EARTH OBSERVATION CLIMATE INFORMATION SERVICE

CHUK Data Standards, v1.1

Date: 18/12/2024

History of modifications / Change Log

Version	Date	Changes	Person
1.0	01/10/2024	1st release	A. Waterfall
1.1	18/12/2024	Added recommendation on chunk size	A. Waterfall

Related Documents / Reference Documents

Document	Author	Reference

Acronyms and/or Abbreviations

Acronym / Abbreviation	Definition
BNG	British National Grid
CF	NetCDF Climate and Forecast Metadata Conventions
CHUK	Climate High resolution United Kingdom
CCI	Climate Change Initiative
CRS	Coordinate Reference System
ECV	Essential Cliamte Variable
EOCIS	Earth Observation Climate Information Service
ESA	European Space Agency
NetCDF4	
STAC	Spatial Temporal Asset Catalogue
WGS84	
WKT	Well-Known Text

General definitions

Term	Definition

Table of Contents

1. Introduction

As part of the UK Earth Observation Climate Information Service (EOCIS) project, new high resolution UK specific climate data products are being produced (CHUK data products). This includes both rapid-response information for climate-linked events (fire early warning and urban flood mapping) and longer term climate data linked to human and ecosystem health and landscape greenhouse gas emissions.

This document describes the data standards that should be met by the CHUK data products produced under EOCIS. A separate document exists for the separate EOCIS Global and Regional data products. Alongside these data format guidelines, there is a specified CHUK 100m data grid that should be used with the data products. The latest version of the CHUK 100m data grid is currently 1.0.

2. Overview of CHUK format

CHUK data should be stored using NetCDF4.

The format for the dataset should contain at least the following data related elements:

i) Coordinate Reference System (CRS) as a Well-Known Text (WKT) string or a PROJ4 string (Section 3.2). The projection for the grid shall be that of the British National Grid (BNG).

ii) Bounding box in BNG and optional in Latitude and Longitude (WGS84) coordinates.

iii) When the data relates to a specific time or time periods, the time dimension should be present in every single file.

iv) Spatial resolution in BNG units (m).

The metadata (embedded or as an auxiliary file) should follow the standard selected by the data producer e.g. the ESA Climate Change Initiative (CCI) metadata standards [1]. The metadata must include:

v) Traceability of files to their origin, and to lower-level input files used.

vi) Propagation of licence conditions and citation requirements to users.

vii) Any additional important considerations for resilience and quality

3. File Format Specification

3.1 NetCDF

CHUK data should be stored using netCDF4.

In general, metadata should be compliant with Version 1.10 of the NetCDF Climate and Forecast (CF) Metadata Conventions [2] (https://cfconventions.org/) with the exception of some flag information (see section 3.3.1) . The online tool cfchecker may be helpful for checking compliance with the CF conventions.

Datasets should only use netCDF features which are compatible with the current version of the CF convention. The use of netCDF-4 groups and the new netCDF-4 data types, such as int64 or unsigned bytes, are not recommended as these can cause issues with some older tools, but could be used if the additional functionality they provide overweighs these issues.

Please use chunking in netCDF files (using a chunk size of 1000 in x and y dimensions) and compression level 5.

3.2 Coordinates

Data should be located on the British National Grid (BNG) Coordinate Reference System (EPSG:27700) at 100m resolution. Exceptions to this should be agreed with the EOCIS project team.

The grid is defined using the following file, which is stored according to the CHUK data format:
https://gws-access.jasmin.ac.uk/public/nceo_uor/eocis-chuk/EOCIS-CHUK-GRID-100M-v1.0.nc

- Data should be organised by dimensions (time, y, x). For data where there is no time dimension only, data may be organised by dimensions (y,x). Single timestep files should still include a time dimension.

- It is recommended to include a time_bnds boundary variable as per the CF-convention, to express the time range covered by each data point. (Note, time variables and time_bnds should **not** use an int64 data type, as this can cause issues with older tools.)

- Data should contain coordinates for the y and x dimensions as northings and eastings on the BNG that match those in the file above. Data should contain a `crsOSGB` variable as included in the file above. These correspond to the centre point of the pixel.

- Data may optionally contain lat, lat_bnds and lon, lon_bnds coordinates (and similarly x and y bounds) and if included, these should match the values in the file above. A 32-bit floating

point representation is sufficient to ensure sub-1m. If given, the lat/lon coordinates should correspond to the centre of the pixel.

3.3 Variables

Variables should be stored in the netCDF file in conformance with the CF convention (http://cfconventions.org) and using the standard name attribute where appropriate. Where an existing standard name exists within the CF convention, this should be included as an attribute on the variable. Where a standard name does not exist, or has not yet been accepted by CF, the standard name attribute should not be included for that variable. Data producers are strongly encouraged to submit new standard names for inclusion in CF as appropriate. Names can be proposed and discussed directly here: https://github.com/cf-convention/discuss. The CEDA team can also provide advice on the process. In addition, the following should be considered:

- Ancillary and coordinate variables should also be identified as specified in the CF conventions:
  - Ancillary variables such as uncertainty or quality flags should be identified by the 'ancillary_variables' attribute of the related primary variable. Ancillary variables that are required to correctly interpret a key primary variable must be identified in this way.
  - Coordinate variables are identified as those with the same name as the corresponding dimensions, whilst related auxiliary coordinate variables can be identified using the 'coordinates' attribute of a variable.
  - For each variable, it is recommended that the attributes 'valid_range' (or alternatively 'valid_min', 'valid_max') and 'actual_range' are provided, if appropriate.
  - The 'valid_range' should define the expected valid range of the variable. Under CF conventions, values outside this range are treated as missing data.
  - The 'actual_range' provides the actual range of the data within the file, within the limits of the valid range.

3.3.1 Flag Variables

Flag variables contain auxiliary information relating to one of the key variables in the file and are typically used to represent data quality or an indicative quantity (e.g. a land/sea mask). Flag variables can be either exclusive (only one condition can be set at a time) or inclusive (zero or more conditions can be set). Exclusive flag variables can be represented by a simple enumerated series (e.g. 0, 1, 2, 3, 4, 5) and defined using the 'flag_values' variable attribute. Inclusive flag variables must be represented using bitwise notation (e.g. 1, 2, 4, 8, 16, 32) and defined using the 'flag_masks' variable attribute. Both 'flag_values' and 'flag_masks' values must be defined alongside the 'flag_meanings' attribute, which consists of a space separated list of the corresponding flag

conditions. The '_FillValue' attribute can be optionally set to define missing/invalid data, if not already included as one of the flag conditions.

Where a flag variable represents data quality, it is recommended that a value of '0' (the number zero or no bits set) is taken to indicate good quality data.

3.4 Global Attributes

We recommend that the following attributes are included in the metadata for consistency with the global and regional data products. These were originally based on the CCI data format [1], which in term are based on those from the Attribute Convention for Data Discovery (ACDD) [3].

Global Attribute	Description
title	Succinct description of the dataset
institution	Where the data was produced.
source	Original data source(s), e.g. MERIS RR L1B version 4.02) Multiple source datasets and ancillary datasets used, with their DOI if available, as a free-text, comma-separated list
history	Processing history of dataset
references	References to algorithm, ATBD, technical note describing dataset
tracking_id	A UUID (Universal Unique Identifier) value
Conventions	Any conventions that have been followed e.g. CF-1.10, ...
product_version	The product version of this data file
format_version	The version of the data format used e.g. "EOCIS CHUK Data Standards vx.x"
summary	A paragraph describing the dataset
keywords	A comma separated list of key words and phrases
id	
naming authority	The combination of the naming authority and the id should be a globally unique identifier for the dataset

keywords_vocabulary	If you are following a guideline for the words/phrases in your "keywords" attribute, put the name of that guideline here
comment	Miscellaneous information about the data
date_created	The date on which the data was created
creator_name	The person/organisation that created the data
creator_url	A URL for the person/organisation that created the data
creator_email	Contact email address for the person/organisation that created the data
project	The scientific project that produced the data: "UK Earth Observation Climate Information Service (EOCIS)"
geospatial_lat_min	Decimal degrees north, range -90 to +90
geospatial_lat_max	Decimal degrees north, range -90 to +90
geospatial_lon_min	Decimal degrees east, range -180 to +180
geospatial_lon_max	Decimal degrees east, range -180 to +180
geospatial_vertical_min	Assumed to be in meters above ground unless geospatial_vertical_units attribute defined otherwise
geospatial_vertical_max	Assumed to be in metres above ground unless geospatial_vertical_units attribute defined otherwise
time_coverage_start	Format yyyymmddThhmmssZ
time_coverage_end	Format yyyymmddThhmmssZ
time_coverage_duration	Should be an ISO8601 duration string
time_coverage_resolution	Should be an ISO8601 duration string. For L2 data on the original satellite sampling it is acceptable to use 'satellite_orbit_frequency'
standard_name_vocabulary	The name of the controlled vocabulary from which variable standard names are taken e.g. 'CF Standard Name Table v82'
license	Describe the data licence / terms and conditions of access.For example: :license = "Creative Commons Licence by attribution (https://creativecommons.org/licenses/by/4.0/)";

platform	Satellite name e.g. Sentinel-5. Separate lists by commas and use angled brackets for a platform series, e.g. 'Envisat, NOAA-<12,14,16,17,18>, Metop-A'.
sensor	Sensor name e.g. AATSR. Separate lists by commas.
spatial_resolution	A free-text string describing the approximate resolution of the product. For example, "1.1km at nadir". This is intended to provide a useful indication to the user, so if more than one resolution is relevant e.g. the grid resolution and the data resolution, then both can be included.
geospatial_lat_units	Geospatial latitude units used
geospatial_lon_units	Geospatial longitude units used
geospatial_lon_resolution	Geospatial latitude resolution used
geospatial_lat_resolution	Geospatial longitude resolution used
key_variables	A comma separated list of the key primary variables in the file i.e. those that have been scientifically validated.
Acknowledgement	This should acknowledge the funders of the data e.g. This work was supported by the Natural Environment Research Council (NERC grant reference number NE/X019071/1, "UK EO Climate Information Service").
program	EOCIS
program_url	https://eocis.org
program_email	EOCIS@reading.ac.uk

4. Filenaming

For consistency with the EOCIS Global and Regional data, the following filenaming convention is recommended, based on one form of the CCI filenaming convention.

4.1 Filenaming convention

The following form of the filenaming is recommended:

EOCIS-<EOCIS Project>-<Processing Level>-<Product Type>-<Product String>[-<Additional Segregator>]-[<IndicativeDate>[<Indicative Time>]]-fv<File version>.nc

The fields in the filename convention are:

<Indicative Date> <Indicative Time>	The identifying date for this data set. Format is YYYY[MM[DD]], where YYYY is the four-digit year, MM is the two-digit month from 01 to 12 and DD is the two-digit day of the month from 01 to 31. The date used should best represent the observation date for the data set. It can be a year, a year and a month or a year and a month and a day. <Indicative Time> The identifying time for this data set in UTC. For data files that include multiple timesteps spanning a significant period of time, this could be expressed as a date range separated by an underscore e.g. <Indicative Start Date><Indicative Start Time>_<Indicative End Date><Indicative End Time>
EOCIS	This is the name of the project producing the data, which in most cases would be EOCIS. Note, this corresponds to RDAC (Regional Data Assembly Centre) in the GHRSST file naming convention.
<Processing Level>	See *Table 1* in the Appendix
<EOCIS project>	For CHUK data this should be 'CHUK_<ECV project>' e.g. CHUK_SST .
<Product Type>	This corresponds to the CCI Product Type and the SST Type in the GHRSST convention and

	shall contain a brief term to describe the main product type in the data set. This should be consistent across data of the same product type. Where new or different product types are produced by EOCIS, the new term shall be added to the list by informing the CEDA team. Example product types as used in the ESA CCI project are given in Table in the Appendix
<Product String>	Each team shall define the product strings that they will use for their data. The Product String field must not include any hyphens but can include underscores.
<Additional Segregator>	This is an optional part of the filename. It must be used if otherwise different data sets would generate the same filename. It can also be used to include in the filename information that doesn't fit elsewhere in the filename convention, but which projects feel is useful for easy identification of different data sets. Each EOCIS team shall define the Additional Segregators they will use for their data. More than one element may be included, separated by an underscore, not a hyphen. Only one additional segregator should be used.
fv<File Version>	File version number in the form n{1,}[.n{1,}] (i.e. one or more digits followed by an optional '.' and another one or more digits).

4.2 Version numbering

The version numbering of any given data product should follow these principles:

- Version numbering should always increase
- The file version used in the filename should uniquely identify the particular instance of the dataset and should always increase with subsequent product versions (if these are not identical).

5. Worked examples with data

Links to examples will be added when available.

6. Supporting Open-Source Library/API: chuk-api

A public GitHub repo will be available (https://github.com/eocis-chuk/chuk-api) to offer the following services to help with the creation and use of CHUK data:

- Load a CHUK format dataset augumented with lat/lon/lat_bnds/lon_bnds
- Check that a file adheres to the CHUK data format (requirements additional to CF compliance)
- Export CHUK data/metadata to other data formats (for example, Geotiff)
- Export CHUK metadata to other metadata formats (possibly, STAC)
- Evolution of the CHUK data format

7. References

[1] ESA Climate Office, 2021, 'CCI data standards v2.3', CCI-PRGM-EOPS-TN-13-0009, available from https://climate.esa.int/media/documents/CCI_DataStandards_v2-3.pdf

[2] B. Eaton et al, 'NetCDF Climate and Forecast (CF) Metadata Conventions, Version 1.10, 31 August 2022, http://cfconventions.org/Data/cf-conventions/cf-conventions-1.10/cf-conventions.pdf

[3] Attribute Convention for Data Discovery, https://wiki.esipfed.org/Attribute_Convention_for_Data_Discovery, accessed 19/10/2023

Appendix

Table 1: Processing level of the data as specified in the ESA CCI Data Standards[1].

Level	<Processing Level> Code	Description	Based on Source
Level 0	L0	Unprocessed instrument and payload data at full resolution. EOCIS does not make recommendations regarding formats or content for data at this processing level.	GHRSST
Level 1A	L1A	Reconstructed unprocessed instrument data at full resolution, time referenced, and annotated with ancillary information, including radiometric and geometric calibration coefficients and georeferencing parameters, computed and appended, but not applied, to L0 data.	GHRSST
Level 1B	L1B	Level 1A data that have been processed to sensor units.	GHRSST
Level 1C	L1C	Level 1b data that have been further processed, e.g. by correcting radiances or by mapping onto a spatial grid, prior to deriving geophysical variables from the data.	SMOS data products definition and ESACCI discussions
Level 2	L2	Retrieved environmental variables at the same resolution and location as the level 1 source	CEOS interoperability handbook
Level 2 Pre-processed	L2P	Geophysical variables derived from Level 1 source data at the same resolution and location as the level 1 data, typically in a satellite projection with geographic information. These data form the fundamental basis for higher level EOCIS products.	GHRSST
Level 3	L3	Level 2 variables mapped on a defined grid with reduced requirements for ancillary data. Three types of L3 products are defined:	GHRSST
	L3U	Uncollated (L3U): L2 data granules remapped to a space grid without combining any observations from overlapping orbits.	
	L3C	Collated (L3C): Observations combined from a single instrument into a space-time grid.	

	L3S	Super-collated (L3S): observations combined from multiple instruments into a space-time grid.	
Level 4	L4	Data sets created from the analysis of lower level data that result in gridded, gap-free products.	GHRSST
Indicator	IND	Indicators derived from satellite data.	ESACCI

Table 2: Example ECV projects and product types from the ESA CCI data standards. Further terms can be added for EOCIS as required. For CHUK data, the EOCIS project name should be 'CHUK_<ECV project>

ECV Project	**Parameter**	**<Product Type>**
Aerosol	Aerosol optical depth	AOD
	Absorbing aerosol index	AAI
	Stratospheric aerosol extinction profile	AEX
	Aerosol type	ATY
	Multiple aerosol products	AER_PRODUCTS
Biomass	Above-ground biomass	AGB
Cloud	Cloud cover	CFC
	Cloud top pressure	CTP
	Cloud top height	CTH
	Cloud top temperature	CTT
	Cloud optical thickness	COT
	Cloud effective radius	CER
	Cloud liquid water path	LWP
	Cloud ice water path	IWP
	Joint cloud physical properties	JCH
	Multiple cloud products	CLD_PRODUCTS
Fire	Burned area	BA
GHG	column-averaged dry air mole fraction of CO2	CO2
	column-averaged dry air mole fraction of CH4	CH4
Glaciers	Glacier area	GA
Ice Sheets	Ice sheet surface elevation change	SEC

	Ice sheet velocity	IV
	Glacier calving front location	CFL
	Glacier grounding line location	GLL
Lakes	Multiple lake products	LK_PRODUCTS
Land Cover	Land cover map	Map
	Condition fire (burned area)	BA
	Condition water (water bodies)	WB
	Condition snow	Snow
	Condition normalised difference vegetation index	NDVI
	Condition albedo	Alb
	Leaf Area Index	LAI
	Surface reflectance	SR
Ocean Colour	Multiple products (chla, nlw, IOPs, etc)	OC_PRODUCTS
	Phytoplankton Chlorophyll-a concentration	CHLOR_A
	Normalised water leaving radiance	NLW
	Remote Sensing Reflectance	RRS
	Spectral attenuation coefficient for downwelling irradiance	K_490
	Total absorption	ATOT
	Total backscattering	BB
	Absorption by coloured dissolved organic matter	ADG
	Backscattering by particulate matter	BBP
	Absorption by phytoplankton	APH
Ozone	Ozone total column	TC
	Ozone nadir profile	NP
	Ozone limb profile	LP
Permafrost	Permafrost extent	PFR
	Ground temperature	GTD
	Active layer thickness	ALT
Sea Ice	Sea Ice Concentration	SICONC

	Sea Ice Thickness	SITHICK
	Sea Ice Extent	SIEXTENT
Sea Level	Corrected sea surface height	SSH
	Sea level anomaly	SLA
	Absolute dynamic topography	ADT
	Maps of sea level anomalies	MSLA
	Mean sea level	MSL
	Mean sea level trends	MSLTR
	Mean sea level amplitude and phase	MSLAMPH
Sea State	Significant wave height	SWH
Sea Surface Salinity	Sea surface salinity	SSS
Snow	Snow water equivalent	SWE
	Snow cover fraction – snow on ground	SCFG
	Snow cover fraction – viewable	SCFV
SST	Sea surface temperature	SSTint
	Sea surface skin temperature	SSTskin
	Sea surface subskin temperature	SSTsubskin
	Sea water temperature	SSTdepth
	Sea surface foundation temperature	SSTfnd
Soil Moisture	Surface soil moisture volumetric absolutes	SSMV
	Surface soil moisture volumetric anomalies	SSMVA
	Surface soil moisture degree of saturation absolute	SSMS
	Surface soil moisture degree of saturation anomalies	SSMSA
	Soil water index volumetric absolute	SWIV
	Soil water index volumetric anomalies	SWIVA
	Soil water index degree of saturation absolute	SWIS
	Soil water index degree of saturation anomalies	SWISA