

Make HDF5 files netCDF4-Compatible and CF-compliant within Groups

Recommendation:

We recommend that all HDF5 Earth Science product files be made netCDF4-compatible and CF-compliant within groups.

Recommendation Details:

Compatibility with netCDF4

Unlike netCDF3, netCDF4 is based on HDF5 and thus allows for the creation of group structures. Therefore, it makes sense to create group structures in netCDF4 directly, or within HDF5 products that can be read through the netCDF4 API. This can be achieved by adding dimension datasets and dimension scales that follow the netCDF data model to the HDF5 products.

Example: A dimension named Time:

1. When setting up the definitions, a dataset called Time is created at the root level with the required values.
2. The dataset at the root level is turned into a dimension scale via the `H5DSset_scale` function. This allows datasets below the root level to attach to the dimension dataset via the `H5DSattach_scale` function.

CF-Compliant Within Groups

The CF conventions are widely employed guidelines for Earth Science data and metadata storage. The purpose of the CF conventions is to require conforming datasets to contain sufficient metadata that they are self-describing in the following ways: Each variable in the file has an associated description of what it represents, including physical units if appropriate; and each value can be located in space (relative to Earth-based coordinates) and time. Thus, adhering to CF guidelines will increase completeness, consistency, and interoperability of conforming datasets.

Currently CF only applies to "flat files" with a single group, and not to files with multiple groups or hierarchical structures that typify modern NASA satellite datasets. Until CF is extended to apply to multiple groups, such NASA datasets can be most CF-compliant by following CF within each group.

To achieve the maximum CF-compliance within each group, we recommend the following:

1. Datasets in the group have the required CF attributes.
2. For the cases when horizontal space coordinates can be described with one-dimensional latitude and longitude arrays the following is recommend:
 - a. Dimension datasets are located at the appropriate level for the group.
 - b. Dimension datasets for the group have the appropriate CF attributes (e.g., no fill value).
 - c. The appropriate dimension scales have been implemented for the group.
3. For the cases when horizontal space coordinates can be described only with two-dimensional latitude and longitude arrays, the CF coordinate attribute that consists of the coordinate dataset names must be used for each corresponding dataset. To avoid ambiguity and to take advantage of some popular CF tools ([Panoply](#), etc.) that already support HDF5/netCDF-4 files with multiple groups, the following is recommended to handle the coordinates:
 - a. For coordinate datasets that are not latitude and longitude, follow 1)-3) under 2. above.
 - b. Make all dimensions associated with the 2-D latitude/longitude arrays pure netCDF dimensions by defining the NAME attribute of the dimension dataset to be "This is a netCDF dimension but not a netCDF variable."
 - c. If only one pair of 2-D latitude/longitude arrays is needed for the whole file, then define them at the root level. This can make more CF tools visualize the physical HDF5 datasets.
 - d. Use the absolute path of HDF5 datasets that store latitude and longitude for the coordinate attributes. For example, consider an HDF5 dataset temperature that is under the group g2, and whose parent group is g1 (i.e., `float /g1/g2/temperature[Dim1][Dim2]`). The two-dimensional latitude and longitude fields that describe this temperature field is under the group g1 (i.e., `/g1/latitude[Dim1][Dim2]`, `/g1/longitude[Dim1][Dim2]`). One should define a coordinates attribute `coordinates="/g1/latitude /g1/longitude"`.