Data Ranges for Quality Control

Martin Juckes

1. Executive Summary

The CMIP5 CMOR tables included values expressing the expected range of data variables. These were intended for use in quality control, but a significant number of false negatives (data be flagged as erroneous because the limits were set too tightly) limited the overall usefulness. The values used were based on a scan of the values in the CMIP3 archive. A brief survey of the CMIP5 archive described below suggests that:

- There are many variables for which robust limits, often close to those used in CMIP5, can be set;
- There are many variables were the CMIP5 ranges give useful information, but the number of models is too limited, or the disagreement among models too high, to provide robust limits;
- In some cases there may be a robust lower or upper limit (e.g. if the variable must be positive), but only limited information on other limits;
- Caution is needed when the vertical extent of the domain has potential to vary;

To deal with this, it is proposed that the guidance be expanded to include a status flag which will allow quality control software to be selective about limits applied in a systematic way. The status flag will have values “ROBUST”, “SUGGESTED”, “TENTATIVE”, “UNSET”.
The objective here is to set generic limits on parameters which can be used for all simulations of the Earth’s climate and all CMIP style sensitivity studies. The range of acceptable values is intended to accommodate the variation expected between experiments.

2. Introduction

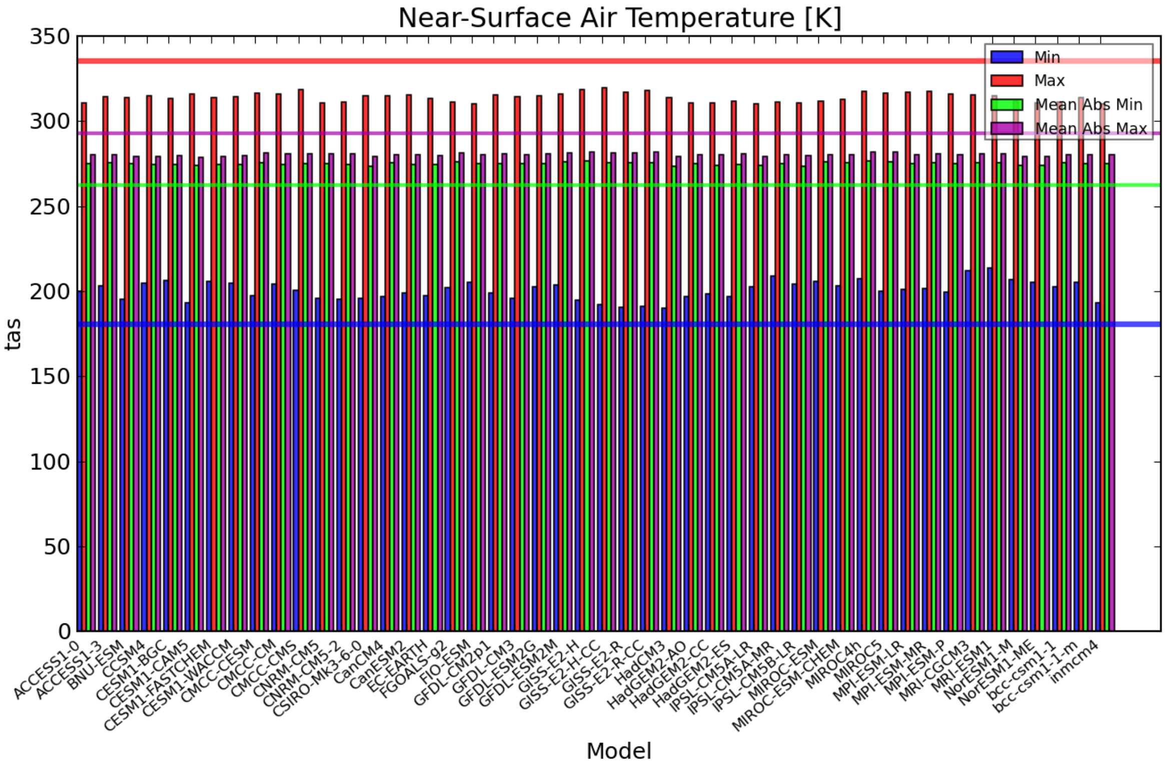
The vast range of data variables in the CMIP archives makes systematic quality control very difficult. The CMIP5 data request included quality control parameters listed below for many variables:

- valid_min: Minimum acceptable data value;
- valid_max: Maximum acceptable data value;
- mean_abs_value_min: Minimum acceptable value of global mean of absolute value;

- mean_abs_value_max: Maximum acceptable value of global mean of absolute value.

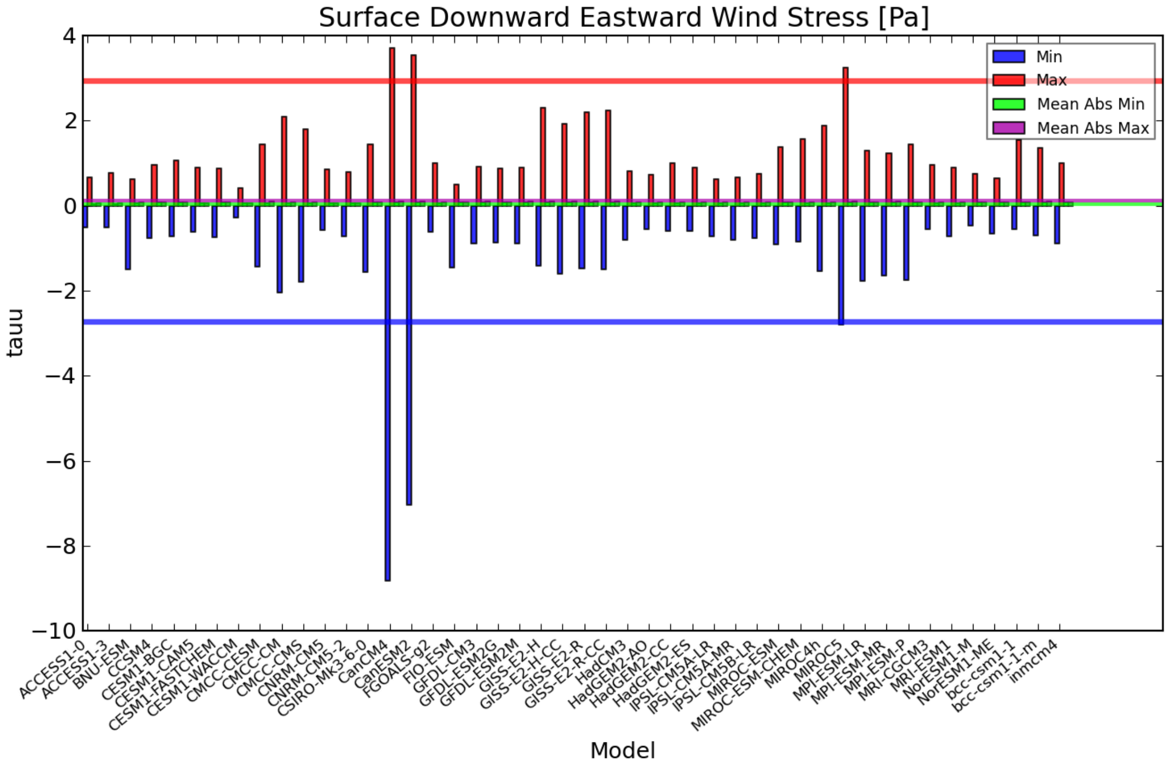
In many cases, such as “Near Surface Air Temperature [tas]” in the “Amon” table, the archived data falls cleanly within these values (see figure below). There are also many examples where the data from one or more models fall outside the limits.

It is clear that temperature in the climate model should approximate the temperature of the actual climate, and it is not surprising that the ranges of plausible values are well understood. In particular, the typical values at the resolved scales of the model are not that different from the typical values which might be measured at a single point in the real world. Many variables represent processes which are represented through parameterization schemes, and there may be differences of many orders of magnitude between a grid scale mean and a local value.



Scaling and offset errors vs. long tail variability

There are many examples in the archive of data which appears to be in error by a fixed scaling factor or offset, possibly associated with data values which are not using the declared units. Examples include temperatures declared as Kelvin but with values expected for temperature in degrees Centigrade, or convective precipitation with mean absolute values 1000 times smaller than expected. The guide values should aid the early detection of such errors. There are also many cases the valid_min/max bounds are exceeded, but 99.9% of values lie within the bounds. In these cases there is either a highly localised error in the data or, possibly, a model representation of processes which exaggerates the long tail of variability in a parameter which has non-Gaussian behaviour. In this case the evaluation of the test may depend on an assessment of the realism of this long-tail behaviour, which should not be the role of the quality control software. The downward eastward wind stress shown below is such an example. The maximum value varies greatly between models, and those that exceed the guide values are not clear outliers. In such cases it is appropriate to expand the the upper and lower limit substantially.



There are also cases where all models fall inside the CMIP5 guide values, but there is still a visible outlier. The figure below shows box plots of the distribution of values of Sea Level Pressure. The 3 IPSL models are consistent with the guide values (this can’t be read directly from this figure) but in comparison with other models they are clear outliers. In such cases it may be worth checking with the modelling groups to see whether it would be useful for them to have such values flagged as errors (IPSL has now resolved the problems associated with this variable).

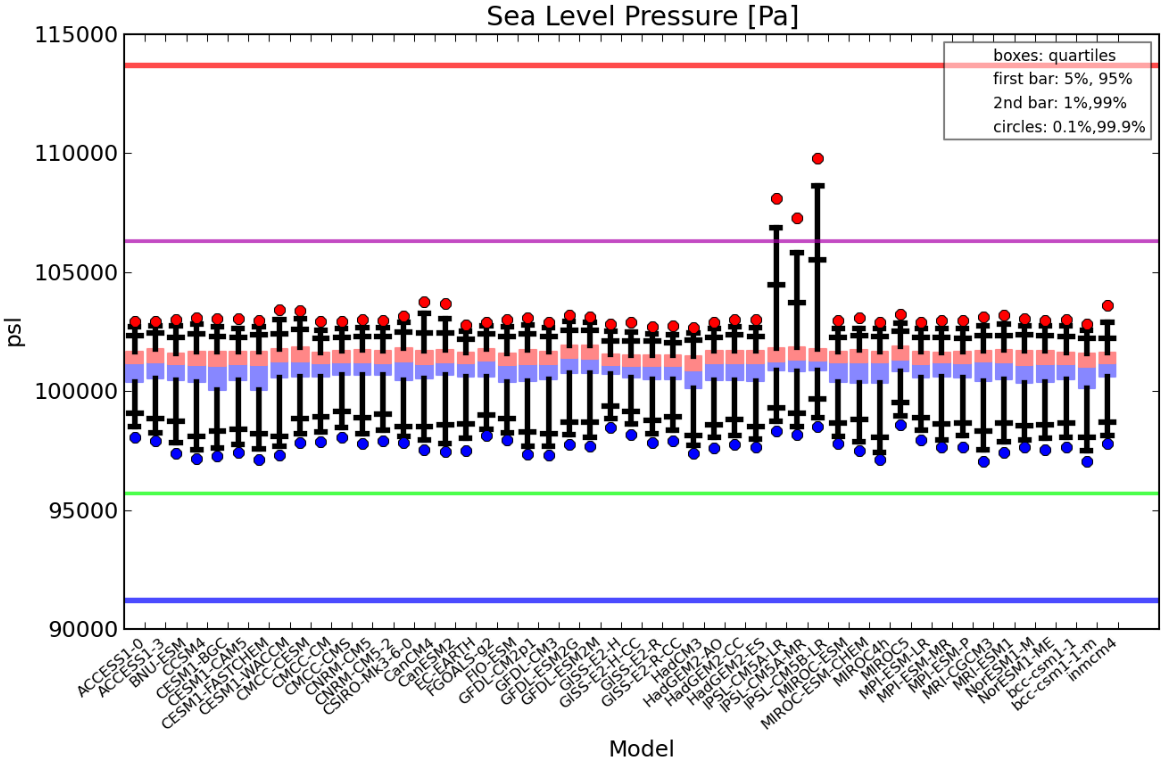


Figure 3: estimated percentiles of sea-level pressure. Percentiles are estimated as the median of the percentiles of each spatial field in a time series. E.g. if there are 120 monthly fields, the percentiles are calculated for each of these fields and then the median of the individual 95th percentiles becomes the estimate of the 95th percentile of the time series.

When an area fraction variables is archived as a percentage (range 0 to 100) instead of a fraction (range 0 to 1) it will be easy to spot, but the converse is harder. For some variables it may be acceptable to have a small or zero values at some time in some experiment. Data incorrectly specified as a fraction might be spotted by looking for constant value areas: there are typically areas that saturate at the upper or lower limit. Regions with value constant at unity would be a clear indication of an error in a area fraction requested as a percentage.

Small Sample Size

With 60 models participating in the CMIP5 exercise, the figures above show data ranges for many models. For some variables, however, the number of models which submitted data is very limited. For example, only 3 models submitted fco2antt (Anthropogenic CO2 Emissions). In such cases the low sample size means that data ranges must be treated with extreme caution.

Varying Vertical Domain

Some diagnostics can have varying domains. For instance, the daily geopotential height data of the MPI-M models extends up to 70,000m2s-2, compared to an upper limit of 35,000m2s-2 for all other models. This discrepancy arises because all other models submitted data on the 8 requested levels, up to 10hPa, while the MPI-M models have additional optional levels up to 10Pa. In such cases it may only make sense to provide data ranges for the requested vertical domain.

Common Errors

There are many datasets with relative humidity greater than 100%. This can occur if the processing of levels which intersect the surface is not done carefully.

3. Approach

A simple approach would be to omit the estimated limits when they are not known with confidence, but this would throw away a lot of information and create an awkward “all or nothing” choice. Instead, a graduated approach is taken with status flags which allow the degree of confidence in limits to be expressed. In some cases there is one limit which is well established and others which are less well quantified, so it is necessary to have a flag for each quality limit.

The 4 quantities used in CMIP5 will be taken into a new “Quality Guidance” record, which will contain additional information to provide a robust basis for quality control.

The quantities “valid_min”, “valid_max”, “ok_max_mean_abs” and “ok_min_mean_abs” have the same meaning as for CMIP5, but extra attributes are added: “valid_min_status”, “valid_max_status”, “ok_max_mean_abs_status” and “ok_min_mean_abs_status”. These attributes will take one of four string values:

- ROBUST: A well characterised limit based on a rigorous constraint (e.g. and area fraction must be between 0 and 1) or on a large ensemble of consistent model results.
- SUGGESTED: A limit which may not be reliable, but which is based on a range of models or plausible arguments.
- TENTATIVE: Very limited information.
- UNSET: No value available.

Values will be set manually .. time consuming but better able to spot unusual behaviour.

4. Options

In order to provide some guidance on the presence of areas of constant data, a “contant_values”/”constant_values_status” pair could be added. The first would be a list of values, e.g. “contant_values=0. 1.”. The 2nd will be a status flag as above, taking values

“ROBUST”, etc. If the status flag is present and the values list is empty or absent, this implies that the data is expected to have no constant values. If both attributes are absent, then no judgement is offered.