

Robust benchmarking of homogenisation algorithms as part of the Surface Temperature Initiative

Katharine Willett[†]; Steve Easterbrook; the Benchmarking and Assessment working group for the International Surface Temperature Initiative

[†] Met Office Hadley Centre, United Kingdom

Leading author: kate.willett@metoffice.gov.uk

The 21st Century requirements on Climate Science call for long-term, 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 allow 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. It will also improve our understanding and quantification of uncertainty. Furthermore, the provision of realistic global scale error models with known truths should enhance methodological development. As part of the Surface Temperature Initiative (www.surface temperatures.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. 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 detection and false alarm rates. Here we describe our progress to date in this area.