

Proceedings of the 2024 CF Workshop

17-20 September, Norrköping, Sweden



Editors:

Lars Barring¹, Ellie Fisher², Alison Pamment³,
Francesca Eggleton⁴, Daniel Lee⁵, David Hassell⁶,
Ethan Davis⁷, Kevin O'Brien⁸, Antonio S. Cofiño⁹

1 <https://orcid.org/0000-0001-7280-2502>

2 <https://orcid.org/0000-0002-5204-7615>

3 <https://orcid.org/0000-0001-5040-4626>

4 <https://orcid.org/0000-0002-8329-7932>

5 <https://orcid.org/0000-0002-1041-232X>

6 <https://orcid.org/0000-0001-5106-7502>

7 <https://orcid.org/0000-0002-8530-2213>

8 <https://orcid.org/0000-0003-4167-3028>

9 <https://orcid.org/0000-0001-7719-979X>

Table of Contents

SMHI web page announcing the Workshop.....	3
Workshop agenda.....	5
Presenters for Workshop part 1.....	7
Opening Session - Notes.....	8
Vocabularies Session - Notes.....	9
Units Session - Notes.....	12
Uncertainty Session - Notes.....	15
Metadata for describing statistical processing - Notes.....	20
New Technologies Session- Notes.....	24
Cross Format Issues Session- Notes.....	28
Cross Domain Session- Notes.....	31
Synthesis Session - Notes.....	36
Overview of CF news and introduction to hackathon sessions.....	39
Good Housekeeping Hackathon - Notes.....	46
HEALPix Hackathon - Notes.....	48
CF Roadmap Hackathon - Notes.....	49
BCP 14 Hackathon - Notes.....	51
Standard Names Hackathon - Notes.....	52
High-quality CF-NetCDF for ML Hackathon - Notes.....	54
Hackathon Wrap-ups.....	56
Workshop Wrap-up.....	57

The workshop is financially supported by

Climateurope2

European Union's Horizon Europe research and innovation

SMHI

Swedish Meteorological and Hydrological Institute,

SMHI web page announcing the Workshop

CF Community Workshop 2024

Last updated Aug 29, 2024 Published Apr 26, 2024

The CF 2024 workshop will be a four day meeting held in person at SMHI, Norrköping, Sweden, Tuesday 17th September - Friday 20th September.

The meeting will be divided into two parts:

Part I: Engaging the cross-domain CF community to gather user needs and develop a high-level CF roadmap

When? Tuesday 17th September - Wednesday 18th September 2024

Part II: CF development including hackathon sessions

When? Thursday 19th September - Friday 20th September 2024

Event Details

Part I

The workshop will bring together participants from a broad range of CF user communities and potential users. There will be a mixture of invited speakers and presentations from the community. We encourage meeting participants to submit a title for a presentation. Oral presentations should be short (5 minutes) and give an overview of the topic. If you wish to present additional technical details, please prepare a poster to accompany your talk. Poster-only presentations are also welcome. The organising committee will create a detailed program for the workshop based on submissions received.

The aim of part I is to stimulate discussion on how the CF conventions fit into the broad landscape of data and metadata standards currently in use in the environmental sciences. The anticipated outcome is a roadmap for the development of the CF conventions and controlled vocabularies over the next 5 years aimed at maximising the usefulness of CF to a wide range of communities, enhancing interoperability with other widely used standards and encouraging participation in the CF community effort. Proceedings of Part I will be published, and the intention is to publish the roadmap in the peer reviewed literature.

Who should attend part I? The workshop is open to anyone with an interest in understanding and progressing data and metadata standards in the environmental and earth system sciences, in particular:

- those who currently use CF but are, or expect to be, limited by what it currently offers,
- those who have considered using CF but found some barriers to using it in their own work,

- those who work with other environmental data standards and have an interest in developing greater interoperability with CF.

Part II

The workshop will be arranged as a series of hackathon sessions to work on specific aspects of CF development. The precise topics of the hackathon sessions will be determined by the attendees, for example, they may be based on current technical discussions in the CF GitHub organisation, or topics raised in Part I of the meeting.

Links to CF GitHub discussions:

- General discussion topics: <https://github.com/orgs/cf-convention/discussions>
- CF conventions proposals: <https://github.com/cf-convention/cf-conventions>
- CF vocabularies proposals: <https://github.com/cf-convention/discuss>

Each hackathon session will have a chair and a rapporteur. Those leading hackathon sessions are asked to prepare a single slide to introduce the topic of their session.

Who should attend Part II? Anyone with an interest in data and metadata standards in the environmental and earth system sciences. The CF community is international and we welcome new participants at any career stage.

Organizing committee

Ethan Davis, NSF Unidata, UCAR, USA

Alison Pamment, UKRI/NCAS/Centre for Environmental Data Analysis, UK

David Hassell, NCAS/University of Reading, UK

Daniel Lee, EUMETSAT, Germany

Ellie Fisher, NCAS/Centre for Environmental Data Analysis, UK

Kevin O'Brien, NOAA Affiliate, PMEL, University of Washington, USA

Lars Barring, Rossby Centre, SMHI, Sweden

Workshop agenda

Time	Session	Speakers
CEST (UTC+2)	Tuesday, 17 September	
8:00 - 9:30	Registration	
9:30 - 10:30	Opening session	Lars Barring: <i>Welcome, practicalities, format/structure/expectations of meeting</i> David Hassell: <i>Setting the scene – overview of CF</i>
10:30 - 11:00	Break	
11:00 - 12:00	Vocabularies	Chair: Alison Pamment Keynote: Gwen Moncoiffe: <i>CF Standard Names as a global semantic resource</i> Lightning talk: Liqing Jiang: <i>Vocabulary needs of the ocean carbon and acidification community</i>
12:00 - 13:00	Units	Chair: Alison Pamment Keynote: Stuart Chalk (virtual): <i>Digital Units of Measurement: The Digital SI and Interoperable Units</i> Keynote: Chris Little (virtual): <i>OGC Temporal Domain Working Group Activities</i>
13:00 - 14:00	Lunch	
14:00 - 15:30	Uncertainty	Chair: Antonio Cofiño Keynote: Charlotte Pascoe (virtual): <i>Communicating (and understanding) uncertainty through the value chain</i> Keynote: David Huard (virtual): <i>Metadata Pipelines in IPCC Assessment Reports</i>
15:30 - 16:00	Break	
16:00 - 17:30	Metadata for describing statistical processing	Chair: Lars Barring Keynote: Christian Pagé: <i>Metadata Challenges to properly address provenance in climate indices</i> Keynote: José Manuel Gutiérrez Llorente (virtual): <i>Provenance for (complex) climate products: The experience from the IPCC Interactive Atlas</i> Lightning talk: Thomas Martin (virtual): <i>Capturing provenance for ML</i>
CEST (UTC+2)	Wednesday, 18 September	
9:00 - 10:30	Metadata requirements for new technologies	Chair: Daniel Lee Keynote: Daniel Lee: <i>Destination Earth Data Lake - Fueling Europe's Digital Twins</i> Lightning talk: Ag Stephens (virtual): <i>Teaching Large Language Models to speak CF-NetCDF</i> Lightning talk: Jesús Fernández (virtual): <i>Metadata requirements for high-resolution urban modelling</i>
10:30 - 11:00	Break	
11:00 - 12:30	Interoperability and cross format issues	Chair: Ethan Davis Keynote: Sebastián Villaume: <i>Mapping between GRIB and CF-netCDF and ECMWF software development plans</i> Keynote: Martina Stockhause: <i>CF in the context of developments in the Research Data Alliance (RDA): RDA recommendations and CF opportunities</i> Lightning talk: Guillermo Tesoro Calvo: <i>NetCDF-Schema: A general schema for defining and validating netCDF products</i> Lightning talk: Ethan Davis: <i>Zarr and Geozarr</i>

Time	Session	Speakers
12:30 - 14:00	Lunch	
14:00 - 15:30	Cross-domain challenges and opportunities	Chair: Ellie Fisher Keynote: Luke Marsden: <i>CF from the perspective of new discipline</i> Keynote: Martin Juckes (virtual): <i>Challenges in the Expanding Scope of the CMIP Data Request</i> Lightning talk: Heiko Gölzer: <i>Use of CF in the Ice Sheet Model Intercomparison Project (ISMIP)</i>
15:30 - 16:00	Break	
16:00 - 17:30	Synthesis session	Chair: David Hassell
End of first part		
CEST (UTC+2)	Thursday, 19 September	
9:00 - 10:00	Overview of CF news and introduction to hackathon sessions	Lars Bärring: <i>Practical information for Part II</i> Jonathan Gregory: <i>Brief introduction to the CF Committees</i> ——: <i>Review of CF-1.11 and preview of CF-1.12</i> ——: <i>New Vocabulary and Discussions repos – discussions vs. issues</i> Ethan Davis: <i>Brief introduction to the CF Governance Panel, and to the CF Information Management and Support Team</i> ——: <i>DOI for the website and the conventions document</i> Hackathon conveners: <i>Brief introduction to hackathon themes</i>
10:00 - 12:30	Hackathon sessions including break	Daniel Lee: <i>"Good housekeeping": website & conventions text</i> David Hassell: <i>HEALPix grids in CF</i> Guillermo Tesoro Calvo: <i>NetCDF-Schema discussion</i>
12:30 - 14:00	Lunch	
14:00 - 17:30	Hackathon sessions including break	Erin Turnbull: <i>Localized metadata & CF/ERDDAP (cancelled)</i> TBD: <i>CF roadmap/white paper</i>
CEST (UTC+2)	Friday, 20 September	
9:30 - 12:30	Hackathon sessions including break	Daniel Lee: <i>BCP-14</i> Alison Pamment: <i>Standard names</i> Sadie Bartholomew: <i>visualisations for standard names data</i> Ag Stephens: <i>To create a high-quality dataset of CF-NetCDF file headers to be used to train Machine Learning models</i>
12:30 - 14:00	Lunch	
14:00 - 15:00	Hackathon wrap-up	
15:00 - 15:30	Workshop wrap-up	

Presenters for Workshop part 1

(In order of appearance)

Lars Barring: Senior Research Scientist, SMHI Rossby Centre, Sweden, and CF2024 Workshop local organiser

David Hassell: Senior computational scientist, National Centre for Atmospheric Science (NCAS) & University of Reading, UK

Gwen Moncoiffé: Senior Marine Data Manager and NERC Vocabulary Server Team Co-Lead, BODC, National Oceanography Centre, Liverpool, UK

Liqing Jiang: Associate Research Scientist, NOAA Cooperative Institute for Satellite Earth System Studies (CISS), USA

Stuart Chalk: Professor of Chemical Informatics, University of North Florida, USA, and Member CODATA Task Group on Digital Representation of Units of Measurement (DRUM)

Chris Little: IT Fellow, UK Met Office, UK

Charlotte Pascoe: Co-Lead CEDA Atmospheric Science, Centre for Environmental Data Analysis (CEDA), UK

David Huard: Research and Innovation Support Specialist, Ouranos, Quebec, Canada, and IPCC TG-Data co-lead

Christian Pagé: Research Engineer, CERFACS, Toulouse, France

José Manuel Gutiérrez Llorente: Research Professor, Instituto de Física de Cantabria (IFCA), CSIC-University of Cantabria, and IPCC TG-Data Member and IPCC Atlas Lead

Thomas Martin: Software Engineer, NSF Unidata, USA

Daniel Lee: Data access systems operations manager, EUMETSAT, Germany

Ag Stephens: Head of Partnerships, Centre for Environmental Data Analysis (CEDA), UK

Jesús Fernández: Research scientist, Instituto de Física de Cantabria (IFCA), CSIC-University of Cantabria, Spain

Sebastián Villaume: Senior Data Scientist, ECMWF, UK

Martina Stockhause: IPCC DDC Manager, German Climate Computing Center (DKRZ), Germany

Guillermo Tesoro Calvo: Technical Manager, CS Communication & Systèmes, France

Ethan Davis: Technical Manager, NSF Unidata, UCAR, USA, and CF Governance Panel Chair

Luke Marsden: Data Manager, Norwegian Meteorological Institute & University Centre in Svalbard, Norway

Martin Jukes: NCAS Senior Research Scientist in Climate Data, University of Oxford & Centre for Environmental Data Analysis (CEDA), UK

Heiko Gölzer: Research Professor, Bjerknes Centre for Climate Research, Norway

Jonathan Gregory: Senior scientist, NCAS, University of Reading, UK and Met Office Science Fellow, Exeter, UK

Opening Session - Notes

Lars Barring: Welcome, practicalities, format/structure/expectations of meeting

- <https://cfconventions.org> - find information and upload your presentation
 - Locate day → find session → upload!
- Part I
 - Keynotes, lightning talks
 - Focus on conversation/discussions
 - Expectations on CF
 - What others are doing → what can we learn → interfaces/collaborations
 - Input for CF roadmap ← wider landscape
- Part II
 - CF “business session”
 - Hackathon groups - hands-on on specific topics
 - CF roadmap hackathon

David Hassell: Setting the scene – overview of CF

Link to presentation: <https://doi.org/10.5281/zenodo.14230798>

- Climate & Forecast conventions give structure to geoscientific & environmental metadata in netCDF (network Common Data Form) datasets - physically, dimensionally, statistically.
- CF-1.0 in 2003, CF-1.11 in Dec 2023, CF-1.12 later this year!
- CF conventions document - 261 pages, 9 chapters.
 - 1-2 NetCDF files and components
 - 3 Description of the data
 - 4-6, 7 Coordinate types, systems & domain, labels & alternative coordinates
 - 7 Data representative of cells
 - 8 Reduction of dataset size
 - 9 Discrete sampling geometries
- Appendix I - CF data model - high-level description of CF without netCDF encoding details.
- Standard name - what a variable represents physically. Essential metadata attribute. Assigned to values from official standard name table, precise definitions & physical dimensions. Answer the question “What does this mean”, not “what do we call this”. Community interoperability.
- Questions:
 - CF Checker, status?: Needs resources directed at keeping up to date.

- Is the CF Checker on GitHub?: Yes - <https://github.com/cedadev/cf-checker>
- NetCDF schema: hackathon session led by Guillermo Tesoro Calvo, Thursday Morning
- CMIP-CF standard name mapping: not straightforward. Best to look at [CMIP data request](#).

Vocabularies Session - Notes

Alison Pamment: Standard name session overview

Link to presentation: <https://doi.org/10.5281/zenodo.13983049>

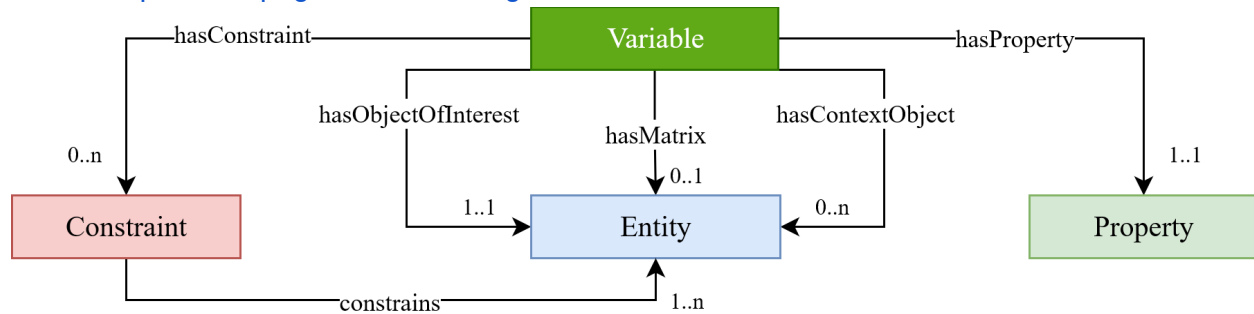
- Current status:
 - Standard names - v86, Sep 2024
 - Area types - v11, Jul 2023
 - Standardised regions - v4, Dec 2018
 - NCEI's seaname list: <https://www.ncei.noaa.gov/data/oceans/nodc/codelists/seanamelist.txt>.
 - XML version: <https://www.ncei.noaa.gov/data/oceans/ncei/vocabulary/seanames.xml>
 - <https://github.com/cf-convention/vocabularies>
 - 100-200 active proposals, 19 open GitHub issues re. CMIP7 Data Request
- 1. How should CF vocabularies develop to cater for new communities?
- 2. Should CF permit use of other vocabularies to identify geophysical variables?
- 3. How to achieve interoperability between CF and other widely-used vocabularies? i.e. FAIR.

Gwen Moncoiffe: Names as a global semantic resource

Link to presentation: <https://doi.org/10.5281/zenodo.14217975>

- Adoption of Semantic Web Standards via web technologies cross-domain interoperability.
CF standard names, CF conventions, W3C standards, I-ADOPT.
- Variable needs:
 - Front facing (understandable, recognisable)
 - Underpinning data workflows (compatibility, consistency)
 One universal standard not possible, different vocabularies, formats, semantic precision etc.

- I-ADOPT ontology - Interoperable Descriptions of Observable Property Terminologies
OWL ontology + recommendations for variable descriptions.
<https://i-adopt.github.io/challenge.html>



- CF standard names as FAIR vocabularies:
 - Consistent, structured, well defined concepts
 - Standardised grammar
 - Strong community governance
- Publication through NVS → BODC Vocab Editor → NVS P07 SKOS collection >>
UID + mapping
inc PREF LABEL, URI, ALT LABEL, DEFINITION and metadata, VERSION metadata, MAPPINGS
- Consider for roadmap:
 - full use of NVS P07 as official source of Standard Names?
 - align with I-ADOPT using a strategic approach
 - key challenge: efficient management of thematic overlaps
- Questions:
 - *What if the UK ceased to exist in its current form, would the domain persist?*
Yes, this has been assured by NERC (also even if the research council does not)
 - *How to capture/accommodate transformations & ancillary information in I-ADOPT?*
Ongoing consideration how to include key elements of CF standard names. The NVC vocab collection/I-ADOPT will never fully replace CF standard names, but complement it.
 - *How does P07 relate to the official standard names table, how do they sit alongside?*
Discussions about using P07 as “canonical source”, encourage users to look here first.

Link to presentation: <https://zenodo.org/records/14150817>

- “Bundled up” CF controlled vocabularies - prefix + variable name + postfix
- Metadata interpretation → instrumental role in many fields including CMIP.
Data findability & (Meta)data exchange - not currently working so well.
- Problems:
 - 1. Prone to errors / duplicates - multiple similar / identical names in table
 - 2. Unnecessary large number of terms - narrow definitions proliferate variations
 - 3. Vocab needs for data submitters - costly to serve needs of data portal
 - 4. Vocab needs for data portal maintenance - data producers need to browse long list
 - 5. Exponentially harder to exchange data → mechanism to map from DACs to unified portal
- Resolution for future - detached vocabularies, with permanent unique IDs, inc. reference & history
- Questions:
 - *Proposing discovery vocabulary, but this needs to be encoded with context information.*
Recommending that these should be managed independently.
 - *Separate specific use cases w. levels of familiarity - portals can be problematic themselves, but a knowledge graph would be an alternative way to capture this information. APIs are queried by the knowledge graph & machine-operable, construct full names from grammar.*
 - *Inefficiency of proposing separate standard names for measurement media vs. an attribute variable (?) to differentiate between these.*
CF grammar is intended for this sort of differentiation, but some semantic/variable overlap.
Possibility of semi-automised tool for construction of standard names (software effort).
W3C ontology SOSA within Semantic Sensor Network
<https://www.w3.org/TR/vocab-ssn/>
SOSA is higher-level than I-ADOPT (targets observable property concept within SOSA).
 - *CF “bundled” standard names can be mapped onto a (searchable) symbolic framework.*
Hierarchy of these may be complex and vary according to domain use. Simple string search querying groups of names is possible on the standard name table

using operators.

- *Balance of / value in prompting users to think about dimensionality of their variables!*

Important to build service closely with research community, what do scientists want?

Units Session - Notes

Alison Pamment: Units session overview

Link to presentation: <https://doi.org/10.5281/zenodo.13983357>

- Current status:
 - Unidata UDUNITS to identify legal units strings (mostly SI) - not required for unit conversion
 - Units_metadata attribute in v85 (?) - aid correct conversion of temperature
 - UDUNITS syntax for time units - seconds since YYYY-MM-D HH:MM:SS.S TZ
e.g. *seconds since 1992-10-8 15:15:42.5 -6:00*
 - Earthbound and current era only... for now.
- Questions for roadmap:
 1. How to achieve interoperability between UDUNITS strings in CF and other units/parameter vocabularies ?
 2. Should we allow units of measure not in UDUNITS database?
 3. Are there alternatives to UDUNITS for time processing?
 4. Can we provide more support for non-standard calendars?

Stuart Chalk: Digital Units of Measurement: Digital SI & Interoperable Units

Link to presentation: <https://doi.org/10.5281/zenodo.14218840>

- Units are ubiquitous, fundamental, for humans only...? Units for machines are not yet solved. Digital representation challenge → CODATA TG Digital Representation of Units of Measurement DRUM. Advocates for unit interoperability, UMIS
<https://umis.stuchalk.domains.unf.edu/>
- <https://codata.org/initiatives/task-groups/drum/>
<https://codata.org/initiatives/task-groups/drum/the-digital-unit-representation-inventory/>

- Digital unit landscape: recommendations, encodings, ontologies, software, scripting, formatting.
- UMIS is written in Python. Combines 20 unit representations, aiming to aggregate & align these.
Identifiers in different ontologies, definitive SI representation at top of unit landing page. UDUNITS is included, links to the external interface. JSON download of representations available.
- Moving to Wikidata as central (long-lasting) repository of UMIS data (host for representations).
- SI Digital Framework - semantic version of SI definitions. <https://si-digital-framework.org/SI>
Forum on Digitalization & Metrology: <https://www.bipm.org/en/committees/fo/forum-md>
- Questions:
 - *UCUM - Unified Code for United of Measure. Related work within UMIS for unit strings.*
Verifying unit strings UCUM vs. QUDT (Quantities, Units, Dimensions & Data Types Ont.)
 - *How UDUNITS fits within larger perspective of OGC & CF community?*
Additional formats beside XML → represent units + metadata to fit different systems.
Administration & maintenance of UDUNITS unclear moving forward. Availability of alternative / complementary software serving same purpose / in multiple languages?
 - *Temperature differences / on scale - functionally different units, relative vs. absolute.*
Community need to address scales in unit quantities, inc. in machinable formats.

Chris Little: OGC Temporal Domain Working Group Activities

Link to presentation: <https://doi.org/10.5281/zenodo.14194151>

- Original temporal geometries had a fixed unit of reference, i.e. they were not calendars. Later implementation of other calendars besides Gregorian.
Best practice document around 2013-2015, joining up with other communities e.g. hydrology.
Code scales, differences between temporal reference systems vs. calendars, relativistic time.

- TimeseriesML - <https://www.ogc.org/standard/tsml/> - XML encoding of OGC Timeseries Profile
- OGC conceptual temporal model -
 - Consistent with W3C ontology
 - Not inconsistent with ISO ("timescales", "timelines")
 - Applying mathematical view of temporal coordinate reference system, not a calendar!
 - Ordinal TRS is a third type - geological/archaeological layers, time-dated events, king lists
- Authoritative vocabularies/taxonomies/ontologies should be in web-resolvable, persistent, human-readable registers. Used outside original context/ownership if useful.
 - WMO and some other bodies not in a position to have a resolvable register. OGC can host authoritative register until reliable host established.
- ISO 34000 Temporal Vocabulary - <https://www.iso.org/standard/77019.html>
 Good = comprehensive, inc. relativistic, planetary & space issues
 Bad = Too many variants, different timescales, 24h split
- MISP Imagery Handbook Chapter 6 - <https://nsgreg.nga.mil/misb.jsp>
 Synchronisation/coordination, quality levels for timekeeping, nanosecond precision
- Leap seconds not abolished before 2035, criteria for insertion/removal may be relaxed before.
- Ongoing/future work
 - Standard temporal CRS (T+hh / T+dd / T+yy), would allow temporal indexing/grid.
 - T-10 "countdown" as another standard temporal CRS. Hold/restart as "datum shift".
 - Liaison with W3C OWL for time, improve liaison with ISO TC154 / TC211.
- Questions:
 - *Integrating satellite data with climate modelling e.g. CMIP, how to capture time difference through satellite swath.*
 Cannot perfectly synchronise clocks if they are moving suitably fast wrt each other. No instantaneous universal time, need to reference "proper time" and relate to ground clocks.
 - *Distinction between temporal reference system and temporal coordinate reference system.*
 Coordinate reference system is locating something with a single value which is monotonic.

Conversion between calendar and reference system depends on the algorithm being used.

- *Model time - not approximation to real world timeline, but using calendars for comparability.*
- *Is it simpler to encode the data in UTC rather than conforming to different calendars?*

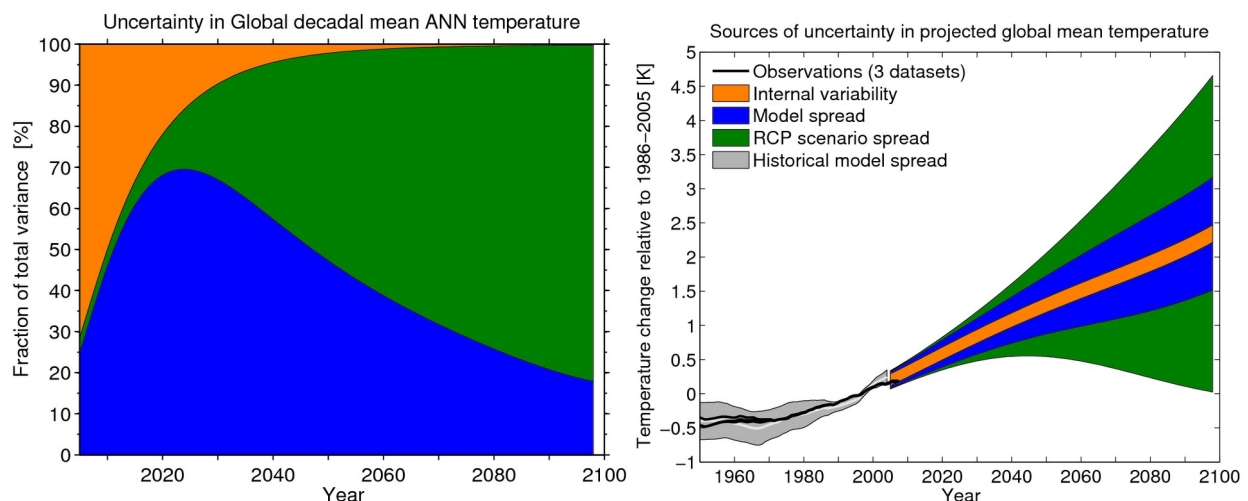
UTC more complex than appears - different precision levels, leap seconds not precisely defined

Uncertainty Session - Notes

Charlotte Pascoe, NCAS/CEDA - Communicating & Understanding Uncertainty Through the Value Chain

Link to presentation: <https://doi.org/10.5281/zenodo.14270790>

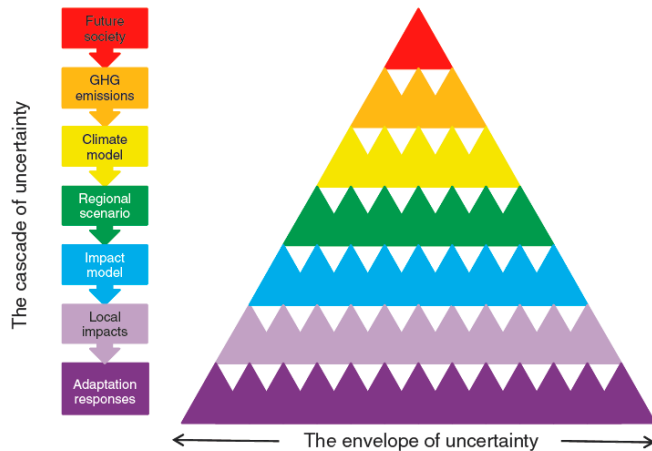
- Diversity of uncertainty within meteorology & climate science. Challenge to communicate nuances to wider community. Potential for SSSOM to structure the communication of uncertainty concepts.
 - How do you communicate uncertainty?
 - How are uncertainty purposes used/managed for decision-making?
- Sources of uncertainty arising from change over time horizons, scenarios and model spread.



Hawkins and Sutton (2009, 2011), Hawkins (2013)

- Conveying uncertainty without using the word itself - i.e. range of responses in possible futures.

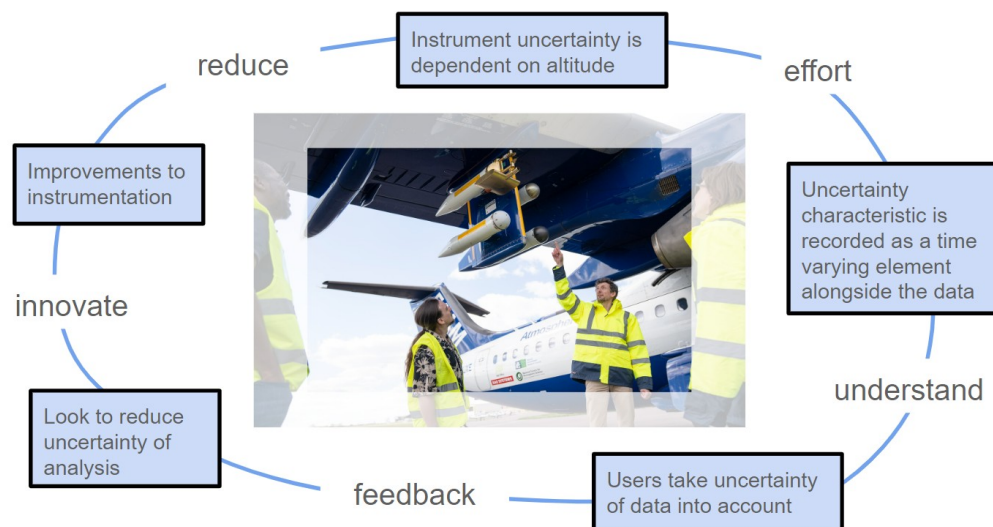
- Language matters - different meanings for scientists vs. public. Focus on what is known first.
- Cascades of uncertainty through communities of practice:



Each step represents different community. Needs to be understood by communities far removed.

- Categorising uncertainty:
 - Location - position in the “cascade” / modelling pipeline / value chain
 - Nature - aleatory, epistemic
 - Level/degree - lies on a spectrum, Complete determinism → Statistical uncertainty → Scenario uncertainty → Recognised ignorance → Total ignorance
- Deep uncertainty - associated with limited knowledge, e.g. future socio-economic / technological developments → climate system responses, better conveyed with “what if” type scenarios.
- Virtuous cycle of uncertainty understanding e.g. FAAM:
 - Reduce → Effort → Understand → Feedback → Innovate

Virtuous Circle of Uncertainty Understanding: FAAM

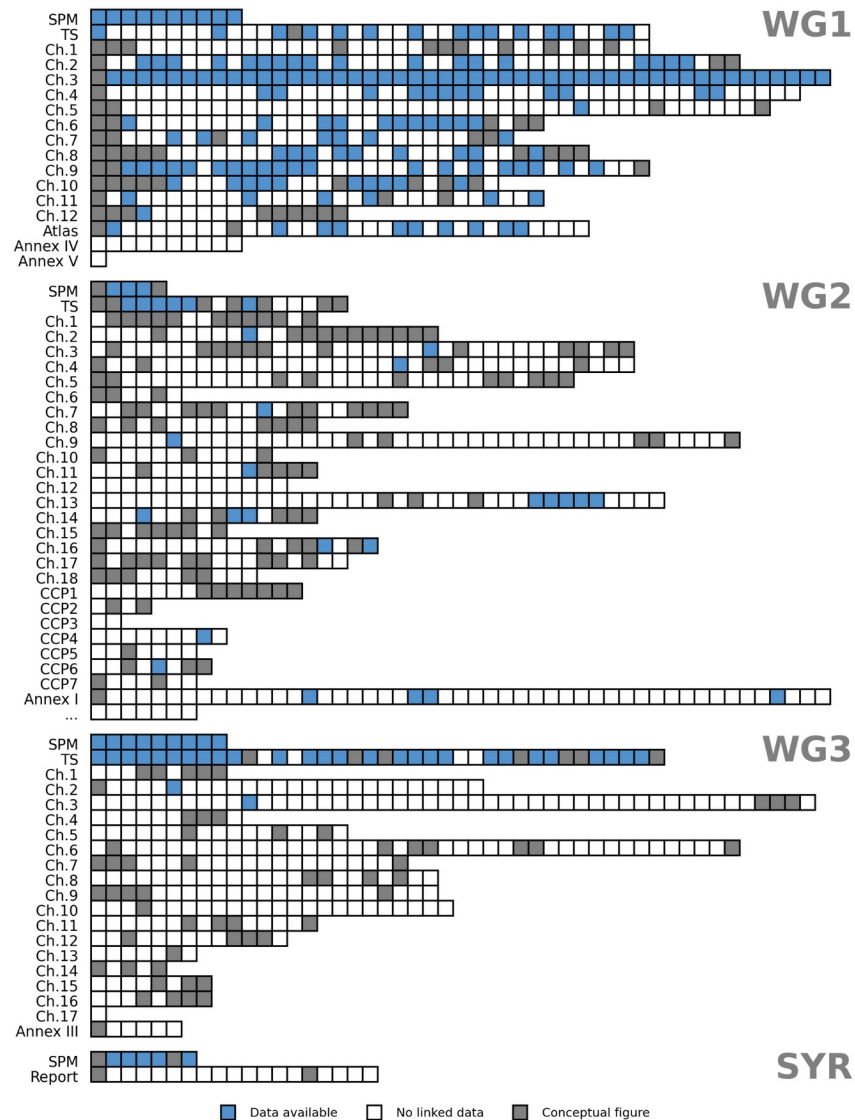


- Questions:
 - *Could CF expand on description/parameterisation of other forms of uncertainty recognised by the users e.g. political, non-numerical, context dependency.*
Assessing relevance of uncertainty for a specific use would help guide how it is described.
Numerical representation has important role in some areas of understanding, not others.
 - *How could SSOM be applied wrt uncertainty*
<https://github.com/mapping-commons/SSOM>
Facilitate information exchange between communities not sharing common nomenclature.
Useful ability to provide some information compared to none - but would need to have ontologies in place to characterise and structure the uncertainty information. Simple = best.
 - *Well established procedures for laboratory & observational communities, less so for modelling.*
Valuable effort for controlled vocabularies for statistical averaging & processing with respect to this.
 - *No language in CF Conventions to describe Monte Carlo simulations. (David Huard)*
Should this be introduced to the Conventions or encoded into CF in some other way?

David Huard: Metadata Pipelines in IPCC Assessment Reports

Link to presentation: <https://doi.org/10.5281/zenodo.14147782>

- Data from IPCC AR6 now - partly - FAIR, but coverage is far from complete across all WGs.
<https://ipcc-data.org> - TG Data, IPCC Data Distribution Centre.



Unequal distribution across chapters and WGs / reports. Most data available for WG1. Over 1100 total data figures, but fewer than 250 have data available. Difference in tools.

- Metadata pipeline in AR6
 - Produce draft figure
 - Internal review & modifications
 - External IPCC review & modifications (many reviewers)
 - Review data & metadata at TSU (Technical Support Unit)
 - Input data licences for redistribution
 - Upload data & metadata to DDC (TSU)
 - QC & catalog record creation
 - Record preview and publicise data record
 - Mint DOI and add to IPCC web page

- Typical climate services workflow e.g. Ouranos:
GHG scenarios → GCM simulation → Subsetting → Downscaling → Impact modelling → Analysis
Emphasis on documentation of provenance in machinable format. Recommended for IPCC.
- AR7 figure data archival
Challenges
 - Timeline compressed by political pressure
 - Data curation casualty of time & resource constraints.
 - Data science expertise among authors - distribution of effort.
 - File format - often stored as CSV for simplicity.
 - WGII and WGIII do not have CF-like community to support.
 Ideas
 - Build provenance tracking plugins.
 - Provide version control repository templates.
 - GitHub action to check NetCDF files for CF compliance
 - Nominate yourself to contribute data science expertise to AR7!
- Questions:
 - *How do you nominate yourself to assist with AR7?*
Separate process by country - call to submit nominations to National Secretariat.
 - *Web applications / tools to make metadata discoverable. What is being done?*
ESMVal <https://github.com/ESMValGroup/ESMValTool>
Climate4R <https://github.com/SantanderMetGroup/climate4R>
 - *Distinction between suitable storage formats vs. use formats. Interoperability.*
BADC-CSV is a suggested formatting for CSV files to include metadata in headers.
 - *TSU had a Figure Manager system for WG1, YAML files used to interface with this.*
Funding needs to be there to provide necessary technical support for tools to IPCC.
 - *CSV implementation in figure data management - problem of lack of standards.*
Textual representation on NetCDF files via. CDL (Common Data Language). Human-readable, could be used as an intermediate “language” to bridge data from CSV → NetCDF.
 - *Where do authors / modellers submit their model data for IPCC Assessment Reports.*
DDC nodes are typically DKRZ (German Climate Computing Centre), CEDA / MetaData Works, University of Cantabria (Interactive Atlas). ESGF has a different set of nodes.
 - *Any feedback on usage of the data from the AR6 visualisations, did this meet the purpose?*
Trackable metrics only for downloads / views - no information on usage afterwards.
IPCC WG1 GitHub repository for scripts: <https://github.com/IPCC-WG1>
 - NetCDF-U - Uncertainty conventions for netCDF datasets:
https://www.researchgate.net/publication/258780653_NetCDF-U_-_Uncertainty_conventions_for_netCDF_datasets

Metadata for describing statistical processing - Notes

Christian Page: Climate Indices Standards and Definition

Link to presentation: <https://doi.org/10.5281/zenodo.14149110>

Several existing standards

Icclim: index calculation for climate

For different applications sometimes necessary to modify definitions of standard indices, e.g. temperature threshold can be adjusted for location when defining “warm days” , “summer days” etc.

Climate indices are based on statistics about an atmospheric field, sometimes draw on multiple variables to calculate the final product. How to record all that information?

Do climate indices and climate indicators mean the same thing?

Are standard indices applicable to your specific region, etc?

Very detailed, e.g., “less than”, “less than or equal to”

Bias correction

Reproducibility cycle(s) - important to reproduce results, be able to test results

Provenance model: Data lineage, knowledge, ...

Provenance challenges: how much domain metadata should be included, what level of granularity to describe complex objects, manage scale of provenance records to be recorded/processed

Climate indices - divide into	an operator	a variable	a threshold,
e.g,	number of days	max temp	> 25 deg C

In CF standard_name, long_name, units, cell_methods all important

Provenance can be in history attribute for file and another history on the variable

Have long list of processing that has been applied, but free text string with a few separators, not useful either for humans or machines!

Need a more standardized and community agreed standards for provenance

1. Clix-meta - definition, units name , basic metadata, vocab
2. W3C-PROV - processing methods, software used, input_files, versions
3. SWIRRL - user interface to benefit from the other standards

SWIRRL-API - provenance aware workspaces

Major gaps and challenges.

Climate indices and indicators means data processing:

How can it be CF compliant? What is scope of CF convention?

How related to FAIR principles

How can we ensure uptake

Where and how should we store all this provenance information?

Chris Little - is this SWIRRL a Scottish company?

Christian - no! (name clash is a coincidence)

Liqing: can we establish a universal metadata framework that can serve both modelling and measurements communities? (Will also need metadata specific to each).

Christian: this is a good idea to reuse common elements

Gwen: have you considered a repo giving unique ids to the code producing the indices?

Christian: There isn't a doi/code for each index - the indices can be calculated from building blocks that can be put together as needed

José Manuel Gutiérrez: Provenance for (complex) climate products: the experience from the IPCC interactive atlas

Link to presentation: <https://doi.org/10.5281/zenodo.14270689>

Curated IPCC-DDC data

Code and data repositories (IPCC-WG1 github)

For atlas needed one solution to represent many products.

Combination of atmospheric, oceanic and other fields to produce derived indices - computed and displayed for regional obs and model projections CMIP5/6 & CORDEX

Need to adjust to different calendars, different data resolutions, etc.

Analysis dimensions: global warming levels, scenarios, baseline dates

Observed trends for modern and paleo periods.

Produce global, regional products, styles of presentation, e.g. climate stripes, graph, maps

Landing page is multi-model CMIP6 temperature change relative to 1850 - 1900 (global map)

User can select other variables and indices (some of which are bias adjusted) and display the maps

For reproducibility of the maps need provenance metadata including processing and final products.

Need machine readable metadata, not just a title on the plot!

METACLIP uses RDF through specific vocabs, written in OWL, describing different aspects in climate product generation.

Extending standard vocabs such as PROV-O

<http://www.github.com/metaclip/IPCC-AR6-Atlas>

This is included in the Atlas - can click to view for a specific plot and produces an RDF file and semantic graph to show data/workflow in producing the graphical product.

Semantic graph is clickable and links to underlying metadata for each step of processing.

Copernicus climate data store - can also build on similar standards needed for the IPCC Atlas

Chris Little:

Can Metaclick vocabularies standards be adapted to taxonomies and ontologies?

José Manuel - in principle can be adapted, time consuming to link vocabularies

Chris - if you wanted to use a vocab from oceanography or chemistry could you do so? Which standards must it comply to? Does it have to be defined in RDF?

José Manuel - need to decide what elements in MetaClip vocab can link to elements in the vocab you want to plug in. Can build RDF from other metadata descriptions of vocabs.

Thomas Martin: Capturing provenance for ML

xarray/metpy/machine learning tutorial - github (Python notebook)

Grab CF-compliant dataset and inspect attributes

Add a flag and some extra attributes to describe it as an ML output.

Working on this at Unidata.

Discussion:

Lars: don't have much provenance metadata in CF. External PROV standards exist. Should we put it in CF or do we link out to it?

Seth: We might need to rely on external files for this information. When you need a lot of provenance the metadata gets very big (bigger than the data!) Don't dump everything in the history attribute. Don't duplicate all the provenance in every file, instead refer to a smaller

number of external files. Add pointer attribute to the external files.

David Hassell: Prefer terminology “external URI resource”! Externalisation of the information is fine.

Daniel: Agree it makes sense to externalize provenance metadata for a given dataset. Provenance is fundamentally a graph so you can point to all the lower level resources - maybe a way of adding a small CF metadata attribute that points to external detailed provenance records.

Lars: “URIs” is good language rather than “files”

Seth: Need a chain of statistical processing, way of linking to e.g. standard names not producing lots of new ones! Need to be able to link smaller elements in multiple orders.

Daniel: agree we can't come up with metadata vocab for every possible combination. Need to be able to point across institutional boundaries with URIs.

José Manuel: Amount of detail in metadata influences the size of the metadata, can get very big. IPCC Atlas has metadata inside file and RDF is outside file, so both are present.

Sadie: What advice can you give to someone who wants to write provenance metadata now? Is it best to dump it all in the history attribute or add a small external tag? Standards are still developing!

Daniel: Quick answer - add it to the history!

Ethan: ML idea seems to be capturing small-scale metadata rather than all the formal provenance? Seems like for now this should also go in history as there isn't an agreed standard yet, but is there any advice on best practice?

Seth: ML is still very unformed, but when putting provenance in a file think of what could go wrong in data processing? How can a data user work out what happened and why did it go wrong?

Thomas: Brief response to Ethan. ML is not formal yet! Need discussion among communities but at least let's capture the low hanging fruit for now and get better as we go along.

Andrew: Do we want to take on provenance as a whole subject in CF? For code, provenance is in github.

Daniel: PROV data model is very simple, so it's appealing. But also possible to produce useless metadata if it's not detailed enough! CF should think about HOW BEST to use PROV or some

other standard, what vocabs and ontologies should we use to provide the PROV content? It would be best to see how other people who have done it - we should learn from the specialists.

Lars: Agree with Daniel. Many users don't want to provide very detailed PROV. Currently we have standard names and cell_methods to provide some provenance information and not much else. What else can we do without making the job of provenance too difficult for the CF community.

David: Similar to grid descriptions - smaller hook to much more detailed external description.

José-Manuel: PROV alone is useless and no standards yet for extensions, add URI to NetCDF file to point to machine readable provenance. Have a specific attribute for the URI, don't just have it in generic history attribute, so you can at least see quickly whether a particular dataset has detailed provenance information available.

David: Agree!

Liqing: Problem of metadata being big not so important, better not to separate the metadata from the data, so feels surprised by this suggestion.

Seth: As provenance metadata gets huge it's just not possible to keep it all in the file with the data because it becomes unmanageable.

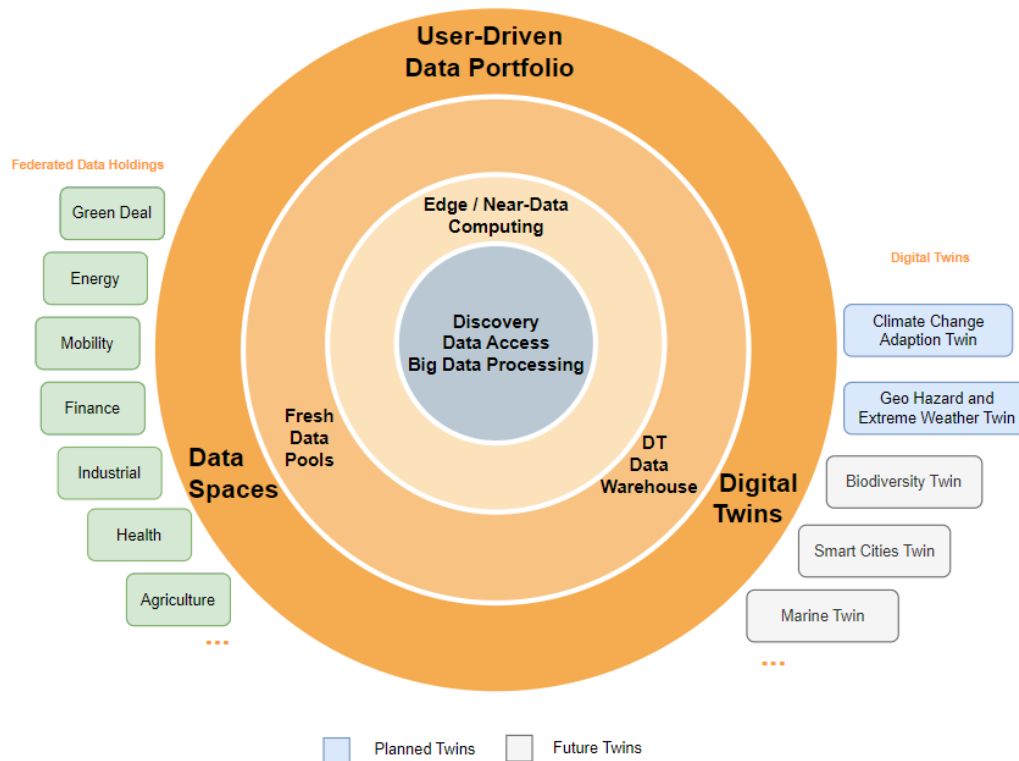
David: Maybe a bit provocative. How much of provenance is about identifying what we have, not providing the full detail? Can we add it in standard names?

New Technologies Session - Notes

Daniel Lee, EUMETSAT - Destination Earth Data Lake - Fueling Europe's Digital Twins

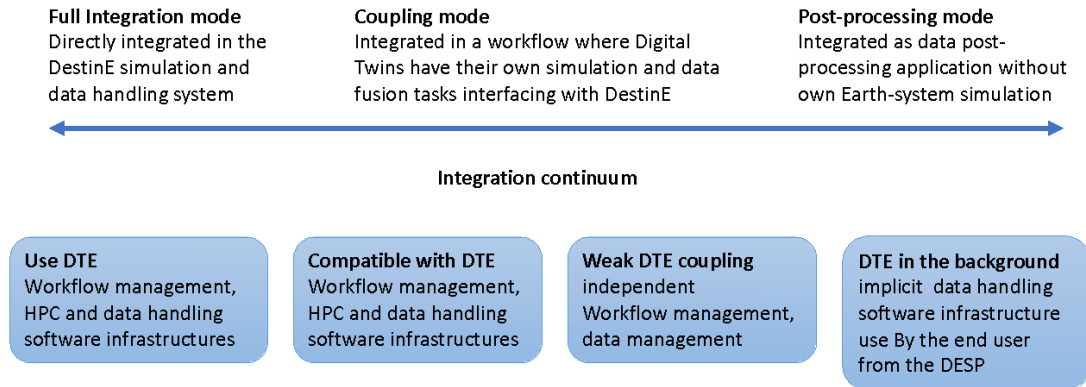
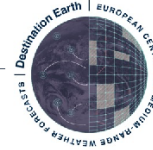
Link to presentation: <https://doi.org/10.5281/zenodo.13982704>

- Destination Earth (DestinE) -
Flagship digital replica of planet to respond and adapt to climate change and extreme events
Novel info system by 3Es: ESA core platform, EUMETSAT data lake, ECMWF digital twin engine
- Data Lake connecting to Data Spaces - federated data spaces → harmonised data access HDA
Geographically distributed physical elements and services. Processing via distributed computing.



- Data space considerations:
 - DestinE, EOSC (European Open Science Cloud), CDSE (Copernicus Data Space Ecosystem), WEkEO (), GREAT (Green Deal Data Spaces), SIMPL (Smart Middleware Platform)
 - Loosely coupled spaces (federated vs. monolithic)
 - Process → transform & stream only what is needed.
- AI “factories” proposed by EU commission:
 - Simplified access to AI supercomputers (HPC)
 - Large-scale model training
 - Feeding GPUs with AI-ready data to optimise use
 - AI ready data = prepared to be usable for AI (fuse, tailor, harmonise)
- ESIP AI-ready checklist:
 - [Checklist to Examine AI-readiness for Open Environmental Datasets \(figshare.com\)](https://figshare.com/Checklist%20to%20Examine%20AI-readiness%20for%20Open%20Environmental%20Datasets)
 - <https://github.com/ESIPFed/data-readiness/tree/main>
- Intertwin (DTE) -
 - HPC sites require (bridges)
 - openStack based IaaS
 - processing near data
 - data exchange (stream)
 - buffer of data
 - distributed infrastructure
 - Data space connection

Different types of Integration



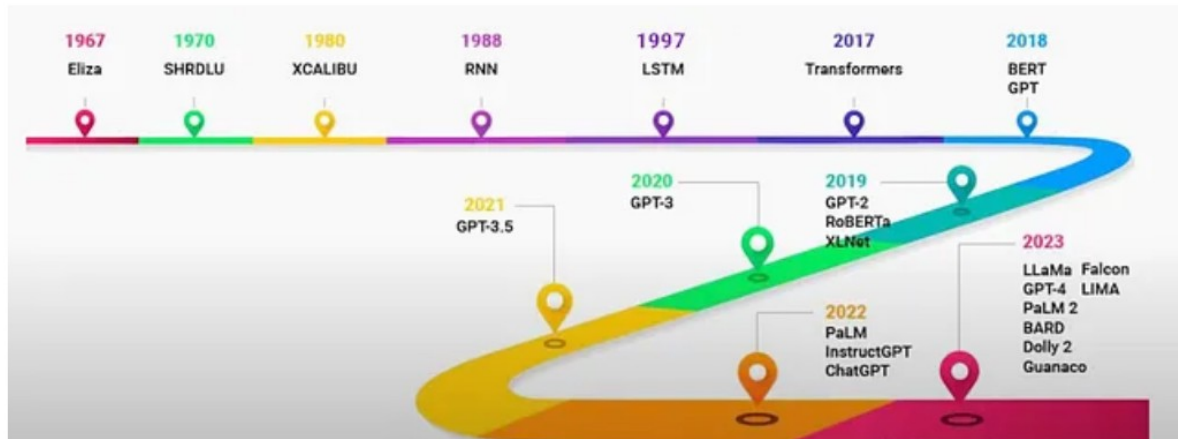
Jesús Fernandez: Metadata requirements for high-resolution urban modelling

Link to presentation: <https://doi.org/10.5281/zenodo.14193316>

- Unprecedented city details are being approached by the climate modelling community.
- Vocabulary/needs vary depending on the community: microscale modelling, different urban parameterization approaches in km-scale modelling
- We plan to propose new CF area (sub-)types relevant for the urban environment
- Specific questions:
 - Can the area_type standard name be used to create custom area types without the need to add them to the CF area type table? How to store the definitions?
 - No, area_type variables still need to have values drawn from the CF area type table. It is not a way to overcome the standard types.
 - Is there an intended hierarchy in the area types? or is it only the result of different, scattered needs over time?
 - No intended hierarchy, although it is known that some underlying relations exist. No a-priori terms are added to consider in advance “all” possible uses, only when requested by needs.
 - What level of granularity is expected for area types? (sub-types can grow wildly)
 - Depends on the demand. Types are intended to bring consensus across model output, rather than reflect particularities from all models.

Link to presentation: <https://doi.org/10.5281/zenodo.14160852>

- Large Language model evolution: 1967 - current. BERT GPT 2018, Chat GPT4 2023. Trained on corpus of internet data. Interesting side effects of this.



Large language model (LLM) timeline. Source: [Brief History of Large Language Models & Generative AI | Evolution of NLP from Eliza to ChatGPT](#)

- Tokenisation - splitting input words/subwords into numeric indices in vector space. This can do the same with code. Tiktokenizer = <https://tiktokenizer.vercel.app/> <https://github.com/openai/tiktoken>

Tiktokenizer

Add message

```
for i in range(10):
    if i % 2 == 0:
        print("Something or other")
    else:
        print(f"I like: {i}")

raise Exception("Crashing out")
```

gpt-4o

Token count
43

```
for i in range(10):
    if i % 2 == 0:
        print("Something or other")
    else:
        print(f"I like: {i}")

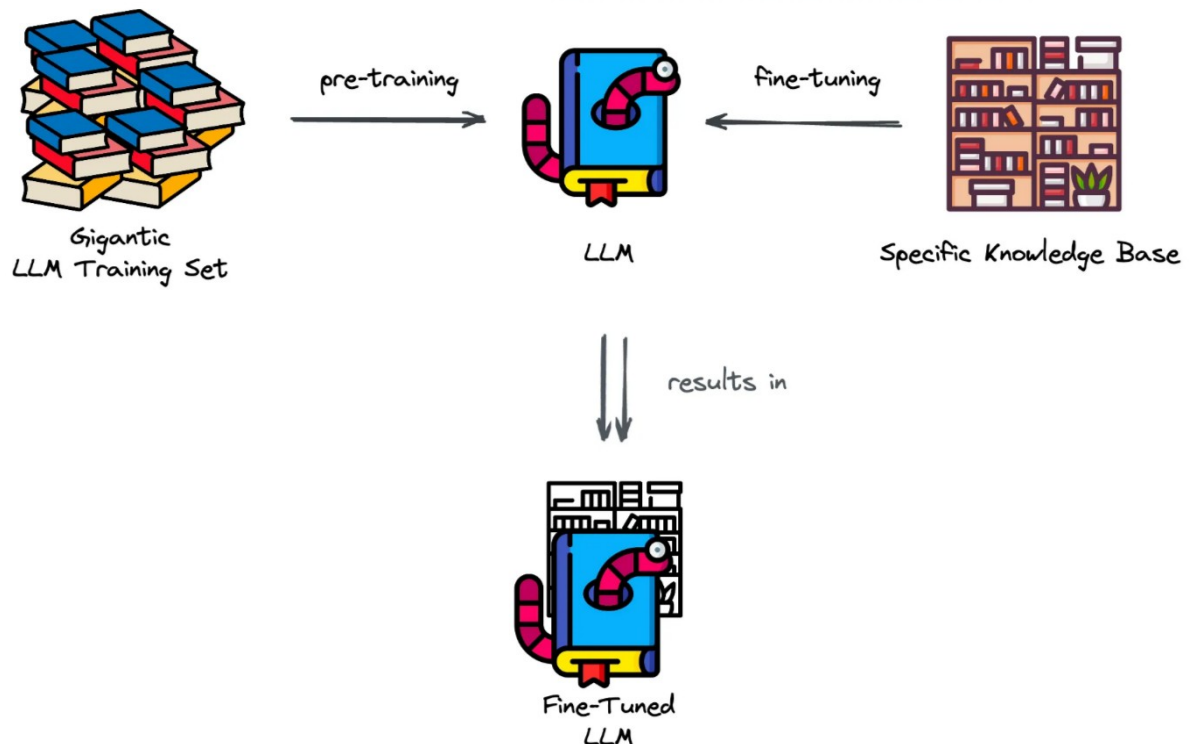
raise Exception("Crashing out")
```

1938, 575, 306, 3352, 7, 702, 1883, 271, 538, 575, 1851, 22
0, 17, 951, 220, 15, 734, 309, 2123, 568, 40330, 503, 1273,
1896, 271, 1203, 734, 309, 2123, 1526, 23796, 1299, 25, 35
4, 72, 119200, 38383, 6851, 568, 14225, 33306, 842, 1405

<https://blog.ml6.eu/leveraging-llms-on-your-domain-specific-knowledge-base->

[4441c8837b47](#)

Open source model candidates - e.g. Meta Llama 3.1. More fine-tuned / pre-trained LLMs.



- Pretrained LLM → curate training set of high quality CF-netCDF headers → gather existing documentation on CF-netCDF → fine-tune model with (2) and (3) → deploy for public use.
- Hackathon idea: creating a high-quality dataset of CF-NetCDF file headers. Knowledge representation problem, training on CF standard names to suggest format.
- Retrieval Augmented Generation (RAG): <https://aws.amazon.com/what-is/retrieval-augmented-generation/>

Cross-Format Issues Session - Notes

Sebasti n Villaume: Mapping between GRIB and CF-netCDF and ECMWF software development plans

Link to presentation: <https://doi.org/10.5281/zenodo.14219169>

- ECMWF mainly producing GRIB and BUFR but now in some big projects with orgs than use NetCDF
- ECMWF CDS produces GRIB but most convert to NetCDF straight away.

- New team in WMO to look at operational NetCDF (TT-NWP)
- GRIB does not represent 3d so any height or ensemble is slices of 2d data. NetCDF to GRIB is hard, vs the other way to convert.
- What ECMWF are trying to do is not map CF standard names to GRIB discipline-category-number – they are mapping standard names and cell method and other attributes into GRIB
- Nemo ocean model is using ORCA grid, IFS model is using HEALPix grib – some grids are not in CF so conversion not yet possible from GRIB
- Example: hurricane track probabilities – give a yes value for defined grid squares which the hurricane passes through in a particular time period. Giving probabilities.
- In GRIB the units are prescribed so you cannot change the units which are set – so if CF has different units, you are forced to translate into GRIB defined units.
- At ECMWF In GRIB you can have multiple forecasts in one file. With when forecast started and how long it runs, bundle them in the same file. How do you do this in CF-NetCDF. – Need two different time coordinates. Where does the time axis go? End up with multiple data points for the same time and date. Few options shown in the presentation.
- How do you do reforecast and hindcast?
- Questions around accumulation
- Potential solutions – cfgrib developed by ECMWF – not best solution
- Would like to develop a CF profile for GRIB mapping – decide a predetermined variable name that is not standard names as GRIB is more complex
- At the moment they are going to use earthkit (BETA) – python module for reading and writing CF and GRIB

Questions

- Alison P - For the examples shown you would have to introduce new standard names - for accumulations etc. Mapping CF standard names to I-Adopt could contribute towards the solution - having the building blocks to producing a more detailed parameter. Previously worked on CF common concept which looking at similar problem - CF experts would like to support ECMWF with this work through I-Adopt
- Chris L - forecast is an observation of the future

Martina Stockhause: CF in the context of developments in the Research Data Alliance (RDA): RDA recommendations and CF opportunities

Link to presentation: <https://doi.org/10.5281/zenodo.14146359>

- RDA has working groups, interest groups and communities of practice...
- I-Adopt is part of this.

- Data versioning interest group, data collections interest group (archiving an records), data granularity group.
- Granularity levels for datasets and citation levels (DOIs)
- FAIRification – data maturity model working group
- Stewardship maturity matrix for climate data (link)
- Metadata 4 machines schema and standard

Guillermo Tesoro Calvo: NetCDF-Schema: A general schema for defining and validating netCDF products

Link to presentation: <https://doi.org/10.5281/zenodo.14196683>

- Requirement is coming from the increased use of NetCDF in EO products
- NetCDF schema in YAML and a validator
- Domain agnostic using pattern matching validation
- Allows for complex relationship an conditional validation
- If we can create a schema to translate CF, we can create CF-NetCDF files easily
- Hackathon to develop idea further
- Open source schema development – aim for schema and validator by end of year.

Questions:

- Have you looked at JSON schemas? - not looking to create new schema definition, same as JSON schema concepts but just writing it in YAML right now - will use JSON for validator - Andrew B
- Chris L: experience from CoverageJSON - schema is the structure, need something else to look at the content for validator, is it correct.

Ethan Davis- lightning talk on Zarr and GeoZarr

Link to presentation: <https://doi.org/10.5281/zenodo.14194167>

- Geozarr is geospatial for Zarr - builds on and uses CF
- Supports GeoTIFF

- Details on GDAL geotransform coefficients

Questions:

- Issues with GeoTIFF supporting projections - trying to find ways around that as not supported in CF

Cross-Domain Session - Notes

Luke Marsden: Expanding CF into new disciplines

Link to presentation: <https://doi.org/10.5281/zenodo.14160704>

- Worked on many projects focussed primarily on polar regions
- Big multi-disciplinary research projects
- Meteorologists, physical oceanographers, marine/terrestrial biologists, chemical oceanographers
- Vary from very structured data to very unstructured, so have to cope with many data types - some go easily into NetCDF, others less so
- Good experience in introducing new users to CF - not easy!
- Biologist, seawater nutrients data, send an excel file, intercept questions that would otherwise go direct to CF community, help with proposing standard names, teach them how to create the files
- Middleman has awareness and competence in using CF conventions, some people can't code, standard names have their own syntax and descriptions so proposals need some work, can't expect all scientists to know all about data management!
- Don't know which bit of document to look at, or even which conventions exist
- More complicated variables, how to encode? How much detail to include? How to include several things in one file? So many degrees of freedom, can be confusing for new users.
- How can we reduce degrees of freedom and provide more tailored advice
- Which attributes to include?
- Rely on ACDD (not much discovery metadata in CF)
- Granularity - e.g., what to do with multiple depth profiles can advise one per file, simpler files, metadata describes data better
- Mostly people want to know time, date, maybe not which cruise

- Advice on granularity in CF would be useful
- Framework for standardising semantics and metadata
- Specify names of dims and coord vars
- Certain optional metadata CF/ACDD
- Wish list
- CF recommended extensions/profiles with documentation
- ACDD governed from within CF
- More support for scientists/data managers
 - Proposing std names
 - Proposing extensions and profiles
 - Developing tools/converters
 - Training:
 - Workshops
 - Training / surgeries - bring your own data
 - Spreadsheet template generator (CF configuration) template available
 - www.nordatanet.no/aen/template-generator
 - You Tube channel

Question: ACDD

Ethan: ACDD is homeless since 2015 - maybe we could take it on? Currently with ESIP but not really moving forward

Profiles and other examples where would we put them? Maybe could go into ACDD - we'd need to work out governance process

Jonathan: Don't know much about ACDD but we tend to point people towards it! We should think about it, we need more effort available to take it on

Fran: How many data managers do you have compared to number of scientists?
6/7 working mostly on DM, 10 on EO. Would be nice to have more!

Link to presentation: <https://doi.org/10.5281/zenodo.14236130>

- Talking about the work of many people, Martin's slides!
- CMIP provides foundations for IPCC assessments of future climate change
- CF has been part of CMIP for a long time
- Co-leading CMIP Data Request Task Team (DR TT)
- DR process expanding, agree list of well specified variables, send to modelling centres and resulting data shared with 1000s of scientists
- CMIP5 970 vars, CMIP6 2062 (23 MIPs) which could each contribute to DR
- Feedback from CMIP6
 - Data request should be better characterized and less extensive
 - Participation should be broader and more inclusive
 - MIP endorsement process will not be repeated for CMIP7, replaced by simple registration
- CMIP7 constrained by IPCC timeline, which is not agreed yet! Aiming fast track simulations for UNFCCC global stock take in 2028
- DR is responding to the timeline with allowing only 3 releases
 - First release Oct 24
 - Update/revision Dec 24
 - Final version Mar 24
- Compromise between checking metadata and need to start model runs in time
- Workflow of DR
 - CMIP IPO now providing much support
 - Strategic approach - allow participants to choose between joining harmonized DR or have a separate project
 - Recruiting community volunteers into 5 thematic author teams
 - Simplifying semantic structure of DR
- Structure of DR
 - Core
 - Harmonized
 - 5 themes
 - Impacts and adaptation
 - Ocean and sea-ice
 - Land and land-ice
 - Atmosphere
 - Earth system

- Unharmonized/a la carte
- Incentive to get volunteers for harmonized effort - publish an academic paper so get credit. Each thematic paper has a scientific steering group with representatives from MIPs, WCRP core projects and lighthouse activities
- Cross-theme working groups (fresh-eyes, technical, CF WG)
- So far 312 new vars in 153 variable groups (down from CMIP6 so far)
- Physical parameter = std name + units
- Variables = physical parameter + area types + coords + dims
- **CF Challenges**
 - **Bringing more people in means more who don't understand CF**
 - **Confusion about roles of DR/CF**
- **Expansion in participation, but limited need for new CF proposals - 17 issues**

Heiko Gölzer: Use of CF in the Ice Sheet Model Intercomparison Project (ISMIP)

Link to presentation: <https://doi.org/10.5281/zenodo.14194187>

- ISMIP global community of sea-ice modellers
- Estimate future sea-level contributions from Greenland and Antarctic ice-sheets, associated uncertainty
- First integrated with CMIP in CMIP6
- Use global models to drive ice-sheet models
- Any CMIP model as forcing for standalone models, but in CMIP7 more focus on coupled AOGCM (-ISM (ice sheet model))
- Ice-sheet specific variables in CMIP DR
- NetCDF only used for ISM outputs since 2015, therefore same with CF
- Ice-sheets use years as unit, not days but have to use days in CF
- Polar projections have to be mapped to lat-lon
- Small community so not so much support from data managers, so data organisation falls on individual modellers to implement in their own models, steep learning curve
- Relying on downscaling from model runs before feeding into ice-sheet models

Jonathan: we could define a special unit of year (different from UDUNITS) for this sort of application so could link to a date for a paleo calendar

Heiko: Would also help with coupling if could relate to a calendar

Ellie: Polar projected coordinates - what are associated problems?

Heiko: Maybe not CF problem but CMIP problem because they want lat-lon coords. Regridding caused loss of resolution. Possibility for storing on native grid - this is done for ocean grid. Prevented data being published on ESGF because nobody available to help! Sometimes you are lucky and find an individual who understands the problem and can help with ice-sheet modelling and global-modelling reprojections.

Lars: This grid problem affects other MIPs

Chris: CF has worked a long time in it's own community, but beyond that there is an ISO standard for map-projections with moving datums, use a coordinate of years. Mapping and coordinate reference systems is geospatial problem, lot of external expertise that we should learn from/link with. Chris can help with contacts.

Martin: Shame the data didn't get into ESGF archive. Hope we can sort it out for CMIP7.

Heiko: Agree we can probably sort this out for CMIP7

Lars: In UDUNITS 1 year=tropical year (close to, but not exactly same as calendar year) so could you use it on geological time-scale?

Heiko: We can define a 360-day calendar and record in days, just makes some big numbers and divide by 360! Works okay, just came as a surprise when began working with global models and CMIP, ice-sheet models don't really work with calendars.

Chris: OGC time WG will register 360 day year as formal CRS. This will give such constructs greater visibility. Analogous to GPS issues - sometimes small differences are significant.

Ellie: Did you need additional CF attributes to define shape of data? Got inspiration for standard names from sea-ice community. Units were okay (needed m year⁻¹ but had to write m s⁻¹).

Heiko: Engagement with CF worked well!

Synthesis Session - Notes

David Hassell: Towards a road map or the development of CF

Link to presentation: <https://doi.org/10.5281/zenodo.14005753>

- Aim
 - Path development for the next five years
 - Detailed ideas or vague ones are valid
 - Write in a form of peer-review paper/letter
- Benefits
 - Real use cases been addressed in a better way
 - Chances for request funding
- Contents
 - Prefer specific features vs just vague directions
 - Recurrent themes not-addressed (neither positively nor negatively)
 - Interactions with non-CF entities
 - Dependencies and/or priorities between items
- Update
 - Items to added, removed or updated by consensus
 - CF annual meeting
 - VCS
- Engagement
 - Having a road map is enough?
 - Interactions with other standards
 - Bring knowledge from other efforts (i.e. provenance) for new features

Questions to discuss

- Is this a road map or a plan with actions and resources?
- Are there parts of CF that other people are doing that we make reference to and not own?
- Could we have more regular in person meeting if we add them onto the end of existing conferences or events that people might be at?
- Define the objective of each of the themes in the list - be specific about what these activities are - provide the use case for why this is needed and that might lead to the priorities.

- Using an online chat like slack really increases the engagement and activity - easier to communicate little and often, ask questions etc.
- cf checker could become open source? Making it available to people to use so they can engage when there are problems
- Tools and CF infrastructure - there are certain things we need to continue to keep CF going
- We need more advertisement of CF, events, announcements, the website, github etc. We don't have dedicate staff for comms and social media so if we can spread the word more and advertise CF better - all announcements happen on the CF GitHub Discussions. Using our other community contacts for advertisement via their mailing lists.
- New user feedback - first interaction with CF gives the impression of an authoritative voice with specific requirements for interaction rather than a small group of people and community standard - this could be better expressed by our website and GitHub - better to get more people involved rather than make it harder to interact and scare new people away!
- Tomorrow hackathon session - plan on a page - flesh out the recurring themes list - also discussing the governance of this plan and roadmap.
- Is a roadmap a wishlist? Big topics vs specific attributes - we could ask user community to add and flesh out the wishlist - invite people in to discuss - having the roadmap will aid discussion.
- Topics from previous workshops that have not been progressed - review the previous years and outstanding items - also can review GitHub Discussions and issues, how far are they along, are these on the roadmap, what's blocking them, are they still current etc.
- What is the timelines for CF - is the roadmap going to reflect whats realistic for a community led standard.
- Could we have a review of the roadmap at particular intervals - what things are moving along, what is dormant and why? Roadmap update meeting every 3 months or something similar??? Starting using the Kanban board in GitHub so we know the status and this can be part of the review.
- Community meetings - open to all
- Dormant issues - if noone is motivated to work on something - maybe it's not that important!
- Humanising the people behind CF - no funding, lack of resources etc - perhaps helps when people first interact
- CF scope is very large!! We should be more specific about things need fixing.
- Making it easier to read the CF conv document - get user questions answered - ChatGPT was interesting - have a core and extensions could be an option
- Wishlist - CF for dummies - CF cheat sheet? CF data model paper does have a summary with images..could review this.

- Meetings - we sometimes have talks at other meetings EGU, AGU but maybe we could also have a stall at things or meetings at existing big events.
- CF remit, here's what CF does for you? Heres some examples of not CF compliant files, here's making it CF compliant etc. Useful for first time users.
- FAQ doc - not up to date - think about how much work required to actually do all these things and maintain them
- Examples, notebooks etc will provide the interaction with community and provide guidance
- Roadmap should contain item on how we manage CF, organisation
- Nature or Frontiers or Data Science for publishing the 'letter' or paper

Recommended paper scope (Liqing):

- Types of CF vocabularies
 - What types of vocabularies the CF community should manage, e.g., observed properties, types of uncertainties, etc.
- Format of the vocabulary names
 - Should we continue using the long and banded up vocabularies or move towards detached and independent vocabularies?
- Related information about each CF term we should manage
 - E.g., stable ID, reference, keyer, editor, history
- User interface
 - What should the ideal user interface look like?
- API
 - What information and format should we adopt?

Guiding Principle: This roadmap should be driven by the needs and preferences of the research community. Thus, it is critical to offer a public comment period, and let the voices from the community be heard.

Overview of CF news and introduction to hackathon sessions - Notes

Jonathan Gregory: Brief introduction to the CF Committees

Review of CF-1.11 and preview of CF-1.12

New Vocabulary and Discussions repos – discussions vs. issues

Link to presentation: <https://doi.org/10.5281/zenodo.14231114>

Ethan Davis: Brief introduction to the CF Governance Panel, and to the CF Information, Management and Support Team

DOI for the website and the conventions document

Link to presentation: <https://doi.org/10.5281/zenodo.14113656>

"Good housekeeping": website & conventions text

Daniel Lee

Have you ever looked at a piece of the CF Conventions technical infrastructure, or the website, and not been overwhelmed by its awesomeness? Then you're probably one of the few people who belong in this hackathon session. We have a list of niggles wanting prioritisation and resolution and maybe you'd like to be the one to take them on. In this session we will

- Review open issues
- Prioritise them
- Work on the ones that are sized where live collaboration is useful, or where we hope to be able to close them by the end of the workshop
- Flag those that have become obsolete so that we have a bit more ease maintaining the backlog moving forward.

We'll concentrate on [the website issues](#) but good first issues such as [String/Char labels \(section 6\) mismatch between Conformance and CF Document](#) and [Concluding discussions](#) are also welcome.

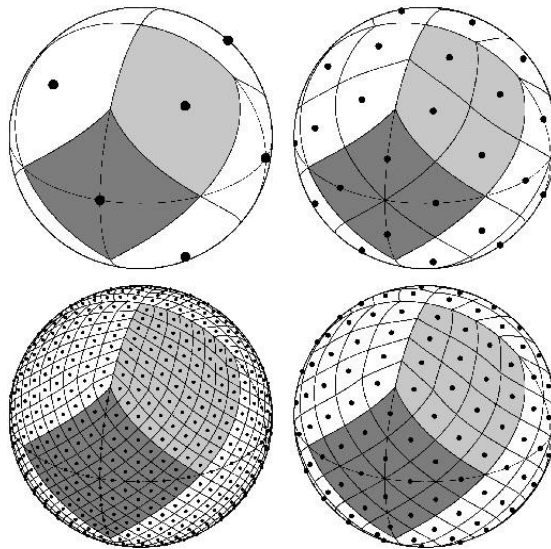
HEALPix grids in CF

David Hassell

HEALPix is an acronym for **H**ierarchical **E**qual **A**rea iso**L**atitude **P**ixelation of a sphere. This pixelation produces a subdivision of a spherical surface in which each pixel covers the same surface area of the globe. A HEALPix grid is such that its resolution is defined by just two scalar parameters, which can take only specific values such that grids of different resolutions are perfectly nested (see the figure, which shows the first four grids: $N = 1, 2, 4$, and 8). This feature makes it useful as a storage grid that is “optimised for hierarchical data exploitation”, meaning that you can easily go to lower resolutions by simple averaging, rather than expensive, and often more destructive, regridding techniques.

These grids are currently not really supported by CF, so **the purpose of this hackathon is to discuss [issue #433](#)**, which proposes a CF-netCDF encoding for a HEALPix grid that uses minimal storage space. The proposed solution has so far gained some support - does that still hold up? - and there are some questions about the scope and the data model that are still open.

References: <https://en.wikipedia.org/wiki/HEALPix>,
<https://iopscience.iop.org/article/10.1086/427976/pdf>



NetCDF-Schema discussion

Guillermo Tesoro Calvo

NetCDF validation is a pervasive issue when dealing with climate products and other datasets encoded using the format. Often this issue is solved several times at different points of the product pipeline, either by manual inspection or the use of bespoke software. Since there is not a clear, standard way of defining a netCDF product, several sources of truth (Documents, Excel files, checker programs, etc) may co-exist, leading to inconsistencies and duplication of efforts.

In this context, netCDF-Schema aims to create a standard way of defining and validating netCDF products, providing a machine- and human-readable single source of truth.

NetCDF-Schema and CF Conventions:

The CF Conventions are a good candidate for a schema to implement for netCDF-Schema. Given how comprehensive and widespread the conventions are when dealing with climate netCDF products, the conventions have been set as the MVP (Minimum Viable Product) for the initial NetCDF-Schema implementation. The rationale is that if the schema has all the necessary features to translate the CF Conventions requirements, it could cover most other use cases.

Current State:

The current work on the NetCDF-Schema is to understand how many features the schema requires to fully support the CF Conventions, and how to implement them as a YAML file. I have laid out a good base for the schema which is coded as a series of **dataclasses** leveraging the validation capabilities of the **marshmallow** library.

Since, in my opinion, the hardest part about this kind of development is to understand how the schema needs to be, there is no validation software yet.

Both the schema and its definition can be found in this [repository](#). Since I'm developing this on company time as part of our R&D activities and I'm waiting for clearance to publish it as open source, the repository is set to private but anybody is more than welcome to request access. Otherwise I will be sharing it during the workshop.

I will be giving a lightning talk on Wednesday about the project with hopefully more details and context.

Hackathon Activities:

Since I'm far from an expert on the CF Conventions, I'd like to use the workshop to go over the schema, making sure that there is support for all CF Conventions requirements, and that no

dark corners are left behind.

Localized metadata & CF/ERDDAP (cancelled)

Erin Turnbull

While English often dominates in the scientific community, it is often beneficial to provide metadata (and, in some cases, data) in other languages so that non-English scientists can understand and use the dataset as well. In some cases, it may even be required, such as the Canadian Government requiring the publication of metadata in both official languages, English and French.

The addition of a standardised methodology for defining the primary language of a CF-compliant NetCDF file's textual metadata/data and for defining alternative versions in different languages will help support the publication of data in different languages. This would enable the development of multilingual tools that can provide a user with the metadata in their language of choice (where it is available).

As an example use case, [ERDDAP](#) is a web application developed by NOAA that makes NetCDF files (among other data sources) available through the Data Access Protocol (DAP), which allows scientists to discover and use the data more easily. It typically leverages the CF convention to define a standardised format for metadata. The web interface for ERDDAP is available in multiple languages, which the user can select between and then browse and download data.

There are currently two gaps in ERDDAP's multilingual feature: (a) the metadata about the dataset cannot be translated if desired and (b) a web accessibility issue when text not matching the language of the page is displayed. Addressing these requires a standardised method for defining the language of the dataset and for making alternative text in other languages available. While ERDDAP could define its own standard, ERDDAP typically borrows from CF and it would be ideal if their approach aligned with the CF conventions.

It is worth noting that a CF standard solution would not just support ERDDAP, but could be applied to other tools as well.

The current proposal is to add an optional feature in the CF standards to allow NetCDF file creators to define the default language for their textual metadata and data, and to allow them to provide alternative versions using a suffix. The proposal has been tested in ERDDAP and shown to address the issues above. In this hackathon, we hope to address any final questions or concerns with the approach and iron out the text to be added to the CF conventions.

CF roadmap/white paper

TBD

[BCP-14](#)

Daniel Lee

The Conventions are legible and easy to read. Nevertheless, they might profit from imperative language like in the Ten Commandments. What is REQUIRED, what SHALL and MUST? What do we RECOMMEND, what is OPTIONAL, and what SHOULD or MAY users do? In this session we'll consider the long-standing proposal to make the normative language in the Conventions explicit, and what steps would be required to get there.

Standard names

Alison Pamment

Visualisations for Standard Names data

Sadie Bartholomew

The number of CF Standard Names has evolved from around 700 in ~2005 to over 4000 by ~2018, and yet more now. It is of interest in terms of tracking and understanding the evolving usage of the names, for example, to be able to see at a glance how the total has evolved over time and per-release, including the names that have been added for a given release and any patterns that may exist in those.

With this motivation, in mid 2020 I created a Python script which extracts names and counts from [the XML defining the names table per release](#) and from that generated a plot of the number over time plotted per-release, as well word clouds of shared phrases and words in the table at given release or set of releases, or added between two given releases (see [Issue 110](#)).

My original intention was to tidy this script and to submit it to the [cf-convention.github.io](#) repository so that it could be run, ideally in an automated way, after any release to add the new release to the totals plot to update it, and to generate a word cloud showing the new names added for the release. But I never have found time to do this. So, in this session, our aims will be to (get through as much as possible of):

- Get the original script working, since it may need updating for e.g. changes in the repo structure;

-
- Figure 1 is a line plot showing the number of standard names in the CF conventions table by date. The x-axis represents the date of table version (marked at Jan 1st) from 2004 to 2020. The left y-axis represents the total number of standard names (0 to 4500), and the right y-axis represents the difference in total number relative to the previous version (0 to 1400). The total number of standard names is shown as a blue line with dots, and the difference in total number is shown as a red dashed line with yellow triangles indicating the change for each version. The total number of standard names increases sharply around 2009 and 2018. The difference in total number is shown as a red dashed line with yellow triangles indicating the change for each version. An inset plot shows the number of standard names in the CF conventions table by date from 2008 to 2020, with blue dots representing the total number of names and yellow triangles representing the difference in total number relative to the previous version.



Creating a high-quality dataset of CF-NetCDF file headers

Ag Stephens

What does a *good* NetCDF file look like? What does a *good* CF file look like? We have examples in the CF-Conventions documentation, and examples of *bad files* that demonstrate non-compliance as part of the CF-Checker test suite. However, many of us that work closely with CF-NetCDF will be able to look at a file header (in human-readable *ncdump* format) and make judgments about whether a file has been well constructed. We can discern the **quality** of a NetCDF file header.

I propose that we start a collaborative effort to curate a dataset of **high-quality CF-NetCDF file headers**, with the following properties:

1. Ensure that all files:
 - a. Are CF-NetCDF compliant
 - b. Meet a minimum standard of metadata requirements (such as the number of global attributes)
 - c. Can be read (and plotted?) by a selection of cross-language tools and libraries
2. Provide examples of data and metadata for all documented flavours/profiles/feature types of CF-NetCDF
3. Provide examples from a range of scientific domains
4. Provide examples from a wide variety of institutions and projects across the world

How could this dataset be used?

1. Test files to improve the robustness of the CF-Checker.
2. Example files to be referenced by the CF-Conventions document.
3. Examples of best practices to guide scientists and data scientists.
4. A published dataset for training Machine Learning models (such as Large Language Models).

Aims of the Hackathon

1. Decide the short-term goals
2. Gather some example data files
3. Explore the requirements for including a file in the dataset
4. Prototype tools for automating gathering and testing data
5. Prototype if/how we might use only file headers and whether they'll work with tools
6. Consider how the data might be streamed into ML tools (PyTorch, TensorFlow)

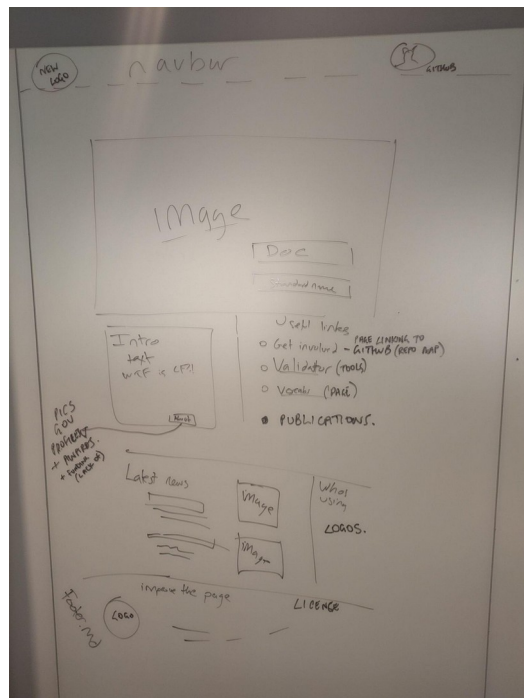
Good Housekeeping Hackathon - Notes

We closed some PRs and set up local test environments, as well as made progress on the link checker! Also improved the instructions for setting up local environments so that it's easier to contribute in the future.

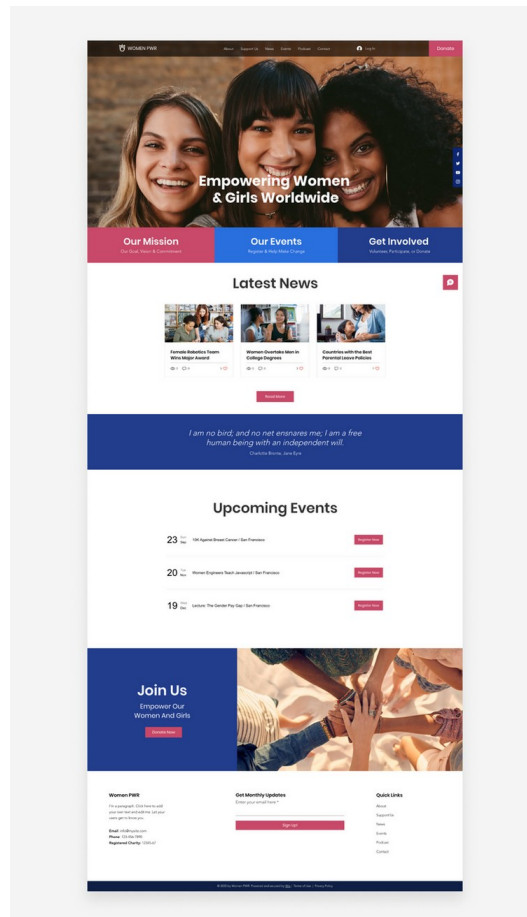
Improving the website homepage idea:

[website working group??? - item in the roadmap]

- Looking at Xarray website as example - templates for website - they are hosting a notebook
- Sphinx - build system - sphinx design shows how we can better layout the website (currently Jekyll)
- Build improvements with auto website build when doing PR for checking before deploying
- Current website using default template
- Name, logo and 'blurb' at top (new logo needed)
- Buttons with quick links
- Images with contributors (profiles with pictures to humanise CF)
- Button to latest version
- We need - access to the assets, contributing 'how to', tools and how to comply,
- What is the purpose and rationale behind each of the headings at the top. Multiple places to access the same thing
- Intent: What do people want to do when they get to the website? Getting started, getting involved, technical docs



- Navigation bar at the top, new logo and logo link to GitHub
- Image sitting transparent and behind the top of the page with the name and button links to our main docs
- Introduction blurb text with button link to 'about' page (this will contain pictures, governance, profiles of real people, awards and how we run (ie. no funding ect.))
- Using columns to give a better look of parts sitting side by side, we can have useful links (Get involved (links to our Discussions page and our github and explaining the repos), links to tools such as cfcheckers, link to vocab page, publications page with link to papers about CF)
- Latest news feed with info about new releases, events, people changes and images of relevant news items for example if CF were speaking at a conference.
- A section with the projects and organisations logos of who is using CF (this gives an idea of the importance and types of big orgs which are using - reputation)
- At bottom (footer section) we have the existing things like 'improve this page' and 'licence information' plus our logo.



HEALPix grids in CF - Notes

- Lars: HEALPix is covered by PROJ:
 - <https://proj.org/en/9.4/operations/projections/healpix.html>
- We should align HEALPix parameters with what in the PROJ.
 - Not all PROJ parameters are in GRIB
- We must have an attribute in the grid mapping variable to specify if the grid_mapping has coordinates (default value) or only dimensions. Attribute name could be: grid_reference
 - Possible values: “coordinate” “dimension” (default if attribute is not there would be “coordinate”).
 - This requires a UML “has a” new connection to be made in the CF data model (between Coordinate Reference and Domain Axis)
- GRIB2 - GRID DEFINITION TEMPLATE 3.150.
 Hierarchical Equal Area isoLatitude Pixelization grid (HEALPix)
https://www.nco.ncep.noaa.gov/pmb/docs/grib2/grib2_doc/grib2_temp3-150.shtml

Created 07/12/2024

Octet No.	Contents
15	Shape of the Earth (See Table 3.2)
16	Scale factor of radius of spherical Earth
17-20	Scaled value of radius of spherical Earth
21	Scale factor of major axis of oblate spheroid Earth
22-25	Scaled value of major axis of oblate spheroid Earth
26	Scale factor of minor axis of oblate spheroid Earth
27-30	Scaled value of minor axis of oblate spheroid Earth
31	Resolution and component flags (see Flag Table 3.3)
32-35	<i>nsides - number of sides within a rhomboid shape</i>
36-39	<i>Lo - Longitude of the centre line of the first rhomboid</i>
40	Grid point position (see Code Table 3.8)
41	<i>Numbering Order (see Code Table 3.12)</i>
42	Scanning mode (see Code Table 3.13)

- Discussed about data defined at edge mid-points and cell vertices, that GRIB allows.

Decided that there is no known use case, and that currently ill-defined how to map data to those locations.

- WKT: Does WKT represent it?
- What about the affine transformation stuff in GeoTiff - is that a projection-type thing without projection coordinates? Or is it more a compression by coordinate subsampling case?

CF Roadmap Hackathon - Notes

Synthesis session slides from Wednesday -- <https://doi.org/10.5281/zenodo.14005753>

Synthesis session notes from Wednesday -- in this document

Publication Timeline

Journal selection

Must be open access

Audience: Broad, access to wider communities not so aware of CF

Broader impact may be harder to get published in

- Nature ... (<https://www.nature.com/sdata/> or [Nature Reviews Earth and Env](#), [Nature Communications?](#))
- Frontiers in ... (<https://www.frontiersin.org/journals/science/>)
- Data Science
- Journal of Geophysical Research: Machine Learning and Computation (AGU)
- EOS
- BAMS ([AMS Publications](#))
- ESSD (<https://www.earth-system-science-data.net/>)
- Journal of Geophysical Research
- Geophysical Model Development
- Open-ended journals?

Journal selection will be continued [on GitHub](#).

Open review process could be beneficial.

Publication Timeline

- Road map elements defined by January 31st 2025
 - Gap analysis, map themes to gaps.
 - Each theme: rationale (use cases); current state in CF, action
 - Various type of action, differing between theme
 - “identify/choose”, or “choose”, or “state a choice”, “CF doesn’t want to do this”, etc.
- Submitted by 31st March 2025
- Published by (well before, ideally) the next CF meeting (September 2025 ???)

This suggested timeline is ignoring the choice of journal - its expected publication lead time may cause us rethink, as might anything else!

Recurrent themes

Scope of CF?

Built in interoperability

Other standards (knowledge map, overlaps and stakeholder map? communication)

Fran: two suggestions for adding to the themes list: tools (cfchecker) was mentioned before and website+engagement (improvements)

Topic	Champions	Home on GitHub
Vocabularies management and development	Alison, Fran, Ellie	https://github.com/orgs/cf-convention/discussions/366
Provenance and Lineage	Christian, Daniel	
Discovery metadata	Andrew, Fran, Ethan	
Recording different types of uncertainty	Lars, Alison, Chris (liaise with Charlotte Pascoe)	
(Horizontal) CRS	Ethan, Lars?, Daniel?	
Time and timelines	Chris L (Met office - via Fran)	
Units ¹⁰	Alison (liaise with Stuart Chalk)	
Identities of complicated variables ¹¹	Lars, Antonio, Christian	
Machine Learning ¹²	Ethan to liaise with Thomas Martin?	
LLM (Large Language Models)	Alison to liaise with Ag Stephens	https://github.com/orgs/cf-convention/discussions/355
NetCDF file formats (serialization formats for CF?)	Sadie	https://github.com/orgs/cf-convention/discussions/359
Interaction with non-netCDF datasets ¹³	Sebastien	
Profiles of use ^{14 15}	Luke Marsden , Ellie Fisher	https://github.com/orgs/cf-

10 *Alison*: and links to vocabularies

11 *Lars*: Related to provenance and lineage (although not necessarily in a straightforward way).

Christian: Does this relate also to climate indicators? What do we mean by "complicated"?

Lars: "complicated" in the sense of CF, meaning that existing CF mechanisms .

12 *Christian*: We may get some insights from the RDA FAIR4ML IG

<https://www.rd-alliance.org/groups/fair-machine-learning-fair4ml-ig/> (comment supported by Chris).

Chris: Other RDA activities may be useful - recommendations on granularity, discovery, citation of dynamic data, etc.

Fran: decision needs to be made by ML community as to how to approach or attributes to add - who is our contact? (ESIP?)

13 *Fran*: such as csv etc.

14 <https://community.wmo.int/en/activity-areas/wis/wmo-cf-extensions>

15 <https://www.ogc.org/standard/netcdf/>

Topic	Champions	Home on GitHub
		convention/discussions/358
Example datasets (training material)	Luke Marsden (Ag Stephens?) , Yiling Liu?	https://github.com/cf-convention/cf-conventions/issues/348
Tools (cfchecker and libraries, schemas)	Luke Marsden, Guillermo Tesoro, Sadie Bartholomew	https://github.com/orgs/cf-convention/discussions/356

Mid october deadline for meeting to give status updates to the items

Do we need to publish workshop summaries?

Short document with concise description and details of CF - could this be separate? [quick win]

1. Introduction to CF
2. Current status 'as-is' and achievements

Structure of roadmap document:

1. Short intro
2. Emerging requirements: Gap analysis (or standards which meet those gaps, does CF need to cover this) - or perhaps not 'gaps' but identified requirements from CF Workshop 2024 (requirements analysis and use cases)
3. 'To-be' and roadmap with detail of each point on roadmap
4. Challenges (preventing the roadmap, risks, resources etc)

BCP 14 Hackathon - Notes

Participants: Antonio, Martin, Jonathan, Lars, David, Daniel, Andrew (prep conversation)

Pros

- clarity
- software developers like it

Cons

- readability - capitals feel like shouting
- entry barrier
- maintenance overhead
- duplication of info with conformance doc

Could we gain clarity just by being careful with language? In 1.11 we have 138 instances of "should"

Could we be "inspired by" BCP 14 without conforming directly to it?

What about words like "deprecated" - what do those mean? Do we want to clarify that in the doc? Lars has found a number of phrases where we might profit from being very unambiguous.

If we're using a controlled vocab, could we have a pre-merge action that flags up our use of it in the new text and asks the merger to verify that those are being used properly? Should be easy to git diff | grep and then pull out all uses of controlled vocab with link to where we say how we stand to them.

Way forward:

- We want to use controlled vocab - BCP14 words + Lars' list after consensus building
Extra words not in Lars's program at present: Suggest, allow, permit, forbid, prohibit
- Do a dry run with caps to see if it makes our brains explode - full caps, small caps, no caps?
- Write down what our stance is towards these words - we might adopt BCP14 or simply be heavily inspired by it in line with our stance vis-a-vis semver
- Update checklist to make it easy to maintain consistency

Note: Our stance toward deprecation in general is not in scope of this issue.

Follow-on in <https://github.com/cf-convention/cf-conventions/issues/546>

Standard Names Hackathon - Notes

Hackathon topics:

Useful links:

<https://cfeditor.ceda.ac.uk/proposals/1>

<https://vocab.nerc.ac.uk/collection/P07/current/>

<https://cfconventions.org/vocabularies.html>

New Tool:

- Guillermo/Gwen will collaborate on a separate tool (Python) to allow quick checking for correct standard names/area types/regions - would check against P07.

How would we propose new mappings between P07 and other vocab collections inside NVS and elsewhere?

Ontology:

- Clarify status of MMI (<https://mmisw.org/ont/cf/parameter>) - mappings are generated in P07 but we don't know if people are using them
- MMI is based on standard names "current" version so this will break if we move where we publish this. Who manages the ontology?

How should we use uRIs in NVS in CF compliant files?

To be FAIR we should definitely include a URI

The opaque URIs are managed robustly in NVS and are part of standard workflow, more concise

The URIs including the standard name are more fragile and more work for NVS to maintain.

Moving to P07 as the canonical source of standard names - what do we need to do?

<https://vocab.nerc.ac.uk/collection/P07/current/>

- Actions for CF community
- Actions for NVS team
- <https://github.com/cf-convention/vocabularies/issues/11>

Justifications:

Impacts and consequences:

- Reference to Standard names xml in the conventions document (appendix B?) - change to p07 NVS
- Tools, python modules and workflows which rely on these files (cfchecker, cfpypthon?, cfplot?, cdo?, iris?)
- Xml could be created automatically from NVS (Andrew Barna)
- Viewing previous version on NVS is currently not working, have had it working previously so will need to bring this back in so we don't need to have them on GitHub and the website (Gwen Moncoiffe)
- Need to consider the KWIK index and how this will be generated in the workflow (Alison) (Need to document this for Antonio).
- Users might want to propose mappings to other vocabularies - how can they do this?
- Investigate UDUNIT-QUDT functionality (<https://qudt.org/schema/qudt/udunitsCode>) as a mean to linking P06 to CF UDUNIT requirement (Andrew Barna)
- Create a new Python package as a standalone checker for standard names and other CF controlled vocabularies (Guillermo Tesoro)

Vocabularies section of CF Roadmap - Open a github discussion to detail the steps (Ellie Fisher)

- What is our longer term vision for interoperability with other controlled vocabularies for parameter naming (not governed by CF)
- Are units part of vocabulary management in terms of providing interoperability (UDUNIT/UCUM/QUDT, etc)?

Work needed on standard names files (these things are for the vocabularies team to work on) (Related to issues started by Lars on improvements to XML and CF committee decisions to reorganise)

- Fix repeated alias entry in V86 HTML
- XSL file - <https://github.com/cf-convention/cf-convention.github.io/issues/471>
- *Create V86 KWIC index (won't do in hackathon!)*
- Move website source files from cf-convention/cf-convention.github.io to cf-convention/vocabularies:

- Steps needed:
 - Clone conventions repo
 - Copy standard names file tree from cf-convention.github.io, also files for area_types, standardized_regions, standard name contributors
 - Commit and push to vocabularies repo
 - Work out what is needed!! to publish the cf-convention vocabularies files to github pages
 - Update links in cf-convention.github.io/vocabularies.md to use the pages published from vocabularies repo
 - Once all of the above are working - remove the standard names directory tree from cf-convention.github.io

High-quality CF-NetCDF for ML Hackathon - Notes

Attendees: Ag Stephens (Remote) (STFC CEDA): ag.stephens@stfc.ac.uk, Luke Marsden (MET Norway), Ethan Davis (Unidata), Carolina Nilsson (SMHI), Yiling Liu (NCI), Alexandra Kokkinaki (NOC-BODC)

The github discussion related to this theme for the roadmap article: <https://github.com/orgs/cf-convention/discussions/355>

Links

Sadie's N-gram poster:

https://figshare.com/articles/poster/Towards_an_explorer_tool_for_visualisation_of_grammatical_patterns_in_the_CF_Standard_Names_through_decomposition_into_ini-grams/24750393

Possible good quality datasets

Data URL	Provider	Brief description	featureType
http://nbstds.met.no/thredds/dodsC/NBS/S2A/2024/09/19/S2A_MSIL1C_20240919T123811_N0511_R095_T33XWF_20240919T162127.nc.html	MET Norway	S2A satellite data	
https://thredds.nci.org.au/thredds/dodsC/uc0/sample/orog_AUS-20i_CMCC-ESM2_historical_r1i1p1f1_CSIRO_CCAM-v2203-SN_v1-r1.nc.html	CSIRO Australia	CORDEX-CMIP6 sample file	Includes CF1.11 features
https://opendap1.nodc.no/opendap/physics/point/cruise/nansen_legacy-single_profile/NMDC_Nansen-Legacy_PR_CT_58US_2021708/CTD_station_P1_NLEG01-1_-_Nansen_Legacy_Cruise_-_2021_Joint_Cruise_2-1.nc.html	Norwegian Marine Data Centre, Nansen Legacy project	Vertical profile CTD data	profile
https://dap.ceda.ac.uk/thredds/dodsC/neodc/esacci/sea_ice/data/sea_ice_concentration/L4/ssmi_ssmis/12.5km/v3.0/NH/2007/09/ESACCI-SEAICE-L4-SICONC-RE_SSMI_12.5kmEASE2-NH-20070915-fv3.0.nc.html	MET Norway	Sea ice concentration maps	

GPT Prompts

Create a CDL for a typical weather station

Add in some random but realistic data in the data section. Include 5 equally spaced time steps.

Reduce the variables to Air Temperature and Relative Humidity. Increase the number of stations to 3, and provide relevant CF-compliant metadata about those stations.

Recreate this CDL file, but include a vertical profile at each of the stations.

Can you add in the relevant feature type metadata

--

Create a CDL for a typical weather station, with only two data variables, include vertical profiles and 5 equally spaced time steps, and the relevant feature type metadata.

What have we learnt?

1. Regarding use of LLMs:

- ChatGPT already has *some* knowledge and skills around CF and NetCDF
 - It has already seen a lot of open-source material that is relevant (GitHub and documentation)
- It would be more efficient to investigate simplest solutions first:
 - ChatGPT free version
 - GPT / Assistant acting as a CF-Agent:
 - Prompted by text, documents, examples, rules etc
 - More involved approaches:
 - RAG
 - Fine-tuning
- Every few months: need to revert back to the simplest solution to check how well it works
- Rules might be better than example datasets
- Synthetic example data might be useful
- Measuring performance is hard:
 - How do we compare different solutions?
 - How do we measure progress/accuracy?

2. Regarding collection of "high-quality" CF-NetCDF files:

- We agree there would be significant value in gathering these, particularly to:
 - guide users in best practice
 - for use with the CF-Conventions documentation
- Synthetic example data might be useful
- Could ChatGPT help us create these examples?

What happens next?

- AS: Start a GitHub discussion to describe the problem/idea and gather a community to work together

Hackathon Wrap-ups

CF Standard Name Hackathon - Making NVS P07 the main source for standard names

- Had discussions about the impacts and consequences.
- Had discussions about the benefits and further developments because of this change
- Created a discussion #366 <https://github.com/orgs/cf-convention/discussions/366>
- Potential change to conventions document - appendix B <https://cfconventions.org/cf-conventions/cf-conventions.html#standard-name-table-format>
- Possible considerations with units

Visualisation of standard names

- Made the visualisation plot as planned!
- More work to do - to add page to website with graph on - will create PR next week

Datasets to train ML models

- Discussing high quality dataset
- Brain stormed how it might work and why needed
- Requirements: useful to have examples or template files integrated with conventions document - snippets in the document can be accessed in full online by users
- Experimented with ChatGPT and AGPT - help construct NetCDF file
- Test out simplest solutions first - what do you want to get out of it - what can't it do
- Set up open AI assistant - given more info
- GitHub Discussion to document and explore the findings and more discussion to take forward

BCP-14

- Should formalise use of these words
- Write something in the conventions to say following a convention
- Either BCP or something similar
- To be BCP compliant - 14 words - must be in CAPS - not many agreed to this.
- One option is to have caps but with small font size
- Also might extend the BCP words list - suggested?
- Going to identify where these types of words are being used in the doc - about 1000 instances
- Actions: look at this highlighted text, evaluate meaning and change where needed, agree on format.
- Aim for 1.13 release of CF document

- Multiple people are needed to work on this and break it down by chapter perhaps
- Conformance document still needed - no changes likely unless consistency of meaning.
- Maintenance: automatic flag/check on pull request review process
- Issue already exists <https://github.com/cf-convention/cf-conventions/issues/546>

HEALPix:

- Class of spherical projections
- The normal ways of doing projections has been discussed
- The issue has been open for a long time
- Different type of projection from others - no coordinates in the projection space - nothing to determine lat lon
- Difficult to represent in CF - mechanism is there so need extending
- Generic grid mapping variable will be added - default to coordinates
- Other polar grids could make use of this
- HEALPix is defined by PROJ
- Will document this discussion in the existing issue and bring this to a conclusion once this is implemented.

Good housekeeping:

- Website homepage improvements
- Having local instance of website set up on machine
- Some people worked on specific tickets - Fran added NVS link to vocab website page whilst the wider change is discussed.

Workshop Wrap-up

Deadline for release 1.12

- Agreed to have a deadline for the next release after CF Workshop meetings:
 - 5 weeks for last-minute contributions
 - 3 weeks for cooling-off contributions
 - Deadline: 11th of November 2024
 - But still dependent to high/important issues

2024 AGU Open Science Recognition Prize

- 2024 Fall Meeting
 - Washington, D.C. 9-13 December 2024
 - Ethan will represent CF at the plenary session recognition
 - Some few spaces available for AGU President's dinner reception
 - Don't wait for your ticket!!!

The CF roadmap for 2030 challenge

- Set path of development of CF over the next five years
- CF Community ownership and engagement
- Publish it!

Homework

- Start to prepare next on-line 2025 CF Workshop
- Volunteers, or ideas, for hosting the next in-person 2026 CF Workshop
- Evangelize CF:
 - *To promulgate or promote (a doctrine or idea, for example) enthusiastically*

Acknowledgments

- To all on-line and in-person participants and presenters
- To the organizing committee for putting the program together
- SMHI to support the logistics CF Workshop,
 - Thumbs-up for the SMHI Canteen food
- Thanks to Lars Barring and colleagues for hosting us

