NetCDF at the 4TU.Centre for Research Data

Maria Cruz and Egbert Gramsbergen, TU Delft, July 2018

With significant contributions to the "Technical Services" section from (listed in alphabetical order) Kees den Heijer (TU Delft), Niels Drost (Netherlands eScience Center), Chris Schubert (Climate Change Centre Austria), and Julia Wagemann (European Centre for Medium-Range Weather Forecasts). With special thanks to Marta Teperek and Alastair Dunning for their many comments and invaluable feedback.

Executive summary

The 4TU.Centre for Research Data (short version: 4TU.ResearchData) was started in 2008 as a collaboration of the libraries of three universities of technology in the Netherlands: Delft University of Technology, Eindhoven University of Technology, and the University of Twente. The data archive, which has been fully operational since about 2010, collects data in science and engineering in a permanent and sustainable manner.

Presently, around 90% of the data (both in terms of volume and number of datasets) stored in the archive are atmospheric and environmental research datasets coded in netCDF – a data format and model that, although generic, is mainly and widely used in climate and atmospheric sciences and oceanography.

4TU.ResearchData has a special interest in this area and it offers specific services and tools to enhance the access to netCDF datasets. In particular, netCDF files can be accessed via the OPeNDAP (Open-source Project for a Network Data Access Protocol) protocol, the main advantage of which is the ability to retrieve subsets of files without the need to download whole datasets.

To better understand netCDF data at 4TU.ResearchData – the datasets and their contributors as well as the services and their users – we conducted desk-based research and a series of semi-structured qualitative interviews with researchers based in the Netherlands who use and produce netCDF data.

This report provides an overview of the current data and services and explores options for 4TU.ResearchData to expand its services related to netCDF data. The analysis is broad in scope, assessing opportunities for creating not just technical services related to storing and archiving netCDF data, but also for advice and guidance, and the advantages that could accrue from building a community of data depositors and users.

Our main conclusions are that the creators and users of the netCDF data stored in 4TU.ResearchData represent heterogeneous research communities within the Earth sciences. They have different views and attitudes to data archiving and data publishing, and store netCDF datasets with very different spatio-temporal characteristics in the archive. Ensuring that any new and current netCDF services continue to be relevant to these communities will require taking this diversity into account.

A need for training and guidance – particularly on data management aspects related to documentation, metadata standards and conventions – is the common thread uniting these communities. This will provide the way forward for 4TU.ResearchData to build a community of data depositors and users.

Expanding technical services beyond what is already provided might be more difficult, given the diversity of the data and the depositors and users, but there are a few services that could help support community building efforts, with the desired goal being higher-quality data and increased rates of data reuse.

Recommendations:

- Organise a workshop on netCDF metadata standards and conventions, focussing on how 4TU.ResearchData could provide training, advice, and guidance in this area. This workshop should include researchers who use and deposit netCDF data in 4TU.ResearchData and other partners in the Netherlands and beyond (from the research community, industry, and similar service providers), who are also interested in this topic. The workshop would serve as a first step to build a national community and a way to continue to build links internationally.

- To help accrue higher-quality data and promote data reuse and to support community building efforts, strengthen and improve communications about 4TU.ResearchData and its netCDF data and services, nationally and internationally.

- Host and keep updating and improving the existing netCDF Kickstarter open source tool, which provides templates for the production of netCDF files conforming to community-defined metadata standards and conventions. In line with the two previous recommendations, promote its use through training sessions and workshops.

- Because there is an important trend in big data, and in particular Earth observation data, for bringing the users to the data, further explore the feasibility of offering a customizable interface to analyse and visualise netCDF data. Such an environment could be created with Jupyter Notebooks/Labs. It might require cooperation from the TU Delft ICT department and some software development expertise.

Aims and scope

The 4TU.Centre for Research Data (short version: 4TU.ResearchData; formerly known as '3TU.DataCentrum') began in 2008 as a collaboration and initiative of the libraries of Delft University of Technology (TU Delft), Eindhoven University of Technology (TU Eindhoven), and the University of Twente.

4TU.ResearchData was originally built "as a data curation facility to meet the diverse needs of heterogeneous research communities" (Rombouts and Princic, 2010). The archive, which has been fully operational since about 2010, collects data from many fields and subjects in science and engineering.

Presently, around 90% of the data (both in terms of volume and number of datasets) in the 4TU.ResearchData archive are atmospheric and environmental research datasets stored as netCDF. Thus, 4TU.ResearchData has a special interest in this area and it offers specific services and tools to enhance the access to netCDF datasets.

This report provides an overview of these data and services and explores options for 4TU.ResearchData to expand its services related to netCDF data. The analysis is broad in scope, assessing opportunities for creating not just technical services related to storing and archiving netCDF data, but also for advice and guidance, and the advantages that could accrue from building a community of data depositors and users.

This work is part of a larger endeavour to ensure that 4TU.ResearchData remains relevant and successful in the long term in a rapidly evolving research data management landscape (Cruz *et al.*, 2018). It is underpinned by two key ideas:

1. Repositories need to have a subject or format focus to remain relevant and be successful in the long term (Cruz, 2018a).
2. Data reuse requires well-informed, sustainable, inclusive, participatory development of data infrastructures (Leonelli, 2017).

Desk-based research was supplemented by a series of semi-structured qualitative interviews with researchers based in the Netherlands who use and produce netCDF data.

What is netCDF?

The Network Common Data Form, or netCDF, is more than just a data format. It is a data format and a data model for scientific data and metadata; it is a set of software libraries that allow storage, access, and sharing of array-oriented data[1]. NetCDF[2] was developed in the late 1980s and is maintained by Unidata[3].

Although the format is generic enough to be used in any discipline or subject where the data can be modelled as an annotated, multi-dimensional array, netCDF is mainly and widely used in oceanography, climate and atmospheric sciences. Although it is not an intrinsically geospatial format, netCDF is widely used for geospatial data[4].

As described in Unidata documentation[5], netCDF data is:

- *Self-Describing*. A netCDF file includes information about the data it contains.
- *Portable*. A netCDF file can be accessed by computers with different ways of storing integers, characters, and floating-point numbers.
- *Scalable*. A small subset of a large dataset may be accessed efficiently.
- *Appendable*. Data may be appended to a properly structured netCDF file without copying the dataset or redefining its structure.
- *Shareable*. One writer and multiple readers may simultaneously access the same netCDF file.
- *Archivable*. Access to all earlier forms of netCDF data will be supported by current and future versions of the software.

Thus, the format is very interesting from an archival and data sharing point of view.

NetCDF and the CF metadata standards and conventions

NetCDF data are self-describing in a fully machine-readable way in that metadata are included together with the data in one single container – the data file itself. However, as noted by Rew and Davis (1997):

---

[1] "Unidata | NetCDF." https://www.unidata.ucar.edu/netcdf/. Last accessed 16 July 2018.
[2] Since version 4 (the latest), netCDF is based on the more generic Hierarchical Data Format (HDF). NetCDF4 is a subset of HDF5, meaning that a HDF5 file that avoids some specific non-netCDF-compatible constructs can be treated as netCDF.
[3] Unidata is a community of education and research institutions with the common goal of sharing geoscience data and the tools to access and visualize that data. Unidata is a member of the UCAR Community Programs, managed by the University Corporation for Atmospheric Research (UCAR) and funded by the US National Science Foundation.
[4] "NetCDF-4 (Network Common Data Form, version 4)." 11 Apr. 2017, https://www.loc.gov/preservation/digital/formats/fdd/fdd000332.shtml. Last accessed 16 July 2018.
[5] "NetCDF: FAQ - Unidata." https://www.unidata.ucar.edu/software/netcdf/docs_rc/faq.html. Last accessed 16 July 2018.

"The extent to which data can be completely self-describing is limited: there is always some assumed context without which sharing and archiving data would be impractical. NetCDF permits storing meaningful names for variables, dimensions, and attributes; units of measure in a form that can be used in computations; text strings for attribute values that apply to an entire data set; and simple kinds of coordinate system information. But for more complex kinds of metadata (for example, the information necessary to provide accurate georeferencing of data on unusual grids or from satellite images), it is often necessary to develop conventions."

The netCDF Climate and Forecast (CF) conventions are an example of a set of community-driven conventions used for the description of Earth sciences data. They are one of the recommended standards by Unidata[6]. The CF conventions[7]

"define metadata that provide a definitive description of what the data in each variable represents, and the spatial and temporal properties of the data. This enables users of data from different sources to decide which quantities are comparable, and facilitates building applications with powerful extraction, regridding, and display capabilities."

NetCDF and FAIR data

A set of 15 guiding principles to make data Findable, Accessible, Interoperable, and Reusable (FAIR) were published in 2016 (Wilkinson *et al.*, 2016). The FAIR principles were almost immediately widely adopted by national funding agencies, including the Netherlands Organisation for Scientific Research (NWO) and the European Commission. In July 2016, the European Commission issued guidelines[8] on FAIR data management in Horizon 2020[9], the Commission's eighth framework programme funding research, technological development, and innovation.

NetCDF files that include extensive (rich) metadata, including both intrinsic and contextual metadata, and comply with domain-relevant standards, such as the CF conventions, already meet many of the FAIR principles regarding Findability (F), Interoperability (I), and Reusability (R). If those files are made available via an archive such as 4TU.ResearchData – which uses standardised and indexed metadata, and releases each dataset with a clearly defined licence and a globally unique and persistent identifier – then netCDF data can meet all of the FAIR principles (Cruz and Gramsbergen, 2018a), including those pertaining to Accessibility (A).

---

[6] "NetCDF Conventions." https://www.unidata.ucar.edu/netcdf/conventions.html. Last accessed 16 July 2018.
[7] "CF Conventions." http://cfconventions.org/. Last accessed 16 July 2018.
[8] "Guidelines on FAIR Data Management in Horizon 2020 - European ...." 26 Jul. 2016, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf. Last accessed 16 July 2018.
[9] "Horizon 2020 - European Commission ...." https://ec.europa.eu/programmes/horizon2020/. Last accessed 16 July 2018.

Overview of netCDF data and services at 4TU.ResearchData

The data

As of 16 July 2018, 4TU.ResearchData stored 7631 datasets, corresponding to 32.8 TB of data (Table 1). Of these datasets, 6526 were stored as netCDF, corresponding to 30.3 TB of data, or 92% of the total data in volume (85.5% of the total of number of datasets). The vast majority of netCDF datasets in 4TU.ResearchData originate from TU Delft, in particular from the Faculty of Civil Engineering and Geosciences. This is consistent with the wide use of netCDF in some areas of the Earth sciences, namely hydrology, climate, ocean, and atmospheric sciences.

Institution	**All data**		**NetCDF data**	
	Number of datasets	Size (TB)	Number of datasets	Size (TB)
TU Delft	7309	29.4	6470	27.3
TU Eindhoven	123	3.0	24	2.9
Univ. of Twente	53	0.2	0	0
Wageningen U&R	21	0.04	10	0.01
Other	125	0.2	22	0.1
Total	**7631**	**32.8**	**6526**	**30.3**

Table 1. Data stored in 4TU.ResearchData as of 16 July 2018 by the institution of the data creator. Other institutions contributing netCDF data to 4TU.ResearchData include the Royal Netherlands Institute for Sea Research (NIOZ), the Royal Netherlands Meteorological Institute (KNMI), the Finnish Meteorological Institute, and Warsaw University of Life Sciences in Poland, among many others.

In terms of volume, most of the netCDF data stored in the 4TU.ResearchData come from one single experiment – the IRCTR Drizzle Radar (IDRA), developed at TU Delft's International Research Centre for Telecommunications and Radar (IRCTR) and installed on top of the Dutch meteorological observatory at Cabauw in the Netherlands (Otto and Russchenberg, 2014). This project has contributed 2325 datasets to date, corresponding to about 27 TB of data (Otto *et al.*, 2010). This is a growing time series of datasets, updated every few months, providing detailed observations of the spatial and temporal distribution of rainfall and drizzle around the radar's location.

The remaining 4201 netCDF datasets, corresponding to a total of about 3.3 TB of data, are either part of much smaller collections (less than 20 GB in size), or are individual datasets that range in size from about 500 KB to 200 GB. With a few exceptions (Voorhoeve and van der Maas, 2016), these datasets can be broadly classified as environmental research data, ranging from river discharge data (Hellebrand, 2004) to measurements of aeolian sediment transport (Hoonhout, de Vries and Cohn, 2016), and from climate projections (Mezghani, Dobler and Haugen, 2016) to local mean sea level models (Gerkema and Duran Matute, 2017) (Figure 1). The spatio-temporal characteristics of these datasets are very heterogeneous. Some are data from a single point (station data) with or without the coordinates in the global metadata; some are data for a collection of stations; some are grid data, with some following the CF conventions, but many not; and all may have a temporal dimension or not.

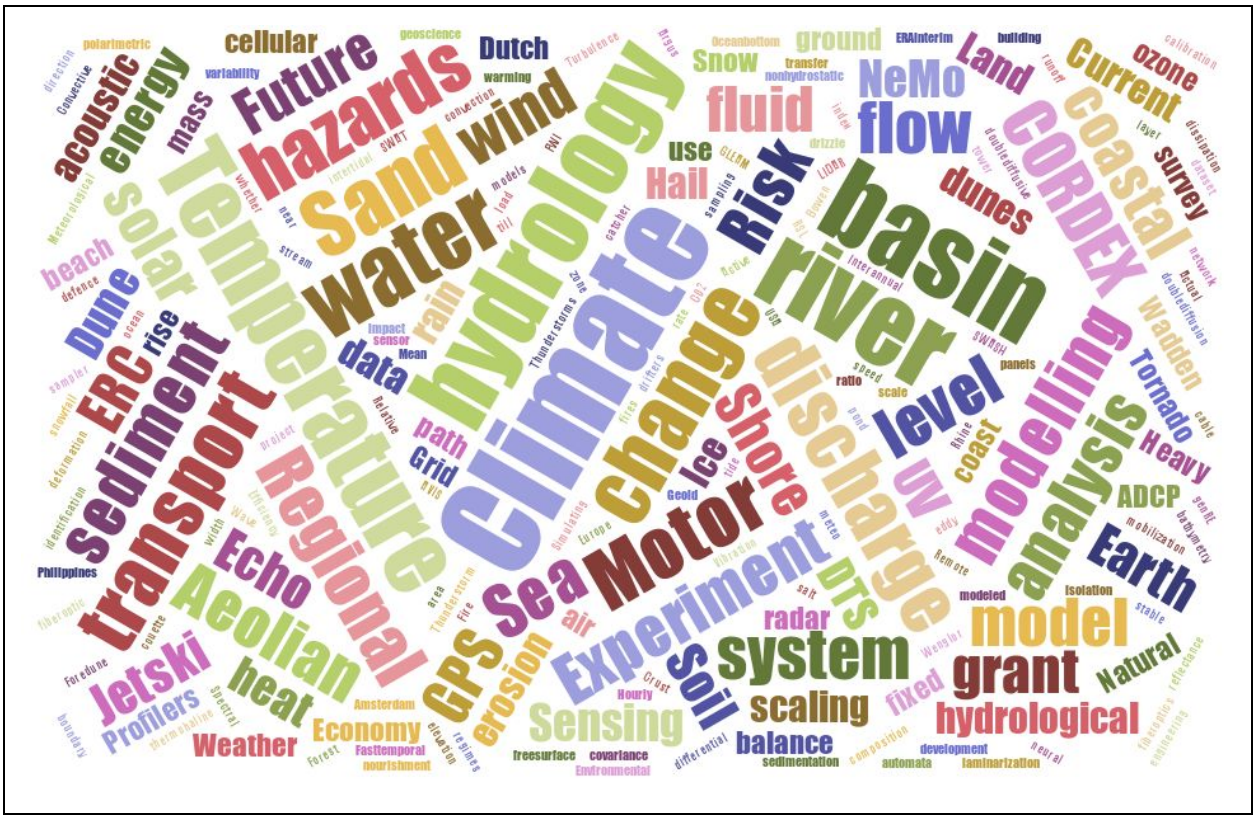

Figure 1. Word cloud based on the 'subject' (key word that describes the topic of each dataset) metadata associated with all the netCDF datasets stored in 4TU.ResearchData as of the end of May 2018. Time series of datasets were only counted once, meaning that the IDRA collection of datasets was only counted once. It is clear that a wide variety of subjects is represented in the netCDF datasets stored in 4TU.ResearchData, with climate change and hydrology featuring most prominently. Overall, with a few exceptions, this list of topics falls broadly under Earth sciences and environmental research.

The services

For all netCDF datasets, besides the usual http download, 4TU.ResearchData offers OPeNDAP (Open-source Project for a Network Data Access Protocol)[10] access since 2011. Some of the benefits of using this protocol include: viewing internal metadata hidden in the data files without having to download the files; accessing slices and subsamples of datasets without having to download the full datasets; access to data with APIs (Application Programming Interfaces) for Java, Python, R, MATLAB, and other programming languages commonly used by researchers. For implementation of the OPeNDAP protocol, 4TU.ResearchData uses Unidata's Thematic Real-time Environmental Distributed Data Services (THREDDS)[11].

For large series of datasets, such as the IDRA collection, 4TU.ResearchData offers the option of making custom agreements about data ingestion, data aggregation, and metadata enrichment.

Data conversion services have been also occasionally provided. For example, in 2011, after 4TU.ResearchData fully embraced the netCDF format, data from the DARELUX (Data Archiving River Environment LUXemburg) project[12] were converted first from the original bespoke xml format to a more netCDF-friendly xml format and then to netCDF. For later additions to DARELUX, csv files were converted to netCDF. Conversion to netCDF increased the ways the data could be interacted with[13] (via the OPeNDAP protocol).

For the Zandmotor project (Stive *et al.*, 2013), 4TU.ResearchData, in collaboration with Deltares[14], created an online environment called Zandmotor Datalab – a single place online where active research data could be stored, shared, edited, processed, and visualised. The DataLab consisted of a THREDDS Data Server with auxiliary components for access and uploading, processing with Python and MATLAB, geographical searching and visualisation, and a database. The complexity of the DataLab, with all its components, required significant effort for maintenance. At the end of the project and upon closure of the DataLab, selected netCDF data were transferred to 4TU.ResearchData's OPeNDAP server (Rijkswaterstaat, 2017). The software was rebranded as OpenEarth DataLab[15] and transferred to Deltares as an open source project[16].

---

[10] "Home | OPeNDAP™." https://www.opendap.org/. Last accessed 16 July 2018.
[11] "Unidata | THREDDS Data Server (TDS)." https://www.unidata.ucar.edu/software/thredds/current/tds/. Last accessed 16 July 2018.
[12] "4TU - Collection: Darelux - River Environment Luxemburg." https://data.4tu.nl/repository/collection:darelux. Last accessed 16 July 2018.
[13] "Researchers about 4TU.ResearchData." Available online: http://researchdata.4tu.nl/fileadmin/editor_upload/Brochure/Brochure__3TU.Datacentrum_2014.pdf. Last accessed 16 July 2018.
[14] "Deltares." https://www.deltares.nl/en/. Last accessed 16 July 2018.
[15] "OpenEarth DataLab - OpenEarth - Deltares Public Wiki." https://publicwiki.deltares.nl/display/OET/OpenEarth+DataLab. Last accessed 17 July 2018.
[16] "GitHub - openearth/datalab: OpenEarth DataLab." https://github.com/openearth/datalab. Last accessed 17 July 2018.

Interviews with netCDF data depositors and users

To better understand 4TU.ResearchData's community of netCDF data contributors and users, we conducted nine semi-structured qualitative interviews with 11 researchers, all based in the Netherlands, who used and produced netCDF data. Most of these researchers deposited netCDF datasets in the 4TU.ResearchData archive; only two of them hadn't done so.

Methodology

We interviewed researchers at all career stages, namely: four PhD students, one post-doc, one senior scientist, three assistant professors, one associate professor, and one full professor. They were all geoscientists, working in areas ranging from atmospheric sciences and remote sensing, to hydrology, oceanography, and coastal engineering. They were mostly affiliated with technical universities in the Netherlands, especially TU Delft, but some were based at national research facilities and industry.

The interviews lasted around 60 minutes each and were all conducted face to face between November 2017 and April 2018. The conversation focussed on netCDF data. The interviewers took notes of key points during the interview and wrote preliminary, more extended reports in the day or so after the interview. All interviewees were informed that the findings were going to be published, but were assured that they wouldn't be named and that no information would be individually attributed to them.

The first results of the interviews were reported at the PV 2018 conference (Cruz, 2018b), held between 15th and 17th May 2018 at the Rutherford Appleton Laboratory, Harwell Space Cluster, in the United Kingdom. The paper for the conference proceedings (Cruz and Gramsbergen, 2018b) was shared with all the interviewees. The results presented in that paper are re-described below.

Main findings

Use of the archive

The results reported in this subsection are mostly from the interviews with the data depositors (9 out of the 11 researchers we interviewed). By use, we mean use of the archive as a netCDF data contributor, not as a data user.

1. What data did they deposit?
    a. Some projects chose to store only raw data for long-term preservation; processed data for comparative analyses were stored elsewhere.
    b. Other projects stored both raw and processed data together with the software and scripts used to process the data.
    c. In some cases, only processed, output data used to produce the figures in a journal publication were archived.

2. Why did they choose to archive their netCDF data?
    a. The need for long-term preservation was particularly important for researchers dealing with climate data and long-term data series.
    b. For many of the researchers, while they appreciated the benefits of long-term preservation and of making their data publicly available, their main motivation was to comply with publisher and journal requirements regarding data availability.
    c. The potential for data reuse and data citation advantage (Piwowar and Vision, 2013) were important motivations for most of the interviewees.

3. Why did they choose 4TU.ResearchData?
    a. Many data depositors chose 4TU.ResearchData because it was locally available at TU Delft; the vast majority of netCDF datasets in the archive come from TU Delft (Table 1).
    b. Most chose 4TU.ResearchData after the recommendation of a colleague, supervisor, or data librarian.
    c. For the most part, the interviewees didn't consider other archives; 4TU.ResearchData was their first and only choice.

4. Were any of the netCDF specific services offered by 4TU.ResearchData key to this choice?
    a. Overall, the OPeNDAP protocol did not seem to have had much influence in the choice of archive for most of the data depositors.
    b. A minority of the researchers were not aware of OPeNDAP and its functionalities; many knew about OPeNDAP, but were just not fully aware that 4TU.ResearchData provided it as a service.
    c. Others knew about this service but did not consider it important, mostly because their datasets were not large enough for them to care about retrieving data subsets or metadata without the need to download entire datasets.

5. What additional services would be useful to have?
    a. Researchers affiliated with projects that chose to deposit raw data mentioned the need and appreciation for processing and visualisation services, but all things considered, they didn't feel this should be a priority or a main role for the 4TU.ResearchData. In their opinion, archiving of (mostly) raw data for long-term preservation should be the focus.
    b. One respondent suggested that 4TU.ResearchData could also store software and initialisation parameters that could then run on a high-performance computing cluster.
    c. Some researchers, particularly early career researchers, mentioned that it would be useful to have templates for the production of netCDF files conforming to specific community standards and metadata conventions, such as the CF conventions.

Use and knowledge of netCDF

In this subsection, we report results from all the interviews.

1. Was netCDF the standard format in their research communities?
    a. For the majority of interviewees, netCDF was the standard data format and model adopted by their communities and it was the primary data format they used and handled.
    b. For a few researchers, netCDF was not a standard in their communities. In some cases, netCDF was used out of choice because of its self-describing properties and interoperability; in other cases, it was simply because it was the output format of commonly used models or software packages.

2. Had the researchers received any formal training?
    a. None of the researchers we interviewed had received formal training on the use or production of netCDF files.
    b. Most of them started using netCDF during their PhD studies and learned by reading manuals and documentation, through advice from peers and colleagues, and just simply by trial and error.
    c. Because most researchers hadn't learned about netCDF in a structured way, they sometimes had gaps in their knowledge.

3. What were those knowledge gaps?
    a. We noticed that some researchers, who would have benefited from the use of OPeNDAP and its functionalities, were not aware of its existence.
    b. With a few notable exceptions, we also noticed a general lack of awareness of the importance of metadata, which can be included in netCDF files. This lack of awareness seemed to sometimes translate to a lack of attention or adherence to metadata standards and conventions.
    c. Some researchers noted that it took them quite a while to learn about useful netCDF tools (e.g. Climate Data Operators) and conventions (e.g. CF conventions) which they wished they had learned about earlier in their careers.

4. Would they be happy to receive training?
    a. There was a general, but not unanimous recognition that there was a need for training, particularly on the research data management aspects of handling netCDF data (e.g. how to include metadata, what metadata to include, and how to adhere to conventions and community standards).
    b. Most of the early career researchers, mainly PhD students, were enthusiastic about receiving formal and in-depth training.

c. The more senior researchers recognised the need for training, but mostly for their PhD students. At their level, they felt that training would put too much of a burden on their already busy schedules.
d. Short training sessions with high-level information about what is possible and available would be the format that would be most welcome by busy senior researchers.

What services and improvements could we offer in the future?

The interviews showed that the creators and users of the netCDF data stored in 4TU.ResearchData represent heterogeneous research communities within the Earth sciences. They have different views and attitudes to data archiving and data publishing, and store netCDF datasets with very different spatio-temporal characteristics in the archive (Figure 1). Ensuring that any new and current netCDF services continue to be relevant to these communities will require taking this diversity into account.

Training, advice, and guidance

As shown by the interviews, a need for training, advice, and guidance – particularly on data management aspects related to documentation, metadata standards and conventions specific to netCDF – may be the common thread uniting the different research communities using and contributing data to 4TU.ResearchData. This may provide the way forward for 4TU.ResearchData to build a community of data depositors and users, which in itself may lead to higher quality datasets and higher levels of dataset reuse. As noted by Leonelli (2017), well-informed, inclusive, and participatory development of data infrastructures is expected to lead to an increase in the quality and re-usability of research data. The challenge will be to find ways to provide high-quality training and advice that is useful and well received by both senior and early career researchers.

Technical services

Besides training, advice, and guidance, we might be able to offer extra technical services in addition to those already provided by OPeNDAP as a standard. These extra services broadly fall into two categories: the production and the consumption of netCDF data. Below we assess the use cases and the feasibility of these services. Some of them are directly related to the results of the interviews; others are services that we evaluated with input from colleagues from other netCDF data providers.

Production of netCDF

General support for the production of netCDF exists in all major programming environments commonly used in science and engineering. This support may be either built-in or provided by easy-to-install extra software packages.

Service: Provision of templates for the production of netCDF files conforming to specific conventions
Namely, the CF conventions, which are widely used.

Use: Because full CF compliance requires a lot of metadata to be included in netCDF files and this is not always easy and straightforward, some of the interviewed researchers indicated that the provision of templates would help them save time and redundant effort.
Feasibility: A service such as this already exists – the "netCDF Kickstarter"[17] started by OpenEarth[18]. This tool builds a skeleton for a program in a number of programming languages. We can build on that. The the service would need to be hosted by the TU Delft ICT department at the 4tu.nl domain; a modest amount of software development would be needed initially to adapt the netCDF Kickstarter. Ideally, a few researchers who are active in the field would be involved in keeping the service in sync with current needs. This service could also provide links to other useful and related tools, such as CF compliance checkers[19] and other template providers[20]. This service would also be suitable for inclusion in training and workshops for researchers.

Service: Conversion of non-netCDF data to netCDF
Use: Occasional. It was done twice (on batches of datasets) in the lifetime of 4TU.ResearchData.
Feasibility: This can still be done very occasionally, but as a service, it is not scalable. A better alternative is to provide guidance, advice, and training so that researchers can produce netCDF data earlier in the research data lifecycle.

Consumption of netCDF

OpenDAP already provides a number of valuable services for netCDF and related file formats, such as an API to access the data without downloading whole datasets, and the aggregation of datasets with a shared dimension, e.g. a time series. Here, we explore extra services that we might be able to offer. The evaluations of expected use are based on current data collections. Future evaluations might differ (generally becoming more positive) if data collections grow and their characteristics evolve (e.g. better CF compliance, more geodata), potentially as a result of training and community building and engagement.

It is also worth noting that there is an an important trend in big data, and in particular Earth observation data, for bringing users and their analytical tools to the data, rather than users downloading datasets for analysis on their own computers or computational facilities (Wagemann *et al.*, 2017; Delgado Blasco *et al.*, 2016). 4TU.ResearchData does not cater for big data (anything larger than a few TB) and could not compete with the large data providers in Earth observation, such as e.g. NASA and ESA. However, the researchers who deposit netCDF in 4TU.ResearchData may also consume data from these large data providers and get used to their tools and any new trends in big data provision. We've seen this during the interviews when one of the researchers showed us how he used Google Earth Engine[21] to handle Earth observation data. Thus, 4TU.ResearchData may also need to provide ways for researchers to bring

[17] "netCDF Kickstarter." http://zandmotor.citg.tudelft.nl/netcdfkickstarter/. Last accessed 17 July 2018.
[18] "netCDF kickstarter - OpenEarth - Deltares Public Wiki." https://publicwiki.deltares.nl/display/OET/netCDF+kickstarter. Last accessed 17 July 2018.
[19] See http://cfconventions.org/compliance-checker.html for two examples. Last accessed 17 July 2018.
[20] For example, the TU Delft Hydraulic Engineering initiative to connect to SeaDataNet, a distributed Marine Data Infrastructure. This initiative will involve the development of templates that are CF and SeaDataNet compliant.
[21] "Google Earth Engine." https://earthengine.google.com/. Last accessed 17 July 2018.

their own analysis tools to the data. We suggest one such service below (see 'Customizable interface to analyse and visualise data').

Service: NetCDF subsetting service (NCSS)

Out of the box, OPeNDAP provides a form of subsetting based on the indices of dimension variables. Example: "give me all the values of x and y for the first 100 values of dimension x". It is not possible to ask directly "give me all the values of x and y where x is between 5.2 and 18.3". For queries such as these, the optional NCSS exists. NCSS biggest advantage lies in its additional features for geographical datasets. Queries may contain easily understandable variables such as North, East, South, West box boundaries and perform on-the-fly transformation between different coordinate systems to evaluate which data to return.

Use: Currently, there are only a few datasets with geographical dimensions[22] in 4TU.ResearchData; thus the use of this service would be limited. But this may change in the future.

Feasibility: This service can simply be switched on. However, making the service user-friendly would require the addition of xml documents describing the data, and this would require significant additional effort. This may be justified if there are a large number of data files that have the same structure. In this case, the production of the xml documents could be automated. There would also likely be performance implications in switching on NCSS.

Service: Visualisation of geographical data

The most straightforward way to visualize geographical data is through the WMS (Web Map Service), an Open Geospatial Consortium (OGC) standard that has been implemented in THREDDS[23]. It can render data visually (colour-coded) on a map. When enabled, this service is available for datasets with geographical dimensions.

Use: Again, currently, there are only a few datasets with geographical dimensions, so the use of this service would be limited. But this may change in the future.

Feasibility: The WMS service can simply be switched on, with a bit of additional configuration. There may be performance implications. Simple visualization as provided via WMS is more suitable for value-added (i.e. processed) data products than it is for raw data. Researchers may not be satisfied with such a simple tool, as they generally want to do their own data processing. For more advanced data visualization and processing, applications on top of WMS, such as the ADAGUC[24] tool developed by KNMI, would be useful. Depending on the amount of suitable data entering the repository, this could be added in a later phase.

---

[22] Datasets may have 'implied' geographical parameters without having explicit geographical dimensions, i.e. geographical coordinates may be obtained by some calculation on the data. A typical example is radar data, such as IDRA data with time and distance dimensions. Knowing the start angle, start time, and sweep velocity, time translates into an azimuth angle which, together with distance, defines polar coordinates centered at the radar location. This can be translated into latitude/longitude coordinates.

[23] "TDS Web Map Service (WMS)." https://www.unidata.ucar.edu/software/thredds/current/tds/reference/WMS.html. Last accessed 24 July 2018.

[24] "ADAGUC - Welcome to the ADAGUC ...." http://adaguc.knmi.nl/. Last accessed 17 July 2018.

Service: Specialised visualisation of specific datasets (e.g. IDRA collection)
Specific types of data may require specific types of visualisation that are not covered by standard services like WMS. Example: radar data such as IDRA with polar coordinates (where the azimuth, in turn, is hidden in the time).
Use: May be a nice thing to have for the IDRA collection, and it could be useful to those outside the weather radar community who may be interested in the data but do not have the tools to visualise it. However, from the interviews, it was not clear this would be a priority.
Feasibility: Requires custom programming. Because of the effort involved, only feasible for large series of same-structured datasets. There will be performance implications.

Service: Customizable interface to analyse and visualise data
Maximum flexibility would be achieved with an online environment for users where they can program their own visualization or analysis. Such an environment can be created with Jupyter Notebooks/Labs. It can be made available for download for users to run on their own machines[25], or run on a server [26] entirely separate from the OPeNDAP server and accessible to logged-in users.
Use and feasibility: This is still under evaluation. The Climate Change Center Austria (CCCA), which is another netCDF data repository, currently provides prepared Jupyter Notebooks[27]. The CCCA's data providers are concerned, particularly for regional climate scenarios, about the open, individual, and dynamic creation of data visualizations (plots, means, etc.) without any involvement of the data creators and their knowledge about limitation or uncertainties. Nethertheless, it's a proper service for data analytics using the same infrastructure without any download and bandwidth consumption.

Service: Dynamic data citation
The Research Data Alliance (RDA) Working Group on Data Citation (WG-DC)[28] recommends that persistent identifiers are generated for every query that results in the creation of a subset dataset. This makes the exact queries citable (e.g. in scientific publications) and allows for the exact same subset datasets to be downloaded, (re-)published, and re-used by others. An implementation of the WG-DC recommendations at the CCCA proved to be successful[29]. For each re-published subset, the CCCA Dynamic Data Citation and Subsetting Tool instantly creates on the CCCA Data Server a landing page with its own unique persistent identifier. The re-published subset inherits metadata from the original dataset, as well as acquiring its own metadata, such as subset creator, bounding box, and time range. This is a first step to automatically keep data provenance information (e.g. relation, version, etc.).

---

[25] This can be done by providing a Docker image of the Jupyter environment.
[26] Jupiterhub and Jupyter Labs provide multi-user servers for Jupyter environments. A Jupyter Labs plugin adding support for THREDDS is available: https://github.com/eWaterCycle/jupyterlab_thredds. Last accessed 17 July 2018.
[27] These Jupyter Notebooks include UK Met Office libraries, such as iris, pandas, and xarray. The service uses OPeNDAP as well as an applied search and filter function.
[28] "Data Citation WG | RDA - Research Data Alliance." https://www.rd-alliance.org/groups/data-citation-wg.html. Last accessed 17 July 2018.
[29] "Implementing the RDA Data Citation Recommendations by the Climate Change Centre Austria (CCCA) for a repository of NetCDF files Webinar." https://www.rd-alliance.org/implementing%C2%A0-rda-data-citation-recommendations-climate-change-centre-austria-ccca-repository-netcdf. Last accessed 17 July 2017.

Use: There are differences between 4TU.ResearchData and CCCA that make it difficult to emulate CCCA's success story. For CCCA, the Austrian Climate Scenarios and their derivatives (e.g. Climate indices) are stored in netCDF files that are CF compliant, have lat/lon coordinates, and time as dimensions, and contain only one variable.[30] Subsetting is based on NCSS, which is more advanced and easier to use than standard OPeNDAP subsetting. Moreover at the CCCA Data Server, each dataset is exactly one netCDF file and all the metadata is in the netCDF itself. All of this makes subsetting easy, conceptually as well as practically. But these are all conditions that, except partially for a few datasets, currently don't exist at 4TU.ResearchData.
Feasibility: Might there be, at some point in time, a large influx of datasets that meet CCCA-like conditions, implementing dynamic citation at 4TU.ResearchData may become worthwhile. Before doing so, there should be an evaluation of the expected use by researchers. At the CCCA data server, the number of downloads of generated subsets has been moderate (around 260 in the first year); the complete dynamic data citation service, including re-publishing of datasets, hasn't been used so far, but the operational service is still young.

Conclusions

4TU.ResearchData was originally built to serve the needs of heterogeneous research communities in science and engineering (Rombouts and Princic, 2010). It is clear that even within the scope of netCDF data, which is mainly used in a limited number of geoscience disciplines, 4TU.ResearchData is still serving heterogeneous research communities, albeit in the Earth sciences, and in particular in environmental research.

Serving these heterogeneous communities is a challenge, but it also creates opportunities. The heterogeneity of the netCDF datasets stored in 4TU.ResearchData together with the diversity of views and attitudes to data archiving and data publishing by the data creators makes it difficult for 4TU.ResearchData to provide technical services. However, the common need for training, advice, and guidance creates the opportunity to develop a community of data depositors and users that could lead to higher-quality data and increased rates of data reuse. To achieve these goals we make the following recommendations.

Recommendations

- Organise a workshop on netCDF metadata standards and conventions, focussing on how 4TU.ResearchData could provide training, advice, and guidance in this area. This workshop should include researchers who use and deposit netCDF data in 4TU.ResearchData and other partners in the Netherlands and beyond (from the research community, industry, and similar service providers), who are also interested in this topic. The workshop would serve as a first step to build a national community and a way to continue to build links internationally.

[30] Nethertheless choosing more than one variable is possible too. Current tests are running on global Radio Occultation Data, i.e. netCDF data which contains the vertical coordinate in addition.

- To help accrue higher-quality data and promote data reuse and to support community building efforts, strengthen and improve communications about 4TU.ResearchData and its netCDF data and services, nationally and internationally.

- Host and keep updating and improving the existing netCDF Kickstarter open source tool, which provides templates for the production of netCDF files conforming to community-defined metadata standards and conventions. In line with the two previous recommendations, promote its use through training sessions and workshops.

- Because there is an important trend in big data, and in particular Earth observation data, for bringing the users to the data, further explore the feasibility of offering a customizable interface to analyse and visualise netCDF data. Such an environment could be created with Jupyter Notebooks/Labs. It might require cooperation from the TU Delft ICT department and some software development expertise.

Acknowledgements

First and foremost, we are extremely grateful to all the researchers who agreed to speak with us for their time and for their responses to our questions. This information was crucial to this work. We are also deeply grateful to a great number of people who provided information, comments, and feedback: Kees den Heijer for sharing his vast knowledge and expertise on netCDF; Jasmin Böhmer for her thoughts and ideas on netCDF and FAIR data; Marta Teperek for her encouragement and making us aware of the work of the RDA Working Group on Data Citation; Madeleine de Smaele for sharing her knowledge of the history of the archive and for putting us in touch with the Climate Change Center Austria (CCCA); Chris Schubert and his colleagues at CCCA for telling us and answering our questions about their implementation of the RDA Working Group on Data Citation; Julia Wagemann (European Centre for Medium-Range Weather Forecasts) for sharing her work on geospatial web services and big Earth data. We also had very fruitful exchanges and discussions with Rolf Hut (TU Delft), Niels Drost (Netherlands eScience Center), Riccardo Riva (TU Delft), Maarten Plieger and Wim Som de Cerff (both KNMI - Royal Netherlands Meteorological Institute), and Trygve Halsne (Norwegian Meteorological Institute). Marta Teperek, Kees den Heijer, Chris Schubert, Julia Wagemann, and Niels Drost read a first draft of the report and provided detailed and extremely valuable comments and feedback. We are most grateful to all of them for doing so. Last but not least, we would like to thank Alastair Dunning, the Head of 4TU.ResearchData, for giving us this exciting task to work on, for his many comments and feedback, and for his unwavering support throughout.

References

Cruz, M (2018a) “How does a data archive remain relevant in a rapidly evolving landscape: the case of the 4TU.Centre for Research Data.” Zenodo. DOI: http://doi.org/10.5281/zenodo.1175238

Cruz, M (2018b) “Adding Value and Facilitating Data Reuse: the Case of the 4TU.Centre for Research Data.” Zenodo. DOI: http://doi.org/10.5281/zenodo.1247903

Cruz, M J, Böhmer, J K, Gramsbergen, E, Teperek, M, de Smaele, M and Dunning, A (2018) “From Passive to Active, From Generic to Focused: How Can an Institutional Data Archive Remain Relevant in a Rapidly Evolving Landscape?” OSF Preprints. DOI: https://doi.org/10.31219/osf.io/jgrkb

Cruz, M and Gramsbergen, E (2018a) “NetCDF data at the 4TU.Centre for Research Data - a review of compliance with the FAIR principles.” Zenodo. http://doi.org/10.5281/zenodo.1316938

Cruz, M J and Gramsbergen, E (2018b) “Adding Value and Facilitating Data Reuse: the Case of the 4TU.Centre for Research Data.” Proceedings of the 2018 conference on adding value and preserving data (PV2018), Harwell, UK, 15-17 May 2018. Persistent URL: http://purl.org/net/epubs/work/37981055. Also available via OSF Preprints. DOI: https://doi.org/10.31219/osf.io/rvfs2

Delgado Blasco, J M, Sabatino, G, Cuccu, R, Rivolta, G and Marchetti, P G (2016) “Research and Service Support: Bringing Users to Data.” Living Planet Symposium, Proceedings of the conference held 9-13 May 2016 in Prague, Czech Republic. Edited by L. Ouwehand. ESA-SP Volume 740, ISBN: 978-92-9221-305-3, p.271. Available online: http://adsabs.harvard.edu/abs/2016ESASP.740E.271D

Gerkema, T and Duran Matute, M (2017) “Annual mean sea level in the Dutch Wadden Sea 2009-2011.” NIOZ Royal Netherlands Institute for Sea Research. Dataset. DOI: https://doi.org/10.4121/uuid:115ef6c5-8c58-4905-91f5-537985fb3b6f

Hellebrand, H (2004) “All data measured by discharge meter in Attert basin.” TU Delft. Dataset. DOI: https://doi.org/10.4121/uuid:0e38dcc8-d524-4abf-ab59-5c9a38075dc3

Hoonhout, B M, de Vries, S and Cohn, N (2016) “Field measurements on aeolian sediment transport at the Sand Motor mega nourishment during the MegaPeX field campaign.” TU Delft. Dataset. DOI: https://doi.org/10.4121/uuid:3bc3591b-9d9e-4600-8705-5b7eba6aa3ed

Leonelli, S (2017) “Towards the European Open Science Cloud: Five Lessons from the Study of Data Journeys.” Zenodo. DOI: http://doi.org/10.5281/zenodo.1043154

Mezghani, A, Dobler, A and Haugen, J H (2016) “CHASE-PL Climate Projections: 5-km Gridded Daily Precipitation & Temperature Dataset (CPLCP-GDPT5).” Norwegian Meteorological Institute. Dataset. DOI: https://doi.org/10.4121/uuid:e940ec1a-71a0-449e-bbe3-29217f2ba31d

Otto, T, Russchenberg, H W J, Reinoso Rondinel, R R, Unal, C M H and Yin, J (2010) “IDRA weather radar measurements - all data.” TU Delft. Dataset. DOI: https://doi.org/10.4121/uuid:5f3bcaa2-a456-4a66-a67b-1eec928cae6d

Otto, T and Russchenberg, H W J **(**2014) “High-resolution polarimetric X-band weather radar observations at the Cabauw Experimental Site for Atmospheric Research.” *Geosci. Data J.*, 1: 7-12. DOI: https://doi.org/10.1002/gdj3.5

Piwowar, H A and Vision, T J (2013) “Data reuse and the open data citation advantage.” *PeerJ* 1:e175. DOI: https://doi.org/10.7717/peerj.175

Rew, R K and Davis, G P (1997) “Unidata’s NetCDF Interface for Data Access: Status and Plans.” Proceedings of the Thirteenth International Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology, Anaheim, California, American Meteorology Society, September 1997. Available online: https://www.unidata.ucar.edu/software/netcdf/papers/netcdf-1996.html

Rijkswaterstaat; Provincie Zuid-Holland; EcoShape (2017). “Zandmotor data.” TU Delft. Dataset. DOI: https://doi.org/10.4121/collection:zandmotor

Rombouts, J and Princic, A (2010) “Building a 'data repository' for heterogeneous technical research communities through collaborations.” International Association of Scientific and Technological University Libraries, 31st Annual Conference. Paper 10. Available online: http://docs.lib.purdue.edu/iatul2010/conf/day2/10

Stive, M J F, de Schipper, M A, Luijendijk, A P, Aarninkhof, S G J, van Gelder-Maas, C, van Thiel de Vries, J S M, de Vries, S, Henriquez, M, Marx, S and Ranasinghe, R (2013) “A New Alternative to Saving Our Beaches from Sea-Level Rise: The Sand Engine.” *Journal of Coastal Research 29*, Issue 5: 1001-1008. DOI: https://doi.org/10.2112/JCOASTRES-D-13-00070.1

Voorhoeve, R J and van der Maas, A (2016) “System identification (SYSID) benchmark for an active vibration isolation system (AVIS).” Eindhoven University of Technology. Dataset. DOI: https://doi.org/10.4121/uuid:494e738d-e2aa-49e4-b076-ac96d3a142e8

Wagemann, J, Clements, O, Figuera, R M, Rossi, A P and Mantovani, S (2017) “Geospatial web services pave new ways for server-based on-demand access and processing of Big Earth Data.” International Journal of Digital Earth, 11:1, 7-25, DOI: https://doi.org/10.1080/17538947.2017.1351583

Wilkinson, M A *et al.* (2016) “The FAIR Guiding Principles for scientific data management and stewardship.” *Scientific Data*. 3, 160018. DOI: 10.1038/sdata.2016.18