The COST-HOME monthly benchmark dataset with temperature and precipitation data for testing homogenisation algorithms

Victor Venema, Enric Aguilar, José A. Guijarro and Olivier Mestre

Content

- COST Action
- Benchmark
  - Why
  - Generation of homogeneous data
  - Insertion of inhomogeneities
- Preliminary results

Benchmark dataset

- Survey: 38 surveys returned
  - Monthly temperature and precipitation
  - Additive vs. multiplicative process
- Test complete homogenisation algorithms

Many solutions (for monthly data)!

- DETECTION
  Visual, Craddock test, Student t-test, Likelihood ratio test (SNHT), Potter test, Bayesian procedures, Local contrast test, Pettitt test, penalized likelihood, MASH...
- CORRECTION
  Composite reference series, interpolated reference series, multiple non-homogeneous series (MASH), ANOVA (Mestre)...

ADVANCES IN HOMOGENISATION METHODS OF CLIMATE SERIES: AN INTEGRATED APPROACH

Olivier Mestre
Enric Aguilar
Ingaborj Auer
Anders Grimvall
Petr Stepanek
Tamas Szentimrey
Victor Venema

Météo-France
U. Rovira i Virgili
ZAMG
Linköping University
CHMI
University Bonn
France
Spain
Austria
Sweden
Czech Republic
Germany

Ingeborg Auer
Olivier Mestre
Anders Grimvall
Petr Stepanek
Tamas Szentimrey
Victor Venema
Benchmark dataset

1) Survey: 38 surveys returned
   - Monthly temperature and precipitation
   - Additive vs. multiplicative process
2) Test complete homogenisation algorithms
   - Statistical detection
   - Correction methods
   - Computation reference (if any)
   - Handle outliers and missing data
   - Complete system

Real data section

1) France (Bourgogne): 9 rain rate
2) France (Brittany): 17 Tmin, 17 Tmax
3) The Netherlands: 11 rain rate*, 9 Tmean
4) Norway: 189 rain rate*, 100 Tmean*
   - Two precipitation networks with 9 and 10 stations
   - Two temperature networks with 7 stations
5) Catalonian region: 40 rain rate, 30 Tmin, 30 Tmax
   - Some meta data available

Outline creation benchmark

1) Start with homogenised data
2) Multiple surrogate and synthetic realisations
3) Mask surrogate records
4) Add global trend
5) Insert inhomogeneities in station time series
6) Published on the web
7) Homogenize by COST participants and third parties
8) Analyse the results and publish

1) Start with homogenised data

- Austria: 43 rain rate, 35 Tmean
- France (Bourgogne): 9 Rain rate
- France (Brittany): 17 Tmin, 17 Tmax
- Catalonian region: 30 Tmin, 31 Tmax

2) Multiple surrogate realisations

- Seven networks; need about 100 networks
  - Solution: Surrogate data
- Multiple surrogate realisations
  - Temporal correlations
  - Station cross-correlations
  - Empirical distribution function
- Advantage compared to autoregressive modelling
  - Non-Gaussian
  - Correlations on long time scales
  - Rust, Mestre & Venema. Fewer Jump, less memory... JGR, Oct. 2008.
The iterative IAAFT algorithm

Time series

Schreiber and Schmitz

Flow diagram

Time series

Distribution

IAAFT algorithm smoothes jumps

One station – with annual cycle

One station – anomalies

Multiple stations – 10 year zoom

Multiple stations – 10 year zoom

Surrogate of Bounded Cascade

Time or space

LWP or LWC

One station – anomalies

Measurement - low cross correlation

Measurement - high cross correlation

Measurement - anomalies


0 20 40 60

0 20 40 60

1900 1901 1902 1903 1904 1905 1906 1907 1908 1909 1910

0 20 40 60

0 20 40 60

1900 1901 1902 1903 1904 1905 1906 1907 1908 1909 1910

0 20 40 60

0 20 40 60
Scatterplot 2 stations monthly rain

Synthetic data
- From the surrogate networks: synthetic networks are computed
  - No temporal correlations
  - Difference or ratio time series are Gaussian
  - Do have almost the same spatial correlations
- Good comparison with surrogate data
  - Test only influence of structure
- Same settings are used to generate inhomogeneities

3) Mask surrogate records
- Beginning of records jagged (rough)
- Linear increase in number of stations
- Three stations in 1900
- Last station after 25% of full time
- End of record all stations are measuring
- Influence of jagged edge on detection and correction

4) Add global trend

5) Insert inhomogeneities in stations
- Random breaks
- Frequency of breaks
  - 2 to 8 / 100a
- Temperature
  - Standard deviation breaks: 0.8 °C
  - Standard deviation seasonal cycle: 0.4°C
- Precipitation
  - Standard deviation breaks: 15 %
  - Standard deviation seasonal cycle: 7.5%
5) Insert inhomogeneities in stations

- Simultaneous breaks
- In 30% of network
- Frequency of breaks: 1 in 30% of stations

Outliers

- Frequency: 1 per station, i.e. per 100a
- Size: 99 and 99.9 percentiles

Local trends (only temperature)

- Linear increase or decrease in one station
- Duration: 30, 60a
- Maximum size: same as breaks
- Frequency: once in 10% of the stations
Inhomogeneous data will be published on the COST-HOME homepage.

Final version: published in May.

Everyone is welcome to download and homogenize the data.

Homogenised by the end of the year.

Returned homogenised data should be in COST-HOME file format.

Also homogenising one or a few networks is fine:
- Start with surrogate data
- Only fully automatic algorithm need to do synthetic

Multiple homogenisation algorithms welcome:
- Influence implementation and operator

Detailed analysis will be performed in the working groups:
- Detection
- Correction

Preliminary analysis of previous version of the benchmark

<table>
<thead>
<tr>
<th>Participant</th>
<th>Algorithm</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. José Guijarro</td>
<td>Climatol</td>
<td>6 Versions with different settings</td>
</tr>
<tr>
<td>2. Péter Domonkos</td>
<td>CM D, MASH D, NSHT D</td>
<td>3 different detection algorithms</td>
</tr>
<tr>
<td>3. Michele Brunetti</td>
<td>Brunetti</td>
<td>Detection Craddock based</td>
</tr>
<tr>
<td>4. Dubravka Rasol &amp; Oliver Mezic</td>
<td>PRODIGE</td>
<td></td>
</tr>
<tr>
<td>5. Matthew Morne &amp; Claudia Williams</td>
<td>Automated pairwise hom.</td>
<td>2 Versions</td>
</tr>
<tr>
<td>6. Christine Grober &amp; Priyanka Auyer</td>
<td>HDCCLIS</td>
<td>Craddock</td>
</tr>
<tr>
<td>7. Gregor Vartacnik</td>
<td>MASH</td>
<td></td>
</tr>
<tr>
<td>8. Petr Stepanek</td>
<td>AnClim</td>
<td></td>
</tr>
<tr>
<td>9. Lucie Vincent</td>
<td>Vincent</td>
<td></td>
</tr>
<tr>
<td>10. Enric Aguilar</td>
<td>NSHT</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Number of homogenised networks per algorithm

<table>
<thead>
<tr>
<th>Homogenisation algo.</th>
<th>All networks</th>
<th>Real rate</th>
<th>Surrogate rate</th>
<th>Synthetic rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Climatol</td>
<td>34</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Brunetti</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>MASH</td>
<td>25</td>
<td>0</td>
<td>26</td>
<td>0</td>
</tr>
<tr>
<td>Vincent</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>HDCCLIS</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Brunetti</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Climatol A</td>
<td>15</td>
<td>0</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>Climatol B</td>
<td>15</td>
<td>0</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>Climatol D</td>
<td>15</td>
<td>0</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>Climatol E</td>
<td>15</td>
<td>0</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>Climatol F</td>
<td>15</td>
<td>0</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>ClimatolG01</td>
<td>40</td>
<td>0</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>ClimatolF</td>
<td>40</td>
<td>0</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>ClimatolE</td>
<td>40</td>
<td>0</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>ClimatolD</td>
<td>40</td>
<td>0</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>AnClim</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>AnClim</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>APHa1</td>
<td>19</td>
<td>18</td>
<td>5</td>
<td>42</td>
</tr>
<tr>
<td>APHa2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>ClimatolG01</td>
<td>40</td>
<td>40</td>
<td>12</td>
<td>92</td>
</tr>
<tr>
<td>ClimatolF</td>
<td>40</td>
<td>40</td>
<td>12</td>
<td>92</td>
</tr>
<tr>
<td>ClimatolE</td>
<td>40</td>
<td>40</td>
<td>12</td>
<td>92</td>
</tr>
<tr>
<td>ClimatolD</td>
<td>40</td>
<td>40</td>
<td>12</td>
<td>92</td>
</tr>
<tr>
<td>ClimatolA</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PRODIGE</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Conclusions

- Algorithm participant
- No. homogenised networks
- Remarks

- Published on the web
- Homogenize by participants
- Analyse the results

- Contributions
- No. homogenised networks - algorithm
### No. homogenised networks – input data

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>624</td>
<td>371</td>
<td>253</td>
</tr>
<tr>
<td>Real</td>
<td>70</td>
<td>40</td>
<td>30</td>
</tr>
<tr>
<td>Surrogate</td>
<td>316</td>
<td>193</td>
<td>123</td>
</tr>
<tr>
<td>Synthetic</td>
<td>238</td>
<td>138</td>
<td>100</td>
</tr>
</tbody>
</table>

#### Mean no. outliers per station

<table>
<thead>
<tr>
<th>Homogenisation alg.</th>
<th>All networks</th>
<th>Real rate</th>
<th>Surrogate rate</th>
<th>Synthetic rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>BROUWELLI</td>
<td>0.0</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>MASH</td>
<td>3.4</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>YEWEN</td>
<td>9.6</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>HOCLIS</td>
<td>6.3</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>ACHBE</td>
<td>5.5</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>ClimaxA</td>
<td>3.6</td>
<td>0.0</td>
<td>4.8</td>
<td>4.2</td>
</tr>
<tr>
<td>ClimaxC</td>
<td>8.6</td>
<td>54.4</td>
<td>54.8</td>
<td>51.9</td>
</tr>
<tr>
<td>ClimaxD</td>
<td>54.4</td>
<td>24.7</td>
<td>58.0</td>
<td>56.7</td>
</tr>
<tr>
<td>ClimaxE</td>
<td>54.2</td>
<td>24.7</td>
<td>67.7</td>
<td>54.4</td>
</tr>
<tr>
<td>ClimaxF</td>
<td>45.8</td>
<td>23.8</td>
<td>47.9</td>
<td>42.9</td>
</tr>
<tr>
<td>ClimaxG01</td>
<td>4.4</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
</tbody>
</table>

#### Mean no. breaks per station

<table>
<thead>
<tr>
<th>Homogenisation alg.</th>
<th>All networks</th>
<th>Real rate</th>
<th>Surrogate rate</th>
<th>Synthetic rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>BROUWELLI</td>
<td>0.0</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>MASH</td>
<td>10.1</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>YEWEN</td>
<td>6.1</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>HOCLIS</td>
<td>6.3</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>ACHBE</td>
<td>3.5</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>ClimaxA</td>
<td>3.6</td>
<td>0.0</td>
<td>4.8</td>
<td>4.2</td>
</tr>
<tr>
<td>ClimaxC</td>
<td>8.6</td>
<td>54.4</td>
<td>54.8</td>
<td>51.9</td>
</tr>
<tr>
<td>ClimaxD</td>
<td>54.4</td>
<td>24.7</td>
<td>58.0</td>
<td>56.7</td>
</tr>
<tr>
<td>ClimaxE</td>
<td>54.2</td>
<td>24.7</td>
<td>67.7</td>
<td>54.4</td>
</tr>
<tr>
<td>ClimaxF</td>
<td>45.8</td>
<td>23.8</td>
<td>47.9</td>
<td>42.9</td>
</tr>
<tr>
<td>ClimaxG01</td>
<td>4.4</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
</tbody>
</table>

### Histogram no. outliers per station – Surrogate 1

#### Histogram no. breaks per station – Surrogate 1
Conclusions

- Interesting benchmark dataset
  - Realistic structures
  - Realistic inhomogeneities

- Preliminary results show large spread
- Automatic algorithms are still less accurate
- Different implementations of algorithm can produce very different results

- Everyone is invited to join the effort

More information

- COST
  - http://homogenisation.org
- Benchmark
  - http://www.meteo.uni-bonn.de/venema/themes/homogenisation
- Surrogate data
  - http://www.meteo.uni-bonn.de/venema/themes/surrogates

- Everyone is invited to join the effort

2) Pre-processing of homogenised data

- Annual cycle removed before, added at the end
- Number of stations between 5 and 20
- Cross correlation varies as much as possible
- Data is few decades long
  - good statistics
- Generated networks are 100 a long
  - Detrended the input stations
  - Mirrored them to more 100 a
  - Cropped to 100 a
  - Larger scale correlations are small

Detected break distribution, tn, tx