Discussion topic summaries

These topics are also being suggested for discussion
- Publish CF Standard Names as an Ontology (Alison Pamment)
  https://github.com/cf-convention/discuss/issues/51
- International effort to create a data standard for chemical oceanographic data?

Standard names: general discussion on any aspect (Alison Pamment)

- The latest standard name table:
  http://cfconventions.org/Data/cf-standard-names/72/build/cf-standard-name-table.html
- The CEDA standard name editor: http://cfeditor.ceda.ac.uk/proposals/1
- A useful reference for background reading is Jonathan Gregory's presentation at the 2018 CF meeting: http://www.met.rdg.ac.uk/~jonathan/talks/CF180620.pdf

Summary of our discussion:

1. Q: How do we get people from other scientific domains involved in CF (given that it has expanded beyond the original Climate-Forecast community)?

   A: We should try to have a CF presence at more scientific conferences e.g. AGU, EGU, ocean conferences. It would be good to have some pre-prepared CF materials available for posters, presentations, training that we can all use as needed.

2. Q: Should we decide to include large sets of new standard names, e.g. to serve the ocean chemistry community, in the standard name table or should we point to external vocabularies as we are now doing for biological taxa?

   A: If a 'suitable' external vocabulary already exists then we should use it. By 'suitable' we mean something that passes FAIRness tests. See e.g. the work being done under FAIRsFAIR essential criteria(WIP) and RDA Interest Group VSSIG (Vocabulary and Semantic Services Interest Group). Every term in external vocabularies should have an individual URI that resolves to something!
   If there is no suitable external vocabulary then the CF community should do the work of agreeing the new standard names so that data producers can get on and write their data.

3. Q: We discussed the issues currently being raised in
   https://github.com/cf-convention/discuss/issues/51 ("Publish CF Standard Names as an
   Ontology").

   A: We think it is possible to have URIs for each standard name in CF "namespace", i.e.
   to include "CF" somewhere in the URI. This could be based on the terms served in NVS
   and would probably need redirection at DNS level. We will arrange a Zoom meeting to
   discuss the technical details and take this forward in the near future. Having a Turtle file
   on the CF website for use with ontology tools is a good idea, and would be even more
   useful if we can set up the redirection.

4. Q: Is it possible to view diffs between standard name versions in a simple way?

   A: At the moment there isn't an easy way to do this on the CF website or in GitHub
   because of the way the files are organised in the repository. A Python tool has been
   developed (by Harry Singh?) for finding diffs between XML versions of the standard
   name table and this could be advertised on the CF website. Also we could tidy up the
   way the standard names files are stored in the website repo so that it would be easier to
   see diffs using the features of GitHub.

   ● EUMETSAT has a tool for viewing vocabularies that is currently restricted to
     WMO formats. It's here: https://vocabulary-manager.eumetsat.int/. You can use it
     to view e.g. the differences between WMO BUFR & GRIB table versions, which
     are in the end pretty similar to CF Standard Names. - user driven tool - could
     request CF to be added

**Cell methods: "within"|"over" "days"|"months" and time axis (Section 7.4) (Lars
Bärring)**
https://github.com/cf-convention/cf-conventions/issues/197

There are two components:
● One is to clarify, and agree on a solution/interpretation of what the "climatology" attribute
  means and its relation to the cell methods "within" and "over": How are they connected?
  What does it mean for different typical cases? Can the "climatology" attribute be
  disconnected from the cell methods? Is in fact "climatology" necessary at all? What to do
  vs. existing CMIP6 data and what to do for CMIP7? What is the status quo, and what
  minimal changes may we want to make? Once these matters have been resolved this
  may come down to rather small changes to Section 7.4.

- The other one, is the more far-reaching need for a new or alternative mechanism that allows for a more flexible description of more complex and/or multi-step temporal processing of data.

Standard way to define subsampled coordinates (Daniel Lee)
https://github.com/cf-convention/discuss/issues/37

The volume of data products is increasing exponentially with the resolution of models and sensing systems. Increased spatial granularity, in particular, has led coordinates to contribute significantly to the cost of encoding, storing, and transmitting data products. As a result, data transfer and storage has become prohibitively expensive in some fields, even when state of the art data compression methods are used.

Data reduction has proven successful in significantly reducing data volumes, and thus costs. The greatest potential is in the field of remote sensing, where observational data is often not on a regular grid and thus explicit coordinates are needed for every observation point. Subsampling coordinate data while providing sufficient information to allow the user to fully recover it can reduce data volumes by up to 40% - *after* off-the-shelf compression techniques have been used.

We propose a method of providing subsampled coordinates to users that would significantly reduce the required data volume while still allowing a full reconstruction of all coordinate data. Our approach is a synthesis of methods used throughout the Earth Observation community and generalised for use in applications beyond remote sensing. It is possible to apply it along an arbitrary number of dimensions and on data that is regularly spaced or warped, as is the case e.g. in satellite observations with increasingly oblique viewing angles across the product. The integration of this method into the CF Conventions is inspired by grid mappings and compression by gathering.

At this stage the proposed method is still being refined and any input we can gather from the Community is greatly appreciated.

"mesh variable" instead of "boundary variable" for contiguous grid cells (Ryan Abernathey)
https://github.com/cf-convention/discuss/issues/5

Most numerical atmospheric and ocean models use some form of staggered curvilinear grid discretization (e.g. Arakawa grids), in which variables are located at some specific point (e.g. cell center, cell face) with respect to a finite volume quadrilateral (grid cell). CF conventions currently recommend encoding the geometry of these cells via the "boundary variable" convention. Boundary variables are extremely general, accommodating arbitrary unstructured,

non-contiguous cell geometry, requiring N boundary vertices to be specified for each grid point for N-dimensional geometry. For contiguous curvilinear coordinates, this representation is inefficient in terms of memory usage, due to the redundancy in the vertices of neighboring contiguous cells. For very large grids, this inefficient representation can become a bottleneck for analysis and visualization.

We propose to extend CF conventions with a new mechanism to represent such cell geometries which lies closer to how models represent their grids internally. We wish to incorporate the concept of a "mesh variable," which describes the boundaries of contiguous grid cells in terms of cell vertices. The sgrid conventions provide an excellent template for such representation, so we will propose to incorporate some or all of sgrid into CF proper. However, as discussed in the github issue, several thorny technical issues would benefit from community discussion.

Embedding provenance information - Metadata handling through processes (David Huard)
https://github.com/cf-convention/discuss/issues/33

When multiple netCDF files are aggregated to compute ensemble statistics, or when chains of algorithms are applied to netCDF files, recording operations and source metadata in the "history" attribute can become very messy. Recording such provenance information is necessary in applications where traceability and reproducibility are critical.

There are standards (e.g. PROV) to encode such provenance or data lineage information in machine readable formats. netCDF files could include a "provenance" attribute in which a machine-readable representation of the provenance would be stored. An example of this can be found at http://metaclip.org/, where provenance information is embedded in figure metadata.

Allow CRS WKT to represent the CRS without requiring comparison with grid mapping Parameters (Alan Snow)
https://github.com/cf-convention/cf-conventions/issues/222

CF guide reference:
http://cfconventions.org/cf-conventions/cf-conventions.html#use-of-the-crs-well-known-text-format

Proposed change:

There will be occasions when a given CRS property value is duplicated in both a single-property grid mapping attribute and the crs_wkt attribute. In such cases the onus is on data producers to ensure that the property values are consistent. *If both a crs_wkt and grid mapping attributes*

exist, the attributes must be the same. As such, information from either one (or both) may be used to represent the CRS of the file, recognizing that the grid mapping parameters should always be completed as fully as possible. If conflicts exist between the representations, you should inform the provider so they can be addressed. ~~However, in those situations where two values of a given property are different, then the value specified by the single-property attribute shall take precedence. For example, if the semi-major axis length of the ellipsoid is defined by the grid mapping attribute semi_major_axis and also by the crs_wkt attribute (via the WKT SPHEROID[…] element) then the former, being the more specific attribute, takes precedence. Naturally if the two values are equal then no ambiguity arises.~~
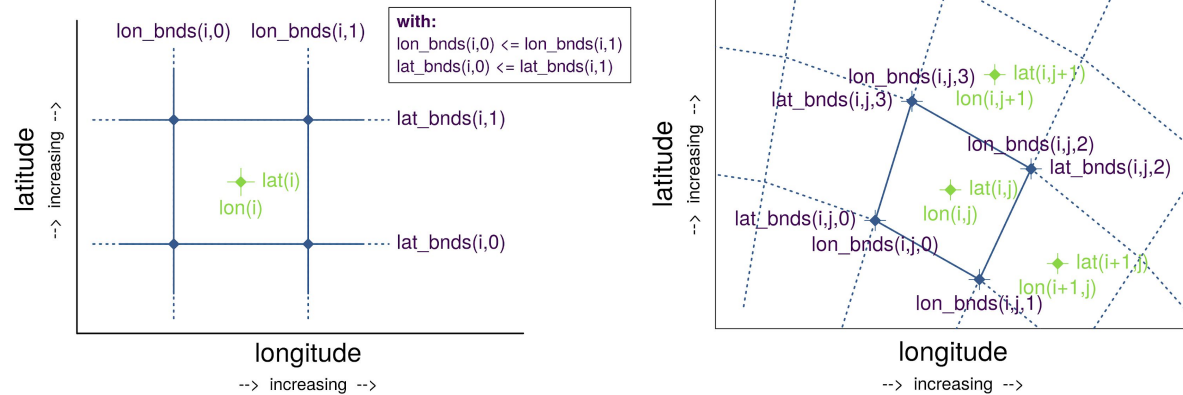
Benefits:

1. The CRS could originate from several different formats such as WKT, PROJ, or SRS Authority Code. If there are errors in the conversion process to the CF or WKT representation, only the provider would have the original CRS representation. As such, if there are conflicts, the provider would be the best source to go to in order to resolve the conflicts.
2. Making this change will simplify the lives of software developers so they can just read in the WKT or grid mapping CF parameters for the CRS without a need to compare the two.

Adding figure to paragraph "Bounds for 2-D coordinate variables with 4-sided cells" in Section 7.1 on bounds (Daniel Heydebreck)
https://github.com/cf-convention/cf-conventions/issues/193

The CF Conventions Section 7.1 "*Cell Boundaries*" contains a description on how to represent grid cell boundaries in `bounds` variables when the grid is defined by two one-dimensional coordinate variables (e.g. `lon(x), lat(y)`) and by two-dimensional auxiliary coordinate variables (e.g. `lon(x,y), lat(x,y)`). In the latter case (subsection "*Bounds for 2-D coordinate variables with 4-sided cells*"), the coordinates of the four vertices consist of four distinct values along each spatial axis (e.g. `lon_bnds(x,y,4), lat_bnds(x,y,4)`) and not of two distinct values as in the first case (e.g. `lon_bnds(x,y,2), lat_bnds(x,y,2)`). The order in which the coordinates have to be provided in the bounds variables is described in the text. However, it might be easier and faster to gasp by the reader by looking at a figure. The proposed figure contains this information.

lon_bnds(i,0)   lon_bnds(i,1)

with:
lon_bnds(i,0) <= lon_bnds(i,1)
lat_bnds(i,0) <= lat_bnds(i,1)

latitude   --> increasing -->

lat_bnds(i,1)

lat(i)
lon(i)

lat_bnds(i,0)

longitude
--> increasing -->

latitude   --> increasing -->

lon_bnds(i,j,3)   lat(i,j+1)
lat_bnds(i,j,3)   lon(i,j+1)

lon_bnds(i,j,2)
lat_bnds(i,j,2)

lat_bnds(i,j,0)   lat(i,j)
lon_bnds(i,j,0)   lon(i,j)

lat(i+1,j)
lon(i+1,j)

lon_bnds(i,j,1)

longitude
--> increasing -->

Moderation of proposals? (Chris Barker)
https://github.com/cf-convention/cf-conventions/issues/151

Both due to a maturing project and new tools it's a good time to discuss and perhaps update and/or clarify how proposals for changes to CF are managed. The CF conventions have been managed in gitHub for a while now, so that we now have a bit of experience with the system. It has worked fairly well in some cases, not so well in others.

We will review the current rules for convention changes:

http://cfconventions.org/rules.html, and discuss what's worked well, what has not, and mull over possible ideas from improvements.

This discussion relates to the following issues in gitHub:

#151 #130 #172

(most of which are closed, but provide context)