CF meeting
9-11 June 2020
Zoom session link

Discussion topic notes

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.

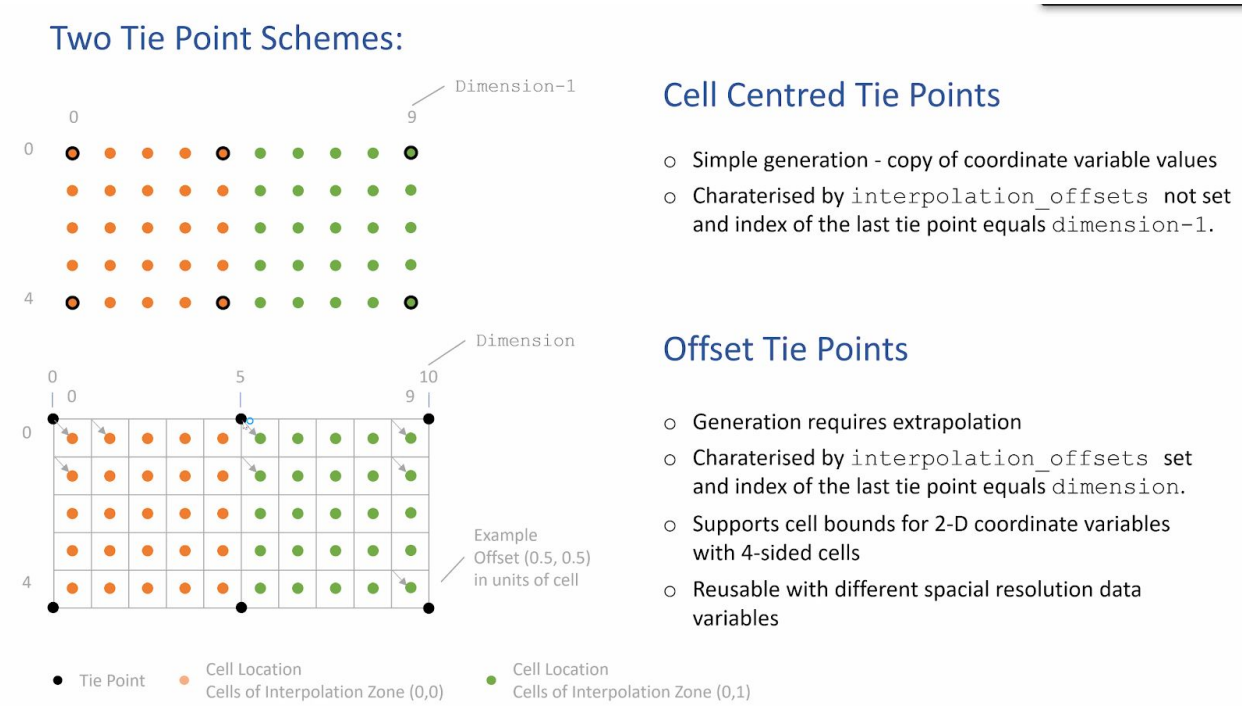
At this stage the proposed method is still being refined and any input we can gather from the Community is greatly appreciated.

Sylvain Herlédan, Alice Bui, Aleksandar Jelenak, Lucile Gaultier, Anders Soerensen, David Hassell, Daniel Lee, Xiang Li, Luc Bujold, Harry Singh

---

Discussion:

- Aleksandar: Are cell offsets part of data variable or should they be on the interpolation container?
    - Anders: There are advantages of doing this per variable because the offsets might be different for different data sets. This also applies to the index mappings. To illustrate:



    - Daniel: The use of offsets needs very precise description in our final proposal. We may consider describing the units for offsets as not to be "cells" but "array positions". The utility is very high because it allows reuse of the tie-points for multiple variables.
- David: Should the lats and lons be attached to the data variable or the interpolation container? This is an important point.
    - Sylvain: If you're only interested in the footprints, there is a concern if the tie points are only attached to the data variables. So you'd need to go "through" the data variable in order to recover the coordinates.
    - David: CF doesn't see coordinates as independent of data variables, and also the array indices are stored in the data variable so that would be another barrier.
- David: I wonder if the interpolation containers need to be mentioned in the context of their dimensions: "`m_radiance:interpolation = m_track m_scan: tp_interpolation m_track m_scan: interpolation time`". In any case, how do we know which indices and offsets apply to which interpolation containers?
    - All that said, I'm a bit confused about the time interpolation in the example: which is the time dimension of the data variable "m_radiance"? Is the "time_scan" variable missing?

Conclusions

- Benefit of proposal is demonstrated.
- Resume biweekly meetings
- Careful consideration needed about what aspects of the approach are considered "immutable" and thus should be stored on the interpolation construct and which should be considered specific to a given data variable