Benchmarking and Assessment (Verification) of Homogenisation Algorithms for the International Surface Temperature Initiative (ISTI)

Kate Willett (Met Office Hadley Centre), Steve Easterbrook (University of Toronto), Claude Williams (NCDC), **Ian Jolliffe (University of Exeter),** Robert Lund (Clemson University), Lisa Alexander (University of New South Wales), Olivier Mestre (Meteo France), Stefan Brönniman (University of Bern), Lucie A. Vincent (Environment Canada), Aiguo Dai (NCAR), Victor Venema (University of Bonn), David Berry (National Oceanography Centre)

OVERVIEW

ISTI aims to facilitate transparent creation of multiple long, high resolution, traceable data-products that are robust to varying non-climatic influences. The Benchmarking group of ISTI will create artificial, but realistic, data sets on which homogenisation algorithms can be tested.

The performance of the algorithms needs be assessed. This is the remit of Team Validation (see 4 below) and is equivalent to forecast verification. Carefully crafted validation will:

- aid objective intercomparison of multiple data-products;
- provide a quantifiable measure of uncertainty;
- facilitate homogenisation algorithm development

1. The Benchmarking and Assessment Program

Temperature benchmarks will replicate the ISTI Databank stations and format. **Analog-known-worlds** are semi-synthetic data, free from inhomogeneity. **Analog-error-worlds** are created from **analog-known-worlds** exploring plausible inhomogeneities (breaks).

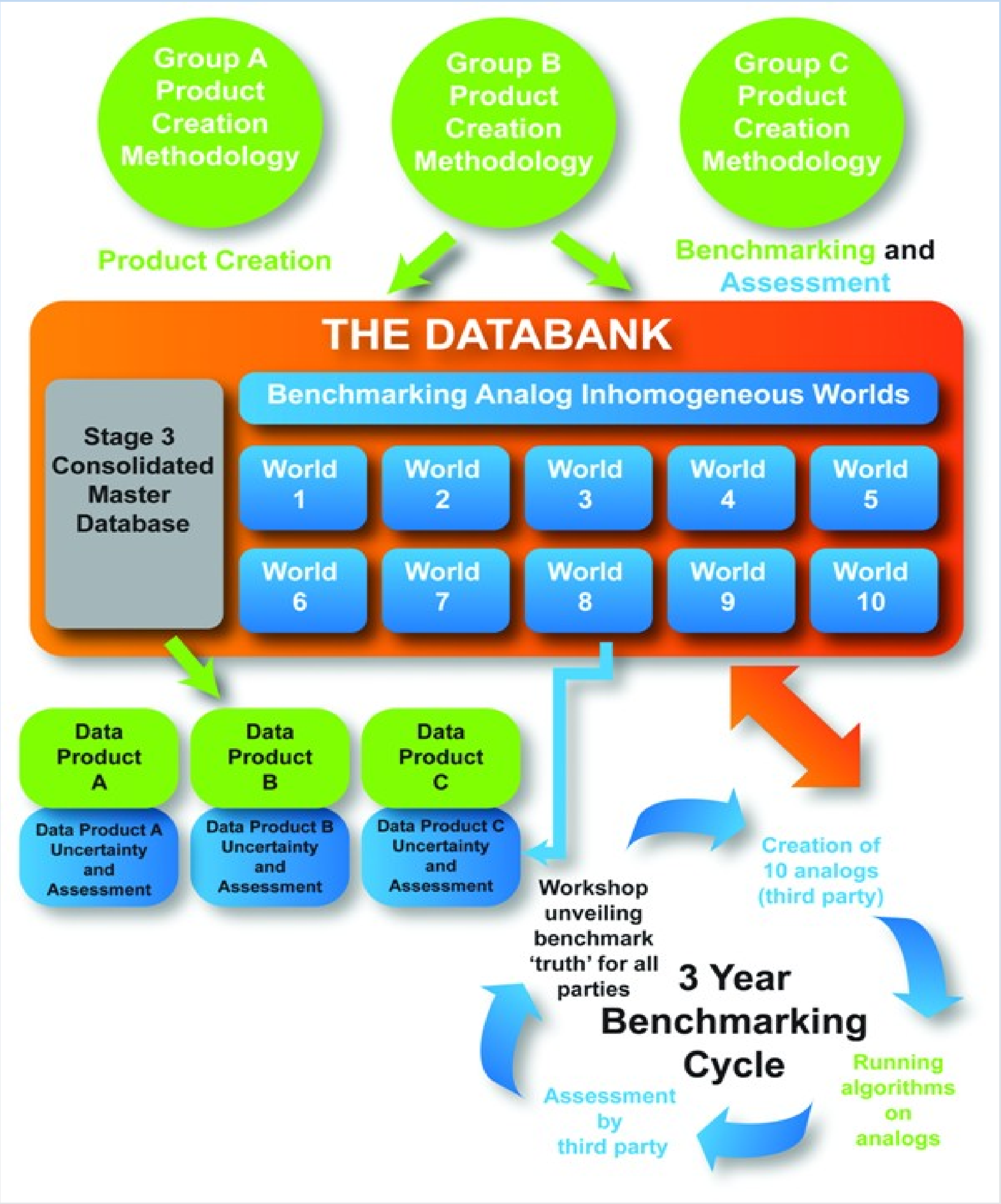
A pilot set of **analog-known-worlds** and **analog-error-worlds** will be made available with the Databank as an immediate resource for algorithm developers rather than waiting for the 3 year cycle to end.

The benchmark cycle will use a different set of **analog-known-worlds** and blind **analog-error-worlds**. The **analog-error-worlds** will be released 8 months after Databank version 1 but **analog-known-worlds** withheld for 2.5yrs to prevent algorithm overtuning.

Data-product creators will have 2.5yrs to use the benchmarks. Product assessments will summarise both the ability to detect and to correctly adjust the breaks. This stage is equivalent to forecast verification.

After 2.5yrs the **analog-known-worlds** will be released and an analysis of the value/success/failure/areas for improvement of the benchmarks will be published. A 'wrap-up' workshop will be held bringing together the benchmark designers and data-product creators.

A new set of benchmarks will be created and the **analog-error-models** released to begin the cycle again.


Group A Product Creation Methodology
Group B Product Creation Methodology
Group C Product Creation Methodology
Product Creation
Benchmarking and Assessment
THE DATABANK
Benchmarking Analog Inhomogeneous Worlds
Stage 3 Consolidated Master Database
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
3 Year Benchmarking Cycle
Creation of 10 analogs (third party)
Running algorithms on analogs
Assessment by third party

Benchmarking and Assessment Working Group website:
www.surfacetemperatures.org/benchmarking-and-assessment-working-group

Benchmarking and Assessment Working Group open blogsite:
surftempbenchmarking.blogspot.com

Website for the Surface Temperature Initiative website:
www.surfacetemperatures.org

Other related project:
COST HOME
www.homogenisation.org

Fig. 1 Schematic of how the benchmarking cycle will work. Benchmarks will be available as part of the Surface Temperature Databank for data-product creators to test their algorithms on.

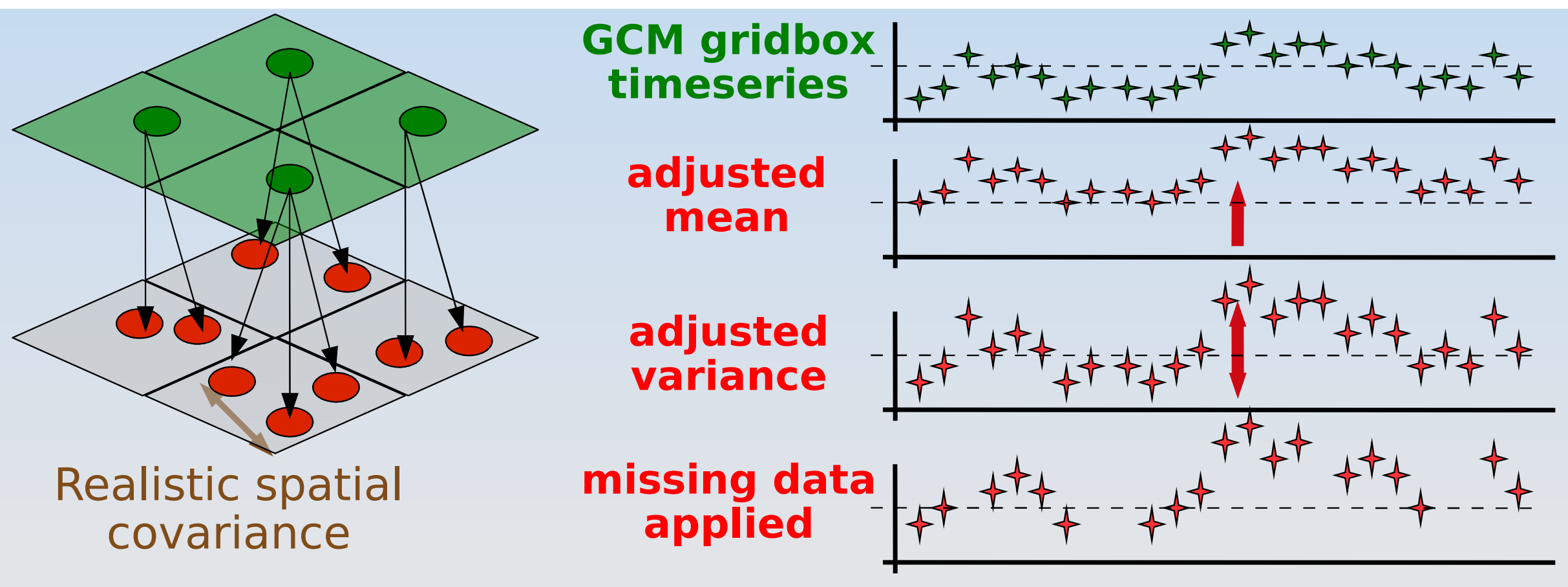
2. Task Team Creation: design and create analog-known-worlds

Create a global station network with real-world properties (climatology, natural variability, autocorrelation, missing data, spatial covariance etc...). Climate Models provide homogeneous data with background trends (T) and seasonal cycle (S). Real-world properties (t,l,h) are obtained from the Databank and white noise error (ε) added.

$$X_{\mathrm{TRUTH}(t,l,h)} = S_{t,l,h} + T_{t,l,h} + \varepsilon_{t,l,h}$$

X = benchmark analog station at time *t*, location *l* and height *h*
S = seasonal cycles
T = trends (long-term signal, local effects, ENSO, NAO, Volcanoes, Solar Cycles etc.)
ε = random error at time/place/height (recording error, instrument error etc)

Fig. 2 Diagram of simple GCM to analog station downscaling. Grid box time series are nudged to match the mean, variance and missing data of real-world stations.



3. Task Team Corruption: design and create the analog-error-worlds

Design a set of breaks (B) - plausible worlds ranging from optimistic (e.g., few large breaks) to pessimistic (e.g., many breaks abrupt and gradual, seasonally varying in mean and variance) (Fig. 3). Replicate the physics of instrument moves/changes/degradation, local environment change, etc. - depends on radiation (hour, date, latitude, cloudiness) and wind speed. Apply to the **analog-known-worlds.**

$$X_{\mathrm{ERROR\_WORLD}(t,l,h)} = X_{\mathrm{TRUTH}(t,l,h)} + B_{\mathrm{ERROR\_WORLD}(t,l,h)}$$

B = break at time/place/height (abrupt, gradual, seasonal, clustered, variance changes etc)

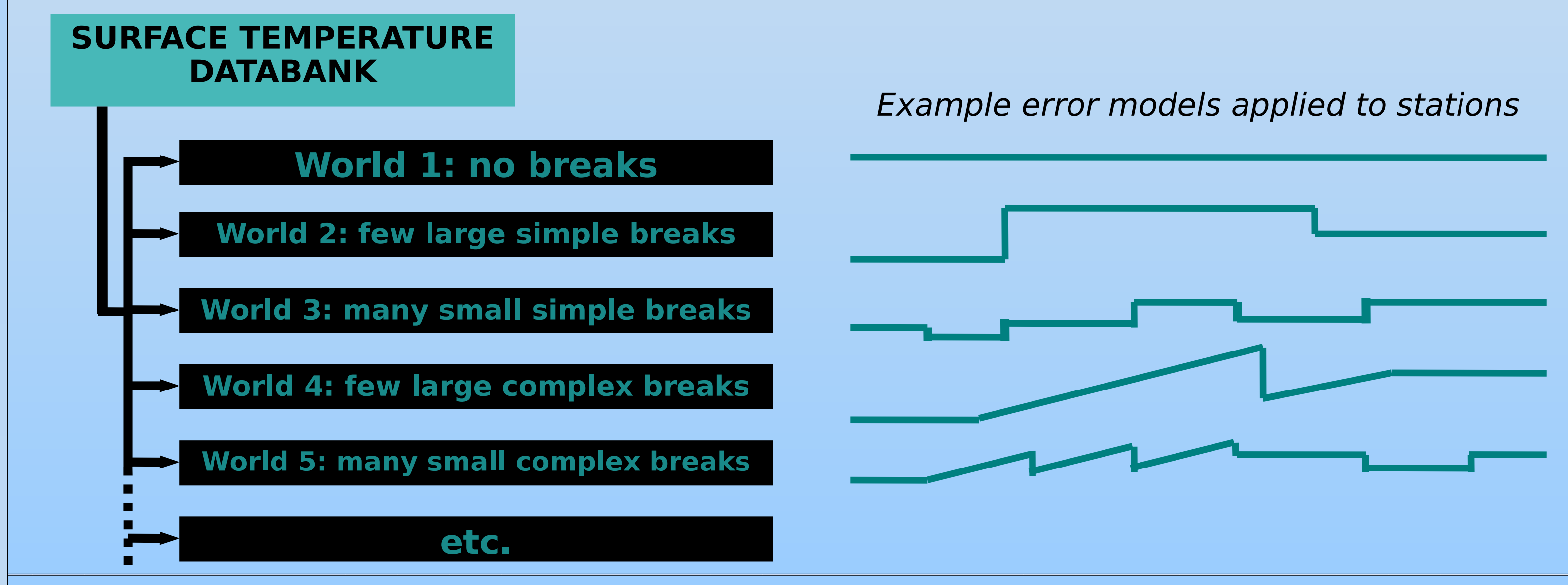

SURFACE TEMPERATURE DATABANK
Example error models applied to stations
World 1: no breaks
World 2: few large simple breaks
World 3: many small simple breaks
World 4: few large complex breaks
World 5: many small complex breaks
etc.

Fig. 3 Diagram of example error structure for the analog-error-worlds.

4. Task Team Validation: design assessment criteria and tools

Benchmarking assessment should test algorithm ability to detect breaks and to 'correct' for non-climatic influences. Comparing homogenised **analog-error-worlds** with **analog-known-worlds** corresponds to forecast verification using existing and novel verification techniques.

Assessing break detection reduces to analysing a (2x2) contingency table, for which standard measures such as hit rate, false alarm rates and many others can be used. However, there are complications such as variable detection rates along a series, differential weighting of outcomes, and the definition of true negatives.

For comparison of individual series, for **analog-known-worlds** with corresponding homogenised **analog-error-world** series, standard measures such as RMSE can be used. However consideration needs to be given on how to combine measures over regions, how to assess the uncertainty associated with values of the measures, and whether measures less sensitive to outliers, such as MAE, might be used.

It is also of interest to compare distributions of data across series. Do the homogenised **analog-error-worlds** successfully reproduce the means, variances, trends, autocorrelations, spatial correlations etc. in the **analog-known-worlds**?

International Surface Temperature Initiative

For further information or expressions of interest please contact: **I.T.Jolliffe@exeter.ac.uk**

www.surfacetemperatures.org
general.enquiries@surfacetemperatures.org
data.submission@surfacetemperatures.org