Planning the next GHRSST Data Specification Document

Chair/Book captain: Jean-François Piollé (Ifremer, France)

Co-chair: Ed Armstrong (JPL/NASA, USA)

Presenter: Jean-François Piollé

Summary

Collaborative editing and reviewing process

Planning evolutions and sharing the work

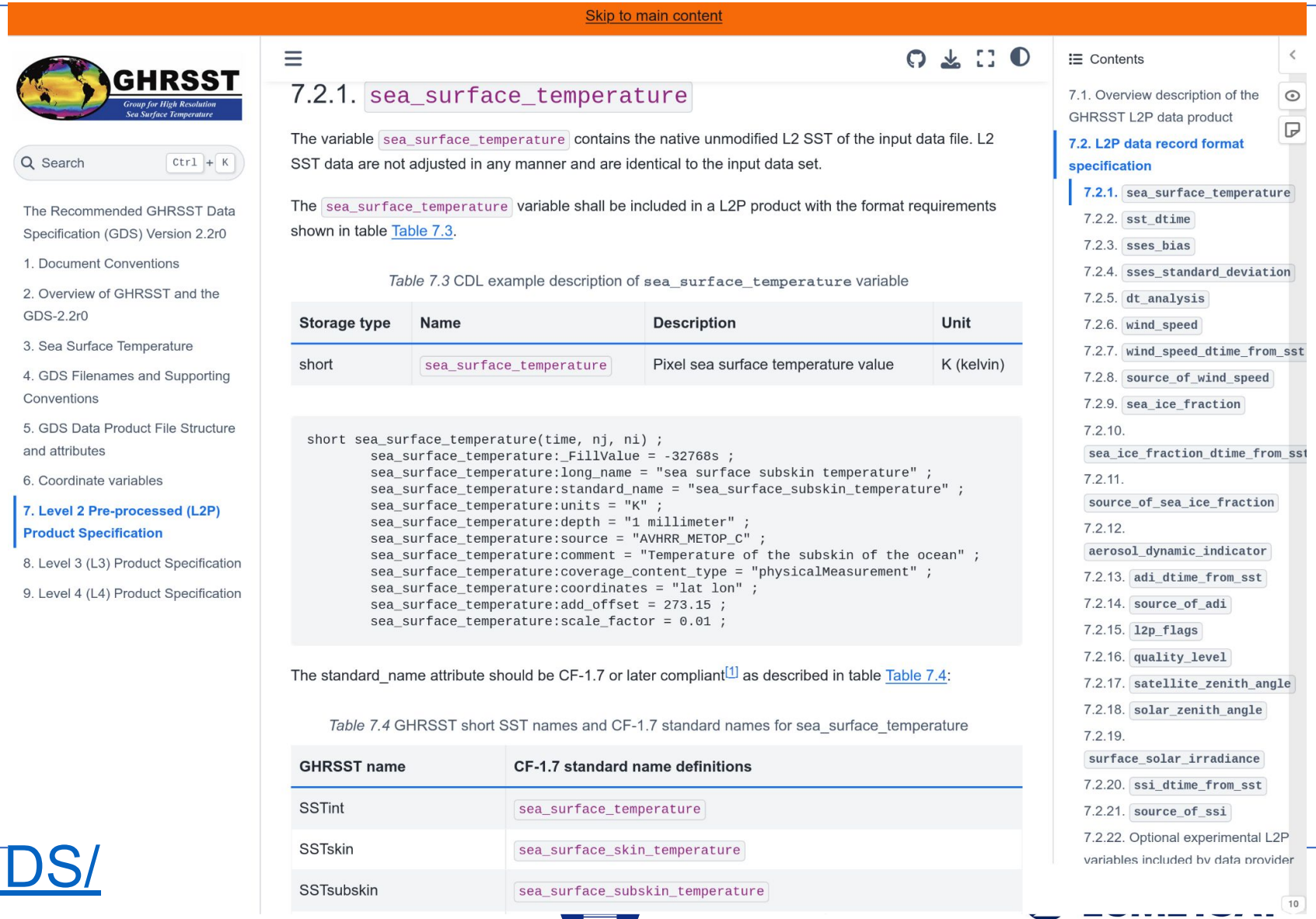
Editing the GDS - the old way

- GDS document became hardly manageable
  - big word document, not editable in libreoffice, gets mixed up through various word editors or human editors, losing content, images, … hard to recover and clean
  - inserting ncdump or code examples is a nightmare - does not help for maintaining consistency
  - history of main changes and versions not properly tracked
  - not practicable for collaborative work => one people to maintain and fix it
  - need to engage GHRSST producers more during inter GHRSST periods
  - updates needs to happen online - we don't want to discuss this in GHRSST workshops
- from user perspective
  - not easy to navigate, difficult to quickly go to relevant examples and templates
  - usage does not fit "sequential" structure of document, lack of links and jumps to cross-references
  - many inconsistencies (between different sections, dead links, obsolete content, examples are not always updated,...)

The new way : jupyter book

- adopting jupyter book (consensus G25)
- generation and publication of document based on release tags
- possibility to revert to or re-generate any older (tracked) version
- avoid formatting and cleaning issues (fonts, …)
- sharing the document, full traceability of changes
- conversion for G26

https://ghrsst.github.io/GDS/



Skip to main content

GHRSST Group for High Resolution Sea Surface Temperature

Search Ctrl + K

7.2.1. sea_surface_temperature

The variable sea_surface_temperature contains the native unmodified L2 SST of the input data file. L2 SST data are not adjusted in any manner and are identical to the input data set.

The sea_surface_temperature variable shall be included in a L2P product with the format requirements shown in table Table 7.3.

Table 7.3 CDL example description of sea_surface_temperature variable

Storage type	Name	Description	Unit
short	sea_surface_temperature	Pixel sea surface temperature value	K (kelvin)

```
short sea_surface_temperature(time, nj, ni) ;
        sea_surface_temperature:_FillValue = -32768s ;
        sea_surface_temperature:long_name = "sea surface subskin temperature" ;
        sea_surface_temperature:standard_name = "sea_surface_subskin_temperature" ;
        sea_surface_temperature:units = "K" ;
        sea_surface_temperature:depth = "1 millimeter" ;
        sea_surface_temperature:source = "AVHRR_METOP_C" ;
        sea_surface_temperature:comment = "Temperature of the subskin of the ocean" ;
        sea_surface_temperature:coverage_content_type = "physicalMeasurement" ;
        sea_surface_temperature:coordinates = "lat lon" ;
        sea_surface_temperature:add_offset = 273.15 ;
        sea_surface_temperature:scale_factor = 0.01 ;
```

The standard_name attribute should be CF-1.7 or later compliant[1] as described in table Table 7.4:

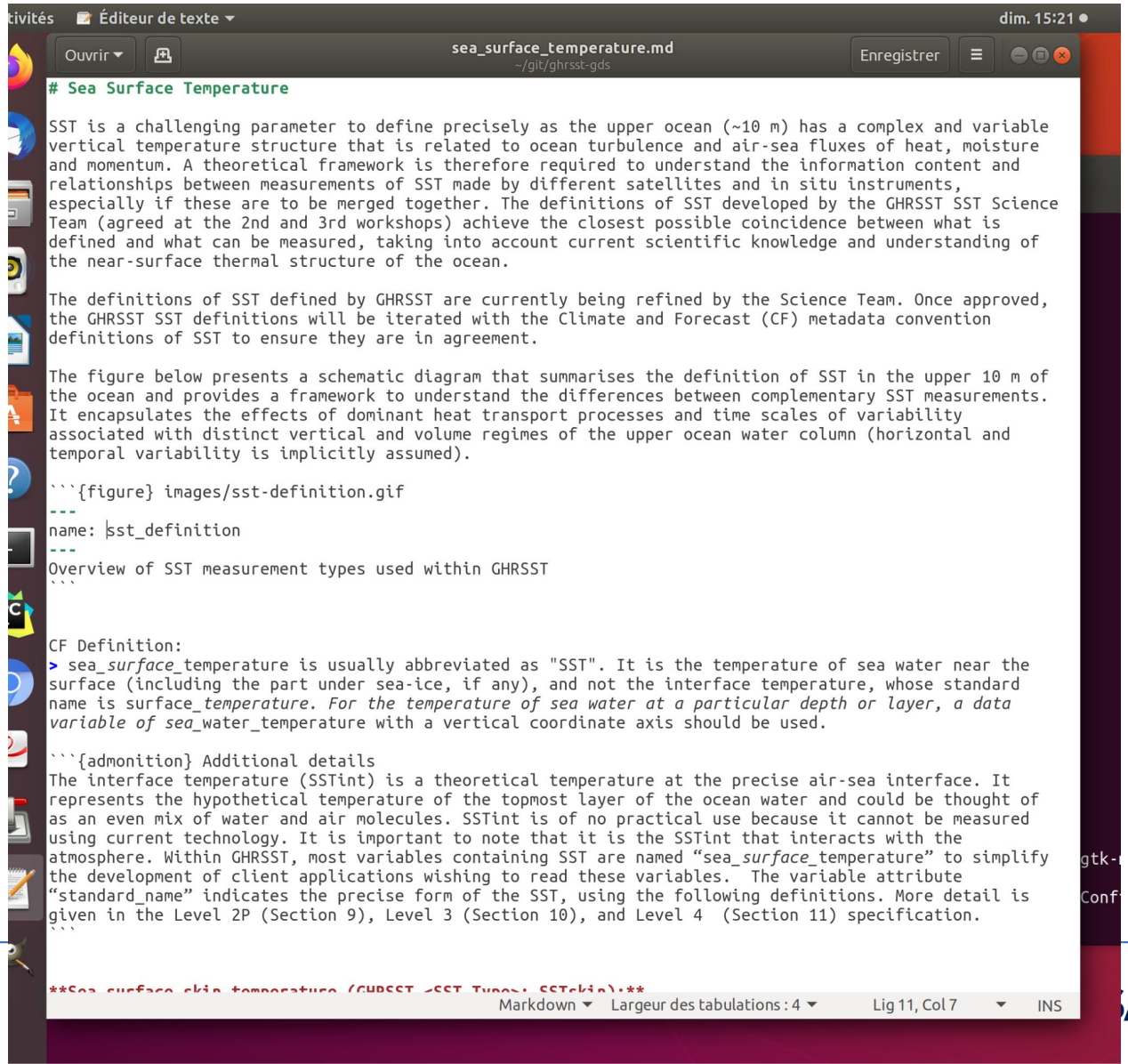
Table 7.4 GHRSST short SST names and CF-1.7 standard names for sea_surface_temperature

GHRSST name	CF-1.7 standard name definitions
SSTint	sea_surface_temperature
SSTskin	sea_surface_skin_temperature
SSTsubskin	sea_surface_subskin_temperature

Contents

editing jupyter books

```
├── _config.yml
├── conventions.md
├── coordinates.md
├── generators
│   ├── l2p.py
│   ├── l3s.py
│   ├── l4.py
│   ├── nonregular_grid_coords.py
│   └── regular_grid_coords.py
├── images
│   ├── ghrsst-logo.png
│   ├── regridding1.png
│   ├── regridding2.png
│   ├── rgts1.png
│   ├── rgts2.png
│   └── sst-definition.gif
├── intro.md
├── l2p.md
├── l3.md
├── l4.md
├── logo.png
├── naming.md
├── overview.md
├── README
├── references.bib
├── requirements.txt
├── sea_surface_temperature.md
├── structure.md
└── _toc.yml
```



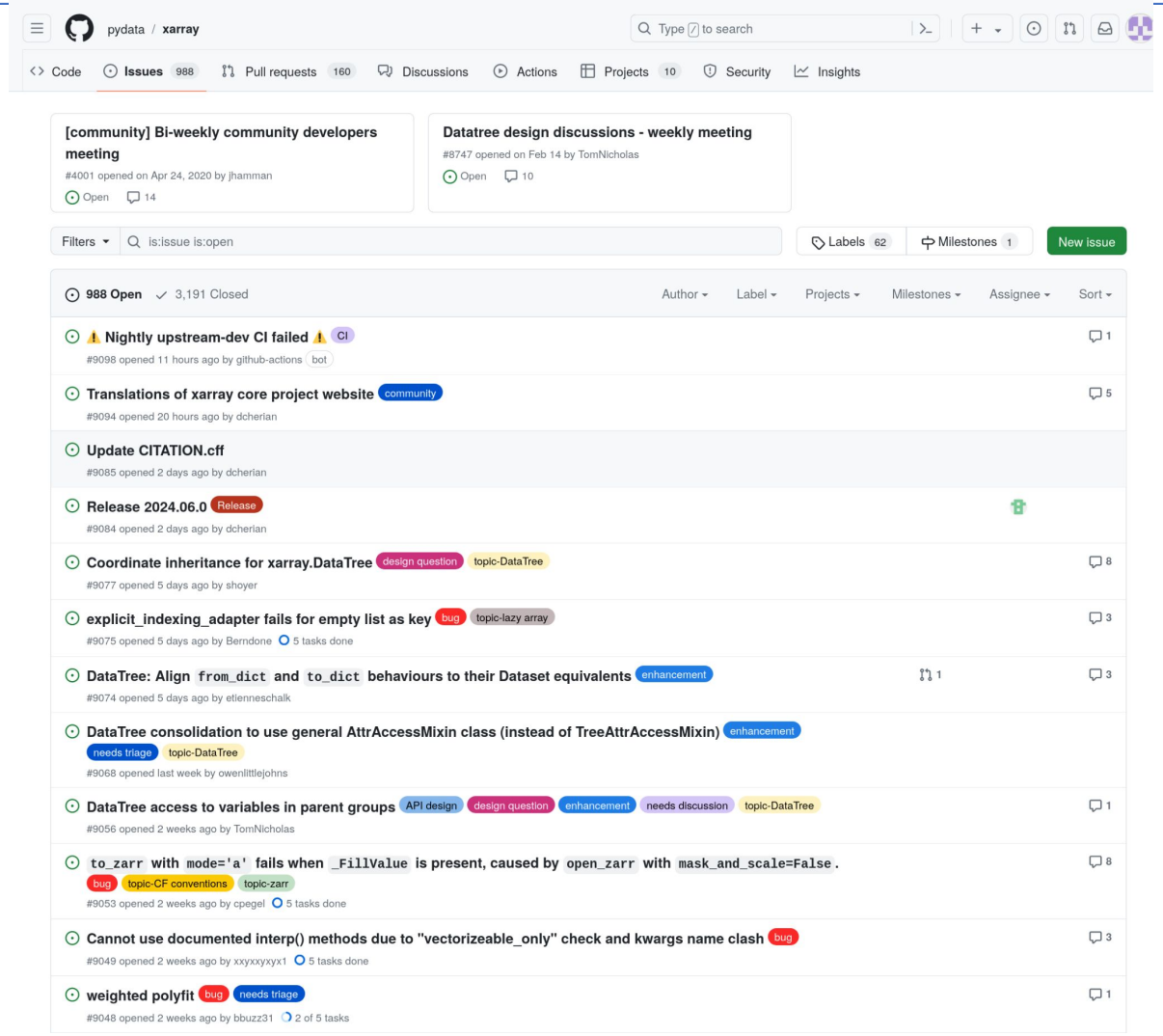
Proposed new editing process

manage changes and revisions "as a open-source software" under github : https://github.com/GHRSST/GDS

github issue tracking to discuss the main changes, assigning sections/updates to editors - can be related to GDS change commits

use **branches** to implement specific new feature or section (e.g. IST, uncertainties,...)

pull request to merge implemented feature



organization of collaborative editing

open to anyone (producer or user) interested (not only ST member) - leverage on Task Teams

get a github account

get permissions and status (contributor, reviewer,...) from GPO (?)

create an issue per new feature - main leader

create new branch for documentation update

exchange through issue tracking system, telecon,...

interact, commit, update, …

pull request when consensus

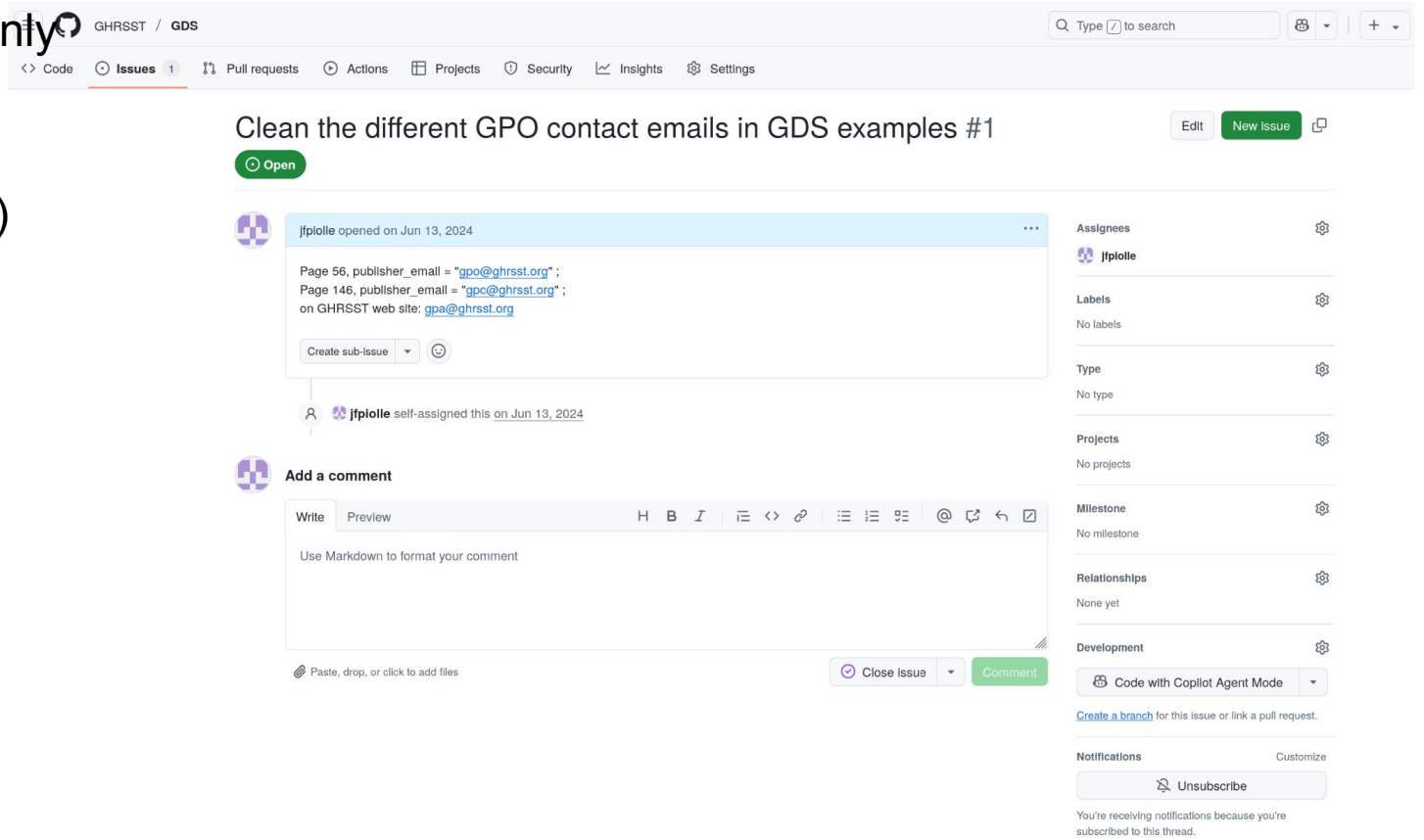
review from book captain(s) merge into master document (GDS v3)

update github pages

review created pages

fix, iterate,...

summary of changes



New GDS features

1. Ice Surface temperature (IST)

More datasets provide both IST and SST

Existing draft specification by EUMETSAT, OSI SAF

Lead: High Latitude TT

2. Spatial and temporal coordinates

Splitting **time** in two variables (*time*+*sst_dtime*) is not really needed for L2P (possibly for L3 to stack data into a cube) => one unique full precision *time* variable ?

better describes projection for gridded datasets (CRS,...)

cell bounds and cell method attribute for averaged pixels (L3, L4)

Coordinates for ultra high resolution :
should we still store full lat/lon matrix ? already optional for reprojected data.
Include specification on GCP/Tie points?
https://cfconventions.org/cf-conventions/cf-conventions.html#compression-by-coordinate-subsampling

Documenting better what lat/lon refer to exactly, what is the accuracy ? what is the accuracy

Lead: TT Coastal SST ? engage agencies (TRISHNA, LSTM,...)

3. Data scaling and quantization

NetCDF quantization (**`scale_factor`** and **`add_offset`**) is confusing for users and source of problems;

- clipping or overflow for saturated values (beyond the possible range, positive values becoming negative,...)
- decoding errors / some people implementing hard coding of unscaling preventing from changing the scale
- was used to reduce file size (float variables)
- NetCDF4 now provides internal compression and number of significant digits
- Cons: number of significant digits result in non “round” numbers

free data storage types for integers (and floats? or not ?)

Removing NetCDF scaling of variables and using true data type (e.g. float instead of integer)

Lead: R/G TS TT

4. Additional fields

ancillary variables:

- total water vapour content
- aerosols : not often provided… need better requirement ?

additional free fields : lift off limits on extra variables

more traceability ?

- processors and processing steps and settings
- used inputs (for instance in L3S/L4)
- **Lead** : L4 TT ?

5. Groups

NetCDF4 introduced **groups of variables**, e.g. in a single file:

- gather related variables in different groups (**Core** : SST + quality + uncertainties, **ancillary** or **experimental** fields, ….)
- several core groups possible, per **sensor view** (e.g. nadir / oblique) or **SST type** (e.g. skin, depth,...) with their own quality_level, uncertainties, etc…
- issue: seamless usage of SST in applications and tools (we need a **primary SST field with consistent naming**)
- gather into different groups variables with **different coordinate systems**, e.g. multi-resolution L3S :
  - separate microwave (0.25°), geostationary (0.05°) and polar orbiting (0.02°) and combined (0.02°) multi-mission grids

6. Uncertainties

- replacing SSES_bias and SSES_standard_deviation
- simple unique uncertainty (**total**) vs separation of random and systematic uncertainty components
- metrological approach ? can it be implemented ?
- specific uncertainties to be defined per processing level (L2P, L3, L4)
- agree first on terms and content (variables) before actual how to ?

Lead ? L4 TT for L4?

7. Regridding

regridding rules - is it followed by all producers ? Is it a "specification" ? is a consensus reachable ?

Several presentations / posters on this topic

Attached uncertainties

Lead: ?

8. Format

New formats used by cloud providers or agencies

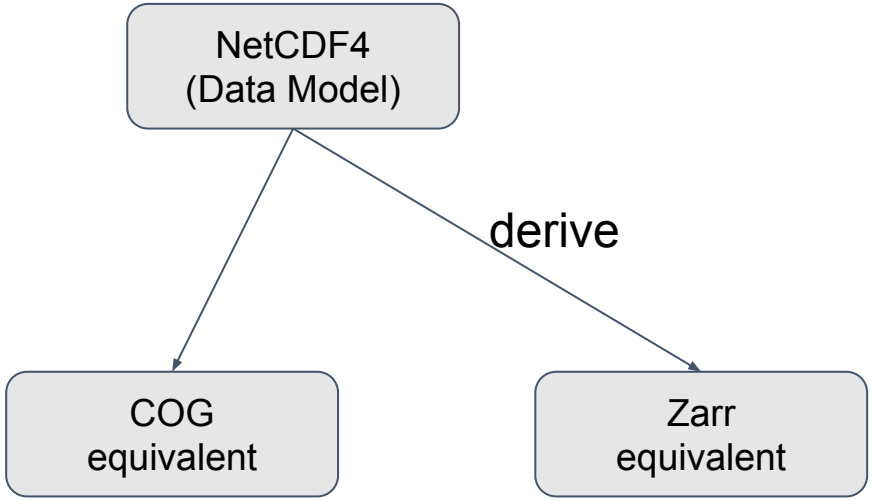
How to ensure GHRSST specification model heritage through new ways of data distribution

keep NetCDF4, change reference format or define equivalencies, conversion rules (and accept different formats)

NetCDF data model is applicable to other formats and could remain the reference description

Unsure about benefit of COG/Zarr over NetCDF for L2P, Zarr emulation is an alternative for gridded data

Lead : R/G TS TT



9. Cloud optimization

Improving NetCDF formatting rules

https://docs.google.com/presentation/d/1CT0pT0UJkUNUkyf2TyW8Mvuxe6962zkj/edit?usp=sharing&ouid=103571070828337788870&rtpof=true&sd=true

10. Fixing GDS issues - Filenaming

Currently:

<Indicative Date><Indicative Time>-<RDAC>-<Processing Level>_GHRSST-<SST Type>-<Product String>-<Additional Segregator>-v<GDS Version>-fv<File Version>.<File Type>

- **<SST Type>** = SSTfnd, SSTskin, SSTsubskin, SSTdepth
  - missing IST, ISST
  - accounting for multiple SST type fields (e.g. skin + depth)
- **<Product String>** = **<platform code>_<sensor code>(_<additional text if needed>)**
- relax requirement on using CEOS mission names, e.g. allow usage of both operational and early mission names (MTG-I1 <> Meteosat-12, Metop-SG-A1 <> Metop-D), or abbreviated names (build reference table?)

11. file_quality attribute

A code value:
0 = unknown quality
1 = extremely suspect (frequent problems, e.g. with known satellite problems)
2 = suspect (occasional problems, e.g. after launch)
3 = excellent (no known problems)

do many producers care about this attribute and is it used ? usually been filled with 3, except for some producer on the first months of a new satellite (2).

rationale for setting this information should be clarified

is this necessarily over a long stable period (manually) or can it be automatically calculated (for example downgraded to 2, from the L1 over_all_quality_flag) ?

If the file quality is automated, should file quality also be lowered if one PDU is missing ?

Maybe better to replace with a percent_missing_data or similar field if this field can be dynamically calculated by the data producer

R/G TS Evolutions

- Data access services
  - file aggregation / subsetting time series : future of THREDDS? replacing with “NetCDF (or whatever) on the cloud”

- STAC in complement / replacement to Opensearch for file inventory / selection

way forward

- finalize editing process and guidelines (GPO with R/G TS)
- list of issues
- enlist contributors
- aim at v3 draft to be reviewed at next GHRSST meeting