Maintaining and Advancing the CF Standard for Earth System Science Community Data

B.N. Lawrence[1], R. Drach[2], B.E. Eaton[3], J. M. Gregory[4], S. C. Hankin[5], R.K. Lowry[6], R.K. Rew[7], and K. E. Taylor[2].

(please send comments to b.n.lawrence AT rl.ac.uk)

Abstract

The Climate and Forecast (CF) conventions governing metadata appearing in netCDF files are becoming ever more important to earth system science communities. This paper outlines proposals for the future of CF, based on discussions at an international meeting held at the British Atmospheric Data Centre in 2005. The proposal presented here is aimed at maintaining the scientific integrity of the CF conventions, while transitioning to a community governance structure (from the current situation where CF is maintained informally by the original authors).

1. Introduction

The Climate and Forecast (CF) metadata conventions are designed to promote the processing and sharing of data stored in files created with the netCDF API[8]. Fundamental features of the conventions are that CF-aware software can automatically determine the space-time location of variables (facilitating analysis and graphical display), and metadata describing each variable is sufficiently detailed to determine whether variables from different sources are comparable. The general principles of CF design are enumerated in Gregory (2003)[9]:

1. Data should be self-describing. No external tables are needed to interpret the file. For instance, CF encodings do not depend upon numeric codes (by contrast with GRIB).

[1] NCAS/British Atmospheric Data Centre, Rutherford Appleton Laboratory, U.K.

[2] Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, U.S.A.

[3] National Center for Atmospheric Research, U.S.A.

[4] NCAS/Centre for Global Atmosphere Modelling, University of Reading, U.K.

[5] NOAA Pacific Marine Environmental Laboratory, U.S.A.

[6] British Oceanographic Data Centre, Proudman Oceanographic Laboratory, U.K.

[7] Unidata, University Corporation for Atmospheric Research, U.S.A.

[8] http://www.cgd.ucar.edu/cms/eaton/cf-metadata/ as of July 5, 2004, describing CF-1.0

[9] http://www.cgd.ucar.edu/cms/eaton/cf-metadata/clivar_article.pdf as of July 5, 2004, document dated November 6, 2003.

2. Conventions have been developed only for known issues. Instead of trying to foresee the future, features are added as required.
3. The convention should be easy to use, both for data-writers and users of data.
4. The metadata and the semantic meaning encoded through the metadata should be readable by humans as well as easily utilized by programs.
5. Redundancy should be minimised as much as possible (so as to reduce the chance of errors of inconsistency when writing data).

CF was initiated in the climate modelling community because of the lack of an adequate existing standard (although COARDS was a good basis, with which CF remains backward compatible). CF is becoming the de facto standard for storing outputs of atmospheric, ocean and climate models. Interest may expand to related "earth system" communities, but few users have thus far embraced the full self-descriptive capabilities enabled by the CF-standard, perhaps because software that can interpret all of this information has yet to be developed. Furthermore, "standard names" have not yet been agreed on for certain quantities, and in some cases the existing conventions are unsuitable or inefficient. To avoid fragmentation of the potential user community, such issues must be addressed more rapidly.

CF use is growing. As CF becomes more important to a diverse range of communities, its boundaries of applicability are being stretched, and the workload on the original authors is growing, such that it is not possible to keep pace with requests for extensions. There have been questions about how (and when) CF can evolve in the future, but it is not possible to answer such questions with confidence given the current situation where authorship of CF is a marginal activity on top of already full workloads, with the authors' personal interests representing only a part of the range of the user community. Moreover the widespread use of CF means that its development must be undertaken with great care and proper consultation.

This paper outlines some of the immediate issues that the CF community is confronting, beginning with how CF transitions to community ownership and maintenance. It is based on an extended discussion held at the British Atmospheric Data Centre in June 2005 as part of the annual Global Organisation for Earth System Science Portals (GO-ESSP) meeting.

2. CF Governance and Development

If CF is to become a widely accepted community standard, it is imperative that the community (a) takes ownership of and responsibility for the future trajectory of the standard and (b) invests substantial resources in its development. To do this we need first a clear definition of the boundaries of the problem (who are the CF community?) and then to set in place a structure for governance and development which:

- Preserves the existing intellectual rigor,
- Achieves rapid response to community requirements, and
- Meets the needs of as much of the community as is possible.

In some circumstances these three “what” requirements will pull in different directions. When this happens, one of the requirements of the governance structure is to impose a set of constraints on CF evolution that minimises this tension. To support the requirements we also need agreement about three “how” issues:

- How should CF maintenance be funded?
- How should CF be maintained from a sociological perspective? (What organizational structures are needed?)
- How should CF be maintained from a technological perspective? (What tools are needed?)

In this section we discuss the definition of “who is the CF community” and these three “how” issues in regard to both what we believe is achievable in the near term, and in terms of the three “what” requirements.

2.1 Who is the CF Community?

Climate modellers initially dominated the CF community, but, recently others (e.g., observational communities) have seized on CF as the solution to their own format description problems. As these communities have expressed their needs, new problems have arisen: for example, different terminologies are used, Earth geometry is real, rather than idealised, and standard names are needed for “raw” instrumental measurements before they are converted to data which are independent of the method of observation. Other problems can be anticipated, particularly as interdisciplinary work progresses. The scope of CF standard names, for example, will likely need to include observational biological oceanography, atmospheric chemistry, forest-type classification, etc.

Given that CF effectively consists of three sets of conventions - vocabulary management, semantic concepts (axes, cells etc), and format specific conventions (netCDF, for now), it is possible that different communities may wish to be involved in different parts of the CF evolution. For example, some users of CF are only interested in the CF standard names, storing their data in other formats (HDF-EOS, NASA-Ames etc), although this is problematic because some parameters need more than just the standard names to be fully described (e.g. the cell methods attribute describes the averaging); i.e., vocabulary and concept are not entirely separated.

For CF to fulfil its mission as a medium of interchange for complex information, the needs of users must be balanced with those of data producers. In addition, the input from developers of software that can interpret and use the CF metadata descriptions is vital. (It was noted at the GO-ESSP meetings in both 2004 and 2005 that only a subset of CF-defined metadata is interpretable and actually used by any real software.) Furthermore, despite the name, forecasters have not been greatly involved in developing CF.

The future of CF will inevitably involve use in earth system models, and so the scope of CF will need to address the requirements of a wide range of communities. Furthermore, CF does not exist in a vacuum. There are other projects that are developing community standards for encoding data and describing it: notably the Open Geospatial Consortium;

the WMO metadata initiatives; and controlled vocabularies in use in other parts of the earth system science community. A significant challenge for the CF community will be finding ways of interacting productively with these communities. Meanwhile a working definition of the "CF community" needs to be "those who choose to use netCDF to store earth system data in a self describing way" even though this initially disenfranchises those who wish to use CF conventions for other data formats.

2.2 Funding

There are four basic funding mechanisms that could support the maintenance and development of the CF conventions:

1. Institutional Subscription Mechanisms: Institutions could form a consortium, funded by membership subscription, which would fund specific activities (with the Open Geospatial Consortium as a possible role model).
2. National Subscription Mechanisms: CF could be sponsored by major international programmes (for example, the World Climate Research Programme) allowing national funding agencies to provide support in accordance with international commitments.
3. Benevolent Organisations: some key interested parties could dedicate staff time to the CF "project" as part (or all) of their normal duties.
4. Marginal Activities: interested parties working on CF in their "spare" time.

Currently all CF development is funded by the fourth of these options, which means that CF development is not responsive enough, and it is leading to untenable workloads for the key players.

Questions that have been raised concerning the first two funding approaches listed above include:

1. How would decisions be made to deploy the funding?
2. In the case of the CF Consortium: who would hold the cash? Could the financial management be separated from the scientific decision-making?
3. How would the CF community become involved with national programmes?

Given the right community management structure, a combination of funding approaches could be used, and indeed given where CF is now, a combination of these approaches is inevitable during what will have to be a transition phase.

In the near future an element of funding from benevolent organisations is expected. Expressions of commitment to support parts of CF development have been received from Unidata, from the Program for Climate Model Diagnosis and Intercomparison (PCMDI), and from the NERC Centres of Atmospheric Science (NCAS): Unidata may assist with the community interface, PCMDI will provide support for the CF development process, and the NCAS/British Atmospheric Data Centre (BADC) will be contributing direct support for standard name maintenance, development and evolution. However, even with this extra support, CF will not be able to evolve as it needs to without visible leadership, accountability and formal encoding of the process of governance.

2.3 Management

In keeping with the past development of the CF standard, the GO-ESSP participants strongly favoured a continued consensus-based governance procedure. Wide participation should be encouraged with input sought from individuals representing a variety of disciplines, perspectives, and geographical locations. Individual influence on CF's future should be commensurate with the value of suggestions made and independent of funding.

There is a strong desire within the existing CF community that, regardless of funding, the evolution of CF should be carried out with due process. Accordingly, we propose 1) that a collection of reference files, illustrating correct CF encoding and suitable for the testing of data-reading applications be maintained at every stage and 2) that CF modifications occur in a manner that conforms to the following sequence:

1. Requests for modification to CF are proposed to the community.
2. A period of community discussion is entered into, including consideration of the relationship of CF to other standards.
3. A provisional resolution is described as a consensual agreement.
4. Reference files incorporating the proposed features are produced.
5. Where the proposed modifications have non-trivial implications for implementation in software, trial implementations will be carried out. Trial implementations will be tested on both existing and new versions of the CF reference files, to check that new features can be correctly interpreted without detriment to backwards compatibility. A public evaluation and assessment of experiences will follow.
6. If problems emerge, the process is repeated from step 2; otherwise …
7. Consensus decision is made to accept the proposed changes as provisional changes to the standard and to advertise these changes to the community (indicating that advanced implementers may want to begin using them).
8. Multiple proposed changes are evaluated as a package in order to implement an orderly version control strategy.
9. Consensus decisions should finalise the release of new versions of CF.
10. The test reference files should then become part of the corpus of test data for future revisions.

The process will be different in detail for modifications (usually additions) to the standard name table on the one hand, and to the CF standard on the other. The former are more numerous, require faster turn-round, but generally do not involve wide community debate or trial implementations; an appropriate turn-round time is one month. For the latter, three months would be more realistic.

We also propose the establishment of two standing CF committees, the membership of which would be open to those with significant interest and time to commit to taking CF forward:

1. Conventions Committee, to be responsible for developing changes to the CF standard, and to include (but not be limited to) representatives of those who have reference implementations, who can provide feedback on the practicality of CF initiatives and validation of tools which wish to be described as "CF-compliant".
2. Standard Name Committee, to be responsible for adding standard names to the CF convention and working towards interoperability with other vocabulary maintainers.

Although CF must remain community-driven, its continuity can only be assured if some formally established board, responsive to community needs, assumes responsibility. Accordingly, it is also proposed that a CF panel be established under the governance of relevant major international programmes, which would be expected to appoint all or nearly all nominees to membership. The next section discusses the establishment of such a panel.

In view of their responsibility to the community, the membership of the committees will be advertised on the CF website. The job of the committee members will be to take an active interest in the community debate on new developments (process outlined above), participating personally where appropriate. On behalf of the community, the committees will take the required decisions within their respective remits by consensus among their own membership in cases where community consensus is not evident because public debate is inconclusive.

Each of the CF committees will include a permanent funded member of staff, viz. the manager of CF standard names and the manager of CF conventions. Both individuals must be skilled technical document editors, as their primary duty will be the maintenance of the CF standard documents. The latter will mostly likely be a software engineer, able to appreciate issues involved in both data processing and models. The former needs an understanding of modern metadata frameworks and a broad scientific interest, since to deal with requests for new standard names (often coming from those who are not themselves the scientists responsible for producing the data) it is necessary to understand what the quantities are and to get to become familiar with the terminology in various fields. If past experience is a guide, the development of CF will depend largely on these two individuals, who will be the first point of call for requests for modifications and who will carry the process forward by their own contributions to discussions and by deploying appropriate technology to facilitate community involvement. In particular, it will be part of their job to provide reference files and implementations. Although the responsibility will lie with the committees corporately, the other members will probably have only part-time involvement and the vagaries of their individual workloads should not be allowed to preclude timely progression of CF. The lack of anyone with time to carry out the role of manager of CF conventions over the last year or so, for example, has meant that several significant developments have been agreed on the CF email list but not yet implemented in the standard (extension to cell_methods, standard_name parameters, relation of formula_terms and bounds, relation of forecast and validity time).

In both committees we recommend that

- Insofar as possible, the principle of “no consensus, no change” should be followed.
- The chair should be elected to serve for a period of three years, although the membership should be self-selected. Where and when possible, the secretarial function should be funded.
- Insofar as possible, a geographical distribution of members should be encouraged

Other ad-hoc groups may be needed to consider various issues such as those listed in the Appendix. Such areas, while heavily overlapping with the CF core, may involve developing branches of CF which may initially be in conflict. Such conflicts will need to be resolved through a community discussion at the committee level.

2.4 Responsibility and Stewardship

As outlined above, a formal governance framework is also desirable. The World Climate Research Programme's (WCRP's) Working Group on Coupled Modeling (WGCM) and Working Group on Numerical Experimentation (WGNE) have been established to work together and with the international climate and weather modeling communities to establish "an integrated approach to climate modelling in the WCRP." Part of their charge is to promote coordinated experimentation, which requires sharing data among centres. For the sharing of climate model output, the format of choice has usually been netCDF, and in some recent WGCM-coordinated projects, the CF-conventions for metadata have been adopted (e.g., simulations in support of the IPCC's Fourth Assessment Report).

The WGCM has expressed interest in data standards for climate model output and has been kept abreast of the CF-developments. Recent discussions at the Ninth Session of the WGCM led to an invitation to the original CF authors to formally request appointment of a panel that under the WGCM would provide oversight for the governance of the CF conventions. It is therefore proposed that the WGCM establish a CF Panel charged with the following responsibilities

- Assure the viability and vitality of the Conventions Committee and the Standard Name Committee by formally appointing members from those self-nominated and by periodically reviewing the membership. This would also provide at least some formal recognition of the voluntary work done by committee members.
- Promote and help integrate CF across WCRP programs and the broader programs of WCRP's sponsors (ICSU, WMO, and IOC). In particular the panel would attempt to influence developing WMO metadata standards so that they gracefully accommodate the CF conventions.
- Encourage continued support of CF by benevolent organizations and explore additional funding mechanisms if necessary.

The WGCM will appoint members of the panel, which would specifically include (but not necessarily be limited to) representatives from: 1) the WGCM, 2) groups contributing significant resources in support of CF (as currently anticipated from Unidata, PCMDI, and NCAS), and 3) the chairs of the Conventions Committee and Standard Name Committee described in section 2.3 above. Although this panel would be responsible for stewardship of CF, it would not have any special responsibility for or influence on its technical content.

3. Next Steps

The two main priorities in the near future are to spread the load of CF management by making better use of technology (including website, issue tracking etc) to exploit the existing in-kind contributions, and to agree upon a strategy for establishing longer term funding stream to support CF activities. The WGCM's CF Panel needs to be established. At the same time, the community needs to buy into the future management structures outlined here (or suggest alternatives). The GO-ESSP meeting proposed a formal timetable to achieve that buy-in and move from the current situation to a new community managed framework:

1. Community discussion should follow the appearance of this paper.
2. One month after this paper appears, community discussion closes, and the existing CF authors decide whether the community response requires a different approach from that outlined here. Assuming the community feedback is not too complicated, this process should take at most a month.
3. If approved, a committee structure based on section 2.3 should be set up with initial chairs selected by the existing CF authors. Potential members should be invited to nominate themselves for appointment by majority vote of the CF authors and committee chairs.
4. Over the following month, membership of the committees should stabilise.
5. The first order of business of the new committees should be to define the changes in content for CF-1.1 and outline the priorities for further work.
6. The managers of CF standard names and conventions should be appointed as soon as funding has been secured.

Throughout this process, interested parties should lobby any or all national or international bodies that could provide ongoing funding for CF maintenance and development.

Appendix: Current and Future Issues for CF

There are a large number of issues which need addressing in the near future, some of which have the potential for requiring divergence in the CF conventions, and this is partly why an effective CF governance structure is needed.
Issues include:

1. **Supporting technology:** To help facilitate CF development, several suggestions have been made, including: 1) integrating the mailing list with forum software to allow better development of discussion threads; 2) assigning version numbers

to the evolving conventions to indicate both major and minor changes; 3) updating and improving the website, including documentation of the reasons for any changes or additions to the CF conventions; 4) developing further rudimentary tools to check for CF compliance and to encourage "best practices".

2. **Staggered and unstructured grids**: While the CF convention supports storing data on staggered grids, and indeed on irregularly spaced grids, it does it in a rather simple manner which permits the data to be used by any software - this is an advantage - but does not record any high-level information about the “construction” of the grids. This is not fully supportive of the wide range of new grids which are being proposed (e.g. the Yin/Yang grid) and may make efficient storage of the data relatively awkward. Proposals now on the table suggest for the first time that grid information may need to be stored in different files from that of the data which, while not strictly in accord with CF’s status as a netCDF convention, may be needed for practical reasons and is an approach already supported by some software. There is also an increasing need to describe unstructured hydrodynamic and watershed models which use meshes consisting of triangles, quadrilaterals and mixtures of polygons. Different issues arise, such as how to describe the indexing methods and connectivity of meshes, and there is a wealth of experience within the structures developed by engineering community to draw upon
3. **GIS information content:** Interoperability experiments are underway in a number of places, aiming to both deliver netCDF data into Geographical Information Systems (GIS) and to store data originating from GIS systems in netCDF. In both cases there are technical issues to resolve about how to deal with differing standard descriptions and how they map onto the other.
4. **Ontologies and nomenclature:** CF parameters are fully described by supplementing their standard names with additional descriptive information stored as attributes of the netCDF variables. Other ontologies categorise their information differently, e.g. by including coordinate information (such as “surface” or “daily maximum”) as part of the parameter description. Such differences complicate the establishment of equivalence tables. Existing standard names identify only those quantities normally produced or inferred from climate models; measured parameters (such as radar signal strength) are currently not supported. It has been suggested that CF could exploit the XML namespace concept to allow parameters to be included from named and linked external vocabularies (or dictionaries) which would help with extending the domain of CF. Such external vocabularies would have to be well managed, comprehensive, and have reliable futures to avoid risks associated with, incomplete, imprecise or duplicate definitions and atrophy. Where CF might come to depend on external vocabularies, it would be important to have establish clear understandings between external vocabulary maintainers and the CF community on how the sets of vocabularies would evolve.
5. **NetCDF-4**: CF-1.0 is based on the netCDF-3 data model, and to some extent reflects the limitations of that model. netCDF-4 introduces an expanded model, allowing, for instance, for ragged dimensions and hierarchical groups of data. The CF-1.0 data model will remain compatible with netCDF-4; however, it may be more natural to express some features of CF using the expanded model. Is it

appropriate or necessary to assume the netCDF-4 data model in subsequent revisions of CF?

6. ***In situ* observations (profiles, time series and trajectories):** The existing CF documents describe how station data might be encoded in CF compliant files in a way which is satisfactory for climate model output. For in-situ observations, it is potentially rather inconvenient and inefficient, and alternatives have been proposed, perhaps using new features of netCDF-4.
7. **Discovery information:** While CF describes the data encoded, it does not include "discovery information", that is, information which allows the user to discriminate between two otherwise identical datasets (for example, two model simulations). There are a large number of potential discovery content standards, and supporting the information to populate one or more of them in CF attributes would be desirable.
8. **Modularisation and compliance**: It is not yet clear what "CF compliance" means, even though it is clear that it will mean different things for datasets and software. Since a great deal of CF is optional, compliance cannot mean that all features are actually used. It has been suggested that CF could be broken into modules each of which could be version controlled, allowing software to claim compliance to specific components of the CF conventions rather than the whole edifice. Since archived data cannot easily be changed once written, backward compatibility for the conventions is a serious issue – should it be an inviolable principle (as it has been so far)?
9. **CF dialects:** As the diversity of uses of CF grows, the need may arise for "varieties" of the standard to be developed for different applications. Because most of CF is optional, this is not generally a problem; it simply means that metadata needed in one discipline will not be used in another. However, a more difficult issue arises where a structure which suits one variety of data is quite unsuitable in another, or where a mandatory feature is too onerous (for instance because it requires too much space for metadata storage). Since so many needs are in common, it would generally be better for interoperability and ease of maintenance to develop a single CF standard, but if it has incompatible dialects, the circumstances in which they should be used will have to be carefully defined.