



# NCEI Data Sets

Brian Nelson

Others: Matt Menne, Dave Wuertz, Jay Lawrimore

August 20, 2019

NOAA Satellite and Information Service | National Centers for Environmental Information



# Overview

- **Hourly Precipitation Data (HPD)**  
–<ftp://ftp.ncdc.noaa.gov/pub/data/hpd/readme.txt>
- **Hydrometeorological Automated Data System (HADS)**  
–<https://hads.ncep.noaa.gov/>
- **Global Historical Climatology Network (GHCN-daily)**  
–<https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/global-historical-climatology-network-ghcn>
- ~~**Integrated Surface Database (Global-hourly)**~~  
–<https://www.ncdc.noaa.gov/isd>
- **U.S. Climate Reference Network**  
–<https://www.ncdc.noaa.gov/crn/>
- **NWS Stage IV gridded precipitation**  
–<https://www.emc.ncep.noaa.gov/mmb/ylin/pcpanl/stage4/>
- ~~**Multi-Radar Multi-Sensor System (gridded precipitation)**~~  
–<https://www.nssl.noaa.gov/projects/mrms/>



# Hourly Precipitation Data (HPD)

# Hourly Precipitation Data (HPD)

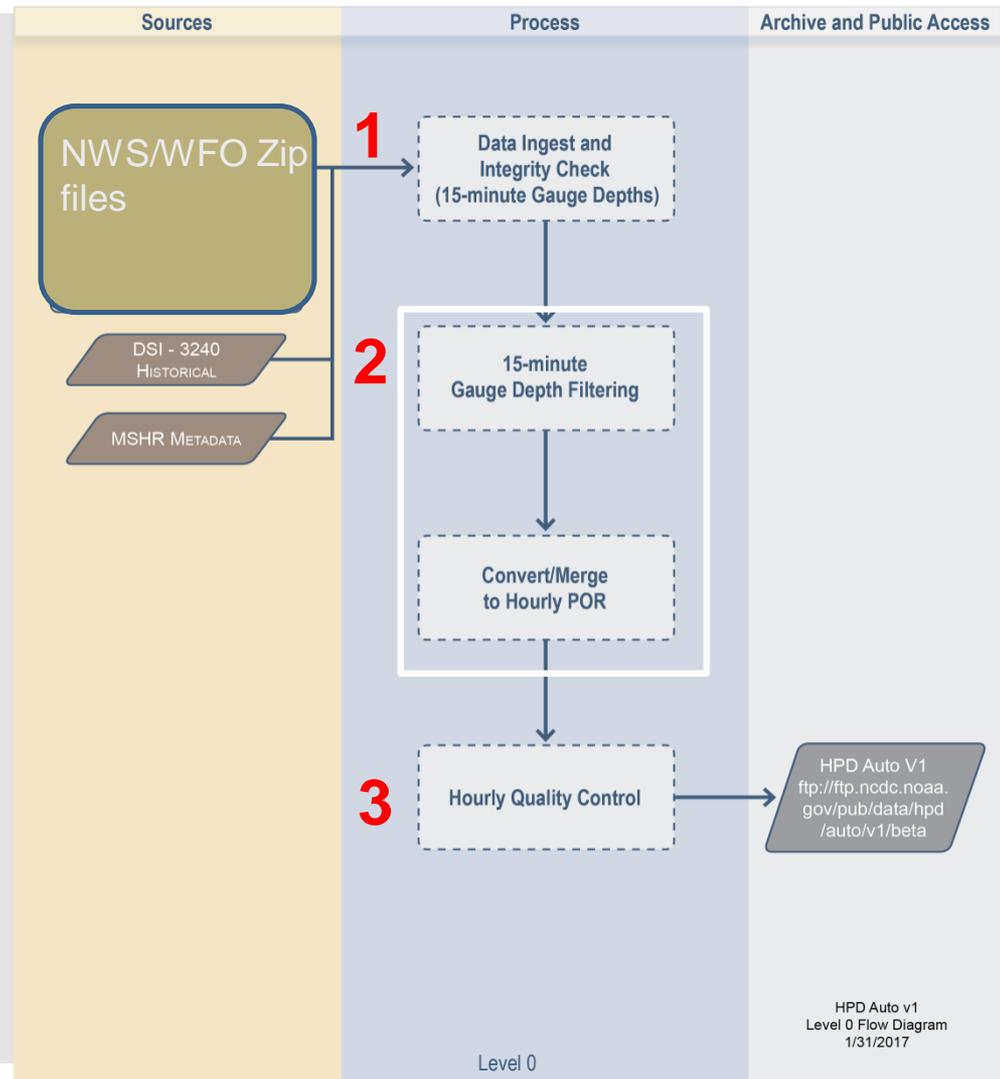
- Dataset known as “HPD” consists of hourly precipitation totals from the NWS Fischer-Porter network of stations across the U.S. and territories
- In operation since the mid-20<sup>th</sup> century
- Provides gauge depths every 15-minutes
  - Summed into hourly totals
  - For decades provided as NCDC’s DSI-3240 dataset
- Punched-paper recording until recently
- NWS “Modernized” the system; 2004-2013
  - Data stored in on-site datalogger and downloaded to a thumb drive during “monthly” site visits.

# Fischer-Porter Rain Gauge



# NCEI HPD Processing Flow

- 3 Main Steps in NCEI process
  - Ingest/Integrity Check
  - Gauge depth QC & conversion to incremental precipitation
  - Period of record hourly QC and data output



# Step 1

- Each WFO consolidates station data into a single Zip file
  - Manual manipulation
  - Delays in transmission
- Each WFO uploads Zip file to NCEI's ftp site
  - File naming conventions not always followed
  - Unrelated files interspersed in Zip file
  - Corrupted data files
  - Duplicated data files
- NCEI conducts several automated checks to identify and resolve the problems.
- In addition NCEI [provides feedback to NWS](#)



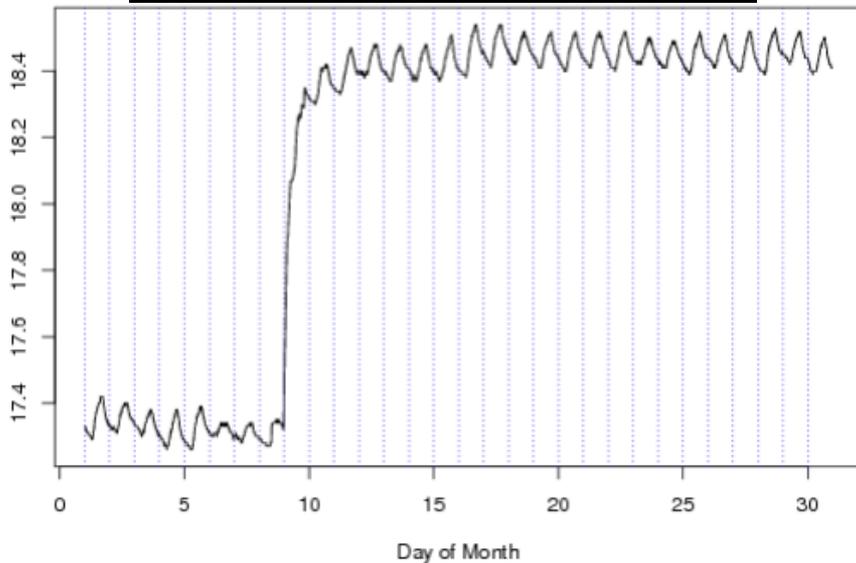
## Step 2: 15-minute Gauge Depth QC and Filtering

- Several algorithms are used to determine the true precipitation signal among what can be a noisy record
  - High frequency oscillations unrelated to precipitation
  - Diurnal variability
  - Malfunctioning gauges

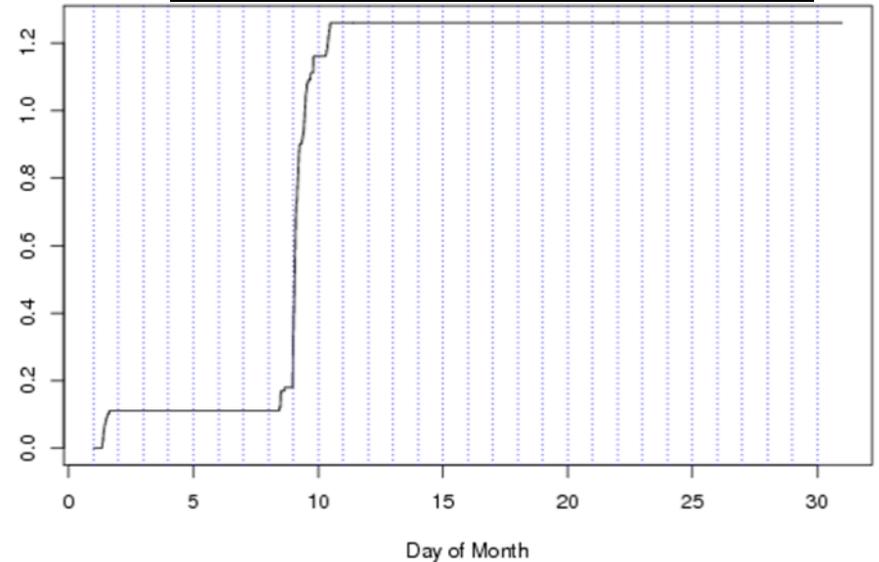
# Step 2: Remove Diurnal Fluctuations

It is often necessary to identify the effect of diurnal heating and reset gauge depths so that precipitation is not computed. Also to identify and remove small negative and positive changes (less than  $\pm 0.03$ " ) within 3-hour windows.

Raw Gauge Depths  
Crown King, AZ 022329; April 2011

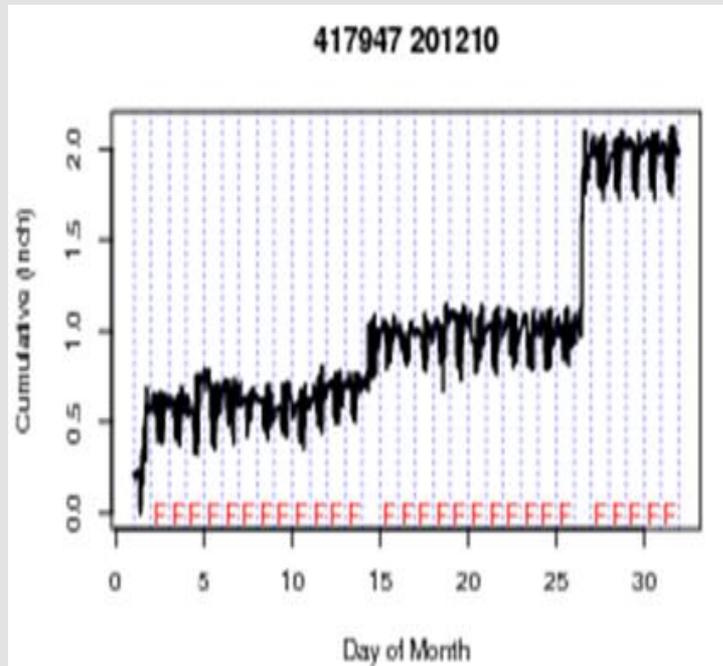


Final Accumulated Precipitation  
Crown King, AZ 022329; April 2011



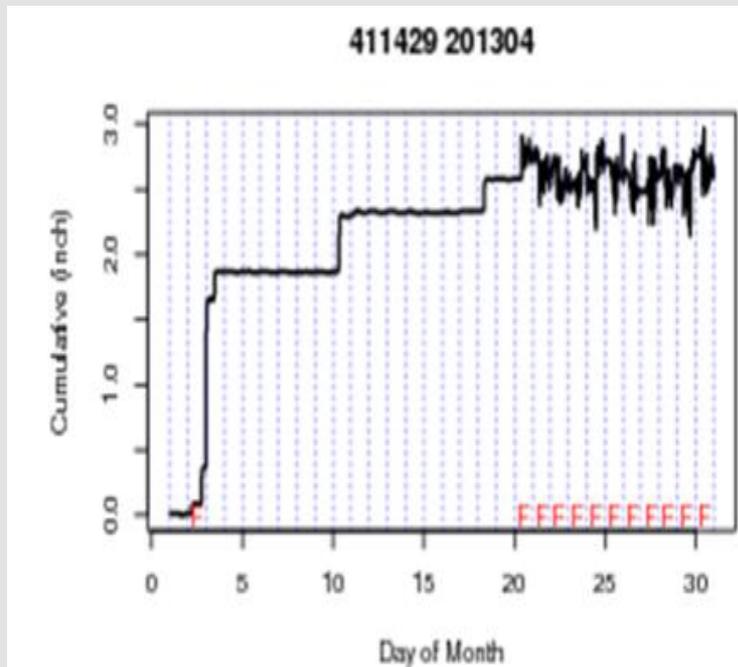
# Step 2: Malfunctioning Gauges

Large fluctuations unrelated to precipitation sometimes occur due to factors such as low voltage power supply or improper placement of bucket and casing during monthly visit.



Latching clip was bent upward slightly on the F&P base causing the lower case assembly to be a little bit more free. 3 latching clips on the base of the F&P and three latching tabs on the lower case assembly.

# Step 2: Malfunctioning Gauges



In some cases erratic behavior will begin after precipitation in the bucket reaches a certain threshold.

Well behaved until about the 20<sup>th</sup> of the month.



# Step 3: Quality Control of Hourly Totals

- Quality control algorithms also are applied on hourly (& daily) timescales
- Algorithms are patterned after those developed for NCEI's GHCN-Daily dataset
- Quality control thresholds were established using the method of Durre and Menne (2008).
- 3 Basic Integrity and 2 Outlier checks



# Hourly QC: Basic Integrity Checks

## Global Hourly Extreme Check

Flag all hourly values that exceed the all-time global record.  
12.0” (Holt, MO, 1947).

## State Daily Extreme Check

Flag all hourly values within the day if the sum daily total exceeds the all-time state precipitation extreme.

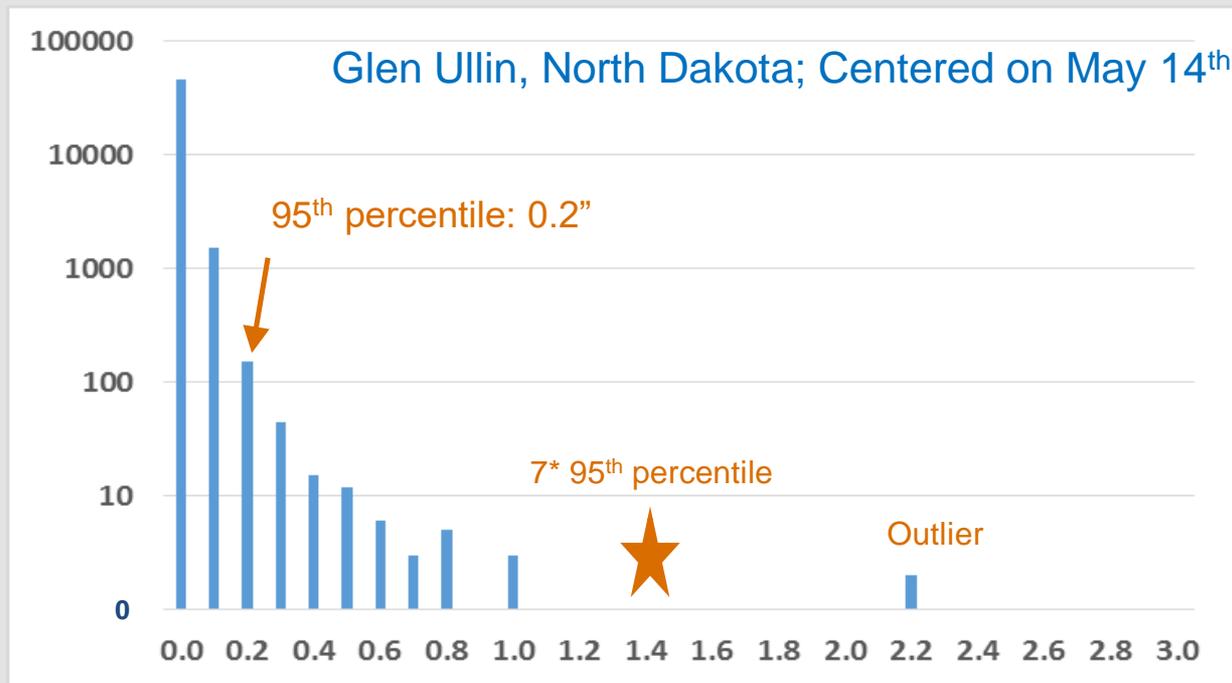
## Streak Check

Flag streaks of 20 or more identical non-zero hours  $\leq 0.3$ ”.  
Flag streaks of 5 or more identical non-zero hours  $> 0.3$ ”.

# Hourly QC: Outlier Check

## Climatological Outlier

Using the distribution of hourly totals within 31-day moving windows, flag hourly values  $>7$  times the station's 95<sup>th</sup> percentile.



Quality control thresholds were established using the method of Durre and Menne (2008).

# Data Access

- HPD data are available
  - Beta Release (3/1/17)
  - <ftp.ncdc.noaa.gov/pub/data/hpd/auto/v1/beta/>
    - Data (ascii fixed format)
    - Readme
    - Documentation
    - Flow Diagrams
- Operational Release in September
  - Will be available in CDO/Common Access

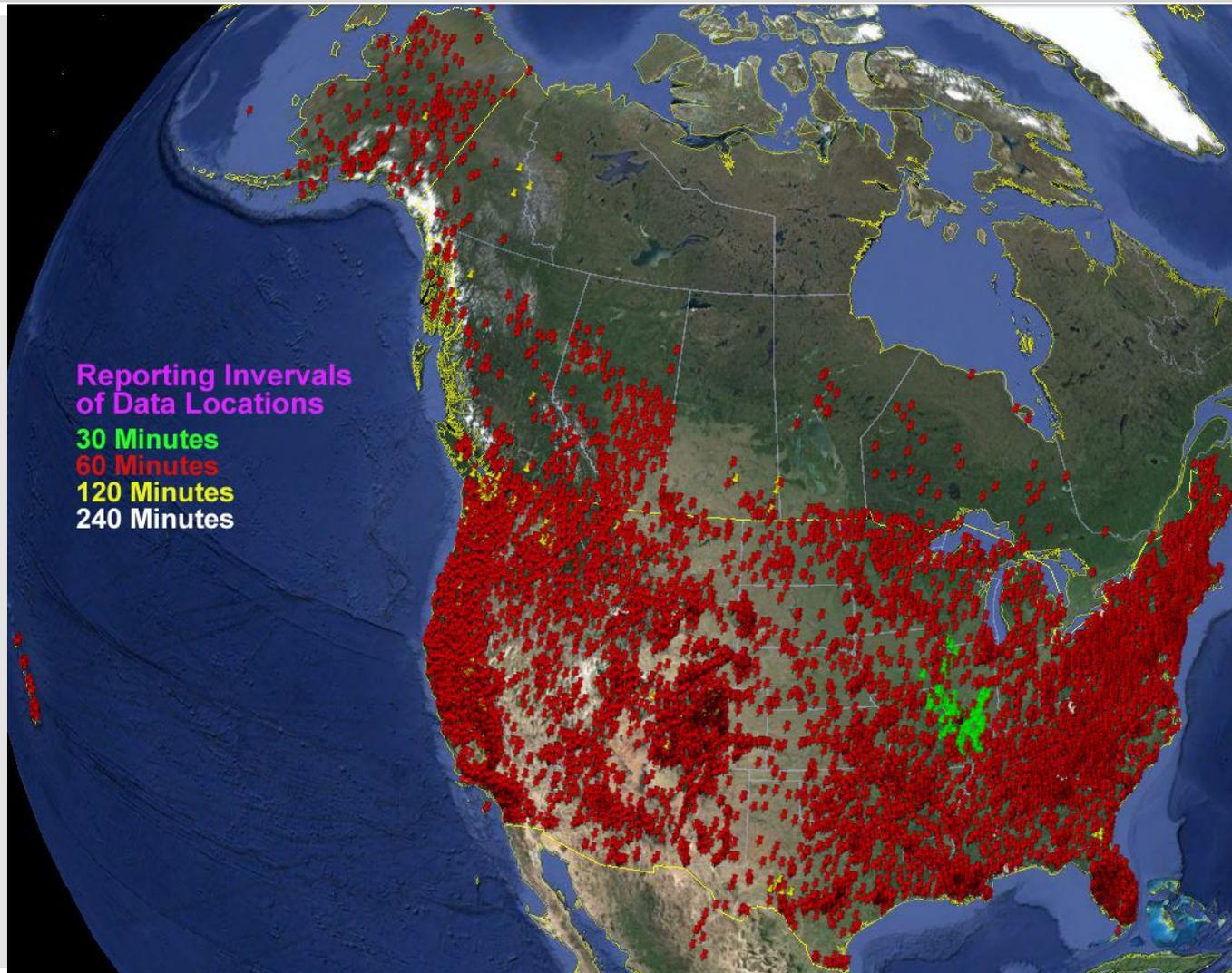


# Hydrometeorological Automated Data System (HADS)

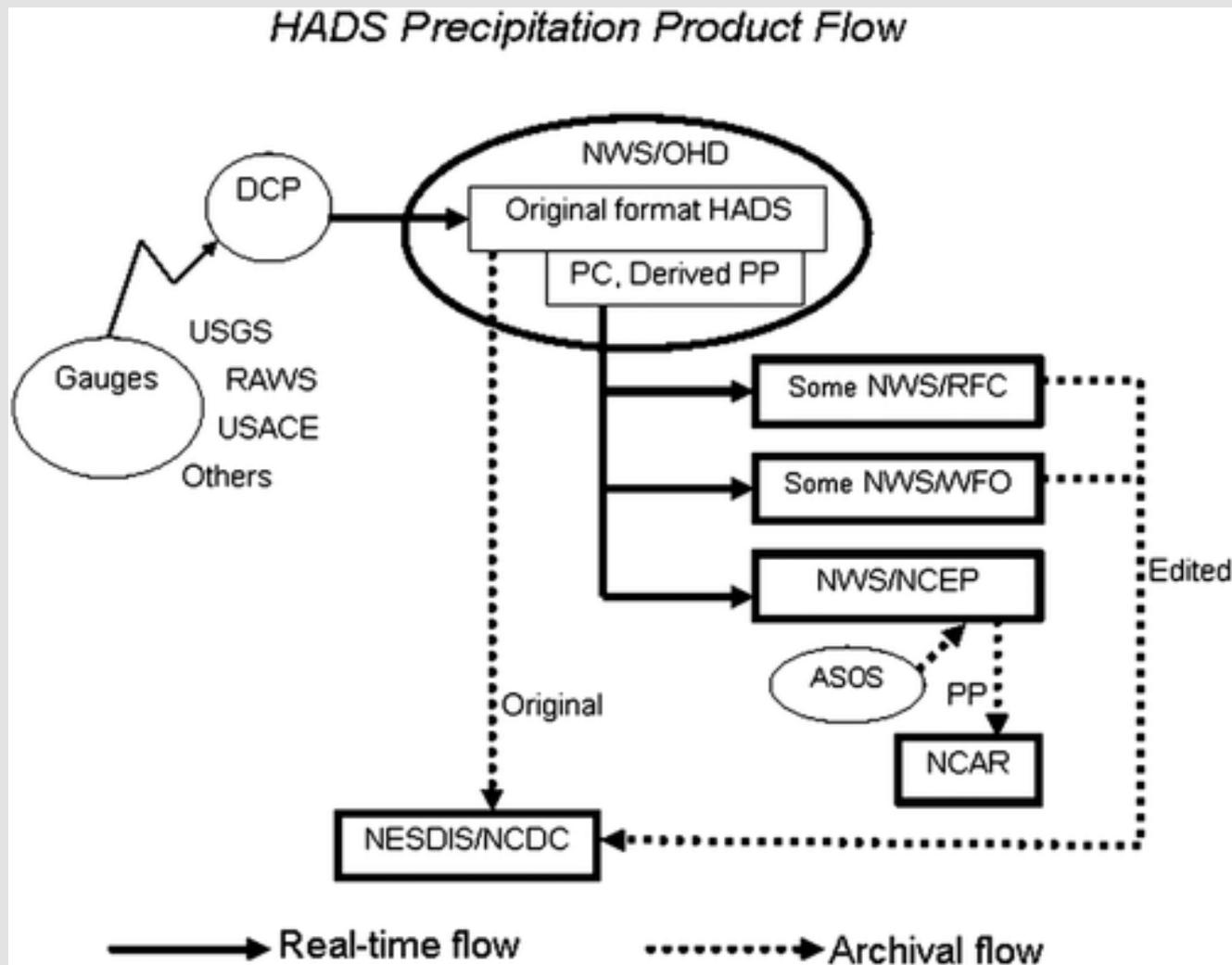
# Hydrometeorological Automated Data System (HADS)

- Real-time and near real-time data acquisition
- operated by the National Weather Service Office of Dissemination
- Raw hydrological and meteorological observation messages from Geostationary Operational Environmental Satellites (GOES) Data Collection Platforms (DCPs)
- Water Resources Division of the U.S. Geologic Survey, the U.S. Army Corps of Engineers, the Tennessee Valley Authority, the Bureau of Land Management, the U.S. Forest Service, the Bureau of Reclamation, and departments of natural resources from numerous state and local agencies throughout the country
- **Archive at NCEI**
- **SHEF format**

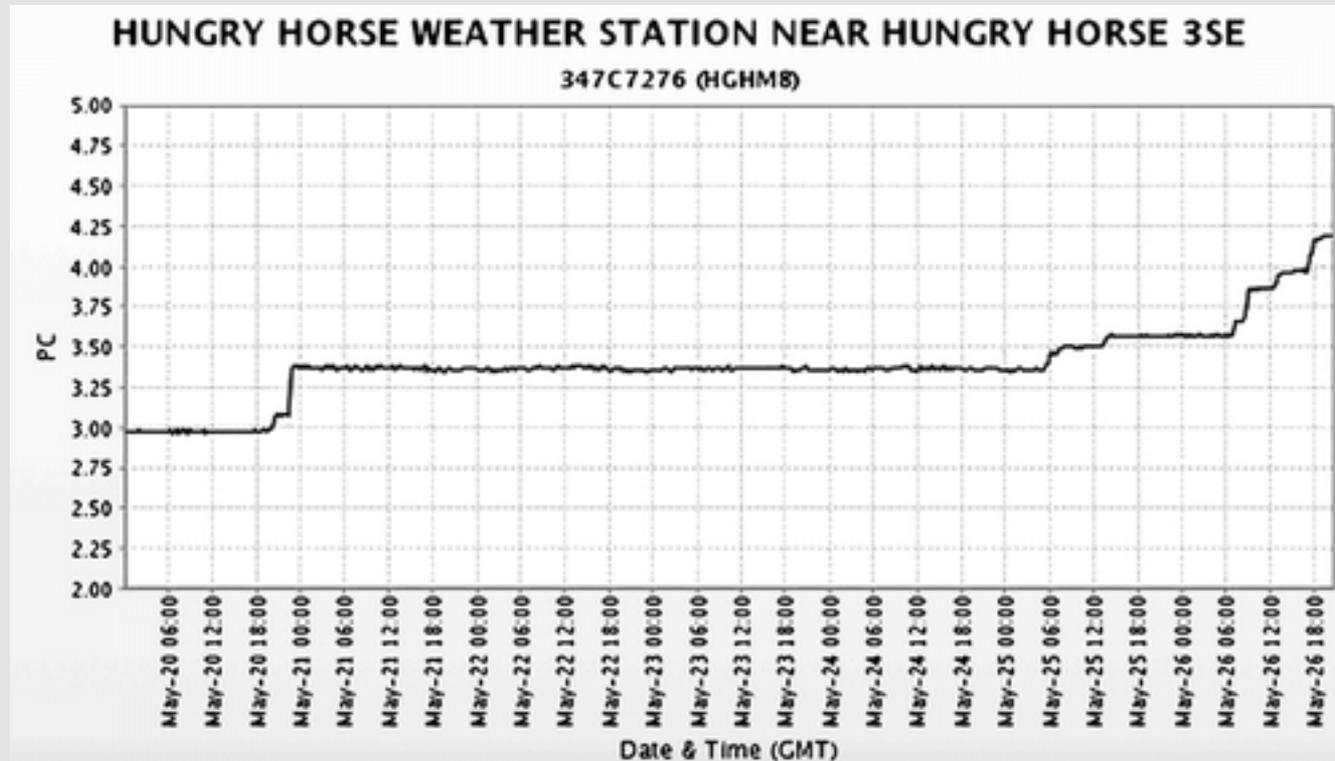
# Hydrometeorological Automated Data System (HADS)



# Hydrometeorological Automated Data System (HADS)



# Hydrometeorological Automated Data System (HADS)





# Global Historical Climatological Network (GHCN)

# The International Collection

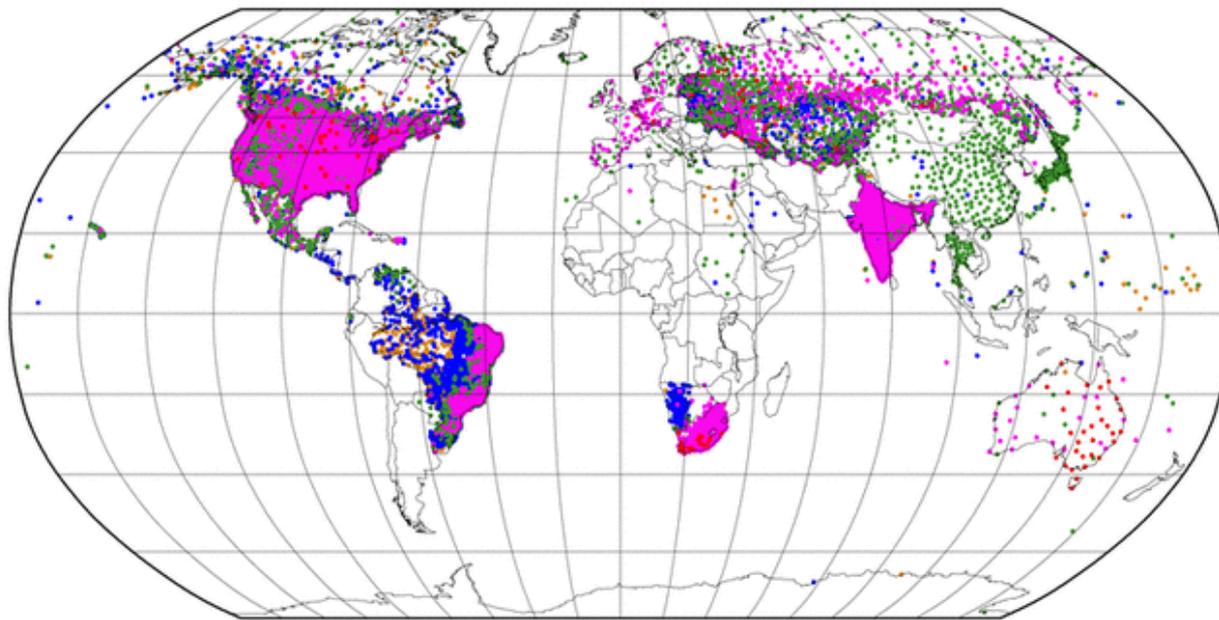
Region/Country	Source/Contact
Countries in West Africa	MeteoFrance
Countries in East Africa	Kenyan Meteorological Department/P. Ambenji
South Africa and Namibia	South African Weather Service/R.S. Vose
China	National Climate Center China Meteorological Administration/D.R. Easterling
India, Japan, Thailand	National Center for Atmospheric Research
Brazil	ANEEL (Agencia Nacional De Energia Electrica)/P.Ya. Groisman
Paraguay, Uruguay, Venezuela	NOAA's Climate Diagnostics Center
Mexico	National Weather Service of Mexico/ A. Douglas
Countries in the Former USSR	Bilateral Exchange/P.Ya. Groisman
Europe	European Climate Assessment and Dataset [Early Version] ( <a href="http://eca.knmi.nl/">http://eca.knmi.nl/</a> )

# The First Global Daily Dataset

- The Global Daily Climatology Network (GDCN)—  
Released on CD in July 2002
- Compiled from data obtained through personal contacts and data from Environment Canada
- Plus two (?) U.S. daily data archives
  - 3200 – Cooperative Observer Summary of the Day
  - 3210 – First Order Summary of the Day

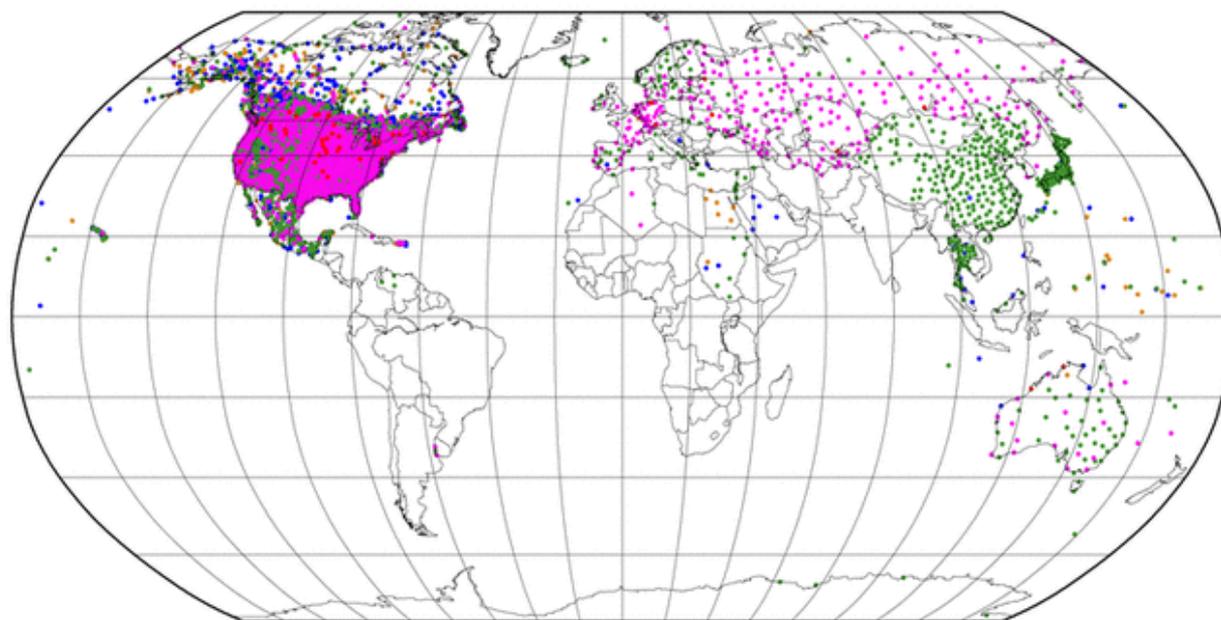


## Precipitation, Period of Record (POR) GDCN V1.0



- 00 yrs < POR ≤ 10 yrs
- 10 yrs < POR ≤ 25 yrs
- 25 yrs < POR ≤ 50 yrs
- 50 yrs < POR ≤ 100 yrs
- 100yrs < POR

## Maximum Temperature, Period of Record (POR) GDCN V1.0



- 00 yrs < POR ≤ 10 yrs
- 10 yrs < POR ≤ 25 yrs
- 25 yrs < POR ≤ 50 yrs
- 50 yrs < POR ≤ 100 yrs
- 100yrs < POR

# Quality Assurance of GHCN-Daily

- 19 different checks most of which are applied to each of the “fab five” elements when appropriately tailored [TMAX, TMIN, PRCP, SNOW, SNWD].
  - Basic integrity checks for other dozens of elements
- Low false positive rate overall (i.e., very limited “collateral damage”)
- Total flag rate equal to approximately 0.24% of all values (highest flag rates for snowfall and snow depth). 1-2% of the flags are estimated to be false positives (i.e., valid values flagged as bad)
  - System is run “unsupervised”
- Uniform QC for full period of record (unset all legacy QC flags)

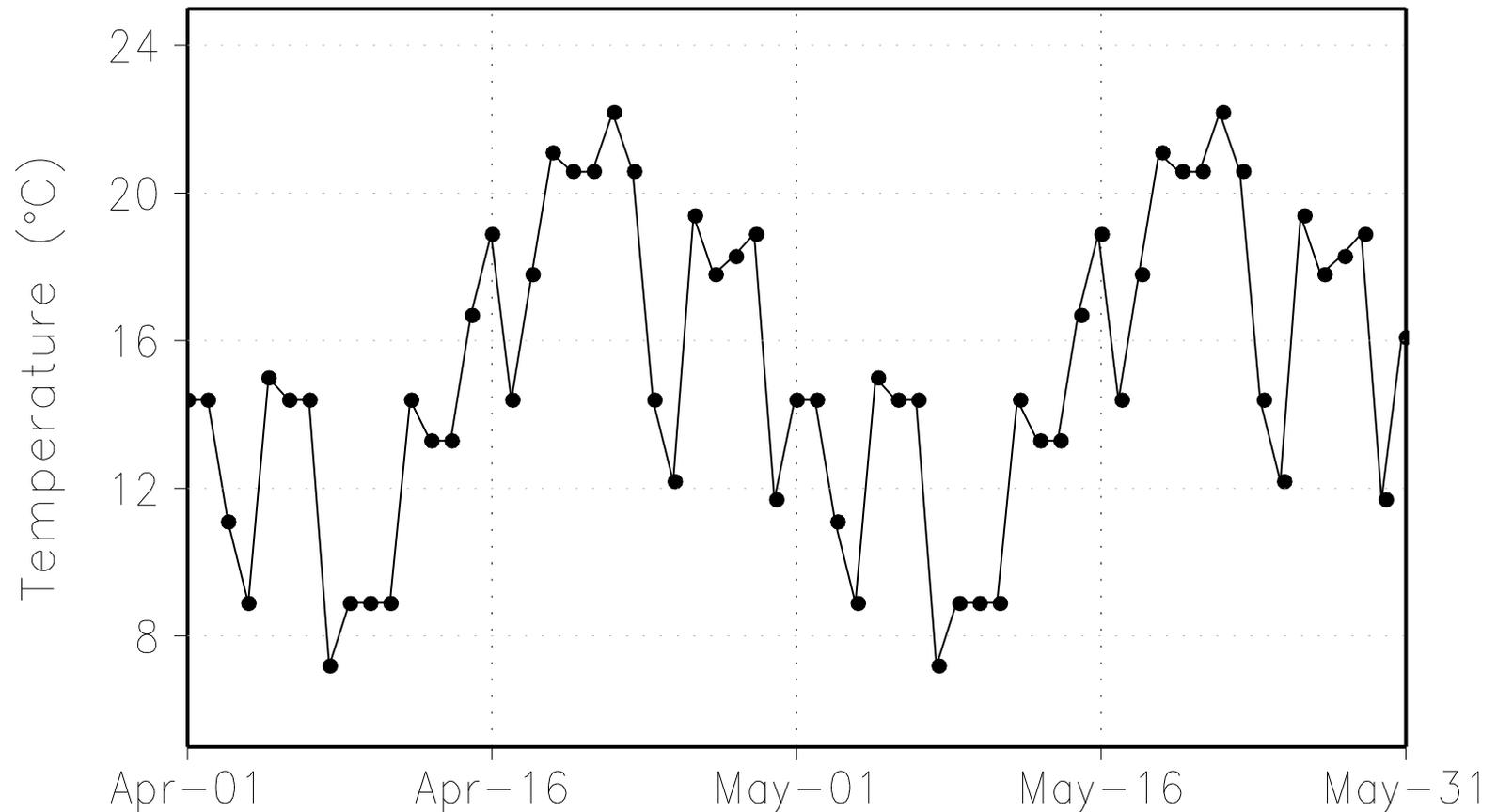


# Quality Assurance

- **Basic integrity**
- **Outlier**
- **Internal and Temporal Consistency**
- **Spatial Consistency**

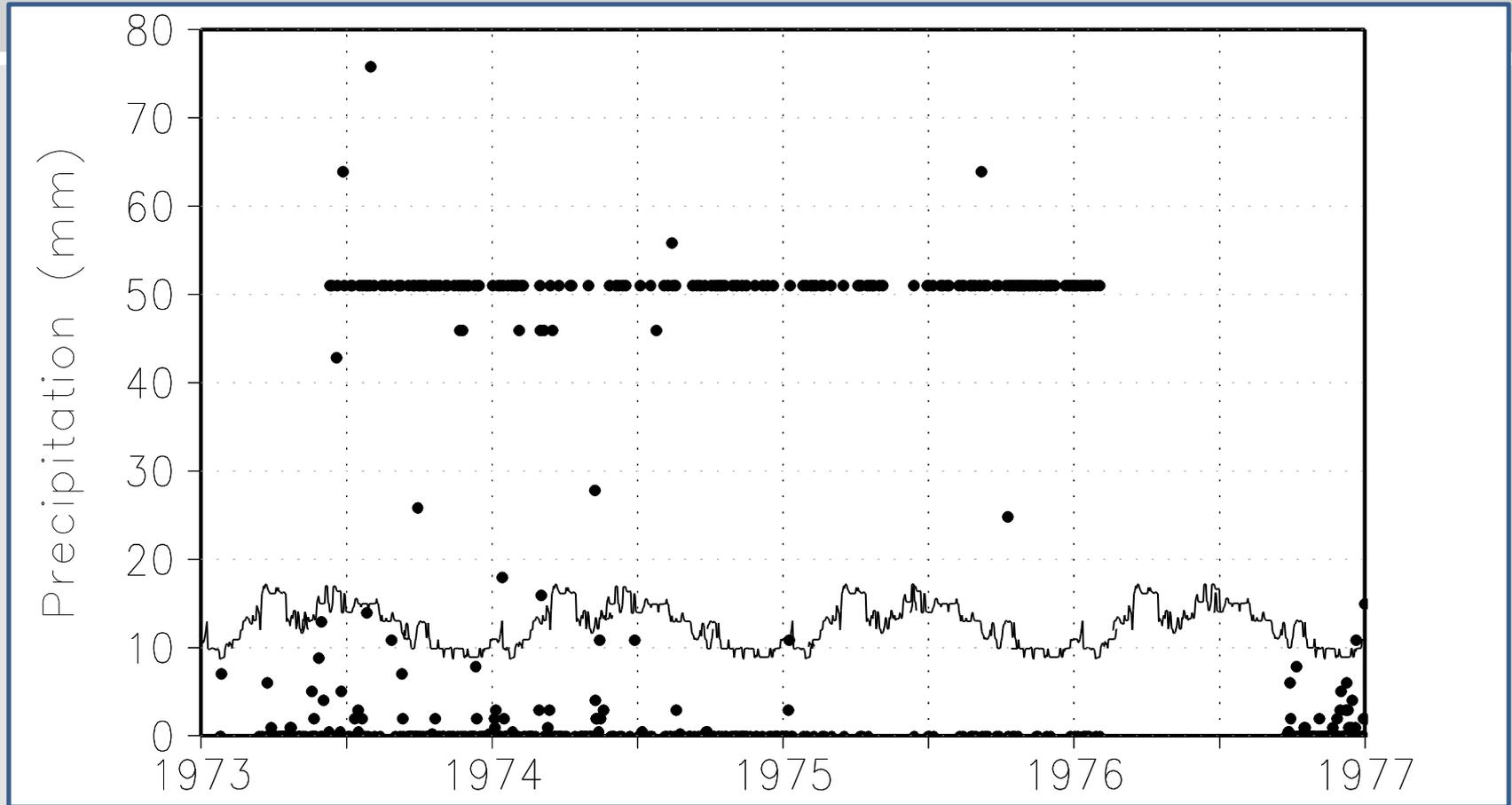
**(Described in *Durre, Menne, Gleason, Houston and Vose 2010*)**

# Example: Duplicated Data



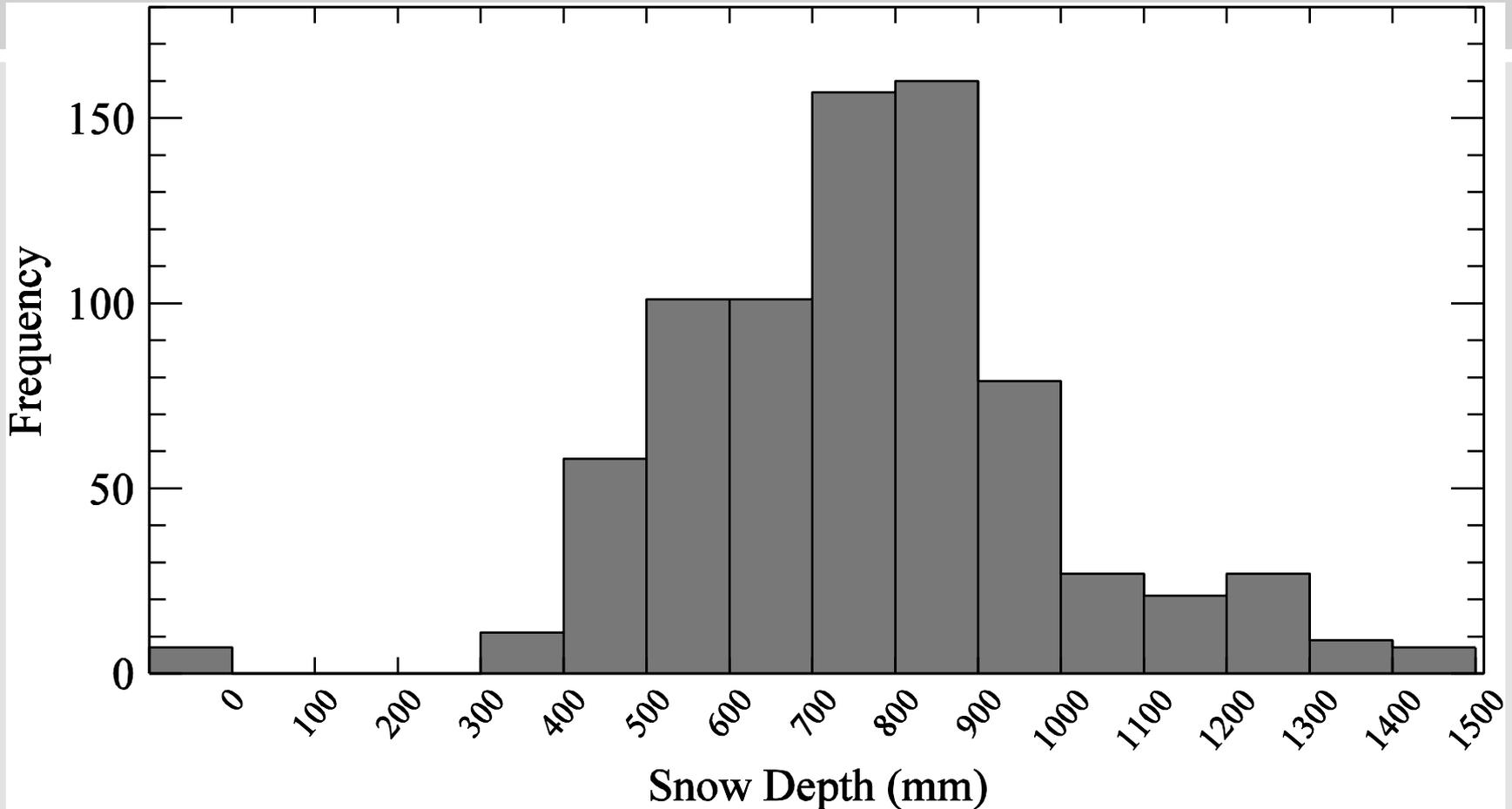
**Daily maximum temperatures during April and May 1967 at Lardeau, Canada (GHCN-Daily station ID = CA001144580), showing an example of data duplication identified by the duplicate check comparing data from different months within a year**

# Example: Frequent Value



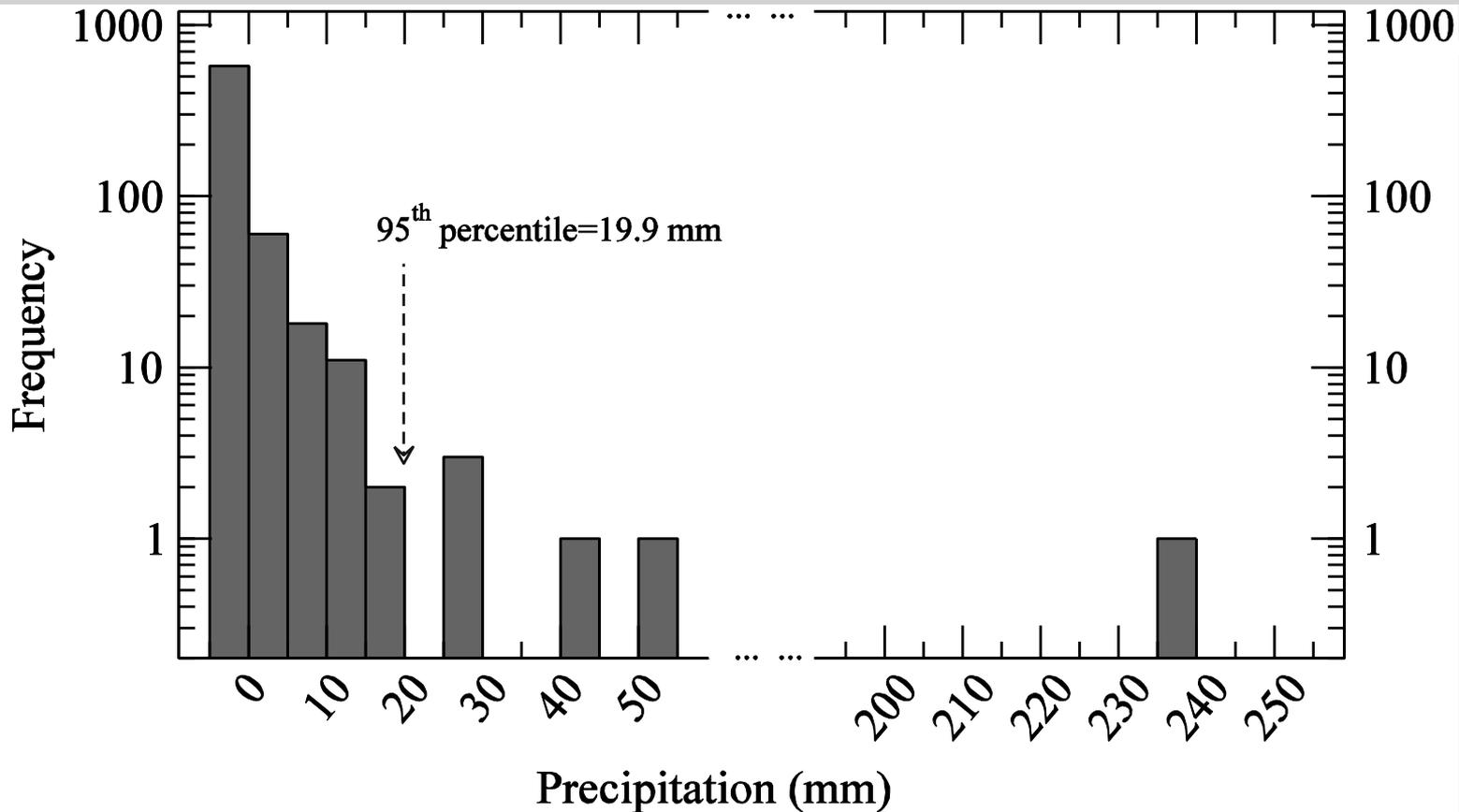
**Time series of daily precipitation totals (solid line) during 1973-1976 at Balmaceda, Chile (GHCN-Daily station CI00085874), containing 162 values of 51.1 mm that are flagged by the frequent-value check**

# Example: Gap in distribution



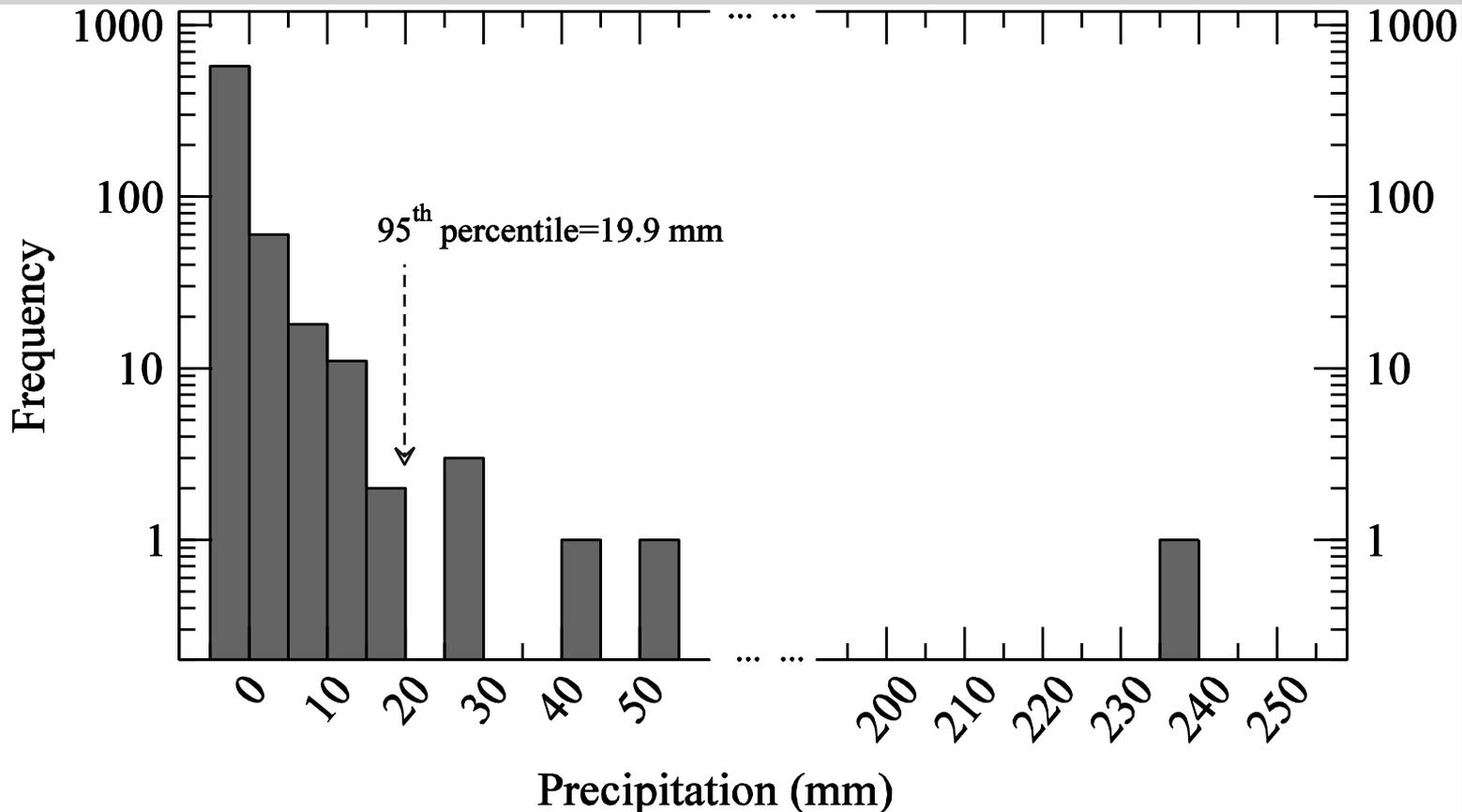
Histogram of all daily snow depths observed in March during the period of record (1975-2008) at Paxson, Alaska (GHCN-Daily station USC00507097), illustrating a data problem identified by the gap check (Table 2). The values of zero (reported in March 1982, 2004, and 2007) are flagged because they differ from the next lowest value ever reported in that calendar month by more than the threshold of 350 mm.

# Example: Climatological Outlier



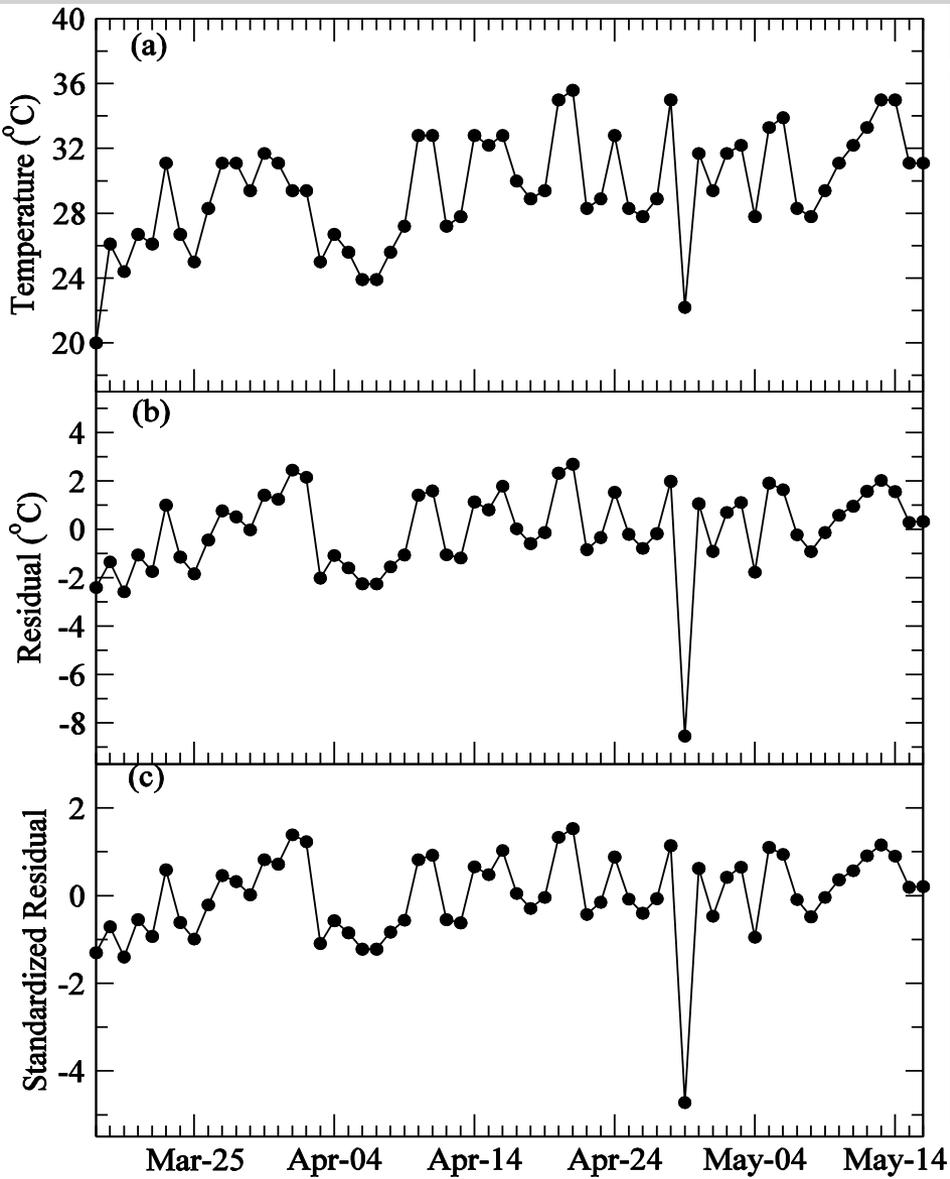
Histogram of daily precipitation totals reported between August 6 and September 3 throughout the 1966-1990 period of record at Gold Hill, Utah (GHCN-Daily station USC00423260), showing an outlier flagged by the percentile-based climatological outlier check

# Example: Climatological Outlier



Histogram of daily precipitation totals reported between August 6 and September 3 throughout the 1966-1990 period of record at Gold Hill, Utah (GHCN-Daily station USC00423260), showing an outlier flagged by the percentile-based climatological outlier check

# Example: Spatial Outlier



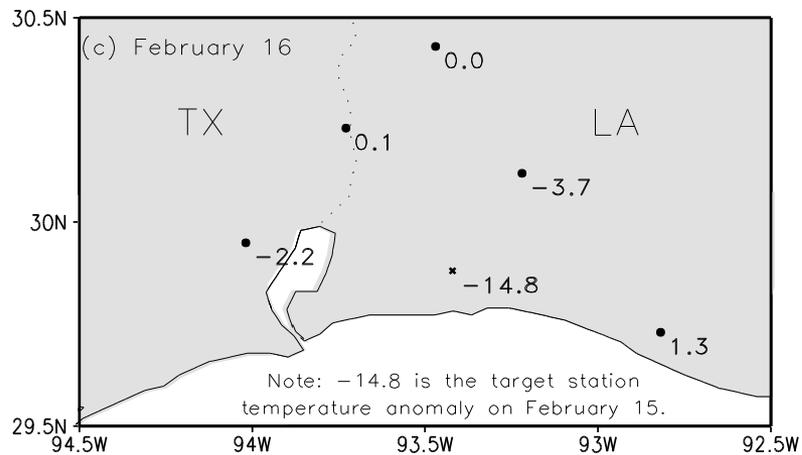
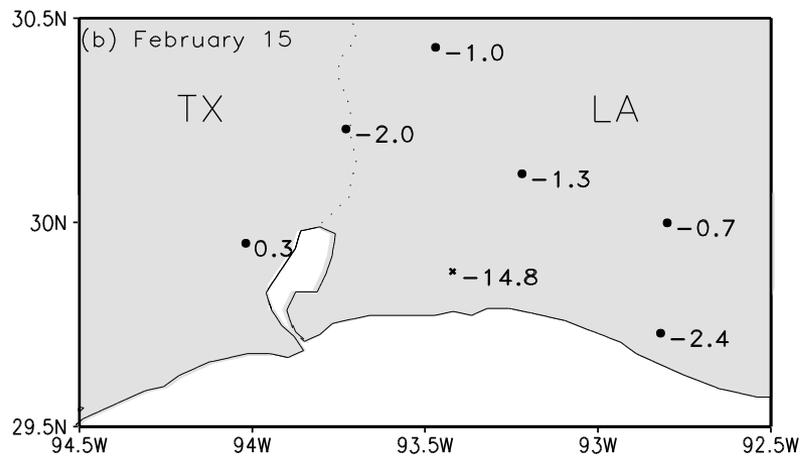
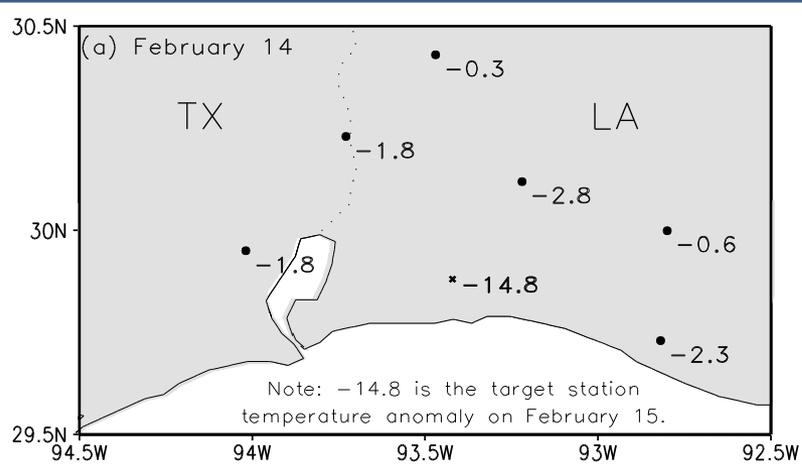
Time series containing a temperature flagged by the spatial regression check

- (a) Daily maximum temperatures at Bracketville, Texas (GHCN-Daily station USC00411007) between 17 March and 15 May, 1991;
- (b) the corresponding residual time series; and;
- (c) the time series of standardized residuals. The temperature of 22.2°C on 28 April is flagged because the residual and standardized residual on that day are greater than 5°C and 4.0 standardized units, respectively

# Example: Spatial Corroboration

Maps illustrating the spatial corroboration check on temperature. Shown are the daily minimum temperature anomaly at Hackberry, Louisiana (GHCN-Daily station USC00163979), on 15 February 2002 and the daily minimum temperature anomalies to which this "target value" is compared:

(a) the six available neighbor anomalies on day -1 (14 February);  
(b) the six neighbor anomalies available on day 0 (15 February); and  
(c) the five neighbor anomalies available on day +1 (16 February). The target value is indicated by an X symbol in each panel, the neighbor values by filled circles. The target anomaly of  $-14.8^{\circ}\text{C}$  is flagged because it is  $11.1^{\circ}\text{C}$  lower than the coldest temperature anomaly among the neighbor values within the three-day window.

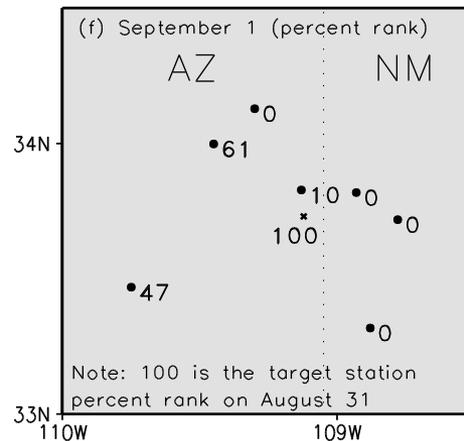
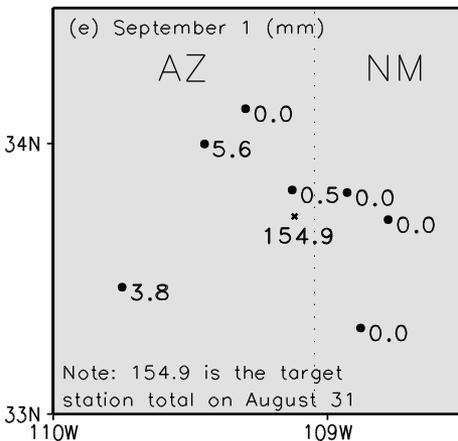
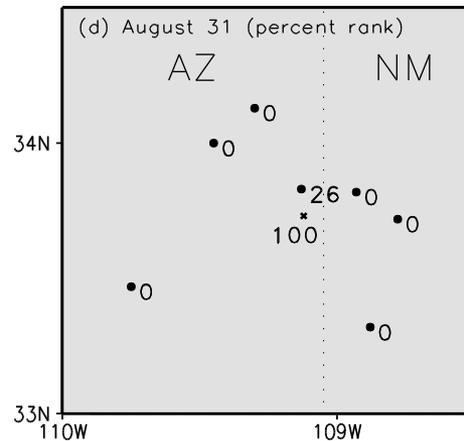
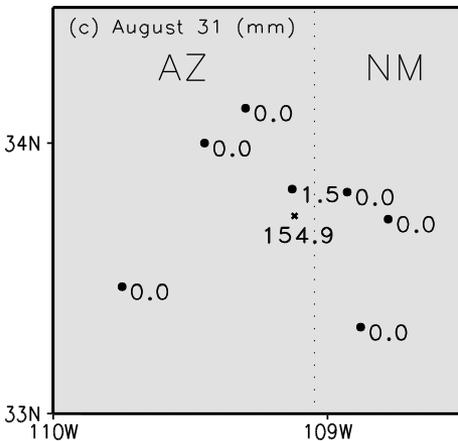
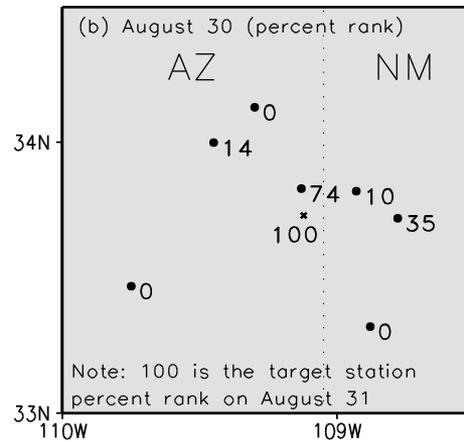
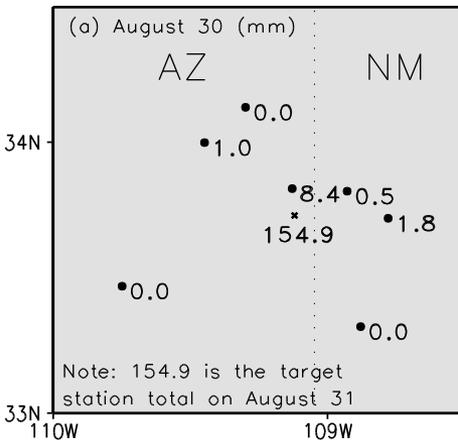


# Example: Spatial Corroboration

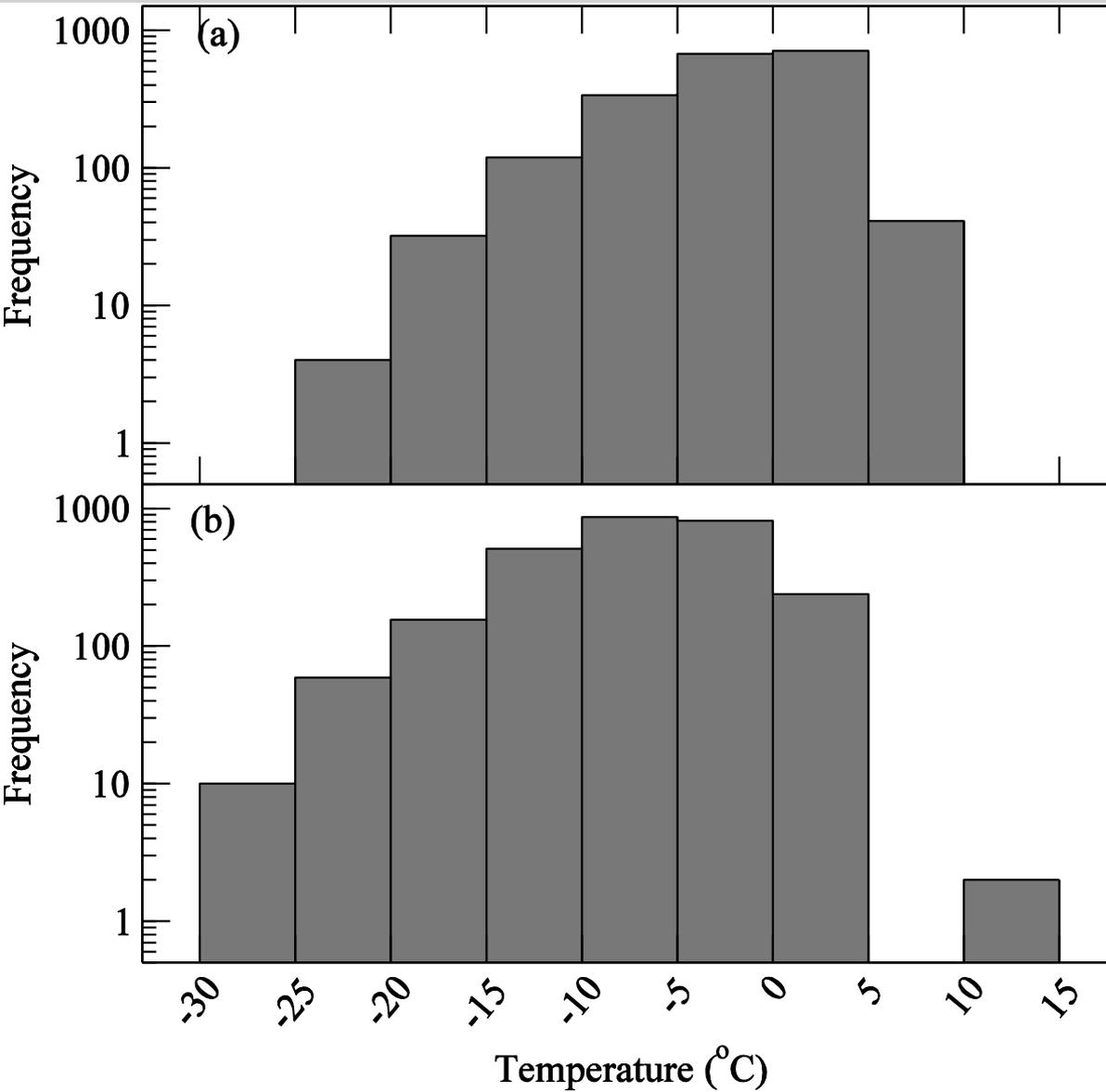
Maps illustrating the spatial corroboration check (Table 4) applied to a 154.9 mm precipitation total at Alpine, Arizona (GHCN-Daily station USC00020174), on 31 August 1996. In addition to this target total or its percent rank (X symbol), the maps show all neighbor information (filled circles) used in the check:

- (a) neighbor totals on day -1;
- (b) neighbor percent ranks on day -1;
- (c) neighbor precipitation totals on day 0;
- (d) neighbor percent ranks on day 0;
- (e) neighbor totals on the day +1; and
- (f) neighbor percent ranks on day +1.

The minimum absolute target-neighbor percent ranked difference is 26, yielding a test threshold of 120.3 mm (Appendix C). The target value is flagged because the minimum absolute target-neighbor difference among totals is 146.5 mm and therefore exceeds the threshold.



# Example: Megaconsistency



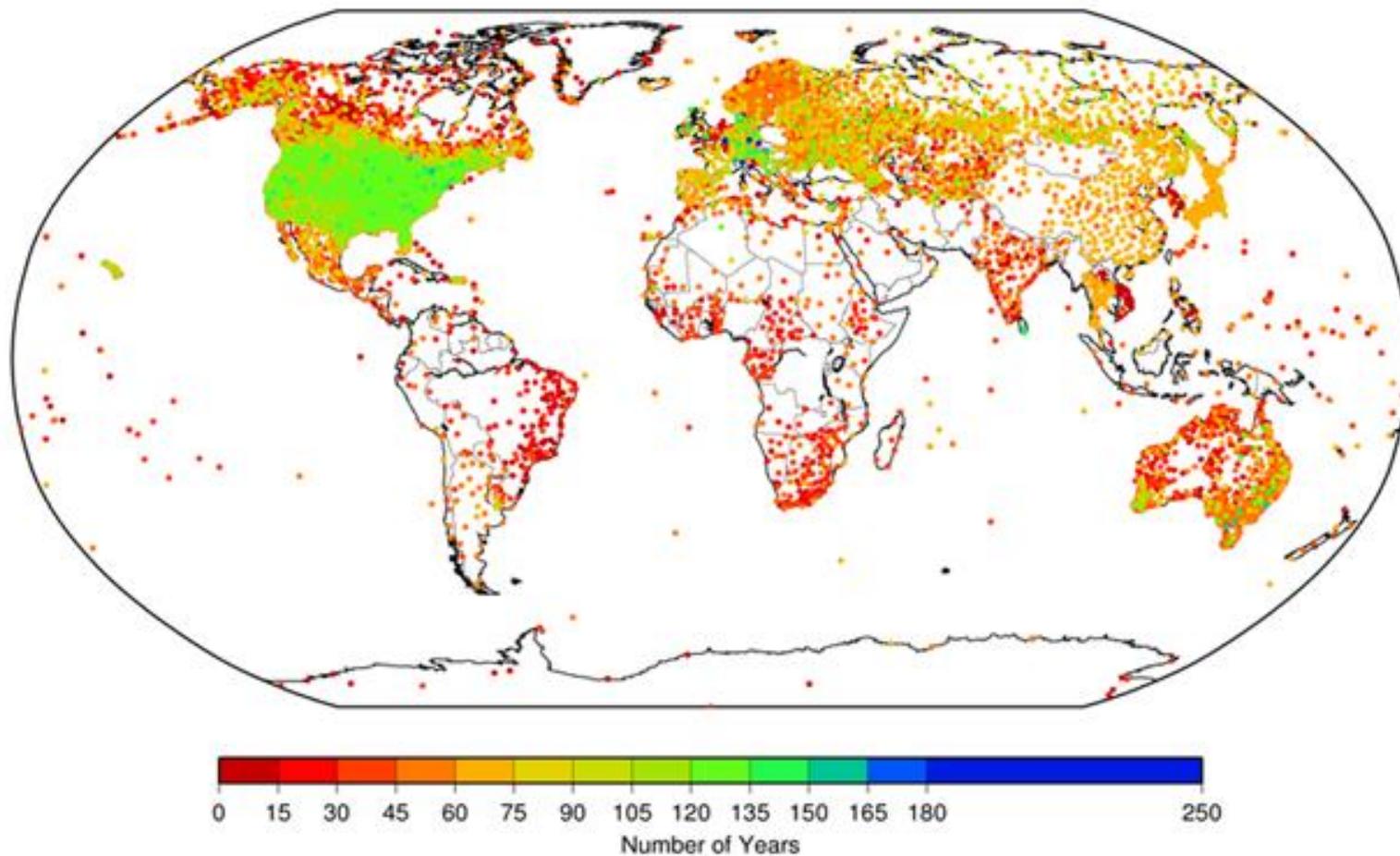
## Histograms of all January

- (a) daily maximum temperatures and
- (b) daily minimum temperatures

reported at Jan Mayen, Norway (GHCN Daily station JN000099950), illustrating the extremes megaconsistency check. The 10.3 $^{\circ}\text{C}$  and 10.6 $^{\circ}\text{C}$  TMINs (both reported in January 1929) are flagged by the check because they exceed the highest unflagged January TMAX (9.5 $^{\circ}\text{C}$ ) reported during the station's 1921-2009 record.

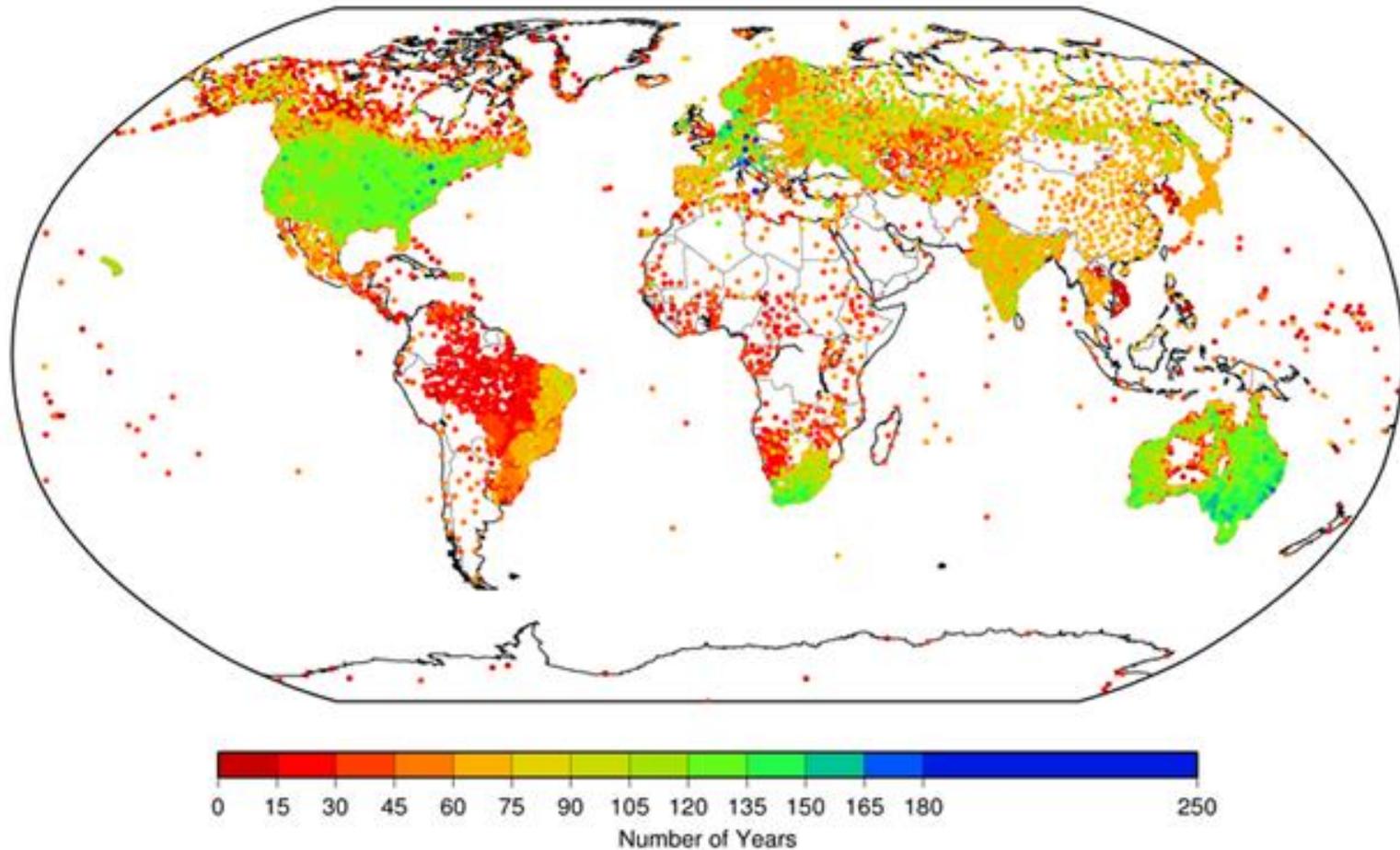
# Daily Temperature

Daily Max/Min Temperature Period of Record [GHCN-Daily Version 3.12-por-2014032113]

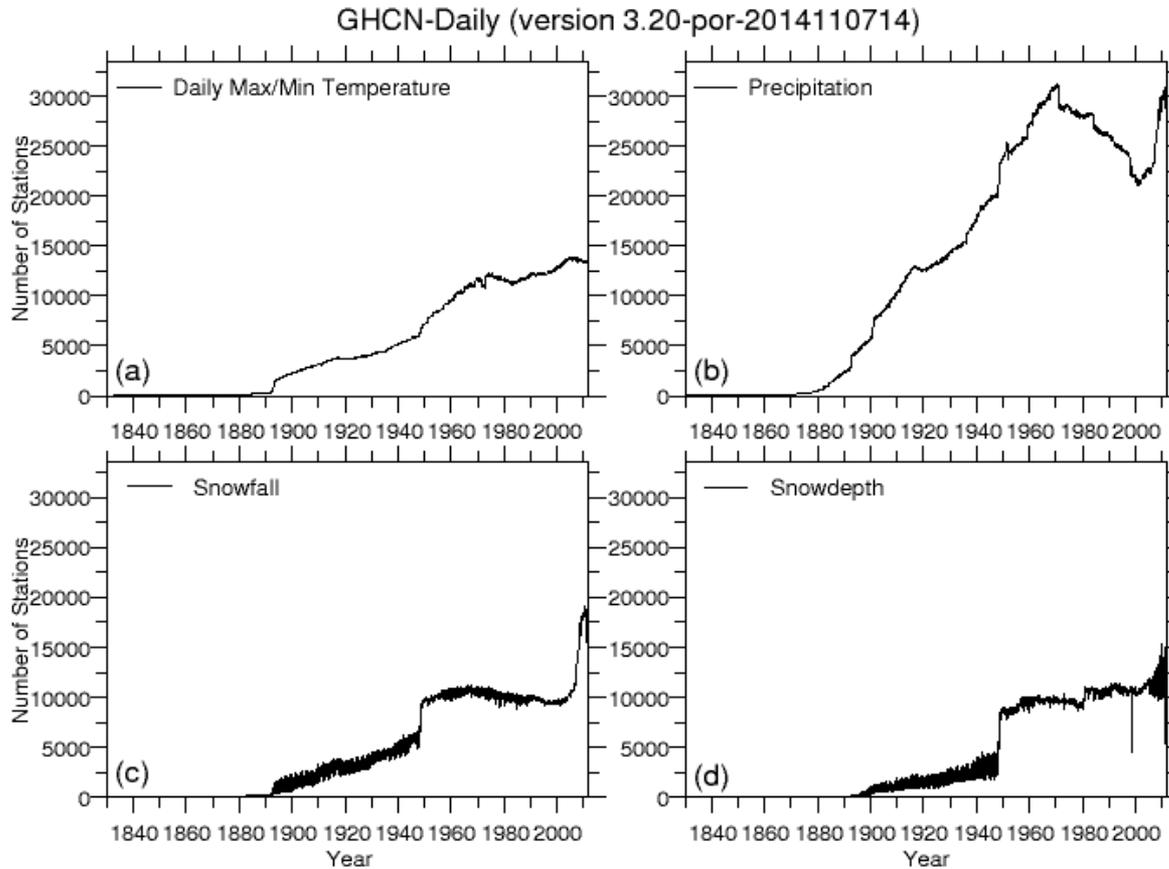


# Daily Precipitation

Daily Precipitation Period of Record [GHCN-Daily Version 3.12-por-2014032113]



# Number of Stations by Year



# Datzilla

Class	Action	Description	Original Value needed in ticket	Alternative Value provided in ticket
1a	Flag It	Original Value is not correct. It should be flagged as a data error. No alternative is provided.	Yes	NA (-9999)
1b*		Original Value is not correct. A previously supplied alternative should also be removed.	Yes (but need OV not EV for GHCN-D)	NA (-9999)
2a	Replace It	Original Value is not correct. It was miskeyed, erroneously transmitted, etc. A new "original" value is provided with this ticket.	Yes	Yes
2b*		Original Value is not correct. A previous alternative value should also be removed. A rekeyed "original" value is provided with this ticket.	Yes (but need OV not EV for GHCN-D)	Yes
3a	Delete It	Original Value is not correct. This value should not be in the database as an original value and needs to be set to missing. No alternative is provided.	Yes	NA (-9999)
3b		Original value is not correct. An alternative value is provided based on an independent source or evidence (e.g., an alternative sensor etc.). The alternative value supersedes the original value.	Yes	Yes
4a	Don't Flag It	Original Value is correct and should not be flagged. It has been flagged by the system at some time in the past, but has been determined to be legitimate.	Yes	NA (-9999)
4b		Original Value is correct and should not be flagged. Although not flagged when this ticket was submitted, the value was unusual enough to be investigated as a potential error. Based on the investigation, the original value is deemed correct.	Yes	NA (-9999)
5	Add It	Original Value was missing. A newly available original value has been keyed via this ticket.	No	Yes
NA		Station identification number was incorrect and needs to be reassigned. Needs to be handled manually. Original source will be updated.	No (but could be -9999)	No (but could be -9999)

# Datzilla Statistics

- **703** Datzilla Tickets have been resolved by NCEI Staff between 2/1/2015 through 1/31/2016 (262 work days per year)
- **55** is the average number of tickets resolved each month during the 12-month period ending on 1/31/2016.
- Datzilla tickets have been resolved at roughly twice the rate since U.S. data processing handled by GHCN-Daily

# HOMR's *In Situ* Station History

<b>Identifiers</b>	Consolidation of IDs over time (ICAO, WBAN, FAA, WMO, COOP, GHCN-Daily...)
<b>Names</b>	Stations can have many aliases
<b>Locations</b>	Latitude/longitude, elevations, topography, obstructions, relocations
<b>Elements</b>	Observation times, reporting methods
<b>Equipment</b>	Types, modifications and siting

Station history management is similar to building a dataset or product – acquire, QA, integrate, manage, provide access.



# Data and Metadata Mapping

## Issues: Example Datzilla Ticket # 5575

From NWS: Station numbers 506166 and 504766 are both mapped to King Salmon, Alaska, WSO, WBAN 25503. Prior to the early 1940s there was a coop at the village of Naknek, about 20km west of the King Salmon Airport, which was built during WW2.

It appears that the name "Naknek" was applied to the King Salmon airport for a few years in 1940s and 50s. In any case, the duplicate stations numbers appear to be causing some users (e.g. RCCs) problems. The coop site at Naknek and the King Salmon airport should be separate climate sites as Naknek is on the coast and King Salmon 20km inland.

[King Salmon AP, AK](#) [Naknek, AK](#)



# U.S. Climate Reference Network (USCRN)

<https://www.ncdc.noaa.gov/crn/>



# NCEP Stage IV

[https://www.emc.ncep.noaa.gov/mmb/ylin/p  
cpanl/stage4/](https://www.emc.ncep.noaa.gov/mmb/ylin/p<br/>cpanl/stage4/)

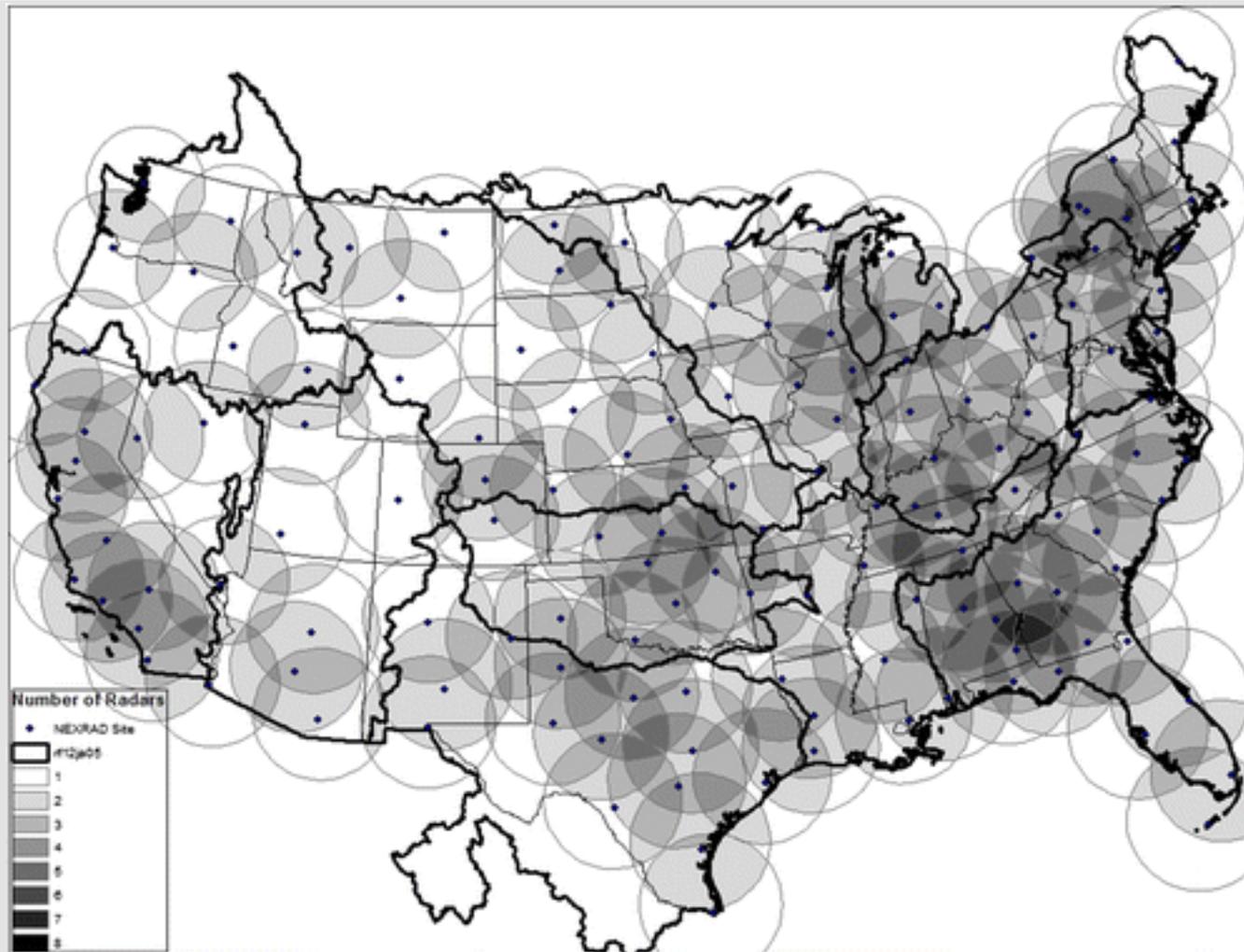


# NCEP Stage IV

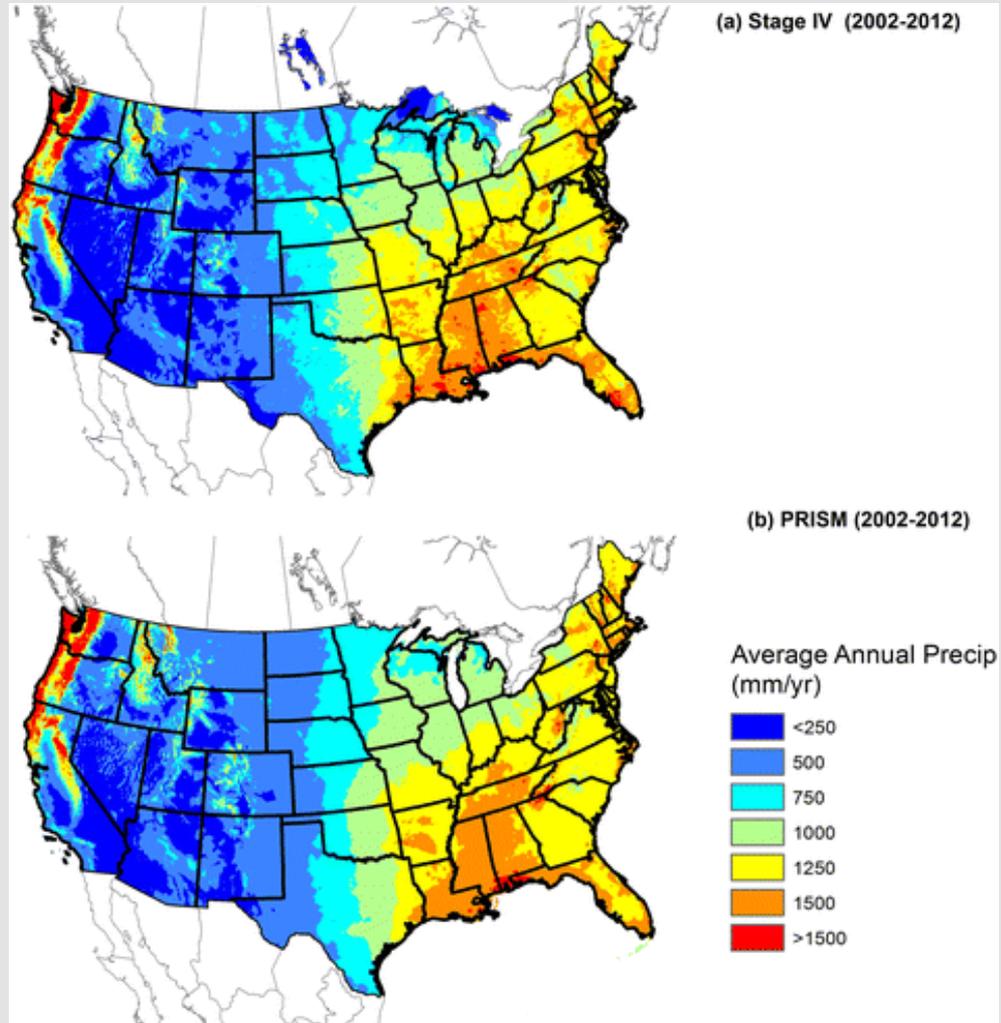
## NCEP Stage IV

Nelson, B.R., O.P. Prat, D. Seo, and E. Habib, 2016: [Assessment and Implications of NCEP Stage IV Quantitative Precipitation Estimates for Product Intercomparisons](#). *Wea. Forecasting*, **31**, 371–394, <https://doi.org/10.1175/WAF-D-14-00112.1>

# NCEP Stage IV



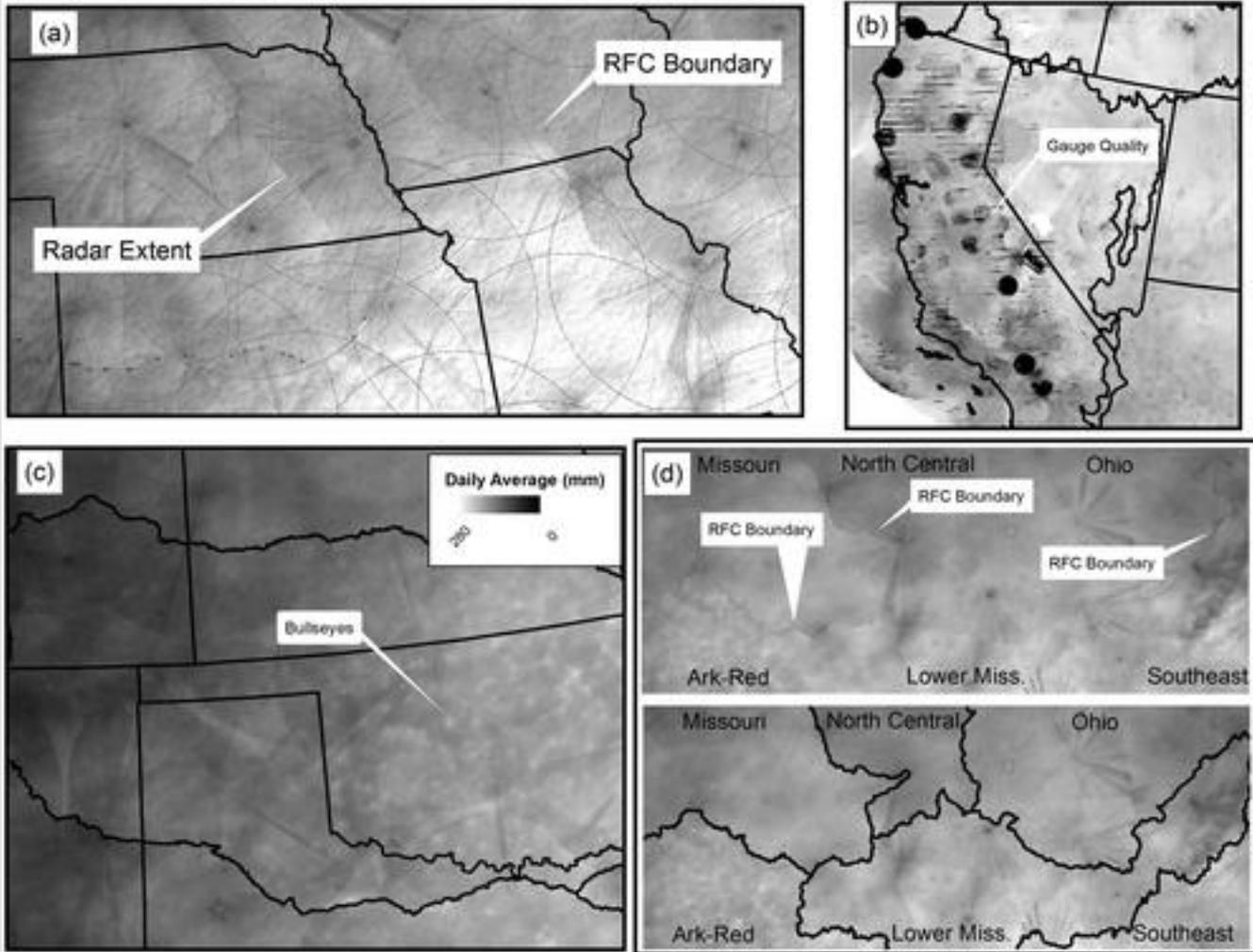
# NCEP Stage IV



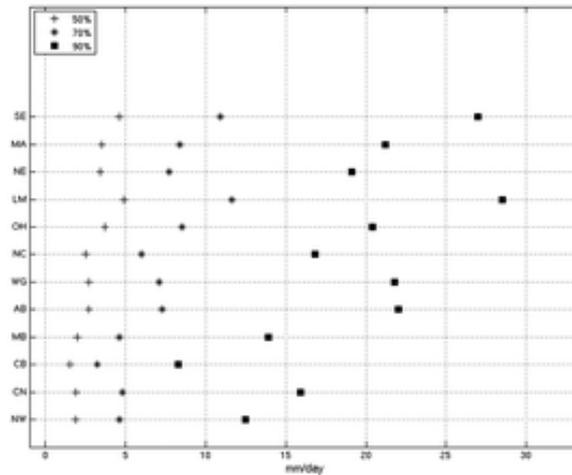
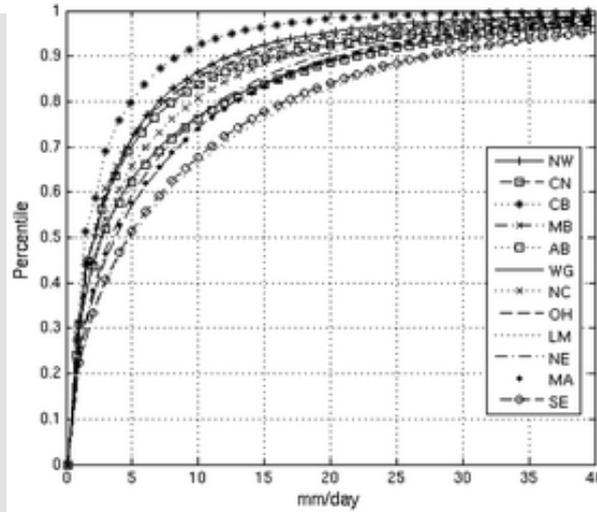
# NCEP Stage IV



# NCEP Stage IV



# NCEP Stage IV





Thank You



# References

Durre, I., M.J. Menne, and R.S. Vose, 2008: Strategies for evaluating quality assurance procedures. *J. Appl. Meteor. Climatol.*, **47**, 1785-1791.

Durre I., M. J. B.E. Gleason, T. G. Houston, and R. S. Vose, 2010: Comprehensive automated quality assurance of daily surface observations. *J. Applied Meteor. and Climatol.*, **49**, 1615-1633, [doi.10.1175/2010JAMC2375.1](https://doi.org/10.1175/2010JAMC2375.1)

Kim, D., B. Nelson, and D. Seo, 2009: [Characteristics of Reprocessed Hydrometeorological Automated Data System \(HADS\) Hourly Precipitation Data](https://doi.org/10.1175/2009WAF2222227.1). *Wea. Forecasting*, **24**, 1287–1296, <https://doi.org/10.1175/2009WAF2222227.1>

Nelson, B.R., D. Seo, and D. Kim, 2010: [Multisensor Precipitation Reanalysis](https://doi.org/10.1175/2010JHM1210.1). *J. Hydrometeor.*, **11**, 666–682, <https://doi.org/10.1175/2010JHM1210.1>

Nelson, B.R., O.P. Prat, D. Seo, and E. Habib, 2016: [Assessment and Implications of NCEP Stage IV Quantitative Precipitation Estimates for Product Intercomparisons](https://doi.org/10.1175/WAF-D-14-00112.1). *Wea. Forecasting*, **31**, 371–394, <https://doi.org/10.1175/WAF-D-14-00112.1>