Creating surface temperature datasets to meet 21st Century challenges

Met Office Hadley Centre, Exeter, UK

7th-9th September 2010

White papers background

Each white paper has been prepared in a matter of a few weeks by a small set of experts who were pre-defined by the International Organising Committee to represent a broad range of expert backgrounds and perspectives. We are very grateful to these authors for giving their time so willingly to this task at such short notice. They are not intended to constitute publication quality pieces – a process that would naturally take somewhat longer to achieve.

The white papers have been written to raise the big ticket items that require further consideration for the successful implementation of a holistic project that encompasses all aspects from data recovery through analysis and delivery to end users. They provide a framework for undertaking the breakout and plenary discussions at the workshop. The IOC felt strongly that starting from a blank sheet of paper would not be conducive to agreement in a relatively short meeting.

It is important to stress that the white papers are very definitely not meant to be interpreted as providing a definitive plan. There are two stages of review that will inform the finally agreed meeting outcome:

1. The white papers have been made publicly available for a comment period through a moderated blog.
2. At the meeting the approx. 75 experts in attendance will discuss and finesse plans both in breakout groups and in plenary. Stringent efforts will be made to ensure that public comments are taken into account to the extent possible.

White Paper on Dataset algorithm performance assessment based upon all efforts

Peter Stott[1], Philip Brohan[1], Dick Dee[2], Alistair Forbes[3], Peter Guttorp[4], Ian Jolliffe[5], Peter Thorne[6]

[1] Met Office Hadley Centre, FitzRoy Road, Exeter, EX1 3PB, UK
[2] ECMWF, Reading, UK
3 National Physical Laboratory
[4] University of Washington, Seattle, USA
[5]University of Exeter, Exeter, EX4 4SB, UK
[6] National Climatic Data Center, USA.

Introduction

This white paper provides a consideration of the steps that need to be taken to provide suitable performance assessment for algorithms used in the development of climate datasets. It discusses how to make use of the benchmarking results described in White Paper no 9 in addition to other methodological decisions made in order to assess suitability of datasets for a particular purpose. Fundamentally this White Paper is therefore about how well derived datasets represent actual real world behaviour and the attendant uncertainties in that assessment. The technical details of the development and testing of benchmarking homogenization algorithms are provided in White Paper no 9.

Considerations for Dataset Algorithm performance assessment

In testing and validating dataset algorithm the first step is to determine what the purpose is for which the dataset is being assessed. Only then does it become feasible to determine how good a particular algorithm is for that particular purpose. Such uses could include, for example, determination of long term mean trends, annual maximum values or diurnal variability of a particular climate variable.

Assessment criteria should be developed independently of algorithm development teams building on suitable expertise gained from a variety of perspectives, ie not just the climate science community. It is strongly desirable that assessment criteria should be determined before any assessments are carried out. Such criteria should be clearly documented and should be approved by the core implementation group who have the task of over-seeing progress on the initiative to develop surface

temperature datasets to meet 21st century challenges as set out in White Paper 15.

It is important that new activities resulting from this initiative are coordinated with on-going national and regional activities to rescue and homogenize data, including for example the European COST-ES0601 initiative to produce a benchmark dataset. Further details of some of these initiatives are given in White Paper no 9.

Possible assessment criteria, depending on the use to which the data are being used, could include an ability to identify long term trends as well as to what extent the algorithm has retained the variance of the climate variable and of the ability of the algorithm to identify and adjust for individual breaks in data series as well as a suitable criteria for the number of acceptable false-positive identification of non existent break points. Further examples of possible assessment criteria include whether simulated data series have non-stationarity in the mean (beyond just seasonal effects), include both short- and longterm "memory" (or temporal correlation) and incorporate realistic continental or global scale spatial correlations. Homogeneity issues are discussed in greater depth in White Paper No 9. They are raised here as examples of assessment criteria that could be appropriate to a particular user and which will therefore need to be assessed.

The assessed quality of an algorithm for a particular purpose would likely depend on the risks associated with making wrong inferences on the basis of the algorithm outputs. If the risks/uncertainties associated with the outputs of an algorithm are deemed to be too high according to the users, it would be useful to determine what new information would reduce the risks optimally, for example would it be better to replace many inexpensive sensors with limited accuracy with a few highly accurate sensors.

A key issue is to determine how well uncertainty estimates in datasets provide a measure of the difference between the derived value and the “true” value, recognizing that any uncertainty statement summarises the uncertainty from all factors, including those that operate over different timescales. Establishing traceability is essential as a first step in combining measurements from different measurement systems. Most data analysis algorithms are designed to operate optimally (unbiased, minimum variation) for data generated according to a particular model. For a fully characterised algorithm, the model predicts the behaviour of the algorithm, allowing uncertainties to be associated with the parameter estimates, for

example. Therefore assessments should be carried out taking account of the sensitivity of algorithm outputs with respect to underlying model assumptions.

It would be worthwhile to consider the future needs for the development of climate services by identifying an appropriate common set of regions or stations that any assessment should include while recognising that not all dataset creators may include these stations or regions. It is suggested that analyses should be carried out to cover densely, moderately and poorly sampled regions as the performance of any algorithm is likely to depend upon network density. Thought would also be needed as to whether it is appropriate to apply a common set of assessment criteria to all regions or whether different assessment criteria are required for each region.

Validation of an algorithm should always be carried out on a different dataset from that used to develop and tune the algorithm. As discussed in White Paper No 9, idealised datasets can be created which contain break points, data gaps, etc, for example by sub-sampling and distorting climate model data. Once an algorithm is frozen, the validation dataset can be used to test it by measuring against the pre-determined application-specific assessment criteria.

White Paper no 9 discussed different approaches for testing homogenization algorithms including in idealized model settings in which potential data gaps, inhomegeneities and other data issues are imposed, reanalyses, and the use of high quality reference datasets. Such model-based tests will always be conditional on the idealized nature of the tests and therefore cannot consider all sources of uncertainty. These uncertainties can be amenable to exploration through the use of reanalyses, which combine observational data with the dynamical constraints of a forecasting model in which multiple climate variables are constrained to be physically consistent. Purely statistical approaches and dynamical approaches (via reanalyses) can be used to validate each other. In addition, a testbed complementary to testing algorithms in models in which deliberate errors have been inserted, is to insert similar such deliberate errors into reference very high quality datasets such as CRN/CRN-HCN dataset which has realistic, irregular grid spacing and variability. Such a test would be good for the smaller scale (country to regional) spatial testing of algorithms as well as for testing spatial-interpolation errors.

Based on these tests, users should be able to determine which are

suitable datasets and approaches for their application. The aim of any assessment should be to determine and flag clearly any datasets that are not fit for a particular purpose and to aid users in determining which are the most suitable datasets for their needs, recognizing importantly that some approaches which may be beneficial for one purpose, eg large-scale trend retrieval, may be detrimental for other features, eg station by station quality. Documentation should aim to provide a clear decision tree for the user, supported by the underlying formal assessments. It should also indicate the extent to which uncertainty assessments provided by the dataset creators provide reliable and robust estimates of the underlying uncertainties for the quantities required by the user for their particular purpose.

Recommendations

- Assessment criteria should be developed entirely independently of the dataset developers and should be pre-determined and documented in advance of any tests.

- It is crucial that the purpose to which a dataset could be put be identified and that a corresponding set of assessment criteria are derived that are suitable for that purpose.

- The output of an assessment should be to determine whether a dataset is fit for a particular purpose and to enable users to determine which are most suitable datasets for their needs. Outputs should be clearly documented in such a form as to enable a clear decision tree for users.

- Validation of an algorithm should always be carried out on a different dataset from that used to develop and tune the algorithm.

- A key issue is to determine how well uncertainty estimates in datasets represent a measure of the difference between the derived value and the “true” real world value.

- It would be worthwhile to consider the future needs for the development of climate services by indentifying an appropriate set of regions or stations that any assessment should include.

- New efforts resulting from this initiative should be coordinated with on-going regional and national activities to rescue and homogenize data.