Met Office

Is it good enough?

Benchmarking homogenisation algorithms and cross-cutting with efforts for land observations

Kate Willett and the Benchmarking and Assessment Working Group

Outline

1) What and Why?

2) The Benchmarking and Assessment Working Group

3) Creating Artificial Data with a Known 'Truth'

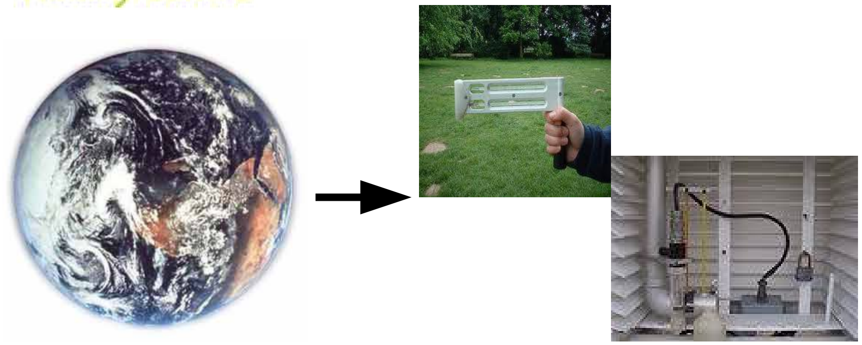
4) Creating 'Error Models' Covering all Known Real-world Nasties

5) Assessing the Benchmarks

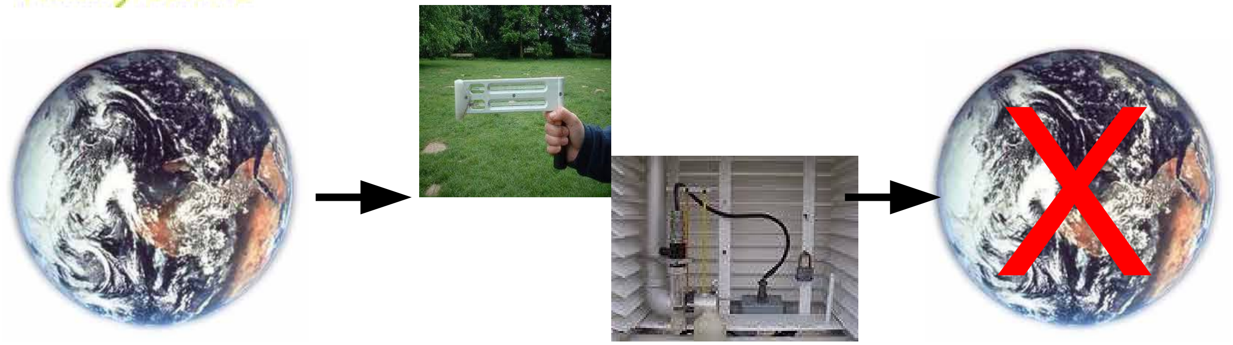
What and Why?

What is Benchmarking?

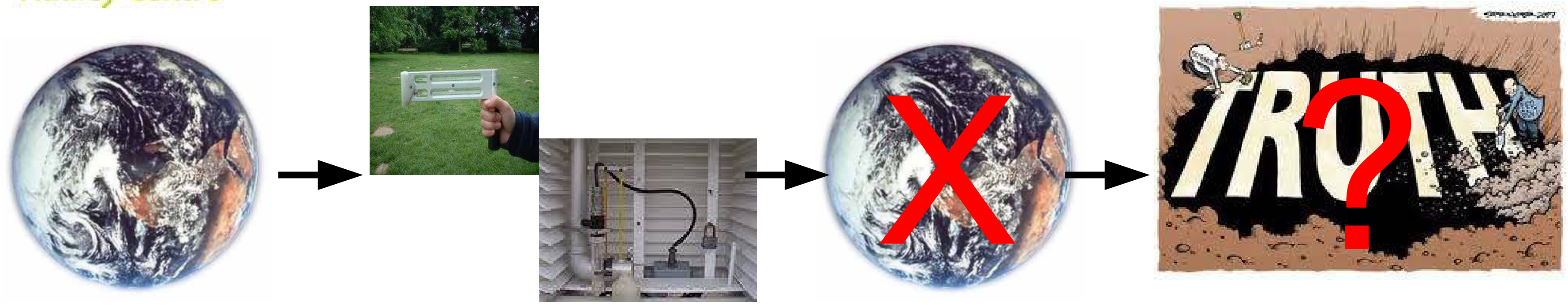
What is Benchmarking?



What is Benchmarking?

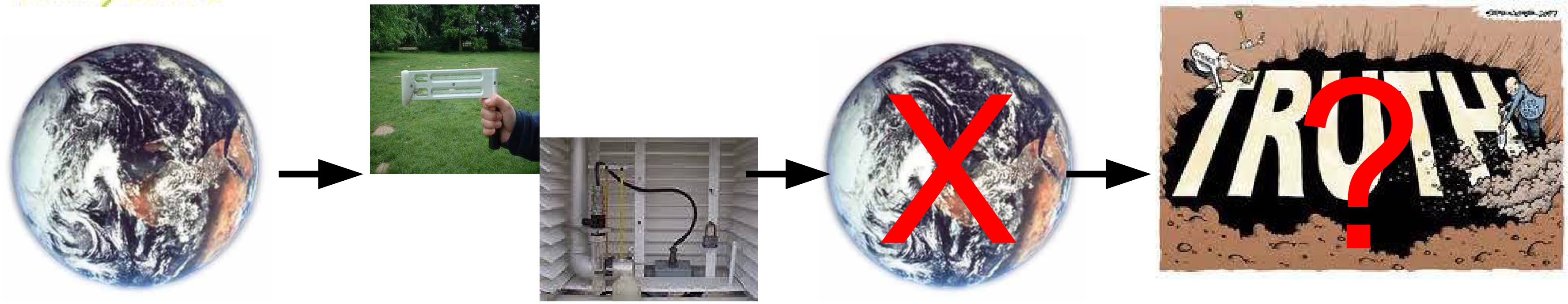


What is Benchmarking?



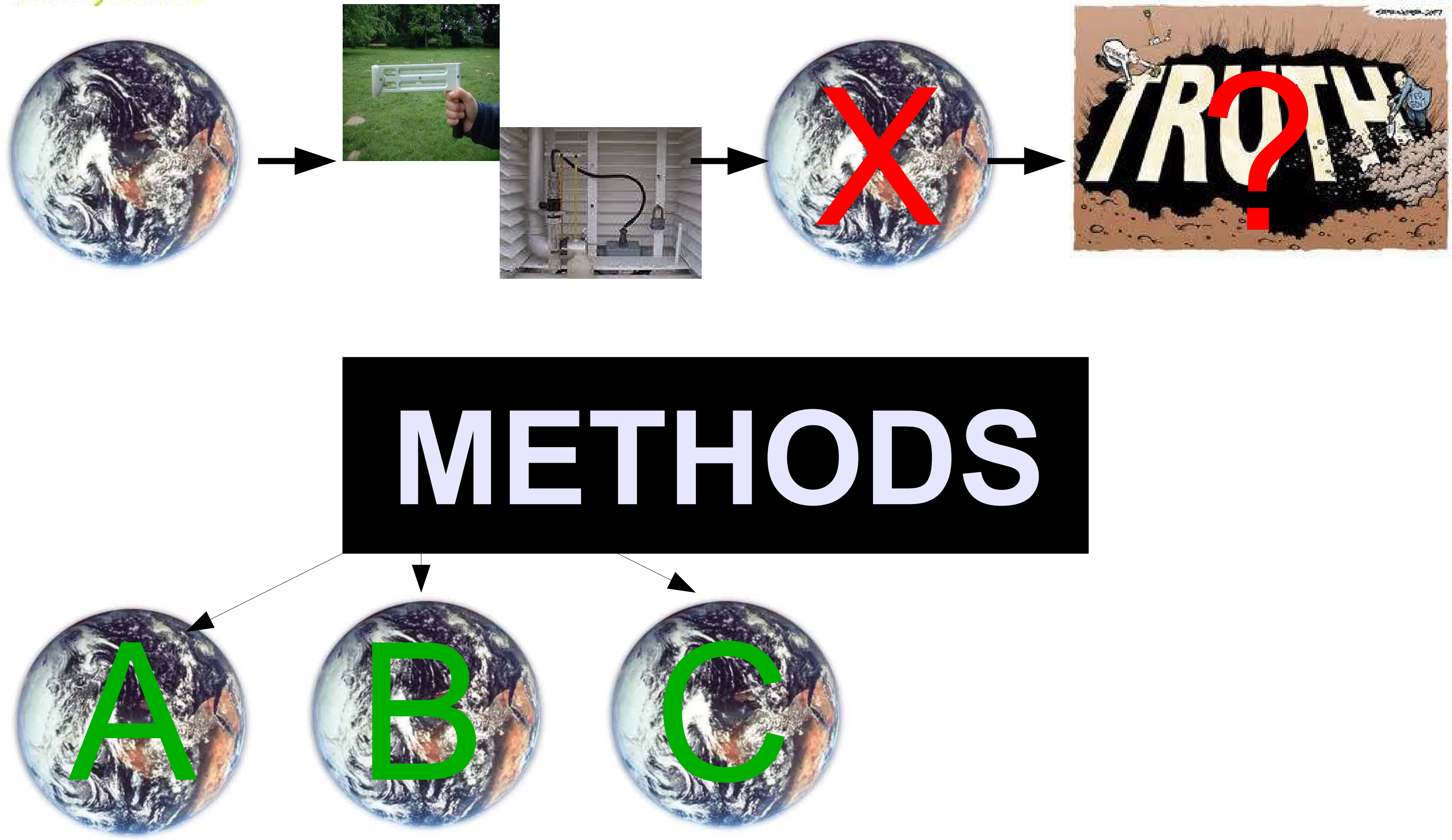
No one-size-fits-all approach

What is Benchmarking?



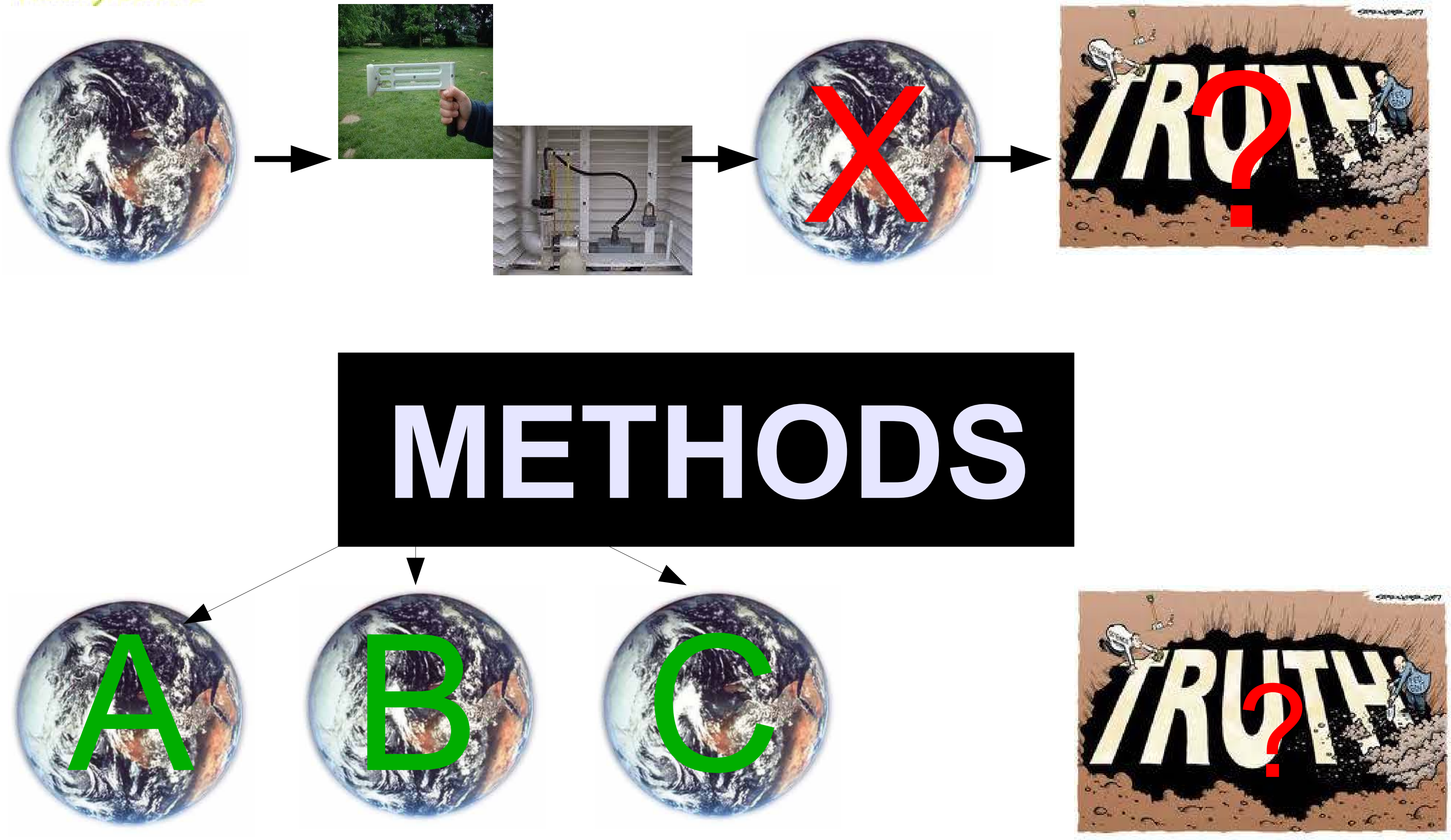
METHODS

What is Benchmarking?



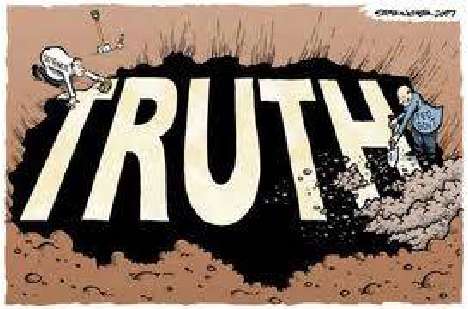
METHODS

What is Benchmarking?



METHODS

What is Benchmarking?

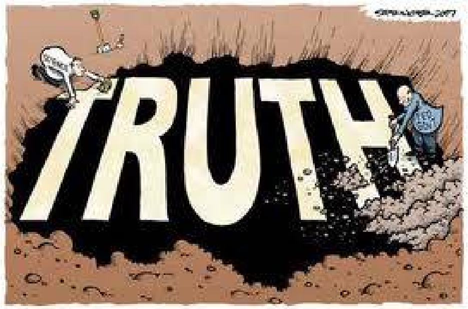


What is Benchmarking?

X_1

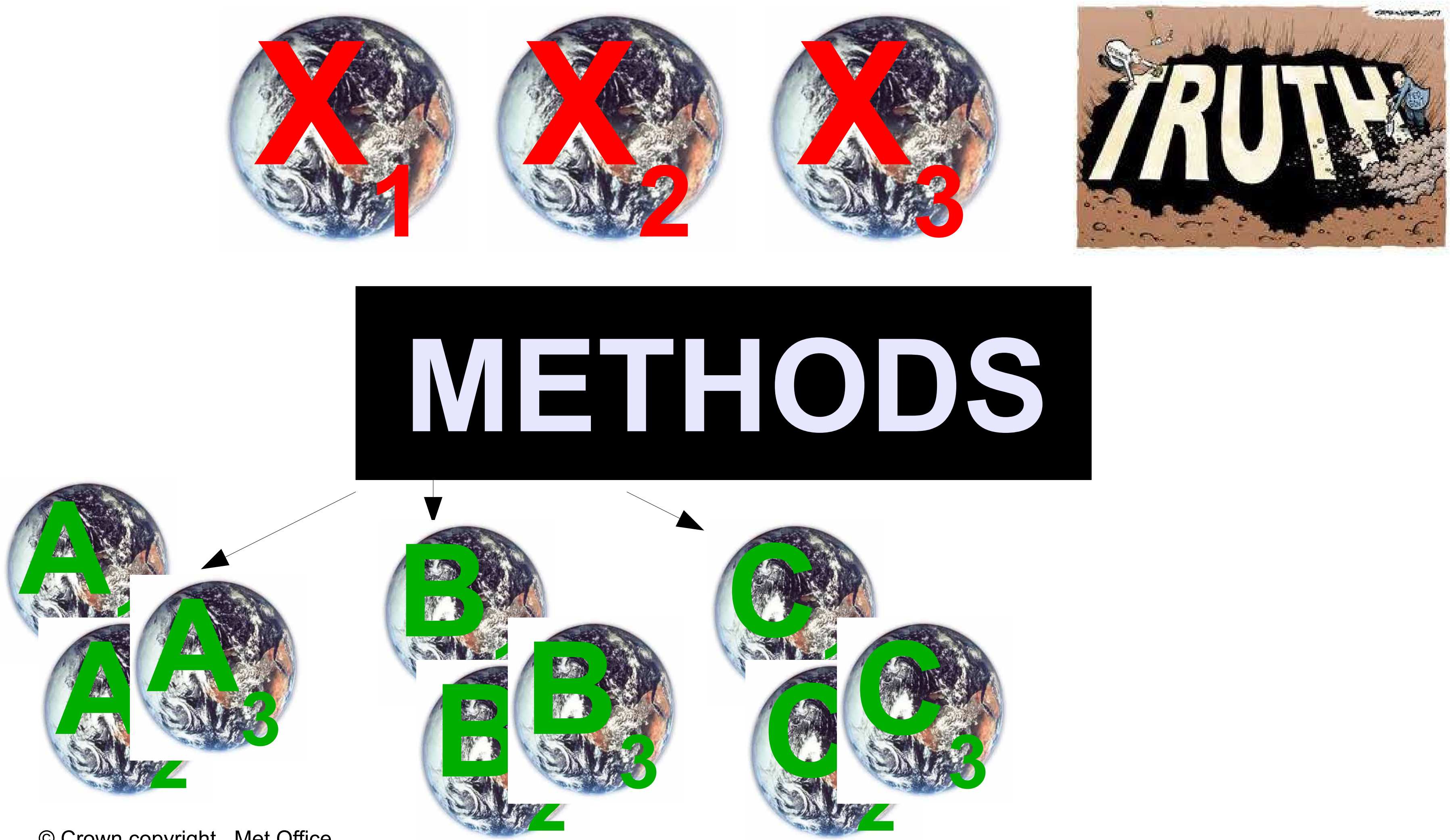
X_2

X_3

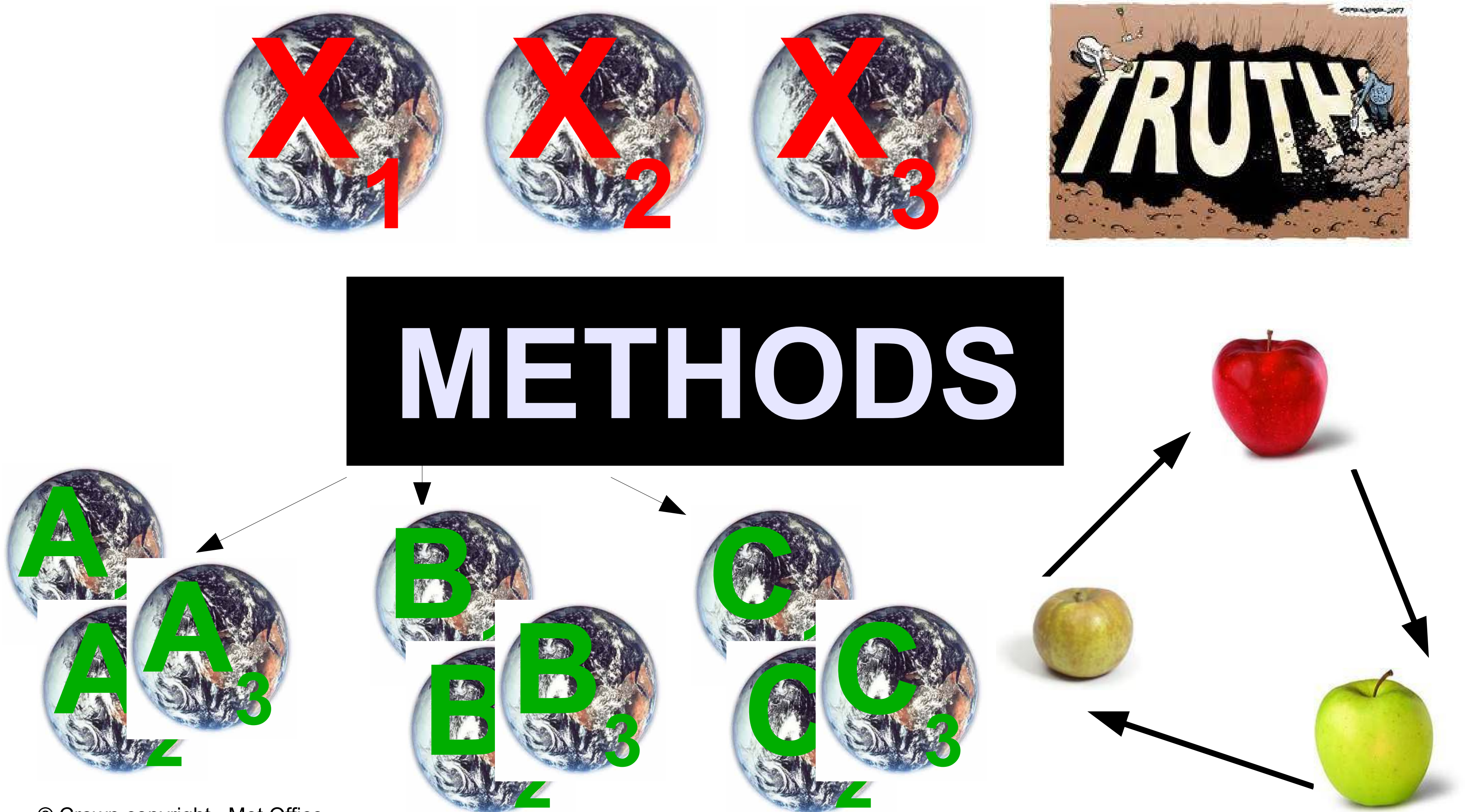


METHODS

What is Benchmarking?



What is Benchmarking?



METHODS

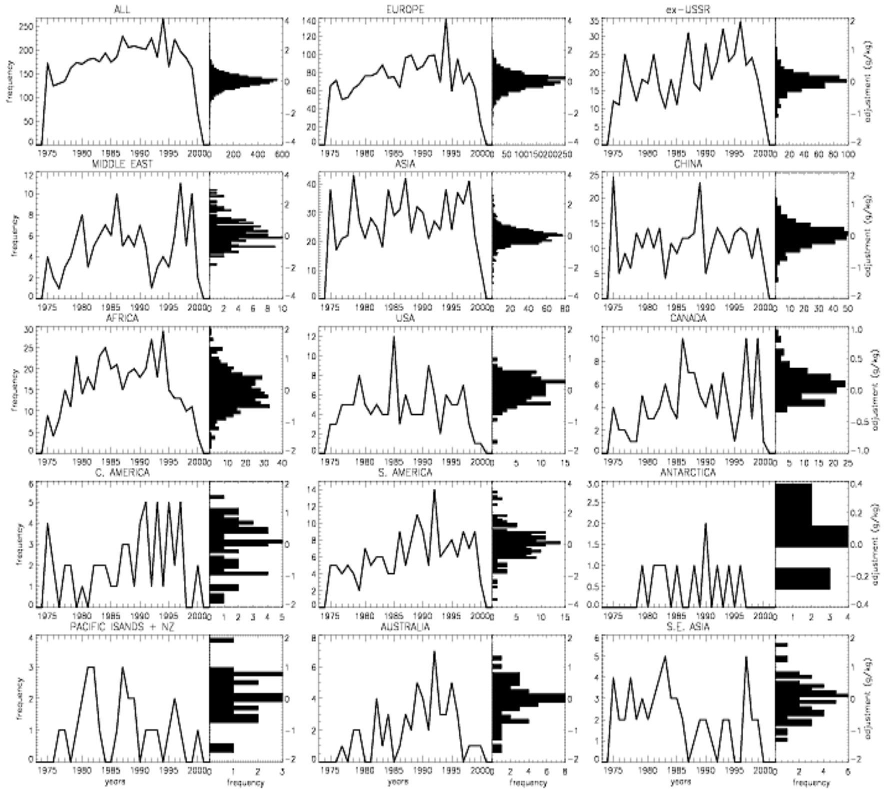
Met Office
Hadley Centre

So why Benchmark?

1) Quantification of methodological uncertainty:

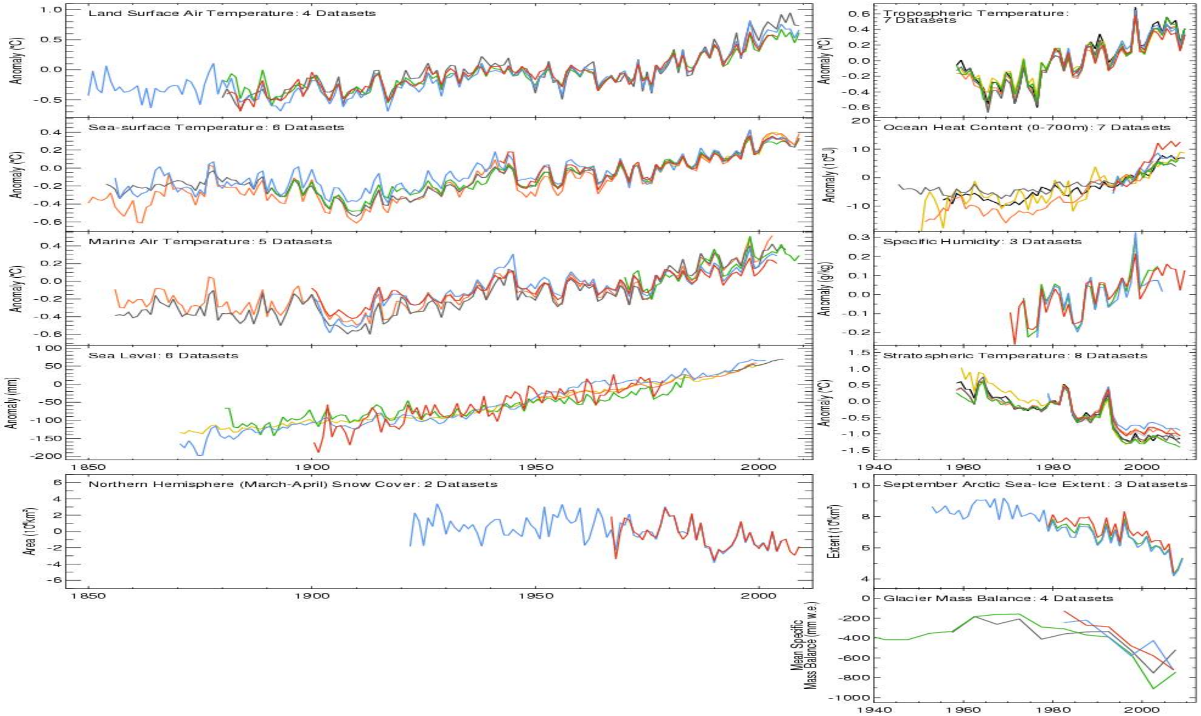
The 'true' climate, free from all random and systematic errors is unknown – therefore we cannot know how close we are to absolute 'truth'.

Understanding the strengths and weaknesses of a data-product methodology against known 'errors' and 'truths' in artificial but realistic data can provide a confidence measure of likely proximity to absolute 'truth' when applied to real data.



So why Benchmark?

2) Informed intercomparison of data-products:



Comparing multiple independent products builds confidence in common features – understanding how and why products differ can provide further confidence

So why Benchmark?

3) Aid advancement of methodologies:

Release of the known 'truth' for the error models will allow data-product creators to test methodologies, understand where weaknesses are and trial improvements

Official benchmarking assessments will be blind to avoid over-tuning but the 'truth' will eventually be released for each benchmarking cycle.

ACMANT
MISH MASH
SNHT
QUANTILE QUANTILE
PMT
MDL
PAIRWISE
CAUSINUS-MESTRE

Met Office

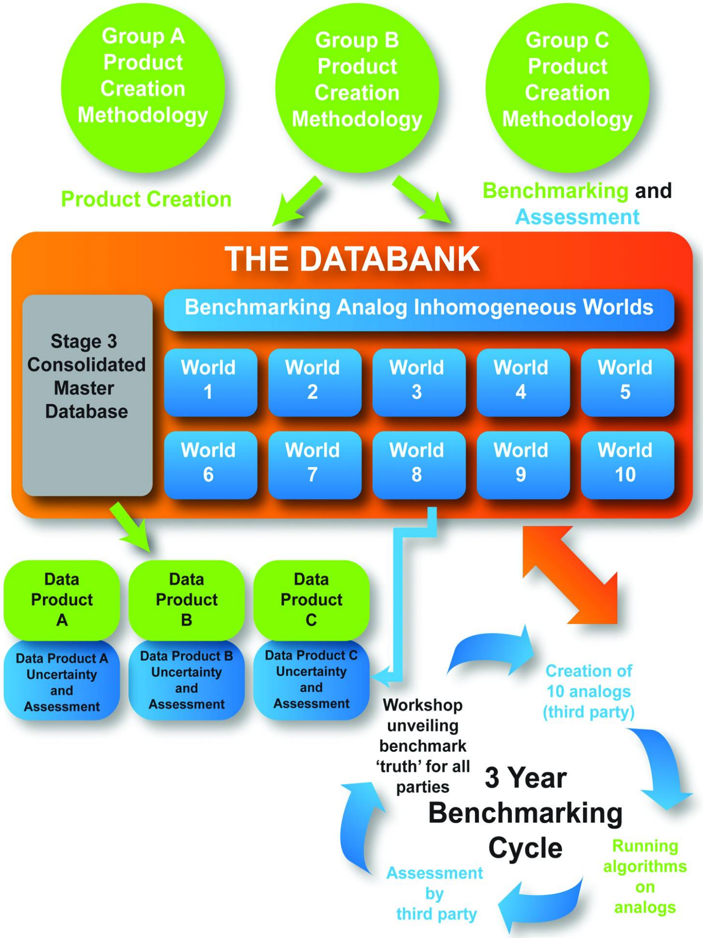
The Benchmarking and Assessment Working Group

Met Office
Hadley Centre

The Benchmarking and Assessment Working Group

Purpose:

To facilitate use of a robust, independent and useful common benchmarking and assessment system for temperature data-product creation methodologies to aid product intercomparison and uncertainty quantification


Group A Product Creation Methodology
Group B Product Creation Methodology
Group C Product Creation Methodology
Product Creation
Benchmarking and Assessment
THE DATABANK
Stage 3 Consolidated Master Database
Benchmarking Analog Inhomogeneous Worlds
World 1
World 2
World 3
World 4
World 5
World 6
World 7
World 8
World 9
World 10
Data Product A
Data Product B
Data Product C
Data Product A Uncertainty and Assessment
Data Product B Uncertainty and Assessment
Data Product C Uncertainty and Assessment
Workshop unveiling benchmark 'truth' for all parties
Creation of 10 analogs (third party)
3 Year Benchmarking Cycle
Running algorithms on analogs
Assessment by third party

BLOGSPOT:
http://surftempbenchmarking.blogspot.com
WEBSITE:
http://www.surfacetemperatures.org/benchmarking-and-assessment-working-group

REVIEW, DEFINE, CREATE, CO-ORDINATE

Kate Willett - Chair (UKMO Hadley Centre, UK), Claude Williams (NCDC, USA), Ian Jolliffe (Exeter Climate Systems, Uni. of Exeter, UK), Robert Lund (Dep. Mathematical Sciences, Clemson Uni., USA), Lisa Alexander (Climate Change Research Centre, UNSW, Australia), Olivier Mestre (Meteo France, France), Stefan Bronniman (University of Bern, Switzerland), Lucie A. Vincent (Climate Research Division, Environment Canada), Aiguo Dai (Climate and Global Dynamics Division, NCAR, USA), Steve Easterbrook (Dep. Computer Science, University of Toronto, Canada), Chris Wikle (Dep. Statistics, University of Missouri, USA), Victor Venema (Meteorologisches Institut, University of Bonn, Germany)

Creating Artificial but Realistic Data with Known 'Truth'

The Artificial Data Must Include Real-World Noise

$$X_{t,l} = S_{t,l} + T_{t,l} + \varepsilon_{t,l}$$

X = Artificial data-point (at TIME t /LOCATION l)

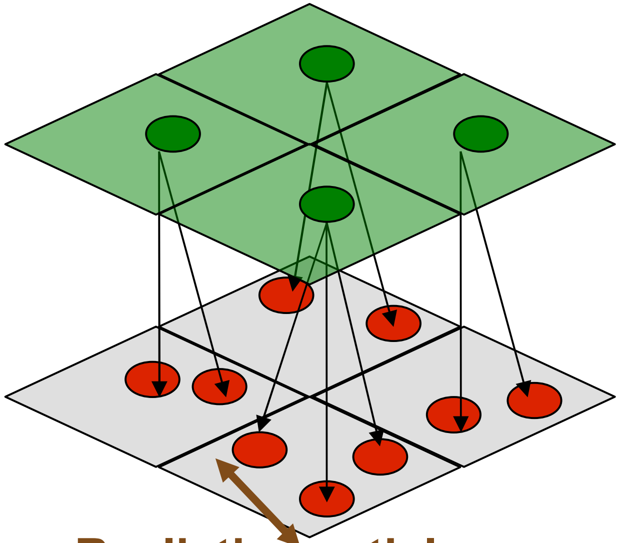
S = seasonal cycles

T = trends (background change, local effects, ENSO, NAO, Volcanoes, Solar Cycles etc.)

ε = random error (recording error, instrument error etc)

With some realistic temporal autocorrelation, spatial covariance structure, data-point characteristics (mean, variance, inter-point correlations)

Downscaling from GCMs to Create Artificial Data-points



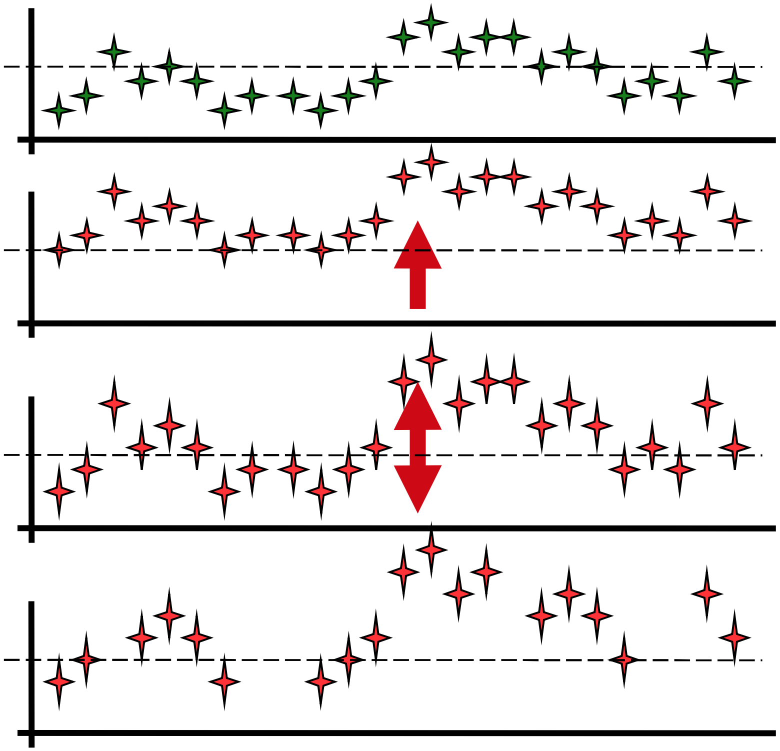
Realistic spatial covariance

GCM gridbox timeseries

adjusted mean

adjusted variance

missing data applied



Met Office

Creating 'Error models' Covering all Known Real-world Nasties

The Artificial Data Must Include Real-World Noise

$$X_{t,l} = S_{t,l} + T_{t,l} + \Sigma_{t,l} + H_{t,l}$$

X = Artificial data-point (at TIME t /LOCATION l)

S = seasonal cycles

T = trends (background change, local effects, ENSO, NAO, Volcanoes, Solar Cycles etc.)

Ʃ = random error (recording error, instrument error etc)

H = inhomogeneity (abrupt, gradual, seasonal, clustered, variance changes etc. - physically governed by radiation and windspeed effects on the specified change)

With some realistic temporal autocorrelation, spatial covariance structure, data-point characteristics (mean, variance, inter-point correlations)

A Suite of Error Models Should Answer A Selection of Big Questions:

Does a background trend (not necessarily linear!) affect inhomogeneity detection/adjustment?

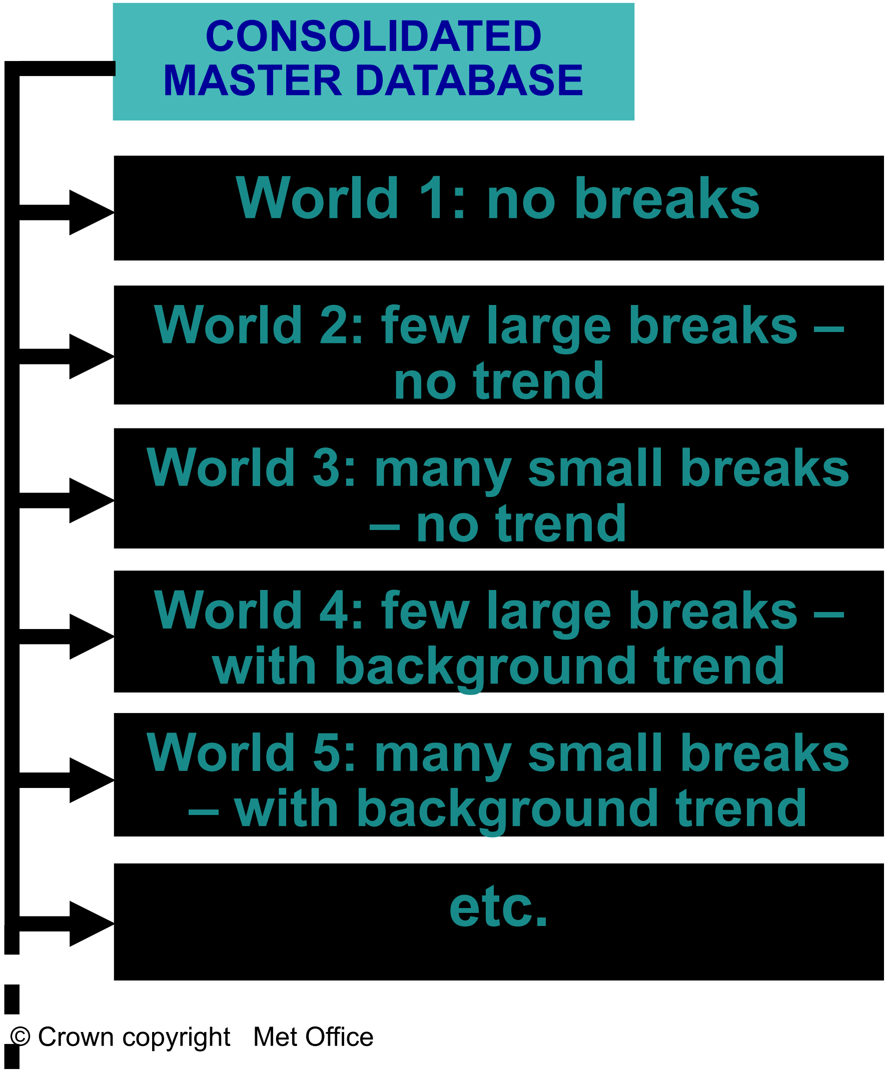
Does metadata provision (null and positive)...?

Does prevalence of many small breaks...?

Does a sign bias...?

Does location of inhomogeneity near record end points...?

Error Worlds



CONSOLIDATED MASTER DATABASE

World 1: no breaks

World 2: few large breaks – no trend

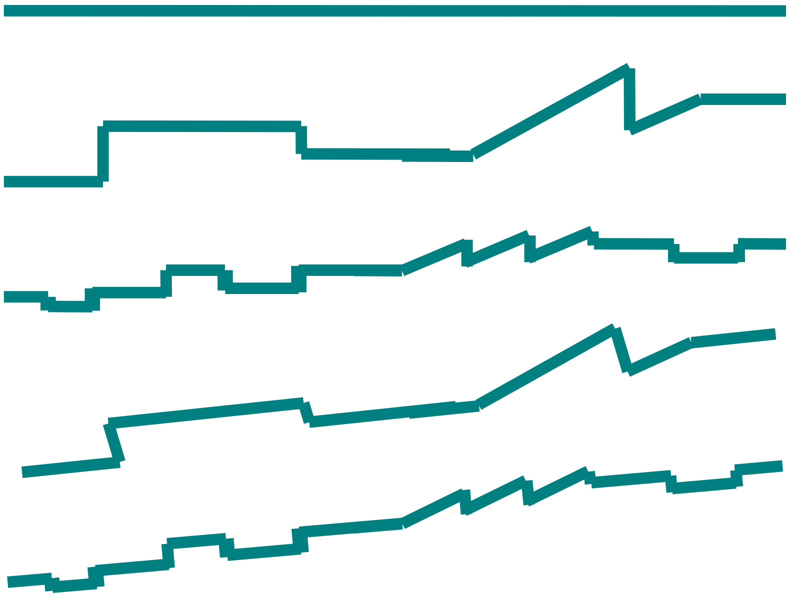
World 3: many small breaks – no trend

World 4: few large breaks – with background trend

World 5: many small breaks – with background trend

etc.

Example error models applied to stations



Assessing the Benchmarks

Assessment

Hit rates and false alarm rates:

Contingency tables:

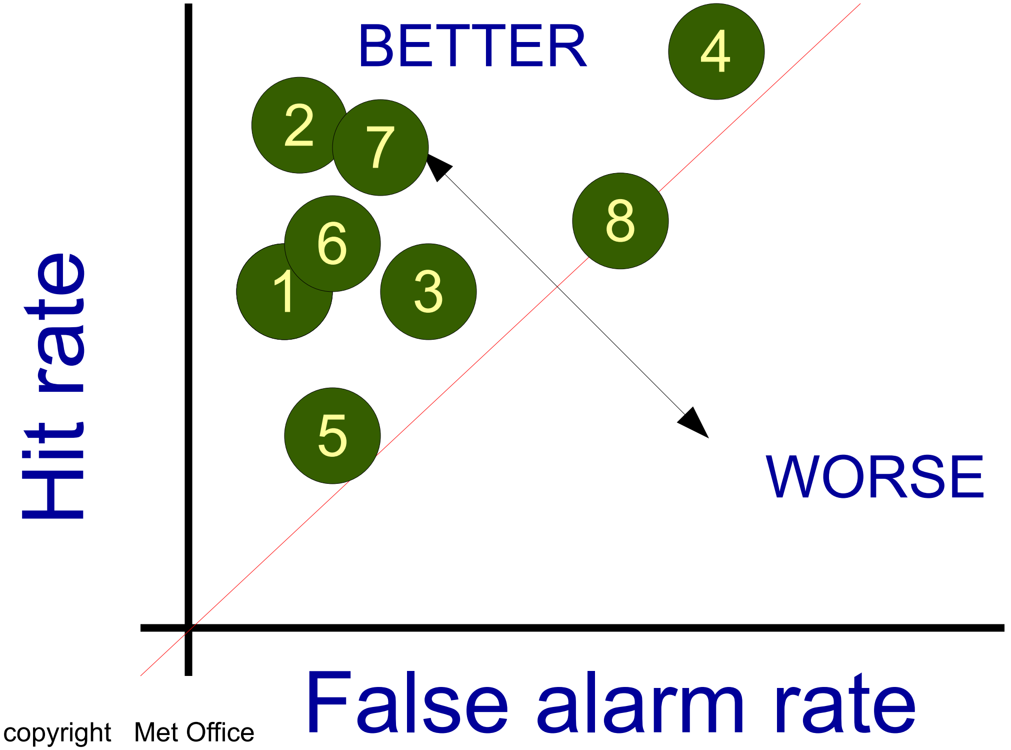
	Changepoint	No Changepoint
Detected (within +/- 3 months)	5	3
Not Detected (within +/- 3 months)	2	42 (potential detections given period of data)

Percent Correct Hit Rate: 90%
Heidke Skill Score = 61%
Probability of Detection hit rate = 71%
False Alarm Rate = 37%

Assessment

Hit rates and false alarm rates:

ROC plots:



Assessment

Closeness to world Truth:

RMSE for:

Climatology

Variance

Trends

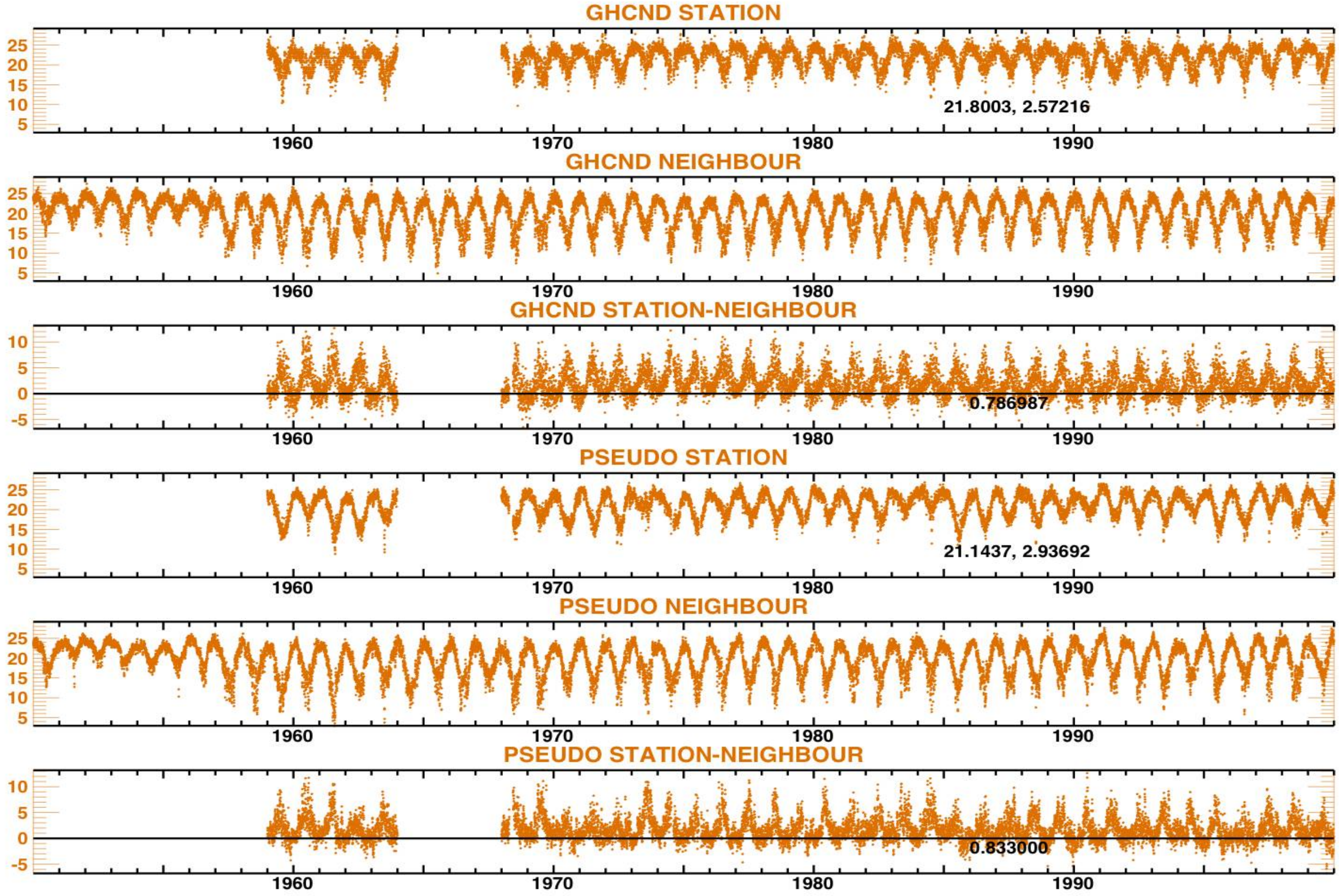
Met Office

Are such techniques useful within the marine community?

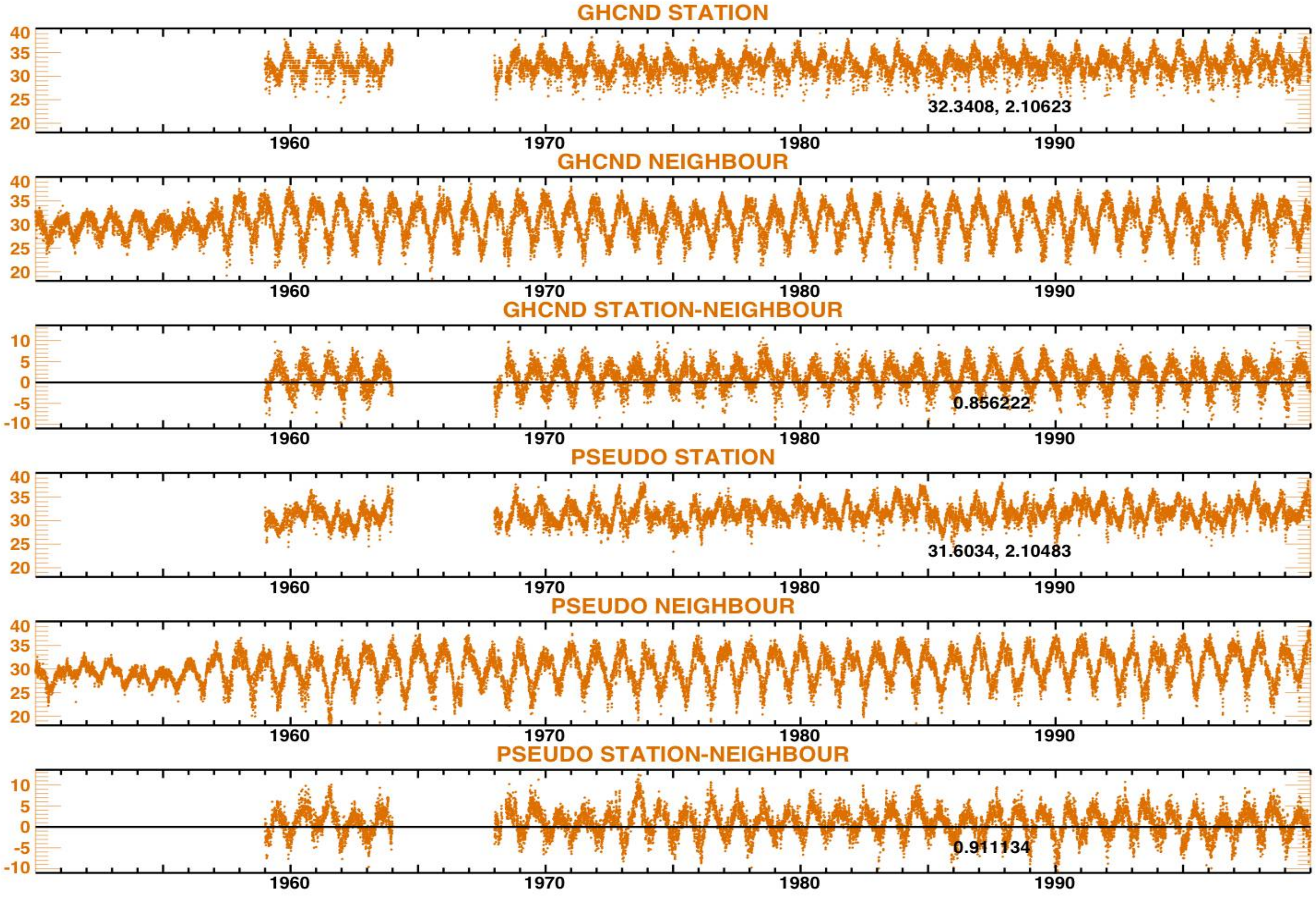
Met Office

My Pseudo-Worlds and Error Models

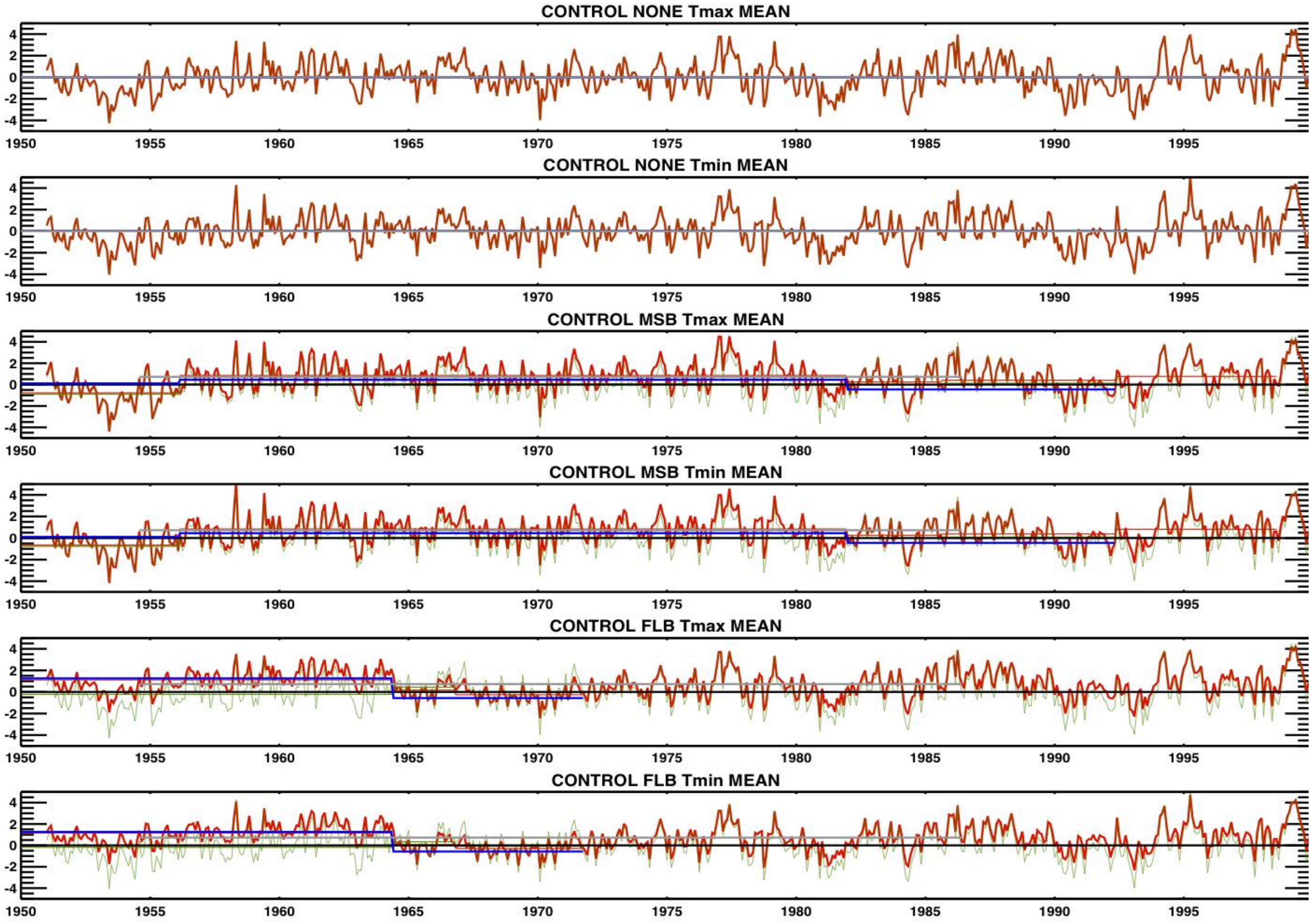
Creating the 'truth'



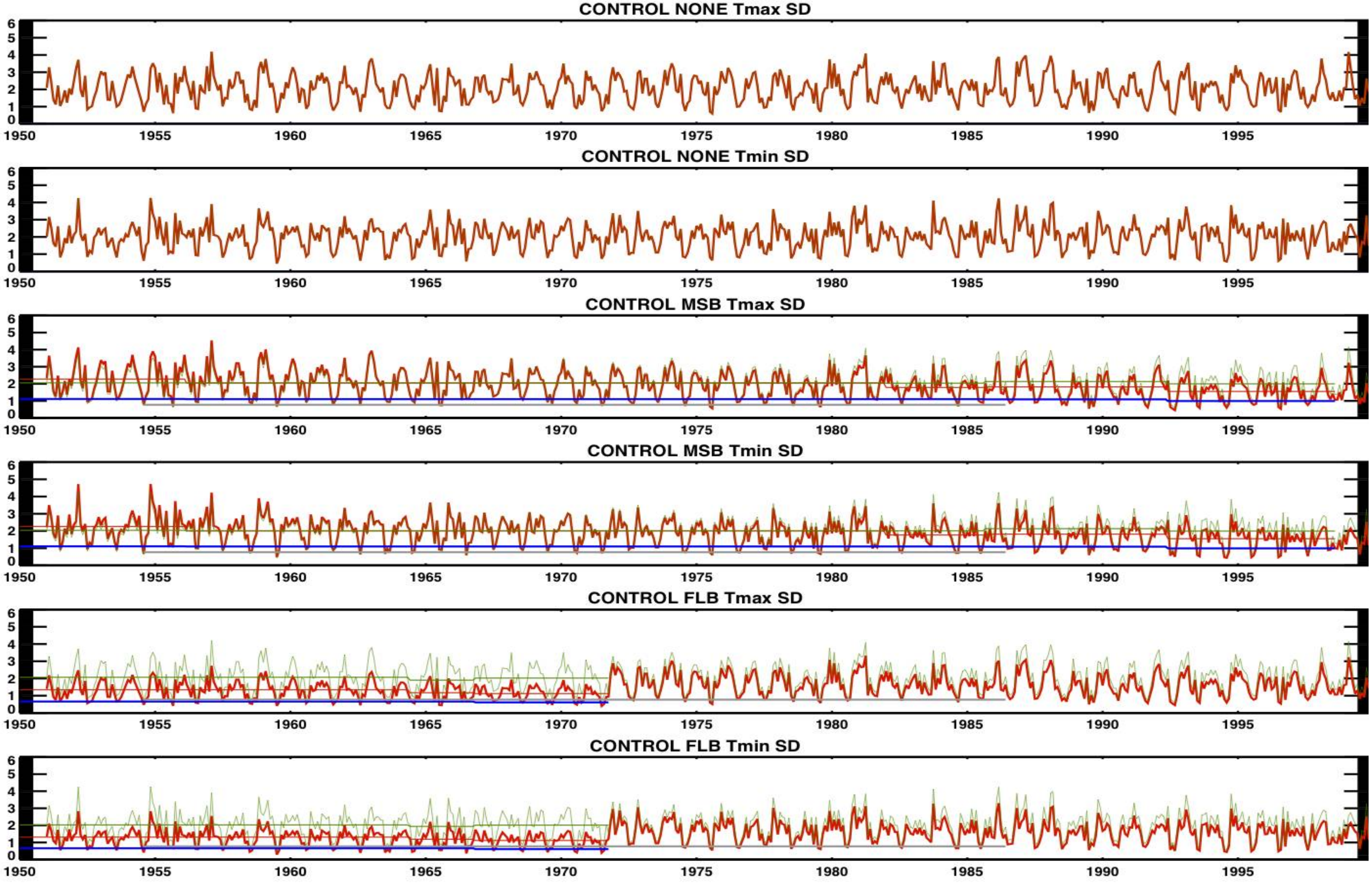
Creating the 'truth'



Creating the 'nastiness'



Creating the 'nastiness'



CONTROL NONE Tmax SD

CONTROL NONE Tmin SD

CONTROL MSB Tmax SD

CONTROL MSB Tmin SD

CONTROL FLB Tmax SD

CONTROL FLB Tmin SD

Met Office

Help!

Real World Nastiness to Include?

Spatial covariance, white noise random error, ENSO etc.

Changepoint Structure

Amount, type, physical characteristics, clustered, metadata, size...

Usefulness of Assessment

Ability to detect changepoints

Ability to adjust timeseries correctly

Ability to cope with/without metadata

etc.

Causes of Inhomogeneity in Marine Data

- Change to predominant observation type over a region (ship, buoy, fixed platform etc.)
- Change to predominant observing instrument type
- Change to observing practices (observing time, rounding practices etc.)
- Change in observation height (bigger ships over time)
- Change in observation density
- Blended Land/Ocean products may see a shift from Land obs to Ocean obs (or vice versa) over time