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

Getting Back to the Data

Understanding the Effects of Gridding

Adjusting for Gridding

Extremes

Conclusion
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

Luke Bornn (SFU)  
Lost Moments
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
As a follow-up comment, Rhines and Huybers (PNAS 2012) argue that it is critical to consider

- normalizations
- trends
- reduction in surface station density
As a follow-up comment, Rhines and Huybers (PNAS 2012) argue that it is critical to consider normalizations, trends, and 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 °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° × 5° 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° × 5° 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

- Gridded Average
- Individual Stations

![Graph showing temperature anomaly and density over latitude and longitude with a map of the region.]
## 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.

<table>
<thead>
<tr>
<th>Moment</th>
<th>Def’n</th>
<th>Cumulant</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean ($\mu$)</td>
<td>$\mathbb{E}(X)$</td>
<td>$\kappa_1$</td>
<td>$\mathbb{E}(\overline{X}) = \mathbb{E}(X_i)$</td>
</tr>
<tr>
<td>Variance ($\sigma^2$)</td>
<td>$\mathbb{E}[(X - \mu)^2]$</td>
<td>$\kappa_2$</td>
<td>$\text{Var}(\overline{X}) = \frac{1}{n} \text{Var}(X_i)$</td>
</tr>
<tr>
<td>Skewness ($\gamma_1$)</td>
<td>$\mathbb{E}[(\frac{X - \mu}{\sigma})^3]$</td>
<td>$\frac{\kappa_3}{\kappa_2^{3/2}}$</td>
<td>$\text{Skew}(\overline{X}) = \frac{1}{\sqrt{n}} \text{Skew}(X_i)$</td>
</tr>
<tr>
<td>Kurtosis ($\gamma_2$)</td>
<td>$\frac{\mathbb{E}[(X - \mu)^4]}{(\mathbb{E}[(X - \mu)^2])^2}$</td>
<td>$\frac{\kappa_4}{\kappa_2^2}$</td>
<td>$\text{Kurt}(\overline{X}) = \frac{1}{n} \text{Kurt}(X_i)$</td>
</tr>
</tbody>
</table>
Gridding’s Effect on Moments

Moment Comparison: Temperature Data (CRU)

Gridded Mean vs. Point Level Mean
Gridded Variance vs. Point Level Variance
Gridded Skewness vs. Point Level Skewness
Gridded Kurtosis vs. Point Level Kurtosis
Gridding’s Effect on Moments

Moment Comparison: Precipitation Data (GHCN)

- Gridded Mean vs. Point Level Mean
- Gridded Variance vs. Point Level Variance
- Gridded Skewness vs. Point Level Skewness
- Gridded Kurtosis vs. Point Level Kurtosis

Luke Bornn (SFU)
Thinking about Correlation

- Stations within a grid box with $n$ samples contain less information than $n$ truly independent stations because of intra-site correlation.
Thinking about Correlation

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

\[
\text{neff} = \frac{n^2}{\sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cor}(x_i, x_j)}
\]  \hspace{1cm} (1)

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

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

\[
\text{neff} = \frac{n^2}{\sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cor}(x_i, x_j)}
\]  

- Correlation can be estimated from historical data and previous research on what affects intra-site correlation.
Thinking about Correlation 2

Adjusted Moment Comparison: Temperature Data (CRU)

Adjusted by True Sample Size

Adjusted by Effective Sample Size

Lost Moments
Thinking about Correlation 2

Adjusted Moment Comparison: Precipitation Data (GHCN)

- Gridded Mean vs. Gridded Mean
- Gridded Variance vs. Gridded Variance
- Point Level Mean vs. (1/n)(Point Level Variance)
- Gridded Skewness vs. Gridded Skewness
- (1/n)Point Level Skewness vs. (1/n)Point Level Kurtosis

+ Adjusted by True Sample Size
\( \diamond \) Adjusted by 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...
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)

<table>
<thead>
<tr>
<th>Variance Adjustment</th>
<th>Thresholds:</th>
<th>Lowest 2.5%</th>
<th>Lowest 5%</th>
<th>Highest 2.5%</th>
<th>Highest 5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unadjusted</td>
<td></td>
<td>0.60</td>
<td>0.33</td>
<td>0.27</td>
<td>0.16</td>
</tr>
<tr>
<td>Adj. by var/n</td>
<td></td>
<td>-13.42</td>
<td>-17.09</td>
<td>-14.63</td>
<td>-17.62</td>
</tr>
<tr>
<td>Adj. by var/n.eff</td>
<td></td>
<td>0.47</td>
<td>-0.01</td>
<td>0.10</td>
<td>-0.21</td>
</tr>
</tbody>
</table>
Extremes: A Conservative Adjustment
### Extremes: A Conservative Adjustment

GCHN Precipitation Data (Observed - Predicted)

<table>
<thead>
<tr>
<th>Variance Adjustment</th>
<th>Thresholds:</th>
<th>Highest 20%</th>
<th>Highest 10%</th>
<th>Highest 5%</th>
<th>Highest 2.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td></td>
<td>1.48</td>
<td>2.00</td>
<td>1.62</td>
<td>1.11</td>
</tr>
<tr>
<td>Adj. by var/n</td>
<td></td>
<td>3.07</td>
<td>-1.96</td>
<td>-3.83</td>
<td>-4.22</td>
</tr>
<tr>
<td>Adj. by var/n.eff</td>
<td></td>
<td>0.38</td>
<td>0.16</td>
<td>0.17</td>
<td>0.21</td>
</tr>
</tbody>
</table>
Extremes: A Conservative Adjustment

Difference in Observed vs. Predicted Precipitation Extremes (GHCN)

Unadjusted

Adjusted by \( \frac{\text{Var}}{\text{n}} \)

Adjusted by \( \frac{\text{Var}}{\text{n}_{\text{eff}}} \)

Sample Size

\( 2 \ 8 \ 11 \ 15 \ 21 \ 26 \ 34 \ 38 \)
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.