Benchmarking and Assessment 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 (to source and known standards) data-products that are robust to varying non-climatic influences.

Multiple independently created and homogenised data-products improve understanding of the strengths and weaknesses of methodological choices and build confidence in common conclusions.

Benchmarking of homogenisation algorithms 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 Land Surface 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 inhomogeneity characteristics.

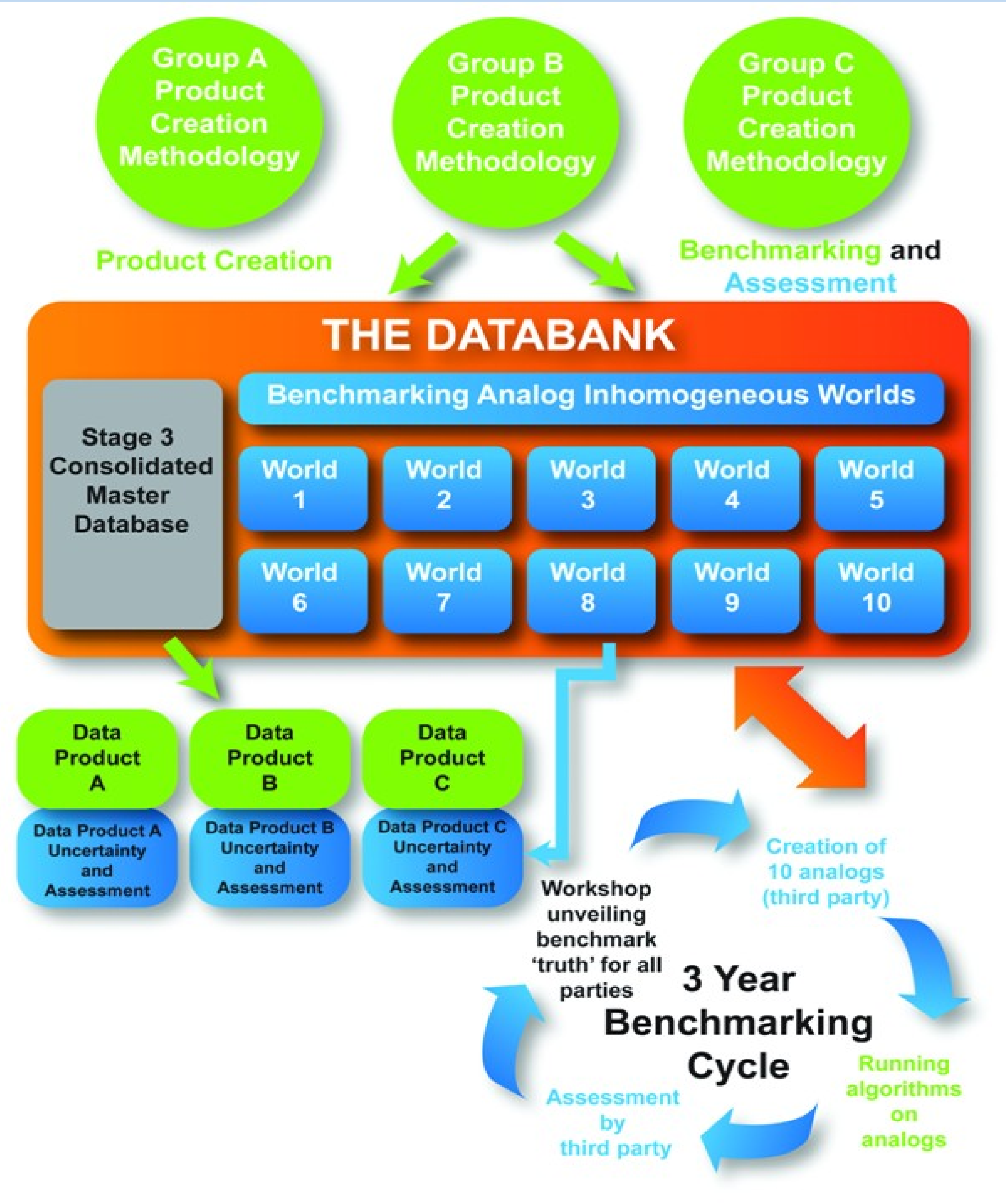
A pilot release of both **analog-known-worlds** and **analog-error-worlds** will be made after the initial databank release to provide an immediate resource for algorithm developers rather than having to wait for the 3 year cycle to end.

For the benchmark cycle a different set of **analog-known-worlds** and **analog-error-worlds** will be created. The **analog-error-worlds** will be released 8 months after the version 1 Databank. The **analog-known-worlds** will be withheld for 2.5 years to prevent algorithm overtuning.

Data-product creators will have 2.5 years to use the benchmarks. An assessment will be provided summarising both the ability to detect and to correctly adjust for inhomogeneities.

After 2.5 years the **analog-known-worlds** will be released and an assessment 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
Creation of 10 analogs (third party)
3 Year Benchmarking Cycle
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 projects:
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

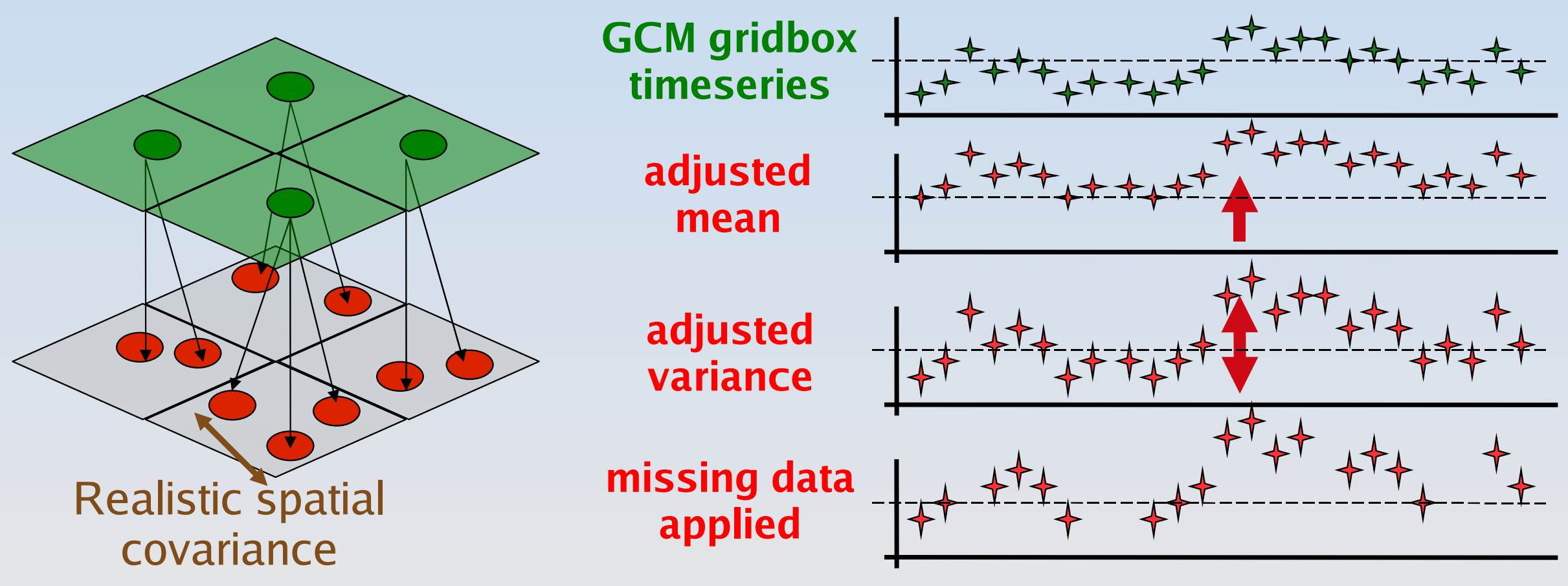
Create a global station network that reflects real-world properties (climatology, natural variability, autocorrelation, missing data and covariance with neighbours) without systematic bias.

GCMs provide homogeneous base series with background trends (T) and seasonal cycle (S). Real-world properties are obtained from the Databank and white noise error (ε) added.

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

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 errors (B) - plausible worlds scaling from overly optimistic (e.g., few large breaks) to overly pessimistic (e.g., many different breaks with gradual changes, seasonally varying changes in the mean and variance) addressing specific questions (Fig. 3). Apply to the **analog-known-worlds.**

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

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

Errors should reflect the physics of instrument/location changes, systematic instrument degradation, local environment change, etc often depending on radiation (hour, date, latitude, cloudiness) and wind speed.

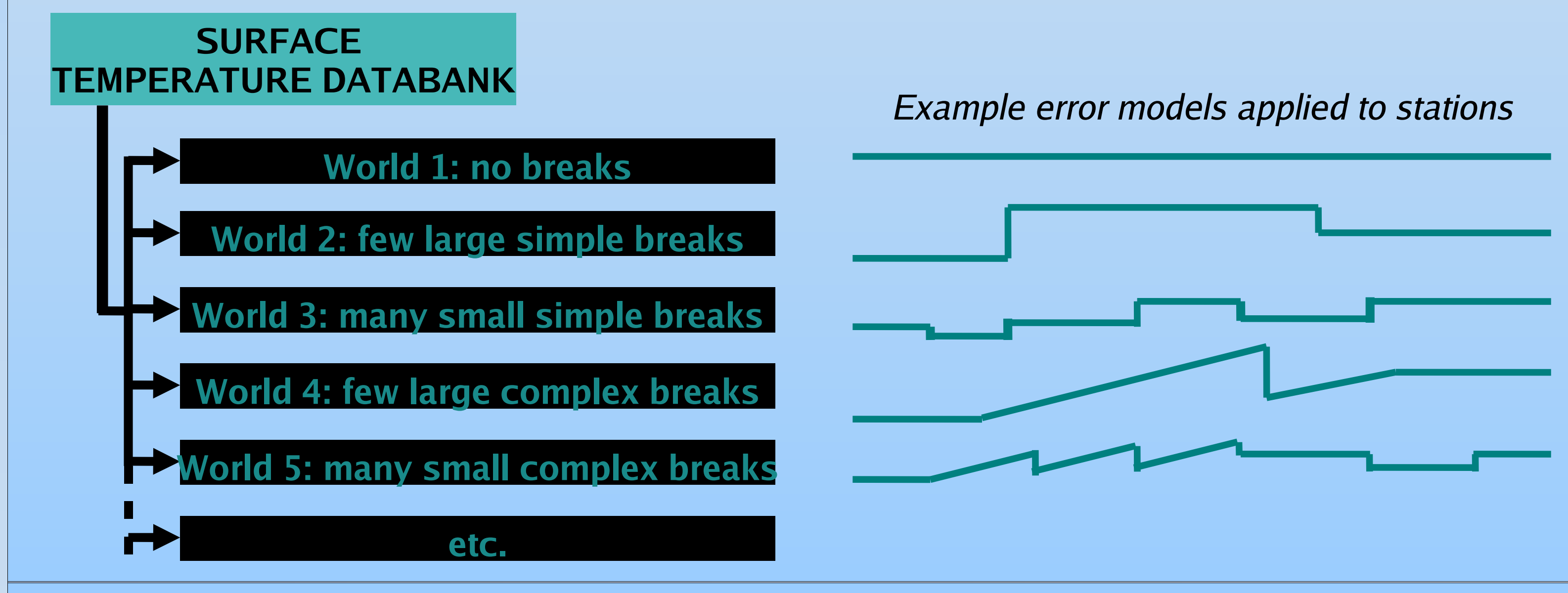

SURFACE TEMPERATURE DATABANK
Example error models applied to stations
World 1: no breaks
World 2: few large simple breaks
World 3: many small simple breaks
World 4: few large complex breaks
World 5: many small complex breaks
etc.

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 the ability of algorithms to detect breaks and the ability to 'correct' the data for non-climatic influences.

Contingency tables could be used to assess **hit rate and false alarm rate** - taking into account correct sign, location and magnitude within an acceptable range of error.

Statistical property comparisons could measure the proximity of each homogenised **analog-error-world** mean state to the **analog-known-world** mean state at both station and region level – how similar are region climatologies, variance, background trends, station autocorrelation, neighbour covariance?

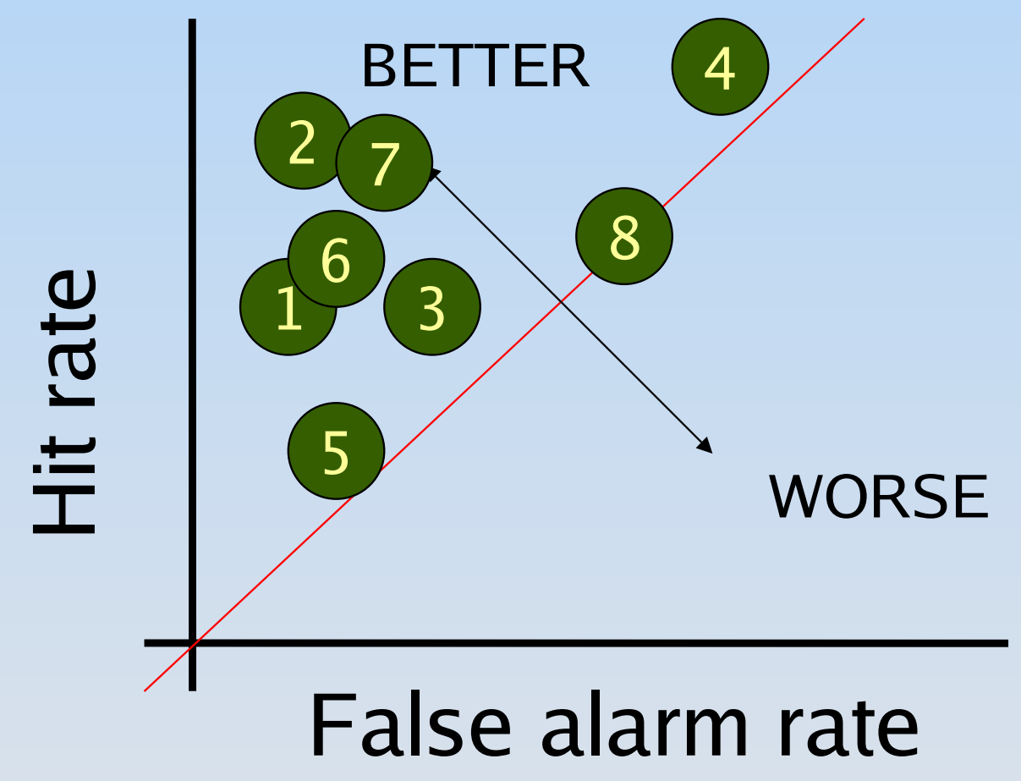


Fig. 4 Diagram of example detection ability assessment for the analog-error-worlds.

International Surface Temperature Initiative

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