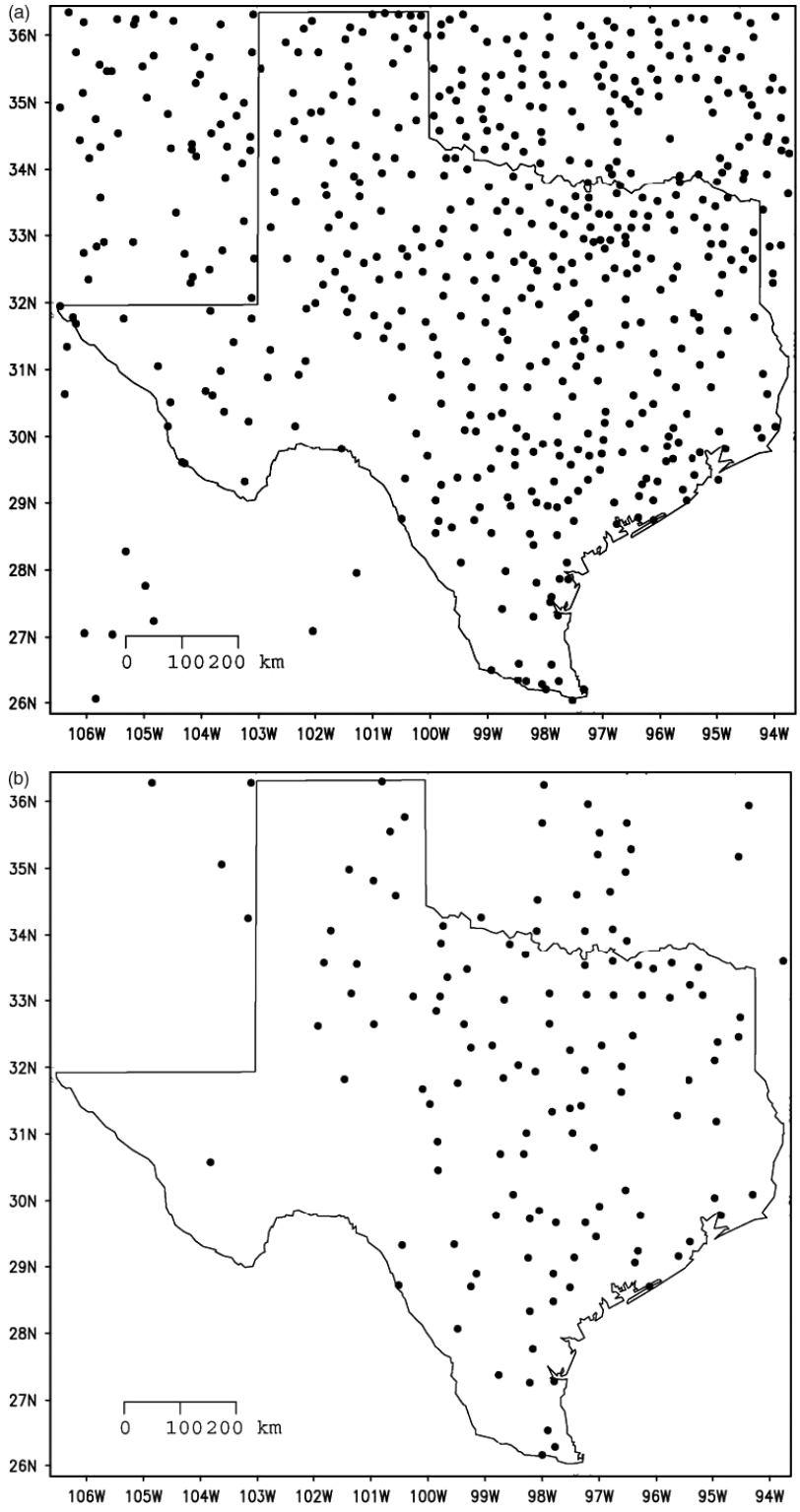

Figure 5. (a) Locations of the stations with a record length of 50% or longer of the entire interpolation period and (b) locations of the stations with a record length of 80% or longer of the entire interpolation period.

is defined as a spatial integration, and hence the numerical integration weight for a station should in general be proportional to the area the station represents. Thus, the weights for the stations over the station-dense eastern Texas region should be less than those for the stations over western Texas. The average by uniform weights puts too much weight on the stations over eastern Texas, where there is usually more precipitation than over western Texas. Hence, the uniform weighted average gives an overestimate in most cases for daily precipitation, and certainly gives an overestimate for annual precipitation.

A crude spatial integration is the area-weighted spatial average, which is equivalent to the average of gridded data that are obtained by the NSA method (Shen *et al*.,

2001). We have tested three cases: uniform weight average *versus* average of gridded data by NSA, hybrid, and IDW methods. The result is that the mean differences between the uniform weight average and each of the three sets of gridded data from January 1, 1901 to December 31, 2000 are the same: 2.4 mm. Of course, this same value is not a general conclusion but rather a coincidence. The standard deviations of the differences are very large: 8.4 mm for the NSA method, 8.4 mm for the hybrid method, and 8.2 mm for the IDW method. Box plots of the differences between the uniform weight average and the average of the gridded data by the hybrid method are shown in Figure 6. The box plots are shown for every 20 years. The box plots show that the mean is not zero but positive. The mean and variances are the largest in the 1901–1920 period because of sparse network coverage. The large positive difference values are due to the events of wide precipitation coverage over Texas and very heavy precipitation in the east but very light precipitation in the west. The negative differences are also due to the events of wide precipitation coverage and heavy precipitation in the west but light precipitation in the east. Apparently, this latter situation occurs less frequently than the former. The first quartile is very close to zero and the third quartile is also very small. This indicates the small differences between the two estimates for most times, because most precipitations are minor events and their differences are small too. The large differences are due to major events of heavy rains, which are not as frequent as the events of light rains. The two boundary ticks are given by 1.5 interquartile range (IQR), where the IQR is defined as the difference between the third quartile and the first quartile. Each circle outside of the 1.5 IQR ticks represents a major precipitation event. The upper ones are the heavy rains in eastern Texas and the lower ones are the heavy rains in western Texas. The amount of the overestimate is not an ignorable value: an annual amount of 876 mm about the annual total precipitation of central Texas. The amount is particularly important to drought monitoring in agriculture. This amount can mean the dramatic change of drought status over western Texas from normal to a D3 category drought (extreme drought with a return period of approximately 10 years). Therefore, it is very important to understand this difference and compute the spatial average from either the properly gridded data or from the optimal spatial average (Smith and Reynolds, 2005).

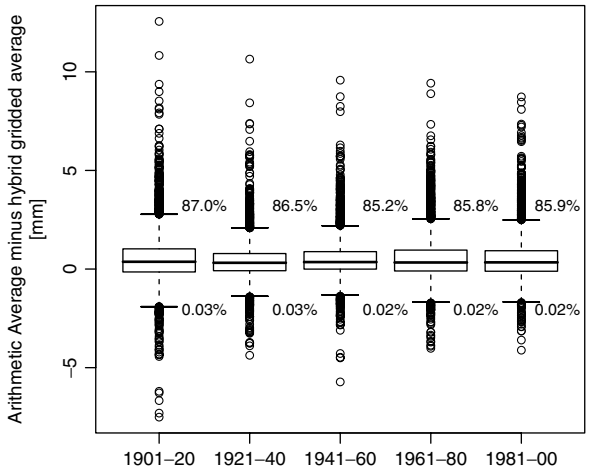

Figure 6. Box plots of the differences between uniform weight average and hybrid gridded average of Texas precipitation from January 1, 1901 to December 31, 2000.

4. Conclusions and discussion

From the GDCN station data, we have generated a $0.2^\circ \times 0.2^\circ$ gridded dataset of daily precipitation over Texas from January 1, 1901 to December 31, 2000 by using the hybrid interpolation method of Shen *et al.* (2001). The hybrid method's characteristic of retaining the precipitation extremes and precipitation frequency in the gridding process is useful in the assessment of climate changes, in applications of weather risk management, and in predicting the future extreme weather conditions. An analysis of the storm on October 18, 1998 showed the hybrid method's advantages of retaining extreme precipitation and maintaining natural spatial continuity. The cross-validation method has been used for analyzing the relative errors of precipitation frequency. The relative error of the precipitation frequency follows a lognormal distribution. We have also found that a simple arithmetic average of station data often overestimates Texas' average precipitation and the mean overestimate is 2.4 mm per day, which is a non-trivial amount comparable to the annual precipitation of western Texas.

Our study provides an approach to examine the quality of precipitation data gridding from perspectives of both spatial and temporal structures. Our results and our hybrid method will help mitigate the problems of significantly higher precipitation frequency of light precipitation and significantly lower annual maximum of daily precipitation found in some gridded data (Ensor and Robeson, 2008). Our results on the probability distribution of the precipitation frequency errors provide a statistical indicator for the precipitation frequency bias. The indicator may be used to determine a correction factor of reducing the precipitation frequency in some gridding procedures. Owning to the importance and difficulty of obtaining the correct precipitation frequency for the practical applications and the researches on climate modelling, every gridded daily precipitation dataset should have a careful statistical analysis of the precipitation frequency errors.

Climate extremes often have a much larger spatial variation than the climate mean. Thus, retaining climate extremes in a gridding method often has a risk of high cross-validation errors (Shen *et al.*, 2001). However, the climate extremes are more important than the climate mean to climate monitoring and applications. Therefore, the climate gridding is useful not only in producing a complete climate field but also in generating realistic climate extremes that can provide guidelines for agricultural applications, despite a possible shift of the locations of

the extremes. The non-stationary nature of the daily precipitation field, however, makes the spatial gridding a challenging statistical and mathematical problem.

The hybrid method has flexibilities and can be a blend of any two or multiple methods. Shen *et al*. (2001) showed the advantages of the hybrid of the simple methods used in this paper over other more sophisticated methods, such as kriging. However, when elevation is taken into account, some other more complicated hybrids that take atmospheric dynamics into account have a higher potential to generate more accurate results. A hybrid example is the EOF method for the monthly total with a multi-station precipitation indicator for the temporal downscaling.

Although our study is for Texas and for a particular dataset GDCN, our methodology is applicable to other regions where the precipitation gauge network has a reasonable spatial and temporal coverage. However, the seasonal variation of the quality of the gridded data measured by various kinds of statistics can be significant for the regions of inland or high latitude where the spatial and temporal scales of the daily precipitation field have large annual cycles.

It is obvious that the errors of the gridded data are related to the station density of a gauge network. Quantifying this relationship will be helpful in searching for improved methods of gridding to raise the quality of gridded data. This, however, seems to be a very challenging problem.

Acknowledgements

This study was supported in part by the US National Oceanographic and Atmospheric Administration's OGP /CCDD program and San Diego State University's start-up research fund for Shen. B. Gleason and D. Easterling of the National Climatic Data Center provided assistance with the daily data extraction.

References

Chen M, Shi W, Xie P, Silva VBS, Kousky VE, Higgins RW, Janowiak JE. 2008. Assessing objective techniques for gauge-based analyses of global daily precipitation. *Journal of Geophysical Research* **113**: D04110, DOI:10.1029/2007JD009132.

Chen M, Xie P, Janowiak JE, Arkin PA. 2002. Global land precipitation: a 50-yr monthly analysis based on gauge observations. *Journal of Hydrometeorology* **3**: 249–266.

Cressie N. 1993. *Statistics for Spatial Data*. John Wiley & Sons, Inc.: New York; 900.

Dunkl CF, Xu Y. 2001. *Orthogonal Polynomials of Several Variables*. Cambridge University Press: London; 400.

Durre I, Menne MJ, Vose RS. 2008. Strategies for evaluating quality-control procedures. *Journal of Climate and Applied Meteorology* in press.

Ensor LA, Robeson SM. 2008. Statistical characteristics of daily precipitation: comparison of gridded and point datasets. *Journal of Applied Meteorology and Climatology* **47**: 2468–2476.

Gandin LS. 1963. *Objective Analysis of Meteorological Fields*. In Russian, Gidrometeor. Isdat., Leningrad. [English translation, 1966, Israel Program for Scientific Translation; 242].

Gleason BE. 2002. *Global Daily Climatology Network*. V1.0, Data Documentation for Data Set 9101,. National Climatic Data Center: Asheville.

Graves CE, Valdes JB, Shen SSP, North GR. 1993. Evaluation of sampling errors of precipitation from space-borne and ground sensors. *Journal of Applied Meteorology* **32**: 374–385.

Griffith DP. 2002. *Processing of Daily Agroclimatic Data*. M.Sc. Thesis, University of Alberta, Edmonton; 104.

Higgins RW, Shi W, Yarosh E, Joyce R. 2000. *Improved United States Precipitation Quality Control System and Analysis*. NCEP/Climate Prediction Center ATLAS No. 7, 40 pp., Camp Springs, MD 20746, USA.

Huff FA, Shipp WL. 1969. Spatial correlations of storms, monthly and seasonal precipitation. *Journal of Applied Meteorology* **8**: 542–550.

Huffman GJ, Adler RF, Morrissey M, Bolvin DT, Curtis S, Joyce R, McGavock B, Susskind J. 2001. Global precipitation at one-degree daily resolution from multi-satellite observations. *Journal of Hydrometeorology* **2**: 36–50.

Johnson R, Wichern DW. 1992. *Applied Multivariate Statistical Analysis* (3rd Ed.). Prentice Hall: Englewood Cliffs; 642.

Jolly WM, Graham JM, Michaelis A, Nemani R, Running SW. 2005. A flexible, integrated system for generating meteorological surfaces derived from point sources across multiple geographic scales. *Environmental Modeling and Software* **20**: 873–882.

Kedem B, Chiu LS, North GR. 1990. Estimation of mean rain rate: application to satellite observations. *Journal of Geophysical Research* **95**(D2): 1965–1972.

Liebmann B, Allured D. 2005. Daily precipitation grids for South America. *Bulletin of American Meteorological Society* **86**: 1567–1570.

North GR, Nakamoto S. 1989. Formalism for comparing rain estimation designs. *Journal of Atmospheric and Oceanic Technology* **6**: 985–992.

North GR, Valdes JB, Ha E, Shen SSP. 1994. The ground truth problem for satellite estimates of rain rate. *Journal of Atmospheric and Oceanic Technology* **11**: 1035–1041.

Robeson SM, Ensor LA. 2006. Comments on "Daily precipitation grids for South America.". *Bulletin of American Meteorological Society* **87**: 1095–1096.

Rupp AJ. 2007. *Gridding Daily Precipitation Data over Texas*. Master Thesis, San Diego State University: San Diego; p 218.

Shen SSP, Basist AN, Li G, Williams C, Karl TR. 2004. Prediction of sea surface temperature from the global historical climatology network data. *Environmetrics* **15**: 233–249.

Shen SSP, Dzikowski P, Li G, Griffith D. 2001. Interpolation of 1961–1997 daily climate data onto Alberta polygons of ecodistrict and soil landscape of Canada. *Journal of Applied Meteorology* **40**: 2162–2177.

Shen SSP, Yin H, Cannon K, Howard A, Chetner C, Karl TR. 2005. Temporal and spatial changes of agroclimate in Alberta during 1901–2002. *Journal of Applied Meteorology* **44**: 1090–1105.

Simpson J, Adler RF, North GR. 1988. A proposed tropical rainfall measuring mission (TRMM) satellite. *Bulletin of the American Meteorological Society* **69**: 278–278.

Smith TM, Reynolds RW. 2005. A global merged land air and sea surface temperature reconstruction based on historical observations (1880–1997). *Journal of Climate* **18**: 2021–2036.

Smith TM, Reynolds RW, Livezey RE, Stokes DC. 1996. Reconstruction of historical sea surface temperature data using empirical orthogonal functions. *Journal of Climate* **9**: 1403–1420.

Stensvand A, Eikemo H. 2005. Use of a rainfall frequency threshold to adjust a degree-day model of ascospore maturity of venturia inaequalis. *Plant Disease* **89**: 198–202.

Thornton PE, Running SW, White MA. 1997. Generating surfaces of daily meteorology variables over large regions of complex terrain. *Journal of Hydrology* **190**: 214–251.

Wahba G. 1990. *Spline Models for Observational Data*, *Volume 59 in the CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM: Philadelphia; 169.

Xie P, Arkin PA. 1997. Global precipitation: a 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bulletin of American Meteorological Society* **78**: 2539–2558.

Xie P, Yatagai A, Chen M, Hayasaka T, Fukushima Y, Liu C, Yang S. 2007. A gauge-based analysis of daily precipitation over East Asia. *Journal of Hydrometeorology* **8**: 607–626.

