



A multi-service data management platform for scientific oceanographic products

Alessandro D’Anca¹, Laura Conte¹, Paola Nassisi¹, Cosimo Palazzo¹, Rita Lecci², Sergio Creti², Marco Mancini¹, Alessandra Nuzzo¹, Maria Mirto¹, Gianandrea Mannarini², Giovanni Coppini², Sandro Fiore¹, and Giovanni Aloisio^{3,4}

¹Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC), Advanced Scientific Computing Division, via per Monteroni, 73100 Lecce, Italy

²Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC), Ocean Predictions and Applications, via Augusto Imperatore 16, 73100 Lecce, Italy

³Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC), Supercomputing Center, via per Monteroni, 73100 Lecce, Italy

⁴University of Salento, Complesso Ecotekne, via per Monteroni, 73100 Lecce, Italy

Correspondence to: Alessandro D’Anca (alessandro.danca@cmcc.it)

Received: 15 May 2016 – Discussion started: 4 July 2016

Revised: 30 December 2016 – Accepted: 4 January 2017 – Published: 13 February 2017

Abstract. An efficient, secure and interoperable data platform solution has been developed in the TESSA project to provide fast navigation and access to the data stored in the data archive, as well as a standard-based metadata management support. The platform mainly targets scientific users and the situational sea awareness high-level services such as the decision support systems (DSS). These datasets are accessible through the following three main components: the Data Access Service (DAS), the Metadata Service and the Complex Data Analysis Module (CDAM). The DAS allows access to data stored in the archive by providing interfaces for different protocols and services for downloading, variables selection, data subsetting or map generation. Metadata Service is the heart of the information system of the TESSA products and completes the overall infrastructure for data and metadata management. This component enables data search and discovery and addresses interoperability by exploiting widely adopted standards for geospatial data. Finally, the CDAM represents the back-end of the TESSA DSS by performing on-demand complex data analysis tasks.

1 Introduction

TESSA (Development of TEchnology for Situational Sea Awareness) is a research project born from the collaboration between operational oceanography research and scientific computing groups in order to strengthen the operational oceanography capabilities in Southern Italy for use by end users in the maritime sector, tourism and environmental protection. This project has been very innovative as it has provided the integration of marine and ocean forecasts and analyses with advanced technological platforms. Specifically, an efficient, secure and interoperable data platform solution has been developed to provide fast navigation and access to the data stored in the data archive, as well as a standard-based metadata management support. This platform mainly targets scientific users and a set of high-level services such as the decision support systems (DSS) for supporting end users in managing emergency situations due to natural hazards in Mediterranean Sea. As example, the DSS WITOIL (De Dominicis et al., 2013a, b) is crucial to face oil spill accidents which could have severe impacts on the Mediterranean Basin and contributes effectively to the reduction of natural disaster risks. Moreover, the DSS Ocean-SAR (Coppini et al., 2016) supports the search-and-rescue (SAR) operations following accidents and EarlyWarning manages alerts in case of

extreme events by providing near-real-time information on weather and oceanographic conditions. Finally VISIR (Manarini et al., 2016a, b) is able to compute the optimal ship route with the aim to increase safety and efficiency of navigation. It relies on various forecast environmental fields which affect the vessel stability in order to ensure that the computed route does not result in an exposure to dynamical hazards. In this context, the developed platform faces the lack of an efficient dissemination of marine environmental data in order to support the situational sea awareness (SSA), which is strategically important for the maritime safety and security (Lecci et al., 2015). In fact, an updated situation awareness requires an advanced technological system to make data available for decision makers, improving the capacity of intervention and avoiding potential damages.

In a “data-centric” perspective, in which different services, applications or users make use of the outputs of regional or global numerical models, the TESSA data platform meets the request of near-real-time access to heterogeneous data with different accuracy, resolution or degrees of aggregation. The design phase has been driven by multiple needs that the developed solution had to satisfy. First of all, nowadays each data management system should satisfy the FAIR guiding principles for scientific data management and stewardship (Wilkinson, 2016): they face the lack of a shared, comprehensive, broadly applicable, as well as articulated guideline concerning the publication of scientific data. Specifically, they identify findability, accessibility, interoperability and reusability as the four principles which need to be addressed by data (and associated metadata) to be considered a good publication. As explained later, the employment of well-known standards concerning protocols, services and output formats satisfies these requirements. In addition, the need for a service that provides information about sea conditions 24/7 at high and very high spatial and temporal resolution has been addressed by exploiting high performance and high availability hardware and software solutions. The developed platform must be able to support the requests of intermediate and common users. To this end, data must be available in the native and standard format (NetCDF) (NetCDF, 2017a, b) as output of the oceanographic models, through a simple and intuitive platform suitable for machine-based interactions, in order to feed user-friendly services for displaying clear maps and graphs. At the end, the system has to provide on-demand services to support decisions; the users must be able to interact with the datasets produced by the models in near-real time, so the platform has to provide services and datasets suitable for on-demand processing minimizing the downloading time and the related input file size.

This paper is organized as follows: Sect. 2 presents a survey on related work, and in Sect. 3 an architectural overview of the main data platform components is provided. The focus is on providing easier and unified access to the heterogeneous data produced in the frame of TESSA project. The implemented data archive and the modules that make the datasets

available to the entire set of services are presented in Sect. 4. In Sect. 5, the Metadata Service features are detailed from a methodological and technical point of view. The module developed in order to serve remote DSS submission requests is described in Sect. 6. Finally in Sect. 7 a few use cases are presented, highlighting the operational chains which exploit the data platform services.

2 Related work

The need for marine and oceanic data management supporting the SSA and the operational oceanography has led to the definition and development of various platforms that provide different types of data and services. In this context, the MyOcean project (MyOcean, 2017; Bahurel et al., 2009) can be mentioned, as this project implementation was in line with the best practices of the GMES/Copernicus framework (Copernicus, 2017).

Specifically, regarding data access, MyOcean provides the scientific community with a unified interface designed to take into account various international standards (ISO 19115 (iso, 2003a), ISO 19139 (iso, 2003b), INSPIRE (INSPIRE, 2017; INSPIRE-directive, 2007)). Concerning the data management, MyOcean relies on OPeNDAP/THREDDS for tasks like map subsetting and FTP for direct download. In addition, a valid solution is represented by the Earth System Grid Federation (ESGF) (Cinquini et al., 2014; ESGF, 2017), a federated system used as metadata service with advanced features, that will be described later. EMODNET MEDSEA Checkpoint (EMODNET, 2017; Moussat et al., 2016) is another solution supporting data collection and data search and discovery: it exploits a checkpoint browser and a checkpoint dashboard, which presents indicators automatically produced from information database. SeaDataNet (SeaDataNet, 2017b) represents a distributed marine data management infrastructure able to manage different large datasets related to in situ and remote observation of the seas and oceans. Through a distributed network approach, it provides an integrated overview and access to datasets provided by 90 national oceanographic and marine data centers. Finally, another efficient solution has been developed within the project CLIPC (CLIPC, 2017). It provides access to climate information including data from satellite and in situ observations, transformed data products and climate change impact indicators. Moreover, users can exploit the CLIPC toolbox to generate, compare, manipulate and combine indicators, create an user basket or launch new jobs for indices calculation.

These systems support a wide range of functionalities such as data access, classification, search and discovery and downloading by means of a distributed architecture. Graphs and maps production are also supported and some of these propose tools to manipulate and combine datasets. The proposed architecture of the TESSA data platform should offer su-

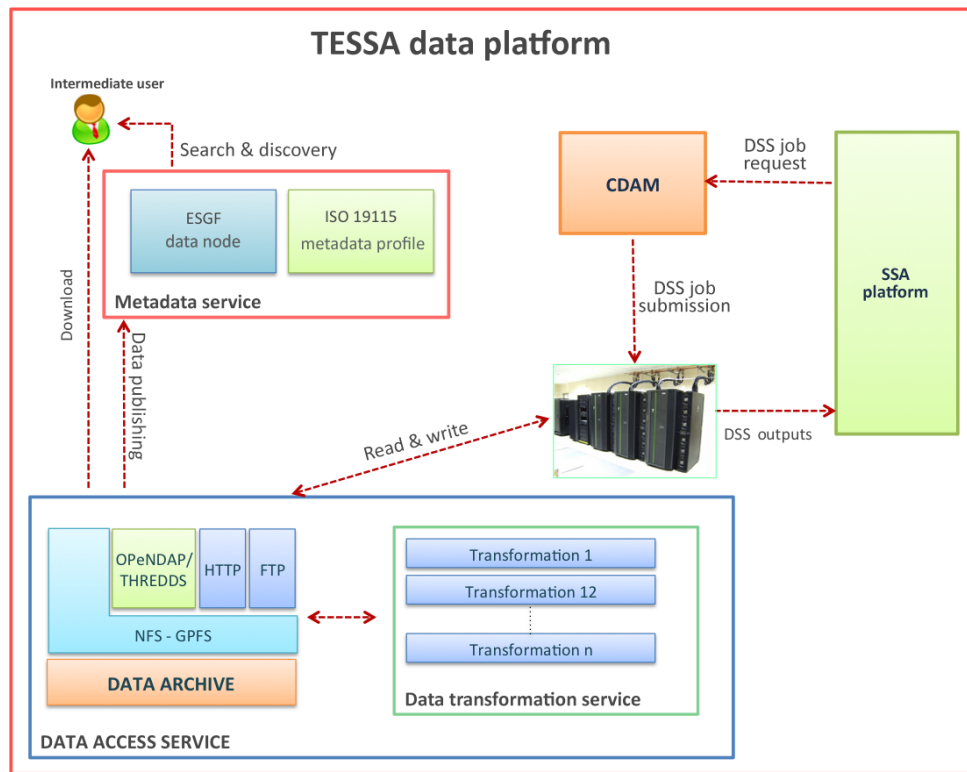


Figure 1. The TESSA data platform.

rior and fast data management functionalities, producing datasets suitable for serving different kinds of services in near-real time: access, search and discovery are just some of the features offered. The developed system, in fact, provides a common interface for submitting complex algorithms in a HPC environment using standard approaches and formats exploiting a unified infrastructure for supporting different scenarios and applications.

3 The data platform architecture

As reported in Fig. 1, the TESSA data platform is composed of three main components: the Data Access Service (DAS), the Metadata Service and the Complex Data Analysis Module (CDAM).

The DAS allows the access to the data archive, the storage of the data produced in the frame of TESSA project. The main consumers of these services are the “intermediate users” (the scientific community) and the automatic procedures fed by produced datasets to perform the needed transformations. In this context, the DAS represents the single interface to access the data archive of TESSA: its design has been driven by the aim to satisfy requirements such as transparency, robustness, efficiency, fault tolerance, security, data delivery, subsetting and browsing. This implies that this component must necessarily be “data-centric” and “multi-

service” in order to offer access to the data, often stored in a single instance on the data archive, through different protocols and, therefore, different levels of functionality.

In addition, the need of interoperability with other data management platforms has driven the design toward the inclusion of well-known services:

- “domain-oriented”, able to provide specific functionalities for the climate change domain;
- “general purpose”, to satisfy general requirements exploiting common services and protocols.

The Metadata Service has been designed as the heart of the information system of the TESSA products, to support the activities of search and discovery and metadata management. Moreover, the Metadata Service addresses the requirements of interoperability, considered a key feature in order to provide to any potential user an easier and unified access to the different types of data produced by the project. These requirements have been satisfied through the adoption of a metadata profile compliant to the ISO standards for geospatial data (ISO, 2009) and the implementation of an efficient information system relying on a specific ESGF data node.

Finally, the CDAM supports the SSA high-level services by performing on demand complex analysis on the datasets stored in the data archive. In particular, a number of intelligent Decisions Support Systems for complex dynamical en-

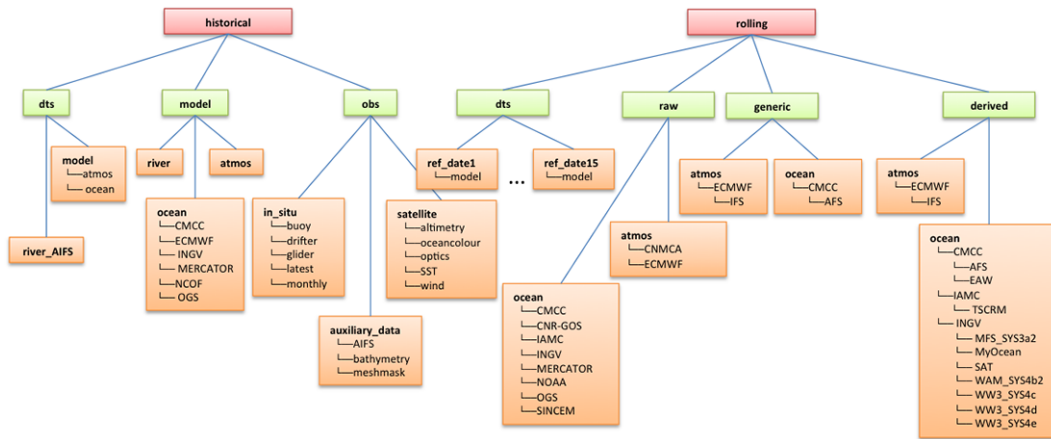


Figure 2. High-level representation of the data archive.

vironments have been designed and developed, exploiting high performance computing functionalities to execute their algorithms. One of the goals is to deploy an infrastructure where a DSS requires the execution of a complex task including (i) accessing data stored in the archive, (ii) executing one or more preliminary tasks for processing them and (iii) storing the results in a specific section of the data archive or publishing them on other sites. Each DSS request interacts with the platform involving a three-layer infrastructure: the user client (web or mobile platform), the SSA platform (Scalas et al., 2016) and the CDAM. Each layer communicates with their linked tiers exchanging messages using well-known protocols and it has been developed in order to implement an high separation of concerns.

The design and the main features of these modules will be detailed in the following sections.

4 The Data Access Service

As stated in the previous section, the DAS represents the single interface to access the data archive of TESSA; specifically, it has been designed with the aim to set up a multi-service platform to access and analyze scientific data. Among the exposed services it is important to mention the THREDDS Data Server (TDS) (THREDDS, 2017), which represents a web server that provides metadata and data access for scientific datasets, using a variety of remote data access protocols. More in detail, it offers a set of services that allows users a direct access to data produced in the context of the TESSA project namely the output of the AFS (Adriatic Forecasting System) (Guarnieri et al., 2008; Oddo et al., 2005, 2006), AIFS (Madec, 2008; Oddo et al., 2009) and SANIFS (Federico et al., 2017) sub-regional models. In particular, the THREDDS includes the WMS – Web Map Service (WMS, 2006) – for the dynamic generation of maps starting from geographically referenced data, the direct downloading using the HTTP protocol. Furthermore, it in-

cludes the OPeNDAP (OPeNDAP, 2017) service, which is a framework that aims to simplify the sharing of scientific information on the web by making available local data from remote connections, providing also variables selection and spatial and temporal subsetting capabilities. To improve performance in terms of load balancing and increase the fault tolerance of the system, a multiplexed configuration for the THREDDS installation based on two different hosts has been used. In addition, a sub-component, the Data Transformation Service (DTS), represents a middleware responsible for applying different data transformations in order to make data compliant to the different protocols offered by the DAS and therefore available to the data consumers (users or services) in a transparent way.

4.1 The dataset repository: the TESSA data archive

The data archive represents the storage for all the data produced within the TESSA project. It has been designed and implemented in order to ease the data search and discovery phase exploiting their classification while, from an infrastructure point of view, data are stored on a GPFS (Schmuck and Haskin, 2002) file system (in RAID 6 configuration) able to provide fast data access and a high level of fault tolerance. Figure 2 shows an high-level schema of the archive.

The data can be subdivided into two categories:

- historical data, which are permanently stored into the archive, including observations and model reanalysis and best analysis;
- rolling data, which are subjected to a rolling procedure, including model worst analysis, simulations and forecasts.

Considering the different kinds of data, the data archive has been created following a tree-based topology where the directory names reflect the data classification (historical as first level, model or observation as second level, atmospheric or

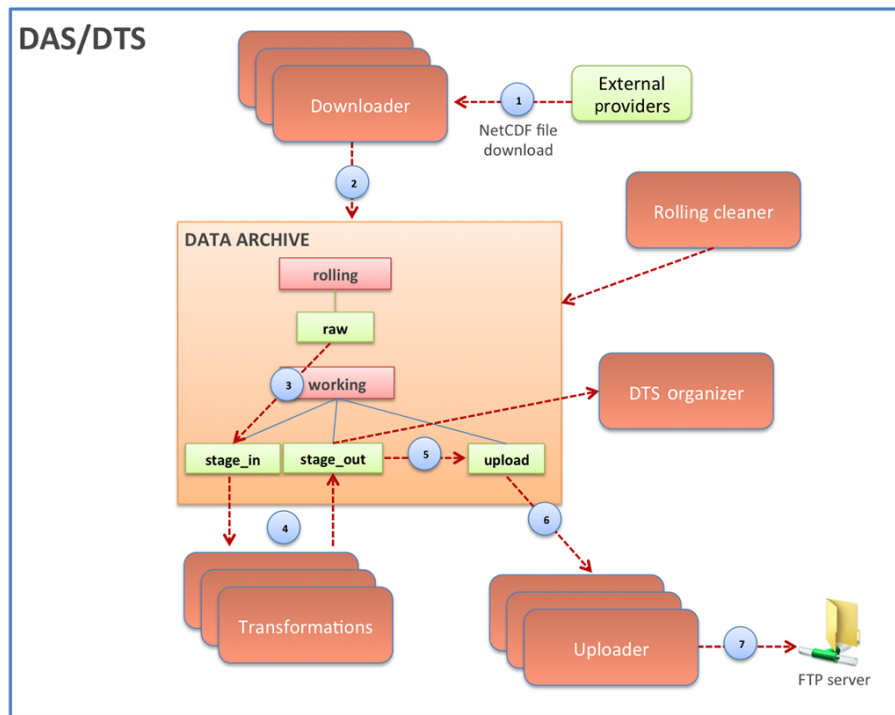


Figure 3. Interactions between the DAS, the DTS and the data archive.

oceanographic as third level, etc.). The directory tree includes also the institute providing the data, the name of the system that produced the data and the temporal resolution of the data. This kind of classification eases and speeds up the search and discovery of data by intermediate/scientific users and automatic procedures. Concerning the “historical” archive, data stored under “obs” include observations retrieved by different instruments (satellites, in situ buoys, vessels, etc.) and consider measurements of the main oceanographic variables (temperature, altimetry, salinity, chlorophyll concentration, etc). The spatial and the temporal resolution and the temporal coverage are different depending on the data type.

Data stored under “model” directory include the best data produced by the numerical models (atmospheric or oceanic) by different external providers. Typically, these data are produced assimilating observational data in order to improve the algorithm functionalities which drive the models. Spatial, temporal resolution and temporal coverage depend on the model considered.

The data stored under “rolling” section exist no longer than 15 days, meaning that they are permanently removed from the archive after that time period since they are superseded by better quality data, e.g., updated forecasts. The “rolling” archive includes also data subjected to further processing operations saved under the proper directories: the processing could include an interpolation to a finer grid in order to increase the spatial resolution (“derived” data), a

subsampling by time or by space (“generic” data) or a transformation in a new format for specific purposes (“dts” data). The access to the data archive is allowed only via the Data Access Service described below.

4.2 Implementation of the DAS-DTS modules

Figure 3 shows a high-level representation of the DAS and its strict interaction with the DTS and the data archive, as implemented in the framework of TESSA project. It is composed by totally automatized cyclic workflows that execute a set of steps on a set of data.

This interchange system has been implemented through some high-parameterized software components which schedule and parallelize the different jobs on an high performance computing environment. The software modules developed for this system are the Downloader module, the DTS_organizer module and the Uploader module. Schematically, these modules download source raw data from a set of external providers and, after a set of transformations, taking into account the appropriate data format, store them in the data archive and publish them through the set of services exposed. Finally the Rolling_cleaner module is used to delete obsolete files and directories in the rolling archive. With the aim to increase the performance and the robustness of the system, all the mentioned components work in a multi-threaded environment: each thread is responsible for managing distinct sets of files. To cope with erroneous input, a MD5 checksum is also used. It is worth noting that the whole pro-

cess is performed in a transparent way with respect to the services that act as consumers of the produced datasets in order to provide data always updated to the latest versions.

5 The Metadata Service and the metadata profile

The analysis of geospatial metadata standards has been the starting point for the definition of a specific metadata profile suitable for the description of the datasets produced in the frame of TESSA project. ISO 19115, “geographic information – metadata”, is a standard of the International Organization for Standardization. It is a constituent of the ISO series 19100 standards for spatial metadata and defines the schema required for describing geographic information. In particular, this standard provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference and distribution of digital geographic data. ISO 19115 is applicable to the cataloguing of datasets and the full description of datasets, geographic datasets, dataset series and individual features and feature properties.

The ISO 19115 standard provides approximately 400 metadata elements, most of which are listed as “optional”. However, it has also drawn up a “core metadata set for geographic dataset” for search, access, transfer and exchange of geographic data; metadata profiles compliant to this international standard shall include this core to ensure homogeneity and interoperability. All metadata profiles based on ISO 19115 should include this minimum set for conformance with this international standard.

The ISO 19115 metadata standard has been widely adopted by a large number of international oceanographic and marine projects, such as the International Hydrographic Organization (IHO, 2017) and SeaDataNet. For its characteristics, it has been considered a valid solution also for TESSA project, as capable of satisfying requirements of simplicity and interoperability. Thus, a specific metadata profile has been designed for describing and cataloging TESSA products, according to the requirements of the international metadata standard ISO 19115 and its XML implementation ISO 19139.

A detailed analysis of TESSA datasets led to the definition of the key aspects about how, when and by whom a specific set of data was collected, how the data are accessed and managed and which data formats are employed in order to obtain a general overview of data and metadata available. On the basis of this information, a minimum set of metadata has been defined able to describe the datasets produced in the frame of TESSA project. This schema includes all mandatory elements of the core metadata set for geographic datasets drawn up by the ISO 19115 standard for search, access, transfer and exchange of geographic data: dataset title, dataset reference date, abstract describing the dataset, dataset language, dataset topic category, metadata point of contact, metadata date stamp and online resource.

The ISO 19115 standard for metadata is being adopted internationally as it provides a built-in mechanism to develop a “community profile” by extending the minimum set with additional metadata elements to suit the needs of a particular group of users.

Hence the generation of the specific TESSA metadata profile composed of the minimum set of metadata required to serve the full range of metadata applications (discovery and access) and other three optional elements that allow for a more extensive description and, at the same time, highlight the main features of TESSA products.

Another important step in the design of TESSA Metadata Service has been the analysis of the information systems most widely adopted, at international level, to support the activities of search and discovery.

In particular, the ESGF has been recognized and selected as the leading infrastructure for the management and access of large distributed data volumes for climate change research. It consists of a system of distributed and federated data nodes that interact dynamically through a peer-to-peer paradigm to handle climate science data, with multiple petabytes of disparate climate datasets from a lot of modeling efforts, such as CMIP5 and CORDEX. Internally, each ESGF node is composed of a set of services and applications for secure data/metadata publication and access and user management. Moreover, the search service offers an intuitive tool based on search categories or “facets” (e.g., “model” or “project”), through which users are enabled to constrain their search results.

For these features, ESGF has been selected as the most appropriate information system for TESSA data dissemination and, above all, its visibility in the climate community.

5.1 Implementation of the Metadata Service

In this phase, the metadata profiles for the description of datasets produced by TESSA models have been formalized; moreover, the installation of a dedicated ESGF data node has been performed, customized on the requirements and specificities of TESSA project.

For its characteristics of simplicity and interoperability, the GeoNetwork metadata editor (GeoNetwork, 2015) has been chosen as metadata editing tool for the description of datasets produced by TESSA models. In fact, GeoNetwork opensource is a multi-platform metadata catalogue application, designed to connect scientific communities and their data through the web environment. This software provides an easy to use web interface to search geospatial data across multiple catalogs, to publish geospatial data using the online metadata editing tools, to manage user and group accounts and to schedule metadata harvesting from other catalogs. In particular, GeoNetwork provides a powerful metadata editing tool that supports the ISO 19115 metadata profile and provides automatic metadata editing and management. Profiles can take the form of templates that are possible to fill in

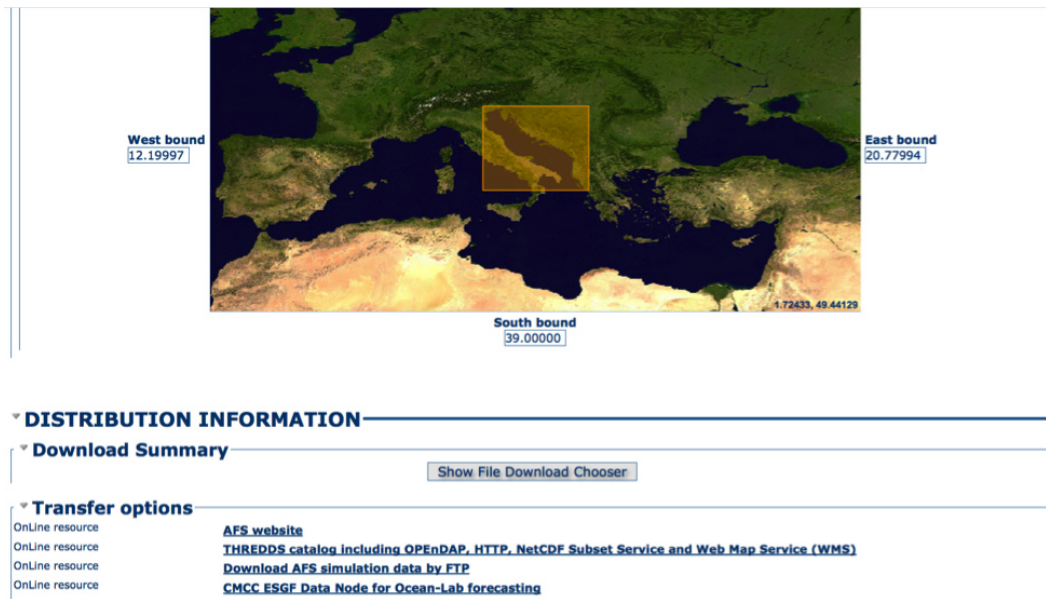


Figure 4. Metadata profile for an example of AFS dataset.

by using the metadata editor. Once the metadata profile has been compiled, it is also possible to generate the related XML schema compliant with ISO 19139, to link data to related metadata description and to set the privileges to view metadata and to download the data.

As an example, Fig. 4 reports an extract of the metadata profile compiled for the description of AFS simulation dataset.

Thus, this metadata profile represents a very powerful but, at the same time, user-friendly tool for the activities of search and discovery of the data produced in the frame of TESSA project. As stated before, the TESSA information system also includes a dedicated ESGF data node, installed and properly configured for the project requirements. First of all, the web interface has been customized by inserting the description of the project and other related information. However, the most important step has been the mapping of TESSA data properties onto the “search categories or facets”, through which users can navigate the archive. In particular, new facets such as “spatial_resolution” and “spatial_domain” have been defined in order to better fit the specificity of TESSA datasets. These new facets have been added to configuration files and so they are now available in the web interface of TESSA data node to support data search. In parallel, specific global attributes have been inserted in all output files of TESSA models in order to provide the metadata values for the corresponding search facets. These features are schematically represented in Fig. 5.

6 The Complex Data Analysis Module

The CDAM component has been designed with the aim to optimize the DSS job submissions in a multi-channel (web or mobile) environment. This component exploits existing technologies, widely spread in client–server architectures, in order to simplify the complexity of the overall system, increasing the flexibility of the involved modules and the separation of concerns.

Figure 6 shows an overview of the workflow related to TESSA DSS and the components involved in the operational chain.

The actors of such a scenario are (i) the user, interacting with the system using his mobile device or his computer, (ii) the SSA platform and (iii) the CDAM, which exploits a cluster infrastructure for the job execution.

The user sends a request to the SSA platform by using a desktop or mobile application related to a specific DSS. The request contains a set of parameters that will be parsed by the SSA platform and sent to the CDAM through a secure communication channel. Once received the request, the CDAM gateway is responsible for preparing the environment for the job submission and for sending the parameter to the underlying cluster infrastructure, which executes the DSS algorithm and, at the end, sends a message to the SSA platform with the result of the processing.

6.1 The CDAM gateway

The CDAM gateway is designated to be the entry point for the job submission. It is responsible for managing the incoming requests on the basis of the different DSS and for interacting with the underlying cluster infrastructure for the

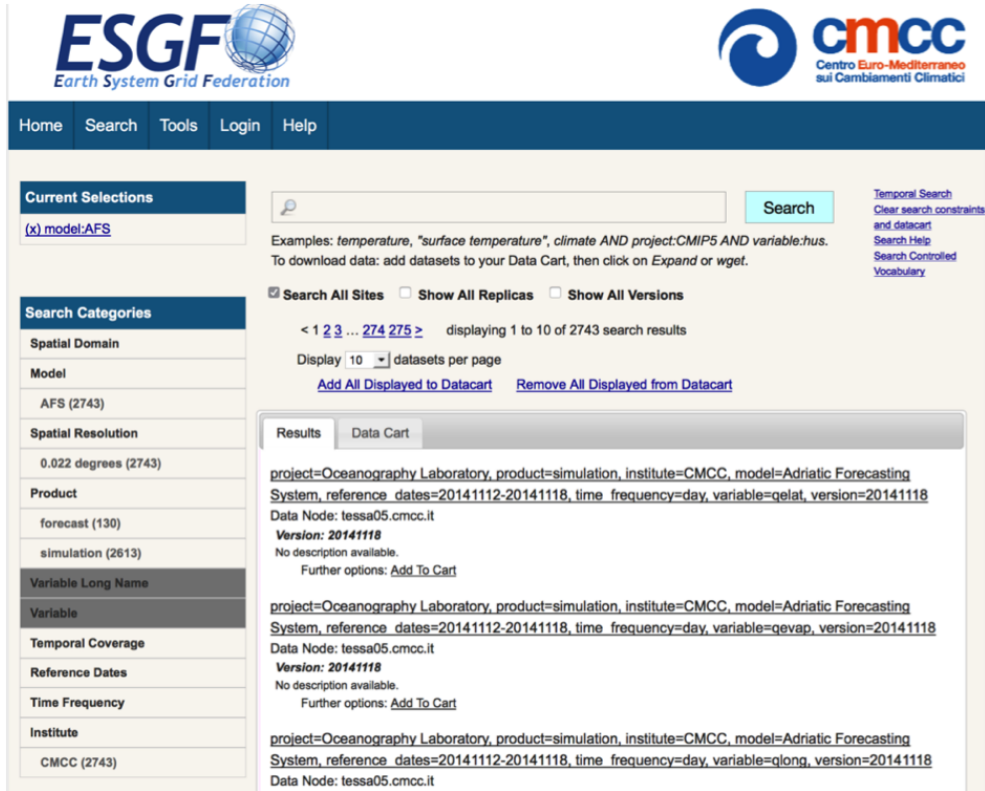


Figure 5. List of facets for AFS products as published in the ESGF data node for TESSA.

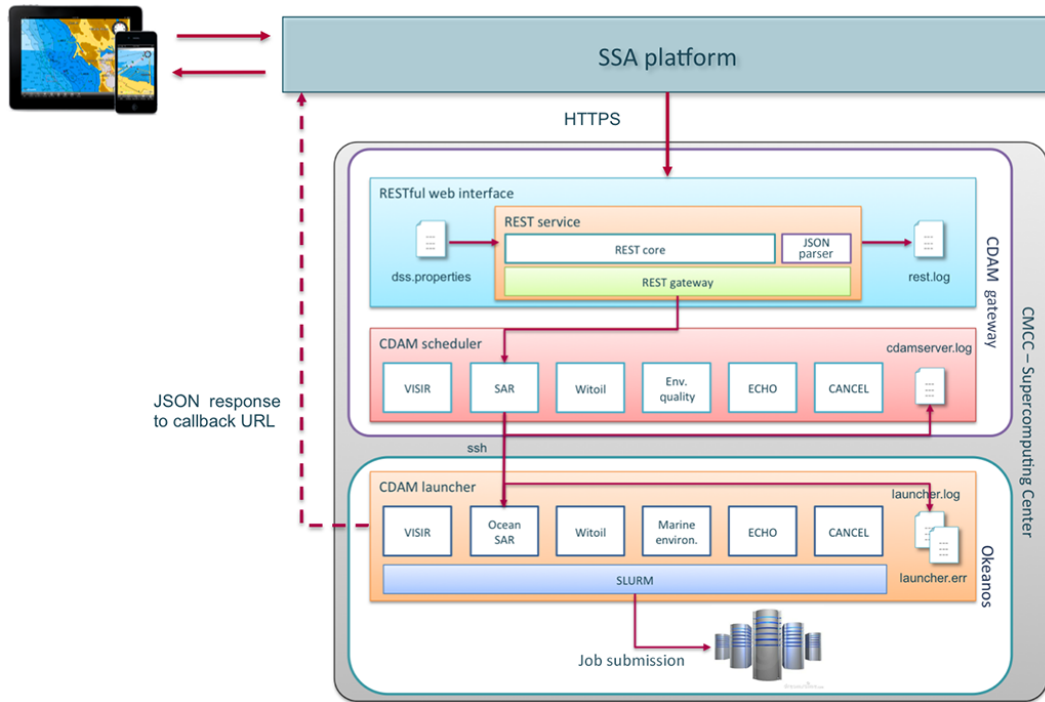


Figure 6. DSS request submission and interactions with the CDAM.

algorithm execution. The CDAM gateway module is constituted by two components, the web interface and the CDAM scheduler.

The web interface provides the SSA platform with a unique and uniform interface since every request is served in the same way. In order to secure the communication channel between the SSA platform and the CDAM gateway, two levels of security constraints have to be satisfied. The former concerns the authentication/authorization method based on the basic authentication mechanism; the sender of the request (in this case the SSA platform) needs to authenticate itself by providing the proper username and password. Moreover, the transfer of such information is protected by the HTTP protocol with SSL encryption (HTTPS). The latter regards a specific OS firewall policy which enables incoming communications from well-known IP range.

The CDAM scheduler is composed of a series of modules, one for each DSS or service that has been implemented. In particular, three types of services has been defined: submission of a job related to a specific DSS, deleting of an active job and an echo service to check if the job submission service is available and to evaluate the response time.

The main tasks of each DSS specific module are

1. to check the parameters received from the web interface;
2. to perform the setup of the environment on the cluster infrastructure, by creating the directories required for the algorithm execution;
3. to contact the execution host with the correct parameters.

Both web interface and CDAM scheduler modules are based on a logging mechanism for keep track of all the operations performed and the errors occurred during the process.

6.2 The CDAM launcher

The CDAM launcher is hosted on the entry point of the cluster infrastructure and has the responsibility to properly manage the DSS submission on the cluster.

As the CDAM scheduler, the launcher is composed of one module for each DSS or service and performs the following operations: it checks the parameters received from the scheduler, it prepares the files that will be used as input for the algorithm execution, it performs the job submission and, finally, it sends the result of the processing to the SSA platform.

It is worth noting that the design and implementation of the CDAM stack is independent from the software modules designated to start the submission; indeed, it provides general interfaces exposing a remote submission service able to hiding the HPC resources for concurrent execution of several jobs. In TESSA, such a task is performed by the SSA platform which manages the incoming requests from the users

and interacts with the CDAM for the DSS algorithms execution.

7 Operational activity: TESSA data platform use cases

The TESSA data platform developed within the TESSA project represents an innovative solution not only as it provides an efficient access to the data but also, above all, as it exploits and integrates different technologies to fully satisfy any potential user's need at operational level. In fact, the three main components of the platform (DAS, Metadata Service and CDAM) are not isolated blocks but interact each other supporting, with a high efficiency level, a number of services such as the SeaConditions operational chain (Sea-DataNet, 2017a), the automatic publication of datasets on the data nodes ESGF and the SSA DSS.

7.1 Automatic publishing procedure on the ESGF data node

An automatic procedure for data publishing on the ESGF data node has been designed and implemented in order to publish updated products on the TESSA ESGF portal.

This procedure is based on the strict interaction between the DTS and the Metadata Service. As first step, DTS applies a pre-processing on the input data that are mainly output of the hydrodynamic sub-regional models and, specifically, daily simulations and forecasts for the Adriatic, Ionian and Tyrrhenian seas with different spatial and temporal resolution. This step is fundamental to make data compliant for the publication on the ESGF data node and involves the insertion of specific global attributes to provide the metadata values for the corresponding search facets. Once the pre-processing of the input data has been performed, the DTS modules alert the DAS component to publish the involved datasets on the ESGF node.

Figure 7 shows the mentioned procedure in terms of a workflow of operational tasks: the delivery of TESSA outputs (simulations and forecasts, hourly and daily) from the production host to the target host, the automatic pre-processing of datasets by DTS operations and the upload of the processed data to the target directory.

On the Metadata Service side, a controller module performs the publication of the different types of data pre-processed by DTS; specifically, the controller is a daemonized module that once a day sets up the directories *map-files/bulletindate* and runs in background the procedures for the publication of the outputs of the models.

7.2 DTS automatic procedures for DSS and numerical models transformations

The DTS is also responsible for pre-processing the input data for the different DSS, a set of tools for supporting users in making decisions or managing emergency situations and sug-

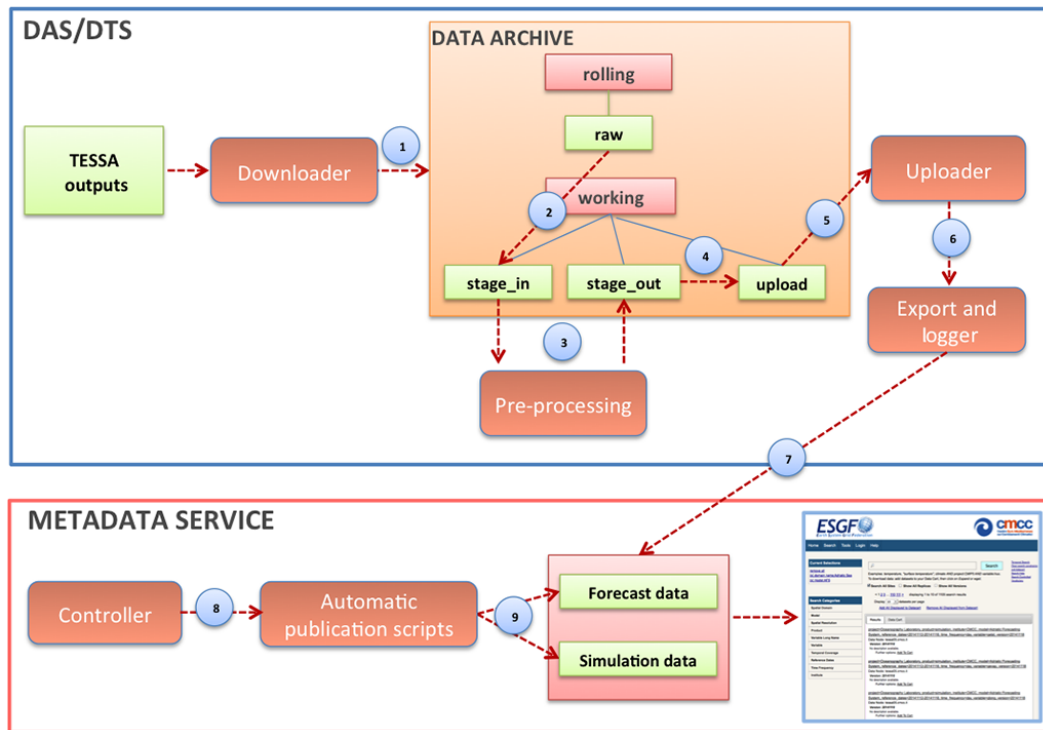


Figure 7. Operational chain for ESGF data publication.

gesting a strategy of intervention, namely WITOIL, VISIR and EarlyWarning. Specifically, WITOIL predicts the transport and transformation of real or hypothetical oil spills in the Mediterranean Sea, VISIR suggests the best nautical routes in any weather–marine condition and EarlyWarning manages alert in case of extreme events by providing a daily update on weather (wind, air temperature) and oceanographic conditions (wave height, sea level, water temperature). In detail, the raw data, downloaded using the DAS routines as described in Sect. 3, are supplied to the automated pre-processing chains of the DTS, responsible for applying a series of transformations on the data in order to suit the specific DSS input requirements. At the end of the transformation process, the system is able to manage the different on-demand DSS requests starting the related algorithms.

As an example, WITOIL requires files of temperature, currents and wind at sea surface; in particular, the marine raw data are subjected to a spatial subsampling selecting only the first 10 levels of depth. For the computation of the least-time nautical route, performed by VISIR, every day a single dataset of 10-day forecasts is provided; in this particular case, the DTS concatenates significant wave height, peak wave period and wave direction fields. For EarlyWarning, the pre-processing procedure applied to the files is very simple and consists of a data unzipping to create a data block useful to the execution of the algorithm. The DTS is responsible for applying all the mentioned processing on the input data in order to let the different DSS operational chains faster

and more efficient exploiting the automation provided by the DAS. Figures 8 and 9 show a nautical route computed by VISIR and a sea level forecast for the city of Venice computed by the EarlyWarning DSS.

The DAS is also extensively used by the DTS automatic processes for activating the sub-regional models production (i.e., AFS, AIFS and SANIFS). The DAS modules are responsible for periodically downloading the source raw data from the external data provider. Once input data are available, the operational forecasting system is activated by specific DTS modules in order to schedule and manage the sub-regional models execution and to properly synchronize the pre- and post-processing operational chains.

The interaction between the DAS and DTS is also important for the automated production of maps created starting from the outputs of the sub-regional model AFS, AIFS and SANIFS by means of dedicated operational chains.

7.3 SeaConditions operational chain

Concerning the SeaConditions operational chain, the main aim of the DTS is to create the properly processed and formatted data to visualize 5 days of weather–oceanographic forecasts with 3h of temporal resolution for the first 2.5 days and 6 h for the others.

Dedicated DAS modules, day by day, download the source raw data from the external providers storing them into the data archive. Raw data are then provided to the SeaCondi-

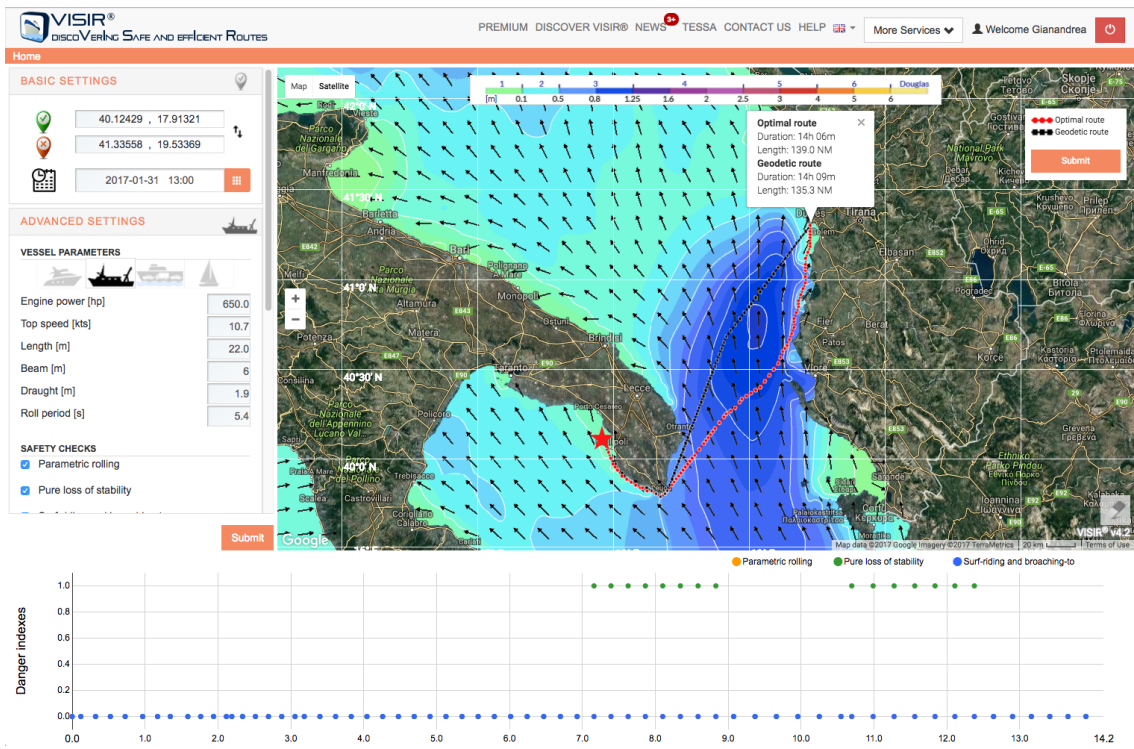


Figure 8. A motorboat route computed by VISIR. The linechart below the map shows the safety constraints that are violated by the geodetic route (black dots on the map) but not by the optimal route (red dots).

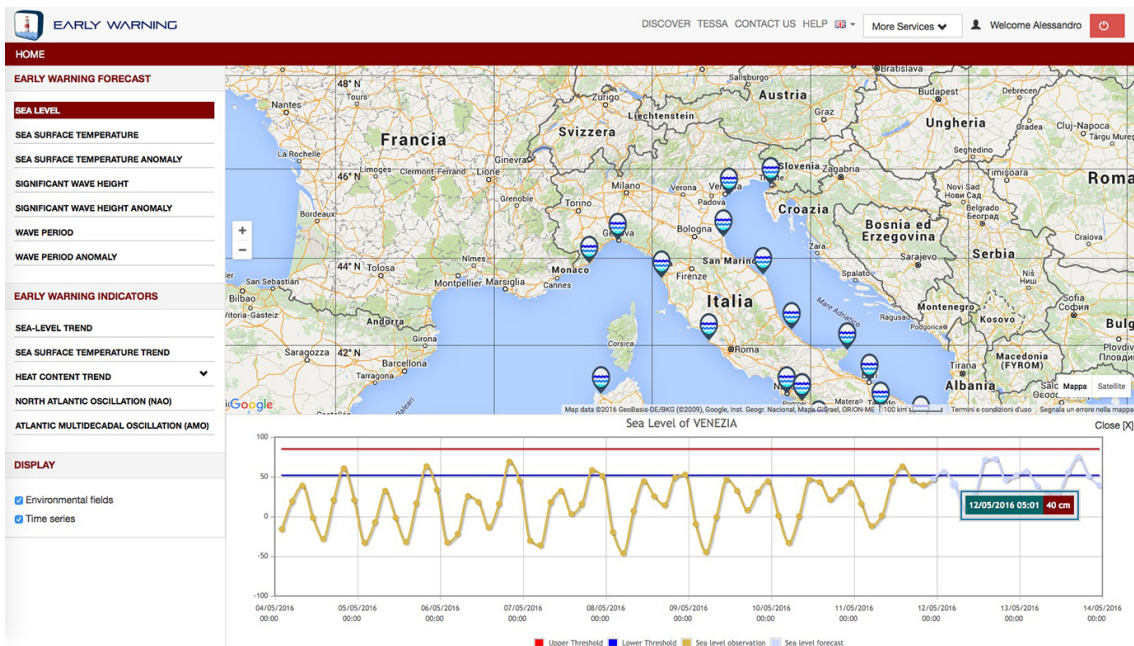


Figure 9. EarlyWarning forecast for the sea level of Venice.

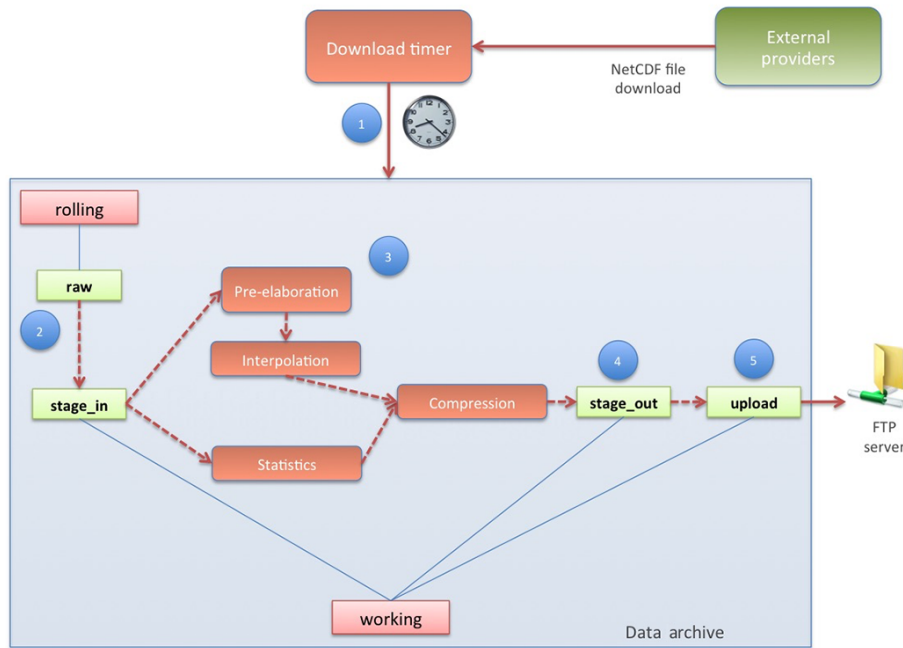


Figure 10. Operational chain for SeaConditions data preparation.

tions operational chains managed by DTS; at the end post-processed outputs are moved to an FTP server. SSA platform procedures are responsible for downloading them as input for the creation of the maps published on the SeaConditions web portal.

Dedicated DAS modules, day by day, download the source raw data from the external providers storing them into the data archive. Raw data are then provided to the SeaConditions operational chains managed by DTS; at the end post-processed outputs are moved to an FTP server.

This operational chain is based on a workflow manager implemented by a daemonized routine called DTS_SCHEDULER. This daemon manages the parallelization of the data processing tasks using stage-in and stage-out folders as interchange layers with the data archive. The different DTS production phases are represented in Fig. 10, including

- pre-elaboration phase, which is involved in the data splitting by time and by variables;
- statistic phase, which aims to compute some statistics as minimum, maximum, mean and standard deviation;
- interpolation phase, in which the split data resulting from the first phase are filtered by time and some are interpolated in order to increase their spatial resolution;
- compression phase, in which the data resulting from the interpolation phase are compressed: Lempel–Ziv deflation, NetCDF4 packing algorithm (from the typical NC FLOAT to NC SHORT reducing the size by about 50 %)

and file zipping with gzip are operations that generally lead to a 70–90 % reduction in file size. If the original NetCDF file occupies 1 GB, at the end of packing and zipping process it will occupy approximately 100 MB.

8 Conclusions

One of the main objectives of the TESSA project was to develop a set of operational oceanographic services for the SSA. To reach this goal, advanced technological and software solutions are the backbone for the real time dissemination of the oceanographic data to a wide range of users in the maritime sector, tourism and scientific community. In this framework, the relevance of the TESSA data platform is apparent through its capacity to provide efficient and secure data access and strong support to high-level services, such as the DSS.

As stated in the introduction, the employment of well-known standards concerning protocols, services and output formats satisfies the FAIR guiding principles for scientific data management and stewardship. Specifically the designed data platform relies on THREDDS for the publication of the TESSA final products (findable and accessible), while the OPeNDAP service exposes a standard DAP interface (interoperable). In addition, the ISO 19115 standard has been adopted for the definition of a metadata profile containing the mandatory elements of the core metadata set for geographic datasets (interoperable and reusable) and the search and discovery of data are supported by the ESGF data node and by the GeoNetwork information system (findable and accessi-

ble). Finally, access to the DSS methods is provided by a machine readable interface exploiting the JSON format (accessible).

In general, the main components of this platform have been designed and developed by taking into account and satisfying a large number of additional requirements such as transparency, robustness, efficiency, fault tolerance and security. Moreover, it is important to emphasize that this platform represents a unique and innovative example of how different components interact with each other to support operational services for the SSA, such as SeaConditions and the other DSS (VISIR, WITOIL, Ocean-SAR, EarlyWarning).

Therefore, these features make the TESSA data platform a valid prototype easily adopted to provide an efficient dissemination of maritime data and a consolidation of the management of operational oceanographic activities.

9 Data availability

Data of subregional models are periodically produced, exploiting the data platform operational chains and published at <http://tds.tessa.cmcc.it>.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This research was funded by the project TESSA (Technologies for the Situational Sea Awareness; PON01_02823/2) funded by the Italian Ministry for Education and Research (MIUR).

Edited by: P. Marra

Reviewed by: two anonymous referees

References

Bahurel, P., Adragna, F., Bell, M. J., Jacq, F., Johannessen, J. A., Le Traon, P., Pinardi, N., and She, J.: Ocean Monitoring and Forecasting Core Services, the European MyOcean Example, in: Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society (Vol. 1), Venice, Italy, 21–25 September 2009, edited by: Hall, J., Harrison, D. E., and Stammer, D., ESA Publication WPP-306, doi:10.5270/OceanObs09.pp.02, 2009.

Cinquini, L., Crichton, D., Mattmann, C., Harney, J., Shipman, G., Wang, F., Ananthakrishnan, R., Miller, N., Denvil, S., Morgan, M., Pobre, Z., Bell, G. M., Doutriaux, C., Drach, R., Williams, D., Kershaw, P., Pascoe, S., Gonzalez, E., Fiore, S., and Schweitzer, R.: The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data, *Future Generation Computer Systems*, ISSN 0167-739X, doi:10.1016/j.future.2013.07.002, 2014.

CLIPC: CLimate Information Portal for Copernicus, <http://www.clipc.eu/home>, last access: 30 January 2017.

Copernicus: Copernicus web-page, available at: <http://www.copernicus.eu/>, last access: 30 January 2017.

Coppini, G., Jansen, E., Turrise, G., Creti, S., Shchekinova, E. Y., Pinardi, N., Lecci, R., Carluccio, I., Kumkar, Y. V., D'Anca, A., Mannarini, G., Martinelli, S., Marra, P., Capodiferro, T., and Gismondi, T.: A new search-and-rescue service in the Mediterranean Sea: a demonstration of the operational capability and an evaluation of its performance using real case scenarios, *Nat. Hazards Earth Syst. Sci.*, 16, 2713–2727, doi:10.5194/nhess-16-2713-2016, 2016.

De Dominicis, M., Pinardi, N., Zodiatis, G., and Archetti, R.: MEDSLIK-II, a Lagrangian marine surface oil spill model for short-term forecasting – Part 2: Numerical simulations and validations, *Geosci. Model Dev.*, 6, 1871–1888, doi:10.5194/gmd-6-1871-2013, 2013a.

De Dominicis, M., Pinardi, N., Zodiatis, G., and Lardner, R.: MEDSLIK-II, a Lagrangian marine surface oil spill model for short-term forecasting – Part 1: Theory, *Geosci. Model Dev.*, 6, 1851–1869, doi:10.5194/gmd-6-1851-2013, 2013b.

EMODNET: EMODnet MedSea Checkpoint, available at: <http://www.emodnet-mediterranean.eu>, last access: 30 January 2017.

ESGF: Earth System Grid Federation, available at: <http://esgf.llnl.gov>, last access: 30 January 2017.

Federico, I., Pinardi, N., Coppini, G., Oddo, P., Lecci, R., and Mossa, M.: Coastal ocean forecasting with an unstructured grid model in the southern Adriatic and northern Ionian seas, *Nat. Hazards Earth Syst. Sci.*, 17, 45–59, doi:10.5194/nhess-17-45-2017, 2017.

GeoNetwork: GeoNetwork opensource, available at: <http://geonetwork-opensource.org> (last access: 30 January 2017), 2015.

Guarnieri, A., Oddo, P., Pastore, M., and Pinardi, N.: The Adriatic Basin Forecasting System new model and system development, in: Coastal to Global Operational Oceanography: Achievements and Challenges, edited by: Dahlin, H., Fleming, N. C., and Petersson, S. E., Proceedings of 5th EuroGOOS Conference, Exeter, 2008.

IHO: International Hydrographic Organization, available at: <http://www.iho.int/srv1>, last access: 30 January 2017.

INSPIRE-directive: Inspire directive web page, available at: <http://eur-lex.europa.eu/JOHtml.do?uri=OJ:L:2007:108:SOM:EN:HTML> (last access: 30 January 2017), 2007.

INSPIRE: Infrastructure for spatial information in Europe, available at: <http://inspire.ec.europa.eu>, last access: 30 January 2017.

ISO 19115:2003 Geographic information – Metadata, Tech. rep., International Organization for Standardization (TC 211), 2003a.

ISO 19139:2007 Geographic information – Metadata – XML Schema Implementation, Tech. rep., International Organization for Standardization (TC 211), 2003b.

ISO: ISO Standards Guide, available at: http://www.isotc211.org/Outreach/ISO_TC_211_Standards_Guide.pdf (last access: 30 January 2017), 2009.

Lecci, R., Coppini, G., Creti, S., Turrise, G., D'Anca, A., Palazzo, C., Aloisio, G., Fiore, S., Bonaduce, A., Mannarini, G., Kumkar, Y., Ciliberti, S., Federico, I., Agostini, P., Bonarelli, R., Martinelli, S., Marra, P., Scalas, M., Tedesco, L., Rollo, D., Cavallo, A., Tumolo, A., Monacizzo, T., Spagnulo, M., Pinardi, N., Fazioli, L., Olita, A., Cucco, A., Sorgente, R., Tonani, M., and Drudi, M.: SeaConditions: present and future sea con-

- ditions for safer navigation, OCEANS 2015, Genova, Italy, doi:10.1109/OCEANS-Genova.2015.7271764, 2015.
- Madec, G.: NEMO Ocean General Circulation Model Reference Manual, Internal Report, LODYC/IPSL, 2008.
- Mannarini, G., Pinardi, N., Coppini, G., Oddo, P., and Iafrazi, A.: VISIR-I: small vessels – least-time nautical routes using wave forecasts, *Geosci. Model Dev.*, 9, 1597–1625, doi:10.5194/gmd-9-1597-2016, 2016a.
- Mannarini, G., Turrisi, G., D'Anca, A., Scalas, M., Pinardi, N., Coppini, G., Palermo, F., Carluccio, I., Scuro, M., Cretì, S., Lecci, R., Nassisi, P., and Tedesco, L.: VISIR: technological infrastructure of an operational service for safe and efficient navigation in the Mediterranean Sea, *Nat. Hazards Earth Syst. Sci.*, 16, 1791–1806, doi:10.5194/nhess-16-1791-2016, 2016b.
- Moussat, E., Pinardi, N., Manzella, G., and Blanc, F.: EMODnet MedSea Checkpoint for sustainable Blue Growth, EGU General Assembly 2016, Vienna Austria, p. 9103, 17–22 April 2016.
- MyOcean: Copernicus Marine Environment Monitoring Service, available at: <http://marine.copernicus.eu>, last access: 30 January 2017.
- NetCDF: OGC NETWORK Common Data Format, available at: <http://www.opengeospatial.org/standards/netcdf>, last access: 30 January 2017a.
- NetCDF: Unidata – NetCDF, available at: <http://www.opengeospatial.org/standards/netcdf>, last access: 30 January 2017b.
- Oddo, P., Pinardi, N., and Zavatarelli, M.: A numerical study of the interannual variability of the Adriatic Sea (2000–2002), *Sci. Total Environ.*, 353, 39–56, doi:10.1016/j.scitotenv.2005.09.061, 2005.
- Oddo, P., Pinardi, N., Zavatarelli, M., and Colucelli, A.: The Adriatic Basin forecasting system, *ACTA ADRIAT.*, 47(suppl), 169–184, 2006.
- Oddo, P., Adani, M., Pinardi, N., Fratianni, C., Tonani, M., and Pettenuzzo, D.: A nested Atlantic-Mediterranean Sea general circulation model for operational forecasting, *Ocean Sci.*, 5, 461–473, doi:10.5194/os-5-461-2009, 2009.
- OPeNDAP: Open-source Project for a Network Data Access Protocol, available at: <https://www.opendap.org/about>, last access: 30 January 2017.
- Scalas, M., Marra, P., Tedesco, L., Quarta, R., Cantoro, E., Tumolo, A., Rollo, D., and Spagnulo, M.: TESSA: design and implementation of a platform for Situational Sea Awareness, *Nat. Hazards Earth Syst. Sci. Discuss.*, doi:10.5194/nhess-2016-166, in review, 2016.
- Schmuck, F. and Haskin, R.: GPFS: A Shared-Disk File System for Large Computing Clusters, in: Proceedings of the 1st USENIX Conference on File and Storage Technologies, FAST '02, USENIX Association, Berkeley, CA, USA, available at: <http://dl.acm.org/citation.cfm?id=1083323.1083349> (last access: 30 January 2017), 2002.
- SeaDataNet: Sea-Conditions web-page, available at: <http://www.sea-conditions.com>, last access: 30 January 2017a.
- SeaDataNet: SeaDataNet web-page, available at: <http://www.seadatanet.org>, last access: 30 January 2017b.
- THREDDS: Unidata THREDDS Data Server, available at: <http://www.unidata.ucar.edu/software/thredds/current/tds/#home>, last access: 30 January 2017.
- WMS: OGC 06-042 – OpenGIS Web Map Server Implementation Specification, Tech. rep., Open Geospatial Consortium Inc., Version 1.3.0, edited by: de la Beaujardiere, J., available at: http://portal.opengeospatial.org/files/?artifact_id=14416 (last access: 8 February 2017), 15 March 2006.
- Wilkinson, M. D.: The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data*, 3, doi:10.1038/sdata.2016.18, 2016.