PROTOCOL FOR THE AC DATA STORAGE IN ESNAS

BSC-AC-2016-002

Data storage

M. Gonçalves, S. Basart, M.T. Pay, M. Guevara, E. Di Tomaso, C. Pérez, P.A. Bretonnière, J. Cuadrado, K. Serradell

Earth Sciences Department

Barcelona Supercomputing Center - Centro Nacional de Supercomputación (BSC-CNS)

16 March 2017

Series: Earth Sciences (ES) Technical Report

A full list of ES Publications can be found on our website under:

http://www.bsc.es/projects/earthscience/ES-CFU/doku.php?id=start

Barcelona Supercomputing Center–Centro Nacional de Supercomputación (BSC-CN)

C/Jordi Girona, 31 | 08034 Barcelona (Spain)

Summary

This document summarizes the structure, filenames, and experiment and variable attributes to be adopted for storing Atmospheric Composition data in the department's network attached storage (esnas and esarchive). The structure has been defined as follows:

```
/esnas/exp/[$PROJECT]/$model[_v$version]/
  |--> constant/
  |--> $expid/
          |--> scripts/
          |--> restart_files/
          |--> docs/
          |--> original_files/
          |--> $outputfreq/
          |    |--> multivar/
          |    |    |--> $expid[-r$ensmemb]_$inittime.nc
          |    |    |--> [$expid[-r$ensmemb]_$inittime_an.nc]
          |    |--> $var/
          |    |    |--> $var[-r$ensmemb]_$inittime.nc
          |    |    |--> [$var[-r$ensmemb]_$inittime_an.nc]
          |    |--> ensemble_mean/
          |         |--> multivar/
          |         |    |--> [$expid_$inittime.nc]
          |         |    |--> [$expid_$inittime_an.nc]
          |         |    |--> [$expid_$inittime_cv.nc]
          |         |    |--> [$expid_$inittime_inc.nc]
          |         |    |--> [$expid_$inittime_oma.nc]
          |         |    |--> [$expid_$inittime_omb.nc]
          |         |    |--> [$expid_$inittime_orej.nc]
          |         |--> [$var]/
```

```
           |                |--> [$var_$inittime.nc]
           |                |--> [$var_$inittime_an.nc]
           |                |--> [$var_$inittime_cv.nc]
           |                |--> [$var_$inittime_inc.nc]
           |                |--> [$var_$inittime_oma.nc]
           |                |--> [$var_$inittime_omb.nc]
           |                |--> [$var_$inittime_orej.nc]
           |--> $timeint_$stat/
                 |--> multivar/
                 |     |--> $expid_$inittime.nc
                 |--> $var/
                       |--> $var_$inittime.nc
/esnas/obs/$institute-$obstype/
  |--> $obsdataset[_v$version]/
           |--> original_files/
           |--> $freq/
           |--> scripts/
/esnas/recon/$origin/
  |--> $dataset[_v$version]/$original_archive_structure
/esnas/oper/$model/ - to be defined
  |--> $domain/$outputfreq/$var/$var_$inittime.nc
```

The experiments inside a `PROJECT` folder are links to the actual data. This means that `/esnas/exp/$PROJECT/$model[_v$version]/$expid` is a link to `/esnas/exp/$model[_v$version]/$expid`.

All the scripts in the any `scripts` folder are links to the files stored in `/esnas/scratch/data_download_scripts`, that is a folder synchronized with a repository on gitlab.

All variable names, units and attributes have been standardized (details included in a table attached)

All experiments have to be identified (expid) and documented in the Earth Sciences Department wiki page:
https://earth.bsc.es/wiki/doku.php?id=working_groups:experiments

Contents

1. Context

Within Atmospheric Composition (AC) we deal with a variety of data coming from different sources and in different formats. This document defines the guidelines for data storage and formatting within the group, in order to harmonize the data structure within the Earth Sciences Department of the BSC.

General guidelines include:

- Data has to be classified according to its source (experiment results, observations, reconstructions or forecast outputs) – *Section 2.1, for details.*

- All the experiment outputs have to have associated a series of attributes that will be included in an experiment identification table – *Section 2.3, for details.*

- Standard variable names and units have to be used for all model outputs (either experiments or forecast) – *Section 3, for details.*

Some constraints have been considered, when defined those guidelines, which result in differences between the data structure in the AC and Climate Prediction (CP) groups. Those are presented in Table 1:

Table 1. Differences between AC and CP data structure for the storage

Data	CP format	AC format	Reason(s)
Model output	One variable per file (2D)	Multiple variables per file (2D, 3D and 4D)	AC uses a large number of variables, normally analyzes groups of variables at once and it is a standard practice on the community to share files with several variables included. Storing one variable per file would involve: Transfer and working with large number of files at once Need for file merging to share data with external users
	Analysis (restart) path outside main expid path	Analyses stored in the same dir as forecasts	It seems tidier

Reconstruction data	It matches the experiment output format (netCDF, one variable per file, 2D)	No formatting will be applied	The reanalysis and analysis data are normally used within AC as inputs for the models (initial and boundary conditions). Reformatting the files would involve modifying the pre-processing modules of the different modelling systems used within AC.
Observations data		Format will depend on the data source	Observations used within AC include data from surface stations and vertical profiles (.csv, .txt or .dat format), satellite data (GRIB, HDF5 or netCDF) and, occasionally, gridded datasets (netCDF or GRIB). Those data are not transformed prior to their use within the model evaluation routines.

2. Folders structure and naming convention

2.1. Overview of the structure

The folder structure within the department’s network attached storage (esnas and esarchive) includes four major data groups: experiment outputs (exp), observations data (obs), analysis and reanalysis (recon) and operational outputs (oper). An overview of the folder’s structure is given below. The naming convention is explained in more detail in section 2.2.

```
/esnas/exp/[$PROJECT]/$model[_v$version]/  - For experiment outputs
  |--> /constant/  - Optional folder to include fixed data for the model runs
  |--> $expid/ - Experiment id (mandatory)
        |--> restart_files/  - Optional folder with the boundary or initial
        |                      conditions
        |--> scripts/ - Optional folder to include post processing scripts
        |--> docs/ - Optional folder to include model/experiment documentation
        |--> $outputfreq/
              |--> multivars/       - For experiment files containing more than
              |     |                 one variable
              |     |--> $expid[-r$ensmemb]_$inittime.nc
              |     |--> [$expid[-r$ensmemb]_$inittime_an.nc]
              |                  - Optional for experiments producing analyses
              |--> $var/         - For experiment files containing only one variable
              |     |--> $var[-r$ensmemb]_$inittime.nc
              |     |--> [$var[-r$ensmemb]_$inittime_an.nc]
              |                  - Optional for experiments producing analyses
              |--> [ensemble_mean]/       - Optional folder for ensemble runs, it
              |     |                       contains optional ensemble stats
              |     |--> multivars/ - For experiment files containing more
              |     |     |                 than one variable
|       |       |--> [$expid_$inittime.nc]
|       |       |            - ensemble average forecast
|       |       |--> [$expid_$inittime_an.nc]
|       |       |            - ensemble average analysis
|       |       |--> [$expid_$inittime_cv.nc]
|       |       |            - normalised ensemble std_dev (ensemble
|       |       |              spread)
|       |       |--> [$expid_$inittime_inc.nc]
|       |       |            - analysis increments (analysis minus
|       |       |              background)
|       |       |--> [$expid_$inittime_oma.nc]
|       |       |            - analysis departures (observations minus
|       |       |               analysis)
|       |       |--> [$expid_$inittime_omb.nc]
|       |       |            - background departures (observations minus
|       |       |               background)
|       |       |--> [$expid_$inittime_orej.nc]
|       |                    - observations rejected by QC
|       |--> $var/ - For experiment files containing only one variable
|               |--> [$var_$inittime.nc]
|               |--> [$var_$inittime_an.nc]
|               |--> [$var_$inittime_cv.nc]
|               |--> [$var_$inittime_inc.nc]
|               |--> [$var_$inittime_oma.nc]
|               |--> [$var_$inittime_omb.nc]
|               |--> [$var_$inittime_orej.nc]
|--> [$timeint_$stat]/
|       |     - Optional folder to include post-processed variables
```

```
            |     |    (e.g. mean or maximum values of a certain variable over
            |     |    time)
            |     |--> allvars/
            |     |     |--> $expid_$inittime.nc
            |     |--> $var/
            |     |     |--> $var_$inittime.nc
/esnas/obs/[$institute[-$obstype]]/   - For observations
 |--> $obsdataset[_v$version]/
       |--> original_files/     - For the original observation files
       |                          (.dat, .txt, .nc, .hdf5, etc)
       |--> $freq/    - For the post-processed observation datasets (if any)
       |--> scripts/  - For the scripts used to download / post-process the
                       observational data
/esnas/recon/$origin/      - For the analysis and reanalysis data used to
 |                          generate initial and boundary conditions
 |--> $dataset[_v$version]/
       |-->$original_archive_structure
/esnas/oper/$model/   - For the operational forecasts outputs
 |--> $domain/$outputfreq/$var/$var_$inittime.nc
```

2.2. Definition of the structure

2.2.1. Experiment data: exp

Experiment data include the results of the simulations performed within the AC group, except those of the operational forecast systems, which have their own folder.

Experiment data, in netCDF format, are classified according to the project, the model, the experiment identifier number, the time frequency used in the output and, in case it is possible, the variable name.

Filenames include the experiment identification and the initial time for the simulation, in case of storing more than one variable per file, and only the variable name and the initial time for the simulation, in case of storing just one variable per file. They are stored as follows:

/esnas/exp/[$PROJECT]/$model[_v$version]/constant/ - For constant initial data and masks

/esnas/exp/[$PROJECT]/$model[_v$version]/restart_files/

- For boundary condition files: Files of boundary conditions with the required format for the specific model

- For initial condition files: Initial conditions in binary format

/esnas/exp/[$PROJECT]/$model[_v$version]/scripts/- Optional folder to include post processing scripts (links to the scripts repository in `/esnas/scratch/data_download_scripts`

/esnas/exp/[$PROJECT]/$model[_v$version]/docs/ - Optional folder to include model/experiment documentation

/esnas/exp/[$PROJECT]/$model[_v$version]/$expid/$outputfreq/

multivar/$expid[-r$ensmemb]_$inittime[_an/_cv/_inc].nc

- Experiment files including more than one variable

multivar/$expid[-r$ensmemb]_$inittime[_oma/_omb/_orej].nc

- Experiment files including more than one variable

$var/$var[-r$ensmemb]_$inittime[_an/_cv/_inc].nc

- Experiment data variables are stored in separated files

$var/$var[-r$ensmemb]_$inittime[_oma/_omb/_orej].nc

- Experiment data variables are stored in separated files

/esnas/exp/[$PROJECT]/$model[_v$version]/$expid/$outputfreq/ensemble_mean/ - For ensemble runs

multivar/$expid_$inittime[_an/_cv/_inc].nc - Experiment files including more than one variable

multivar/$expid_$inittime[_oma/_omb/_orej].nc – Experiment files including more than one variable

$var/$var_$inittime[_an/_cv/_inc].nc – Experiment data variables are stored in separated files

$var/$var_$inittime[_oma/_omb/_orej].nc – Experiment data variables are stored in separated files

/esnas/exp/[$PROJECT]/$model[_v$version]/$expid/$timeint_$stat/ - For specific diagnostics including additional post-processing (e.g. O3 8-h maximum concentration)

multivar/$expid_$inittime.nc – Experiment diagnostics including more than one variable

$var/$var_$inittime.nc – Experiment diagnostics including one variable

Table 2 describes the appropriate naming convention for each folder and file within exp.

Table 2. Naming convention for the folders and files to store experiment data (names between brackets are optional)

Tag	Description	Options
$[PROJECT]	Project name to include only if available	Acronym of the project associated to the simulations (upper case)
$model[_v$version]	Model name (version if desired). Hyphens (-) allowed, not underscores within the name (_)	nmmb-bsc-ctm_v23 wrf-hermes-cmaq bsc-dream8b
$expid	Experiment identification associated to a list of attributes (see section 3.2)	expid: code with format x000, character plus 3 digits (*)
[$domain]	Optional tag to identify the domain associated to the boundary conditions stored	d01, d02
$outputfreq	Output time frequency (more than one time step per file is allowed, but all have to have the specified frequency)	monthly, daily, 12hourly, 6hourly, 3hourly, hourly
$timeint_$stat	Diagnostic frequency and type of operation	8hourly_max, daily_mean, daily_max, monthly_mean ...
$inittime	Initial date of the data included in the file	YYYYMMDDHH, i.e. 2016030100
$var	Variable short name	See section 3

r$ensmemb	Number of ensemble	r01, r02, ...
$member	Name of ensemble member, to be used only for runs in ensemble mode. Format: 3 digits in numeric ascending order with leading zeros.	000, 001, 002, ...

(*) Currently this code will be generated by the experiment creator, in the future it will be provided by autosubmit. The code must be unique; therefore, before creating a new code, check the list of already used experiment ids.

2.2.2. Observations: obs

Observations include all data used for model evaluation and verification purposes. Sources range from surface measurement stations, vertical retrievals (i.e. from LIDARS or ozonosondes), satellite data, gridded verification datasets, etc. They are classified according to their origin. The storage structure will be as follows:

`/esnas/obs/[$institute[_$obstype]]/$obsdataset[_v$version]/`

`original_files/` – Includes raw data from the source (.dat, .txt, .nc, .hdf5, etc)

`$freq/` – Includes data that have been post-processed (if any) and a README file describing the type of post-process applied (i.e. averaging in time, regridding, filtering, etc.)

`scripts/` – Optional folder including the scripts used for downloading or treating the data (if any)

Table 3 includes the naming convention for some datasets commonly used within AC.

Table 3. Naming convention for the folders to store observations data

Tag	Description	Options
$institute	Acronym of the institution that provides the observational data	nasa, ncep, noaa, ecmwf, esa, eea, emep, cru […]
$obstype	Optional to classify type of data	gas-pollutants;meteo;aerosols;…
$obsdataset[_v$version]	Acronym for the observational dataset (including the source and the version)	calipso, eobs, cruts_v3.2 […]
$freq	Temporal frequency of the treated data	monthly, daily, hourly, etc. For satellite data with variable time-steps, use “satellite”

2.2.3. Reconstructions: recon

Reconstructions includes analysis and reanalysis data, normally used for initialization purposes and as driver for limited area simulations (boundary conditions). Due to the constraints posed by the initialization and pre-processing of each modelling system, the storage structure of the reconstruction data will be kept as in the original source(s). Data will be classified as a function of their origin. Sometimes, reanalysis data are also used as verification datasets, in those cases additional folders can be added with treated datasets and scripts (following the same criteria as in the observational datasets).

```
/esnas/recon/$origin/$dataset[_v$version]/$original_folder_structure/$degree_res
```

Table 4 describes the appropriate naming convention for each folder within reconstruction.

Table 4. Naming convention for the folders to store reconstruction data

Tag	Description	Options
$origin	Name of the institution that provides the data	ncep, ecmwf […]
$dataset[_v$version]	Acronym defining the dataset and including the version if needed.	fnl, gfs, era_interim, gldas
$original_folder_structure	Maintain the currently used structure	fnl: fnl_YYYYMMDD_HH_00 files gfs: archive_025_00 (folders YYMMDD00 with wafs.HH.0P25DEG) archive_025_12 (folders YYMMDD12 with wafs.HH.0P25DEG) archive_05_00 (folders YYMMDD00 with wafs.HH.0P5DEG) archive_05_12 (folders YYMMDD12 with wafs.HH.0P5DEG) archive_sst (folders YYMMDD00 with rtgssthr_grb_0.083_awips.grib2 rtgssthr_grb_0.083.grib2 sst2dvar_grb_0.5.grib2) […]
$degree_res[_initial_hour]	Spatial resolution of the dataset and initial hour for the analysis or reanalysis	xpyy[_HH], for x.yy degrees and hh initial hour.

2.2.4. Operational: oper[1]

The oper folder contains the outputs of the operational forecast systems. They are classified according to the modelling system, the modelled domain and the initial date of the forecast.

/esnas/oper/$model/$domain/$outputfreq/$var/$var_$inittime.nc

Table 5 describes the appropriate naming convention for each folder within oper.

Table 5. Naming convention for the folders to store operational data

Tag	Description	Options
$model	Model or modelling system name	bsc_dream8b, nmmb-bsc-ctm, wrf-hermes-cmaq, sds-was, bdfc
$domain	Domain name	asia, med, global, name, eu12, ip4, can2 […]
$outputfreq	Frequency of the output	hourly, 3hourly, 6hourly, monthly […]
$var	Variable name	See section 3
$inittime	Initial date	YYYYMMDDHH

The guidelines for its formatting and transfer to ESNAS have to be further defined.

2.2.5. General considerations

- All names must be in lowercase. The only exception is the folder names of the projects, which all must be capitalized
- There are no hyphens (‘-’) in the names in a path, with the only exception in the case that a model name is written explicitly with a ‘-’ inside
- Datasets without information about the version in their names refer to the first version of the dataset
- General scripts that operate on several subsets of a model or dataset can be placed at the top level folder inside the model/dataset, more particular scripts can be stored in the specific subfolder inside the model/dataset
- A README file with a description of the model or dataset available at the root

[1] The format for the operational forecasts storage will be kept as it is currently done until there will be resources to implement the CALIOPE system making use of NMMB/BSC-CTM, to modify and verify that all necessary evaluation/verification routines work properly with the new structure.

directory of that model/dataset. Datasets should be also documented in the department wiki.

- For more information on general conventions and rules, check: https://earth.bsc.es/wiki/doku.php?id=data:data_repo_conventions

2.3. Experiment identification table

Every newly created experiment has to have associated an experiment identification code. The code will be related to an entry in the experiments list table in the ES-BSC department wiki, where the following information has to be included:

1. Experiment ID (x000, check that the ID does not match any already in use)
2. Model
3. Version
4. Model set-up (description of model physics/chemistry options)
5. Initial and boundary conditions sources
6. Git Branch (whenever available)
7. Mode (CL: climate model, AN: analysis, FC: forecast)
8. Ensemble (Data assimilation experiment yes or no)
9. Domain
10. Resolution (horizontal, vertical - layers or pressure levels)
11. Temporal resolution of the original output
12. Initial simulation time
13. Final simulation time
14. Variables included in the output
15. Creator/Owner

In the near-future an enhanced web interface for experiment tracking will be available and the creation of the above table will be automatised at experiment creation time.

3. Variable naming convention

There is no general standard within the air quality, aerosols and meteorology communities for the variable naming convention. We have therefore defined our own convention, which is based in the National Centers for Environmental Prediction and the European Center for Medium Range Weather Forecasts naming conventions, as well as protocols from international model intercomparison projects (i.e. AEROCOM for aerosols).

A variable table (attached) has been created defining the following attributes:

1. Variable name
2. Standard name
3. Long name

4. Variable description
5. Units
6. Type of data (integer, real)
7. Dimensions (number and kind of dimensions, which for AC can include: time, x,y,z or pres, aerosols' bin)

New variables can be created if needed, following the rules below:

- Avoid variables with more than 4 dimensions (time,x,y,levels)
- Follow the existing standard for aerosols' and gas phase species naming
- Avoid including underscores in the variable names