Geophysical Research Abstracts
Vol. 15, EGU2013-11157, 2013
EGU General Assembly 2013

Quality control in exascale data archives

Martin Juckes
SSTD, BADC, Chilton, Didcot, Oxon, United Kingdom (martin.juckes@stfc.ac.uk)

Quality control of data can and should occur at many stages of the data life cycle. Data producers will generally conduct their own tests prior to release of data. Further tests may be done by within consortium projects, by archive centres or, after publication of the data, by independent investigators. The results of such tests are often used only by the individual or group carrying out the tests. Sharing of quality control information is restricted by many factors: this presentation addresses the lack of a common terminology to define quality control tests. A framework is proposed, based on abstract tests (e.g. a "prescribed range test"), specific tests (e.g. "variable tas is in the range 200-350K") and test results. Test results also need to be linked to the data robustly to ensure that they are available at all subsequent stages of the data lifecycle, and the appropriate mechanisms of linkage will differ at different stages of the lifecycle. The framework should support references to the data from the test results, references to the results from the data or discovery by association. Performing tests may generate useful by-products (e.g. the maximum and minimum values of the data): these ancillary results should be stored and made accessible with test results. In many cases tests will refer to multiple files, and groups of tests will be of more interest than individual tests. Mechanisms for describing tests suites will be discussed with particular reference to the challenges of applying tests to climate model data.