Recommendations on data formats and management

Joel R. Mitchelson, John Kennedy, Colin Morice,   
Finn Lindgren, Ag Stephens, Alison Waterfall

*Revisions*

|  |  |  |
| --- | --- | --- |
| **Date** | **Author** | **Comment** |
| 06/10/2015 | Joel Mitchelson | Created. |
| 09/10/2015 | Joel Mitchelson | Shared within Met Office for comment. |
| 23/10/2015 | Joel Mitchelson | Update after feedback from JK, and re-circulated for comment. |
| 27/10/2015 | Joel Mitchelson | Modify station data to use record variables for more efficient daily update, remove use of mean in station data. |
| 28/10/2015 | Joel Mitchelson | Correct cell methods max -> maximum and min -> minimum. |
| 30/10/2015 | Joel Mitchelson | Add reference to CMIP5, change tasmean -> tas for CMIP5 consistency, add version to filenames and use underscore as separator, suggest use of the term ‘sample’ as alternative to ‘ensemble’, add suffix to ensemble variables to allow simultaneous use of ensemble and single ‘best estimate’ output. Correct section cross-references in section 3. |
| 03/11/2015 | Joel Mitchelson | Add station elevation. |
| 04/11/2015 | Joel Mitchelson | Reorganise station status variables and dimensions. Move suggestions about station uncertainty information to separate section. |
| 20/11/2015 | Joel Mitchelson | Correction to syntax of valid\_range in station data specification. |
| 10/12/2015 | Joel Mitchelson | Recommend 16-bit integers following feedback from Met Office colleagues and evidence from users via KNMI. Updated expressions for variable types for clarity.  Added clarification that section 3 refers to WP2 intermediate data in infilling, not WP1 outputs. |
| 07/03/2016 | Joel Mitchelson | Global fields parameterised by time dimension, time dimension set to UNLIMITED, \_ensemble suffix provisionally changed to sample (underscore removed for ESGF compatibility, and name changed to indicate this is not a perturbation of initial conditions). Dimension names changed to equal variable names where appropriate: lat -> latitude, long -> longitude. |
| 12/04/2016 | Joel Mitchelson | time bounds added for cfchecks compatibility, sample suffix changed to ensemble. |
| 11/07/2016 | Joel Mitchelson | Revised file naming convention using CLIP-C recommendations. |
| 26/08/2016 | Joel Mitchelson | Updated data management section to refer to EUMOPPS. |
| 26/08/2016 | Joel Mitchelson | Revise time/date stamps to use ISO convention. |
| 04/10/2016 | Joel Mitchelson | Updated valid\_range, flag\_values, flag\_meanings attributes in station status file. |
| 13/12/2016 | Joel Mitchelson | Modify DRS to allow satstace to have one file per surface type, and for fullstace to have a separate file to contain ensemble information. Correction tassample - > tasensemble. |
| 14/12/2016 | Joel Mitchelson | Express global fields with respect to local solar time, using ancillary variable for time offset. Implemented requests from user feedback: add X, Y, T axis declarations; remove units from long names in global fields. |
| 30/06/2017 | Joel Mitchelson | Change long name from mean to average as appropriate for the surface-air model output in satstace. |

Contents

[1 Global field data product format 2](#_Toc469497278)

[1.1 Container format 3](#_Toc469497279)

[1.2 Resolution 3](#_Toc469497280)

[1.3 Compliance 3](#_Toc469497281)

[1.4 Filenames 3](#_Toc469497282)

[1.5 NetCDF global attributes 4](#_Toc469497283)

[1.6 NetCDF conventional variables 5](#_Toc469497284)

[1.7 NetCDF dimensions and dependent variables 6](#_Toc469497285)

[1.8 Inclusion of uncertainty fields in NetCDF 6](#_Toc469497286)

[1.8.1 Statistical sampling 6](#_Toc469497287)

[1.8.2 Uncertainty parameters 6](#_Toc469497288)

[2 Station data product format 7](#_Toc469497289)

[2.1 CF-compliant global attributes 7](#_Toc469497290)

[2.2 Variables consistent with work package 2 7](#_Toc469497291)

[2.3 Additional variables for station data temperature file 8](#_Toc469497292)

[2.4 Additional variables for station data status file 8](#_Toc469497293)

[2.5 Dimensions and dependencies for station temperature file 9](#_Toc469497294)

[2.6 Uncertainty variables for station temperature file 10](#_Toc469497295)

[2.7 Dimensions and dependencies for station status file 10](#_Toc469497296)

[2.8 Signal uncertainty information in station status file 11](#_Toc469497297)

[3 Internal data format 11](#_Toc469497298)

[4 Data management 12](#_Toc469497299)

[4.1 Project archive 12](#_Toc469497300)

[4.2 Provenance 12](#_Toc469497301)

[4.2.1 The EUMOPPS System 12](#_Toc469497302)

[4.2.2 Recording provenance in NetCDF files 13](#_Toc469497303)

[4.2.3 Correspondence between code and mathematics 13](#_Toc469497304)

[References 13](#_Toc469497305)

# Global field data product format

The recommendations in this section apply to the following planned EUSTACE products (outputs from work package 2):

* Surface air temperature estimates (with estimates of uncertainty) for all surfaces of Earth, derived from satellite surface skin temperature retrievals
* Globally-complete daily analyses of surface air temperature (with estimates of uncertainty) for the whole of Earth since 1850, based on combined information from satellite and in situ data sources

## Container format

NetCDF [1] is proposed as the universal file container format for the major products of EUSTACE work package 2. This self-describing container format is an established medium for communicating scientific data sets, with large existing user base and a large number of freely available tools for using the data [2]. A EUSTACE data product may comprise one or more NetCDF files, with specified filenames. Specified directory hierarchies may also be used, for example to organise files by year.

## Resolution

We are working on the basis that a minimum viable product requires daily records presented on a 0.25 degree resolution (latitude,longitude) grid. The requirement for daily records is specified in the EUSTACE project proposal, and the resolution is chosen to match that of the established HadISST2 product [3], taken as an initial indicator of user expectations. Further multi-resolution intermediate data sets are described in section 3.1.

## Compliance

Since NetCDF is a container format with infinitely flexible internal structure, we also make recommendations regarding choice of that structure. For clarity of communication, and to maximise the number of compatible tools, we recommend that established standards be adopted for the internals of NetCDF files, wherever relevant standards exist. In particular the climate and forecast (CF) metadata conventions [3] should be applied. De-facto conventions potentially expected by users should also be applied, such as the protocol established for short variable names by the Coupled Model Intercomparison project from the World Climate Research Programme [5].

Compliance with naming conventions may be verified by running the cfchecks program against any NetCDF created. The installation of this program on the CEMS system is maintained by STFC.

The project goals require consistency with conventions established on the CLIP-C project. These place constraints on directory layout and filenames. These are defined in CLIP-C project documentation [6] which generalises the data reference syntax (**DRS**) originally introduced by Taylor *et al.* [7].

## Filenames

Using the HadObs recommendation section from [6] we find that directory names for final storage should be of the following form:

*project/product/institute/framework/collection/frequency/table/realization/product\_version/version/*

A strict interpretation of [6] would indicate that filenames specify a start date and an end date. But in our case one NetCDF file per day is used, so as to facilitate future daily update. Hence we propose a simplified filename:

*variable*\_*collection*\_*framework*\_*realization*\_*YYYYmmdd*.nc

The meaning of each predicate is as follows:

|  |  |
| --- | --- |
| **Predicate** | **Meaning and possible values** |
| *project* | eustace |
| *product* | The type of product. No vocabulary for this is available. We recommend this is set to satellitederived for satstace outputs or combined to indicate a variety of sensor types for fullstace outputs. |
| *institute* | The institute producing the data, e.g. mohc for Met Office Hadley Centre. |
| *framework* | eustace |
| *collection* | This may specify a convenient subset of the data such as one surface type, or otherwise a generic name for the result such as global or ensemble. |
| *frequency* | This must be set to day as EUSTACE is daily data. Although this option is not given in [6] there is an established convention in [7]. |
| *table* | Unused hence set to 0. |
| *realization* | This is unused because multiple realisations (if required) are not stored as separate files but rather inside a single file, hence it should always be 0. |
| *product\_version* | Native version taken from EUSTACE provenance system (see section 4.2 below). |
| *version* | Date of launch of processing. |
| *variable* | Must be set to tas indicating surface air temperature. |
| *YYYYmmdd* | 4-digit year *YYYY*, 2-digit month *mm*, 2-digit day *dd*. This is the date corresponding to local solar time at 0E in the contained data. |

Hence an example of a valid file path is:

eustace/combined/mohc/eustace/fullstace/day/0/0/V000339/20160711/  
tas\_fullstace\_eustace\_0\_18530207.nc

## NetCDF global attributes

The following global attributes specified by CF conventions should exist:

title: EUSTACE Surface Air Temperature Estimates (*collection)*

institution: (*institution)*

source: EUSTACE Catalogue (*identifier)*

history: (*timestamp*) (*module*)

Conventions: CF-1.6

See section 4.2 below for more information on the source and history attributes.

## NetCDF conventional variables

The conventions for variable names within NetCDF files are listed below. CF conventions are used wherever applicable, as well as CMIP5 conventions for variable names.

According to technical requirement PRDREQ015 introduced on 11/04/2016 the observations provided will be with respect to local solar time. UTC values specified by the time variable are therefore only correct at zero longitude and this is indicated in the long name. Temporal offsets for each longitude are provided via an ancillary variable. An alternative approach using a longitude-dependent time variable was tested with users but found to cause problems in some software packages.

32-bit floating point representation may be used for all data. However, the use of 16-bit integers for global data fields would significantly reduce file size (as compared with 32-bit), thereby reducing download time for users. Evidence from users suggests that faster downloads of data are preferred, and so 16-bit is recommended here.

**Variable: time (float32)**

standard\_name: time

long\_name: Time at zero longitude

units: days since 1850-01-01T00:00:00Z

calendar: gregorian

axis: T

bounds: timebounds

ancillary\_variables: timeoffset

**Variable: timebounds (float32)**

**Variable: timeoffset (float32)**

long\_name: Local time offset from UTC (days)

units: days

**Variable: latitude (float32)**

standard\_name: latitude

long\_name: Latitude (deg)

units: degrees\_north

axis: Y

**Variable: longitude (float32)**

standard\_name: longitude

long\_name: Longitude (deg)

units: degrees\_east

axis: X

**Variable: tas (int16)**

standard\_name: air\_temperature

long\_name: Average daily surface air temperature

units: K

cell\_methods: time: mean

**Variable: tasmin (int16)**

standard\_name: air\_temperature

long\_name: Minimum daily surface air temperature

units: K

cell\_methods: time: minimum

**Variable: tasmax (int16)**

standard\_name: air\_temperature

long\_name: Maximum daily surface air temperature

units: K

cell\_methods: time: maximum

## NetCDF dimensions and dependent variables

The following layout of dimensions is recommended per daily file at 0.25 degree resolution:

Dimensions

latitude: 720

longitude: 1440

time: **UNLIMITED**

bounds: 2

Dependencies

time(time)

timebounds(time, bounds)

timeoffset(time, longitude)

latitude(latitude)

longitude(longitude)

tas(time, latitude, longitude)

tasmin(time, latitude, longitude)

tasmax(time, latitude, longitude)

## Inclusion of uncertainty fields in NetCDF

Selection of feasible and useable methods for expression of uncertainty in EUSTACE output fields is a topic of ongoing research. It is recommended that any such expression be added to the NetCDF format using one of the following approaches, or a combination of them.

### Statistical sampling

Uncertainty may be expressed using an ensemble of samples taken from a statistical model. In this case the ensemble dimension can be added, and the ensemble suffix given to variables, so ensemble temperature variables and dependencies are expressed as:

tasensemble(time, ensemble, latitude, longitude)

tasminensemble(time, ensemble, latitude, longitude)

tasmaxensemble(time, ensemble, latitude, longitude)

### Uncertainty parameters

One or more parameters may be used to describe uncertainty on a per-location-per-time basis. An additional variable must therefore be added to represent each one, of the form:

*uncertaintyparameter*(time, latitude, longitude)

If the parameter represents the standard deviation of an uncertainty, the NetCDF attributes associated with each component of the uncertainty should describe what it is and how the errors that it relates to are correlated in space and time, if this is known and easily expressible. *e.g.* for a standard deviation the attributes should indicate locally-correlated with time scale 1-day and spatial scale of 100km.

For parameters describing summary statistics, NetCDF attributes of the variable should be used to represent the nature of the statistic e.g. mean, standard deviation, skew, percentile.

# Station data product format

The recommendations in this section are made in relation to the following planned EUSTACE products which are outputs from work package 1:

* Homogenised meteorological station records of surface air temperature measurements for Europe
* Global data set of surface air temperature measurements from meteorological stations with discontinuities in each station record identified, where possible

We suggest the data set comprises two files:

eustace\_stations\_*region*\_*version*\_temperature.nc

eustace\_stations\_*region*\_*version*\_status.nc

The value of *region* is either europe for the European data set or global for the global one.

A temperature file contains daily temperature time series from all stations. The corresponding status file contains signal status information. This includes break detections, which may depend on time but which occur at irregular intervals; and break detection performance information, which updates at regular time intervals (e.g. every year) but not daily.

The reason for suggesting two separate files is to permit efficient daily update.

Given that only two files are involved, it may be convenient to place *version* identifiers into the filenames as described; though it is also proposed that comprehensive versioning be carried out as described in section 4.2.

## CF-compliant global attributes

Global attributes should be as for work package 2 products, plus inclusion of:

featureType: timeSeries

## Variables consistent with work package 2

The temperature file should contain the following variables with attributes as described in section 1.6 for work package 2:

time, latitude, longitude, tasmin, tasmax

## Additional variables for station data temperature file

In addition to variables already mentioned, the following variables are recommended for expression of station temperatures. All are compatible with CF.

**Variable: station\_name (char)**

long\_name: Station name

cf\_role: timeseries\_id

**Variable: elevation (float32)**

long\_name: Height above the geoid (m)

standard\_name: surface\_altitude

units: m

**Variable: tasmin\_qc (uint8)**

long\_name: Quality control flags

standard\_name: air\_temperature status\_flag

valid\_range: 0 , *maxrange*

flag\_masks: 1, 2, 4, ...

flag\_meanings: *bit0meaning* *bit1meaning* *bit2meaning* ...

**Variable: tasmax\_qc (uint8)**

long\_name: Quality control flags

standard\_name: air\_temperature status\_flag

valid\_range: 0b , *maxrange*

flag\_masks: 1, 2, 4, ...

flag\_meanings: *bit0meaning* *bit1meaning* *bit2meaning* ...

## Additional variables for station data status file

In addition to variables already mentioned, the following variables are suggested ways to express station status, based upon the information made available from EUSTACE work package 1. All are compatible with CF, though not all are CF conventions.

**Variable: detection\_time (float32)**

standard\_name: time

long\_name: Start time of period for break detection status report (days)

units: days since 1850-01-01T00:00:00Z

**Variable: tasmin\_detection\_qc (uint8)**

long\_name: Quality control flags for break detection

standard\_name: air\_temperature status\_flag

valid\_range: [0, 2]

flag\_values: [0, 1, 2]

flag\_meanings: not\_possible possible\_but\_unreliable reliable

**Variable: tasmax\_detection\_qc (uint8)**

long\_name: Quality control flags for break detection

standard\_name: air\_temperature status\_flag

valid\_range: [0, 2]

flag\_values: [0, 1, 2]

flag\_meanings: not\_possible possible\_but\_unreliable reliable

**Variable: tasmin\_break\_station (int32)**

long\_name: Index of station at which break was detected

**Variable: tasmax\_break\_station (int32)**

long\_name: Index of station at which break was detected

**Variable: tasmin\_break\_amplitude (float32)**

long\_name: Inhomogeneity in minimum surface air temperature

units: K

**Variable: tasmax\_break\_amplitude (float32)**

long\_name: Inhomogeneity in maximum surface air temperature

units: K

**Variable: tasmin\_break\_time\_bounds (float32)**

long\_name: Relative time bounds of break evaluation period (days)

units: days

**Variable: tasmax\_break\_time\_bounds (float32)**

long\_name: Relative time bounds of break evaluation period (days)

units: days

**Variable: tasmin\_break\_time\_affected\_bounds (float32)**

long\_name: Relative time bounds of period affected by break (days)

units: days since 1850-01-01T00:00:00Z

**Variable: tasmax\_break\_time\_affected\_bounds (float32)**

long\_name: Relative time bounds of period affected by break (days)

units: days since 1850-01-01T00:00:00Z

## Dimensions and dependencies for station temperature file

Dimensions

station: *number of stations*

name\_strlen: *maximum length of any station name*

time: **UNLIMITED**

Dependencies

time(time)

station\_name(station, name\_strlen)

latitude(station)

longitude(station)

elevation(station)

tasmin(time, station)

tasmin\_qc(time, station)

tasmax(time, station)

tasmax\_qc(time, station)

## Uncertainty variables for station temperature file

If additional uncertainty information is available, variables may be added to the station temperature file, with dimensions of the form:

*uncertaintyparameter*(time, station)

If an ensemble is available, an ensemble\_member dimension is added, and the temperature variable dependencies become:

tasmin(time, ensemble\_member, station)

tasmin\_qc(time, ensemble\_member, station)

tasmax(time, ensemble\_member, station)

tasmax\_qc(time, ensemble\_member, station)

## Dimensions and dependencies for station status file

Time dimensions, with suffix \_time corresponding to equal time increments are used to index those variables which occur at regular intervals, whereas ‘break’ dimensions with suffix \_break are used as integer indices for irregularly occurring break events.

Dimensions

station: *number of stations*

detection\_time: **UNLIMITED**

tasmin\_break: **UNLIMITED**

tasmax\_break: **UNLIMITED**

bounds:2

Dependencies

detection\_time(detection\_time)

tasmin\_detection\_qc(detection\_time, station)

tasmax\_detection\_qc(detection\_time, station)

tasmin\_break\_time(break\_tasmin)

tasmin\_break\_station(tasmin\_break)

tasmin\_break\_time\_bounds(tasmin\_break, bounds)

tasmin\_break\_time\_affected\_bounds(tasmin\_break, bounds)

tasmin\_break\_amplitude(tasmin\_break)

tasmax\_break\_time(tasmax\_break)

tasmax\_break\_station(tasmin\_break)

tasmax\_break\_time\_bounds(tasmin\_break, bounds)

tasmax\_break\_time\_affected\_bounds(tasmin\_break, bounds)

tasmax\_break\_amplitude(tasmin\_break)

## Signal uncertainty information in station status file

The station status file must additionally provide information regarding uncertainty of measurements. The nature of this is still under discussion. It may take the form of regular updates along a time dimension. In this case would introduce one or more trios of variables of the form:

**Variable: uncertainty\_time (float32)**

standard\_name: time

long\_name: Start time of period for uncertainty evaluation (days)

units: days since 1850-01-01T00:00:00Z

**Variable: tasmin\_uncertainty\_maxmimum (float32)**

standard\_name: air\_temperature

long\_name: Maximum uncertainty in minimum daily surface air temperature (K)

units: K

cell\_methods: time: maximum

**Variable: tasmax\_uncertainty\_maxmimum (float32)**

standard\_name: air\_temperature

long\_name: Maximum uncertainty in maximum daily surface air temperature (K)

units: K

cell\_methods: time: maximum

And new time dimension:

uncertainty\_time: **UNLIMITED**

With dimensional dependencies for variables:

uncertainty\_time(uncertainty\_time)

tasmin\_uncertainty\_maximum(uncertainty\_time, station)

tasmax\_uncertainty\_maximum(uncertainty\_time, station)

# Internal data format

This section describes data formats used internally within work package 2. These are internal parts of the infilling system and ***not*** directly intended for any of the outputs produced by algorithms developed in work package 1. The different infilling analyses performed by Bath and Met Office each require input data to be placed on multi-resolution pseudo-equal-area triangular grids (hereafter referred to more simply as **triangular grids**). At present it is unknown whether the same layouts of triangular grids are applicable to both methods. Where possible it would be desirable for the same triangular grids to be used, for economy of processing and storage time. This is a matter for ongoing discussion between work package 2 project partners.

The above NetCDF conventions are recommended where applicable. However for efficient storage it is also recommended that the (longitude, latitude) coordinates of points on triangular grids are not stored inside the files containing processed input data. Instead they should be referred to by integer index, with the coordinates corresponding to each integer index given in a separate file. Provenance tracking may be used as to provide quality assurance amongst grid versions as described in section 4.2.

Formats for statistical model data files produced by Met Office and Bath will be determined as research in these areas progresses. For these it is also recommended to use provenance tracking as described in section 4.2.

# Data management

## Project archive

In addition to output products, archiving by STFC of the following is necessary for full traceability of EUSTACE data sets, and for demonstration that a viable end-to-end system has been created:

* Multi-resolution triangular meshes of input data as required by Met Office and Bath for input to infilling analysis
* Statistical models with all necessary parameters for reproduction of outputs

Without these it would be almost impossible for an external scientist to verify independently the content of the EUSTACE products; nor to see clear evidence that an end-to-end system has truly been demonstrated.

## Provenance

We have a goal that any data set (and so by implication any intermediate data set) can be traced back to the data sets and code that were used in its creation, and that code can be traced to the corresponding mathematics submitted for peer review. This is to allow later independent verification of EUSTACE results, and is also necessary during the lifetime of the project to know when data is out of date relative to the latest code version or mathematics. We also seek to facilitate simple tests of data integrity.

### The EUMOPPS System

To facilitate simultaneous versioning of code and data as well as data integrity tests, Met Office will provide the EUSTACE Met Office Processing and Provenance System (EUMOPPS). This allows automatic generation and update of a **catalogue file** to consistently and reliably record the following: input pathnames and checksums, operations carried out on the data, output product pathnames, processing environment, and code revision used. Each catalogue file has a globally unique identifier, which unambiguously represents this complete provenance chain.

### Recording provenance in NetCDF files

The *identifier* string in the specification of the source global attribute (section 1.5) for product files should be a string which refers to the unique identifier of the corresponding catalogue file. This is similar to the Federation of Earth Science Partners (ESIP) attribute convention for data discovery, which suggests that audit trails written into NetCDF attributes may contain a unique identifier of a processing scheme [8] documented elsewhere. The processing tool which created the output file is recorded in the history attribute along with the time of processing.

### Correspondence between code and mathematics

This is by expert inspection. Written reports containing mathematics should refer to a specific code repository and revision number. Correspondence is subject to internal code review with a step involving cross-checking mathematics against the specified code version. To permit independent verification it is recommended that EUSTACE code be made open source at the end of the project, and mathematics published, allowing it to be cross-checked.

# References

|  |  |
| --- | --- |
| [1] | University Corporation for Atmospheric Research, “NetCDF Documentation,” [Online]. Available: http://www.unidata.ucar.edu/software/netcdf/docs/index.html. [Accessed 06 10 2015]. |
| [2] | University Corporation for Atmospheric Research, “Software for Manipulating or Displaying NetCDF Data,” [Online]. Available: http://www.unidata.ucar.edu/software/netcdf/software.html. [Accessed 7 10 2015]. |
| [3] | “Hadley Centre Sea Ice and Sea Surface Temperature data set (HadISST) version 2,” Met Office Hadley Centre, [Online]. Available: http://www.metoffice.gov.uk/hadobs/hadisst2/. [Accessed 23 10 2015]. |
| [4] | B. Eaton, J. Gregory, B. Drach, K. Taylor and S. Hankin, “NetCDF Climate and Forecast (CF) Metadata Conventions: Version 1.6,” 2011. [Online]. Available: http://cfconventions.org/Data/cf-conventions/cf-conventions-1.6/build/cf-conventions.pdf. [Accessed 6 10 2015]. |
| [5] | Lawrence Livermore National Laboratory, “CMIP5 - Modeling Info - Producing Model Output,” [Online]. Available: http://cmip-pcmdi.llnl.gov/cmip5/output\_req.html. [Accessed 2015 10 30]. |
| [6] | R. Petrie, V. Bennet and M. Juckes, “Metadata and controlled vocabularies specification for data, quality control and uncertainties,” CLIP-C Deliverable 52, European Commission, 2016. |
| [7] | K. E. Taylor, V. Balaji, S. Hankin, M. Juckes, B. Lawrence and S. Pascoe, “CMIP5 Data Reference Syntax (DRS) and Controlled Vocabularies,” Science and Technology Facilities Council, UK, 2012. |
| [8] | Federation of Earth Science Partners (ESIP), “Attribute Convention for Data Discovery,” [Online]. Available: http://wiki.esipfed.org/index.php/Attribute\_Convention\_for\_Data\_Discovery. [Accessed 26 08 2016]. |