TEAMx Data Management Plan

Authors[1]

Manuela Lehner, Helen Ward, Stefano Serafin, Marco Arpagaus, Lorenzo Giovannini, Martin Juckes, Peter Knippertz, Martin Kohler, Katharina Loewe, Stephen Mobbs, Corinna Rebmann, Greg Stossmeister, Stephanie Westerhuis

TEAMx (Multi-scale transport and exchange processes in the atmosphere over mountains – Programme and experiment) is an international research programme focused on improving our understanding of transport and exchange processes in the atmosphere over mountains. The goals of TEAMx are summarized in a *Memorandum of Understanding*[2], while the scientific objectives are described in the *TEAMx White Paper*[3] and an *Experimental Plan,* consisting of a *Field Observations Plan*[4] and a *Numerical Modeling Plan*[5]. The practical implementation of the *TEAMx Observational Campaign (TOC)* is described in the TEAMx *Implementation Plan*. This *Data Management Plan* summarizes all provisions regarding the storage and exchange of TEAMx data, including near real-time data exchange during the TOC, long-term storage and access to TEAMx data for the TEAMx community, as well as the wider scientific community. The associated *TEAMx Data Policy* in Appendix A of this document sets out the duties and rights of data providers and users in order to best accommodate the expectations of both, while Appendix B contains a data publication checklist for data providers.

1. Data sharing needs

TEAMx is a bottom-up financed international research initiative that brings together many different groups. Sharing and exchange of data between groups is thus central to the success of TEAMx and needs to happen for different purposes before, during, and after the TOC. The following non-exhaustive list summarizes some of the needs for data storage and exchange within TEAMx. A list of different data types to be produced within TEAMx is given in Appendix C.

- During the TOC, data need to be shared in near real-time to facilitate coordination of the campaign, to check for instrument issues, and to assess local weather conditions to make campaign decisions. Sharing of these data within the TEAMx community can be largely limited to images (e.g., quicklooks, satellite and radar images, and forecast products) and specialist field planning tools. However, automatic generation of quicklooks from data collected during the TOC needs to be ensured.
- During the TOC, observational data need to be transferred in (near) real-time for data assimilation purposes. This requires data files that follow strict data formatting guidelines and pre-processing steps, and therefore collaboration between field scientists and the modelling community.
- During and shortly after the TOC, raw observational data collected by TEAMx investigators need to be stored and backed up long-term. These raw data may or may not be published depending on the quantity of the data, usefulness to others, and the preference of the responsible investigator, but it needs to be ensured that the data are not lost. Exchange of these data may be necessary among individual TEAMx investigators.

[1] The members of the Task Team Data Management are responsible for the content of the present version.
[2] http://www.teamx-programme.org/files/TEAMx-MoU.pdf
[3] https://www.uibk.ac.at/iup/buch_pdfs/10.1520399106-003-1.pdf
[4] http://www.teamx-programme.org/files/TEAMxFieldObservationPlan_v6-02.pdf
[5] http://www.teamx-programme.org/files/NMP_v1.pdf

- During and for several years after the TOC, processed and quality-controlled observational data need to be accessible to the TEAMx community and the wider scientific community. These data may change as calibrations are improved, bugs are discovered, and new retrievals are applied. Version control and good documentation are thus required for traceability.
- Numerical model simulations are run throughout the duration of the TEAMx programme, including, for example, forecast runs for TOC planning, reanalyses of the TOC, model intercomparison studies, individual case studies, and high-resolution idealized simulations. Model output needs to be shared within the TEAMx community throughout the duration of the programme. Model input data to reproduce simulation results, in particular published ones, need to be stored long-term (e.g., model version information or model code, model namelists, and model initialization data).
- In agreement with requirements from funding agencies and many journals, data used in publications need to be publicly accessible, assigned a DOI, and should be archived long-term.

2. Near real-time data exchange during the TOC

Observations and forecasts for planning TOC activities

During the TOC, current weather information (e.g., satellite and radar images and data from station networks) and operational weather forecasts provided by National Weather Services specifically for the TOC must be made available to the field scientists for campaign decisions (e.g., planning of Intensive Observational Periods, IOPs). In addition, quicklooks need to be produced for the instruments deployed by the TEAMx investigators to provide information about the instrument status and to provide local atmospheric data at the sites of interest. These products will be made accessible to the TEAMx community through online services hosted and maintained by the UK Centre for Environmental Data Analysis (CEDA) for data quicklooks and by the University of Innsbruck, Department of Atmospheric and Cryospheric Sciences (ACINN) for forecast products, respectively:

A. ACINN is operating a weather portal, which provides visualizations of operational numerical weather prediction runs and current weather information. Access will be granted to the TEAMx community during the TOC.
B. The TEAMx community will be able to produce *quicklooks* for instruments deployed by TEAMx investigators on demand using Jupyter Notebooks on the JASMIN[6] server operated by CEDA. TEAMx investigators are asked to set up an automatic data transfer to the JASMIN server at least once per day and to provide existing quicklook scripts that can be converted to Jupyter Notebooks. Depending on the available resources and the data size, data may only be stored on the JASMIN server for a limited amount of time.

Data exchange among TEAMx investigators

TEAMx investigators are responsible for providing their own solutions for near real-time exchange of data during the TOC and other TEAMx measurement campaigns.

Data exchange for data assimilation purposes

Near real-time data assimilation (DA) during the TOC is envisaged by multiple institutions. Each of these institutions has stringent and differing file format requirements for their respective model code. The TEAMx investigators interested in DA are responsible for working together and with the field scientists to find a solution for data transfer and for formatting observational data according to the DA needs, for example, by providing scripts for data conversion to the field scientists.

[6] https://jasmin.ac.uk

3. Long-term data storage and access

Processed and quality-controlled observational and model data need to be stored long-term and made accessible to the TEAMx and the wider scientific communities. Version control is necessary to allow the publication of updated datasets as new calibrations or data retrievals become available or potential bugs are discovered. DOIs need to be assigned to published datasets according to funding-agency and scientific journal requirements.

Long-term storage of TEAMx data will be achieved through a Distributed Data Centre (DDC), consisting of a small set of existing data repositories, where data can be stored permanently. Access to all TEAMx data will be provided through a Central Data Portal (CDP).

TEAMx Central Data Portal (CDP) – Access to the TEAMx datasets will be provided through the existing Earth Data Portal[7] (EDP) hosted and maintained by the Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research. EDP *is developed by scientists for scientists in order to facilitate FAIR research data management to publish, visualize, browse and access research data from interdisciplinary research collaborations as well as individual research initiatives*[7].

The CDP will allow access to all data products collected as part of the TEAMx programme and published through the DDC. This includes observational data collected from instruments deployed as part of the TOC and other TEAMx measurement campaigns (e.g., pre-campaigns), observational data from long-term monitoring networks that are part of the TEAMx programme, data required to run numerical simulations, and potentially numerical model output.

TEAMx Distributed Data Centre (DDC) – A small group of existing data repositories will be used to store and publish TEAMx data products with some standardisation of metadata and data formats. To allow data access via the CDP, data repositories need to fulfil one of the following criteria:

- The repository is already integrated in EDP, that is, datasets published at the respective repository are already harvested by and accessible from the EDP.
- The repository is not yet integrated in EDP but provides an OAIPMH[8] (Open Archives Initiative Protocol for Metadata Harvesting) standard for metadata.
- Other, for example, institutional repositories can be made accessible if the EDP Data Management Plan is used to provide all the required information.

In addition, some predefined keywords need to be used to assign the respective dataset to TEAMx on the EDP. These will be communicated to the TEAMx investigators in a future, updated version of the *TEAMx Data Policy* (Appendix A) and the data publication checklist in Appendix B.

TEAMx investigators must ensure that their data products are published according to the *TEAMx Data Policy* (Appendix A). Some TEAMx participants may be required to use institutional repositories or may need to follow specific requirements provided by the respective funding agencies when selecting a suitable data repository. To increase the quality, consistency, and visibility of TEAMx datasets, the use of curated data repositories is, however, strongly encouraged. A number of suitable data repositories are listed in Appendix D. If no other requirements exist, the recommendation is to use PANGAEA[9], which is already integrated in EDP.

[7] https://earth-data.de
[8] https://www.openarchives.org/pmh/
[9] https://www.pangaea.de

Appendix A – TEAMx Data Policy

This *TEAMx Data Policy* refers to all long-term storage and access of TEAMx data via the *TEAMx Central Data Portal* and *Distributed Data Centre*. The document sets out the duties and rights of data providers and users in order to best accommodate the expectations of both. It has been developed in line with the World Meteorological Organization (WMO) guidelines[10] which state that the exchange of data products should be free, timely, and unrestricted, and follow the FAIR principles[11] (data sets should be findable, accessible, interoperable and reusable).

A1. Definitions

TEAMx Central Data Portal (CDP) – the long-term data portal that provides access to all TEAMx data.

TEAMx Distributed Data Centre (DDC) – the collection of data repositories, in which TEAMx datasets are stored that can be accessed via the CDP.

Data repository – any storage infrastructure that is used to host a TEAMx dataset, either public or institutional, such as a database or storage server, and that can be accessed via the CDP. A list of recommended data repositories is given in Appendix D.

Data product – any specific dataset or subset of data archived in the DDC (e.g., measurements collected with a specific instrument or simulations performed with a specific model).

Data provider – any individual, group of individuals, or institution providing a data product to the DDC and acknowledged as responsible for the collection, processing, quality control, and documentation of the data product.

Data user – any individual, group of individuals, or institution accessing TEAMx data from the DDC during or after the embargo period, either through the CDP or directly through the respective data repository in the DDC.

TEAMx project – a research initiative that has been endorsed by TEAMx and addresses one or more of the goals of TEAMx. There are two types of TEAMx projects: (1) TEAMx core projects (i.e., those focused on the TEAMx study region and highly aligned with TEAMx goals) and (2) TEAMx related projects (i.e., those addressing TEAMx objectives and conducting complementary observations or simulations outside the main study region or main study period). For the TEAMx core projects the data collected must be shared with the TEAMx community. For the TEAMx-related projects it is requested (but not required) that the data are made available to the TEAMx community.

TEAMx investigator – any active member of a TEAMx project.

TEAMx Observational Campaign (TOC) – The main TEAMx observational field phase between September 2024 and September 2025.

TEAMx measurement campaign – Any measurement campaign conducted as part of TEAMx, including the TOC as well as other campaigns, for example, pre-campaigns to test instruments, measurement sites, and measurement strategies in advance of the TOC.

A2. Embargo period

An embargo period exists for one year after the end of the TOC or other TEAMx measurement campaign, during which the data have been collected. During the embargo period, data access may be restricted to TEAMx investigators. Numerical modelling data have to be published at the time of

[10] https://community.wmo.int/resolution-40
[11] https://www.nature.com/articles/sdata201618

publishing scientific results based on the respective simulations in a scientific journal (Section A3), independent of the time that has elapsed since running the model simulations. Publication of numerical modelling data after presenting results based on the respective simulations at a conference is desirable, but not mandatory. Numerical modelling data related to simulations that do not lead to scientific publications do not have to be published. The data providers should nevertheless follow the guidelines in this section with respect to sharing their data before publication.

Some data repositories, like PANGAEA, offer embargo periods, which allows the data providers to prepare the dataset for publication before the end of the embargo period. The following provisions apply during this period:

a) It is left to the discretion of the data provider to publish their data product in the DDC and thus make it publicly accessible before the end of the embargo period. The data product must, however, be published in the DDC at the end of the embargo period, following the guidelines in A3.
b) All TEAMx investigators have equal access to all TEAMx data. This means in particular that data providers are required to share their data with other TEAMx investigators even during the embargo period upon request. However, in order to safeguard the efforts of data providers and their research group (in particular PhD students), data providers can reserve the right to refuse access to a data product during the embargo period if the proposed work is too similar to their own TEAMx project goals. In case the potential data user and the data provider are unable to come to an agreement, the Coordination and Implementation Group (CIG) will decide.
c) No public release of a TEAMx data product (sharing with colleagues, conference presentations, publications, commercial and media use, etc.) is allowed without the permission of the data provider.
d) Commercial use of TEAMx data products is prohibited, unless authorized by the data provider.
e) The data providers must provide an example acknowledgement that can be used by the data users to cite the respective dataset. The use of the acknowledgement in A4 is recommended.

A3. Data publication

Data providers are required to publish a first version of their quality-controlled and/or processed data product no later than the end of the embargo period (see A2), ensuring that the selected data repository is compatible with the CDP and that all guidelines regarding keywords, etc. are followed to ensure that the dataset is included as a TEAMx dataset in the CDP. The processing and applied quality control must be specified in detail in the accompanying metadata. Data providers are encouraged to choose a data repository that assigns a DOI to the dataset to facilitate referencing, even if not explicitly required by their funding agency. Data providers are further responsible for their own storage of raw, unprocessed data.

For numerical modelling studies, all data required to run simulations must be published at the time of publishing scientific results based on the respective simulations in a scientific journal. This includes model namelists, information regarding model version, model code in the case of code modifications, and input data or links to input data if they are publicly available (e.g., ERA5). To keep data storage requirements manageable, model output does not need to be published, but is encouraged for simulations that are likely of further use to other TEAMx investigators.

All datasets should be published in netCDF[12] format, unless there are justified exceptions (e.g., model code or namelists). While not strictly enforced, data providers are encouraged to use the Climate and Forecast[13] (CF) netCDF standard, the standard recommended by Unidata[14], who developed and maintains the netCDF libraries. Adhering to a common standard will increase the usability of datasets across the TEAMx community, thus increasing the visibility of datasets through their use in multiple publications. The CF standard contains a list of recommended attributes to describe the file content, including, for example, variable units, missing value information, and information about the data provider and instrument. Variable names should follow the list of CF standard names[15] and the guidelines for construction of CF standard names[16] for variables not included in the existing list whenever possible.

All data must be provided under a Creative Commons (CC) licence[17]. The data provider can select the licence from the available CC Licence options.

The data providers must provide an example acknowledgement that can be used by the data users to cite the respective dataset. The use of the acknowledgement in Section A4 is recommended.

A4. Data use

Any use of TEAMx data products, during or after the embargo period, must include an acknowledgement (i.e., citation) of the source and the data provider. The following acknowledgement is recommended:

> The *[description of the data product]* was collected/produced as part of the TEAMx programme and provided by *[name of the data provider, institution of data provider]*. *[name of the data provider]*'s contribution to TEAMx was funded by *[name of the funding agency, information about grant]*. The data are archived at *[name of data repository]* and are accessible at [*URL/DOI of data product*].

Data users must inform the respective data providers if a data product is to be shared with other parties via journal articles, presentations, and research proposals. If the data product constitutes a substantial part of the work, the data provider should be offered co-authorship and the opportunity to collaborate (both during and after the embargo period). In the case of co-authorship, an additional acknowledgement of the data provider is not required.

A5. Data protection

Depending on the data repository selected by the data provider for publishing their data product in agreement with the requirements in Section A3, data providers and data users may be required to provide personal information (e.g., name, institution, and email address) to publish and access data products. It is the responsibility of the data provider to collect information on the respective data repository's data protection standards and consider this information in their choice of data repository if desired.

[12] https://www.unidata.ucar.edu/software/netcdf/
[13] http://cfconventions.org/Data/cf-conventions/cf-conventions-1.7/cf-conventions.html
[14] https://www.unidata.ucar.edu/software/netcdf/conventions.html
[15] https://cfconventions.org/Data/cf-standard-names/current/build/cf-standard-name-table.html
[16] Cfconventions.org/Data/cf-standard-names/docs/guidelines.html
[17] https://creativecommons.org/share-your-work/cclicenses/

Appendix B – Checklist for data providers

The checklists provided here for observational and model data are intended as an aid for data providers. Please refer to the full text of the above Data Management Plan and Data Policy (Appendix A) for details on the individual points.

B1. Observational data

Before and during the measurement campaign

- ✓ Ensure that the raw data are stored safely.
- ✓ Update the instrument status every day in the online document that will be provided for this purpose.

Optional, but recommended:

- ✓ Set up an automatic data transfer to the JASMIN data server with a frequency of at least once per day (see Section 2).
- ✓ Provide a visualization script to produce quicklooks from your data on the JASMIN data server (see Section 2).

During and after the measurement campaign until the end of the one-year embargo period

- ✓ Share the data with TEAMx colleagues if requested, following the guidelines in Section A2.
- ✓ Publish your dataset no later than the end of the embargo period (see below).
- ✓ Produce netCDF files (preferably CF standard) of your quality-controlled and post-processed datasets.

At the end of the embargo period (see Section A3)

- ✓ Select an appropriate data repository that is compatible with the CDP and that fits your needs. See Appendix D for a list of recommended repositories.
- ✓ Select a CC licence that fits your needs.
- ✓ Provide measurement details, quality-control and post-processing steps, and an example acknowledgement text in the metadata.
- ✓ Use the following keywords when publishing the datasets: *[WILL FOLLOW]*.

B2. Model data

At the time of publishing scientific results based on the respective simulations (see Section A3)

- ✓ Prepare and publish all data required to reproduce the simulations for publication, including model namelists, information regarding model version, model code in the case of code modifications, and input data or links to input data if they are publicly available. If netCDF is a viable option, use netCDF (preferably CF standard).
- ✓ Select an appropriate data repository that is compatible with the CDP and that fits your needs. See Appendix D for a list of recommended repositories.
- ✓ Select a CC licence that fits your needs.
- ✓ Use the keyword TEAMx when publishing the datasets.

Optional:

- ✓ Publish the model output in netCDF format (preferably CF standard) in addition to the data required to reproduce the simulation.

Appendix C - Types of TEAMx data

TEAMx involves a wide variety of observational and modelling tools, spanning different interests in atmospheric science across a range of spatial and temporal scales. The following non-exhaustive list summarises different types of data involved:

- Observational data collected during the TOC. These data come from many different projects, including major infrastructure deployments and many smaller deployments. The total data amount is expected to be on the order of 100 TB.
- Observational data from operational medium- to long-term existing sites which are a key part of TEAMx (e.g., existing eddy covariance and remote sensing stations, weather and air quality networks).
- Observational data from medium- to long-term existing sites which would be very relevant for TEAMx but have restricted access (e.g., radar data, some emissions data). It is currently unclear how or if these data could be shared. Access to the data is usually possible for research purposes but permissions may need to be granted on a case-by-case basis.
- Data retrieved from satellites during the TOC (e.g., cloud cover, land-surface temperature, soil moisture).
- Various auxiliary datasets such as topographic maps, land cover, soil type, vegetation characteristics.
- Data associated with modelling activities of various projects. It will not be possible to store the vast amount of data associated with all TEAMx modelling activities, and in most cases storing a large volume of model output is not useful or necessary. Only data needed to run the model (e.g. model code, namelists, input data if not accessible elsewhere) will be stored long-term.
- Data associated with central TEAMx modelling activities which will be of value to many groups (e.g. reanalysis, forecasts during the campaign, collaborative model comparisons). There are clear benefits to being able to share this type of data, even though the data volume could be very high. An estimate of data volume for collaborative model comparisons where the intention is to share only a subset of model output is on the order of 10 TB.
- Scripts used to process and visualize observational and model data (e.g., scripts to prepare observational data for data assimilation purposes, quicklook scripts)

Appendix D – Recommended data repositories

Data repository	Weblink	Data curation	Embargo period	Data volume limits	Notes
PANGAEA	www.pangaea.de	yes (may take up to 8 weeks)	yes	10 GB/submission 1 submission/day 10 submissions/month	
zenodo	zenodo.org	no	yes	50 GB/record (higher quotas can be requested)	
CEDA	archive.ceda.ac.uk	yes (level depending on anticipated data re-use)			NERC-funded projects
GFZ Data Services	bib.telegrafenberg.de/dataservices	yes	yes		
World Data Centre for Climate	www.wdc-climate.de/ui	yes		8 GB/file 1000 files/dataset	Approval required prior to data submission
RADAR	https://radar.products.fiz-karlsruhe.de/en				German institutions Service agreement between institution and FIZ Karlsruhe required annual fee

The above list is not complete and contains only repositories that provide services to more than just a single institution. A full list of OAIPMH-compliant repositories that can be harvested by the CDP can be found under https://www.openarchives.org/pmh/. If another data repository has to be used that is not listed, e.g., an institutional repository, the EDP Data Management Plan Tool *[LINK WILL FOLLOW]* can be filled out to provide the required information to make the datasets findable.