Computational Procedures for the 1981-2010 Normals: Precipitation, Snowfall, and Snow Depth

Preliminary Documentation
July 1, 2011

Imke Durre, NCDC; Michael F. Squires, NCDC; Russell S. Vose, NCDC, Anthony Arguez, NCDC; Scott Applequist, NCDC; and Xungang Yin, STG, Inc.

1. Overview

The 1981-2010 U.S. Climate Normals being released by NOAA's National Climatic Data Center (NCDC) during 2011 include a suite of descriptive statistics based on precipitation, snowfall, and snow depth measurements at several thousand stations from across the United States and its Caribbean and Pacific territories. Three groups of statistics are provided: 30-year averages, frequencies of occurrence, and percentiles (Table 1). Thirty-year averages of month-to-date, year-to-date, monthly, seasonal, and annual totals of precipitation and snowfall serve as basic descriptors of a location's climate. Daily relative frequencies and average number of days per month for precipitation, snowfall, and snow depth exceeding various thresholds (e.g., precipitation >= 0.01 inches) provide a starting point for estimating the present-day and future likelihood of those threshold exceedances (Table 2). Finally, percentiles of both daily values and monthly totals offer distributional information that can be useful when placing a particular amount of precipitation or snow into historical perspective.

Table 1. Statistics Produced for the 1981-2010 Normals	
Precipitation	Average monthly/seasonal/annual totals
	Average month-to-date totals
	Average year-to-date totals
	Average number of days per month exceeding various thresholds
	Daily relative frequencies exceeding various thresholds
	25th, 50th, and 75th percentiles of monthly totals
	Daily 25th, 50th, and 75th percentiles
Snowfall	Average monthly/seasonal/annual totals
	Average month-to-date totals
	Average year-to-date totals
	Average number of days per month exceeding various thresholds
	Daily relative frequencies exceeding various thresholds
	25th, 50th, and 75th percentiles of monthly totals
	Daily 25th, 50th, and 75th percentiles
Snow Depth	Average number of days per month/season/year exceeding various thresholds
	Daily relative frequency exceeding various thresholds
	Daily 25th, 50th, and 75th percentiles

Table 2. Element Dependent Thresholds for *Relative Frequency* and *Number of Days per Month* Statistics	
Precipitation	>= .01”, .10”, .50”, 1.00”
Snowfall	>= 0.1”, 1.0”, 3.0”, 5.0”, 10.0”
Snow Depth	>= 1”, 3”, 5”, 10”

The above-mentioned suite of statistics are referred to as "traditional normals" and are calculated directly from the available data. For active stations whose records are too short for such calculations, average monthly, seasonal, annual, month-to-date, and year-to-date totals are estimated and provided as "quasi-normals".

The remainder of this document describes the procedures used to compute the various parameters. Section 2 provides a brief description of the data and the station selection criteria for traditional normals. The computational procedures for traditional normals are explained in section 3, and a brief description of the quasi-normals is provided in section 4. Figures and other material will be incorporated into a forthcoming Journal article on all of these computations.

2. Data

A. Source data

The statistics are computed from the Global Historical Climatology Network - Daily (GHCN-Daily) data set (Menne et al., submitted). The data in GHCN-Daily originate from more than 20 sources and are processed through a carefully designed suite of quality-control (QC) procedures (Durre et al. 2010). The U.S. component of GHCN-Daily is an integrated version of NCDC's holdings of daily surface observations, including those from the U.S. Cooperative Observer Network (Coop) and the Automated Surface Observing System (ASOS), among others, and thus represents the most complete historical record of daily data for the United States.

For the calculations described herein, 24-hour precipitation totals, 24-hour snowfall totals, and once-a-day snow depth measurements are considered. Data from NCDC's Global Summary of the Day, a GHCN-Daily source based on synoptic messages, are not included in the normals calculations because they tend to be less accurate and apply to 24-hour periods different from the reporting periods in other data sources. Also excluded are data flagged as erroneous by the GHCN-Daily QC system and multi-day accumulations of precipitation and snowfall extending over more than three days.

B. Station selection for traditional normals

In order to be included in the calculations, a station must meet the following criteria:

(1) It must be part of a network operated by the National Weather Service (NWS).

(2) It must be located in one of the 50 U.S. states or on one of the Caribbean or Pacific Islands where the NWS operates stations.
(3) Its record of precipitation measurements not flagged as erroneous by the GHCN-Daily QC system must be sufficiently long and complete to allow for each of the statistics included in the normals to be based on at least 10 years of data during 1981-2010; i.e.,
  (a) each of the 12 calendar months must be complete in at least 10 of the 30 years and
  (b) for each day of the year, there must be at least 10 years in which at least 20 values are available within plus or minus 14 days of that day.
(4) Its entire suite of calculated normals parameters must pass various consistency checks intended to identify stations with significant biases in the data (e.g., snowfall reported as zero rather than missing).

The above data completeness requirements are based on recommendations in WMO (1989) as well as on sensitivity tests conducted for each of the statistics computed.

Since records of snowfall and snow depth observations are not sufficiently complete at all stations that meet the data completeness requirements for precipitation, statistics for these two elements are provided for appropriate subsets of the chosen precipitation stations. As a result, statistics for all three variables are available at approximately 5300 stations, precipitation and snowfall statistics are available for another 1100 stations, and statistics for precipitation alone are provided for a further 1100 stations.

3. Computational procedures for traditional normals

A. Averages

1) Averages of monthly, seasonal, and annual precipitation and snowfall totals

The average monthly, seasonal, and annual totals of precipitation and snowfall require as a first step the computation of monthly totals for each station/year/month. Following WMO (1989), a monthly total is calculated for every station/year/month in GHCN-Daily that is complete when daily values, two-day accumulations, and three-day accumulations are considered. Accumulations that extend from the end of one month to the beginning of another are excluded. February 29 is included in the monthly totals for February in leap years. For precipitation, an attempt is made to fill in monthly totals that are missing during the normals period using a previously developed method based on median absolute deviation regression relationships between stations and qualifying neighbors. Monthly snowfall totals are not estimated because of the large number of zeros that degrade the quality of the regression relationships in many locations and because of the temperature dependence and larger spatial variability of snowfall compared to precipitation. A description of the estimation procedure is provided in the appendix.

For each calendar month, the average monthly total then is the arithmetic mean of all observed and, for precipitation, estimated totals available during the 30-year period. Thus, any given average can be based on a complete record of observed monthly totals, an incomplete record of observed monthly totals when estimates could not be produced, or, in the case of precipitation, a record that is complete when both observed and estimated values are considered.

Also provided are average totals for the year as well as for the seasons of December-February, March-May, June-August, and September-November. Seasonal and annual averages are obtained by summing the appropriate average monthly totals.

2) Average month-to-date and year-to-date totals

Average month-to-date and year-to-date totals of precipitation and snowfall are calculated from daily data in complete months, smoothed to reduce spurious variations in day-to-day increases, and scaled to match the corresponding average monthly totals described in the previous section. Specifically, the procedure works as follows:

First, for each calendar month, those years are identified during which daily precipitation observations are available for the entire month. For each day (*i)* of the month, "raw" average month-to-date totals are obtained by summing the daily precipitation measurements between day 1 and day *i* of the month across all qualifying years and dividing the resulting grand sum by the number of qualifying years. (February 29 is excluded from this calculation.) The results directly reflect the available data, but have two shortcomings:

> (a) Sampling variability results in significant short-term variations in the day-to-day increases in month-to-date totals.
>
> (b) The end-of-month totals may not match the corresponding average monthly totals described in the previous section which are based not only on daily precipitation observations, but also on two-day and three-day accumulations and estimated monthly totals.

These issues are addressed by the subsequent smoothing and scaling steps.

In preparation for smoothing, raw average year-to-date totals are formed by summing the average month-to-date totals accordingly. For example, the year-to-date total for March 26 is the sum of the end-of-month total for January, the end-of-month total for February, and the month-to-date total for March 26. The resulting average year-to-date totals are then further aggregated into three-year-to-date totals in order to allow for smoothing across the beginning and end of the year. A 29-day running mean is then applied to the three-year series before the middle of the three years is disaggregated again into smoothed average month-to-date totals. The 29-day filter was chosen after testing various shorter running means and was found to remove variations that are likely to be the result of sampling variability while preserving changes in the rate of increase on the time scale of weeks such as those seen during the onset of the southwest monsoon. Although the smoothing may significantly change the average end-of-month total, this effect will be eliminated by the subsequent scaling step.

Finally, the smoothed average month-to-date totals are scaled to match the corresponding average monthly totals in order to achieve consistency between the two statistics. In most cases, this is done by multiplying the month-to-date total for each day of the month by the ratio of the average monthly total to the month-to-date total on the last day of the month. However, two cases need to be addressed in which this "ratio scaling" does not apply:

> (a) If the average month-to-date total at the end of the month and the average monthly total both equal zero, no scaling is necessary because the statistics already match. This situation most commonly arises for snowfall during the snow-free season. For precipitation, it only arises at a few locations in the Southwest during summer months.
>
> (b) If the average month-to-date total at the end of the month is zero, while the average monthly total is greater than zero, the average month-to-date total is set equal to the average monthly total on the last day of the month and set to zero on all other days of the month. This was found to occur only once for snowfall when the average monthly total was equal to 0.1 inches due to the inclusion of months containing two-day and three-day accumulations which are not considered in the month-to-date totals.

B. Frequencies of occurrence

The frequencies of occurrence of several types of events (Table 2) are expressed in terms of two quantities: (1) average (or expected) number of days per month, season, and year on which the event occurs and (2) relative frequencies of occurrence (in percent of available values) during 29-day windows centered on each day of the year. The daily relative frequencies are intended as a supplement to the monthly frequencies that provides an indication of major changes in relative frequency within a month, such as the increase in the frequency of precipitation during the onset of the southwest monsoon at stations in the southwest.

1) Average number of days per month, season, and year

For a particular station and calendar month, the expected number of days per month on which precipitation, snowfall, or snow depth exceed a specific amount is calculated as follows:

First, all years are identified in which the daily observations of the element of interest are missing on nine or fewer days of the month. In other words, up to approximately 1/3 of the month is allowed to be missing. Next, for each of the qualifying years, the number of days on which the variable of interest is greater than or equal to the specified threshold (e.g., precipitation >= 0.01 inches) is counted. Third, the counts for each year are summed, and the result is divided by the total number of days with data in the qualifying years to obtain the probability of the threshold exceedance for the calendar month. This relative frequency is then multiplied by the number of calendar days in the month to obtain the corresponding expected number of days per month on which the threshold is exceeded. (For February, the number of calendar days in the month is set to 28+7/30 to account for the seven leap years during 1981-2010.)

Assume, for example, that the average number of days with precipitation >= 0.01 inches during January is to be calculated. If only 25 days are available in one of the years during January, and January is complete in all other 29 years, the total number of days available would be 29 years*31 days + 25 days = 924 days. If 308 of those days have precipitation >= 0.01 inches, then the probability of this event during January would be 308/924, or 0.33, and the event would be expected to occur on an average of 0.3333*31, or 10.33, days during the month.

2) Daily relative frequencies of occurrence

The relative frequencies for each day of the year are calculated in a manner that is consistent with the computation of monthly frequencies. In this case, however, the values are chosen from a 29-day window centered on each day of the year, and the frequency is expressed as a percent of available values. Specifically, the computation works as follows:

> a) For each day of the year except February 29, an empirical relative frequency is first calculated using data from all years that have values on at least 20 of the 29 calendar days within the applicable window. For example, if the relative frequency of measurable precipitation is to be computed for January 1, data for 2010 are used if there are at least 20 values available during January 1-15 and December 17-31 of that year. The pool of values considered then consists of all available values on those 29 days during all qualifying years, and the relative frequency is equal to the percentage of values within the pool that are greater than or equal to 0.01 inches.
>
> b) Due to the limited sample size, the relative frequency for February 29 is not calculated directly. Rather, it is set to the average of the frequencies for February 28 and March 1.
>
> c) In order to reduce short-term fluctuations in the resulting frequencies that are likely to be associated with sampling variability, the empirical frequencies are smoothed with a 29-day running mean. After several other types of filters as well as shorter running means had been tested, this particular filter was found to yield the desired level of smoothing while retaining variations on the time scale of weeks.

For a particular station and element, relative frequencies are produced only if there are at least 10 qualifying years for each of the 365 windows. The 29-day window was chosen to increase the sample size from what it would be if only the values reported on the calendar day of interest were considered; with a minimum of 20 values in at least 10 years, each frequency is calculated from at least 200 values.

C. Percentiles

Since precipitation, snowfall, and snow depth follow distributions that are positively skewed, the medians and quartiles of both monthly totals and daily observations are also provided as an indication of the variability of these values. The percentiles of monthly precipitation and snowfall totals are calculated for each calendar month. The same percentiles are also computed within moving 29-day windows centered on each day of the year for precipitation on days with measurable precipitation, for snowfall on days with measurable snowfall, and for snow depth on days with snow on the ground.

1) Medians and quartiles of monthly totals

For each calendar month, all available monthly totals (including estimates in the case of precipitation) are sorted from lowest to highest, and the lower quartile (i.e., 25th percentile), median (50th percentile), and upper quartile (75th percentile) are calculated following standard procedures. If values from all 30 years are available, then the lower quartile is the eighth lowest value, the median is the average of the 15th and 16th lowest values, and the upper quartile is the 23rd lowest value. The calculation is performed only if there are at least 10 values available for the calendar month, a condition that is ensured by the station selection criteria described in section 2b above.

2) Medians and quartiles of nonzero daily observations

Percentiles are calculated for a station, element, and day of the year using the same pool of values from which the daily relative frequencies are produced, and the same level of smoothing is also applied. However, since the percentiles are based only on nonzero values, some additional considerations are required with a sample of these values is limited as is the case when the climate is dry or snow is rare. A full description of the procedure follows.

> a) For each day of the year except February 29, an empirical percentile is first calculated using data from all years that have values on at least 20 of the 29 calendar days within the applicable window. The nonzero amounts within this set of values are sorted from lowest to highest, and the median and quartiles are identified as described for monthly totals in the previous subsection. If not at least 10% of the chosen values are nonzero, the median and quartiles for that day of the year are set to missing at this stage.
>
> b) The median and quartiles for February 29 are set to the average of the corresponding percentiles for February 28 and March 1.
>
> c) The resulting sets of medians and quartiles are each smoothed with a 29-day running mean. Since percentiles may only have been calculated for parts of the year, the running mean is not always based on the full 29 days: In order to allow for the smoothing of even those percentiles that directly precede or follow a time

of the year during which percentiles are missing, the running mean is calculated whenever percentiles are available on at least 15 of the 29 days; percentiles that cannot be smoothed in this manner are set to missing.

d) Some additional steps are taken to address the potentially discontinuous nature of the resulting percentiles in relatively dry and relatively snow-free regions. First, gaps in the percentiles that are shorter than 15 days are filled in using linear interpolation between the corresponding percentiles immediately preceding and following the gap. Second, for a particular station and element, the resulting set of percentiles is retained only if there is a continuous stretch of (empirical and interpolated) percentiles that is at least 30 days long. Third, any remaining continuous stretches of percentiles shorter than 15 days are removed. Two final cosmetic steps are taken to address some special cases. The entire set of precipitation percentiles for a particular station are retained only if they contain no more than two gaps after the preceding three steps. In addition, at locations in the northern Great Plains, mid-winter gaps in snowfall percentiles are filled in using linear interpolation even when they extend over more than two weeks since the percentiles before and after the gap typically do not differ much from each other.

As a result of all of the above steps, precipitation percentiles are provided at approximately 5900 locations year-round as well as at another 1600 stations where precipitation is sufficiently frequent only during part of the year. Percentiles of snowfall in snow depth are available during the central portions of the snow season at approximately 2500 and 2600 stations, respectively.

4. Quasi-normals

A subset of the statistics described in the previous section are provided for approximately 1800 active stations whose precipitation records did not meet the completeness requirements for the traditional normals. Although the statistics are not expected to be as accurate as those based on more complete records, they are intended as guidance for a basic assessment of the precipitation climatology at NWS and Climate Reference Network (CRN) stations that were operational in 2010 end for which no statistics would otherwise be available.

The "quasi-normals" are available for precipitation only (not snowfall or snow depth) and include the following parameters:

- Estimated average monthly totals, or "pseudo-normals," computed closely following the neighbor-based estimation approach of Sun and Peterson (2005, 2006);
- seasonal and annual totals calculated from the monthly pseudo-normals as described in Section 3a above; and
- average month-to-date and year-to-date totals computed from the daily data and scaled by the monthly pseudo-normals in a manner analogous to the method described in Section 3b.

For quasi-normals, a less restrictive completeness requirement applies than for traditional normals: All 12 calendar months must have complete precipitation records in at least two, rather than 10, of the years between 1981 and 2010. However, the station must also report at least one precipitation total during 2010 in order to be considered "active." a final selection criterion is imposed by the pseudo-normals methodology: In order to be able to calculate monthly pseudo-normals, a station must have at least 10 neighbors within 500 km for which traditional normals are available. This number is close to the 11 neighbors used by Sun and Peterson (2006). As a result of these criteria, quasi-normals are available for 1741 NWS stations and 84 CRN stations.

Appendix: procedure for estimating monthly precipitation totals

The estimation procedure works as follows:

Regression relationships are developed separately for every station/calendar month for which such estimation is necessary, using all available monthly totals during the normals period that overlap with qualifying neighbors. Neighboring stations used are not required to be located in the United States or its territories and do not have to meet the network affiliation criteria for inclusion in the normals, thus increasing the possibility for estimates in border areas, on islands, and in data-sparse regions. However, when estimates are required for a particular target station and calendar month, qualifying neighbors have to meet the following criteria: (1) they must be located within 500 km of the station for which the estimates are to be produced; (2) they must have at least 10 monthly totals for that calendar month during the normals period that overlap with totals available at the target station; and (3) they must have monthly totals during all years for which an estimate is needed at the target station for that particular calendar month.

Qualifying neighbors are sorted in order of descending index of agreement between their monthly totals for the calendar month of interest and overlapping monthly totals at the target station. The index of agreement provides a measure of similarity in terms of both correlation and amplitude.

A regression model and accompanying estimates are then developed as follows. The neighbor with the highest index of agreement is used to develop a regression model. Next, the neighbor with the second highest index of agreement, if available, is added, and a two-neighbor regression model is established. If the estimates based on the two-neighbor model has a lower index of agreement with the available monthly totals at the target location than the one-neighbor model, the one-neighbor model is used. If the two-neighbor model is in better agreement than the one-neighbor model, the neighbor with the third-highest index of agreement is added, and a three-neighbor model is developed. This iterative procedure continues until no more qualifying neighbors are available or the maximum allowable number of neighbors has been reached. The maximum number of neighbors allowed depends on the number of observations at the target but cannot exceed five. (Any negative precipitation estimates are set to zero for each model developed before its agreement with the observations is evaluated.)

Estimates are not produced for a particular station and calendar month if any of the following conditions apply: the monthly totals at the target station are all equal to zero, thus making the development of a regression relationship impossible; there are no qualifying neighbors; or no valid regression model can be produced because all possible combinations of input data results in degenerate or non-unique solutions. Therefore, there are some stations whose record of monthly totals is not complete for all calendar months even after the attempt to estimate the missing monthly totals.

The requirement of a minimum number of 10 years of overlap for the development of regression models is chosen to be consistent with the minimum completeness requirement for inclusion in the normals products. Increasing the minimum number of years required to 15 decreases the number of stations for which estimates can be produced by about 15% and yields only a slight improvement in the error statistics.

The neighbor search radius of 500 km leaves only a few island stations for which no suitable neighbors can be found. Increasing the radius to 1000 km does not yield estimates for those stations. At the same time, sensitivity tests suggest that smaller radii (75, 150, or 225 km) would increase the number of stations and calendar months without estimates and slightly increase the overall error in the estimates.

References

Durre, I., M. J. Menne, B. E. Gleason, T. G. Houston, and R. S. Vose, 2010: Comprehensive automated quality assurance of daily surface observations. *Journal of Applied Meteorology and Climatology,* **8**, *1615-1633.*

Menne, M., I. Durre, B. G. Gleason, T. G. Houston, and R. S. Vose, 2011: An overview of the Global Historical Climatology Network-Daily database. Submitted to the Journal of Atmospheric and Oceanic Technology.

Sun, B., and T. C. Peterson, 2005: Estimating temperature normals for USCRN stations. *International Journal of Climatology,* **25**, *1809-1817*.

Sun, B., and T. C. Peterson, 2006: Estimating precipitation normals for USCRN stations. *Journal of Geophysical Research, 111, D09101, doi:10.1029/2005JD006245.*

World Meteorological Organization (WMO), 1989: Calculation of monthly and annual 30-year standard normals. WMO-TD 341, Geneva.