The ISTI Benchmarks – Their Construction and Characteristics

SAMSI/IMAGe workshop, July 2014

Dr Kate Willett
(Met Office Hadley Centre, UK)

www.surfacetemperatures.org
Talk Outline

- Benchmarking basics
- Creating a 'clean' synthetic world
- Creating a set of error filled worlds
- Assessing homogenisation algorithm skill against the benchmarks
- Other progress
- Help!...
Talk Outline

- Benchmarking basics
- Creating a 'clean' synthetic world
- Creating a set of error filled worlds
- Assessing homogenisation algorithm skill against the benchmarks
- Help!...
Benchmarking Basics
Confidence in Adjustments Made?

Type I Error
Do not detect or adjust when there has been a changepoint

Type II Error
Detect and adjust when no actual changepoint occurred

Missed adjustments vs false alarms: which is worse?

What about adjustments in the wrong direction?

Adjustments that are the wrong size/length or do not correctly adjust across the seasonal cycle?
Benchmarking Cycle

Create c.10 analog-error-worlds

- Simulate 'clean' spatio-temporal characteristics of actual stations underpinned by low frequency variability from a climate model to maintain plausible spatial correlation

- Add abrupt and gradual changepoints to approximate our best guess real world error structures

- Run homogenisation algorithms on the test data and assess ability to recover original 'clean' data

- Useful for further improvement of algorithms

Example use of benchmark data for USHCN
Benchmarking cycle
'TRUTH' UNKNOWN

\[ \text{XTRUTH}_{t,l} = S_{t,l} + L_{t,l} + V_{t,l} + M_{t,l} \]

XTRUTH = a climate element at time \( t \) and location \( l \)
- \( S \) = seasonal cycle
- \( L \) = long-term trends
- \( V \) = variability (ENSO, NAO, Volcanoes, Solar Cycles...)
- \( M \) = microclimate (topography, proximity to coast, prevailing wind, local environment...)

\[ \text{XOB}_{t,l} = \text{XTRUTH}_{t,l} + \varepsilon_{t,l} + \lambda_{t,l} \]

XOB = observation at time \( t \), location \( l \) and height \( h \)
- \( \varepsilon \) = random error at time/place/height
  (recording error, instrument error etc.)
- \( \lambda \) = systematic error at time/place/height *possibly correlated*
  (station move, exposure change, instrument change, observing practice change, urbanisation etc.)
Creating a 'Clean' Synthetic World
Team Creation: How To...

- Real Station climatology = seasonal cycle (S)
- Real Station standard deviation = variability (V) and microclimate (M)
- GCM interpolated to station location = long-term trend (L) and variability (V)
- Vector Autoregressive (VAR) model = variability (V) and microclimate (M)

IMPORTANT:
- Retain serial correlation at at least lag 1
- Retain cross-correlations with neighbours
- Do not introduce inhomogeneities
The Plan

- Use VAR to simulate standardised anomalies
  - For each time point:
    - For each station:
      - station at t=0 from 40 nearest neighbours + station at t-1
      - station simulated shock at t=0 (spatially correlated)

HELP: Ensuring real spatial correlations requires either huge matrices, some type of gibbs sampling or iterative VAR and hoping for the best

- Multiply by real station standard deviations
- Add interpolated smoothed GCM trend
- Add back real station climatology

VOILA!
The Plan

Station 1 $t-1$

Station 1 $t=0$

Station 2 $t-1$

Station 2 $t=0$
Why use a GCM?

- Global, spatially correlated trend and variability fields
- Ability to play with different background trends (no warming vs burn everything)

Interpolate monthly means
Calculate climate anomalies
Fit loess (0.4 vs 0.15?)

USE LOESS LINE
interpp(gcmlons,gcmlats,gcmmonth,stationlons,stationlats)

HELP: I need a better way to interpolate accounting for the dateline
\[ A_t = \Phi_1 A_{t-1} + Z_t \]

- \( A \) = matrix of 1 to J stations for all times 1 to T
- \( \Phi_1 \) = autoregressive parameter matrix
- \( Z_t \) = matrix of error terms/shocks (residuals)

\[
\text{vec}(\Phi_1) = (\mathbf{1} \otimes \Gamma(0))^{-1} \text{vec}(\Gamma(1))
\]
- \( \Gamma(0) \) = covariance matrix of all stations
- \( \Gamma(1) \) = covariance matrix of all stations at lag 1

For simulation, use exponential decay function to proxy correlation by distance (get fit from real data)

\[ f <- \text{function}(x,a,b)\{ a*\exp(-b*x) \} \]

Cross-Correlation: \( a=0.9162, b=0.0005 \)

Cross-Correlation at lag 1: \( a=0.4162, b=0.0005 \)

HELP: Fit for lag 1 correlation not strong. Is this sensible?
Obtaining realistic spatially correlated shocks

\[ Z_t = A_t - \Phi_1 A_{t-1} \]

Shock terms need to follow the observed distribution and be spatially correlated

- Create global networks and neighbourhoods across the globe
- Create covariance for each network+neighbourhood using distance as a proxy
- Run a factorisation algorithm to create global fields of spatially correlated shocks that are MVN
- Transform (QQ) to observed distribution for each station
Obtaining realistic spatially correlated shocks

\[ Z_t = A_t - \Phi_1 A_{t-1} \]

Shock terms need to follow the observed distribution and be spatially correlated

- Create global networks and neighbourhoods across the globe
- Create covariance for each network + neighbourhood using distance as a proxy
- Run a factorisation algorithm to create global fields of spatially correlated shocks that are MVN
- Transform (QQ) to observed distribution for each station
Obtaining realistic spatially correlated shocks

193 Networks, Neighbours within 5 deg of boundaries (or less if network is small)
Obtaining realistic spatially correlated shocks

t0 shock terms with spatial correlation
(dark to light ranges from -3 to 3 degrees C)
HELP!

GCM interpolation
VAR parameters:
Need to be derived as function of distance - better to add elevation and other factors that affect correlation
Using a function may lead to drastic over correlation
Better to use multivariate regression over matrix algebra?

Building the world:
Instability due to neighbour disconnect = rubbish
Creating a set of Error-filled Worlds
Effects of Changes that are not of Climate Origin

**STATION MOVE:** EXPOSURE AND MICROCLIMATE = abrupt change in mean and diurnal extremes - may affect seasonal cycle extremes

**SHELTER CHANGE:** EXPOSURE = abrupt change in diurnal extremes - may affect seasonal cycle extremes

**OBSERVING PRACTICE CHANGE:** SAMPLING = abrupt change possible in mean and extremes

**INSTRUMENT CHANGE:** CALIBRATION = abrupt change in mean and possibly extremes

**LANDUSE CHANGE:** EXPOSURE AND MICROCLIMATE = gradual change in mean and diurnal extremes - may affect seasonal cycle extremes
Team Corruption

\[ \text{XERRORWORLD}_{t,l} = \text{XTRUTH}_{t,l} + \lambda \text{ERRORWORLD}_{t,l} \]

SURFACE TEMPERATURE DATABANK

- World 1: no breaks
- World 2: few large breaks
- World 3: many small breaks
- World 4: abrupt and gradual breaks
- World 5: many small complex breaks
- etc.

Example error models applied to stations
Adjustment

Adjust mean

Adjust variance

By month? By season? By other variables e.g., humidity, wind, solar radiation?

Spatially clustered breaks

Seasonal cycles/features

Biases
Assessing Homogenisation Algorithm Skill Against the Benchmarks
Levels of Assessment

Level 1: Return of original climate features e.g., trends, climatology, variability (UNC)

Level 2: Ability to detect changepoints and characterise them (ALG)

Level 3: Detailed assessment of strengths and weaknesses against specific types of inhomogeneities (ALG)

Level 4: Assess performance on benchmarks alongside results from real world homogenisation (how realistic are the benchmarks?) (BEN)

RETURN:
Homogenised stations
List of changepoints and mean shift applied relative to reference period (present day)
Team Validation

**Contingency Table**

<table>
<thead>
<tr>
<th></th>
<th>Inhomogeneity present</th>
<th>No inhomogeneity present</th>
<th>TOTALS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Trend recovery</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>percentage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ROC Curve</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Inhomogeneities detected within +/- 3 months**
- Adjusted value must be correct sign (+/-) and within +/-1 °C
- HITS: 5 (4)
- FALSE ALARMS: 3 (3)
- TOTALS: 8 (7)

**Inhomogeneities not detected within +/- 3 months**
- Adjusted value incorrect sign and not within +/- 1 °C
- MISSES: 2 (3)
- CORRECT MISSES: 42 (42) (potential detections)
- TOTALS: 44 (45)

**Heidke Skill Score**: 61%
**Probability of Detection Hit Rate**: 71%
**False Alarm Rate**: 7%
Other Benchmarks
Daily Benchmarks for the USA using a GAM

Stations with temperature records from 1970 to 2011 in the contiguous USA:
Blue = Focus stations with no more than a quarter of the record missing
Monthly T and Precip benchmarks for small networks

COST HOME
see Victor Venema
Monthly USHCN Tmax, Tmin and Tmean benchmarks
Help!

Stage Three (Recommended Merge)
Number of Stations: 31999

~30000 stations, 10 blind worlds, 4 open worlds, monthly mean T
kate.willett@metoffice.gov.uk
@Kate_M_Willett
Methods

$$A_t = \Phi_1 A_{t-1} + Z_t$$

$A$ = matrix of 1 to J stations for all times 1 to T
$\Phi_1$ = autoregressive parameter matrix
$Z_t$ = matrix of residuals (random shocks)

$$\Phi_1 = \Gamma_{(1)} \Gamma_{(0)}^{-1}$$
$\Gamma_{(0)}$ = covariance matrix of all stations
$\Gamma_{(1)}$ = covariance matrix of all stations at lag 1

$$\text{Corr}(x,y) = a \ast \exp(-b*\text{distance})$$
Cross-Correlation: $a=0.9162$, $b=0.0005$
Cross-Correlation at lag 1: $a=0.4162$, $b=0.0005$

$$Z_t = A_t - \Phi_1 A_{t-1}$$

$$\Sigma_Z = \Gamma_{(0)} - \Phi_1 \Gamma_{(1)}^T$$
$\Sigma_Z$ = Covariance matrix of the residuals (random shocks)
193 Networks, Neighbours within 5 deg of boundaries (or less if network is small)
The Random Shock Part: Spatially Correlated

$t_0$ shock terms with spatial correlation (dark to light ranges from -3 to 3 degrees C)
The VAR Part: Spatio-temporally Correlated

Station 1 t-1

Station 2 t-1

Station 1 t=0

Station 2 t=0
Adding Long-term Trend, Climatology and Station St. Dev.

Real station standard deviation applied.

Long-term trend from HadCM3 A1B 1970-2083 (Interpolation improved by Finn)

Real station climatology added back.
Results: Station AR(1) correlation

Station Autocorrelation at lag 1:

CLEAN too high compared to REAL world

Real world contaminated by random error, inhomogeneity, missing data.

AREA FOR IMPROVEMENT: DISTANCE FUNCTION
Results: Station to Neighbours cross-correlation

Station Cross correlation with 40 nearest neighbours:

CLEAN too high compared to REAL world

Real world contaminated by random error, inhomogeneity, missing data.

AREA FOR IMPROVEMENT: DISTANCE FUNCTION
Results

Mean SD and AR(1) Corr of Difference series: 0.69, 0.42

Mean SD and AR(1) Corr of Dirty Difference series: 0.71, 0.47
Remaining Issues

St Dev of Difference series too low: Compare with USCRN which is clean but very short

AR(1) of station is too high: adjust distance function

Cross-correlations possibly too high: adjust distance function (maybe vary with latitude?)

Have not used all 30000+ stations: figure out how

Errors

Validation
Future Plans

Two papers on building a clean world: Single network (Kate and Robert Lund), Whole world (Kate, Robert and Richard Chandler)

Continue to work with the ISTI BAWG on the benchmarks (Kate, Victor, Claude, Matt, Robert L., Colin G., Renate, Peter plus others)

Stay in contact with other groups to help them use the benchmarks if useful

Meet up with Victor to work on error worlds

At least three talks planned for next few months on ISTI and Benchmarks (Edinburgh, Imperial, Met Office)
Future Plans

Complete within next month or so

Single network VAR method paper with Robert Lund

Whole World VAR method paper with Robert Lund and Richard Chandler

Possible meet up with Victor to work on error worlds

Carry on working with BAWG