

Review of Climate and Forecast Metadata Conventions Implementation and Operational Suitability

NASA's Earth Science Data Systems Standards Process Group (SPG) is considering the Climate and Forecast (CF) Metadata Conventions, for adoption as a community standard. You are invited to review this Requests For Comment (RFC) in the context of your **implementation experience** with this specification and its **suitability for operational use**. You only need to answer questions that are applicable to you. Please send completed review to:

spg-rfc-021@lists.nasa.gov.

Implementation Experience questions:

1. *(Your background)* Describe in a sentence or two your overall implementation experience related to the proposed specification. (e.g., *specification implementer, tools developer, data provider, scientific analyst, science user, etc.*) Have you directly implemented the CF metadata conventions? Did you use pre-existing software, and if so, what did you use?

We are primarily data providers, with some work in tools development. Our DAAC has a range of expertise which we use in applying the CF convention, including Java developers, metadata experts, and science experts. We are a node in the Earth System Grid (ESG), which uses netCDF-CF heavily, which provides us direct access to the CF team at PCMDI. We have implemented CF data using a range of tools, including NCO and the CF checker.

2. *(Completeness)* Does the specification (the online documents referenced) provide all the detail you need to implement it in software? (e.g., *to read or write a data file; to implement or modify a profile or extension; or develop a tool such as a metadata translator*) If not, describe what is missing in the specification.

Yes, the specification is generally complete enough for implementation. There are some exceptions with details of standard names, as defined below. The specification is much more complete for CF compliance in netCDF format. The CDL format (CF in XML, often used for documentation and from dump utilities) is somewhat less well described and much harder to find. For smaller datasets, such as field ecology studies, it may be appropriate to store data using an XML format, as well as using XSLT to enable translations from other formats with controlled vocabularies for variable names.

3. *(Accuracy)* Do any parts of the specification contain inaccuracies, or internal inconsistencies? If so, please provide details.
We have not identified any at this point.

4. *(Clarity)* Is any part of the specification ambiguous, or poorly explained? If so, please provide details.

Generally, yes. We have significant experience with netCDF-CF and notes that one part was unclear: the specification document, p. 10, says: "Unless it is dimensionless, a

variable with a `standard_name` attribute must have units which are physically equivalent (not necessarily identical) to the canonical units..." A specific example in our case related to whether our temperature variables in degC would need to be converted to the CF canonical unit K. After several Web searches we think we determined the answer is no (see <http://www.mail-archive.com/cf-metadata@cgd.ucar.edu/msg00454.html>). Follow-ups in this thread suggested that the specification wording would benefit from clarification of this concept of "physically equivalent" units; we support providing such clarification. In addition, as suggested by another thread, the Uunits table should be Web viewable without having to download and install the full package. Our conclusion about not needing units conversion is based in part on the earlier version of the Uunits material, which is available on the Web if you search for it. Easy accessibility of the standard and its components, to both the data provider and the data user communities, is an important precursor to acceptance and success.

As a further point, there are other concerns with standard names, particularly for chemical entities and for other variables not commonly used in the climate modeling framework. See our response to question 9 for further details.

5. (*Balance*) Does the standard describe the right set of concepts and attributes and enable the appropriate operations for its intended users? In particular, have the guiding principles outlined in section 5.2 been followed in the development of standard names?

Generally, yes. We have some concern, however, that the 4th principle (i.e., the need to be both human-readable and machine-readable) is generally implemented in a way that causes the name itself to carry too many meanings, causing some degree of confusion. As one example, in `surface_carbon_dioxide_mole_flux`, "surface" is defined (the lower boundary of the atmosphere), in some contrast to the field measurements in the fluxnet community (called FC). In flux measurements, surface is defined as "right above vegetation layer" or canopy height, or with specific measurement of tower height. So, while CF makes a lot of progress in defining terms in ways that they can be machine interpreted, there are still some ambiguities and challenges for integration that could render names even more cumbersome. In the surface CO₂ flux example, there could be comparable flux measurements at multiple elevations, and it will be difficult to name the variables in ways that allow for direct comparison with remote sensing or model data.

6. (*Usefulness*) How well does this specification meet your information sharing needs? (*e.g., Does it properly represent and describe your datasets? What are the pros and cons of these metadata convention attributes?*)

The specification is generally useful, particularly where there is clear overlap between our data and climate modeling variables. It lacks standard names for a number of field ecology variables, which is mostly an issue where we need to integrate that data with regional and ecosystem modeling results, and potentially for integration with remote sensing and reanalysis products. As noted above, there are places where the standard naming convention is problematic where the ecology community makes measurements at multiple different levels (offsets) and those offsets are difficult to incorporate into the variable name (and perhaps should not be incorporated into the variable name). A common flux convention, for example, is that there is a variable named soil temperature, which has an attribute of "offset" that describes the depth in the soil at which the

temperature is measured.

7. *(Implementation)* What implementation challenges does the proposed standard present? (e.g., does it conflict with other metadata requirements for your data? Is it compatible with the data formats you use?)

A substantial challenge is working with our backfile. We are evaluating new datasets for CF appropriateness, and are making that data format available where it is appropriate. Evaluation and conversion of the backfile of already-published data would require a substantial effort, and we are evaluating what datasets are sufficiently value-added to convert to CF. ORNL DAAC data sets are generally not in self-describing formats, so conversion can be a very manual process. This is also an issue for other archives at ORNL and elsewhere. We anticipate some problems where we will wind up needing to create variable names for which standard names do not exist, but where a name different from what we choose will get added to the list at a later point in time.

It is also important to recognize that CF metadata is largely at the granule level, and does not represent sufficient information to be able to generate full ECHO or GCMD collection-level metadata records. There is a need to store and represent the collection level metadata in ways that are outside of CF-compliant files.

8. *(Flexibility)* In what software environment(s) have you used the CF metadata conventions (e.g., Solaris, Linux, Windows, Mac OS X)?

Primarily Windows and Linux (Red Hat, Solaris, and Ubuntu). Some work in Mac OS X (growing).

9. *(Standard Names)* In your opinion, does the standard name table provide an adequately comprehensive set of names for the metadata representation?

The standard names need to be better organized and a hierarchical structure may be appropriate, potentially with semantic relationship expressions and “about this variable” metadata. If the names are not well organized, with more and more names defined in the table, it will be harder and harder for users to choose the appropriate names.

It is also clear that a large number of additional names are needed for ecology, field atmospheric science, and remote sensing quantities.

For names relating to chemical quantities, the situation is particularly complex.

“Conventions may also add value to scientific formats by providing higher-level abstractions such as coordinate systems, and by supporting capabilities needed by specific communities, such as standard names for physical quantities to determine whether data from different sources are comparable and to distinguish variables in archives.”

A. The conventions for standard names are lacking for physical and chemical quantities.

In the standard name table, entries typically use a common name in the “standard name” and might include a standard nomenclature / convention name in the description. See alpha_pinene example below. But for benzene, see example below, no standard convention name is mentioned. The value of embedding this info in the

description is limited as far as I can tell.

Source: <http://cf-pcmidi.llnl.gov/documents/cf-standard-names/standard-name-table/14/cf-standard-name-table.xml>

```
<entry id="atmosphere_mass_content_of_alpha_pinene">
  <canonical_units>kg m-2</canonical_units>
  <description>"Content" indicates a quantity per unit area. The "atmosphere content"
    of a quantity refers to the vertical integral from the surface to the top of the
    atmosphere. For the content between specified levels in the atmosphere, standard
    names including content_of_atmosphere_layer are used. The chemical formula for
    alpha_pinene is C10H16. The IUPAC name for alpha-pinene is (1S,5S)-2,6,6-
    trimethylbicyclo[3.1.1]hept-2-ene.</description>
</entry>
```

mass_concentration_of_alpha_pinene_in_air

Mass concentration means mass per unit volume and is used in the construction mass_concentration_of_X_in_Y, where X is a material constituent of Y. A chemical species denoted by X may be described by a single term such as 'nitrogen' or a phrase such as 'nox_expressed_as_nitrogen'. The chemical formula for alpha_pinene is C10H16. The IUPAC name for alpha-pinene is (1S,5S)-2,6,6-trimethylbicyclo[3.1.1]hept-2-ene

Source: <http://cf-pcmidi.llnl.gov/documents/cf-standard-names/standard-name-table/14/cf-standard-name-table.html>

```
= <entry id="atmosphere_mass_content_of_benzene">
  <canonical_units>kg m-2</canonical_units>
  <description>"Content" indicates a quantity per unit area. The "atmosphere content"
    of a quantity refers to the vertical integral from the surface to the top of the
    atmosphere. For the content between specified levels in the atmosphere, standard
    names including content_of_atmosphere_layer are used. The chemical formula for
    benzene is C6H6. Benzene is the simplest aromatic hydrocarbon and has a ring
    structure consisting of six carbon atoms joined by alternating single and double
    chemical bonds. Each carbon atom is additionally bonded to one hydrogen atom.
    There are standard names that refer to aromatic_compounds as a group, as well
    as those for individual species.</description>
</entry>
```

mass_concentration_of_benzene_in_air

Mass concentration means mass per unit volume and is used in the construction mass_concentration_of_X_in_Y, where X is a material constituent of Y. A chemical species denoted by X may be described by a single term such as 'nitrogen' or a phrase such as 'nox_expressed_as_nitrogen'. The chemical formula for benzene is C6H6. Benzene is the simplest aromatic hydrocarbon and has a ring structure consisting of six carbon atoms joined by alternating single and double chemical bonds. Each carbon atom is additionally bonded to one hydrogen atom. There are standard names that refer to aromatic_compounds as a group, as well as those for individual species.

Source: <http://cf-pcmidi.llnl.gov/documents/cf-standard-names/standard-name-table/14/cf-standard-name-table.html>

B. Picking a convention is not straightforward. If IUPAC is going to be the default convention (?) then a cross-reference capability to CAS RNs and Chemical Abstracts Index Names (and others?) should be provided and vice-versa.

CAS_RN	Chemical Abstracts Index Name (CAS_9CI)	Name_IUPAC
108-38-3	Benzene, 1,3-dimethyl-	1,3-Methylbenzene

Note that both of these chemical nomenclature standards lack names for physical quantities.

C. Consider a more comprehensive convention, for example, the EPA Substance Registry. Identifies substances with a Systematic name, registry name, CAS RN; and assigns an “EPA Identification Number” when a CAS RN is not applicable.

http://iaspub.epa.gov/sor_internet/registry/substreg/home/overview/home.do

U.S. EPA Substance Registry Services

The system provides a common basis for identification of, and information about:

- Chemicals
- Biological organisms
- Physical properties
- Miscellaneous objects

The SRS provides a range of services to users:

- Search and retrieval of:
 - Single substances
 - Programmatic, statutory or other lists of substances
 - Groups of substances
- **Information about creating machine-to-machine integration between the SRS and other systems**
- Outreach and education material to gain a better understanding of the SRS and its services
- Links to related regulatory information within EPA and other federal agencies and states

Substance Details Report

May 24, 2010

Core Metadata

Substance Type	Chemical Substance	Internal Tracking Number	8425
EPA Registry Name	.alpha-Pinene	Substance Status	Approved
EPA Registry Name List	Chemical Identification		
Systematic Name	Bicyclo[3.1.1]hept-2-ene, 2,6,6-trimethyl-		
Systematic Name List	Chemical Abstracts Index Name		
Preferred Acronym			
CAS Number	80-56-8		
EPA Identification Number			
TSN Number			
Kingdom			
ICTVDB			
Molecular Formula	C10H16		
Molecular Weight	136.24		
Comment/Description			
Definition			
Classifications			
Associated Identifiers			
Former CAS Numbers	2437-95-8		
Structure			
Notation	C1(=C)C2CC1C2(C)C	Notation Type	SMILES

Page 1 of 4

Note physical substances have been assigned a systematic name and an “EPA Identification Number”. For example,

Substance Details Report

May 24, 2010

Core Metadata

Substance Type	Physical Property	Internal Tracking Number	1647635
EPA Registry Name	Particulate matter - PM2.5	Substance Status	Approved
EPA Registry Name List	Chemical Identification		
Systematic Name	Particulate matter (PM2.5)		
Systematic Name List	Environmental Protection Agency Data Systems		
Preferred Acronym			
CAS Number			
EPA Identification Number	E1647635		
TSN Number			
Kingdom			
ICTVDB			
Molecular Formula			
Molecular Weight			
Comment/Description	E1647635 has replaced E624671245. Refer to 40 CFR Part 58 and Appendix A, 5.5.		
Definition	Particulate matter where particles are less than or equal to 2.5 micrometers in size.		
Classifications			
Associated Identifiers	There is no information about former or incorrectly used identifiers.		
Structure	There is no structure information for this substance.		

Page 1 of 3

D. A convention or conventions need(s) to be identified as soon as possible so that existing archives have a clear target for any translations that they may need to undertake.

Additional fields may need to be added to the standard-name-table to identify “alternative IDs” or to specify the nomenclature “thesaurus / convention / standard”.

Some existing data archives have implemented internally consistent naming conventions that are “supporting capabilities needed by specific communities”.

Using existing resources when possible. Consider the EPA registry, ARM archive (<http://www.archive.arm.gov>), and on a smaller scale, check NARSTO resources for specific examples of atmospheric constituent cross references (<http://cdiac.ornl.gov/programs/NARSTO/epasupersites.html>) and some shortened names for a community. See “Guidelines for Using Consistent Names...Accepted Names” in (http://cdiac.ornl.gov/programs/NARSTO/MILAGRO_ICARTT_Names_20060316.pdf).

E... A standard reference table for common names is desirable if the community decides not to adopt / implement a more rigorous nomenclature standard. This would permit others to create value-added cross-reference products, but they would not be under the purview of the CF Conventions.

Operational Suitability questions:

10. Do you currently use or plan to use CF conventions in a production setting? What types of applications do you use with CF Conventions? Does the metadata model work well with the data types and data manipulations in your application?
Yes, we use CF for some datasets (such as doi:10.3334/ORNLDAAC/909). We also use CF in the context of a THREDDS data server, where some of the datasets are CF-compliant. We are evaluating the role of CDL for some datasets and applications. We use CF for ESG in other parts of our operation. ORNL has a number of climate modelers who routinely generate data in CF compliant forms.

We do not typically consume CF files, so we are less familiar with the tools for reading and manipulating the data. We make some use of these as tests, to validate that the files will work with common tools, such as IDL.
11. Why do you choose to use the CF metadata conventions for your applications?
The ORNL DAAC provides terrestrial ecology data to a variety of communities, including the climate modeling community. CF is an important format to that user base. Further, a self-describing and persistent format is highly desirable from an archive-stable perspective, enabling future use and lower costs for future technology migrations.
12. Have you or your users encountered any difficulty when using some of the data access or visualization tools (e.g., IDL, GrADS, etc.) on files with CF metadata? If you have, please provide a brief description of your experience.
None to date.
13. Does the CF metadata conventions meet your requirements for discovering, accessing, providing interoperability of data and metadata? (e.g., *Can it handle the data types in your applications? Do you provide catalog services that utilize CF conventions?*)
Generally, yes, with the exceptions of otherwise mentioned needs for additional standard

names and discovery/collection-level metadata.

14. What operational challenges or limitations do the CF metadata conventions present? (*e.g., Does it take a long time to learn how to use it? Does it require advanced processing power, large amounts of memory, complex configuration, etc.*)

As noted above, CF is insufficient for discovery level metadata, particularly at the collection level. It is potentially appropriate to replicate some of this metadata down into the CF granules themselves, though this runs the risks of minor updates in collection metadata being confused with actual updates to data. Putting this metadata into the CF granules has the advantage of enabling the granules to stand alone, but has the disadvantage of potentially duplicating information, with all of the duplicates needing to be updated.

The ORNL DAAC THREDDS server (<http://daac.ornl.gov/thredds.shtml>) does provide cataloging information, and it is necessary to augment the CF metadata somewhat for discovery.

ORNL DAAC team members have also been involved in separately-funded efforts (from DOE) to integrate observational data into ESG. That work has highlighted further the need to potentially augment the granule-level metadata in CF files with collection and higher-levels of metadata, and the best practices for doing that are not currently established.

There is also a need for better tools (particularly for smaller datasets) for creating CF-compliant files.

15. What benefits do CF conventions present? Do the benefits of CF conventions outweigh the challenges? (*e.g., Do the conventions offer the flexibility you want to package the data types in your applications? Do they facilitate interdisciplinary studies?*)

The key benefits for us are:

a) The extensive use of CF in the modeling community means that making some fraction of our data available in CF makes that data much more immediately useful to the climate modeling community.

b) Broader use of CF enables better data integration, lowering the barriers for many people for data use, and enabling a wide range of client tools for data manipulation and visualization.

c) The format is self-describing, which has value for future archive evolution, as well as for current tool development.

d) netCDF-CF, as a specific implementation, is built on the netCDF framework, which is widely used throughout a number of communities, and there are a large number of tools already existing that can read/write netCDF files. That format is also generally well-described and has effective compression and high-performance access tools.

e) it is possible (though not always desirable) to use the netCDF dump utility to create an

XML file, which is then completely human-readable and CF compliant. This is of benefit from an archive preservation perspective.

16. How much data do/ will you provide using these CF metadata specifications? (*number of distinct data products or data sets, total data volume, number of files.*)

Of order 200 datasets (200 GB, tens of thousands of files) across the various archives in ORNL Environmental Sciences Division. For the ORNL DAAC, it is of order 50 datasets (30GB, 2000 files). Future volume could be in the millions of files and terabytes of data, but that is to be determined.

17. How many users and user-groups do you have or expect to have for data using CF metadata conventions, and what is your expected user community?

The ORNL DAAC has about 200 users of the THREDDS Catalog service. The MAST-DC project (also NASA-funded) has a comparable number, partially overlapping with the ORNL DAAC. Future use depends on the availability of tools and other factors. If we make more use of CF convention for data, the usage could grow to the majority of ORNL DAAC users, but many of those would only be using CF as a happenstance of ORNL DAAC usage, not because they necessarily want or need to use it.

18. (*User comments*) Any additional comments, observations or criticisms of CF metadata conventions and the RFC can be provided here.

The document ESDS-RFC-021v0.02 (Russ Rew, April 2010) is concise and very informative. We did not find a FAQ for CF conventions; this could have saved us time with some of our questions.

Overall, we see CF as a very positive convention and strongly encourage the active adoption of CF within NASA.