Lost Moments: The Effect of Pre-processing on Environmental Data

Luke Bornn

Department of Statistics, Simon Fraser University

w/ Hannah Director (Harvard -> LANL -> UW)

May 13, 2015

Outline

A Look at Climate Data

- Historical climate data is fraught with changing measurement methods and inconsistent spatial and temporal coverage

A Look at Climate Data

- Historical climate data is fraught with changing measurement methods and inconsistent spatial and temporal coverage
- To compensate, measurements within a geographic area are often averaged to create an aggregated, gridded data set

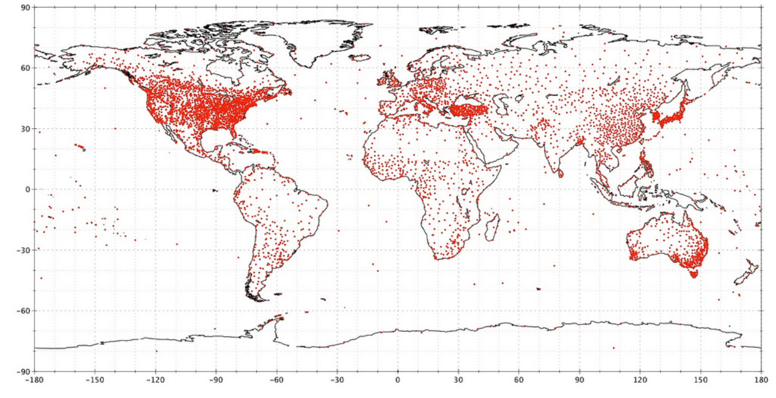
A Look at Climate Data

- Historical climate data is fraught with changing measurement methods and inconsistent spatial and temporal coverage
- To compensate, measurements within a geographic area are often averaged to create an aggregated, gridded data set
- While aggregation generally preserves the mean, the distribution of the raw measurements is drastically changed

A Look at Climate Data

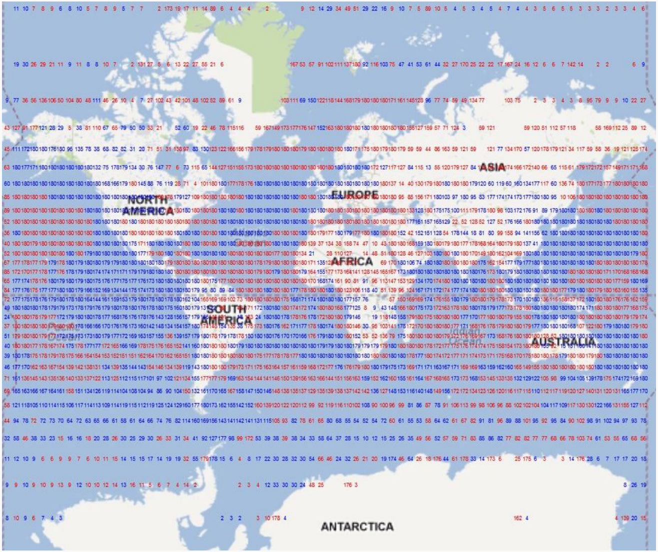
- Historical climate data is fraught with changing measurement methods and inconsistent spatial and temporal coverage
- To compensate, measurements within a geographic area are often averaged to create an aggregated, gridded data set
- While aggregation generally preserves the mean, the distribution of the raw measurements is drastically changed
- Failure to distinguish between raw/gridded data can significantly affect the scientific validity and real world impact of an analysis

Raw Climate Data



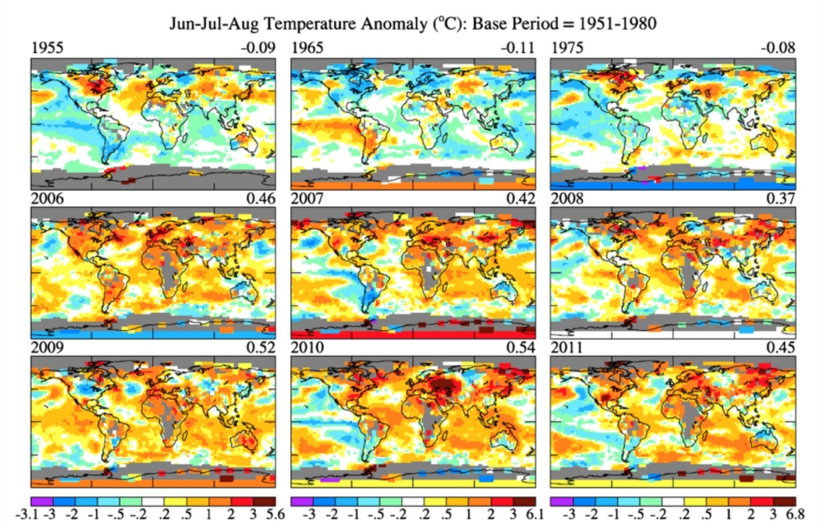
Source: http://employee.heartland.edu/rmuench/tempdata.htm

Gridded Climate Data



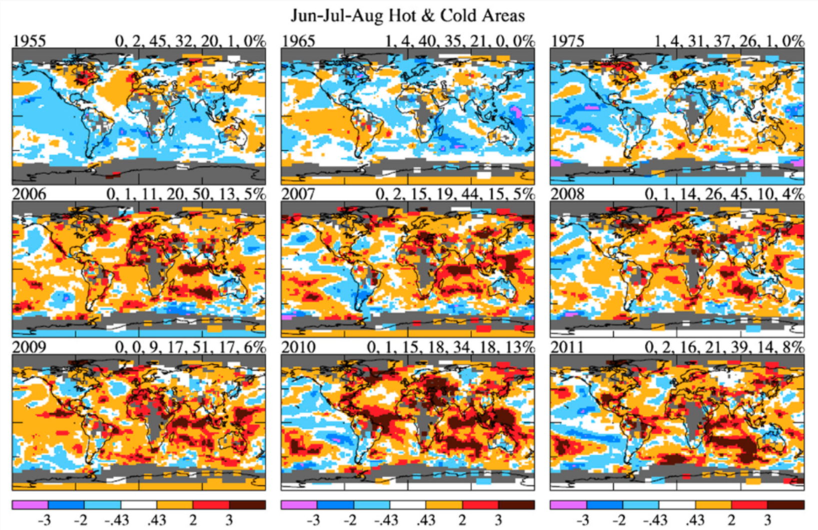
Source: https://sunshinehours.files.wordpress.com/2012/09/hadcrut3_gridded_180.jpg

An Example



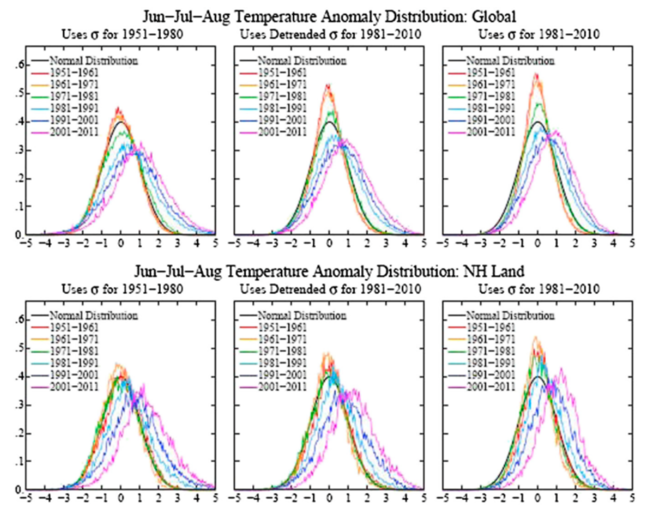
Hansen, Sato and Ruedy (PNAS 2012), Figure 1

An Example, continued



Hansen, Sato and Ruedy (PNAS 2012), Figure 3

An Example, continued



Hansen, Sato and Ruedy (PNAS 2012), Figure 4

An Example, continued

- As a follow-up comment, Rhines and Huybers (PNAS 2012) argue that it is critical to consider
  - normalizations
  - trends
  - reduction in surface station density

An Example, continued

- As a follow-up comment, Rhines and Huybers (PNAS 2012) argue that it is critical to consider
  - normalizations
  - trends
  - reduction in surface station density
- Between 1951-1980 and 1981-2010, there is a 35% decrease in number of stations reporting monthly averages

An Example, continued

- As a follow-up comment, Rhines and Huybers (PNAS 2012) argue that it is critical to consider
  - normalizations
  - trends
  - reduction in surface station density
- Between 1951-1980 and 1981-2010, there is a 35% decrease in number of stations reporting monthly averages
- Rhines and Huybers (PNAS 2012) assume a $1\,^{\circ}\mathrm{C}$ variance within grid box, homogeneity, normality, and independence between stations
- Their conclusion is that after these adjustments, there is no obvious increase in variance

Data Under Study

- Climate Research Unit (CRU) Monthly Temperature Anomaly Data (1950-2010):

Data Under Study

- Climate Research Unit (CRU) Monthly Temperature Anomaly Data (1950-2010):
  - Temperatures are expressed as monthly anomalies from a base period of 1961-1990

Data Under Study

- Climate Research Unit (CRU) Monthly Temperature Anomaly Data (1950-2010):
  - Temperatures are expressed as monthly anomalies from a base period of 1961-1990
  - Each station's time series is reported along with a mean for each $5^\circ \times 5^\circ$ grid box.

Data Under Study

- Climate Research Unit (CRU) Monthly Temperature Anomaly Data (1950-2010):
  - Temperatures are expressed as monthly anomalies from a base period of 1961-1990
  - Each station's time series is reported along with a mean for each $5^\circ \times 5^\circ$ grid box.
- Global Historical Climate Network (GHCN) Monthly Total Precipitation Data (1950-2010):

Data Under Study

- Climate Research Unit (CRU) Monthly Temperature Anomaly Data (1950-2010):
  - Temperatures are expressed as monthly anomalies from a base period of 1961-1990
  - Each station's time series is reported along with a mean for each $5^\circ \times 5^\circ$ grid box.
- Global Historical Climate Network (GHCN) Monthly Total Precipitation Data (1950-2010):
  - Total monthly precipitation measured at stations in North America and the former Soviet Union

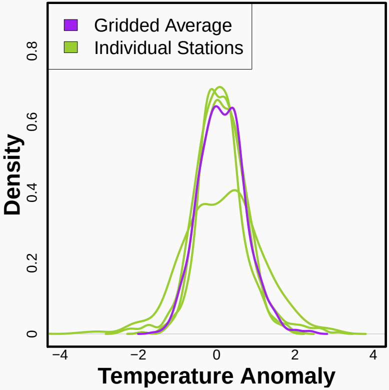
Data Under Study

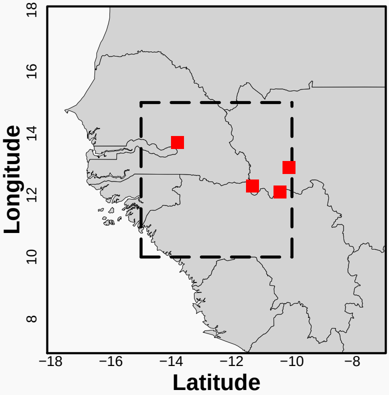
- Climate Research Unit (CRU) Monthly Temperature Anomaly Data (1950-2010):
  - Temperatures are expressed as monthly anomalies from a base period of 1961-1990
  - Each station's time series is reported along with a mean for each $5^\circ \times 5^\circ$ grid box.
- Global Historical Climate Network (GHCN) Monthly Total Precipitation Data (1950-2010):
  - Total monthly precipitation measured at stations in North America and the former Soviet Union
  - We averaged station data to form a $5^\circ \times 5^\circ$ spatially gridded product

Data Under Study

- Climate Research Unit (CRU) Monthly Temperature Anomaly Data (1950-2010):
  - Temperatures are expressed as monthly anomalies from a base period of 1961-1990
  - Each station's time series is reported along with a mean for each $5^\circ \times 5^\circ$ grid box.
- Global Historical Climate Network (GHCN) Monthly Total Precipitation Data (1950-2010):
  - Total monthly precipitation measured at stations in North America and the former Soviet Union
  - We averaged station data to form a $5^\circ \times 5^\circ$ spatially gridded product
- Stations missing greater than 10% of measurements were omitted to ensure a relatively constant sample size

Gridding's Effect on Moments



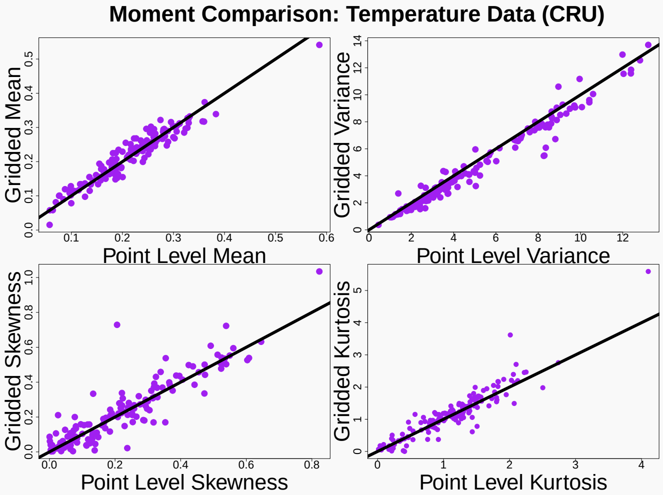


Gridding's Effect on Moments

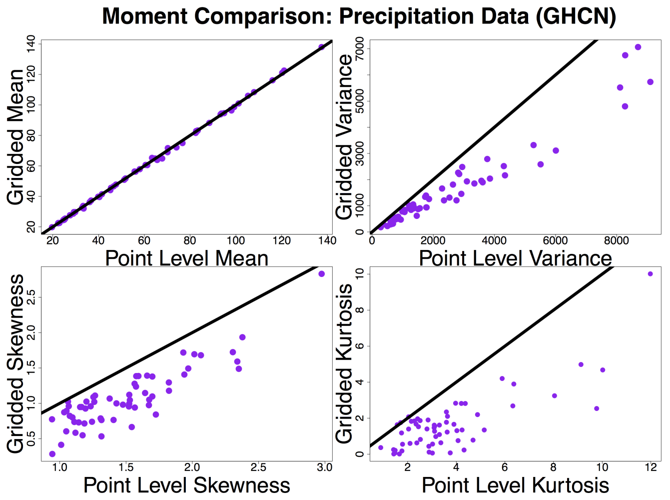
Table : *Mathematical definitions of the first four moments where X_i represents a single observation and $\overline{X}$ represents the mean of a group of observations and the relationships between these individual and averaged values.*

Moment	Def'n	Cumulant	Relationship
Mean (μ)	$\mathbb{E}(X)$	κ_1	$\mathbb{E}(\overline{X}) = \mathbb{E}(X_i)$
Variance (σ^2)	$\mathbb{E}[(X-\mu)^2]$	κ_2	$\mathbb{V}ar(\overline{X}) = \frac{1}{n}\mathbb{V}ar(X_i)$
Skewness (γ_1)	$\mathbb{E}[(\frac{X-\mu}{\sigma})^3]$	$\frac{\kappa_3}{\kappa_2^{3/2}}$	$\mathbb{S}kew(\overline{X}) = \frac{1}{\sqrt{n}}\mathbb{S}kew(X_i)$
Kurtosis (γ_2)	$\frac{\mathbb{E}[(X-\mu)^4]}{(\mathbb{E}[(X-\mu)^2])^2}$	$\frac{\kappa_4}{\kappa_2^2}$	$\mathbb{K}urt(\overline{X}) = \frac{1}{n}\mathbb{K}urt(X_i)$

Gridding's Effect on Moments



Gridding's Effect on Moments



Thinking about Correlation

- Stations within a grid box with n samples contain less information then n truly independent stations because of intra-site correlation

Thinking about Correlation

- Stations within a grid box with n samples contain less information then n truly independent stations because of intra-site correlation
- Effective Sample Size (ESS) corrects for this:

$$n_{eff} = \frac{n^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} Cor(x_i, x_j)} \tag{1}$$

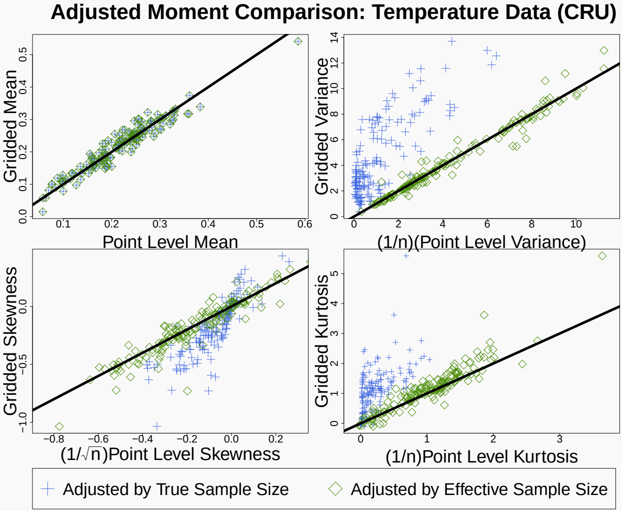
Thinking about Correlation

- Stations within a grid box with n samples contain less information then n truly independent stations because of intra-site correlation
- Effective Sample Size (ESS) corrects for this:

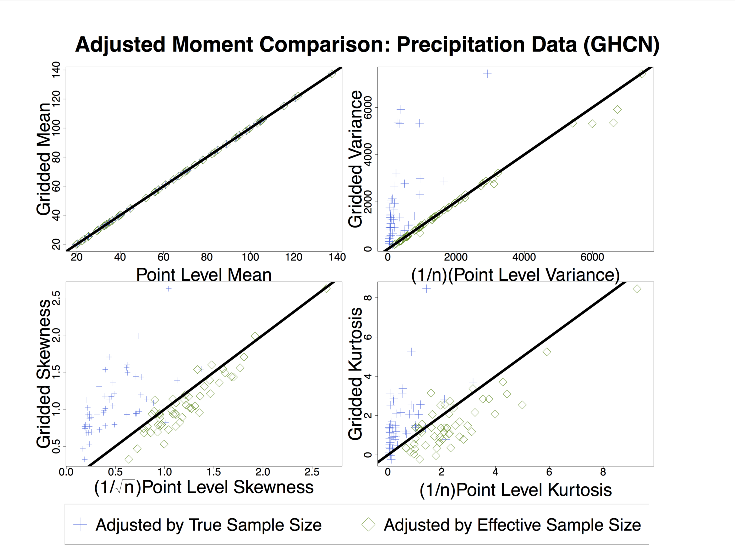
$$n_{eff} = \frac{n^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} Cor(x_i, x_j)} \tag{1}$$

- Correlation can be estimated from historical data and previous research on what affects intra-site correlation

Thinking about Correlation 2



Thinking about Correlation 2



Extremes: A Simple Example

- Extremes of the grid box average are not of practical interest, but estimates of extremes from individual station data are extremely noisy...

Extremes: A Simple Example

- Extremes of the grid box average are not of practical interest, but estimates of extremes from individual station data are extremely noisy...
- So, we adjust the empirical moments of the gridded data to point-level using factors of the effective sample size

Extremes: A Simple Example

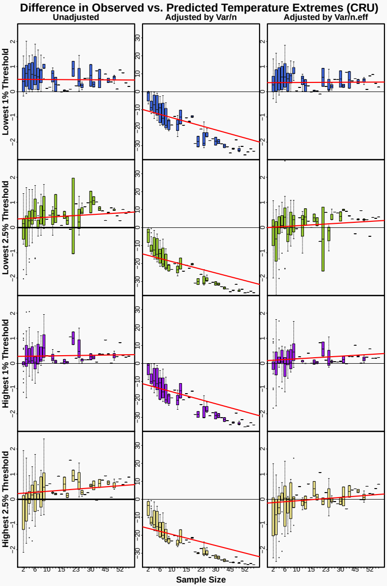
- Extremes of the grid box average are not of practical interest, but estimates of extremes from individual station data are extremely noisy...
- So, we adjust the empirical moments of the gridded data to point-level using factors of the effective sample size
- These adjusted moments can be used to estimate the point-level distributional parameters and the corresponding distributions can be used to estimate what percent of the data is above or below extreme thresholds underlying data

Extremes: A Conservative Adjustment

CRU Temperature Data (Observed - Predicted)

Variance Adjustment	**Thresholds:** Lowest 2.5%	Lowest 5%	Highest 2.5%	Highest 5%
Unadjusted	0.60	0.33	0.27	0.16
Adj. by var/n	-13.42	-17.09	-14.63	-17.62
Adj. by var/n.eff	0.47	-0.01	0.10	-0.21

Extremes: A Conservative Adjustment

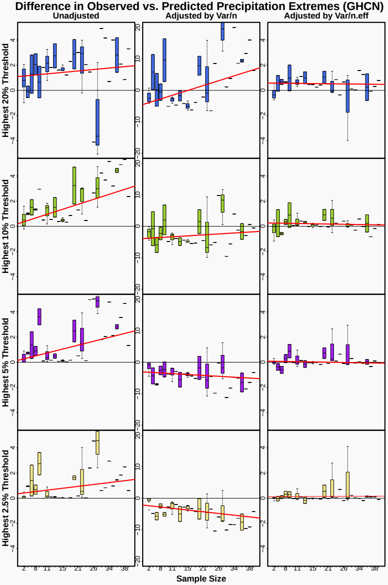


Extremes: A Conservative Adjustment

GCHN Precipitation Data (Observed - Predicted)

Variance Adjustment	**Thresholds:** Highest 20%	Highest 10%	Highest 5%	Highest 2.5%
Unadjusted	1.48	2.00	1.62	1.11
Adj. by var/n	3.07	-1.96	-3.83	-4.22
Adj. by var/n.eff	0.38	0.16	0.17	0.21

Extremes: A Conservative Adjustment



Conclusion

- Averaging fundamentally changes a measurement's distribution which matters for answering pertinent questions in climate science

Conclusion

- Averaging fundamentally changes a measurement's distribution which matters for answering pertinent questions in climate science
- Reporting information on original sample sizes and intra-site correlation would make gridded products more interpretable and useful

Conclusion

- Averaging fundamentally changes a measurement's distribution which matters for answering pertinent questions in climate science
- Reporting information on original sample sizes and intra-site correlation would make gridded products more interpretable and useful
- Similar issues likely exist for gridded climate model outputs and addressing them may be an area of future work