––––

Upper Air Database v0

Issued by:

Date:

Ref:

Official reference number service contract:

This document has been produced in the context of the Copernicus Climate Change Service (C3S).

The activities leading to these results have been contracted by the European Centre for Medium-Range Weather Forecasts, operator of C3S on behalf of the European Union (Delegation Agreement signed on 11/11/2014). All information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose.

The user thereof uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission

and the European Centre for Medium-Range Weather Forecasts has no liability in respect of this document, which is merely representing the authors view.

Contributors

UNIVIE

F. Ambrogi

M. Blaschek

L. Haimberger

# Introduction

This document provides details regarding the Upper Air Database v0 .

With the term database we define a set of netCDF files, one for each observation station, that include radiosondes observation data from several original sources.

While several sources offer data for the same station, datasets are often incomplete,

i.e. they might lack of some observation time intervals, or some specific data might be missing. The aim of this work is indeed to collect the largest amount of information from several available dataset, and build a comprehensive database of all the available information. This include the observation data as well as any auxiliary metadata that can help interpret the meaning of the data, for example details about the sensors of the sondes or of the location of the observing stations.

This database will be soon delivered to the Copernicus Data Store, so that the necessary adjustments and interfaces with the CDS webpage can be implemented in order to make the dataset available for download and use.

The files are structured according the Common Data Model (CDM) that can be consulted on the GitHub page:

<https://github.com/glamod/common_data_model/>

In the rest of the document, we will describe the source of the data, how the data have been merged together in one unique file, the proposed structure of the output file, a brief description of the extensions and improvements we envisage for version v1, and a quick guide on how to access the data from the netCDF files.

# Data Sources

The data were contained in seven different original sources, that for brevity we labelled as ERA5\_1, ERA5\_1759, ERA5\_1761, ERA5\_3188, BUFR, IGRA2, NCAR.

ERA5\_1:

ERA5\_1759:

ERA5\_1761:

ERA5\_3188:

BUFR:

NCAR:

IGRA2:

… description of the datasets …

In general the structure of the file of each dataset is similar. In particular, we identify as record each different set of data that was acquired during each different launch of the radiosonde. Each record contains measurements of several variables at different pressure levels. In the CDM language, this information is stored in the ‘report\_id’ variable e.g. in the ‘header\_table’ or in the ‘observations\_table’. This number can be retrived in the seqno@hdr variable in the ODB files form the ERA5 datasets. The NCAR dataset, composed of is in text files, this information is stored in the header of each record, with a specific assigned number.

For the IGRA2 and BUFR file however this explicit information is missing. However it can be reconstructed by assigning a progressive number to each different record that is looped through following the temporal order of the observations in the file. Ina similar fashion, also a counter identifying each single observation of each single variable, at each pressure level, at each date and time can be defined. This constitute the variable ‘observation\_id’ in the CDM tables. Note that, as it will be discussed in Section xxx, these two numbers are necessary to identify the source of each data in the merged files.

# Merging Procedure

Each of the dataset described in the previous section potentially contain data for a certain station. The aim of what we call merging procedure is to combine information taken from different sources, and different input files, into a single output file in netCDF format. The aim of this Section is to extensively describe the merging procedure we use in order to combine the observation data from the different datasets.

As a trivial example, we consider the case where only one dataset include data (i.e. selection of a given measured variable, for a given pressure level) for a specific observation date and time. Note that with the term “data” we refer to the value of a meteorological variables (*temperature, wind speed, wind direction, relative humidity, geopotential*), reported at a specific date and time, for a specific value of the pressure (called *pressure level*).

For this trivial case, the merging procedure simply reduces to the selection of the data from the only possible original dataset. Now let's consider the case of two datasets which have data for the same variable, i.e. they report a measured value for a certain observable, for the same pressure level, acquired on the same date and time. Ideally both datasets will report the same value; however, it might happen than the values were, for example, rounded and the data from a certain dataset reports more significant digits than the other. Another possibility is that data were removed from a dataset since the corresponding measurements were flagged as not accurate, while the data might still appear in other datasets. Moreover, data might have been interpolated to standard pressure levels, so that in one dataset there might be more pressure levels available.

It is crucial to define in an unambiguous way how the data to enter the final merged file are selected, and allow to reconstruct their original source. Note that the merging procedure only concerns data belonging to the same report, i.e. for each available report only one dataset will be chosen, or in other words, there will be no combination of different dataset for different pressure levels, or different measured values for different variables in a given report: the data belonging to a defined report must come only from a single dataset.

We know describe how the merging procedure is performed.

**Merging of the observation date and time**

The first parameter to merge is the date and time ("date\_time" in the observations\_tables). Every record from each dataset has its own time stamp with the date and time. This might be slightly different from one dataset to the other; for the v0 of the database, we neglect any time difference and classify as different observations also those one which fall within a modest time difference, e.g. 1 or 2 hours, which in principle are the same observation (since typically one have only one launch per station every day, or at most every 6 hours). This is a check left for future versions of the database.

We then extract the list of “date\_time” for each dataset, and from those we create a combined list, that constitute the global set of available temporal observations.

**Merging of the pressure levels observations**

For each distinct *date\_time*, we have data from several dataset. We then loop over the available pressure levels, which might be different for each dataset. For each pressure level, a set of measurements for the different meteorological variables is available.

**Merging criterion**

The total count of distinct observed values for each pressure level, in each report i.e. each distinct ascent is taken as a highest level criterion for the choice of the dataset to be used in the merged file. This means that the dataset proving the largest number of measured data for a single report, will be taken as “best” dataset. Whenever several dataset report the same number of measured data,

We apply the following hierarchical selection:

IGRA2 > NCAR > ERA5 > BUFR.

**Observation id renumbering**

The “observation\_is” variable must be unique for each observation value in the “observations\_table”. To be consistent with this requirement, but at the same time to allow to retian the original observation\_id value from the original dataset, we implemented a convertion from the original value to the merged one in the following way. We define a dataset-numbering mapping as:

IGRA2: 1 , NCAR: 2 , BUFR: 3, ERA5\_1759: 4 , ERA5\_1759: 5, ERA5\_1761: 6, ERA5\_3188 :7

We then multiply this number times 1 billion, and we add it to the original ‘observation\_id’ . This way, the first digit on the merged id allows to identify the original dataset, while the digits following the zeros allow to obtain the original id.

# Output File

In this Section we describe the structure of the files that form the v0 of the database. The files will be extended and will include more metadata in the future versions. In the present form, they contain all the relevant information regarding the observation data, as well as the necessary auxiliary information to identify the original source of each entry.

We take as an example the "*merged\_chuadb\_windc\_82930.txt.nc*" file.

Each station file contains three CDM tables, i.e. the “*station\_configuration*”, the “*source\_configuration*” and “*observation\_table*”.

These tables are stores as netCDF groups. For the former two tables, there are distinct groups for each of the different dataset that contain observation data for that particular station, and that were merged together in the *observation\_table*.

As an example, the example file contains can be accessed like:

*import os,sys*

*import netCDF4 as ncdate\_time*

*F ='merged\_chuadb\_windc\_82930.txt.nc'*

*ds = nc.Dataset(F)*

and printing the variable ds will show the included groups:

*bufr\_source\_configuration, bufr\_station\_configuration,*

*era5\_1759\_source\_configuration, era5\_1759\_station\_configuration,*

*era5\_1\_source\_configuration, era5\_1\_station\_configuration,*

*igra2\_source\_configuration, igra2\_station\_configuration,*

*ncar\_source\_configuration, ncar\_station\_configuration,*

*observations\_table*

This means that data for this observing station is found in the original datasets: bufr, era5\_1, era5\_1759, igra2 and ncar, and the corresponding *station\_configuration* and *source\_configuration* tables can be found.

These tables contain unique information for the station, as collected from the original source files. They do not provide metadata varying with

each observation, so they are not repeated for each different observation.

The *source\_configuration*, as of now, contains only the useful variable *'source\_file'*, which contain the name of the data source file.

Most important is certainly the observations\_table, containing the following variables:

*“date\_tim”e, “latitude”, “longitude”, “observation\_id”, “observation\_value”, “observed\_variable”, “source\_id”, “z\_coordinate”*

For example,

*ds.groups[‘observations\_table’][:1]*

will access the first entry of the “observations\_table”, that looks like:

*1966-08-14 12:00:00 -9.37 -54.900002 5000022158 100.0 106 era5\_1759 85000.0*

This must be interpreted as the observation, at date and time *1966-08-14 12:00:00*, of the wind direction (number 106) equal to 100.0 [m/s], at the station at latitude and longitude *-9.37* and *-54.900002 .* The record\_id 5000022158 is unique for the merged file, and tells that the data in the original dataset (era5\_1759: 5) are identified by the 22158 record id.