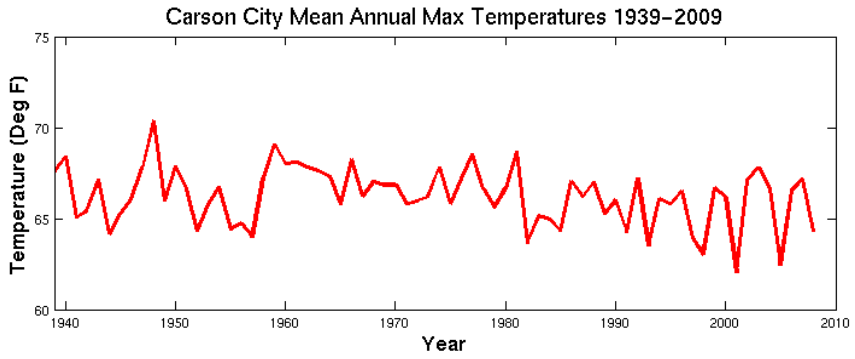
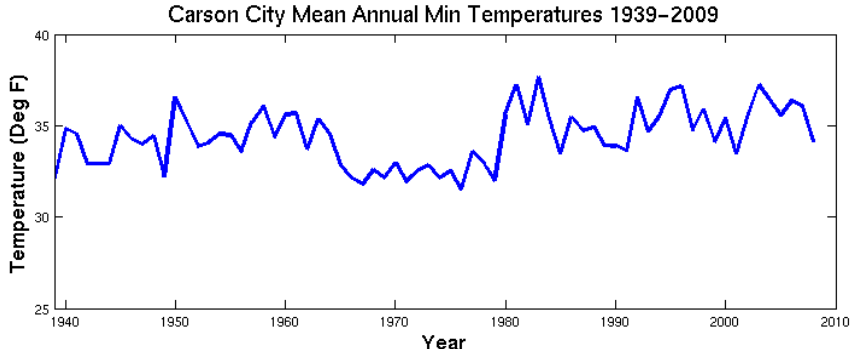
Historical Climate Data: The Never-Ending Battle For Acquisition

Ben Hatchett, John Mejia, and Darko Koracin

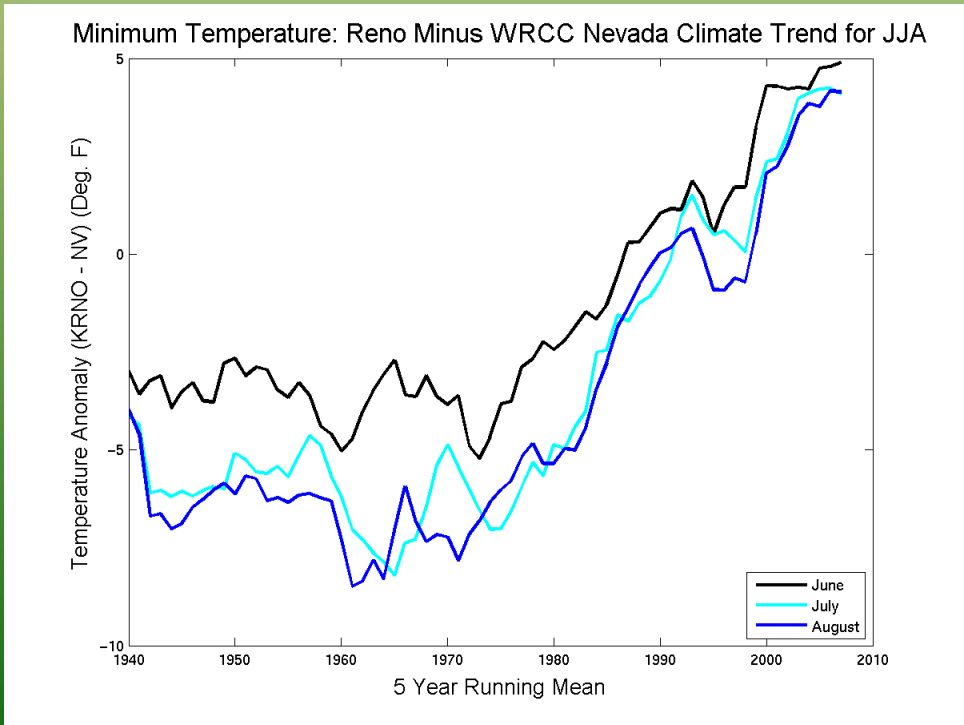
Division of Atmospheric Sciences, DRI, Reno NV





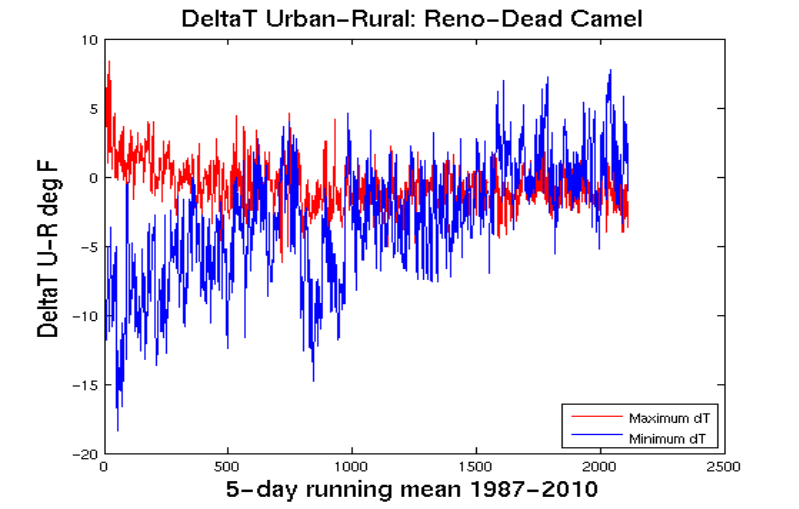
Outline

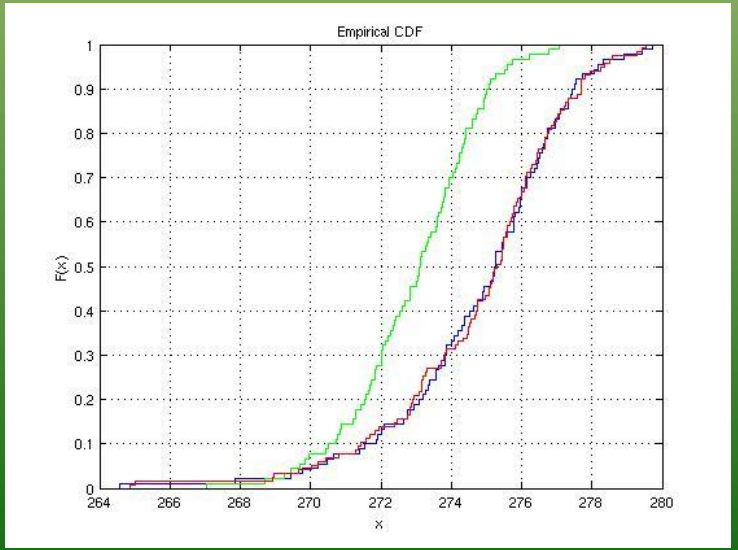
- Why do we need historical climate data?
- Where to acquire data?
- Data Issues
- Formatting For Everyone
- Conclusions
- Continuing Work



Why do we need historical data?

- Hard to understand present or future without knowledge of the past
- Calibration of Models (Downscaling Products)
  - E.g. CDF technique
- Verification of Model Hindcasts
  - Force with even years, verify against observed odd years
- Who gets to deal with acquiring the data?
  - Usually graduate students

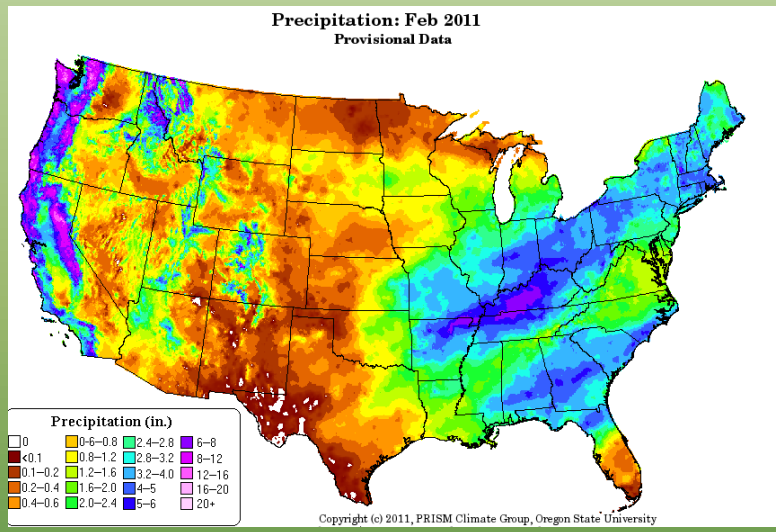


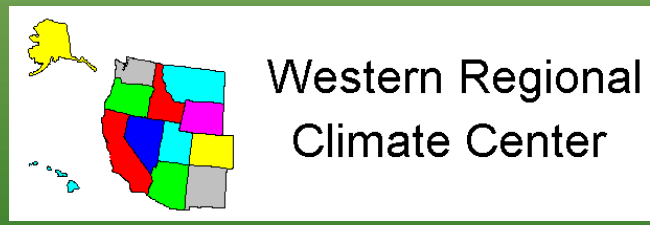


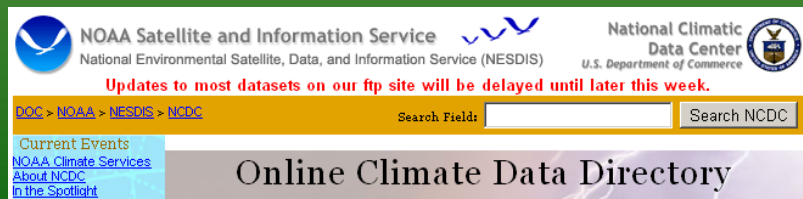
Where to Acquire Data?

Obvious Questions To Guide Your Search:

- What temporal scale needed?
  - 10-Minutes? Hourly? Daily? Monthly? Seasonal?
- What spatial scale(s)?
  - Station (point)-based
    - Single station or interpolation?
  - Gridded
    - Further downscaling necessary?
- **EXAMPLE DATA SOURCES**
  - Western Regional Climate Center (stations)
    - http://www.wrcc.dri.edu/index.html
  - National Climatic Data Center (both)
    - http://lwf.ncdc.noaa.gov/oa/climate/climatedata.htm
  - PRISM Climate Group (gridded)
    - http://www.prism.oregonstate.edu/
  - Earth System Research Laboratory (gridded)
    - http://www.esrl.noaa.gov/psd/data/gridded/







NOAA
Earth System Research Laboratory
Physical Sciences Division

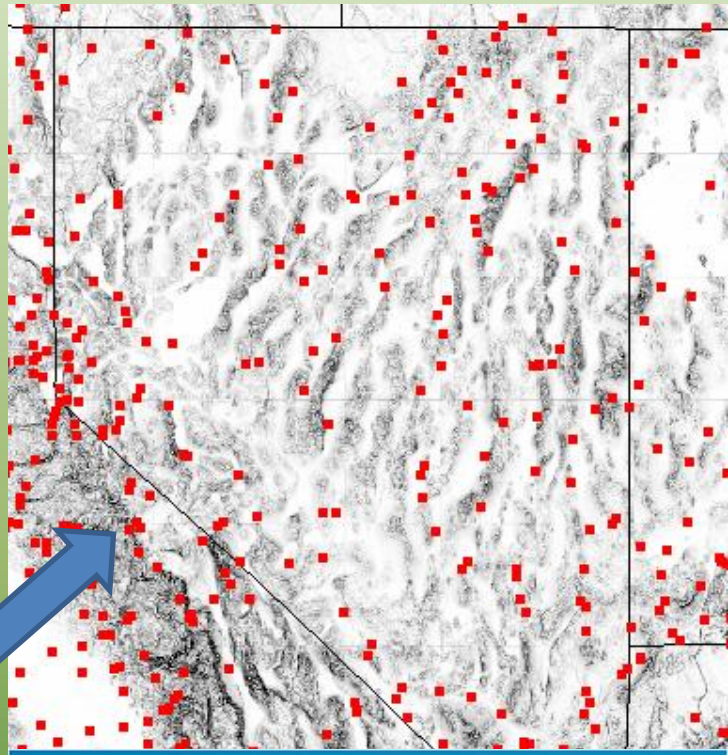
Data Issues

- *Simply downloading and using data is recipe for trouble.*
  - Understanding problems and learning to recognize them can be useful tool in all disciplines.
- Types of Problems:
  - Erroneous Data
    - Example: Carson City, Nevada

    Maximum T missing, but MaxT reported in MinT column! Solutions often simple yet creative.
  - Missing Data
    - What proportion is missing?
      - Weighting stations based on missing data proportion
    - How are missing data represented?
      - More importantly, is it easy to alter to fit your needs?
      - Scripting to eliminate/replace strange representations
    - Replacing Missing Data
- Note to Students: Learn to script. Bash, Perl, Python, whatever, it will make your life easier! It is never too late!

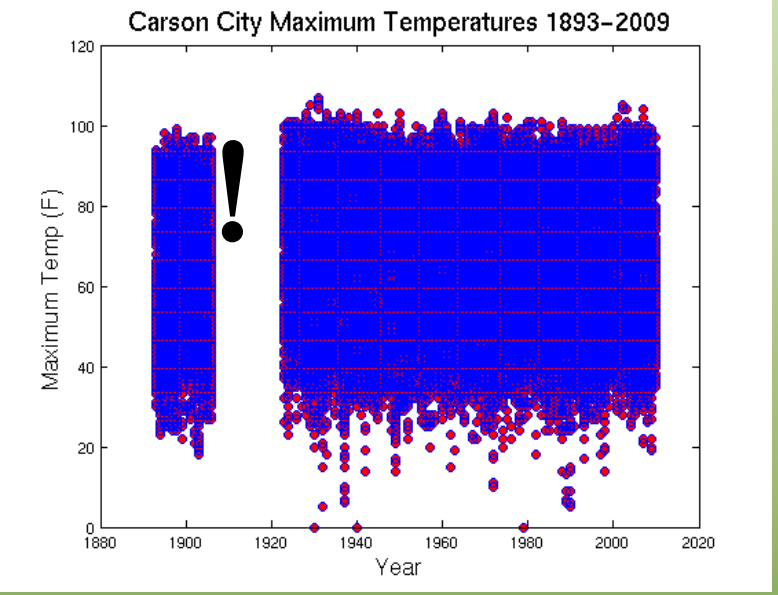
Some Bigger Problems…

- Station moves
  - Check the metadata!
- Local Land Use Change
  - Photographs, site visits, land use maps
- Station Locations
  - Automated sensors vs. COOP stations
    - Remote vs. Populated: Topographical Bias
    - Length of Record
      - Most mesonet/RAWs stations >1980
      - COOPs <1890
    - Temporal
      - COOP: daily, automated: hourly/10-minute
    - Quality of data
      - Tipping buckets freeze → Precipitation Bias
      - Anemometers icing up
      - RH <100%
      - Human inconsistencies → measurement time

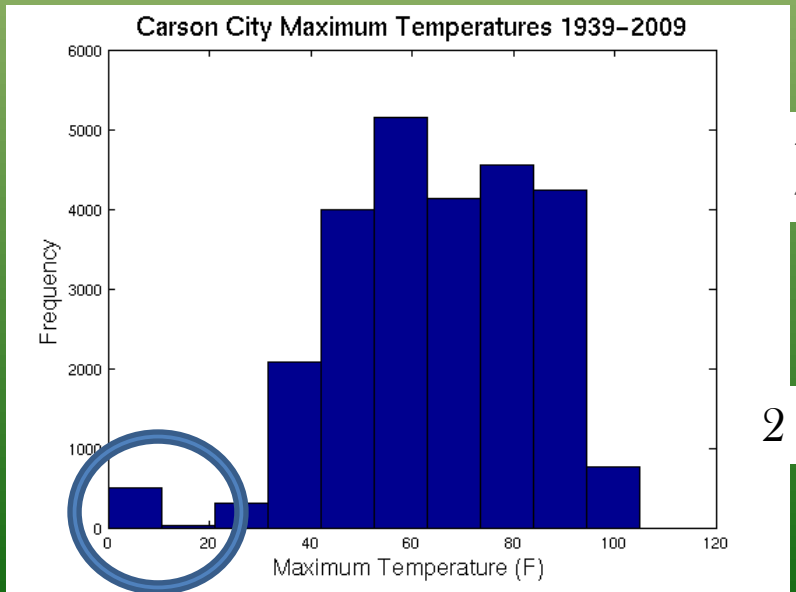


http://www.nap.edu/books/0309103878/xhtml/images/p200131deg339001.jpg

Tips For Quality Checking

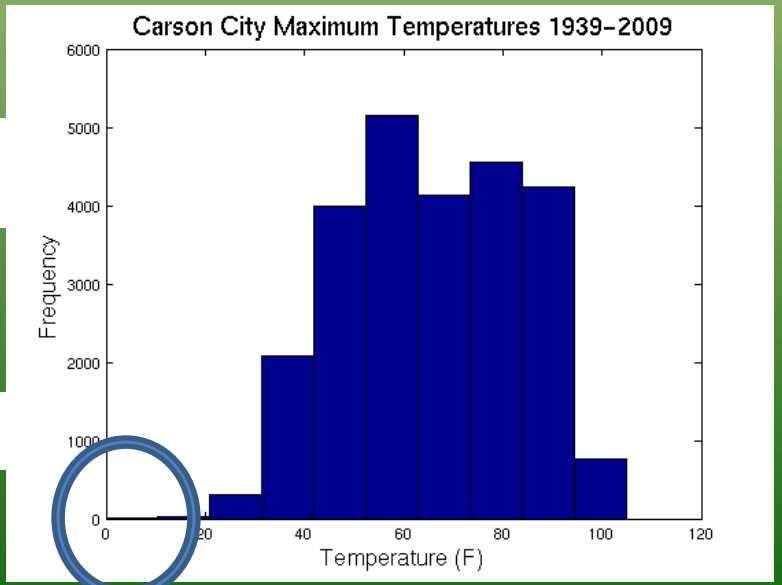


- **At each step, make a plot!**
- Missing data?
- Strange values/outliers →More investigation!
  - Find the outliers, compare with nearby stations.
  - 0F in Carson, 33F in Reno?
- *Simple, creative solutions will be much less troublesome than over-coding*
  - Helps to count errors as you go
- NCDC tries to flag errors; occasionally miss.
- **Up to the users to report strange data!**



<3σ DTR→NaN
40 values found

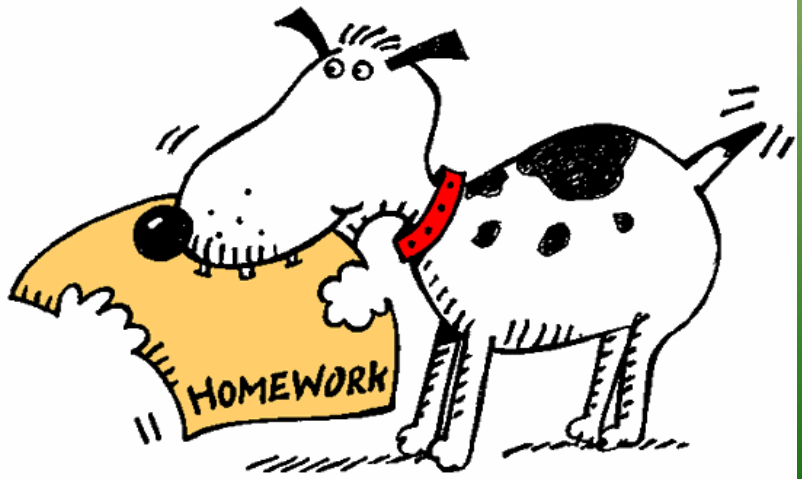
2 errors remain



Filling in the Blanks…

- Many methods to replace missing data
  - Threshold?
- Simplest method: Replace with mean values
  - Changes distribution
- Regression method (Smith et al. 1996)
- Expectation Maximization (Schneider 2000)
  - Yields better results than non-iterative regression method

http://www.gps.caltech.edu/~tapio/imputation/



http://www.math.hmc.edu/~kindred/math104/graphics/dog-hmwk.gif

Formatting For Everyone!

- This is somewhat of a fantasy of science; hard to please everyone
- Main uses: Model Input, GIS, Timeseries Analysis
- Ideas to Reduce Downstream Formatting Issues
  - Discuss with anyone you might work with in future what format they would prefer.
  - Format data initially in manner that can easily be manipulated/loaded into analysis programs
  - Create Metadata!
    - Mention data problems, methods, explanations, etc
    - The better the metadata, the fewer emails those using your data will pester you with

Good and Bad Formatting

- Bad Formats…
  - Use one or (worse) many of these: - / , | „
  - No metadata, converted units
  - MM/DD-YYYY,MaxT|MinT,Precip
- Good Formats…
  - Can easily be read into analysis/visualization tools
    - No fun to spend time manipulating
  - Have well-written metadata (units, locations, scales)
  - Use familiar units for user group (WRCC in degF)
  - Example (ours, so of course its best)
    - Output as ASCII, also write to NetCDF
    - YYYY MM DD MaxT MinT Prec \n
    - Why is this good?
      - It's Simple!!!

Conclusions

- Climate data is often neither simple nor well-organized
  - Significant time investments may be necessary to achieve state of data desired.
- Think simple and creative
  - Finding outliers or weird data points
  - Removing/replacing erroneous values
- Investing time into manipulating/formatting is not a waste…
  - The downstream cascade of well-formatted, quality data will save time, reduce headaches, and result in higher quality research
  - Everyone benefits from data sharing!
  - Don't hesitate to ask your users about formats/information they desire (preemptive strike!)

Present and Continuing Work

- Formatted 45 RAWS and 35 COOPs (finishing ALL Nevada COOPs soon) to simple daily format: YYYY MM DD Tmax Tmin Precip
  - Easy for most impact studies to utilize
- Stations are being statistically downscaled and used in hybrid statistical-dynamical downscaling method developed at DRI
  - Part of Intermountain Climate Change Assessment
    - Historical Analysis/Modeled Future for Western U.S.

References

- Smith, T. M., R. W. Reynolds, R. E. Livezey, and D. C. Stokes, 1996: Reconstruction of historical sea surface temperatures using empirical orthogonal functions. *J. Climate,* **9,** 1403–1420.
- T. Schneider, 2001: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*, **14**, 853-871.