



D.4.7.2 Report on high level science products based on polar data



Deliverable number:	D4.7.2
Work package:	WP4 – Atmospheric Domain
Intermediate Objective:	IO4.5
Deliverable type:	<input checked="" type="checkbox"/> Document, report
	<input type="checkbox"/> Websites, patent filings, videos, etc.
	<input type="checkbox"/> Other: please specify
Dissemination level:	<input checked="" type="checkbox"/> Public
	<input type="checkbox"/> Restricted
Estimated delivery (bimester):	B10
Actual delivery date:	25/06/2024
Author(s) (Partner-OU):	Alice Cavaliere, Vito Vitale, Mauro Mazzola, Giulio Verazzo (CNR-ISP-BO)
Reviewed by:	Lucia Mona, Ermann Ripepi, Gianluca Di Fiore (CNR-IMAA)
Note:	

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System - CUP B53C22002150006 (D.D. n. 130/2022)
 Funded by EU - Next Generation EU
 Mission 4 “Education and Research” - Component 2: “From research to business” -
 Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Table of contents

1	<i>INTRODUCTION</i>	5
2	<i>POLAR DATA REPOSITORY ARCHITECTURE</i>	5
3	<i>DATA FLOW TO POPULATE POLAR REPOSITORIES</i>	6
3.1	ERDDAP data server	7
3.2	ERDDAP Content management system	8
4	<i>APPLICATION OF PROCESSES</i>	9
4.1	Example of procedure automated into IADC system	10
4.2	Example of procedure not automated into IADC system	11
4.3	Example of dataset obtained from manual procedures	12
5	<i>DATA INTEGRATION AND ADDED VALUE</i>	13
5.1	Tools for data integration	14
5.1.1	<i>A Jupyter notebook for CCT data from IADC ERDDAP</i>	14
5.2	Added value and products from data integration	15
5.2.1	<i>Jupyter Notebook example with CCT data</i>	16
5.3	Visualisation as first powerful tool for data integration	17
5.3.1	Streamlit Dashboard	17
6	<i>CONCLUSIONS</i>	18
7	<i>REFERENCES</i>	18

Index of figures

Figure 1:	Polar data System overall Architecture	5
Figure 2:	Data and metadata flow	6
Figure 3:	How duties and main responsibilities will be distributed in the polar Data Management system (model SOURCE: Strasser et al. Promoting Data Stewardship Through Best Practices, DataONE, 2011).....	7
Figure 4:	ERDDAP IADC web page	7
Figure 5:	ERDDAP CMS linked to ERDDAP IADC datasets.	8
Figure 6:	The Data journey from observations to the Data Repository	9
Figure 7:	The Journey of our data from the measurement site (or laboratory) to the correct insertion in a data center or data repository (DEPOSIT/INGEST, PRESERVE) includes a fair number of steps and the use of different devices/tools/software.	10
Figure 8:	Schema of data flow for CCT data on IADC ERDDAP.....	10
Figure 9:	CCT D2 meteorological dataset on IADC ERDDAP.....	11
Figure 10:	Schema of data flow for not continuous Ozone and UV data on IADC ERDDAP.....	12
Figure 11:	SQM sensors installed on Gruvebadet (Ny-Alesund).	13
Figure 12:	Schema of data flow for not continuous SQM data on IADC ERDDAP	13
Figure 13:	Jupyter Hub for IADC (CCT data example).....	14
Figure 14:	Example of Erddapy request form IADC ERDDAP data center.....	15
Figure 15:	Example of Windrose plots for some recent years data request form retrieved from ERDDAP CCT data (Wind speed average and direction at 2m (m s-1)).	15

Figure 16: Linear model vs Xgboost 2023 CCT temperature based on daily mean temperatures from 2009 to 2022. 16
Figure 17: Stremlit dashboard to select ERDDAP node and dataset for data exploration and plot.. 17
Figure 18: Histogram ad Correlation Heatmap for CCT data retrieved from IADC ERDDAP. 17

Index of tables

Table 1: Statistical results of linear model vs Xgboost for 2023 CCT temperature. 16

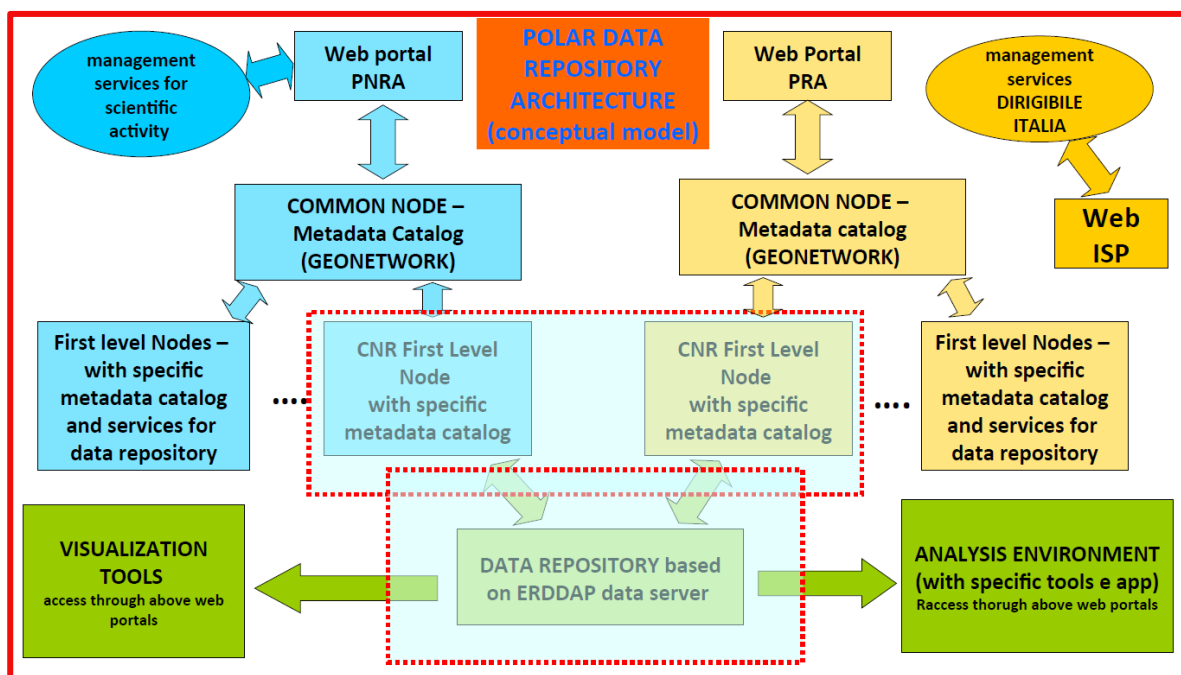
1 INTRODUCTION

Since 1985, Italy has maintained a presence in Antarctica, conducting research under the National Antarctic Research Program (PNRA). The Ministry of Scientific Research (MUR) funds the program, while the National Research Council (CNR) and the Italian National Agency for New Technologies, Energy, and Sustainable Economic Development (ENEA) handle scientific and logistical operations, respectively. The PNRA is inherently multidisciplinary, encompassing a wide range of scientific fields interested in Antarctic research, including marine, atmospheric, astrophysical, biological, geological, glaciological, and geophysical sciences. In parallel, Italy is also engaged in Arctic research through the Svalbard Integrated Earth Observing System (SIOS). This involvement led to the creation of the Italian Arctic Data Center (IADC), which aims to compile and make available data, particularly from the Arctic station “Dirigibile Italia” to the SIOS Data Management System (SDMS). This system also adheres to the principles of distributed networks and interoperability. The general objective is to use the resources made available by ITINERIS to primarily strengthen Data Management of the observations collected in the Arctic, with a special focus on long-term activities carried out in all domains (atmosphere, marine, cryosphere, ecosystems), and the Italian Arctic Data Center (IADC). The implementation plan illustrated below leverages the experiences and successes from developing both the NADC and IADC. It details the data flow processes and highlights some key applications that utilise core data.

2 POLAR DATA REPOSITORY ARCHITECTURE

The Italian Polar Data Repository is a scientific and technological infrastructure designed to gather, handle, publish and provide access to scientific data and metadata regarding Polar regions. The research activities in the Polar Area are promoted and supported by two government funded research programs: the Italian Antarctic National Research Program (PNRA) and the Italian Arctic Research Program (PRA). Figure 1 presents the overall conceptual model of Polar Data Repository Architecture.

Figure 1: Polar data System overall Architecture

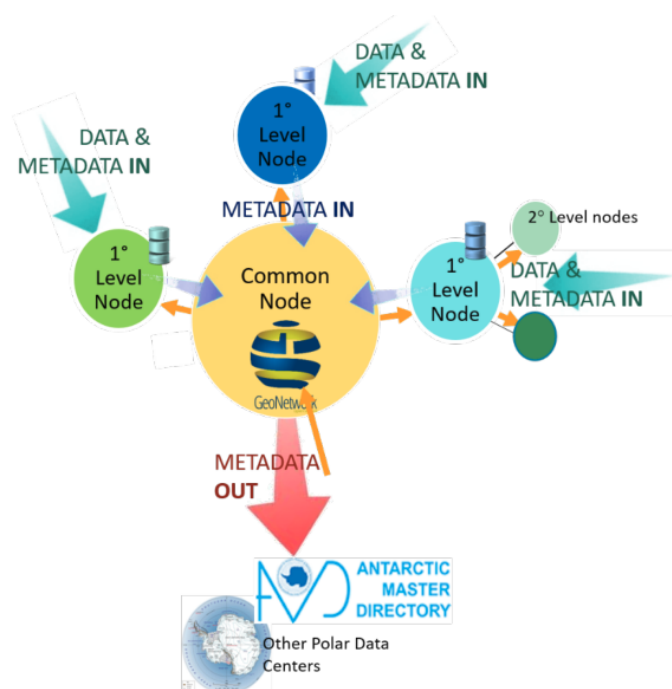


The presence of two separated programs makes it necessary to keep separate the front-end through which NADC and IADC data infrastructure and repositories can be reached. Figure 1 clearly shows how data services will be generally accessed through the web portals that both the PNRA and the Arctic (PRA) have set up. Figure 1 illustrates how the backend structures, particularly those with data repository functions, will be developed collaboratively. This approach optimizes hardware and software resources and facilitates the future integration of Arctic and Antarctic data repositories. Currently, metadata catalogues are available for both NADC and IADC. Tools for populating the metadata catalogue are available (see www.pnra.aq/dati). However, data repository functionalities using Environmental Research Division's Data Access Program (ERDDAP) and the ERDDAP content management system are currently implemented only for the IADC.

3 DATA FLOW TO POPULATE POLAR REPOSITORIES

The flow of data and metadata in the infrastructure is described in Figure 2. The Common Node manages a harmonised copy of all the dataset metadata shared by the First Level Nodes. However, the "master" copy of the metadata is maintained and updated by the First Level Nodes. The Common Node executes a periodical harvesting from all First Level Nodes in order to ensure synchronisation and update.

Figure 2: Data and metadata flow



In terms of data, each First Level Node is responsible for storing, maintaining, and updating their respective datasets. The data is kept within these nodes to ensure its integrity and security. While the Common Node does not store the actual data, it provides access points and summaries that facilitate data discovery and interoperability. This decentralised approach enables robust data management, optimising resource usage, and enhancing data accessibility.

While Common Node and First Level Nodes manage metadata, data are in charge of Second Level Nodes. Partners of PNRA/PRA that do not have a custom solution for data management are opting to use ERDDAP, an open-source software made by NOAA. This strategic decision reflects our commitment to adopting established technologies that adhere to industry standards and best practices.

First level node is committed to support those researchers that are not able to implement a fair complainant second level node. With this strategy, work of first-level node is optimized to secure the fairness of the whole system of systems without being sub-merged from a hardware and software point of view by a high amount of different data sources. Secure the data flow require a strong engagement of the research community and a strong interaction with data manager experts.

Figure 3 provides a graphical representation of the data life cycle, showing how roles and tasks between data providers and data manager should be distributed in the system.

Figure 3: How duties and main responsibilities will be distributed in the polar Data Management system (model SOURCE: Strasser et al. Promoting Data Stewardship Through Best Practices, DataONE, 2011)

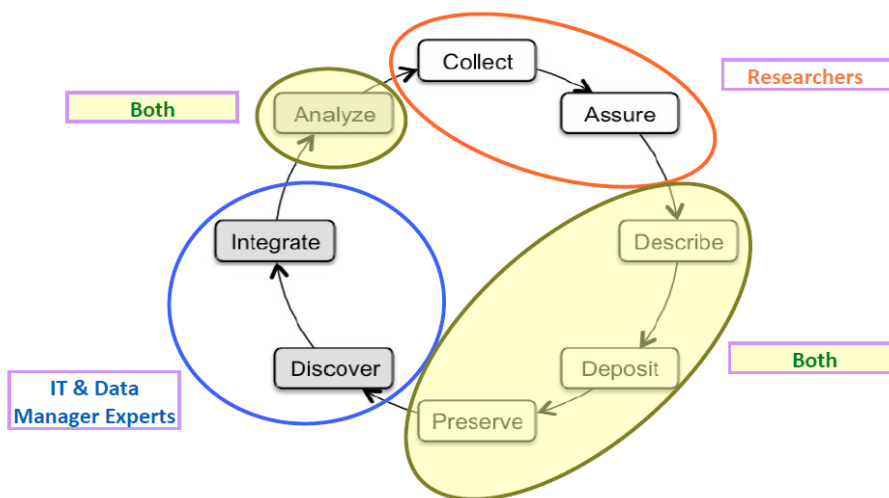
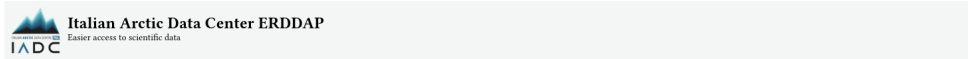


Figure 3 clearly indicate that cooperation between the two communities is necessary in crucial parts of the data life cycle if we desire to secure a continuous and robust data flow, and at the same time, creating optimal conditions for rapid growth of the catalogue and available resources. Significant effort is dedicated to provide dataset Quality Assurance (QA) documentation. This documentation includes comprehensive guidelines and checklists to ensure the accuracy, reliability, and consistency of datasets integrated into the repository. By maintaining high standards for data integrity and usability, this documentation plays a crucial role. Complete documentation is essential not only for the phases of data insertion into the repository but also for tracking all actions undertaken on the original data, including details about the experimental setup underlying the data acquisition process.

3.1 ERDDAP data server

ERDDAP is a data server that gives a simple, consistent way to download subsets of scientific datasets in common file formats and make graphs and maps, ideals of the OPeNDAP, WCS, SOS and OBIS standards. Figure 4 displays the ERDDAP page of the IADC portal. ERDDAP supports both human interaction (e.g. OPeNDAP requests) and machine-to-machine interoperability. ERDDAP data server supports several common data file formats (html table, netcdf, csv, txt, mat, json, etc.) and output files are created on-the-fly in any of these formats. ERDDAP implements FGDC Web Accessible Folder (WAF) with FGDC-STD-001-1998 and ISO 19115 WAF with ISO 19115-2/19139. ERDDAP tries to solve the problem of managing different data formats and sharing protocols.

Figure 4: ERDDAP IADC web page



ERDDAP > List of All Datasets

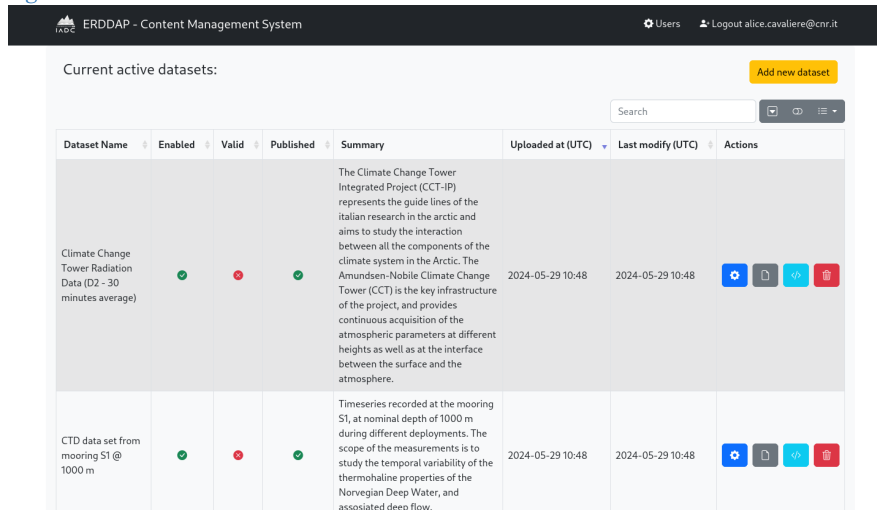
33 matching datasets, listed in alphabetical order.

Grid DAP	Sub-set	Table DAP	Make A	W M	Source Data Files	Access-ible	Title	Sum-mary	FGDC, ISO, Metadata	Back-ground Info	RSS	E mail	Institution	Dataset ID
						public	' The List of All Active Datasets in this ERDDAP '						CNR-ISP	allDatasets
						public	Aerosol scattering and absorption coefficients at the Gruevbadet Aerosol Laboratory (Svalbard)		F I M	background	6.855		CNR	aerosol_optical_gvb
						public	Aerosol scattering and absorption coefficients at the Gruevbadet Aerosol Laboratory (Svalbard), from 2022		F I M	background	6.855		CNR	aerosol_optical_gvb_2022
set						public	AIRQino arctic stations		F I M	background	3.355		CNR-IBE	airqino
						public	CH4 and CO2 turbulent flux at Ny Alesund		F I M	background	3.355		CNR	cct_output_g5y9m
						public	Climate Change Tower Meteorological Data (D2 - 30 minutes average)		F I M	background	3.355		CNR	cct_meteo_d2
						public	Climate Change Tower Radiation Data (D2 - 30 minutes average)		F I M	background	3.355		CNR	cct_radiation_d2
						public	Climate Change Tower Soil Temperature Data (D2)		F I M	background	3.355		CNR	cct_soil_d2
set						public	CTD (data from NISKIN Bottle) LRE1 ARCTIC Cruise Italian Arctic project CASSANDRA		F I M	background	3.355		OCS	ctd_cassandra_bottle_lqk7hb
set						public	CTD (DOWNCAST) LRE1 ARCTIC Cruise Italian Arctic project CASSANDRA		F I M	background	3.355		OCS	ctd_cassandra_downcast_lytlq
						public	CTD data set from mooring MD4 @ 35m and 85m (Kongsfjorden)		F I M	background	3.355		CNR	md4_ctd
						public	CTD data set from mooring S1 @ 1000 m		F I M	background	3.355		OCS	s1_ctd
						public	Daily equivalent black carbon from aerosol absorption coefficient data for ACPD submission		F I M	background	3.355		CNR-ISP	gilardini_acpd_2018_2021
						public	EGUphere-2023-1376 Equivalent Black Carbon Data		F I M	background	3.355		CNR-ISP	gilardini_acp_ebc_2023
						public	EGUphere-2023-1376 Meteo Data		F I M	background	3.355		CNR-ISP	gilardini_acp_mct_2023
						public	Equivalent black carbon from aerosol absorption coefficient		F I M	background	3.355		CNR-ISP	ebc_2010_2020
						public	EXAODEP-2020 ozone column at Barentsburg Svalbard station		F I M	background	3.355		CNR-ISP	ozone-barentsburg
						public	EXAODEP-2020 ozone column at Ny-Alesund Svalbard station		F I M	background	3.355		CNR-ISP	ozone-ny-alesund
						public	EXAODEP-2020 surface UV irradiance at Hornsund Svalbard station		F I M	background	3.355		CNR-ISP	uv-hornsund
						public	EXAODEP-2020 surface UV irradiance at Longyearbyen Svalbard station		F I M	background	3.355		CNR-ISP	uv-longyearbyen
						public	EXAODEP-2020 surface UV irradiance at Ny-Alesund Svalbard station		F I M	background	3.355		CNR-ISP	uv-ny-alesund

3.2 ERDDAP Content management system

One of the critical aspects of ERDDAP is the creation of dataset XML schemas, which contain essential metadata and links to dataset files. However, this process can be challenging due to its technical complexity and the requirement for specialized skills. To address the challenge of managing dataset XML schemas, an ERDDAP Content Management System (CMS) was implemented, as detailed in the following section. ERDDAP (Environmental Research Division's Data Access Program) Content Management System (CMS) is a versatile tool developed by CNR and designed to efficiently manage environmental data access.

Figure 5: ERDDAP CMS linked to ERDDAP IADC datasets.



The primary functionality of ERDDAP CMS is to simplify the creation of XML schemas, broadening accessibility to a wider user base and diminishing reliance on advanced technical skills. By providing a user-friendly interface and automating certain tasks, the CMS enhances the efficiency and effectiveness of managing dataset metadata within ERDDAP empowering members of the polar research community to independently contribute polar datasets to the system. This functionality is developed to enhance data accessibility within the system by leveraging common file formats compatible with the OPeNDAP standard. Initially centered on managing tabular data, the implementation conforms to the NetCDF Climate and Forecast (CF) Metadata Conventions Tabular. Additionally, Quality Assurance (QA) metadata is incorporated into the system to ensure the

accuracy and reliability of the datasets, further enhancing their usability and integrity. The system's architecture is grounded in an ERDDAP tabular dataset type capable of ingesting data from various sources, including local repositories like NetCDF files and ASCII files, as well as virtual resources such as databases and HTTP sources. Each dataset, whether continuous or discontinuous, employed a CF discrete sampling geometry, defined by the relationships among its spatiotemporal coordinates, referred to as its feature type. Notably, the implementation included time series and time series profile datasets, with the latter being more complex due to the presence of two element dimensions (profile and depth). Various analyses were conducted to develop effective methods for handling and integrating also trajectory-based and grid datasets into the system. This effort formed part of a broader study on new dataset typologies beyond tabular data, aimed at enhancing the system's capacity to accommodate diverse data formats commonly used in research environments.

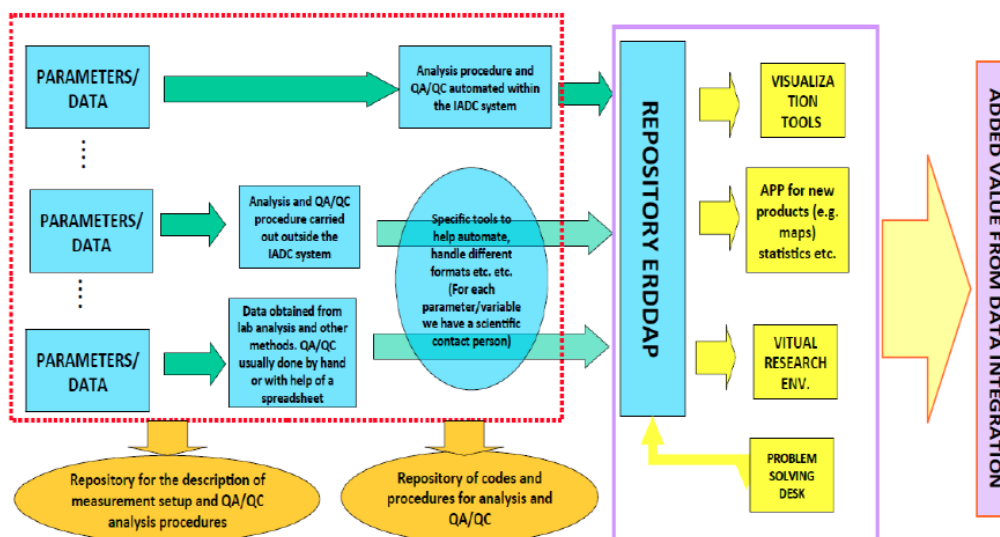
4 APPLICATION OF PROCESSES

Currently, our attention in implementing the scheme outlined in Figure 3 is directed towards the datasets and parameters, starting with the concept of "core data" as defined by SIOS. These datasets serve as fundamental resources for addressing identified key research questions in Earth System Science (ESS).

These datasets may be continuous or discontinuous, potentially associated with the scope of the measurement campaign or limitations inherent in the instrumentation used. Figure 6 further illustrates a general categorization of datasets based on analysis and QA/QC procedures:

- Analysed and processed automatically within the IADC system.
- Analysed and underwent QA/QC procedures outside the IADC system.
- Obtained from laboratory analysis or other methods and require manual QA/QC procedures.

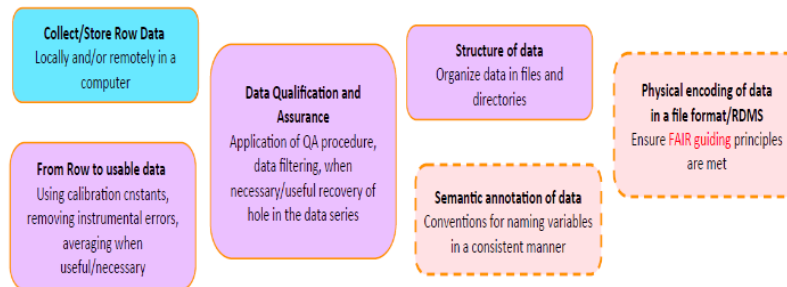
Figure 6: The Data journey from observations to the Data Repository



Crucially, associating specific information regarding acquisition specifics with each dataset is essential. As schematized in Figure 7, this information should detail the measurement procedures undertaken and any quality assurance measures implemented, such as outlier or anomaly removal

techniques. For these reasons, the sections below present example procedures for both continuous and discontinuous of some core datasets as defined by SIOS in the IADC.

Figure 7: The data flow from the measurement site (or laboratory) to the correct insertion in a data center or data repository (DEPOSIT/INGEST, PRESERVE) involves several steps and the use of various devices, tools and software.



4.1 Example of procedure automated into IADC system

The Amundsen-Nobile Climate Change Tower (CCT) is a scientific platform dedicated to studying the thermodynamic characteristics of the atmospheric boundary layer and the exchange processes between the surface and the lower layers of the atmosphere. The structure is composed of 17 modules equipped with patch boxes to provide a power supply and data connection that ends in a dedicated hut at 40 m from the tower, where the acquisition systems are located. The CCT provides continuous profiles of meteorological parameters at four levels up to 34 m, measurements of turbulent fluxes of momentum heat and moisture at two levels as well as of radiation balance components (visible and infrared). Measurements of the characteristics of the snow layer (depth and temperature) are also provided in conjunction with the atmospheric parameters.

Continuous data from the Climate Change Tower (CCT) are ingested into IADC ERDDAP through a direct connection with the relational database containing the raw data collected by the automatic system in Ny-Alesund. Java drivers (JDBC) serve as intermediaries between ERDDAP and the database, facilitating seamless data transfer. The raw data (D0), sampled at a temporal resolution of 1 minute, may contain errors. As illustrated in Figure 8, after undergoing quality control procedures (D1), the CCT dataset is transformed to average values over 30 minutes (D2) and transferred to the main node database and made publicly accessible on ERDDAP. These datasets undergo monthly updates on ERDDAP, providing 30-minute averages (μ) and standard deviations (σ) (Figure 9).

Figure 8: Schema of data flow for CCT data on IADC ERDDAP

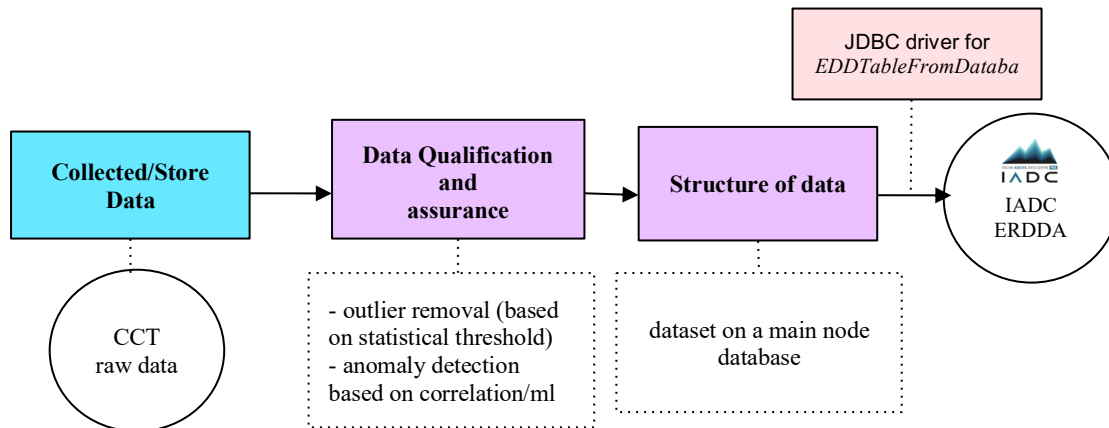


Figure 9: CCT D2 meteorological dataset on IADC ERDDAP

Italian Arctic Data Center ERDDAP
Easier access to scientific data

ERDDAP > info > cct_meteo_d2

Grid Sub-Set	Table Data	Make Graph	W Data Files	Source M	Access Data Files	Title	Summary	FGDC, ISO, Metadata	Background Info	RSS	Email	Institution	Dataset ID
data	graph	files	public			Climate Change Tower Meteorological Data (D2 - 30 minutes average)		F I M	background			CNR	cct_meteo_d2

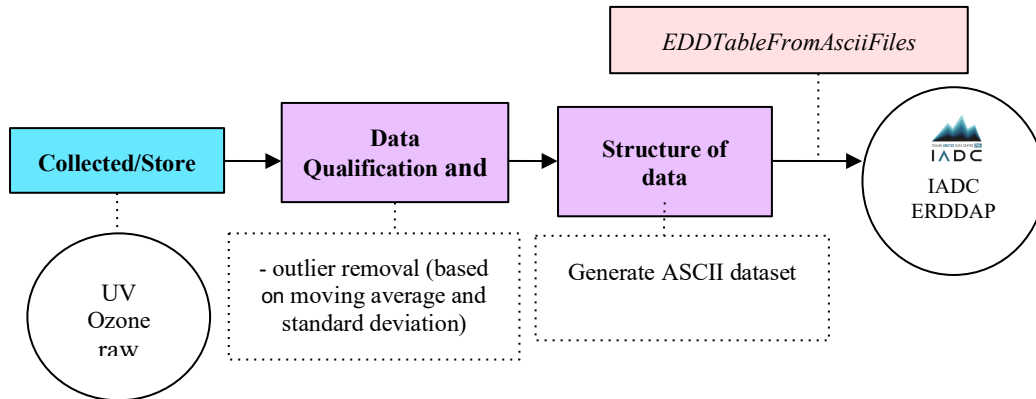
The Dataset's Variables and Attributes

Row Type	Variable Name	Attribute Name	Data Type	Value
attribute	NC_GLOBAL	cdm_data_type	String	TimeSeries
attribute	NC_GLOBAL	cdm_time_series_variables	String	station_id,latitude,longitude
attribute	NC_GLOBAL	Conventions	String	COARDS, CF-1.6, ACDD-1.3
attribute	NC_GLOBAL	Eastmost_Easting	double	11.86587
attribute	NC_GLOBAL	ENVRI_demonstrator	String	true
attribute	NC_GLOBAL	ENVRI_platform_long_name	String	Amundsen-Nobile Climate Change Tower in Ny-Ålesund
attribute	NC_GLOBAL	ENVRI_platform_short_name	String	CCT
attribute	NC_GLOBAL	ENVRI_platform_URI	String	https://www.tsp.cnr.it/index.php/en/infrastructures/observation-facilities/tower/
attribute	NC_GLOBAL	featureType	String	TimeSeries
attribute	NC_GLOBAL	geospatial_lat_max	double	78.92136
attribute	NC_GLOBAL	geospatial_lat_min	double	78.92136
attribute	NC_GLOBAL	geospatial_lat_units	String	degrees_north
attribute	NC_GLOBAL	geospatial_lon_max	double	11.86587
attribute	NC_GLOBAL	geospatial_lon_min	double	11.86587
attribute	NC_GLOBAL	geospatial_lon_units	String	degrees_east
attribute	NC_GLOBAL	infoUrl	String	https://metadata.iadc.cnr.it/geonetwork/srv/api/records/c0b0a6e-5e64-4e19-b4e9-b1a7b737824
attribute	NC_GLOBAL	institution	String	CNR
attribute	NC_GLOBAL	keywords	String	AIR TEMPERATURE, ATMOSPHERIC PRESSURE, ATMOSPHERIC WINDS, BOUNDARY LAYER TEMPERATURE, BOUNDARY LAYER WINDS, HUMIDITY, SURFACE WINDS

4.2 Example of procedure not automated into IADC system

The UV-ozone dataset, provided by WG6 of the Atmospheric Flagship Programme, is formed from observations of solar ultraviolet irradiance (UV) and ozone columns measured at Hornsund, Longyearbyen and Ny-Ålesund stations in the spring 2020 and its effect on the ultraviolet solar radiation reaching the ground. UV data concern spectral measurements and integrated values like erythemal, UV-B and UV-A both irradiances and doses.

Figure 10: Schema of data flow for not continuous Ozone and UV data on IADC ERRDAP



As illustrated in Figure 10, the quality of data is controlled by applying a procedure aimed to isolate and remove the doubtful measurements that present large deviations from the actual data trends. In all cases such a procedure performs (i) a smoothing of the corresponding data as a function of time through running average with a window adopted for each instrument, (ii) calculating the standard deviations of the data within the window at each step of the running average and (iii) removing the points that show large deviations from the window means taking into account the standard deviations. Before applying the smoothing, data are controlled for the gaps which are filled through interpolation. The temporal width of the window is usually between 5 and 15 minutes depending on the measurement frequency of the instrument. The criterion of the point removal is between 2- and 3-time standard deviation in case of UV data and 1 standard deviation in case of ozone column. The larger value for UV is chosen to prevent the removal of eventual correct measurements since the solar irradiance at Svalbard can be subject to significant variations due to the sharply changeable cloud conditions.

4.3 Example of dataset obtained from manual procedures.

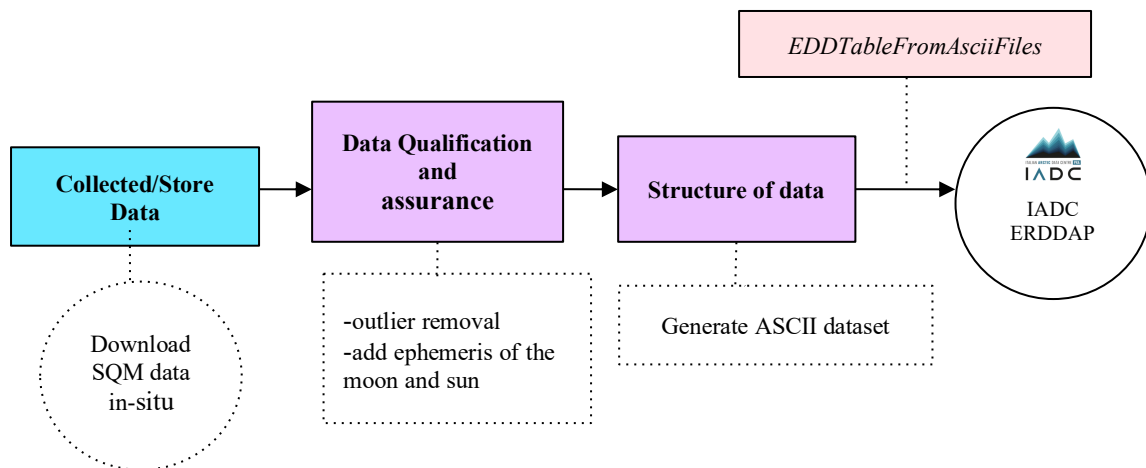
Sky-Quality Metres (SQM) datasets are continuous datasets collected on IADC, which is part of the AUXMON project. The AUXMON (AUXiliary MONitoring) project, developed by IBE-CNR, focuses on providing continuous monitoring of auxiliary variables to support investigations carried out during polar night. Among its core infrastructure are the Sky-Quality Metres (SQM) sensors, designed as low-cost solutions for measuring winter night sky brightness (Figure 11). These sensors were deployed on Gruvebat (Ny-Ålesund) and Antonsverk (Ny-Ålesund) in February and provide uninterrupted measurements exclusively focusing on night-time observations.

Figure 11: SQM sensors installed on Gruvebadet (Ny-Alesund).



As illustrated in Figure 12 SQM provides ERDDAP not continuous data because they remain non-operational from April to September.

Figure 12: Schema of data flow for not continuous SQM data on IADC ERDDAP



5 DATA INTEGRATION AND ADDED VALUE

Once the data has been collected and described, there is the need to extract value from it. The possibility to obtain new information from the analysis and integration of large amount of data provided by different domains and/or different measurements sites, gives the meaning and reason to invest in Data Management and in implementing complex and expensive data infrastructures. In approaching the issue of the added value extraction from data repositories, we can consider two parallel but different lines: (i) make it possible to extract new knowledge and information by integrating and comparing even very different data, and (ii) increase data use by an ever-increasing number of users from the most disparate categories.

Relatively to this point, the intention is to develop a wide range of services and tools by working along three precise lines of action:

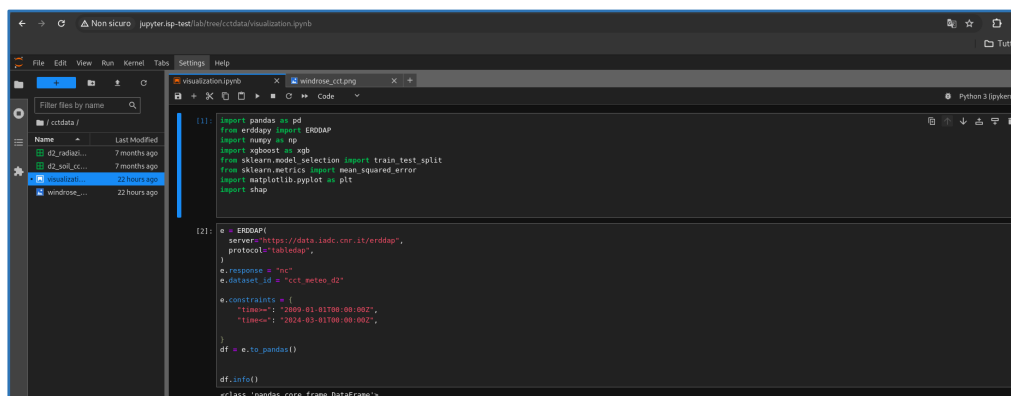
- 1 - through the development and implementation of an analysis Environment that allows expert users to access data and process it freely according to their needs,
- 2 - designing a series of tools and instruments for simple and/or medium complexity processing available to less experienced users, tools that help and guide the user where needed, including in the selection of suitable data sets and parameters,

3 - developing graphic tools that help carrying out data integration, taking advantage of their characteristics as tools that intuitively allow us to understand/glimpse relationships and can lead to subsequent in-depth, more quantitative analyses.

5.1 Tools for data integration

The Italian Polar Data Repository integrates a test environment based on a JupyterHub developed to facilitate simple retrieval from ERDDAP datasets. This integration aims to streamline access and analysis, enabling researchers to efficiently interact with and analyse polar data. JupyterHub is a powerful open-source platform designed to facilitate collaborative and interactive computing environments. It serves as a centralised hub that enables multiple users to access and work with Jupyter Notebooks, a popular tool for creating and sharing documents containing live code, equations, visualisations, and narrative text. With JupyterHub, organizations, educational institutions, and research teams can provide a shared computing environment where users can collaborate on data analysis, machine learning experiments, and scientific research. This platform simplifies the deployment and management of Jupyter Notebooks, making it easier for users to leverage the benefits of reproducible and data-driven workflows while maintaining security and scalability. During this initial phase, we developed several Jupyter Notebooks within the JupyterHub environment (Figure 13). These notebooks share a standardised approach for accessing ERDDAP data. This approach leverages *Erddapy*, an open-source Python library. *Erddapy* simplifies data retrieval from ERDDAP by utilising its RESTful Web Services. *Erddapy* can generate the appropriate URL for any ERDDAP request, including searching for datasets, acquiring metadata, and downloading the data itself. These procedures are designed to streamline data retrieval and analysis, enhancing the efficiency and effectiveness of working with datasets in the Italian Polar Data Repository.

Figure 13: Jupyter Hub for IADC (CCT data example).



```
(1): import pandas as pd
from erddapy import ERDDAP
import numpy as np
import xgboost as xgb
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
import matplotlib.pyplot as plt
import shap

(2): e = ERDDAP(
    server="https://data.iadc.cnr.it/erddap",
    protocol="tabledap",
)
e.response = "nc"
e.dataset_id = "cct_meteo_d2"

e.constraints = {
    "time": "2009-01-01T00:00:00Z",
    "time2": "2024-03-01T00:00:00Z",
}

df = e.to_pandas()

df.info()
<class 'pandas.core.frame.DataFrame'>
```

5.1.1 A Jupyter Notebook for CCT data from IADC ERDDAP

A Jupyter Notebook has been developed to retrieve Climate Change Tower (CCT) data through ERDDAP and plot key meteorological variables such as wind speed and temperature. As illustrated in Figure 14, utilising the *Erddapy* package, the notebook accesses the ERDDAP server and constructs the appropriate URLs for data requests. *Erddapy* also allows users to specify constraints such as time ranges or specific variables of interest, making the data retrieval process highly customizable.

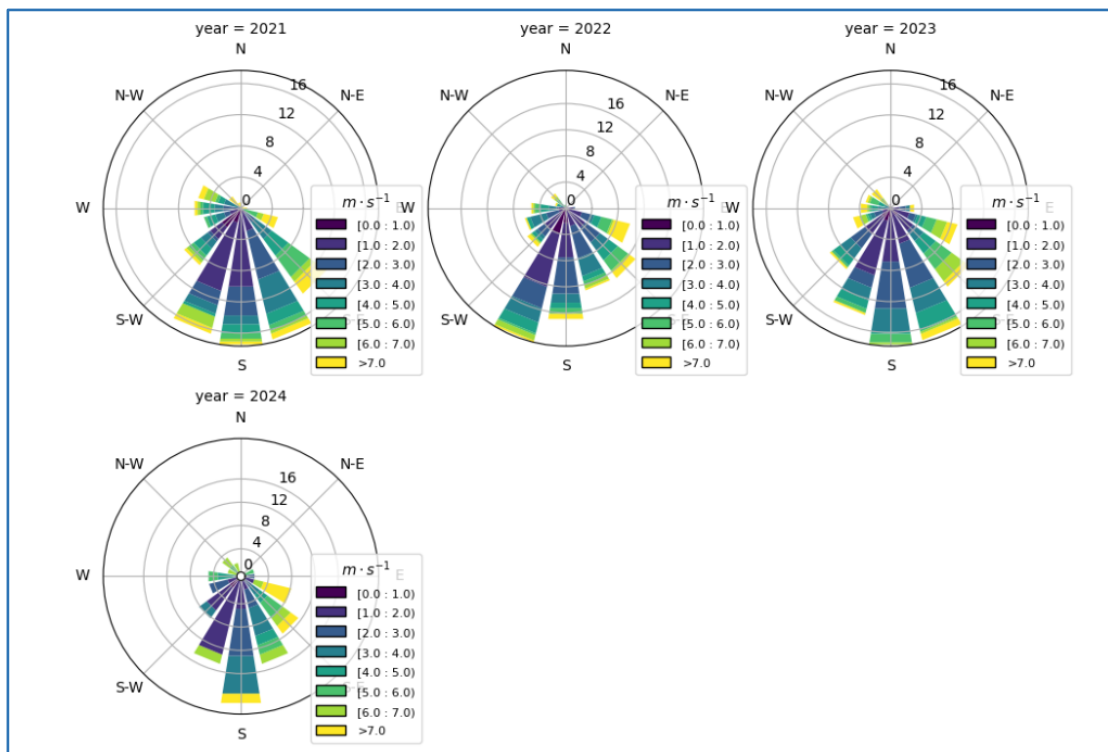
Figure 14: Example of Erddapy request form IADC ERDDAP data center.

```
e = ERDDAP(
    server="https://data.iadc.cnr.it/erddap",
    protocol="tabledap",
)
e.response = "nc"
e.dataset_id = "cct_meteo_d2"

e.constraints = {
    "time>=": "2009-01-01T00:00:00Z",
    "time<=": "2024-03-01T00:00:00Z",
}
df = e.to_pandas()
```

Once the data is retrieved, the notebook employs Python libraries like Pandas for data manipulation and Matplotlib for visualisation. Users can easily generate time-series plots of wind speed and temperature, providing valuable insights into the meteorological conditions recorded by the CCT. Figure 15 illustrates an example Windrose plot for a recent year's data request, showcasing the wind speed average and direction at 2 metres above the ground, retrieved from ERDDAP CCT data. This tool not only facilitates data analysis but also enhances the accessibility and usability of CCT datasets within the Italian Polar Data Repository.

Figure 15: Example of Windrose plots for some recent years data request form retrieved from ERDDAP CCT data (Wind speed average and direction at 2m (m s-1)).



5.2 Added value and products from data integration

In many cases, users need to perform standard or relatively simple analyses on the data they retrieve. The added value of the infrastructure lies in its ability to provide a vast amount of data, both in terms of quantity and, perhaps more importantly, in terms of variety and multidisciplinary content. By offering tools and services for these basic calculations (such as simple statistical averages and regressions), the infrastructure simplifies the user's experience. Users of the Polar Data Repository

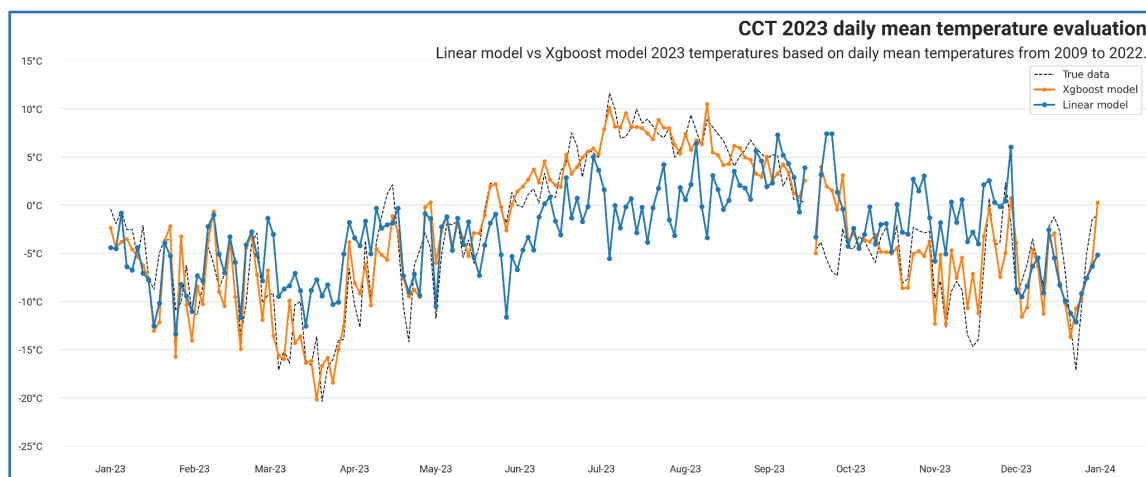
can conduct analyses directly within the system, eliminating the need to download datasets and process them locally. They can then download the results, saving time and effort. Practical examples include computing the erythemal dose from UV radiation data, deriving surface albedo from solar radiation data, calculating heat fluxes from eddy-covariance data, or generating vertical profiles of weather parameters. For each variable or group of variables, derived products are routinely used more extensively than the original data, facilitating further analyses and process descriptions. Additionally, employing state-of-the-art algorithms, such as machine learning, can automate the completion of datasets when there are substantial missing values due to file corruption or failure to record data points. This approach ensures more comprehensive and accurate datasets for analysis. Thus, these routines can be used and exploited to try to equip the Polar Data Repository with a large number of calculation "apps" allowing non-expert users to obtain the products and results they need, in a simple way and with little effort. Jupyter Notebook can be a resource to achieve this result.

5.2.1 Jupyter Notebook example with CCT data

While simple interpolation techniques like moving averages can work for time series data with small gaps, methods that are more accurate are needed for complex missing value patterns. Machine learning models excel at inferring missing values by learning the underlying relationships between variables in the data. We demonstrate this concept with a specific example: inferring missing CCT temperature at 2 metres data for 2023. Here, we hypothesise gaps in the CCT dataset and aim to predict these missing temperatures using a machine learning model.

To achieve this, we build a Jupyter Notebook that retrieves CCT data from ERDDAP using the *Erddapy* library. The core analysis involves comparing two models trained on data spanning from 2009 to 2022: linear regression and XGBoost. We incorporate additional variables into these models to improve prediction accuracy. These variables include relative humidity, wind direction, wind speed, and temporal attributes like season and month.

Figure 16: Linear model vs Xgboost 2023 CCT temperature based on daily mean temperatures from 2009 to 2022.



The results of this analysis are detailed in the table below, providing insight into the performance and effectiveness of each modelling approach.

Table 1: Statistical results of linear model vs Xgboost for 2023 CCT temperature.

MODEL	R-squared	Root Mean Squared Error	Mean Absolute Error
LINEAR	0.38	5.6°C	2.6°C
XGBOOST	0.77	3.5°C	1.3°C

By leveraging XGBoost's ability to capture complex relationships, this approach has the potential to significantly improve the accuracy of inferred CCT data compared to simpler methods. This data-driven strategy, utilising historical information and environmental factors, offers a promising solution for filling missing values in time series datasets, potentially leading to more complete and reliable data for further analysis.

5.3 Visualisation as first powerful tool for data integration

Plotting the data, and through this graphing making comparisons, and starting to understand if and what type of relationships can exist between the different variables in relation to the phenomenon and process that one is trying to investigate. In recent years, many steps forward have been made in the direction of making graphical tools available to users of databases and data repositories to effectively visualise the selected dataset. The Polar Data Repository, in addition to implementing tools and functions will try to develop a small graphic library that makes it possible to compare the temporal trend of a parameter in different stations, or the temporal trend of different parameters in the same station/platform, comparing the trend of the same quantity in different years.

5.3.1 Streamlit Dashboard

Our initial focus was to develop a user-friendly dashboard tool for exploring datasets within the IADC and NADC data centers. This tool seamlessly integrates with each ERDDAP instance, providing direct access to all variables within each dataset (Figure 17). Users can define temporal queries by selecting their desired range from the available data. Streamlit, an open-source Python library, served as the foundation for developing this interactive dashboard. It empowers users to not only explore the data but also visualise it in various ways. Supported visualisations include time series trends, histograms, and correlation matrices. These interactive plots facilitate a deeper understanding of relationships between variables and can help identify potential issues within the data (Figure 18).

Figure 17: Streamlit dashboard to select ERDDAP node and dataset for data exploration and plot.

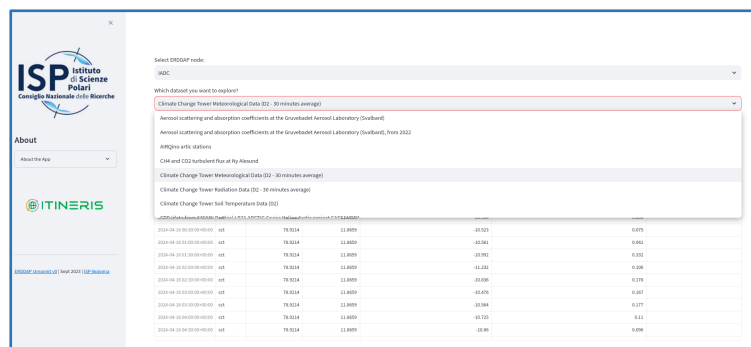
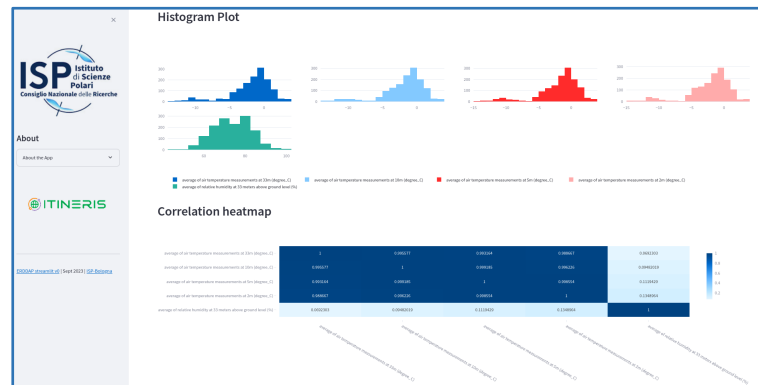


Figure 18: Histogram and Correlation Heatmap for CCT data retrieved from IADC ERDDAP.



6 CONCLUSIONS

The preceding sessions focused on the implementation plan for polar data repositories. This plan explored the system's architectural principles and data flow, demonstrating how these carefully chosen technical elements pave the way for the development of high-level science products. A critical role in this process is played by dedicated visualisation tools. The tools developed thus far have served as a powerful asset throughout the process, particularly for data integration. Their functionalities, such as interactive data dashboards and correlation matrices, facilitated the visual exploration of relationships between diverse datasets. Looking ahead, these visualisation tools hold immense potential. By enabling even more complex visual exploration, they will empower researchers to intuitively grasp underlying connections and patterns between dataset variables. This intuitive understanding will then serve as a springboard for subsequent in-depth and quantitative analyses.

7 REFERENCES

1. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm international conference on knowledge discovery and data mining (pp. 785-794)
2. Chiarelli C., Longo S., Principato A., Verazzo G. and Vitale V., 2020, NADC – Italian Antarctic Data Centre. 2020, Data Science Journal.
3. ERDDAP documentation, URL: <https://github.com/ERDDAP/erddap>
4. *Erddapy* documentation, URL: <https://ioos.github.io/erddapy/>
5. European Commission DG-R&I. (2015, October 01). “Guidelines on Data Management in H2020”. Retrieved from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
6. European Commission, 2016, “Open Science (Open Access)”, URL: <https://ec.europa.eu/programmes/horizon2020/en/h2020-section/open-science-open-access>
7. GEO. (2015, October 01). GEOSS Data Management Principles. Retrieved from http://www.earthobservations.org/documents/dswg/201504_data_management_principles_long_final.pdf
8. GEO. (2015, October 01). The GEOSS Data Sharing Principles. Retrieved from https://www.earthobservations.org/geoss_dsp.shtml
9. Jupyter hub project, URL: <https://jupyter.org/hub>

10. INSPIRE, 2013, “Guidelines for the encoding of spatial data” available at: http://inspire.ec.europa.eu/documents/Data_Specifications/D2.7_v3.3rc3.pdf
11. Nativi S., Craglia M., Pearlman J., 2012, The Brokering Approach for Multidisciplinary Interoperability: A Position Paper. International Journal of Spatial Data Infrastructures Research, Vol.7, 1-15. Available at: <http://ijsdir.jrc.ec.europa.eu/index.php/ijsdir/article/view/281/319>
12. Nativi,S., Craglia M., Pearlman J., 2013, “Earth Science Infrastructures Interoperability: The Brokering Approach”, Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of , vol.6, no.3, pp.1118,1129, June 2013.
13. Nativi S., Mazzetti P., Santoro M., Papeschi F., Craglia M., Ochiai O., 2015, Big Data challenges in building the Global Earth Observation System of Systems, Environmental Modelling & Software, Volume 68, June 2015, Pages 1-26.
14. Stremlit framework: <https://streamlit.io/>