AdriaClim

Climate change information, monitoring and management tools for adaptation strategies in Adriatic coastal areas

Project ID: 10252001

D 4.3.1 Implementation of the Network Data Repository

PP9 – CMCC

Final version

Public document

Project Acronym:	AdriaClim
Project ID Number:	10252001
Prject Title:	Climate change information, monitoring and management tools for adaptation strategies in Adriatic coastal areas
Priority Axis:	2 - Climate change adaptation
Specific objective:	2.1 - Improve the climate change monitoring and planning of adaptation measures tackling specific effects, in the cooperation area
Work Package Number:	4
Work Package Title:	Information system and products
Activity Number:	3
Activity Title:	Big data repository and networking services
Partner in Charge:	CMCC
Partners involved:	All partners
Status:	Final
Distribution:	Public
Date:	31/12/2021

Deliverable:	D4.3.1 [Implementation of the Network Data Repository]
Due month	M16 [December 2021]
Delivery Date	31/12/2021
Document status	V0.1
Authors	CMCC Foundation, ARPAE, ARPA FVG, CNR ISMAR, University ofBologna, IOF
Reviewers	

Table of contents

1. Aims and content of the document

The aim of the 4.3.1 Deliverable is to describe the “data repository” component of theAdriaClim Information System and its implementation.

The document is organized as follows:

- **section 2** gives a rough picture of the data repository architecture;
- **section 3** introduces ERDDAP, an open source data server, and describes how it has been leveraged to build the data repository, providing some additional implementation details;
- **section 4** provides the list of nodes composing the distributed data repository,and the most relevant datasets (modelling and observational) for each node.

2. The data repository

Given the large amount of data involved in the AdriaClim project, and since activities are performed in almost all the coastal regions of the Programme area (both on the Italian and the Croatian sides), an efficient access to datasets and a distributed approach has been considered.

The architecture is based on a federated data management system, where nodes of data are held and controlled at a remote location. The principles of federation keep asingle data source for each partner’s data and remove the need to keep multiple replicated sources synchronised.

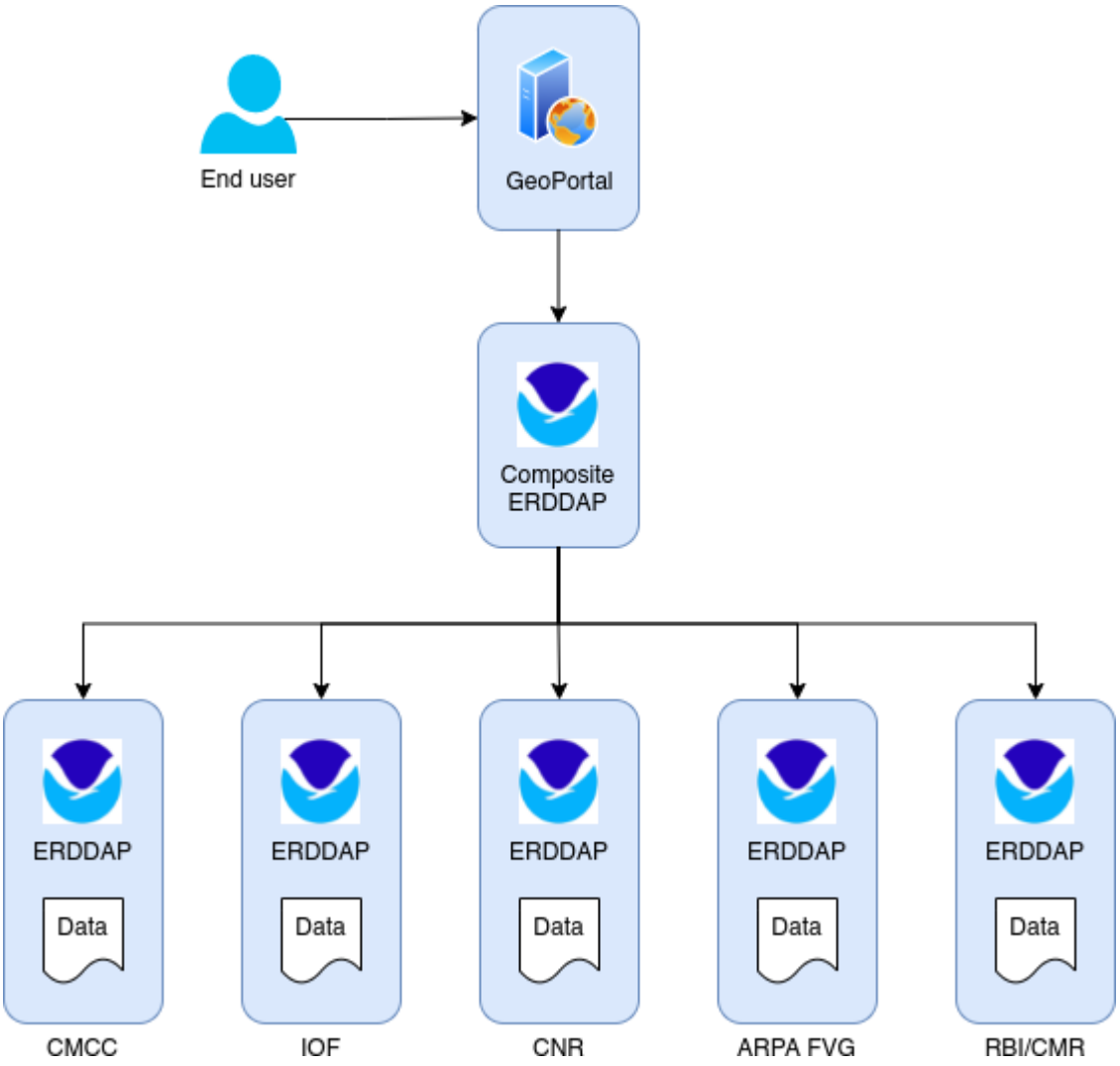
Instead of a single partner hosting the sum total of the project output data (this generally would involve significant duplication of data and a significant monetary

cost), the use of a federation of nodes, deployed at each partner data center, guarantees that each contributing organisation maintains full control of their data while also contributing to the whole dataset, which is not dependent on any single partner to keep publicly available.

When a user searches for specific data on the GeoPortal, the latter interacts with a “central” node that only stores metadata, routing queries to remote nodes which actually contain the data.

As described in the next session, the architecture is based on ERDDAP, the tool used to store, visualize and download the datasets.

The following diagram depicts a high level architecture of the distributed data repository.


End user
GeoPortal
Composite ERDDAP
ERDDAP
Data
CMCC
ERDDAP
Data
IOF
ERDDAP
Data
CNR
ERDDAP
Data
ARPA FVG
ERDDAP
Data
RBI/CMR

The way how to do that and some different scenarios are explained in the next section.

3. Implementation via-ERDDAP

What is ERDDAP

ERDDAP is a free and open source data server written in Java and developed by NOAA that provides a simple, consistent way to visualize and download subsets of gridded and tabular scientific datasets in common file formats, acting as a “middleman” between the end users and the various - local and remote - data sources.

It supports temporal and spatial subsetting or, for tabular data, via other constraints.

A variety of data types can be distributed on ERDDAP: in situ, satellite, or model

data among others.

The data can be downloaded in different formats (netcdf, csv, ESRIcsv, JSON, text and more).

ERDDAP also allows users to create customizable maps and graphs, and then to generate images in PNG format, transparent PNG, PDF and more.

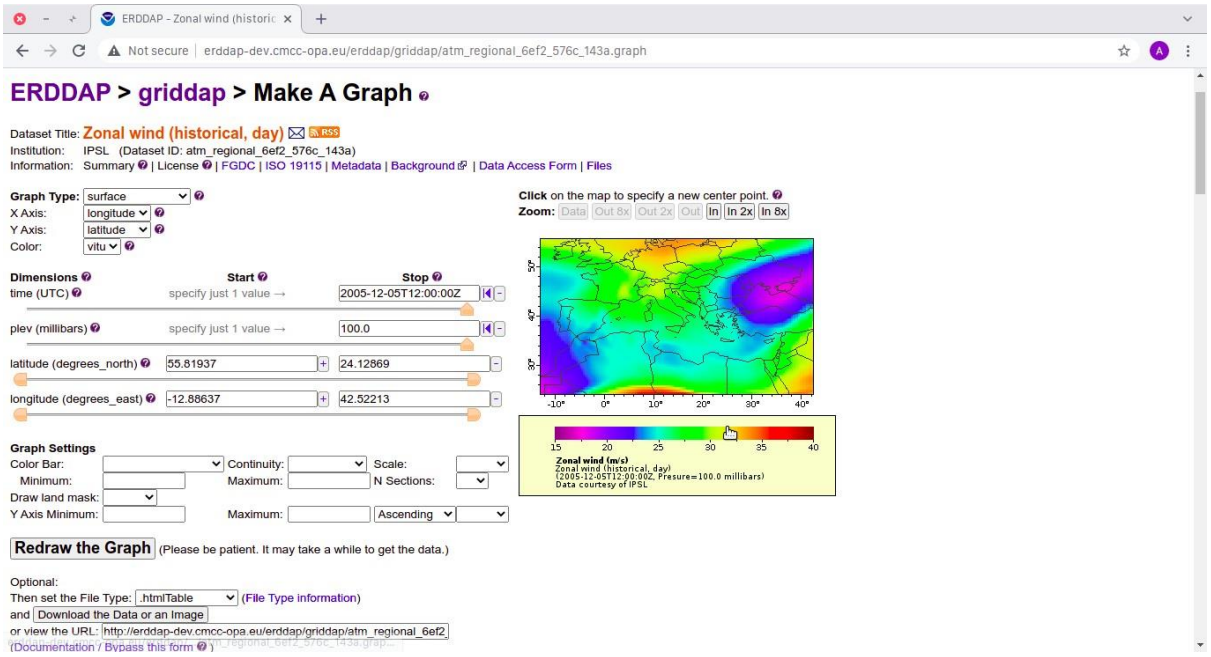
ERDDAP is not only a web application, but also a RESTful web service: indeed, for every web page with a form, there is a corresponding ERDDAP web service that is designed to be easy for other applications to use. In this way, it can return user-interface results as a table of data in different - computer-program friendly - file types, the most common of which are:

- CSV;
- JSON;
- mat;
- NetCDF-3 binary;
- htmlTable (an HTML web page with the data in a table);
- Google Earth kml.

Additionally, ERDDAP is compatible with the current WMS 1.3.0 standard (GetCapabilities, GetMap - opaque, GetMap - transparent), and then it is able to provide a basic WMS service for each dataset configured.

It also offers RSS and email/URL subscriptions services.

In the following image, the “griddap” service is shown. It allows requests for a data subset, graph, or map from a gridded dataset (for example, sea surface temperature data from satellite), via a specially formed URL. It uses the OPeNDAP Data Access Protocol (DAP) and its projection constraints.



In this case the dataset is related to the “Zonal wind” variable, historical data from 1970 to 2005, daily timestep.

Types of datasets in ERDDAP

There are two main categories of datasets in ERDDAP: **EDDGrid** datasets handle gridded data, and **EDDTable** datasets handle tabular data.

In EDDGrid datasets, data variables are multi-dimensional arrays of data (for example, for satellite data and model data).

There must be an axis variable for each dimension, and axis variables must be specified in the order that the data variables use them. All data variables must use all of the axis variables. Each dimension must be in sorted order (ascending or descending), and can be irregularly spaced.

The main EDDGrid dataset types are:

- EDDGridFromAudioFiles, aggregating data from a group of local audio files;
- EDDGridFromDap, handling gridded data from DAP servers;
- EDDGridFromEDDTable lets you convert a tabular dataset into a gridded dataset;
- EDDGridFromErddap, handling gridded data from a remote ERDDAP;

- EDDGridFromNcFiles, aggregating data from a group of local NetCDF (v3 or v4) .nc and related files;
- EDDGridCopy can make a local copy of another EDDGrid's data and serves data from the local copy.

In EDDTable datasets, tabular data can be represented as a database-like table with rows and columns. Each column (a data variable) has a name, a set of attributes, and stores just one type of data. Each row has an observation (or group of related values). Examples of tabular data are in-situ buoy, station, and trajectory data.

The main EDDTable dataset types are:

- EDDTableFromAsciiFiles, aggregating data from comma-, tab-, semicolon-, or space-separated tabular ASCII data files;
- EDDTableFromAudioFiles, aggregating data from a group of local audio files;
- EDDTableFromAwsXmlFiles, aggregating data from a set of Automatic Weather Station (AWS) XML files;
- EDDTableFromCassandra, handling tabular data from one Cassandra table;
- EDDTableFromDatabase, handling tabular data from one database table;
- EDDTableFromEDDGrid lets you create an EDDTable dataset from an EDDGrid dataset;
- EDDTableFromErddap, handling tabular data from a remote ERDDAP;
- EDDTableFromHttpGet is ERDDAP's only system for data import as well as data export;
- EDDTableFromJsonlCSVFiles aggregates data from JSON Lines CSV files;
- EDDTableFromMultidimNcFiles aggregates data from NetCDF (v3 or v4) .nc files with several variables with shared dimensions;
- EDDTableAggregateRows can make an EDDTable dataset from a group of EDDTable datasets;
- EDDTableCopy can make a local copy of many types of EDDTable datasets and then re-serve the data quickly from the local copy.

ERDDAP federations

“ERDDAP federation” is the built-in feature used to set up the distributed data repository: in this way it is possible to avoid data duplication and to minimize costs, while each partner still manages its own data in its own data center.

ERDDAP is designed to work well within a federation of “ERDDAPs”. Instead of moving all the datasets from different locations to a centralized data server, ERDDAP's design works with the world as it is: each data center which produces data can continue to maintain, curate, and serve their data, and yet, with ERDDAP, the data can also be instantly available from a centralized ERDDAP, without the need for

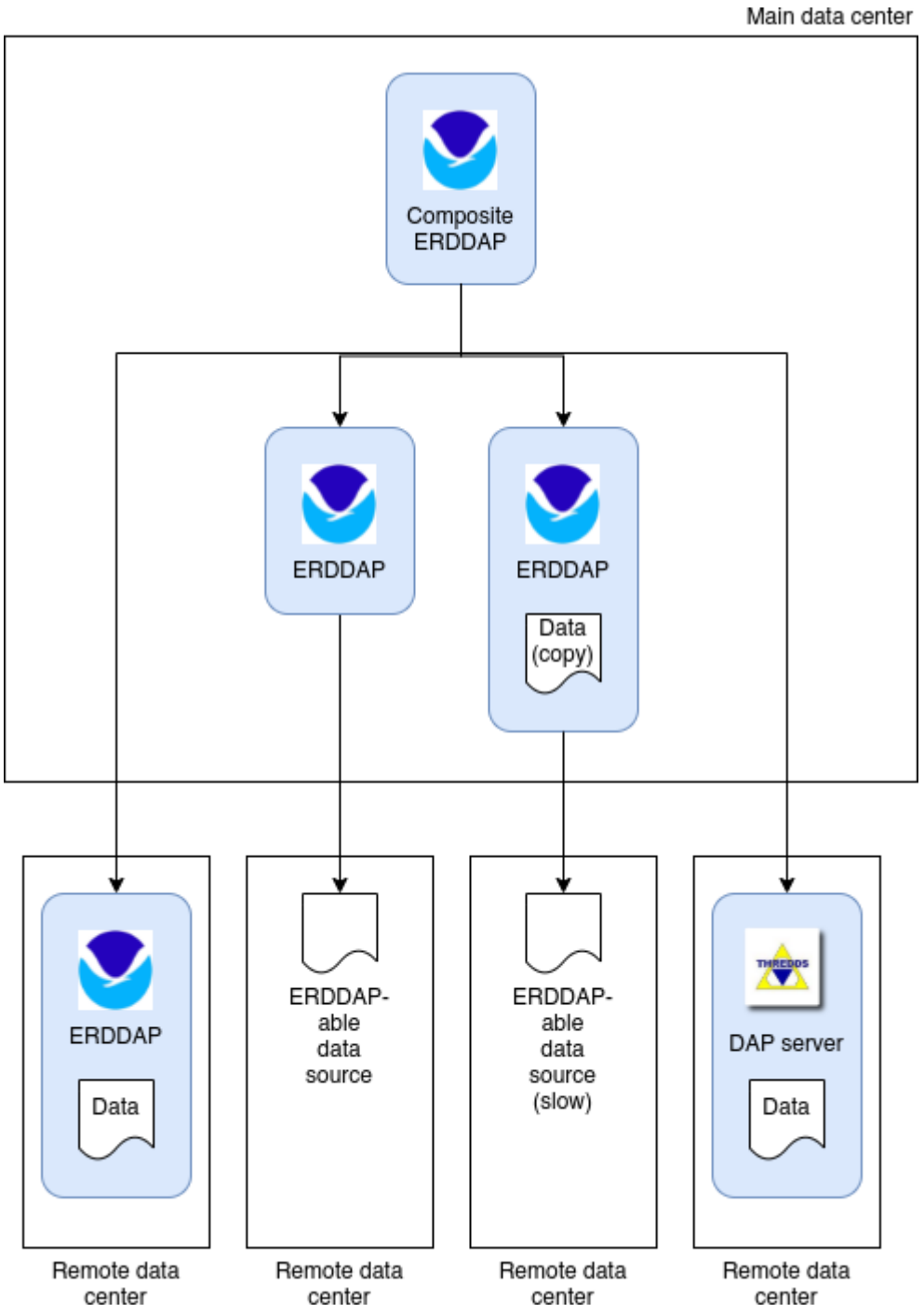
transmitting the data to the centralized ERDDAP or storing duplicate copies of the data.

An ERDDAP federation therefore consists of the following components, some of them are optional:

- a centralized “composite” ERDDAP (in some cases more than one in high availability, for fault tolerance);
- remote ERDDAP/(Open)DAP servers;
- non-ERDDAP but “ERDDAP-able” remote data sources;
- slow ERDDAP-able remote data sources, for which a copy of the datasets is required.

The **composite ERDDAP** is a regular ERDDAP except that it just serves data from other ERDDAPs. When it receives a request for data, it redirects the data request to the remote ERDDAP server actually hosting that data. The redirect is transparent to the user regardless of the client software.

Remote ERDDAP and **(Open)DAP servers** can be directly connected to the central composite node through EDDGridFromErddap or EDDTableFromERDDAP dataset types. If the remote server is some other type of DAP server (THREDDS, Hyrax, or GrADS), it can be connected via-EDDGridFromDap dataset type.



An **ERDDAP-able data source** is a data source from which ERDDAP can read data.

In this case, it is possible to set up another ERDDAP in the grid which is responsible for serving the data from this data source.

If the **ERDDAP-able data source** is delivered by a **slow** service or over a slow network, another ERDDAP can be deployed, storing a copy of the dataset through EDDGridCopy and/or EDDTableCopy dataset types.

4. Data repository nodes and datasets

CMCC node

The CMCC node is based on a virtual machine equipped as follows:

- Processors: 8 vCPU
- Memory: 12GB RAM
- Storage: 512GB for ERDDAP data (the “bigParentDirectory” in ERDDAP terminology)
- OS: Ubuntu 20.04 LTS
- ERDDAP v2.14, deployed as a Docker container.
  Allocated memory (min/max): 4/8 GB

The storage doesn't include the data sources (mainly NetCDF files), they are provided via-NFS from a remote - read-only - storage.

Currently, it provides the following datasets:

1. MedCordex oceanographic historical data (1970 - 2005);
2. MedCordex oceanographic projections data RCP 8.5 (2006 - 2050);
3. MedCordex atmospheric-land historical data (1970 - 2005);
4. MedCordex atmospheric-land projections data RCP 8.5 (2006 - 2050).

The mentioned MedCordex datasets are a temporal subsampling of the original ones covering 1951-2100 years and considering only RCP8.5 of the two available (RCP4.5 and RCP8.5).

In the next future the CMCC node will also provide the output data of the AdriaClim subregional earth system model downscaled from MedCordex. Data will be related to five modeling components: the WRF based atmospheric component, the WRFHydro based hydrology component, the NEMO based ocean component, the WWIII based wave component and the BFM based marine biochemistry one.

Regarding the Apulia Pilot area, the CMCC node will provide (i) observational data from in situ available monitoring stations and new instruments which are planned to be installed (a multispectral radiometer and sensors for water quality), (ii) modeled data from the foreseen sub-regional and coastal ocean downscaling.

Additionally the CMCC node hosts the following datasets produced by University of

Bologna:

A) COPERNICUS MARINE ENVIRONMENT MONITORING SERVICE (CMEMS)

A.1) Physical Reanalysis in the Adriatic Sea: Sea surface height, Sea water potential temperature, Sea water Salinity and Horizontal Velocity: Eastward and Northward components from 1987 to 2019.

A.2) Biochemical Reanalysis in the Adriatic Sea: Nitrate, Phosphate, Ammonium, Chlorophyll, Oxygen and Ocean pH from 1999 to 2019.

A.3) Sea Waves Reanalysis in the Adriatic Sea: Spectral Significant Wave Height, Wave Period at spectral moments wave period Tm10 and Mean Wave Direction from Spectral Significant Wind direction from 1993 to 2019.

B) COPERNICUS MARINE ENVIRONMENT MONITORING SERVICE (CMEMS) REPROCESSED OCEAN COLOUR

B.1) Reprocessed Satellite Chlorophyll in the Adriatic Sea: REP- Monthly averaged Multi sensor mass concentration of Chlorophyll- A in seawater at 1 km of resolution, REP- Daily Interpolated Multi sensor mass concentration of Chlorophyll-A in seawater 1 km of resolution and REP- Daily Climatology of mass concentration of Chlorophyll-A in seawater at 1 km of resolution.

B.2) Reprocessed Monthly Mean satellite wind: Eastward wind speed and Northwardwind speed

C) COPERNICUS CLIMATE CHANGE SERVICE (C3S)

C.1) ERA5 Reanalysis (Middle and Upper Air) in the Mediterranean Sea: Air temperature, Specific Humidity, Geopotential, Wind Component U and Wind Component V from 1979 to 2020.

C.2) ERA5 Reanalysis (Near Surface) in the Mediterranean Sea: Air temperature, Specific Humidity, Geopotential, Wind Component U and Wind Component V from 1979 to 2020.

C.3) ERA5 Reanalysis (Surface Level) in the Mediterranean Sea: Temperature at 2 m, Dewpoint temperature at 2 m, Mean Sea Level Pressure, 10 m Wind Component U, 10 m Wind Component V, Total Precipitation, Cloud area fraction, Surface net solar radiation, Surface downward solar radiation, Surface net thermal radiation and Surface downward thermal radiation from 1979 to 2020

C.4) European Daily Gridded Observational (EOBS): Relative Humidity, Air pressureat sea level, Surface downwelling shortwave flux in air, Thickness of rainfall amount, Air temperature – Daily Mean Temperature, Air Temperature – Daily Minimum Temperature and Air Temperature – Daily Maximum Temperature from 1950 to 2020.

D) COPERNICUS EMERGENCY MANAGEMENT SERVICE

D.1) European Flood Awareness System (EFAS): Mean discharge in the last 6 hoursfrom 1991 to 2020.

IOF node

The IOF node is currently under construction. Node will be based on ERDDAP, deployed as a container on bare metal and hosted on a dedicated server located in IOF server room. It will provide the following datasets:

I. In situ monitoring stations in Split-Dalmatia pilot site:

a) Parameters at **ST101** (central part of the Kaštela Bay, 43°51′N; 16°38′E, depth 38 m) at 0, 5, 10, 20, 30 m:

- Temperature
- Salinity
- Chlorophyll *a*
- Diatom abundance
- Dinoflagellate abundance
- Nanoflagellate abundance
- Total phytoplankton abundance
- production of heterotrophic bacteria (with different DNA content, i.e. High-DNA bacteria and Low- DNA bacteria)
- abundance of heterotrophic bacteria (with different DNA content, i.e. High-DNA bacteria and Low- DNA bacteria)
- abundances of two cyanobacteria groups (Prochlorococcus and Synechococcus)
- abundances of pico-eukaryotic algae
- abundances of protistan grazers (heterotrophic nanoflagellates).

b) Parameters at **CJ 009** (located near the island of Vis, central Adriatic Sea, 43°N; 16°33′E, depth 103 m) at 0, 10, 20, 30, 50, 75, 100 m:

- Temperature
- Salinity
- Chlorophyll *a*

- Diatom abundance
- Dinoflagellate abundance
- Nanoflagellate abundance
- Total phytoplankton abundance

- production of heterotrophic bacteria (with different DNA content, i.e. High-DNA bacteria and Low- DNA bacteria)
- abundance of heterotrophic bacteria (with different DNA content, i.e. High-DNA bacteria and Low- DNA bacteria)
- abundances of two cyanobacteria groups (Prochlorococcus and Synechococcus)
- abundances of pico-eukaryotic algae

abundances of protistan grazers (heterotrophic nanoflagellates).

Measurements and samplings have been carried out since January 2020 and willlast until the end of the project.

c) Parameters determined from sediments sampled at **ST101, ST103, CJ 009, Pantan1, Pantan2**:

- Grain size (gravel, sand, silt and clay content in %),
- Carbonate content (%)
- LOI (lost of ignition)-organic matter (%),
- Total N (%),
- Organic C (%),
- P (%).

Samplings have been carried out since August 2020 and will last until the end of the project at stations ST101, ST103, CJ 009, while samplings at stations Pantan1 and Pantan2 will end in December 2021.

II. Outputs from ROMS and BFM models applied to Kaštela Bay.

III. Continuous measurements from meteo-ocean station in the Neretva River estuary:

- Air temperature
- Air humidity
- Air pressure
- Sea level
- Sea temperature
- Salinity
- Hydrostatic pressure

IV. Continuous measurements from autonomous sensors in the Neretva River estuary:

- Sea temperature (six locations)
- Salinity (six locations)
- Dissolved oxygen (three locations)

V. Outputs from ROMS model applied to the Neretva River estuary.

CNR node

The CNR node is currently under construction. Final node will be based on ERDDAP,deployed as a Docker container and hosted on ISMAR-CNR data infrastructure.

It will provide the following datasets:

- Data from In situ monitoring stations for Veneto pilot area sea surface height and waves (significant wave height, mean wave period and mean wave direction). An historical series is already available from Piattaforma Acqua Alta and we're working on licensing issues to make release series from other stations .
- Veneto pilot area sea surface height and waves (significant wave height, mean wave period and mean wave direction) from SHYFEM-WWM simulations for the MedCordex and AdriaClim historical and climate scenarios.

ARPA FVG node

The ARPA FVG ERDDAP node will provide the following datasets:

1. Pilot Area temperature, salinity, sea surface height fields from benchmark yearly simulation (selected set of SHYFEM model node hourly outputs for year 2018) netCDF format files, CF convention.
2. Pilot Area temperature, salinity, sea surface height fields from sensitivityclimate change cases yearly simulation (selected set of SHYFEM model nodehourly outputs) netCDF format files, CF convention.
3. Pilot Area temperature, salinity, dissolved oxygen and chlorophyll measured profiles (2014-2021 monthly

cruises) ASCII CSV files

4. Pilot Area macrozoobenthos measures (2008-2018 seasonal cruises) ASCII CSV files
5. Pilot Area meteorological measures (2000-2021 hourly records) ASCII CSV files

RBI/CMR node

The RBI/CMR node is fully based on ERDDAP, deployed as a Docker container.It will provide the following datasets.

In situ meteorology (atmospheric data) 1nm off the coast In situ meteorology (atmospheric data) 6nm off the coastIn situ SST 1nm off the coast

In situ SST 6nm off the coast