NetCDF seasonal-to-decadal data handling strategy in ENSEMBLES: File structure and metadata

F. J. Doblas-Reyes, ECMWF

Version 22/10/2007

1. Introduction

The ENSEMBLES seasonal-to-decadal (s2d) ocean data should be written in CF-compliant (http://www-pcmdi.llnl.gov/cf) NetCDF format. The atmospheric data has also been written in NetCDF (from the original GRIB format) using a GRIB to NetCDF convertor. The data is publicly disseminated via an OPeNDAP interface, which can handle NetCDF files.

CF-compliant NetCDF provides a language that can be used for describing the data content of a file. However, no rules are yet available for describing data from a multi-forecast system (such as a multi-model ensemble), as it is the case with GRIB. Further, it does not (as yet) provide a standard logical structure for operational data. For example, a set of six fields with specific attributes could be described with those attributes in several structurally different ways with CF-compliant NetCDF. In order to be able to create files with data generated by different forecast ensemble systems in a homogeneous and consistent way, it is necessary to give very detailed instructions on how the data should be written. The requirements to encode data from IPCC simulations (http://www-pcmdi.lnl.gov/ipcc/IPCC_output_requirements.htm) are a good example of this sort of instructions, although it does not define the metadata that allow to code multi-forecast system simulations. Therefore, there is a need to describe the structure and metadata of multi-forecast system ensemble s2d data using well-defined NetCDF headers that could be also valid for simulations and forecasts of different time scales.

This proposal aims at extending the CF convention to avoid ambiguities in files containing multi-forecast system ensemble data. It takes account of established practice at operational centres and usual practice in the research community of climate variability at different time scales. It is based upon a more general proposal under

discussion by WCRP. The proposal tries to specify and agree on the required metadata, give rules and guidance on how the metadata is to be encoded in CF-compliant NetCDF, and, finally, decide how files should be structured for data exchange. A companion proposal offers a list of appropriate CF standard attributes to describe the physical variables of the ENSEMBLES common dataset.

In which form the proposals will succeed and become part of the CF convention is not yet known, so comments on these ideas are welcome. At present, the whole proposal is subject of a discussion in the CF mail list: http://www.cgd.ucar.edu/pipermail/cf-metadata/2006/thread.html. Follow the threads "CF and multi-forecast system ensemble data", "CF and multi-forecast -- provisional standards" and "Getting back to ensembles".

2. NetCDF metadata

NetCDF files describe their content using three types of metadata:

- Dimensions, which define the size of the different variables
- Variables, specifying the type of data the file contains
- Global attributes, which provide additional information that applies to all the data in the file

The variables consist of the coordinates in space and time plus the physical field itself. Some other variables with additional information can be included. In addition, CF-compliant NetCDF mandates or recommends the following global attributes to document the overall nature of the data:

- Conventions, such as CF-1.0
- Title, such us the project at the origin of all the data
- Institution
- Source
- History, including modifications to an eventual original file
- References, especially web links

- Comment

The above fields are often filled in as lengthy, human-readable strings, sometimes with multiple pieces of information under one heading. These *human-readable* metadata are intended purely for human use and should not be used for characterizing the large variety of data the file might contain. As an example, consider that we attempt to characterize a forecast from the EUROSIP multi-model seasonal forecast operational system using only global attributes in the following way:

- Title: Météo-France seasonal forecast data
- Institution: “Model created and run by Météo-France. Data processed by ECMWF. Data distributed by ECMWF.”
- Source: “Data generated by the ARPEGE model coupled to OPA, run by Météo-France at ECMWF.”
- History: “Simulation created in real time.”
- References: “http://www.ecmwf.int/products/forecasts/seasonal/documentation.html”
- Comment: “Part of EUROSIP multi-model forecast system. Use of data subject to EUROSIP data policy; see web link for details”

Ideally the above would contain additional information, such as version numbers, system numbers, original resolution, etc. However, it would be impossible to include data from another EUROSIP forecast system, as the global attributes would not be appropriate to distinguish between the fields produced by each forecast system. In addition, CF has not considered up to know how to describe the elements of an ensemble of simulations in a file. Therefore, a CF-compliant NetCDF file of an ensemble of simulations requires the definition of a set of variables, for which no standard names exist yet, allowing the use of some of the above mentioned global attributes as variables for the sake of appropriately describing the ensemble. This document discusses a proposal to solve this problem by creating rules to specify:

- The metadata content needed to describe the data of a specific multi-forecast system s2d experiment

- How these metadata is to be specified in CF-compliant NetCDF files

3. Metadata content

In addition to the dimensions and coordinate variables (longitude, latitude, level, time, etc) that usually describe the physical field (say temperature), an additional dimension and a set of metadata are proposed below. The metadata should be machine-readable and properly distinguish different fields in a way that enables the data to be handled and archived independently of the file where they are contained. They should also be useful for data searching and automatic plot labelling.

In a multi-forecast system context, a physical field might include data from multiple forecast systems, either from a single centre (in an initial-condition ensemble for example) or from several institutions (for instance, in the case of a multi-model). The ECMWF GRIB attributes origin/expver/system/method (http://www.ecmwf.int/publications/manuals/mars/guide/Language_hints_for_ECMWF_fields.html) make a natural *tuplet* to define a particular homogenous ensemble forecast. The ensemble of fields is spanned in the NetCDF file by a new dimension (called *ensemble* henceforth) that is identified by the coordinate variable with standard_name “realization”. A multi-model ensemble forecast, an initial condition ensemble or a perturbed-parameter ensemble are all made of a collection of such tuplets.

Many of the metadata are logically independent, in the sense that specifying one does not fix the value of another. They identify uniquely the data, while others provide information about the individual values of other data variables that help identify the source of the data. For example, we might want to provide the name of the institution at the origin of the data, or describe certain characteristics of a given experiment used to produce the ensemble. Such logical connections are noted below, since in some representations of the data, they may affect how the metadata can or should be coded. In the following, it is assumed that data to be exchanged are raw model output. The list of standard names (apart from the spatial coordinates defining the longitude, latitude and level/depth) required in the definition of the fields are:

- **forecast_reference_time** (FLOAT). Start date of the forecast.

- **forecast_period** (INTEGER). Elapsed time since the beginning of the forecast.

- **realization** (INTEGER). Number identifying the integrations made with a given source, institution and experiment_id but with different initial-condition perturbations, which form a homogenous and statistically indistinguishable ensemble. In other words, each ensemble member in the multi-forecast system forecast/simulation has to be individually identified. This way, the metadata caters for ensembles from different forecast systems with different ensemble sizes.

- **experiment_id** (STRING). The producing centre is responsible for assigning unique experiment identifiers for the different experiments created, and should (ideally) provide documentation of each experiment. It is possible for common experiment identifiers to be agreed between different centres, if they are carrying out a common experiment. But there is no *a priori* guarantee that identical identifiers from different centres could refer to scientifically equivalent experiments.

- **institution** (STRING). Institution with scientific responsibility for the forecast system, eg Met Office or ECMWF. It takes the value of the GRIB attribute “origin”.

- **source** (STRING). It combines information about the forecast system, including the name of the model and the variations introduced such that different integrations have different properties (for instance, the different climate drift of the members of a perturbed parameter ensemble forecast). This information should be used to identify the models, distinguish between different prediction systems used by a specific centre. It takes the values of the GRIB attributes “system” and “method” as defined in the RT1 web site (http://www.ecmwf.int/research/EU_projects/ENSEMBLES/data/atmosphere_archiving_GRIBheaders.html) For instance, the Met Office will have to choose a different value for the GloSea forecasts and for DePreSys (as both are based on the HadCM3 coupled model). A table online should provide detailed information of the systems.

A file containing forecasts from several consecutive start dates might have fields that, while having been produced from different start dates, verify at the same time. The variables “forecast_reference_time” and “forecast_period” are independent time variables employed to define the two time axes of a forecast with several start dates and lead times that specify a unique forecast field. One of the time axes spans the start dates of the forecasts, while the second one ranges from the beginning to the end of a given forecast. Both share the same dimension, which amounts to the total number of time steps for all the start dates contained in the file. This implies that the “forecast_reference_time” is repeated as many times as forecast time steps are available in the file. This prevents having forecasts with the same verifying date but produced from a different start date (and, hence, intrinsically different). Software reading the files should use both coordinates to estimate the verification time of a particular field. Note that “forecast_period” cannot have time units referenced to a specific date, as “forecast_reference_time” does. In other words, every field is identified by the start date of the forecast, referred to a fixed date in the past (usually, the first of January 1950) and by the forecast time referred to the beginning of the forecast (i.e., not referred to fixed dates). We believe this is an appropriate way to code the time dimension of the forecasts in CF-compliant NetCDF, even though the units for “forecast_period” do not match the specification used in other instances (eg, IPCC data). Feasible alternatives can be found in the CF documentation.

The ensemble dimension is used to identify a source (a multi-model ensemble produced by a specific forecast system or a member of a perturbed-parameter ensemble), a realization (a member of an initial-condition ensemble), an institution and an experiment_id. The four variables are dimensioned with the number of ensemble members in the file, which has to be a full hypercube. This dimension is usually called “ensemble”. A multi-forecast system ensemble dataset uses these variables to distinguish the elements of a given experiment or operational forecast.

Although not actually needed for distribution and archive purposes, additional variables with the ensemble dimension can be also suggested (although they are not used with the ENSEMBLES data):

- **original_distributor** (STRING). Centre with responsibility for distribution of data, i.e. the centre who first made the data available, and to whom queries of data integrity should be sent (eg ECMWF).

- **production_status** (STRING). Operational, research or a user defined project identifier. The value “research” should be used for general research at a specific centre, while project_id should be used for specified international research projects.

- **sst_specification** (STRING). It describes the use of the SSTs in the specific experiment and can take values such as “coupled”, “observed”, “predicted”, “persisted anomaly” or “persisted absolute”.

- **real_time** (CHARACTER). It takes the values “true” or “false”, according to whether the forecast was or not made in real-time. It is an attribute of the individual forecasts.

- **archive_date** (INTEGER). It takes the format “YYYYMMDD” or “unknown” to describe when the data was archived or published. The aim is to provide an approximate timestamp, to easily distinguish between recent experiments and much older ones. Also, in the case that data need to be corrected in a globally distributed data system, the archive_date could be used to distinguish between the older, original data and the newer, corrected data. This is an attribute of the individual forecasts.

These variables offer flexibility in describing multi-forecast system ensemble experiments or operational forecasts, and are intended to allow a straightforward mapping from existing metadata practice in the global seasonal forecasting community. However, CF intends to be parsimonious and these variables have not been included in the proposal under discussion.

For ENSEMBLES, a full description of the experiment identifiers is available from http://www.ecmwf.int/research/EU_projects/ENSEMBLES/table_experiments/. The use of a web link is much more appropriate than trying to include large amounts of detail in the NetCDF file itself, and also allows relevant information to be kept up to date.

Note that the rules described are CF compliant in that it does not introduce any new standard_name attributes for physical variables. However, if we want to standardize the usage that we propose here, such that application software can unambiguously interpret these data files, it would be desirable to ask for the CF convention to be extended. Such a request might be made separately for the simple concept of the ensemble dimension (necessary for any sort of ensemble forecast to be represented and presumably not controversial as a concept) and the more complex "origin/expver/system/method" tuple needed to represent multi-forecast system ensemble data. If CF approval were to be given, the above layout of variables would be unchanged.

4. Examples of metadata in CF-compliant NetCDF

Below you can find an example of CF-compliant metadata to construct fully informative headers for a hindcast of 20 start dates of 7-month long daily data on 10 pressure levels from 5 different forecast systems using 9-member ensembles each:

- dimensions:
    - latitude=180 ;
    - longitude=360 ;
    - level=10 ;
    - time=4200 ;
    - ensemble=9 ;
    - string15=15 ;
    - string8=8 ;
    - string4=4 ;
- variables:
    - float longitude(longitude);
        - longitude:data_type = "float" ;
        - longitude:units = "degrees_east" ;

        - longitude:axis = "X" ;
        - longitude:standard_name = "longitude" ;
        - longitude:topology = "circular" ;
        - longitude:modulo = 360 ;
        - longitude:valid_min = 0. ;
        - longitude:valid_max = 359. ;
    - float latitude(latitude);
        - latitude:data_type = "float" ;
        - latitude:units = "degrees_north" ;
        - latitude:axis = "Y" ;
        - latitude:standard_name = "latitude" ;
        - latitude:valid_min = -89. ;
        - latitude:valid_max = 89. ;
    - float level(level);
        - level:data_type = "float" ;
        - level:units = "hPa" ;
        - level:axis = "Z" ;
        - level:standard_name = "air_pressure" ;
        - level:positive = "up" ;
    - float reftime(time);
        - reftime:units=”days since 1950-01-01 00:00:00” ;
        - reftime:standard_name=”forecast_reference_time” ;
        - reftime:long_name=”Start date of the forecast” ;
    - long leadtime(time);

        - leadtime:units = "hours" ;
        - leadtime:standard_name = "forecast_period" ;
        - leadtime:long_name = "Time elapsed since the start of the forecast " ;
    - char institution(ensemble,string15) ;
        - institution:standard_name = "institution" ;
        - institution:long_name = "Institution responsible for the forecast system” ;
    - char experiment_id(ensemble,string15) ;
        - experiment_id:standard_name = "experiment_id" ;
        - experiment_id:long_name = "Experiment identifier" ;
    - int source(ensemble,string15) ;
        - source:standard_name = "source" ;
        - source:long_name = " Method of production of the data" ;
    - int realization(ensemble) ;
        - realization:units = "1" ;
        - realization:standard_name=" realization" ;
        - realization:long_name=”Number of the simulation in the ensemble” ;
    - float physical_field(time, ensemble, level, latitude, longitude) ;

For simplicity, two tags are used for the variables with standard_name “forecast_reference_time” (reftime) and “forecast_period” (leadtime).

As explained above, information such as that contained in “institution” needs to be coded in the form of variables rather than global attributes because global attributes are a file-based concept and would restrict the data that could be written in the file. The “ensemble” dimension is used to reference similarly-dimensioned

variables (“institution” through to “realization”) to be used as *auxiliary variables* within the meaning of the CF convention. In other words, “ensemble”, as a new dimension, is intended to allow for multi-forecast system data in the same file by indexing the forecast-defining tuples within the file. It should be preferably omitted from the NetCDF file if it contains a single ensemble member. Note that the use of this new dimension is one of the features that are delaying the introduction of these rules in the CF convention.

Two more simple examples are as follows. First, for an ENSEMBLES stream 1 coupled ECMWF 9-member ensemble experiment:

- institution = ECMWF (repeated 9 times);
- experiment_id = 1004 (repeated 9 times);
- source = IFS29R2/HOPE-E System 1 Method 1 (repeated 9 times);
- realization = 0, 1, 2, 3, 4, 5, 6, 7, 8 ;

While for a Met Office perturbed parameters experiment with 9 versions of the HadCM3 model:

- institution = ECMWF (repeated 9 times);
- experiment_id = 1504 (repeated 9 times);
- source = DePreSys System 51 Method 10, DePreSys System 51 Method 11, DePreSys System 51 Method 12, DePreSys System 51 Method 13, DePreSys System 51 Method 14, DePreSys System 51 Method 15, DePreSys System 51 Method 16, DePreSys System 51 Method 17, DePreSys System 51 Method 18 ; (for the 9 versions of the same model with different parameters)
- realization = 0 (repeated 9 times);

5. Additional requirements

CF-compliant NetCDF offers a great deal of flexibility to encode the data. However, homogeneity is required in a unified dataset. Below you can find a few issues that you should bear in mind when creating your files:

- Monthly means and accumulated variables are encoded using the “cell_methods” attribute with the forecast_period variable. The “cell_methods” attribute allows any sort of time averaging to be described in the NetCDF header. The main inconvenient is that some applications may be unable to decode it. As an example, the following attributes are required to encode monthly means:
    - long leadtime(time) ;
    - leadtime:units = "hours" ;
    - leadtime:standard_name = "forecast_period" ;
    - leadtime:long_name = "Time elapsed since the start of the forecast" ;
    - leadtime:bounds = "time_bnd" ;
    - long time_bnd(time,time_bnd) ; with time_bnd dimension= 2
    - time_bnd:units = "hours" ;
    - float physical_variable(ensemble, time, level, latitude, longitude) ;
    - physical_variable:cell_methods = "leadtime: mean (interval 1 day)" ; where the interval points out at the use of daily data to compute the monthly mean

  with the following information in the data section:

    - time_bnd = 0, 744, 768, 1464, 1488, 2208, 2232, 2952, 2976, 3672, 3696, 4416; for monthly means over 6 months

  More information on the use of cell_methods can be found in the accompanying document on standard names.

- The use of the additional dimension “ensemble” can make datasets impossible to use with those applications (eg Grads, Ferret) that do not properly handle multi-dimensional (more than four dimensions) datasets. This is a limitation that, hopefully, will be corrected in the future once an agreement to encode multi-forecast system ensemble data has been reached. At present, Grads is

working on creating a version of GDS that handles five dimensions, although it is not expected to be CF-compliant.

- ENSEMBLES requires the preparation of files containing monthly means of the ocean analyses. Analyses only require the use of a time axis (versus the double time axis described for the forecasts). This axis takes the standard_name “time”, is dimensioned with the number of monthly means in the file and has units similar to “forecast_reference_time”.