GAIA-CLIM Report / Deliverable D7.1

Gap Analysis for Integrated Atmospheric ECV CLImate Monitoring:
Data Management Plan

A Horizon 2020 project; Grant agreement: 640276
Date: 31 July 2015
Lead Beneficiary: NERSC
Nature: R
Dissemination level: PU

Work-package	WP 7 (Management)
Deliverable	D7.1
Title	Data Management Plan
Nature	R
Dissemination	PU
Lead Beneficiary	Nansen Environmental and remote Sensing Center, Norway
Date	31st July 2015
Status	First version
Authors	Anna Mikalsen, NERSC
Editors	Peter Thorne (NUIM), Nico Cimini (CNR), Karin Kreher (BK Scientific), Justus Notholt and Matthias Buschmann (Uni Bremen), Jörg Schulz and Marie Doutriaux-Boucher (EUMETSAT), Fabio Madonna and Gelsomina Pappalardo (CNR)
Reviewers	
Contacts	anna.mikalsen@nersc.no
URL	http://www.gaia-clim.eu/

This document has been produced in the context of the GAIA-CLIM project. The research leading to these results has received funding from the European Union's Horizon 2020 Programme under grant agreement n° 640276. All information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission has no liability in respect of this document, which is merely representing the authors' view

Table of Contents

D.7.1: Data management plan commensurate with Pilot on Open Research Data, initial version, July 2015

Project Name: Gap Analysis for Integrated Atmospheric ECV Climate Monitoring (GAIA-CLIM)

Funder: European Commission (Horizon 2020)

Grant Title: No 640276

Project description

The Gap Analysis for Integrated Atmospheric ECV Climate Monitoring (GAIA-CLIM) Project will establish sound methods for the characterisation of satellite-based Earth Observation (EO) data by surface-based and sub-orbital measurement platforms - spanning Atmosphere, Ocean and Land observations. GAIA-CLIM shall add value by:

- Improving traceability and uncertainty quantification on sub-orbital measurements;
- Quantifying co-location uncertainties between sub-orbital and satellite data;
- Use of traceable measurements in data assimilation; and
- Provision of co-location match-up data, metadata, and uncertainty estimates via a 'virtual observatory' facility.

The project is not envisaged to directly collect primary data, i.e. make measurements for the sole purpose of the project. Rather it will provide added value to existing and forthcoming measurements, taken by both consortium members under separate funding support and by third party institutions participating in various national and international measurement programs.

GAIA-CLIM shall primarily use metrologically reference quality measurements that are traceable and have well quantified uncertainty estimates. At the global scale, currently envisaged potential contributing networks include the Global Climate Observing System (GCOS) Reference Upper-Air Network, the Network for Detection of Atmospheric Composition Change (NDACC) and the Total Column Carbon Observing Network (TCCON). At the European level, these include networks such as MWRNET and ACTRIS. A full listing of contributing observations will become apparent upon completion of task 1.2, envisaged in year 2 of the project. Importantly, GAIA-CLIM will only make use of those primary observations to which no academic restrictions to use, re-use, and re-distribute any longer apply. The providers of primary data from these networks shall implicitly or explicitly agree to release their data according to this data management plan and the 'virtual observatory' data policy. At the time of writing, the 'virtual observatory' and respective data policy do not exist, yet. However, this data policy will be in compliance with the H2020 Pilot on Open Research Data (s. next section). The usage of satellite data has to follow the data policies prescribed by the satellite operators, although GAIA-CLIM will only use those data where the rights for re-use and re-distribution in the 'virtual observatory' can be attained. In reality this constitutes the vast majority of satellite data. Furthermore, re-analysis and Numerical Weather Prediction (NWP) data may also become part of the forthcoming 'virtual observatory'. Such data will generally arise from within the consortium (ECMWF and MO partners under WP4) and no restrictions are envisaged.

Project parts dealing with enhancing existing primary data streams are:

- Preparation and assessment of reference-quality sub-orbital data (including in global assimilation systems) and characterisation of key satellite datasets

  a. Assessment of several new satellite missions, using data assimilation of reference-quality sub-orbital measurements, targeting temperature and humidity (under work package 4).

  b. Development of infrastructure to deliver data dissemination for reference data co-locations with satellite measurements (under work packages 3 and 5).

  c. Development of a software infrastructure for preparation, monitoring, analysis and evaluation of reference data (under work packages 2 and 5).

  d. Development of a general methodology for using reference-quality sub-orbital data for the characterisation of EO data (under work packages 4 and 5).

- Creation and population of a 'virtual observatory'

  a. Creation of a collocation database between EO measures and reference-quality measurements.

  b. Preparation of data to enable comparisons, including relevant uncertainty information and metadata for users to understand and make appropriate use of the data for various applications.

  c. Creation of data interrogation and visualization tools, building upon existing European and global infrastructure capabilities.

  d. Planning for the potential transition of the resulting 'virtual observatory' from research to operational status in support of the Copernicus Climate Change Service and Copernicus Atmospheric Service.

Pilot on Open Research Data

GAIA-CLIM participates in the H2020 Pilot on Open Research Data. Knowledge generated during the project will be shared openly. Any milestones, deliverables or technical documents produced will, following appropriate internal-to-project review procedures involving at least an expert and a management-based review, be published online and made discoverable. Peer-reviewed publications will by policy be to journals that are either open access or allow the authors to pay for the articles to be made open access (for such instances, the additional charges will be paid).

Dissemination and Exploitation of Results

A core facet of GAIA-CLIM is the 'virtual observatory' of visualization, subsetting, and analysis tools, which will constitute the primary means by which end-users will be able to access, visualize and utilize the outputs of the project. The 'virtual observatory' will be build upon and extend a number of existing facilities operated by project partners, which already undertake subsets of the desired functionality such as the Network of Remote Sensing Ground-Based Observations in support of the

Copernicus Atmospheric Service (NORS), the Cloud-Aerosol-Water-Radiation Interactions (ICARE) Project and the US National Oceanic and Atmospheric Administration (NOAA) Products Validation System (NPROVS). The resulting 'virtual observatory' facility will be entirely open and available to use for any application area. Significant efforts will be made to build an interface that is easy to use and which makes data discovery, visualization and analysis effortless. The 'virtual observatory' work package includes a specific task dedicated to documenting the steps required to transition this facility from a research to an operations framework with a view to constituting a long-term infrastructure.

Primary source datasets envisaged to be used within GAIA-CLIM

For the initial version of this data management plan, a number of datasets that are envisaged to contribute primary data streams to be used in GAIA-CLIM are documented here. Upon completion of Task 1.2 in year 2 some further datasets will likely be added. Where networks have data policies that place restrictions on near-real-time use, GAIA-CLIM shall only use the open delayed-mode data. Note that GAIA-CLIM will respect the data policy of the data originators and that the documentation herein should not be taken to imply advocacy for changing existing policies. Rather, it is important to understand and document the policies and practices that pertain to the source data.

1. GRUAN

Data set reference and name

GCOS Reference Upper Air Network (GRUAN)

Data set description

A group of stations coordinated by the GRUAN Lead Centre, hosted by the German Meteorological Service, DWD. Data products that meet necessary conditions of traceability and uncertainty quantification, documentation and publication are served via the US National Oceanic and Atmospheric Administration's National Centers for Environmental Information (NOAA NCEI) in Asheville, North Carolina, USA.

Standards and metadata

Data and comprehensive metadata must be undertaken according to stated requirements (documented through a technical document), shared with a central processing facility, and traceable to either SI or community accepted standards. The processing is open and transparent.

Data sharing

Data are shared without restriction or delay via NOAA NCEI.

Archiving and preservation (including storage and backup)

The archive is on a secure backed-up service and a copy is retained at the GRUAN Lead Centre. Entire data streams are periodically reprocessed when new insights on instruments accrue. Such reprocessing always incurs a change in version number and associated documentation.

2. NDACC

Data set reference and name

Network for the Detection of Atmospheric Composition Change (NDACC)

Data set description

The NDACC is composed of more than 70 high-quality, remote-sensing research stations[1] for observing and understanding the physical and chemical state of the stratosphere and upper troposphere and for assessing the impact of stratospheric changes on the underlying troposphere and on global climate. While the NDACC remains committed to monitoring changes in the stratosphere with an emphasis on the long-term evolution of the ozone layer, its priorities have broadened considerably to encompass issues such as the detection of trends in overall atmospheric composition and understanding their impacts on the stratosphere and troposphere, and establishing links between climate change and atmospheric composition. A wide variety of trace gases is measured[2].

Standards and metadata

NDACC is organized in several working groups, which are predominantly based on the applied measurement techniques: i.e. Brewer & Dobson, FTIR, Lidar, Microwave, Satellite, Sondes, Spectral UV, Theory, UV/Vis and Water Vapor. To ensure quality and consistency of NDACC operations and products, a number of protocols have been formulated covering topics such as measurement and analysis procedures, data submission, instrument inter-comparisons, theory and analysis, validation, and Cooperating Networks[3]. Regular working group meetings and instrument inter-comparisons are held to safeguard a continued high standard of the network's products.

Data sharing

All NDACC data over two years old is publicly available[4]. However, many NDACC investigators have agreed to make their data publicly available immediately upon archiving. The public record is available through anonymous ftp[5]. The use of NDACC data prior to its being made publicly available (i.e., for field campaigns, satellite validation, etc.) is possible via collaborative arrangement with the

[1] A map of the NDACC stations can be found on: http://www.ndsc.ncep.noaa.gov/; and a list of all the stations, including a description of the instrumentation and data products, is available on: http://www.ndsc.ncep.noaa.gov/sites/.

[2] The observational capabilities can be displayed by a chart on: http://www.ndsc.ncep.noaa.gov/

[3] The NDACC protocols are accessible under http://www.ndsc.ncep.noaa.gov/organize/protocols/

[4] A directory summarizing the operational status of the NDACC is available at: http://www.ndsc.ncep.noaa.gov/data/madir/. Long-term NDACC measurement activities are listed in Sections Ia and Ib, while Sections IIa and IIb include NDACC measurement activities conducted intermittently or during limited duration campaigns.

[5] Regular NDACC data: ftp.cpc.ncep.noaa.gov/ndacc/station

appropriate PI(s). Rapid delivery data, which will likely be revised before entry in the full database, is also available for some instruments[6].

In all cases when NDACC data is used in a publication, the authors agree to acknowledge both the NDACC data center and the data provider. Whenever substantial use is made of NDACC data in a publication an offer of co-authorship will be made through personal contact with the data providers and/or owners. Users of NDACC data are also expected to consult the on-line documentation and reference articles to fully understand the scope and limitations of the instruments and resulting data, and are encouraged to contact the appropriate NDACC PI (listed in the data documentation on the web page) to ensure the proper use of specific data sets. Those using NDACC data in a talk or paper are asked to acknowledge its use, and to inform the 'Theory and Analysis Working Group' PIs of any relevant publications.

Archiving and preservation (including storage and backup)

All data are released to the public and available on the anonymous ftp site no more than two years after measurement date. Data and comprehensive metadata is accessible via the NDACC data table[7] and clicking on the station name will take the user to the associated public data site.

3. TCCON

Data set reference and name

Total Carbon Column Observing Network (TCCON)

Data set description

TCCON is a network of ground-based Fourier Transform Spectrometers that takes direct solar absorption spectra at about 20 sites around the globe. From these, column averaged mole fractions of trace gases (CO_2, CH_4, N_2O, HF, CO, H_2O, and HDO) are inferred with a retrieval software. The HF and HDO retrievals are uncalibrated and hence preliminary. Each site contributes their dataset as an extending series for the current version of the retrieval software. Data are updated monthly and are publicly available no later than one year after the measurement; however, many sites choose to release their data much sooner.

Standards and metadata

TCCON products are calibrated against in-situ WMO values[8]. In this way, the long-term stability is checked continuously. All data are delivered with an extensive metadata overhead.

Data sharing

[6] NDACC rapid delivery data: ftp://ftp.cpc.ncep.noaa.gov/ndacc/RD

[7] http://www.ndsc.ncep.noaa.gov/data/data_tbl/

[8] Atmospheric measurements are only useful if calibrated to a common reference scale. For this reason the atmospheric in-situ measurements are anchored to the WMO Mole Fraction Scale. For the remote sensing measurements, this is not possible. Therefore the aircraft profiles, based on in-situ measurements are applied. Calibrations have been performed at several sites by high-flying aircrafts, covering the altitude region from about 200 m to 12 km. Internal test measurements by calibration gas cells ensure to detect any change in the alignment of the instruments.

Data is openly accessible and hosted at the Carbon Dioxide Information Analysis Center (CDIAC)[9] at Oak Ridge National Laboratory, USA. The data is made freely available to everyone. Acknowledgement and/or co-authorship in case of heavy use cases is expected. The data are stored in NetCDF format and each file has a DOI assigned to it (one per site and retrieval version). It is envisaged that each dataset will be described in a data publication paper.

Archiving and preservation (including storage and backup)

Archiving and preservation are ensured by the World Data Center (WDC) for Atmospheric Trace Gases'[10] standard implemented by CDIAC. In the near future, the data will be mirrored at the PANGAEA[11] data center, hosted by the Alfred-Wegener-Institute in Bremerhaven/Germany.

4. ACTRIS

Data set reference and name

ACTRIS (Aerosols, Clouds, and Trace gases Research InfraStructure Network)[12]

Data set description

ACTRIS is a European Project aiming at integrating European ground-based stations, equipped with advanced atmospheric probing instrumentation for aerosols, clouds, and short-lived gas-phase species. ACTRIS will have the essential role to support building of new knowledge as well as policy issues on climate change, air quality, and long-range transport of pollutants. The networks provide consistent datasets of observations, which are made using state-of-the-art measurement technology and data processing. Many of the stations from the different networks are co-located with or close to remote-sensing and in-situ instrumentation. The data is available through the ACTRIS data portal[13].

Standards and metadata

At the time of writing, there is no unified standard for all measurements and no metadata made available, yet.

Data sharing

The ACTRIS Data Centre web portal allows to search and analyse atmospheric composition data from a multitude of data archives through a single user interface. For some of the databases, the interface furthermore allows to download data.

ACTRIS data is freely available for non-commercial use. Use of this data implies an agreement to reciprocate.[14]

---

[9] http://tccon.ornl.gov/
[10] http://cdiac.ornl.gov/
[11] http://www.pangaea.de/
[12] http://www.actris.eu
[13] Go to http://actris.nilu.no to access the latest map (updated automatically) with additional information about the ACTRIS sites and variables or download data from the ACTRIS Data Centre.
[14] The ACTRIS data policy can be found at: http://actris.nilu.no/Data/Policy/?referrer=about

Archiving and preservation (including storage and backup)

The ACTRIS database is maintained by the Norsk Institutt for Luftforskning (NILU). The ACTRIS-2 project runs until 2019. Attempts are made to achieve long-term preservations by making the network an European Research Infrastructure.

5. MWRNET

Data set reference and name

An International Network of Microwave Radiometers (MWRnet)

Data set description

MWRnet links together a group of stations operated by independent institutions and running Microwave Radiometers (MWR) operationally. MWRnet activities are coordinated by the MWRnet chairs. Data products from the independent member institutions are collected and harmonized occasionally to foster the participation to international experiments and projects.

Standards and metadata

Data products from MWRnet members are collected and harmonized for providing uniform datasets to large-scale international experiments and projects. The resulting data and metadata have been tailored case by case according to the needs.

For the MWR data assimilation experiment performed within the HYdrological cycle in Mediterranean EXperiment (HyMeX)[15] preparation phase, the OBSOUL ascii format was used to comply with the Météo France ARPEGE/ALADIN/AROME system.

For the contribution to the HyMeX Special Observing Period 1 (SOP1), data and associated metadata were provided in NetCDF format[16].

For the contribution to the TOPROF[17] Observation minus Background (O-B) experiment, it has been adopted the observation data product standard defined for the High-Definition Clouds and Precipitation for advancing Climate Prediction (HD(CP)2) project, which follows to the possible extent the principles given in the NetCDF Climate and Forecast Metadata Conventions 1.6[18].

Data sharing

The policy for data sharing is agreed with the MWRnet members case by case. For the HyMeX preparation phase and SOP1 field experiment, the MWR data have been released according to the HyMeX Data and Publication Policy[19]. For GAIA-CLIM, the MWRnet members shall agree to release their MWR data according to this data management plan and the Virtual Observatory data policy.

Archiving and preservation (including storage and backup)

[15] HyMeX: http://www.hymex.org

[16] NetCDF format: http://www.unidata.ucar.edu/software/netcdf/

[17] TOPROF: http://www.toprof.eu

[18] NetCDF Climate and Forecast Metadata Conventions: http://cfconventions.org/

[19] HyMeX Data and Publication Policy: http://floodscale.irstea.fr/illustrations/hymex-datapolicy.pdf

The policy for data archiving and preservation is decided by the MWRnet chairs case by case. For the HyMeX SOP1 field experiment, the MWRnet data has been gathered on the HyMeX common backed-up database for secured, facilitated, and enhanced availability. The entire data streams are periodically reprocessed when new insights on instruments accrue. Such reprocessing always incurs a change in version number and associated documentation.

For GAIA-CLIM, the MWRnet data archiving and preservation policy is still to be decided.

Scientific research data should be easily:

1. Discoverable

Are the data and associated software produced and/or used in the project discoverable (and readily located), identifiable by means of a standard identification mechanism (e.g. Digital Object Identifier, DOI)?

Data and metadata will mainly be made available through the 'virtual observatory' facility. This online tool will make the data discoverable and also provide mapping, comparison and visualization functions. Data versioning, source locations, and any DOIs from the primary data sources will be retained. The possibility of creating data and software DOIs for the 'virtual observatory' shall be investigated, but it is not yet decided. For instance, DOI-registration works well for static data sets but remains mostly unexplored for regularly updated (changed) data. Thus, a decision for or against usage of DOIs depends very much on the final operation mode of the 'virtual observatory', which needs to be developed during the project. The 'virtual observatory' facility will be hosted by EUMETSAT and made discoverable.

2. Accessible

Are the data and associated software produced and/or used in the project accessible and in what modalities, scope, licenses?

As GAIA-CLIM participates in the Pilot on Open Research Data, knowledge generated during the project is shared openly. Any milestones, deliverables or technical documents produced are, following appropriate internal-to-project review procedures, published online and made discoverable. Commensurate with the Pilot on Open Research Data, all work explicitly produced by GAIA-CLIM will be open. However, GAIA-CLIM work in many cases will build upon pre-existing capabilities of the partners. In a restricted subset of these cases, Intellectual Property Right (IPR) restrictions relate to these background materials. Such background material IPR is covered within the consortium agreement (cf. Annex 1). The policing of this aspect is the responsibility of the Technical Coordination Group.

The 'virtual observatory' facility will be entirely open and available to use for any application area. However, following the results of the user survey, the 'virtual observatory' will contain online applications. The underlying software will be openly shared to the extent useful, but GAIA-CLIM will not provide software usage support for users. This is beyond the scope and resources of the project.

Peer-reviewed publications will by policy be to journals that are either open access or allow the authors to pay for the articles to be made open access.

3. Assessable and intelligible

Are the data and associated software produced and/or used in the project assessable for and intelligible to third parties in contexts such as scientific scrutiny and peer review?

Research is undertaken within GAIA-CLIM to improve observational traceability for a number of broadly used methods of observation and the quantification of the co-location mismatch uncertainties. The software resulting from GAIA-CLIM that shall constitute input to the 'virtual observatory' shall be shared openly and without restriction and shall be well documented.

The novel approach of GAIA-CLIM is to demonstrate comprehensive, traceable, EO Cal/Val for a number of metrologically mature ECVs, in the domains of atmospheric state and composition, that will guarantee that products are assessable and intelligible to third-party users.

4. Usable beyond the original purpose for which it was collected

Are the data and associated software produced and/or used in the project useable by third parties even long time after the collection of the data?

Data served will be available for any use regardless of whether it is within the currently envisaged end-uses or otherwise. Significant efforts will be made to build an interface that is easy to use and which makes data discovery, visualization and analysis effortless. . All software that underlies the 'virtual observatory' and is created using GAIA-CLIM resources shall be made available. The 'virtual observatory' work package includes a specific task dedicated to documenting the steps required to transition this facility from a research to an operations framework in support of Copernicus services.

Once the project will be completed, the 'virtual observatory' and its underlying software will remain available, but in a "frozen state" with the aim of becoming further developed and integrated into the emerging Copernicus Climate Change Service and Copernicus Atmospheric Service. If continued in this way, Copernicus data and software distribution policies will be applied in the long-term.

5. Interoperable to specific quality standards

Are the data and associated software produced and/or used in the project interoperable allowing data exchange between researchers, institutions, organisations, countries, etc.?

The project will only deal with both EO and sub-orbital (including in-situ and ground-based remote-sensing) data, which are available for academic use without restriction to simplify issues over dissemination of added value products derived by the project. These added value products will be made available immediately after they are produced and quality controlled without restriction. Data are accompanied by conversion tool that enable likely two different output formats that are in broad use within the recognised primary stakeholder communities, e.g. CF-compliant NetCDF. The data will be made available along with reading routines and visualisation tools through the 'virtual observatory' facility, which will allow data discovery and data usage for calibration and validation of level 1 and level 2/3 EO observations. The expectation is that new software written will use open-source software to the extent possible and useful and use of existing software shall have a preference for using programming languages that are open source or have open source compilers available such as e.g., C++, Fortran or python.

Annex 1: Specific limitations and/or conditions of background material covered within the GAIA-CLIM consortium agreement

In concordance with the GAIA-CLIM consortium agreement, the following background is hereby identified and agreed upon for the Project. Specific limitations and/or conditions, shall be as mentioned hereunder:

Describe Background	**Specific limitations and/or conditions for implementation (Article 25.2 Grant Agreement)**	**Specific limitations and/or conditions for exploitation (Article 25.3 Grant Agreement)**
Satellite Data for the purposes of WP4. (a) Within scope of WMO Resolution 40 (b) Outside scope of WMO Resolution 40	The Data described is third-party background and for the purposes of; (a) it can be used on a royalty-free basis and in compliance with any terms associated with WMO Resolution 40. (b) can be used subject to the terms and conditions of the original owners	
Software and Coding for the purposes of WP4 (which includes associated algorithms)	These will be fully documented and included as part of the Deliverables of the Project and as such will not be subject to any specific restrictions on use, beyond those agreed in the Consortium Agreement.	
Data to be used for Mapping Geographical Capabilities in WP1	The Data described contains third-party Background, which is split into two categories; (a) AMDAR data, the use of such information being subject to the terms and conditions imposed by AMDAR. (b) Information of Airline Communication Equipment and Flight Routes will also form part of the Data, however all this Information is publically available.	
Data, Modelling and Software used for Met Office contribution to WP2	Mode-S data, observations and comparisons with the UKV model will be Met Office owned Background, to which no special conditions beyond those agreed in the Consortium Agreement shall apply.	