**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 to be assessed. This is the remit of Team Validation (see 4 below) and is equivalent to forecast verification. Carefully crafted validation will assist 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 formats. Analog-known-worlds are semi-synthetic (i.e., 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 overfitting.

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.

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.

\[ X_{\text{truth}}(t,l,h) = S(t,l,h) + T(t,l,h) + \varepsilon(t,l,h) \]

\[ X = \text{benchmark analog station at time } t \text{, location } l \text{ and height } h \]

\[ S = \text{seasonal cycle} \]

\[ T = \text{trends (long-term signal, local effects, ENSO, NAO, Volcanoes, Solar Cycles etc.)} \]

\[ \varepsilon = \text{random error at time/place/height (recording error, instrument error etc)} \]

**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 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.

\[ X_{\text{error-world}}(t,l,h) = X_{\text{truth}}(t,l,h) + B(t,l,h) \]

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

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

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-worlds, 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 mean, variances, trends, autocorrelations, spatial correlations etc. in the analog-known-worlds?

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

---

**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)**