Climate and forecast (CF) metadata convention

Jonathan Gregory

CGAM, Department of Meteorology, University of Reading, UK
Met Office Hadley Centre, UK

CF has been developed by Brian Eaton, Jonathan Gregory, Bob Drach, Karl Taylor and Steve Hankin.

Goals

Locate data in space–time and as a function of other independent variables, to facilitate processing and graphics

Identify data sufficiently to enable users of data from different sources to decide what is comparable, and to distinguish variables in archives

Framed as a netCDF standard, but most CF ideas relate to metadata design in general and not specifically to netCDF, and hence can be contained in other formats such as XML

Basis of design

Intended for

use with climate and forecast data

atmosphere, surface and ocean

model-generated data and comparable observational datasets

Backward-compatible with COARDS so that

applications which understand CF can also process COARDS datasets

CF datasets will not break applications based on COARDS

Hence

COARDS is a subset of CF

where COARDS is adequate, CF does not provide an alternative

extensions to COARDS are all optional and provide new functionality

But some COARDS features are deprecated in CF

General principles

Data should be self-describing—no external tables needed to interpret it

Conventions have been developed only for things we know we need

Avoid being too onerous for data-writers and data-readers

Metadata readable by humans as well as easily parsed by programs

Minimise redundancy and possibilities for silly mistakes

NetCDF-specific principles

Avoid multiplicity of attributes

Information is generally provided per-variable, not per-file

Nothing depends on the names of variables (except the Unidata coordinate variable convention).

Origin of the data

CF provides some basic “discovery” metadata in global attributes

title	What’s in the file
*institution	Where it was produced
*source	How it was produced *e.g.* model version, instrument type
history	Audit trail of processing operations
*references	Pointers to publications or web documentation
*comment	Miscellaneous

*These ones can also be attributes of variables containing data

Description of the data

`units` is mandatory for all variables containing data other than dimensionless numbers. The units do not identify the physical quantity. `units` can be udunits strings *e.g.* `1`, `degC`, `Pa`, `mbar`, `W m-2`, `kg/m2/s`, `mm day^-1`
or COARDS specials `layer`, `level`, `sigma_level`.
Udunits doesn't support `ppm`, `psu`, `dB`, `Sv`.

`standard_name` identifies the quantity. Units must be consistent with standard name and any statistical processing *e.g.* `variance`.
Standard name does not include coordinate or processing information.

`long_name` is not standardised.

`ancillary_variables` is a pointer to variables providing metadata about the individual data values *e.g.* standard error or data quality information.

Numeric data variables may have `_FillValue`, `missing_value`, `valid_max`, `valid_min`, `valid_range`. CF deprecates `missing_value`.

Variables containing "flag" values need `flag_values` and `flag_meanings` to make them self-describing.

Standard name table

Currently 366 entries. Expanded on request. Many expected from PRISM.

Machineable XML or human-readable HTML with help text and guidelines.

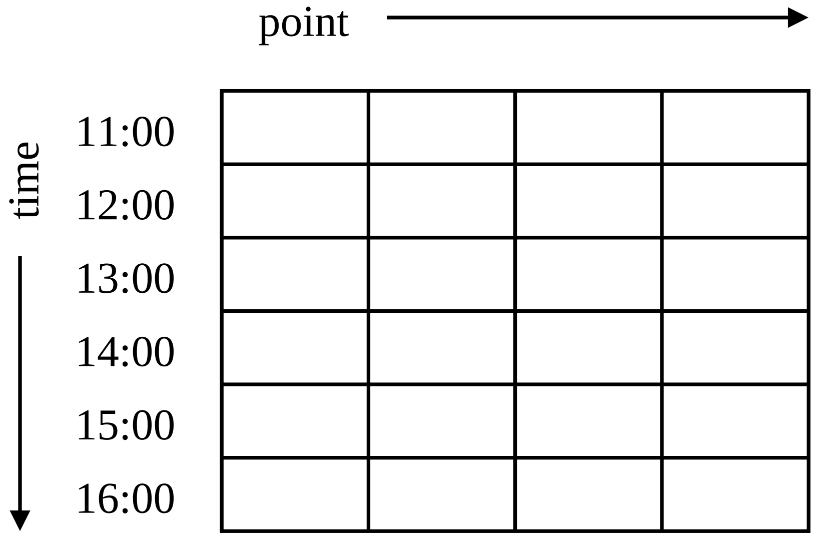
Units	GRIB	PCMDI	Standard name
K	13	theta	air_potential_temperature
1	71 E164	clt	cloud_area_fraction
kg m-2	79		large_scale_snowfall_amount
kg m-2 s-1			large_scale_snowfall_flux
m s-1			lwe_large_scale_snowfall_rate
K Pa s-1		mpwapta	product_of_omega_and_air_temperature
1			region
1	91	sic	sea_ice_area_fraction
1e-3	88	so	sea_water_salinity
W m-2		rlds	surface_downwelling_longwave_flux
W m-2		rls	surface_net_downward_longwave_flux

Dimensions and coordinates

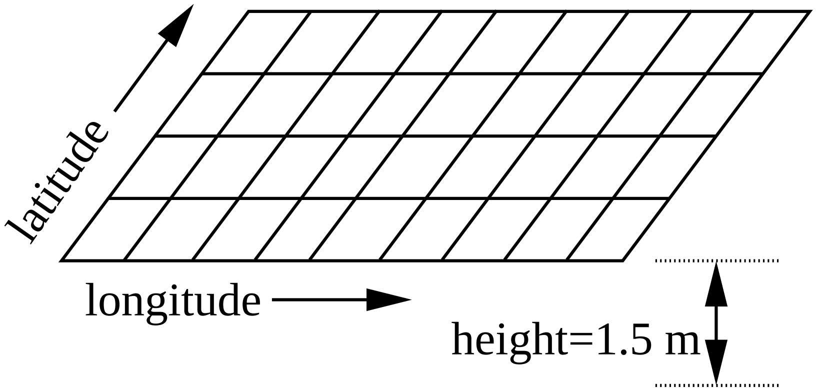
Dimensions establish the index space of data variables
e.g. `temperature(lat,lon)` has an element `temperature(46,0)`.

Coordinates are the independent variables, data the dependent variables
e.g. temperature is 252.2 K at 0°E and 25.0°S.

COARDS requires CDL dimension order `tzyx`. CF recommends it.



`point` is a dimension without coordinates



`height` is a coordinate without dimension or of size one

Variables containing coordinate data

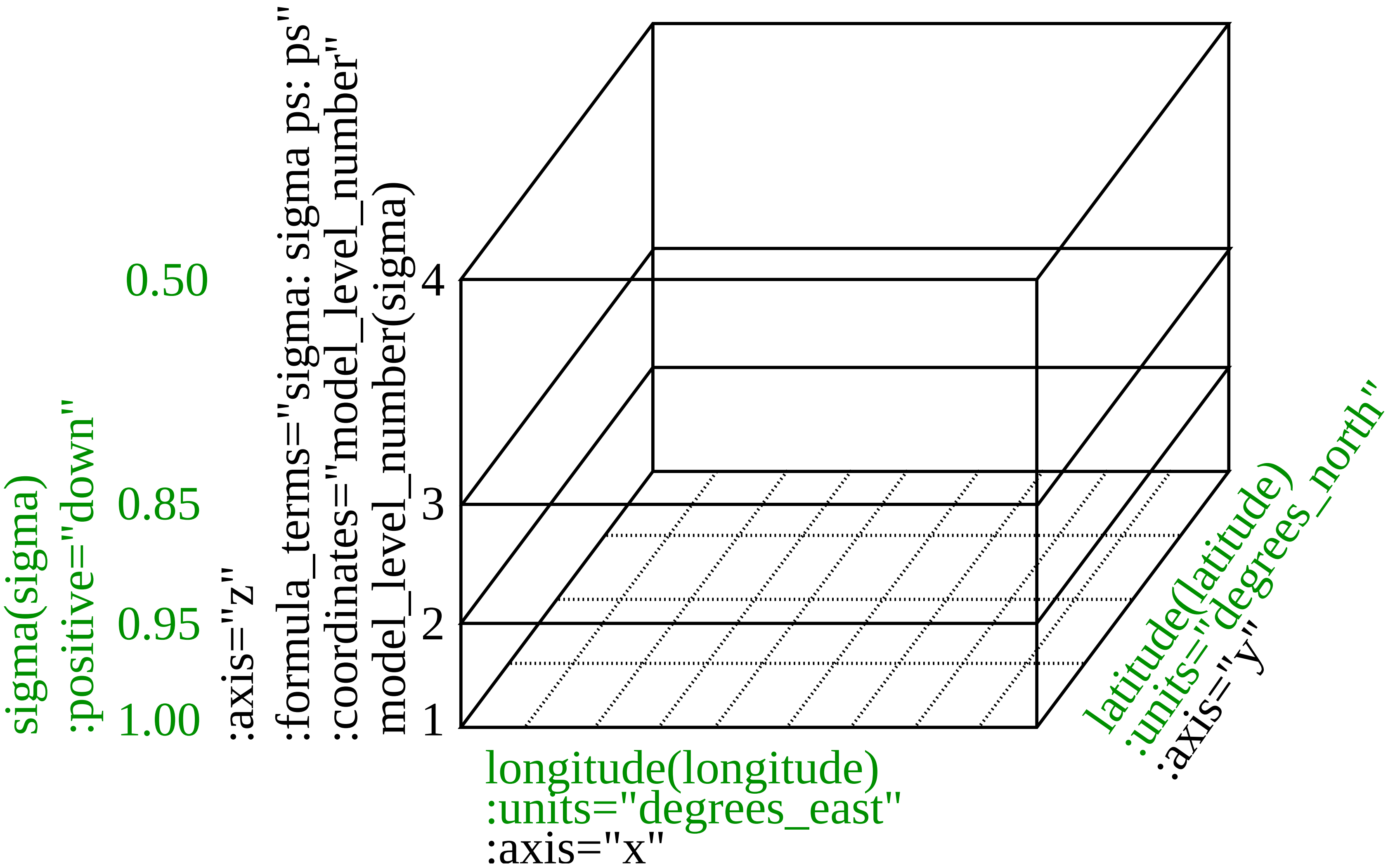
The Unidata netCDF standard associates one coordinate variable with each dimension, by identity of name *e.g.* `lat(lat)`.

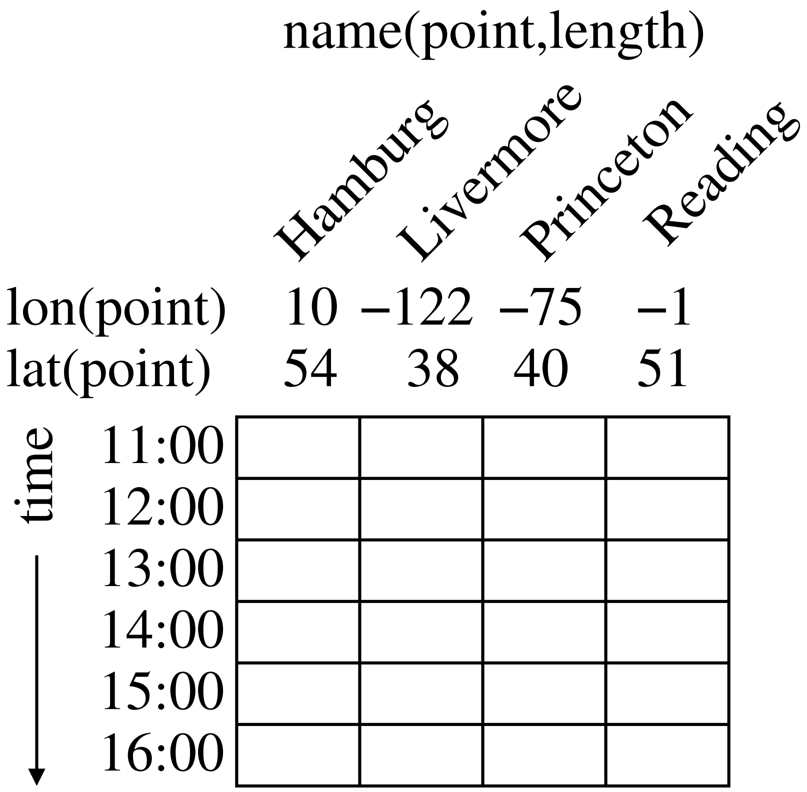
The coordinate variable distinguishes the elements along the axis
$\Rightarrow$ its values must be distinct. By convention they are strictly monotonic.

The CF standard introduces

- Scalar coordinate variables: zero-dimensional variables or single strings.
- Auxiliary coordinate variables: have any subset of the dimensions of the data variable, not necessarily monotonic.

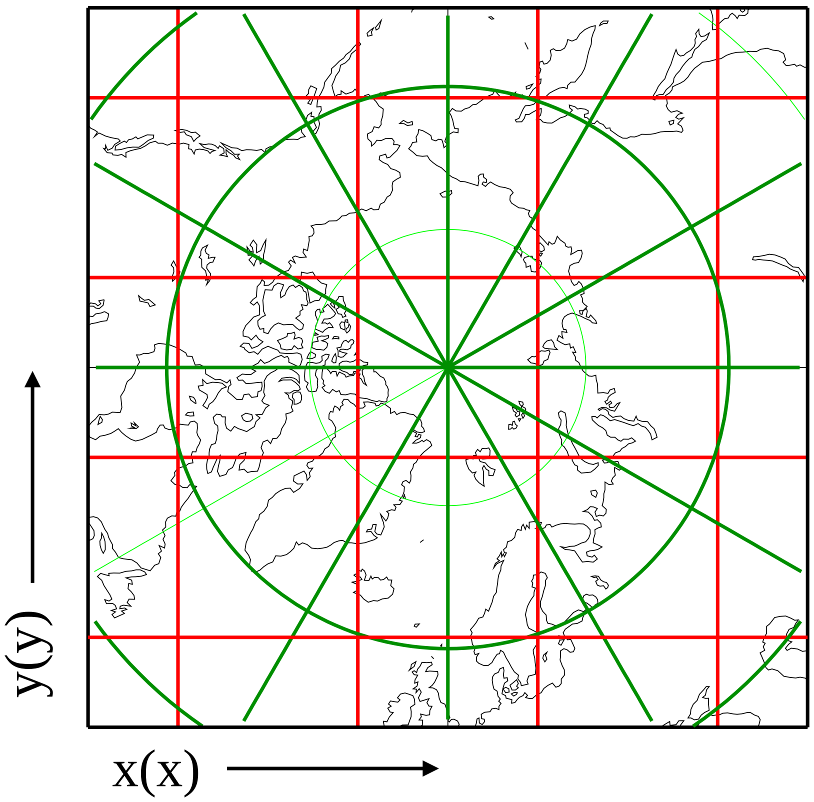
Associated with the data variable through the `coordinates` attribute.


sigma(sigma)
:positive="down"
0.50
0.85
0.95
1.00
:axis="z"
:formula_terms="sigma: sigma ps: ps"
:coordinates="model_level_number"
model_level_number(sigma)
4
3
2
1
longitude(longitude)
:units="degrees_east"
:axis="x"
latitude(latitude)
:units="degrees_north"
:axis="y"



	Hamburg	Livermore	Princeton	Reading
lon(point)	10	−122	−75	−1
lat(point)	54	38	40	51
11:00				
12:00				
13:00				
14:00				
15:00				
16:00				

float air_temperature(time,point)
:coordinates="lat lon name"



float air_pressure_at_sea_level(y,x)
:coordinates="lat lon"
:grid_mapping="mappinginfo"
float lat(y,x)

Time

Time (year, month, day, hour, minute, second) is encoded with units "*time_unit* `since` *reference_time*".

The encoding depends on the calendar, which defines the permitted values of (year, month, day). *e.g.* `2003-8-31` exists in the standard calendar, but not in the 360-day calendar.

```
                     36583.625                             in standard
2000-2-29 15:00:00 =            days since 1900-1-1
                     36058.625                             in 360-day
```

COARDS supports only the standard calendar.

Beware year and month units!

`year=365.242 days`, `month=year/12`.

If not default, the calendar is indicated by the `calendar` attribute of the time coordinate variable. Possibilities are

- `gregorian` or `standard` (the default)
- `noleap` or `365_day`
- `360_day`
- `proleptic_gregorian`
- `all_leap` or `366_day`
- `julian`

Also `none`, for perpetual season. Arbitrary calendars can be defined *e.g.* for palæoclimate in terms of month lengths.

Unfortunately udunits supports only the standard calendar at present. CDMS can handle others.

Perhaps `ncgen` and `ncdump` will be extended to support “formatted time”:

```
data:
  time=2000-2-29 15:00:00, 2000-2-29 16:00:00, ...
```

instead of

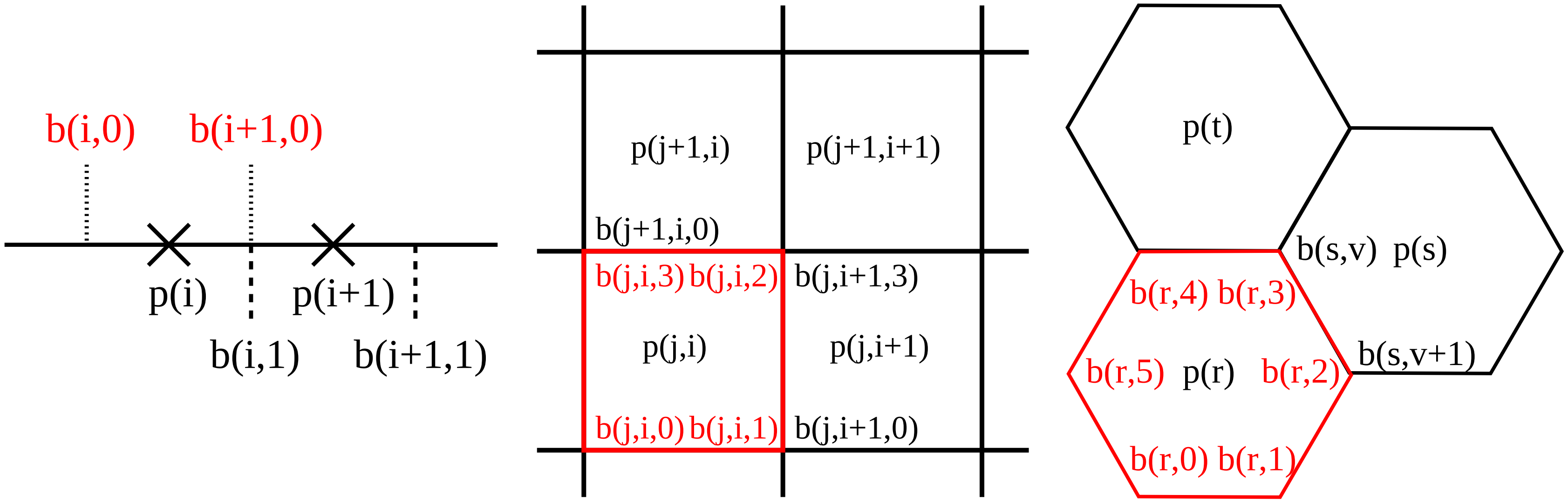
```
  time=36583.625, 36583.667, ...
```

Bounds

It is often necessary to know the extent of a cell as well as the grid point location *e.g.* area of lon–lat boxes, thickness of a vertical layer, length of a time-mean period.

Bounds can be particularly important for size-one coordinate variables.

Boundary variables can be attached to any variable containing coordinate data. They have an extra trailing dimension and assume all attributes of their parent. They can be used to test contiguousness of cells.



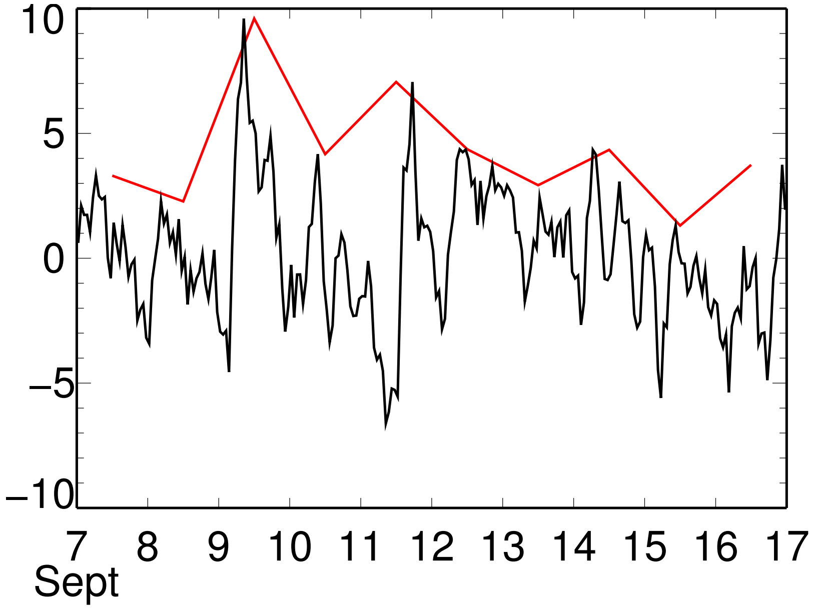
Cell methods

The `cell_methods` attribute of a data variable indicates how variation within the cells is represented. The method may be different for each axis.

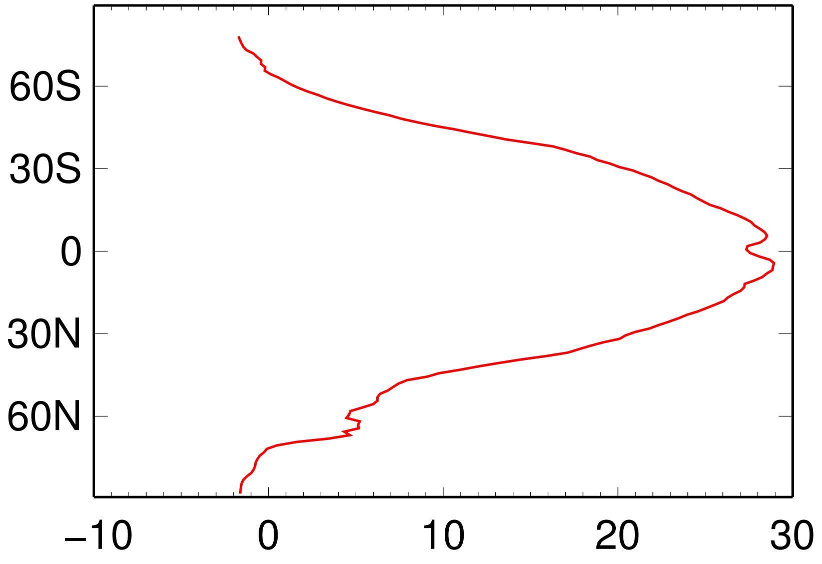
Default:

`point` for intensive quantities *e.g.* temperature

`sum` for extensive quantities *e.g.* precipitation_amount in time



:cell methods="time: maximum"



:cell_methods="longitude: mean"

Climatological statistics

Climatological statistics may be derived from

- Corresponding portions of the annual cycle in a set of years, *e.g.* average January temperatures in the climatology of 1961–1991.
- Corresponding portions of a range of days *e.g.* the average diurnal cycle in April 1997.
- Both concepts at once.

COARDS supports the first kind with the “year 0” convention.
CF deprecates this because (a) it doesn’t indicate the range of years, (b) there is no convention for how dates in year 0 should be encoded, (c) year 0 may be a real year in non-standard calendars.

Climatological statistics have a time coordinate with climatological bounds.

An interval of climatological time represents a set of subintervals which are not necessarily contiguous.

The `cell_methods` indicates the interpretation.

NB bounds are shown translated into (year, month, day, etc.)

Average winter-minimum temperature for 1961–1991

```
cell_methods="time: minimum within years time: mean over years"
climatology_bounds=1961-12-1, 1991-3-1
```

Average temperature for 1–2 a.m. in April 1997

```
cell_methods="time: mean within days time: mean over days"
climatology_bounds=1997-4-1 1:00, 1997-4-30 2:00
```

Monthly-maximum daily precipitation total for June 2000

```
cell_methods="time: sum within days time: maximum over days"
climatology_bounds=2000-6-1 6:00:00, 2000-7-1 6:00:00
```

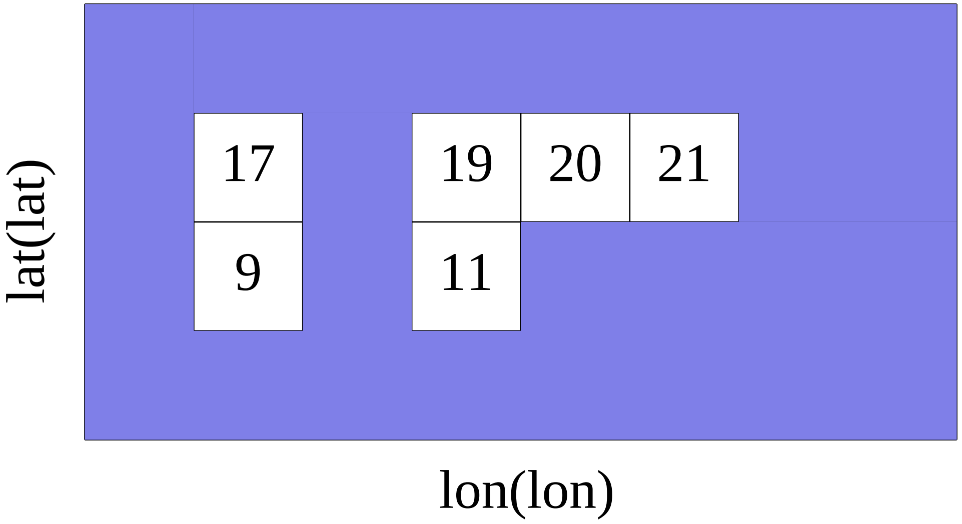
Reduction of dataset size

Packing (lossy):

packed_value = (*unpacked_value-add_offset*)/*scale_factor*

```
byte short int      float double
```

Gathering:



float data(lat,lon)

9	11	17	19	20	21

int land(land)

float data(land)
:compress="lat lon"

Built-in per-variable compression in netCDF would be useful!

Evolution of the standard

Non-beta CF-1.0 about to be released.
CF and the standard name table will continue to evolve.

Work will immediately begin on 1.1, to include:

- statistical operations which combine data variables *e.g.* covariance.
- handling of forecast and analysis time.
- more grid mappings.
- support for spectral representation.

Currently no recognition of:

- related horizontal grids *e.g.* Arakawa B-grid, C-grid.
- relations between vector and tensor components.

Do we need these?

CDML and NcML introduce some features and concepts not present in CF.

CF in the real world

CF has a web site
`http://www.cgd.ucar.edu/cms/eaton/cf-metadata`
and an email list for requesting and discussing extensions.

Adopted by:
PCMDI and *MIP, PRISM, ESMF, NCAR, Hadley Centre, GFDL, various EU projects.

Supporting software:
CF-checker. Planned PCMDI output routine. VCDAT/CDMS. NCO? Ferret? Simple utilities?

CF has been developed by volunteers. It may become too much for us.