Robust benchmarking of homogenisation algorithms for the Surface Temperature Initiative

Katharine Willett (1), Peter Thorne (2), Lisa Alexander (3), and Robin Chadwick (1)

(1) UK Met Office Hadley Centre, FitzRoy Road, Exeter, EX1 3PB, UK, (2) Cooperative Institute for Climate and Satellites – North Carolina, NOAA's National Climatic Data Center, Asheville, NC, USA, (3) Climate Change Research Centre, University of New South Wales, Sydney, Australia

The 21st Century requirements on Climate Science call for longterm, spatially widespread observational datasets that are robust to varying non-climatic influences over time. While observed data cannot be restored to absolute ground-truth, we can improve our understanding of the strengths and weaknesses of our many methodologies and it is essential that we do so. Benchmarking our homogenisation algorithms against a known reference will improve our understanding and quantification of uncertainty. It also allows some meaningful intercomparison of methodologically independent datasets – where by the nature of their methodological choices, some will be more suited to certain applications than others.

As part of the Surface Temperature Initiative (www.surfacetemperatures.org) a Benchmarking and Assessment working group was set up to co-ordinate efforts in this area. The group aims to learn from and work with existing efforts with the goal of organising a benchmarking and assessment program. This will run alongside data-product creation resulting from the Surface Temperature Initiative. The ultimate aim of the Initiative is to aid the creation of robust surface temperature data-products by providing a comprehensive data-bank where data are: version controlled; traceable to known origin; associated with observation metadata; and freely available.

To aid progress within this area, we are investigating a more involved benchmarking method than has been undertaken previously – utilising spatially complete GCM data downscaled to match real observing network characteristics (climatology, variance, autocorrelation with neighbouring stations). This gives us a known 'truth' to which a number of 'pseudo-worlds' of likely random and systematic errors can be added. It allows us to test each homogenisation method by comparing the resulting trends, mean state and other characteristics to the known 'truth' and quantify the percentage of detection and false alarm rates. Here we describe our progress to date in this area.