BioClimate: A Science Gateway for Climate Change and Biodiversity research in the EUBrazilCloudConnect project

Fiore, S.; Elia, D.; Blanquer, I.; Brasileiro, F.V.; Nuzzo, A.; Nassisi, P.; Rufino, I.A.A.; Seijmonsbergen, A.C.; Anders, N.S.; Galvão, C. de O.; de B.L. Cunha, J.E.; Caballer, M.; Sousa-Baena, M.S.; Canhos, V.P.; Aloisio, G.

DOI
10.1016/j.future.2017.11.034

Publication date
2019

Document Version
Final published version

Published in
Future Generation Computer Systems

Future Generation Computer Systems 94 (2019) 895–909

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

BioClimate: A Science Gateway for Climate Change and Biodiversity research in the EUBrazilCloudConnect project

Sandro Fiore [a,*], Donatello Elia [a], Ignacio Blanquer [b], Francisco V. Brasileiro [c], Alessandra Nuzzo [a], Paola Nassisi [a], Iana A.A. Rufino [c], Arie C. Seijmonsbergen [d], Niels S. Anders [d], Carlos de O. Galvão [c], John E. de B.L. Cunha [c], Miguel Caballer [b], Mariane S. Sousa-Baena [e], Vanderlei P. Canhos [e], Giovanni Aloisio [a,f]

[a] Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici, Lecce, Italy
[b] Universitat Politecnica de Valencia, Valencia, Spain
[c] Universidade Federal de Campina Grande, Campina Grande, PB, Brazil
[d] IBED, University of Amsterdam, Amsterdam, Netherlands
[e] Centro de Referência em Informação Ambiental, Campinas, SP, Brazil
[f] University of Salento, Lecce, Italy

HIGHLIGHTS

- Climate and biodiversity systems are closely linked across a wide range of scales.
- User-centric research environment built on top of a federated cloud infrastructure.
- A multi-scale, integrated approach to investigate the climate-biodiversity system.
- The BioClimate Science Gateway represents the high-level interface for scientists.
- BioClimate supports different kinds of scientific data analysis.

ARTICLE INFO

Article history:
Received 14 June 2017
Received in revised form 8 September 2017
Accepted 18 November 2017
Available online 24 December 2017

Keywords:
Science gateways
Scientific data management and analytics
Biodiversity and climate research

ABSTRACT

Climate and biodiversity systems are closely linked across a wide range of scales. To better understand the mutual interaction between climate change and biodiversity there is a strong need for multidisciplinary skills, scientific tools, and access to a large variety of heterogeneous, often distributed, data sources. Related to that, the EUBrazilCloudConnect project provides a user-oriented research environment built on top of a federated cloud infrastructure across Europe and Brazil, to serve key needs in different scientific domains, which is validated through a set of use cases. Among them, the most data-centric one is focused on climate change and biodiversity research. As part of this use case, the BioClimate Science Gateway has been implemented to provide end-users transparent access to (i) a highly integrated user-friendly environment, (ii) a large variety of data sources, and (iii) different analytics & visualization tools to serve a large spectrum of users needs and requirements. This paper presents a complete overview of BioClimate and the related scientific environment, in particular its Science Gateway, delivered to the end-user community at the end of the project.

1. Introduction

Climate and biodiversity systems are closely linked across a wide range of scales. To predict the effects of climate change on the biodiversity system, which is essential towards sustainable landscape and eco-services management, there is a need to further investigate the interaction between the climate system and biodiversity. Presently, researchers and professionals are burdened by scattered data sources, wealth of analysis tools to master and implement, and computational limitations to upscale their analysis.

The "*EU-Brazil Cloud infrastructure Connecting federated resources for Scientific Advancement*" (EUBrazilCloudConnect) [1] is a project funded under the third EU-Brazil coordinated call. It is a preliminary step towards providing a user-oriented environment for scientific research communities to test the execution of

* Corresponding author.
E-mail address: sandro.fiore@cmcc.it (S. Fiore).

https://doi.org/10.1016/j.future.2017.11.034
0167-739X/

challenging applications exploiting a federated cloud infrastructure. The project addresses the scientific challenges of a set of multidisciplinary and highly complementary scenarios. The one on the analysis of biodiversity, natural resources, and climate change is the most challenging from the scientific data management standpoint, and represents the main focus of this paper. The use case involves multiple heterogeneous data sources and several processing pipelines running on top of a trans-Atlantic federated cloud infrastructure between Europe and Brazil seamlessly integrated through the BioClimate Science Gateway. With regard to existing approaches and tools that are mainly client-side/desktop based, the use case delivers an integrated environment for climate change and biodiversity research, accessible through a Science Gateway [2–5], with cloud-based infrastructure and high performance, server-side analytics capabilities. This paper focuses, in particular, on the BioClimate Science Gateway, its main scientific challenges, architectural design and implementation details.

This paper is an extended version of a preliminary work presented at the *8th International Workshop on Science Gateways* (IWSG'16) [6]. In particular, it significantly extends the previous contribution by providing more detailed and technical information about involved data sources, infrastructural and software components, related work, as well as impacts and benefits for the scientific community.

The remainder of this paper is organized as follows. Section 2 provides an overview of the EUBrazilCloudConnect project. Section 3 describes the EUBrazilCloudConnect use case on biodiversity and climate change, and Section 4 discusses the overall use case architecture. Section 5 presents in detail the BioClimate Science Gateway in terms of requirements, capabilities and implementation. Section 6 discusses key related work, whereas Section 7 presents main benefits and impact of the proposed solution. Finally, Section 8 draws the conclusions and highlights the future work.

2. The EUBrazilCloudConnect project

EUBrazilCloudConnect is a research project funded by the European Commission (FP7) and the National Council for Scientific and Technological Development of Brazil (CNPq). EUBrazilCloudConnect proposes a user-oriented approach focusing on efficient computing and storage resources, federated systems, programming frameworks and tools that meet the requirements of the proposed use cases and that can be more widely adopted by the scientific community. EUBrazilCloudConnect focuses on interoperable, standards-based solutions for cloud computing by advancing cloud technologies and federation. The overarching objective of EUBrazilCloudConnect is to drive cooperation between Europe and Brazil by strengthening the scientific and knowledge-based society as key to sustainable and equitable socioeconomic development. The core of this collaboration is defined through three scientific uses cases, which require the collaboration between Brazil and Europe in the provision of data, services, tools, and the needed expertise. The proposed scientific scenarios require access to the project e-infrastructure to run complex workflow pipelines, as well as access to heterogeneous and large datasets for data analysis and visualization.

The three use cases of EUBrazilCloudConnect aim, respectively, at:

1. advancing the molecular analysis and identification of parasites and vectors of Leishmaniasis [7] through the integration of processing pipelines on Brazilian and European biological databases and geo-referenced data;
2. improving the exploitation of high-level heart simulation data through the integration of Alya [8] and ADAN [9] heart simulators, leading up to a multiscale simulator combining electrophysiology and the modeling of the fluid dynamics for the whole cardio-vascular system;
3. advancing the knowledge in the climate change and biodiversity domains through the integration of multiple analysis pipelines and data sources (satellite images, high-resolution Light Detection And Ranging –LiDAR [10], climate records and future scenarios, and species distribution models).

This paper presents the third use case and the related infrastructure that has been set up to properly fulfill all the end-user requirements, with a particular focus and emphasis on the BioClimate Science Gateway.

3. The EUBrazilCloudConnect biodiversity and climate change use case

The climate and biodiversity systems are very complex and closely interlaced across a wide range of spatial and temporal scales. Scientists working in this area have to face key challenges related to, among other things, the integrated analysis of both observed and simulated data, the manipulation and access to large, distributed, and heterogeneous data sources, and the lack of advanced, user-friendly, and integrated tools/environments for scientific data analysis and visualization.

Moreover, direct measurements of climate and biodiversity are often difficult and time-consuming to obtain. In this regard, intensive in situ measurements are feasible only at a few locations, but models and proxies detectable by remote sensing can be used to extrapolate from point locations, where in situ measurements are available, to the regional scale [11]. Then, it has been a common practice to use climate and biodiversity indicators from remote sensing products.

For example, the land use/land cover (LULC) system is frequently used as proxy of biodiversity system changes. These interactions can be studied at various scales, ranging from microscopic scales, and on (genomic, taxonomic, ecosystem) scales of level of individual plant and animal species. These different datasets require different analysis tools and describe only parts of the climate or biodiversity systems. As a result, *a multi-scale and integrated approach is required to investigate the climate-biodiversity system as a whole*. However, currently researchers and professionals are hindered by scattered data sources, wealth of analysis tools they need to master and implement, and computational limitations to upscale their analysis. From the perspective of the end-users, hence, the availability of various data sources jointly with data analysis and processing tools seamlessly integrated into a single platform, is a major step forward in performing scientific research in this area.

To face such critical challenges, a use case on climate change and biodiversity was set up in the EUBrazilCloudConnect project. The use case is mainly data-centric and it aims at better understanding the interactions between the biodiversity system and the climate system (see Fig. 1).

The end-users targeted by the proposed use case include, but are not limited to, the climate change and biodiversity research communities, as well as policy makers and professionals from other fields, including hydrologists, agriculturalists, and environmentalists.

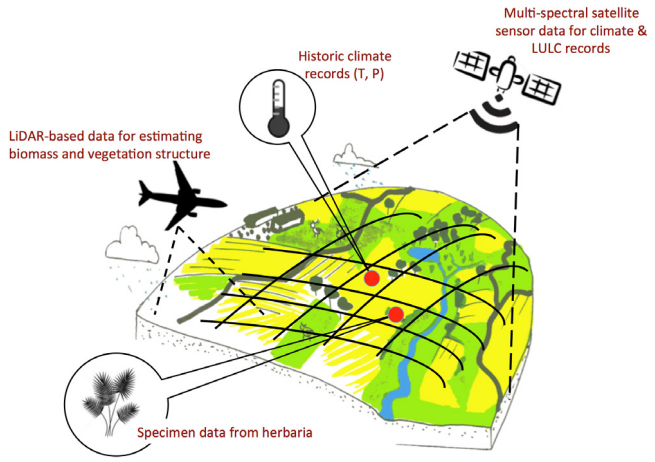


Fig. 1. Schematic representation of the, different data sources involved in the use case.

4. The BioClimate use case system architecture

The BioClimate use case infrastructure addresses scientific computation, data analytics and heterogeneous data storage, leveraging a federated cloud infrastructure for delivering high performance. To face all the scientific challenges, it joins together heterogeneous data sources, on-premises cloud infrastructures, multiple data services, tools, and a central hub, namely the BioClimate Science Gateway (advertised and accessible from the main EUBrazilCloudConnect project website) into a single, federated trans-Atlantic environment, to allow an integrated approach for climate and biodiversity research.

This section provides a detailed explanation of all the components running into the BioClimate environment providing an overview about the full system. With this background, an in-depth discussion on the BioClimate Science Gateway is then presented in Section 5.

The BioClimate use case system architecture (see Fig. 2) consists of the following main building blocks:

- The *BioClimate Science Gateway* represents the central hub with high-level user interface provided by the use case.
- The *Elastic-job engine* (Section 4.4) takes care of the execution of the users' requests submitted through the BioClimate Science Gateway. Driven by the workload, the Elastic-job engine interacts with the Infrastructure Manager (IM) [12] (Section 4.3) to deploy/undeploy multiple *Parallel Data Analytics* cluster instances on-demand (Section 4.2).
- The *Parallel Data Analytics (PDAS)*, [13,14] provides support regarding scientific data analytics applied to large-scale datasets. In the back-end, it properly handles different domain-specific libraries to deal with several scientific data formats (*e.g.* Network Common Data Form - NetCDF [15]), tools (*e.g.* Geospatial Data Abstraction Library — GDAL [16]) and services (*e.g.* OpenModeller [17]) regarding both simulated and observed data, thus representing the proper interface between the domain-specific software and the general-purpose part of the system.
- The *System catalog* (Section 4.5) is a central data repository used by both the front-end and the back-end components to store all the dynamic information regarding the BioClimate system.
- The *BioClimate ClearingHouse* (Section 4.5) is a database accessible from the BioClimate Science Gateway interface where scientific users can (i) persistently publish experiments, (ii) search for them in the database (through a facet-based search functionality), and (iii) retrieve the related outputs.

- The lowest layer of the diagram relates to the *private cloud infrastructures*, running either OpenNebula or OpenStack at the Infrastructure as a Service (IaaS) level, and the *data sources* (Section 4.1) made available by the project partners or already available from national and international agencies. Data sources are permanently stored on shared disks attached to the virtual machines (VMs) deployed in the private cloud infrastructure and made available via Network File System (NFS). From a data perspective standpoint, it is important to outline that the approach in BioClimate is not to perform pure data integration at the data sources level, but rather to integrate the analysis of different data sources into the same scientific experiment. To summarize, the data processing relies on virtual machines running PDAS, which are deployed and configured by IM on top of private cloud resources, as required by the Elastic-job engine. In this way, the infrastructure is isolated from the application logic, for the sake of portability. Seamless access to the different private clouds available in the federated infrastructure was provided by Fogbow's Open Cloud Computing Interface **(OCCI)** API. While the Fogbow instance, along with the IaaS environments available at the partners' premises, have been shared across different EUBrazilCloudConnect use cases, specific instances of IM and PDAS have been deployed and dedicated to serve the BioClimate environment only. The following subsections describe more thoroughly the main components of the architecture, following a bottom-up approach, starting from the infrastructural layer up to the System Catalog and ClearingHouse, while the BioClimate Science Gateway is depicted in Section 5.

4.1. Data sources

As mentioned before, the climate and biodiversity systems deal with a wide range of spatial and temporal scales. From a *data* perspective, the proposed use case faces strong challenges like management, integration, analysis and visualization, since the involved data sources are different in nature, spatial resolution and sampling frequency. The data sources included into BioClimate are the following:

- *SEBAL datasets*. Output of satellite images series (Landsat [18]) processed by the SEBAL [19,20] algorithm to produce estimates of energy balance and evapotranspiration of water to the atmosphere, with 30 m-grid resolution and 16-day regular sampling frequency over the whole Earth. Remote sensing data are provided by the United States Geological Survey (USGS) and the National Aeronautics and Space Administration (NASA). In particular, the infrastructure allows processing of Landsat data coming from the Brazilian Semi-arid region. SEBAL dataset consists of 9 very high spatial resolution NetCDF files of 12 GB each. The files provide historical information regarding various climate and vegetation indicators (Normalized Difference Vegetation Index (NDVI), Surface Albedo, Surface Temperature, Net Radiation, Sensible Heat Flux, Evapotranspiration, *etc.*) for a time range spanning from 1984 to 1995.
- *LiDAR data*. LiDAR is an optical remote-sensing technique that uses laser light to densely sample the surface of the earth, producing highly accurate x, y, z measurements. These data are produced on-demand from airborne sensors and, therefore, their sampling frequency is not regular. The spatial resolution is very high, with about 16 points per sq-meter useful to characterize the terrain and 3D vegetation structures. For the areas where hyper-spectral imagery

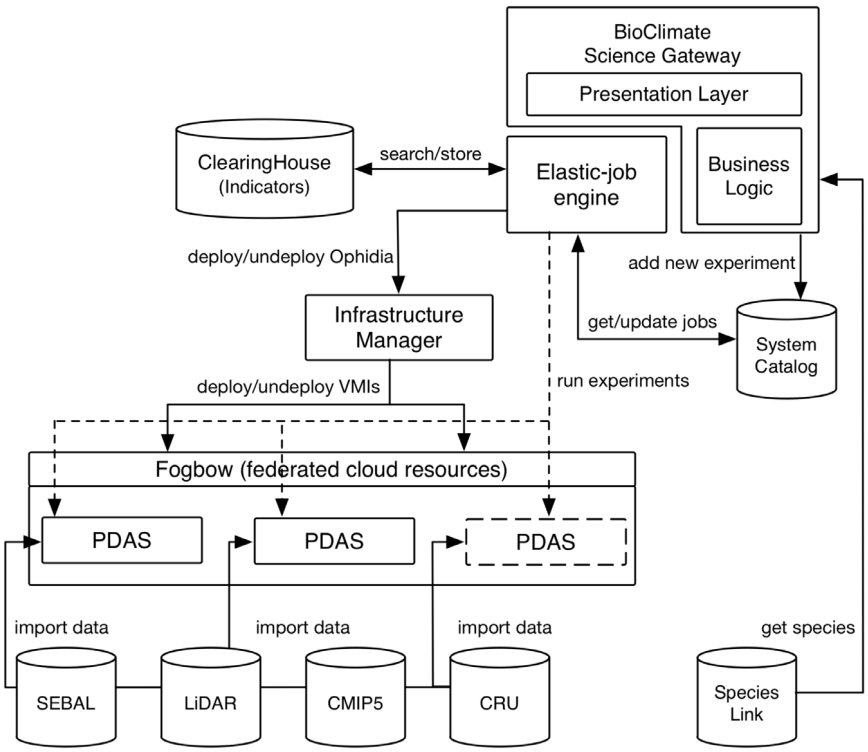


Fig. 2. BioClimate use case system architecture.

is apparently absent, EUBrazilCloudConnect leverages the available LiDAR data, such as the Ducke Reserve near Manaus in Brazil, provided by EMBRAPA [21] (Brazilian Agricultural and Livestock Research Corporation). Vegetation and terrain metrics represent the key indicators that can be inferred from these datasets. Input dataset consists of 9 LiDAR tiles in LASer compressed format (LAZ) for a total of around 800 MB (5.6 GB in non-compressed ASCII textual format).

- *Biodiversity data sources*. Biodiversity data are available for numerous locations worldwide, not regularly sampled over space and time. Our current implementation uses hundreds of locations in two regions in Brazil with data from the *speciesLink* datasets [22], which are provided by CRIA (Centro de Referência em Informação Ambiental). The SpeciesLink provides free and open access to several datasets comprising 7.3 million primary research-grade data, derived from the federation of 350 Brazilian valuable biodiversity datasets, gathered from 150 institutions in Brazil and abroad.
- *Climate data from the CMIP5 Federated Data Archive through the Earth System Grid Federation (ESGF)* [23]. The Coupled Model Intercomparison Project (CMIP) provides a community-based infrastructure in support of climate model diagnosis, validation, intercomparison, documentation and data access. CMCC provides about 100TB of data related to three different models in NetCDF format, Climate and Forecast (CF) metadata conventions [15]. Starting from these datasets, multiple climate indicators can be computed. Input datasets integrated into the platform consist of 8 NetCDF files for a total of 2.3 GB. In particular climate variables providing future prediction (for 2006–2100 time range) for minimum and maximum temperature (from different climate models and emission scenarios) at a global scale are considered.
- *Observed data*. Observations from the Climatic Research Unit (CRU) used in the project are high-resolution gridded time-series datasets (CRU TS v.3.23 [24], released and made available under the Open Database License by CRU, University of East Anglia). Input datasets consist of two NetCDF files of 2.7 GB each, including selected variables such as temperature and precipitation. The datasets relate to the global domain (all land areas excluding Antarctica at 0. 5° resolution) with historical, monthly values from 1901 to 2014.

As it can be easily inferred from the spatial coverage of the different datasets reported before, data availability has been another important aspect to consider in the project to implement the target climate and biodiversity experiments.

4.2. Parallel data analysis service

The PDAS is a core component of the Ophidia project [25] and provides a framework for parallel I/O and data analysis. PDAS relies on an array-based storage model and a hierarchical storage organization to partition and distribute multidimensional scientific datasets. Since the storage model does not rely on any specific scientific data format, it can be exploited in different scientific domains and with very heterogeneous sets of data. As a matter of fact, today PDAS supports the following data formats: NetCDF [15], FITS [26], SAC [27], GRIB [28], thus being able to deal with data from both climate and weather domains as well as from the marine and astronomical contexts.

In the context of the BioClimate system, the PDAS targets scientific challenges for both batch and interactive data analysis on NetCDF, LiDAR and remote sensing data. Python scripts, integrated into the PDAS back-end (through a wrapper-based approach) provide additional functionalities to process LiDAR products by interacting with external tools (*e.g.* GDAL [16]) as well as services (*e.g.* OpenModeller [17]). To address the data analytics requirements and support the processing pipelines of the use case, several new

features and mathematical functionalities have been developed during the project lifetime. Some key analytics operators exploited in the experiments implemented for this use case are: (i) data subsetting and aggregation; (ii) statistics computation, predicate evaluation and linear regression; (iii) script execution to integrate external tools in Ophidia; (iv) dataset import and export; (v) on-the-fly exploration of time series for interactive analysis. So far, the PDAS provides about 100 array-based functionalities and 50 analytics operators. In both cases a specific API is available to help end-users extending the PDAS mathematical and statistical capabilities exploiting a plugin-oriented approach.

All the outputs of the PDAS processing are stored in JSON format. This makes easy the integration of the results into web contexts like the BioClimate Science Gateway and the parsing of the outputs from JavaScript and Python-based applications.

A detailed description of the PDAS architecture, its main features and capabilities can be found in previous papers [13,14]. Moreover, [29] reports in technical detail a former implementation of the automated and cloud-based deployment of PDAS instances using the IM component and its Resource and Application Descriptive Language (RADL [12]).

4.3. Infrastructure manager

The Infrastructure Manager (IM) is an open source "dev-ops" service and application that enables deploying and configuring complex and multi-instance virtual appliances in a wide range of cloud IaaS, such as on-premises, public and scientific Clouds, and container orchestration platforms. IM eases the access and the usability of IaaS clouds by automating the Virtual Machine Images (VMI) selection, deployment, configuration, software installation, monitoring and update of Virtual Appliances [12], on multiple cloud backends (OpenNebula, OpenStack, Amazon EC2, Microsoft Azure, Google Cloud platform, OCCI, Docker, Kubernetes, libvirt, EGI Federated Cloud and, thanks to EUBrazilCloudConnect, Fogbow). In combination with Elastic Compute Clusters in the Cloud (EC3) [30], it also manages the horizontal elasticity of the infrastructure (adding/removing nodes) and supports the TOSCA Simple Profile in YAML Version 1.0 [31] for infrastructure description.

TOSCA specifications are translated into the IM native Resource and Application Description Language to create and to get information about the infrastructure. Through a declarative approach (merging standard specifications, such as Open Virtualization Format (OVF) [32] with the contextualization language derived from Ansible [33]), RADL allows defining the requirements of the target resources in terms of hardware specification, software requirements, OS configuration, *etc*. In the overall use case architecture, the Elastic-job engine exploits IM to instantiate on the EUBrazilCloudConnect federated cloud infrastructure VMIs through the PDAS recipes in RADL.

In the frame of EuBrazilCloudConnect, Infrastructure Manager deploys the BioClimate Virtual Appliances on top of Fogbow [34]. Fogbow is a lightweight and extensible middleware to federate IaaS clouds. It provides an API that implements the OCCI standard, and extensions to this standard to address federation functionalities, such as membership management and asynchronous instantiation of resources. These asynchronous requests, called orders, provide a more appropriate way to issue the remote creation of resources in a federation, and avoid failures due to timeouts and resource pre-emption. The middleware also provides extra functionalities useful for the federation of private clouds, such as a reverse tunneling service to allow access to VMs with only private IPs, and mechanisms to deploy private clouds backed up by desktops exploited opportunistically and with power save capabilities. Finally, Fogbow can be used as a simple way to seamlessly integrate in the federation resources that can be acquired from public IaaS providers, i.e. cloudbursting.

4.4. Elastic-job engine

The Elastic-job engine (a multi-threaded daemon based on the GNU C libraries) has a twofold goal: (i) it translates the user requests into tasks and it properly schedules the resulting jobs on the different PDAS running on the EUBrazilCloudConnect federated cloud infrastructure; and (ii) it interacts with IM to dynamically and elastically manage the set of available PDAS instances.

The management of the workload is performed exploiting a smart scheduling algorithm, which dynamically assigns the jobs over a set of queues. A job queue is associated to each PDAS cluster running on the infrastructure. To horizontally scale the infrastructure, a new PDAS instance is deployed automatically on the private cloud resources when the number of pending jobs on all queues exceeds a configurable threshold. On the other hand, when a PDAS instance is idle for a given (configurable) time period, then the instance is undeployed. A more detailed description of the queue policy adopted and its rationale, are provided in [29]. It is worth mentioning that interactions among the gateway interface and the Elastic-job engine are based on an asynchronous approach (jobs are first sent by the front-end to a relational database that serves as a common queue and then – when a slot becomes available in one of the PDAS instance queues – are handled by the Elastic-job engine). Along with the deployment and configuration of a new node, IM can reconfigure all the nodes according to the new scenario (with more or less instances).

4.5. System catalog and clearingHouse database

This section describes in more technical details –highlighting key features and main differences – two databases exploited in the BioClimate environment: the *System Catalog* and the *ClearingHouse*. From a very high-level categorization, while the former is more dynamic and operational, the latter is more static and historical. In the following, the most relevant technical aspects are discussed.

The System Catalog stores the management information regarding the whole system and represents a key component for the interactions between the Science Gateway and the rest of the infrastructure. The System Catalog is implemented as a relational database and it runs on top of a MySQL RDBMS. In particular, it keeps track of the following information:

- Experiment configuration, runtime status (e.g. pending, running, success, error, abort), type, owner/user, input arguments, submission and completion times, output format (e.g. chart, table), etc.
- PDAS clusters status, which includes current and past activity such as the status of the nodes and the number of jobs executed by each node;
- Data source metadata, which includes both infrastructural settings related to the setup of each data source, and scientific metadata like variable name, time range, spatial domain, coordinate system, etc.

The information related to the runtime status (e.g. experiments and PDAS instances) is updated by the Elastic-ob engine. In particular, once the experiment is submitted from the BioClimate interface, the business logic of the Science Gateway adds the new experiment in the System Catalog with a *pending* state and periodically checks for its completion. In parallel, the Elastic-job engine asynchronously queries the System Catalog to spot new experiment requests, dispatch and submit the related processing pipelines/workflows on a selected PDAS (as explained in Section 4.4) and updates their execution status. The PDAS cluster usage information is periodically updated by the Elastic-job engine during the experiment execution. When an experiment is

completed, the information related to its output is stored in the *temporary storage area* of the System Catalog for a limited period of time (e.g. 5 days). Such area is regularly cleaned by another Elastic-job Engine thread looking at the expiring date of each experiment.

On the other hand, the BioClimate ClearingHouse system allows a scientist to publish and share with other end users, her relevant experiments and their output results. Similarly to the System Catalog, it has been implemented exploiting a MySQL RDBMS to store information about the experiments, output products and the associated metadata (e.g. spatial and temporal domains, type of experiment, variables or indicators considered, etc.).

The mechanism by which an experiment is moved from the temporary storage area of the System Catalog to the persistent one in the ClearingHouse is called the *publication process*.

5. The BioClimate Science Gateway

This section presents in detail the BioClimate Science Gateway, the central hub of the use case, which provides user-friendly access to a set of analytics experiments combining the analysis of the different data sources to study the impact of climate change in regions with high interest for biodiversity conservation, such as the Brazilian Amazon and the semi-arid Caatinga regions in Brazil.

The BioClimate Science Gateway is open to all users (data policies and licensing information are also available on the informative section of the gateway) and provides (after a simple registration phase) a user-friendly and highly interactive web interface to access and process (i) historical temperature and precipitation records, (ii) different climate model scenarios with predictions of future temperature and precipitation, (iii) Landsat satellite imagery for climate and biodiversity indicators, (iv) LiDAR 3D forest metrics and biodiversity indicators at a very high resolution, and (v) plant occurrences data for ecological niche models for the prediction of future plant distribution based on different climate scenarios.

It is worth to mention that the application-domain partners of the EUBrazilCloudConnect project, which represent a broad set of scientific disciplines including ecology and species niche modeling, biodiversity, remote sensing, climate change, and environmental modeling, (i) have actively contributed to the definition of a comprehensive set of experiments, and (ii) have been strongly involved into the design (requirements elicitation, testing, and validation phases) of the BioClimate Science Gateway, providing also key input from an end-user perspective regarding the definition of the user interfaces associated to each experiment.

Through BioClimate, scientific users can also (i) submit multiple types of experiments, (ii) visualize the output of the analysis related to remote sensing images that provides 3D information concerning the structure of the vegetation, (iii) get access to the ClearingHouse system to search for and retrieve previous experiments and contribute to it with new results, among other tasks.

At a very high-level, the key requirements regarding the BioClimate Science Gateway can be grouped into the following categories:

- *Support for heterogeneous data sources*. BioClimate *must* provide a unified interface to access and process (i) satellite images (from Landsat), (ii) meteorological/climate data (from CMIP5), biodiversity data (from speciesLink) and (iv) LiDAR datasets related to some target areas. Furthermore, the gateway *must* provide also metadata information describing these data sources.
- *Integration of multiple analysis and visualization tools*. BioClimate *must* allow data analysis and visualization by supporting a set of domain-specific tools and algorithms to (i) calculate 3D vegetation products based on LiDAR data [10], (ii) run Ecological Niche Modeling (ENM) over species data, (iii) process climate models datasets to calculate indicators, and (iv) run time series analysis, data reduction, data subsetting and data transformation on SEBAL data.
- *Clearinghouse system.* The gateway *must* provide access and visualization to persistently stored products for future reference and download, exploiting a ClearingHouse system. The system *should* allow a fast and easy access to cached experiments through online search & discovery capabilities.
- *Usability*. The BioClimate interface *must* support the execution of scientific data analysis experiments, data visualization, data exploration, access to data in multiple formats/outputs, and interactivity.

The following sections provide a description of the main views and interfaces made available by the BioClimate Science Gateway, showing how the requirements presented above have been addressed.

5.1. Interactive analysis

The *Interactive analysis* panel supports real-time, exploratory analysis of time series from the climate data available through the gateway, i.e. CRU historical data (temperature and precipitation variables) and future simulated data from the CMIP5 experiment (maximum and minimum temperatures from different climate models and scenarios).

As shown in Fig. 3, the interactive analysis requires the selection of a dataset and a variable from the list of datasets/variables available and a point from the map. The bottom section of the BioClimate Science Gateway displays the result of the analysis in terms of (i) a chart with the time series and its trend line and (ii) a table with a comprehensive set of aggregated statistical information, which includes mean, variance, standard deviation and autocorrelation, among others.

5.2. Batch analysis

The *Compute* panel provides a set of features to define and submit complex batch analysis experiments regarding the available data sources. The experiment execution and its status are tracked on the interface, which constantly provides a real-time feedback to the end-user (see Fig. 4).

A set of experiments is provided through the Compute panel. For each of them, a map for spatial domain selection and a form to set the input parameters are also provided. The following experiments have been defined and implemented:

- The *Interannual analysis of SEBAL output* (see Fig. 4) provides information about interannual trends and statistical information of a specific SEBAL variable. The BioClimate Science Gateway integrates data processed by the SEBAL algorithm and provides functionalities to analyze several variables produced by this algorithm (*e.g.* Enhanced Vegetation Index, Leaf Area Index, Normalized Difference Vegetation index, Land Surface Albedo, Ground Heat Flux, Net Surface Radiation and Land Surface Temperature). The interface allows both spatial and temporal selection. The results of the experiment show the distribution of the selected SEBAL variable for each month and some statistical values computed on each year (as in Fig. 5).
- The *Climate and SEBAL variables intercomparison* allows the comparison of the behavior of climate and SEBAL variables. In particular it supports the analysis of the variables produced by the SEBAL algorithm and the ones (precipitation and temperature) from the historical climate data. From a

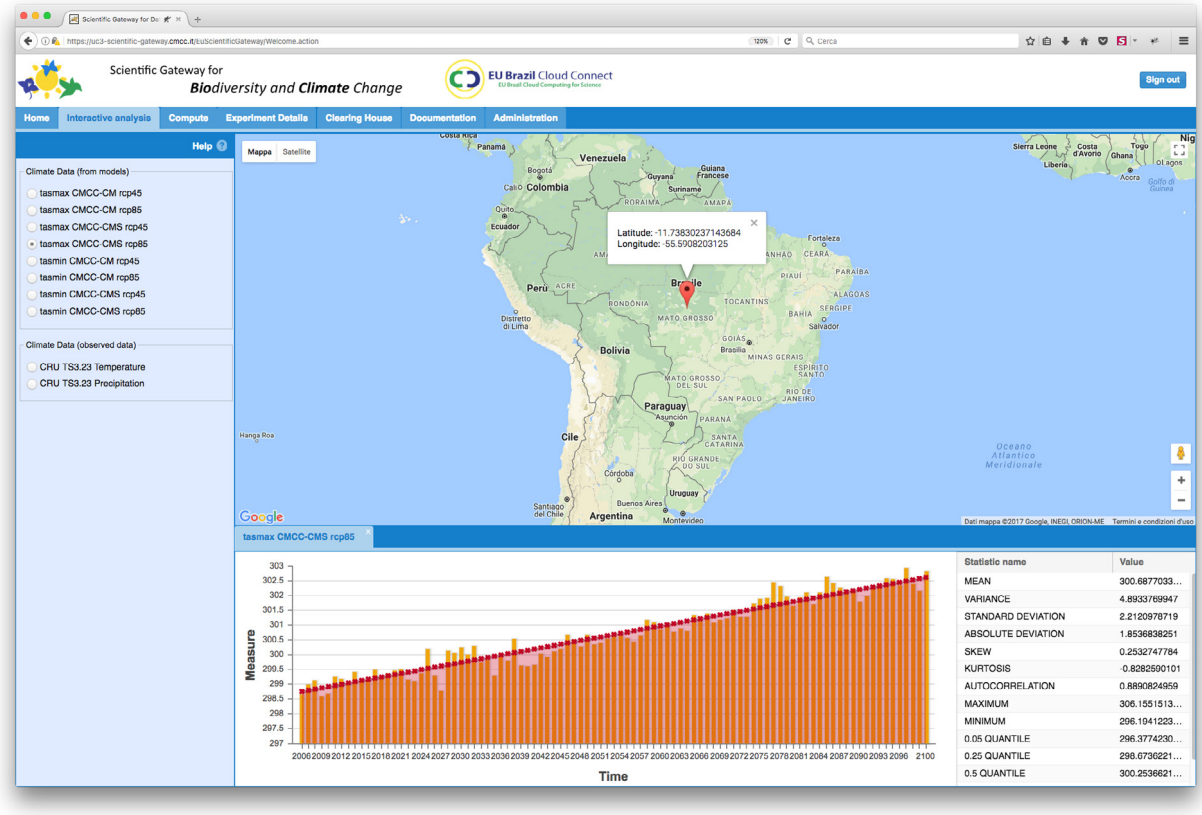


Fig. 3. Interactive analysis.

scientific standpoint, this experiment provides useful information about the relationship between climate and vegetation indices. Like in the previous experiment, a bounding box and the time range can be selected. The output of the experiment (see Fig. 7) shows the trend of the climate and SEBAL variable in a line chart, whereas their correlation is represented in a scatter plot.

- The *Climate indices intercomparison* allows the comparison of indicators computed on CMIP5 datasets related to different climate models and for future emission scenarios (RCP4.5 and RCP8.5 [35]). Four well-known indicators based on maximum and minimum temperature are available for comparison (over a user-defined spatial domain and time-range boundaries) [36], namely:
  - TXx: annual maximum temperature computed over monthly maximum temperatures;
  - TNx: annual maximum temperature computed over monthly minimum temperatures;
  - TXn: annual minimum temperature computed over monthly maximum temperatures;
  - TNn: annual minimum temperature computed over monthly minimum temperatures.
- The *Ecological Niche Modeling* (ENM) experiment integrates the functionalities available through the OpenModeller Web Service API to create and project models defined over occurrences of biodiversity data. This experiment allows the comparison of the projections of models into three different environmental scenarios (*present*, *future optimistic 2070* and *future pessimistic 2070*). The models are created with the maximum entropy algorithm [37] and are based on the species occurrences selected by the user. After the projection of the model is completed, the experiment output shows a map for each environmental scenario as well as the metadata related to the experiment. The maps can be also downloaded in raster format (ByteHFA – Byte Erdas Imagine).
- The *LiDAR products intercomparison* allows the comparison and evaluation of the statistical relationship between LiDAR products. More specifically, the following LiDAR products are available through the BioClimate Science Gateway:
  - Vegetation metrics: Digital Surface Model (DSM), Canopy Height Model (CHM), Forest Cover, Above-ground Biomass and Relative Height (at 50%);
  - Terrain metrics: Digital Terrain Model (DTM), Aspect and Slope Angle;
  - LiDAR data characteristics: Point Density.

  In this case, a LiDAR tile can be selected from the map. The output of the experiment shows images representing the LiDAR products (Fig. 6(a)), a scatter plot with the correlation between the two selected variables (Fig. 6(b)) and a table with statistical values (i.e. maximum, minimum, mean and standard deviation) for each product.
- *Relative Height analysis of LiDAR data* provides information about relative height at different percentiles (25%, 50%, 66%, 75% and 90%) of the points in a LiDAR tile. The relative height analysis provides insight about the vertical distribution of vegetation points, and indirectly also on the vertical distribution of biomass which is an important biodiversity indicator. For each percentile, the images representing the LiDAR products and the relative height point distribution histogram, jointly with some statistical information, are shown in the output results, similarly to the previous experiment.

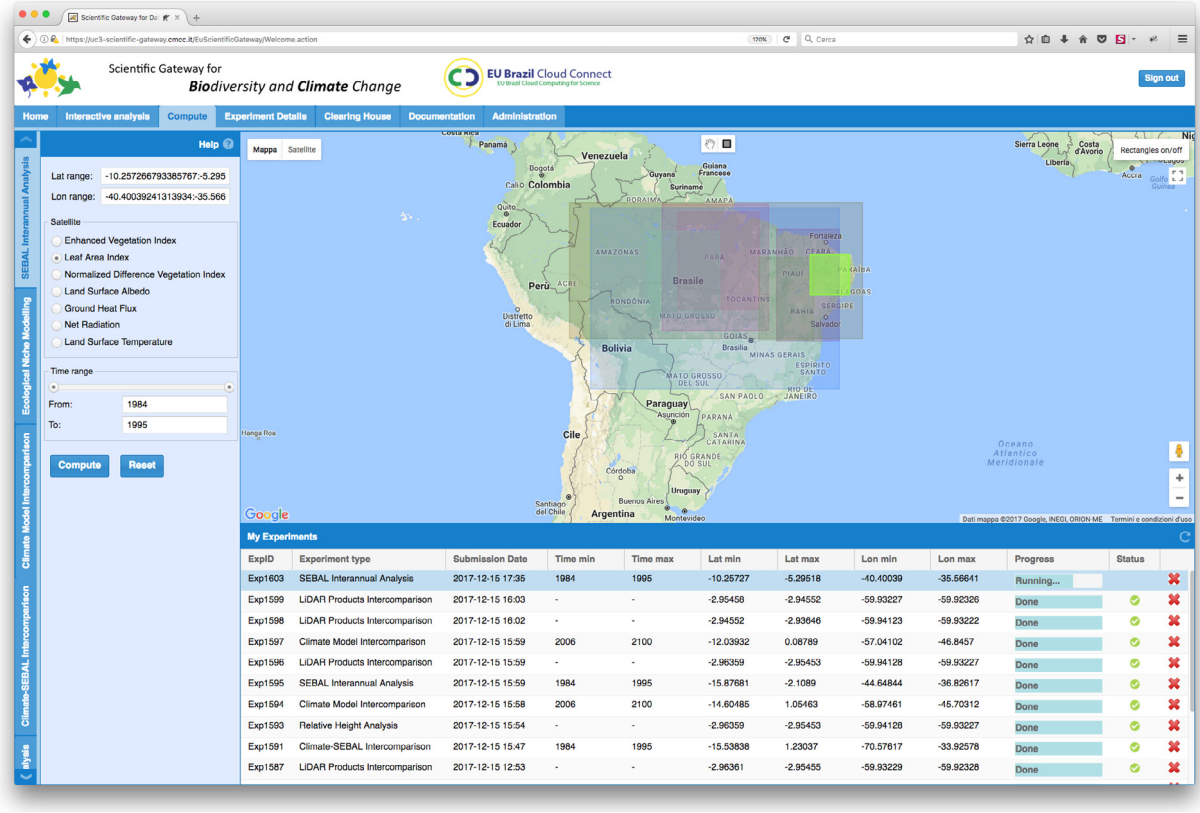


Fig. 4. SEBAL Interannual analysis compute interface. The interface allows the selection of a bounding box from the central map, and one SEBAL variable and the time range boundaries from the column on the left. After the experiment is submitted, the bottom table is updated with a new row displaying information about the status of the experiment.

5.3. Experiment visualization & download

Once the computation of the experiment is completed, the details about the output are available through the *Experiment Details* section. Fig. 5 displays the output produced by a SEBAL interannual experiment, whereas Figs. 6 and 7 display the output produced respectively by a LiDAR intercomparison experiment and Climate-SEBAL intercomparison experiment.

In particular, to better suit the experiment peculiarities, a specific detail view is provided for each experiment defined above. Hence, various gadgets organized in different fashions are used to display the results; among these are: line charts to display statistical values and trend lines; scatter plots to evaluate variable and indicators correlation; tables to show the results and statistical values; maps with the environmental scenario; images of the LiDAR products; and, finally, point distribution histograms.

Most of the information provided through the gadgets is also available for download in CSV, raster, GeoTIFF, and PNG formats (according to the experiment type). Furthermore, metadata information regarding the experiment is available in the same view.

5.4. BioClimate Science gateway ClearingHouse interface

The BioClimate ClearingHouse aims at becoming a community database where scientists can store all their key experiments and make, according to the FAIR principles [38], their research data findable, accessible, interoperable and reusable. The BioClimate Science Gateway ClearingHouse interface provides a specific view showing the experiments performed and saved by the end-users. The experiments are also visualized on a map with different colors in order to quickly identify those related to a specific area. Additionally, a facet-based search feature is available to discover the experiments of interest stored into the ClearingHouse. In this regard, examples of search criteria are: (i) spatial domain, (ii) experiment type, and (iii) submission date. The BioClimate Science Gateway ClearingHouse interface provides interactive *read-only* views (related to the presentation and visualization of the experiments output in terms of data and metadata) which are exactly the same of those described in the previous Section 5.3. The output data are also available for download.

5.5. Infrastructure monitoring

The execution of data analytics jobs in this highly dynamic and elastic cloud-based scientific environment can be affected by the scheduling policies and the time related to deploying/undeploying the infrastructural resources. Hence, a visual interface to track job execution is very useful to understand (i) how effectively the resources are used and (ii) how well the different job policies are performing. To this end, the BioClimate Science Gateway includes two administrative interfaces that allow, respectively, the management of users and their privileges and (ii) the monitoring of the resources exploited dynamically by the gateway (*i.e.* PDAS cluster instances) as well as some statistics regarding the experiments executed, classified according to their type and status (i.e. pending, running, completed successfully and with error). This dashboard (see Fig. 8) provides charts displaying real-time monitoring information regarding the number of experiments running or pending, and the current status of the resources. In particular, a histogram shows the set of experiments executed on the PDAS instances active in the last 5 min, while a pie chart shows the distribution of the experiments across the clusters currently running at least one job.

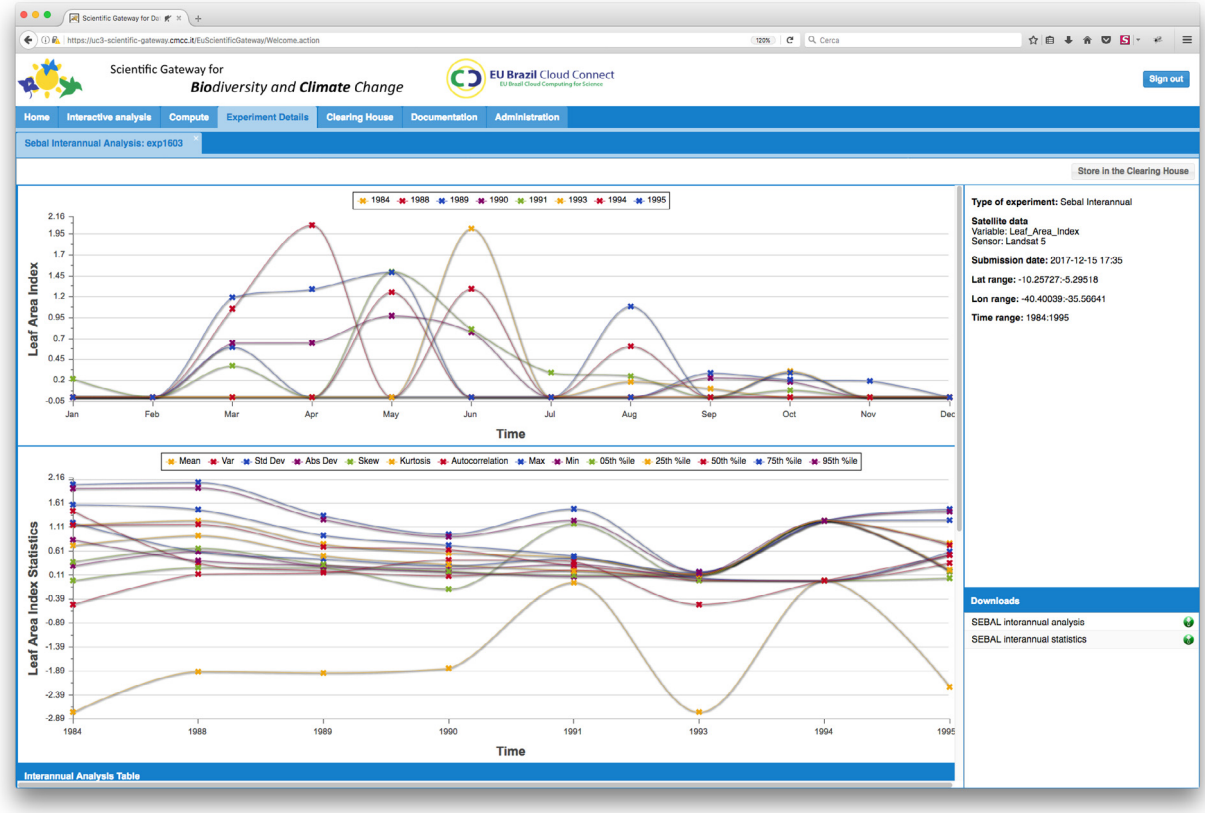


Fig. 5. SEBAL Interannual analysis details interface. The results show the distribution of the selected SEBAL variable by year and the statistical values for that variable computed on each year. This information is displayed both in charts and tables and can be downloaded in CSV format. Moreover, metadata of the submitted experiment is available on the right side of the page.

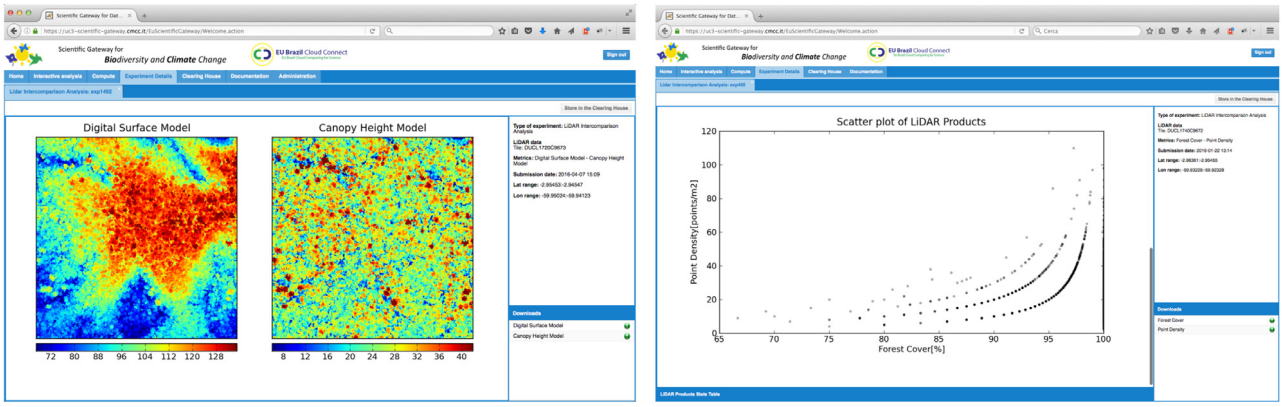


Fig. 6. LiDAR intercomparison details interface. The output of the experiment shows images representing the LiDAR products (Figure 6.a) and a scatter plot with the correlation between the two selected variables (Figure 6.b). Additionally, the images representing the products can be downloaded in GeoTiff format. Metadata related to the experiment is reported on the right side of the page.

5.6. Portability, usability, extensibility, and performance

To address *portability* and *separation of concerns* between the presentation layer and the business logic, the BioClimate has been implemented according to the Model-View-Controller (MVC) pattern. The presentation layer, running on the client side (*i.e.* a browser), provides a rich user interface to submit data analysis tasks and visualize their results. It is implemented as a JavaScript web application based on the Ext JS library [39]–which offers a number of gadgets such as panels, charts and grids – and Google Maps API [40] for the visualization of geo-referenced data. The server side of the Science Gateway implements the business logic to manage users, handle the requests and the post-processing of the results, and it is based on Java and Apache Struts2 framework [41].

To increase *performance* and make the output visualization faster, by design, the heavier tasks, related to the post-processing of the outputs, are performed on the server side and the ready-to-use results are consumed by the JavaScript library on the presentation layer. The full environment is highly efficient, due to high-performance capabilities offered by the PDAS, the possibility to run multiple PDAS instances in parallel and the job scheduling policy of the Elastic-job engine which takes into account data locality. Caching mechanisms running on the back-end of the BioClimate Science Gateway are out of the scope of this paper and will be presented in a future work.

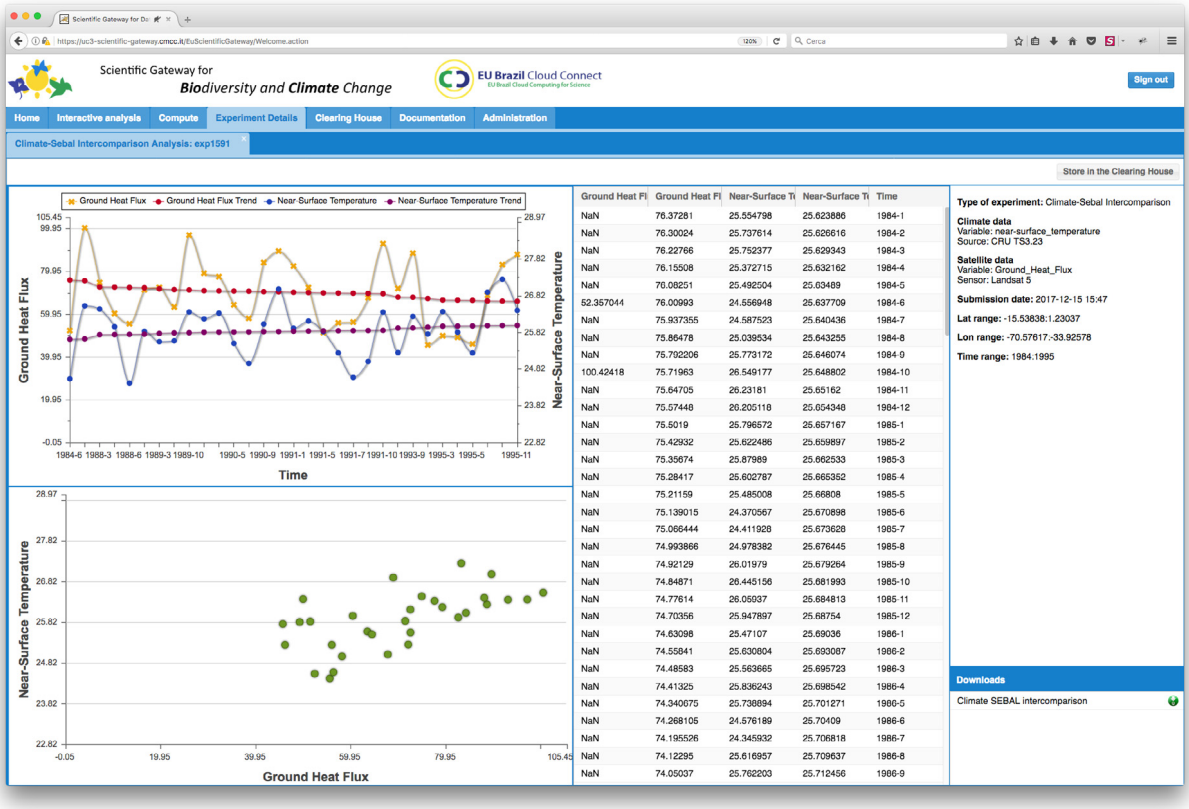


Fig. 7. Climate-SEBAL intercomparison details interface. The output of the experiment shows, in a line chart, the trend of the climate and satellite variable and, in a scatter plot, their correlation. The trend and the values of the variables are also displayed in a table and are available in CSV format for download. Metadata regarding the experiment is available on the right column.

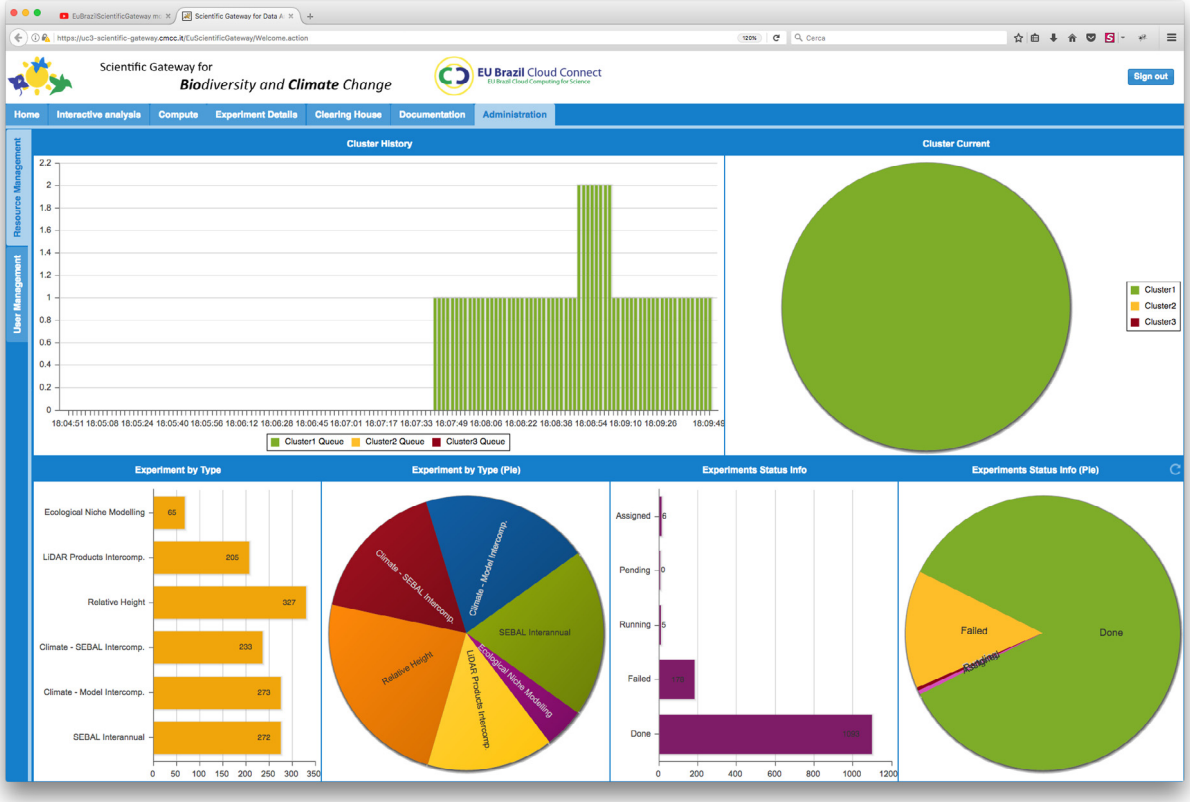


Fig. 8. Monitoring dashboard.

Usability has been addressed by defining and validating the BioClimate Gateway by target users using a set of predefined experiments regarding the different data sources and types of analysis. Each experiment is associated to a customizable template to address data analytics tasks on climate and biodiversity data; it requires a specific pipeline of operations, including sub-setting, data reduction and mathematical/statistical functions.

The user interface has been implemented starting from a core set of templates and gadgets for scientific data analysis (e.g. output visualization, input form, experiment submission) implemented in the first part of the project and used as building blocks to be combined for the development of all the experiments. Through this approach, *extensibility* has been also addressed at the Science Gateway level, since additional types of experiments could be added in the future as new modules, without changing or impacting on the overall implementation. This also relies on the fact that the PDAS framework offers a broad set of analytics operators that can be easily combined together to design a scientific experiment and submitted using the same PDAS *experiment* web service interface (WS-I compliant).

5.7. Security

Security is implemented across the whole architecture taking into account several levels. Regarding the front-end, security is implemented in terms of user authentication. In order to avoid potential attacks that aim at stealing passwords, the system employs a technique based on salted password hashing, using a Java implementation of a Cryptographically Secure Pseudo-Random Number Generator, called Password-Based Key Derivation Function 2 (PBKDF2) [42]. Additionally, HTTPS is used to provide encryption for the communications between client and server.

At the Elastic-job engine level, the PDAS terminal is used to send requests to a PDAS server interface. It exploits the X509v3 digital certificates-based authentication and the Virtual Organization Membership Service (VOMS) based authorization. Different levels of privileges are defined to distinguish user roles locally at each PDAS server or globally at the VOMS server. For this purpose, a Grid Security Infrastructure (GSI) and VOMS enabled interface, supporting both X.509 certificates and VOMS-based authorization, has been implemented. It also addresses the interoperability with the EGI FedCloud environment [43].

6. Related work

In the Biodiversity and Climate change domain there are various tools used by scientists today. However, there are several limitations and issues that relate to the data availability, data analysis approaches, performance, usability, etc. This section presents some of the most relevant related work in the area that represent the state of the art for these research topics.

LiDAR online [44] provides an example of platform to market LiDAR, GIS data on the Internet in an intuitive user interface. Such an approach provides an intuitive map-based search window for the end-user, and easy to find products under the Geo-Services tab, forestry. However, it is difficult to add data/find metadata of input products and tutorials are unclear and not intuitive, especially for non-specialists.

NASA LiDAR Access System (NLAS) provides an example of the future role of an application to serve world-wide products [45]. It focuses on LiDAR data only.

NDVI changes (Esri) [46] offers time series analysis through a user-friendly interface. However, the analytical part of the products are limited, since multi-temporal analysis cannot be customized, and it does not allow downloading the product and overlapping layers as vector type files.

Series View (by LAF/INPE) [47] allows downloading graphs and specific-location data. However, it provides a small number of products for analysis and it does not provide the images used in the application.

The Brazilian Semiarid NDVI Viewer (by INSA/UFAL) [48] gives the possibility of running temporal data analysis of socioeconomic data. On the other hand, it provides a low number of land surface information: it only uses one vegetation index, does not allow crossing of socioeconomic and environmental information.

World Evapotranspiration Web Viewer (Esri) [49] provides an important environmental variable (evapotranspiration) for all land surfaces. However, the information provided is summarized at catchment scale. It does not allow the user to obtain more detailed information on higher resolutions. The information on evapotranspiration refers to annual totals, thus presenting limited application for environmental studies. It is not possible to download the data used in the application.

Concerning climate data, the most relevant production-level tools for data analysis are Climate Data Operators (CDO) [50] and NetCDF Operators (NCO) [51]. In both cases, they are client-side and sequential. These are key limiting factors to tackle (near) real-time data analytics. Other software like Live Access Server (LAS) [52] have a server-side support in terms of data visualization, but still they rely on a sequential analytics back-end.

One of the major innovations of BioClimate is to allow flexible user-defined integrated analyses of very distinct datasets — in their nature, temporal and spatial resolution and geographical coverage. One basic obstacle to these analyses was the lack of knowledge of one scientific community on another field's data availability and formats. BioClimate makes available data from various sources in a way that can be handled easily by users of different backgrounds. Another difficulty is the access to data analytics tools that can handle the wide variety of data sources for biodiversity and climate change studies, now dealt under BioClimate. The above examples of existing services show that they are mostly information providers of particular environmental variables and do not allow for user-defined cross variable analyses.

7. Main benefits, impacts, and user experience

The BioClimate Science Gateway represents a major step forward for climate change and biodiversity research studies. Indeed, it (i) integrates several scientific data analysis tools into the same environment, (ii) provides input from a wide variety of scientific datasets, (iii) implements specific analysis experiments addressing climate and biodiversity indicators calculation, (iv) combines analysis tools (interactive and batch) using a multi-disciplinary approach, (v) provides interactive visualization tools for rapid interpretation of results (dashboard approach), and (vi) implements a large set of experiments (e.g. SEBAL Interannual Analysis, Climate Model Intercomparison, Climate-SEBAL Intercomparison, Relative Height Analysis, LiDAR Products Intercomparison, and Ecological Niche Modeling).

7.1. Main benefits and impacts

In terms of benefits, with regard to the existing tools, BioClimate exploits the EUBRazilCloudConnect cloud infrastructure to speed up the processing of satellite images, allowing the execution over bigger areas and with multiple scenes in parallel. Through the PDAS it supports the analysis of very long time series, avoiding huge data downloads, which represents a strong limitation for current desktop-based data analysis approaches.

Yet, scientific users found key having into the same environment a set of tools they commonly rely on, already available and

transparently accessible, just with a browser. There is no longer need to deal with multiple sequential scripts (e.g. Python or bash) due the availability of high-level concepts like the ***experiments***, which come with an already set up environment exploiting, behind the wall, the high performance data analytics support offered by the PDAS to speed up performance. Of course this is limited to the data sources available in the current infrastructure and to the set of experiments implemented so far. However, it is also true that the modularity of the entire system allows for an easy implementation of new experiments as well as the setup of additional data sources.

The opportunity for cross-related analysis (several experiments target this aspect) is another advantage which was very positively evaluated by the scientific users during the validation phase performed in the EUBrazilCloudConnect project. Indeed most of the tools available today are strictly related to a specific type of data and opportunities to cross-relate different and large data sources is very limited.

7.2. User experience and scientific validation

The first users of BioClimate are the participants of the EUBrazilCloudConnect project, which took part in the various phases of the use case design, development and validation. Hands-on and training sessions have been organized with the application-level users during the development stages in order to gather feedback about the features implemented and the usability of the system. As a result, a set of how-to guides have been embedded into the Science Gateway interface to assist beginner users. Additionally, the users undertook an internal assessment after the first and second release of the system assuring its compliance with respect to the requirements specification. The results of this assessment demonstrated not only the adherence to the requirements, but also the reliability of the software, the effectiveness of the interface, its performances and usability (with a gentle learning curve). Finally, the output produced by the BioClimate Science Gateway have been also validated from a scientific standpoint.

8. Conclusions and future work

This paper presented the BioClimate Science Gateway, a data-centric and user-oriented scientific environment for climate change and biodiversity research. The proposed solution (i) leverages on a solid background regarding cloud computing, big data analytics, web and database technologies, (ii) has been deployed in a real federated infrastructure across Europe and Brazil, and (iii) provides a user-friendly, seamless and highly interactive Science Gateway supporting end-users scientific research on biodiversity and climate change.

As described in this paper and based on the end-user feedback and evaluations, the BioClimate platform implemented in the EUBrazilCloudConnect project provides a novel, unique, and integrated environment with multiple, heterogeneous data sources and several types of analytics capabilities offered to scientists through user-friendly interfaces. Such elements have represented for the target scientists a major step forward to perform scientific data analysis and visualization in this domain. The user experience has been good and the change of paradigm (process the data on the server-side) has been evaluated as the key added value with regard to existing approaches. One limitation to the current implementation is that all kinds of data are not available on the entire Brazilian region, so only two target areas have been taken into consideration to demonstrate the feasibility of the approach. However it is important to remark that, based on national and international agencies efforts, new data (*e.g.* LiDAR) with a larger coverage will be available over the next years.

Extending the set of available data sources, jointly with new analysis and visualization capabilities will be part of the future work; accordingly, new analytics experiments will be included into the BioClimate Science Gateway. Security support, based on OAuth2 and OpenIDConnect standards will be also added to the system.

As a final remark, a lot of interest has been also raised by governmental & environmental agencies, both research & education, especially in Brazil. As future work, a set of actions addressing sustainability, even beyond the end of the project, will be implemented in close connection with the relevant stakeholders.

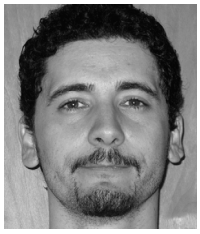
Acknowledgment

This work was supported by the EU FP7 EUBrazilCloudConnect Project (Grant Agreement 614048), and CNPq/Brazil (Grant Agreement n° 490115/2013-6).

References

[1] EUBrazilCloudConnect. Online: http://eubrazilcloudconnect.eu. Last accessed: 24 August 2017.
[2] Kacsuk Péter (Ed.), Science Gateways for Distributed Computing Infrastructures: Development Framework and Exploitation By Scientific User Communities, Springer.2014 http://dx.doi.org/10.1007/978-3-319-11268-8.
[3] U. Becciani, E. Sciacca, A. Costa, P. Massimino, C. Pistagna, S. Riggi, F. Vitello, C. Petta, M. Bandieramonte, M. Krokos, Science gateway technologies for the astrophysics community, Concurr. Comput.: Pract. Exper. 27 (2015) 306–327. http://dx.doi.org/10.1002/cpe.3255.
[4] N. Wilkins-Diehr, D. Gannon, G. Klimeck, S. Oster, S. Pamidighantam, TeraGrid science gateways and their impact on science, Computer 41 (11) (2008) 32–41. http://dx.doi.org/10.1109/MC.2008.470.
[5] Sílvia Delgado Olabarriaga, Nancy Wilkins-Diehr, Concurrency, Computation: Practice, Experience, GCE15 special issue conference publications, 28 (7) (2016) 1949–1951.
[6] D. Elia, A. Nuzzo, P. Nassisi, S. Fiore, I. Blanquer, F.V. Brasileiro, I.A.A. Rufino, A.C. Seijmonsbergen, N.S. Anders, C.de O. Galvao, J.E.de B.L. Cunha, M. de Sousa-Baena, V.P. Canhos, G. Aloisio, A Science Gateway for Biodiversity and Climate Change Research. in: Proc. of IWSG 2016 (8th International Workshop on Science Gateways), 8–10 June 2016, Rome, Italy. CEUR-WS.org, online http://ceur-ws.org/Vol-1871/.
[7] I.P. Llanes-Acevedo, G.E. Ferreira, E. Torres, J. Cala, C. Arcones, F.P. Shimabukuro, et al., A leishmaniasis virtual laboratory to contribute to leishmaniasis surveillance, Trop. Med. Int. Health 20 (2015) 219.
[8] Alya Red System. Online: https://www.bsc.es/computer-applications/alya-red-ccm.Last accessed: 24 August 2017.
[9] S.M. Watanabe, P.J. Blanco, R.A. Feijóo, Mathematical model of blood flow in an anatomically detailed arterial network of the arm, ESAIM Math. Model. Numer. Anal. 47 (4) (2013) 961–985.
[10] M.A. Lefsky, W.B. Cohen, G.G. Parker, D.J. Harding, Lidar remote sensing for ecosystem studies, BioScience 52 (1) (2002) 19–30 Online: http://bioscience.oxfordjournals.org/content/52/1/19.short.
[11] H.M. Pereira, et al., Essential biodiversity variables, Science 339 (2013) 277LP–278 (80-.).
[12] M. Caballer, I. Blanquer, G. Molto, C. Alfonso, Dynamic management of virtual infrastructures, J. Grid Comput. 13 (1) (2014) 53–70.
[13] S. Fiore, A. D'Anca, C. Palazzo, I.T. Foster, D.N. Williams, G. Aloisio, Ophidia: Toward big data analytics for escience, in: Proceedings of the International Conference on Computational Science, ICCS 2013, Barcelona, Spain, 5–7 June, 2013, 2013, pp. 2376–2385.
[14] D. Elia, S. Fiore, A. D'Anca, C. Palazzo, I. Foster, D.N. Williams, G. Aloisio, An in-memory based framework for scientific data analytics. in: Proceedings of the ACM International Conference on Computing Frontiers (CF '16), May 16–19, 2016, Como, Italy, pp. 424–429.
[15] R.K. Rew, G.P. Davis, The Unidata NetCDF: Software for scientific data access, in: Sixth International Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology, 1990, pp. 33–40.
[16] GDAL - Geospatial Data Abstraction Library. Online: http://www.gdal.org/.Last accessed: 24 August 2017.
[17] M.E. Souza Muñoz, R. Giovanni, M.F. Siqueira, T. Sutton, P. Brewer, R.S. Pereira, D.A.L. Canhos, V.P. Canhos, openModeller: A generic approach to species' potential distribution modelling, GeoInformatica 15 (1) (2009) 111–135.
[18] The Landsat program. Online: http://landsat.gsfc.nasa.gov/.Last accessed: 24 August 2017.

[19] W. Bastiaanssen, M. Menenti, R. Feddes, A. Holtslag, A remote sensing surface energy balance algorithm for land (sebal). 1. formulation, J. Hydrol. 212 (1998) 198–212.
[20] W. Bastiaanssen, H. Pelgrum, J. Wang, Y. Ma, J. Moreno, G. Roerink, T. Van der Wal, A remote sensing surface energy balance algorithm for land (sebal).: part 2: validation, J. Hydrol. 212 (1998) 213–229.
[21] Embrapa. Online: https://www.embrapa.br/. Last accessed: 24 August 2017.
[22] Centro de Referencia em Informacao Ambiental. speciesLink service. Online: http://splink.cria.org.br/. Last accessed: 24 August 2017.
[23] K.E. Taylor, R.J. Stouffer, G.A. Meehl, An overview of cmip5 and the experiment design, Bull. Am. Meteorol. Soc. 93 (4) (2012) 485–498.
[24] I. Harris, P. Jones, T. Osborn, D. Lister, Updated high-resolution grids of monthly climatic observations –the cru ts3. 10 dataset, Int. J. Climatol. 34 (3) (2014) 623–642.
[25] The Ophidia Project. Online: http://ophidia.cmcc.it/. Last accessed: 18 November 2017.
[26] FITS (Flexible Image Transport System). Online: https://fits.gsfc.nasa.gov/iaufwg/. Last accessed: 07 September 2017.
[27] SAC –Seismic Analysis Code. Online: http://ds.iris.edu/ds/nodes/dmc/software/downloads/sac/. Last accessed: 07 September 2017.
[28] FM 92-XII Ext. GRIB–General Regularly-distributed Information in Binary form. Edition 2–Version 2–05 November 2003. http://www.wmo.int/pages/prog/www/DPS/FM92-GRIB2-11-2003.pdf.
[29] S. Fiore, Big data analytics for climate change and biodiversity in the EUBrazilCloudConnect federated cloud infrastructure, in: Proceedings of the 12th ACM International Conference on Computing Frontiers, CF'15, Ischia, Italy, May 18–21, 2015, 2015, pp. 52:1–52:8.
[30] A. Calatrava, E. Romero, G. Moltó, M. Caballer, J.M. Alonso, Self-managed cost-efficient virtual elastic clusters on hybrid cloud infrastructures, Future Gener. Comput. Syst. 61 (2016) 13–25.
[31] OASIS Standard, TOSCA Simple Profile in YAML Version 1.0, 21 December 2016. Online: http://docs.oasis-open.org/tosca/TOSCA-Simple-Profile-YAML/v1.0/os/TOSCA-Simple-Profile-YAML-v1.0-os.pdf.
[32] DMTF Open Virtualization Format Specification. Online: https://www.dmtf.org/standards/ovf. Last accessed: 07 September 2017.
[33] Ansible. Online: https://www.ansible.com. Last accessed: 07 September 2017.
[34] F. Brasileiro, G. Silva, F. Araújo, M. Nóbrega, I. Silva, G. Rocha, Fogbow: a middleware for the federation of iaas clouds, in: 2016 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), IEEE, 2016, pp. 531–534.
[35] D.P. van Vuuren, J. Edmonds, M. Kainuma, K. Riahi, A. Thomson, K. Hibbard, G.C. Hurtt, T. Kram, V. Krey, J.-F. Lamarque, et al., The representative concentration pathways: an overview, Clim. Change 109 (1) (2011) 5.
[36] Climate Change Indices. Definitions of the 27 core indices. Online: http://etccdi.pacificclimate.org/list_27_indices.shtml. Last accessed: 07 September 2017.
[37] Maximum Entropy algorithm. Online: http://openmodeller.sourceforge.net/algorithms/maxent.html. Last accessed: 07 September 2017.
[38] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, et al., The fair guiding principles for scientific data management and stewardship, Scientific Data 3 (1) (2016) 160018. http://dx.doi.org/10.1038/sdata.2016.18.
[39] Ext Js library. Online: http://docs.sencha.com/extjs/. Last accessed: 24 August 2017.
[40] Google Maps API. Online: https://developers.google.com/maps/. Last accessed: 07 September 2017.
[41] Apache Struts2 framework. Online: https://struts.apache.org/. Last accessed: 07 September 2017.
[42] B. Kaliski, Pkcs #5: Password-based cryptography specification version 2.0, RFC 2898, September 2000. Online: http://tools.ietf.org/html/rfc2898. Last accessed: 24 August 2017.
[43] EGI FedCloud. Online: http://www.egi.eu/infrastructure/cloud/. Last accessed: 24 August 2017.
[44] LiDAR Online, Online: http://www.lidaronline.com/tools/maps/. Last accessed: 07 September 2017.
[45] NASA LiDAR Access System. Online: http://www.opentopography.org/nlas. Last accessed: 07 September 2017.
[46] NDVI Changes (ESRI). Online: http://changematters.esri.com/compare. Last accessed: 07 September 2017.
[47] Series View. Online: https://www.dsr.inpe.br/laf/series/index.php. Last accessed: 07 September 2017.
[48] The Brazilian Semiarid NDVI Viewer. Online: http://www.insa.gov.br/. Last accessed: 07 September 2017.
[49] World Evapotranspiration Web Viewer. Online: http://www.arcgis.com/home/item.html?id=b1a0c03f04994a36b93271b0c39e6c0f. Last accessed: 07 September 2017.
[50] Climate Data Operators (CDO). Online: https://code.zmaw.de/projects/cdo. Last accessed: 07 September 2017.
[51] C.S. Zender, Analysis of self-describing gridded geoscience data with netcdf operators (nco), Environ. Model. Softw. 23 (10–11) (2008) 1338–1342.
[52] L. Cinquini, D. Crichton, C. Mattmann, J. Harney, G. Shipman, F. Wang, R. Ananthakrishnan, N. Miller, S. Denvil, M. Morgan, Z. Pobre, G.M. Bell, C. Doutriaux, R. Drach, D. Williams, P. Kershaw, S. Pascoe, E. Gonzalez, S. Fiore, R. Schweitzer, The earth system grid federation: an open infrastructure for access to distributed geospatial data, Future Gener. Comput. Syst. 36 (2014) 400–417. http://dx.doi.org/10.1016/j.future.2013.07.002.



Sandro Fiore, Ph.D., Data Scientist and Director of the Advanced Scientific Computing Division at the Euro — Mediterranean Center for Climate Change (CMCC). His research activities focus on High Performance and Distributed Computing, with specific regard on distributed data management, big data analytics and high performance database management. In 2006, he joined as data scientist the CMCC Scientific Computing Division leading the Scientific Data Management Research Group. In November 2014, he has been appointed as Director of the Advanced Scientific Computing. He has been involved in several EU Projects EGEE, EGI-InSPIRE, IS-ENES, EUBRAZILCC, EXARCH, ORIENTGATE, TESSA, CLIP-C, EUBRA-BIGSEA, INDIGO-DataCloud, EESI, EXDCI. Since 2010, he is the Principal Investigator of the Ophidia project, a research project on high performance data analytics and mining for eScience. Since 2011, he has been Visiting Scientist at Lawrence Livermore National Laboratory. He is co-author of more than 50 papers in refereed books/journals/proceedings on distributed and grid computing and holds a patent on data management topics. He is editor of the book "Grid and Cloud Database Management" (Springer, 2011). He is ACM Member.

Donatello Elia holds a M.Sc. degree in Computer Engineering (2013) from the University of Salento in Italy. In 2013 he joined the Advanced Scientific Computing (ASC) division at the Euro-Mediterranean Center on Climate Change (CMCC) Foundation. His main research interests include high performance, cloud and distributed computing, data analytics and mining, and data management. He has been involved in various European projects from FP7 and Horizon 2020 programme. He is also a member of IEEE and IEEE Computer Society.

Ignacio Blanquer, associate professor of the Computer System Department at UPV since 1999, has been a member of GRyCAP since 1993. He has been involved in Parallel Computation and Medical Image processing, participating in more than 60 national and European Research Projects, has authored and co-authored 32 articles in indexed journals and book chapters and in more than 80 papers in national and international journals and conference proceedings. He has served as coordinator of the application area in the Spanish Network for e-Science, including his role in the managerial board, participates in the user support of the Spanish National Grid Initiative and was Community Manager in VENUS-C, where he collected and evaluated user requirements and provided hands-on support for migration to the cloud. He has been the project coordinator of EUBrazilCloudConnect (FP7) and CLUVIEM (national research project) and currently is the project coordinator of EUBra-BIGSEA (H2020) and co-principal investigator in the BigCLOE national research project.

Francisco Brasileiro is a Full Professor at the Federal University of Campina Grande, Brazil. He received a B.S. degree in Computer Science from the Federal University of Paraba, Brazil, in 1988, an M.Sc. degree from the same University in 1989, and a Ph.D. degree in Computer Science from the University of Newcastle upon Tyne, UK, in 1995. His research interests are in distributed systems in general, with focus on federated and cooperative systems. He has co-authored more than 100 papers published in refereed journals and conference proceedings. He has been the Principal Investigator of over 20 projects, funded by public funding agencies in Brazil, and by industry. He is a member of the Brazilian Computer Society, the ACM, and the IEEE Computer Society.

Alessandra Nuzzo obtained a Master of Science, cum Laude, in Computer Engineering at the University of Salento in October 2011. In 2013, she joined the Advanced Scientific Computing division of CMCC and worked in the following projects: ORIENTGATE, TESSA, OFIDIA, EUBrazilCC, MARSOP4, IS-ENES2, INDIGO-DataCloud. Her research activities mainly focus on the distributed data management and data analytics, the design and the development of tools and software for climate scientists. From October to December 2016 she visited the Lawrence Livermore National Laboratory (LLNL), within the Program for Climate Model Diagnosis and Intercomparison (PCDMI) working in the context of the Earth System Grid Federation (ESGF). She leads one of the ESGF Working Groups on the data usage and demographics statistics of the ESGF federation.

Paola Nassisi received her Master's Degree, cum laude, in Computer Engineering from the University of Salento (Lecce) in 2011. In 2012 she joined the Advanced Scientific Computing division at the Euro-Mediterranean Center on Climate Change (CMCC). Since then, she was involved in national and international research projects like TESSA, OFIDIA and the EU FP7 IS-ENES2 project. Her research activities focus on visualization and data analytics within the Climate Change context with particular reference to the design, development and maintenance of scientific data archives as well as access and search&discovery services. Currently, she is collaborating for the realization of a distributed monitoring framework to collect and visualize data usage statistics about the Earth System Grid Federation infrastructure. She is actively involved in the H2020 INDIGO-Datacloud project for the development of a data analytics gateway for e-Science.

Iana A.A. Rufino has joined to Natural Resources Center of UFCG in 2006 where now she is a Full Professor. She holds a Ph.D. in Natural Resources at UFCG in 2004, a M.Sc. degree in Architecture and Urbanism at USP in 1996, and a Bachelor's degree in Civil Engineering at UFPB in 1994. She had developed her post-doctoral studies at the Center for Geospatial Technology/Texas Tech University in Lubbock, TX (USA) from 2012 to 2013. Nowadays, she teaches classes for undergraduate courses (Civil Engineering and Architecture and Urbanism) and for Graduate courses (Environmental and Civil Engineering; Natural Resources). She is also an Advisor and Researcher collaborating with many national and international projects. Her main interests are Geospatial Technologies applied to Environmental Studies in Urban and Regional Areas; Remote Sensing supporting Land Use/Land Cover Changes studies in semi-arid regions.

Arie C. Seijmonsbergen holds an MSc (1988) and PhD degree (1992) in Physical Geography from the University of Amsterdam (UvA) in The Netherlands. Currently he is assistant professor in the Theoretical Computational Ecology (TCE) department within the Institute for Biodiversity and Ecosystem Dynamics (IBED) at the UvA. His research is focused on the functioning of Geo-Ecosystems by analyzing the 3D structure of both landscape and the vegetation using air-born and terrestrial LiDAR-based high resolution elevation data as well as geodiversity mapping at multiple scales.

Niels Anders received a M.Sc. degree in Computational Bio- and Physical Geography (2008) and PhD degree in Earth Sciences (2013) from the University of Amsterdam, The Netherlands. His passion lies in connecting state-of-the-art computing and data technology with geosciences, particularly related to alpine and Mediterranean landscape development. After two postdoctoral fellowships at Wageningen University & Research, The Netherlands and University of Amsterdam he is currently primarily active as software developer and geodata analyst at Geodan BV, Amsterdam

Carlos de O. Galvão is an Associate Professor in Water Resources Engineering at the University of Campina Grande, Brazil. He has been working for over 30 years on water management, particularly in semi-arid context. He chairs the Water Resources Team of the Brazilian Research Network on Global Climate Change, and is a member of the Leadership Teams of the Committees on Water Management and Climate Change of the International Association of Hydro-Environment Engineering and Research (IAHR). He is also a member of IEEE, of the American Geophysical Union (AGU) and International Association of Hydrological Sciences (IAHS).

John de B.L. Cunha holds a M.Sc. in Environmental and Civil Engineering at the Federal University of Campina Grande (UFCG). He is a Professor at the Semiarid Sustainable Development Center of the UFCG since 2012. His main research interests focus on the application of remote sensing techniques for energy and water fluxes and balance, and land cover change studies in the Brazilian Semiarid. John has been involved in several international research projects with Africa, Europe and USA institutions in the frame of his research in the Brazilian Semiarid. During the past 6 months, John has been a visiting researcher at the Forest Research Center of the University of Lisbon (Portugal) and at the Department of Cartography, GIS and Remote Department of the University of Goettingen (Germany)

Miguel Caballer obtained the B.Sc. and M.Sc degree in Computer Science from the Universitat Politecnica de Valencia (UPV), Spain, in 2000 and 2012. He is member of the Grid and High Performance Computing group of the Research institute of Instrumentation for Molecular Imaging (I3M) since 2001. He has participated in several European and National research projects about the application of Parallel, Grid and Cloud computing techniques to several areas of engineering. Other fields of interest include Green Computing.

Mariane S. Sousa-Baena is a plant biologist who holds a M.Sc. degree in Plant Anatomy (2005) from the University of So Paulo (Brazil) and a Ph.D. degree in Plant Systematics and Evolution from University of So Paulo and the University of California Davis (UCDavis). Her main research interests include development and evolution of Neotropical plants. She worked at the Centro de Referncia em Informao Ambiental (CRIA) from 2012 to 2015 analyzing online biodiversity primary data for Brazilian plant species to investigate the status of knowledge of the Brazilian flora. She just joined the GaTE Lab (Genomics and Transposable Elements) at University of So Paulo and her current project is in the field of Evolutionary Developmental Biology.

Vanderlei P. Canhos is the Director of the Centro de Referncia em Informao Ambiental CRIA, a non-governmental organization established with the mission to make biodiversity data freely and openly available via Internet. Holds a Ph. D. degree in Food Science and Technology from Oregon State University (1980). His early work at the University of Campinas (UNICAMP) lead to the development of the Brazilian network of microbial culture collections and biobanks. As a member of the Organization for Economic Cooperation and Development (OECD) Biological Resource Centers Task Force (20012006), he contributed to the development of the Best Practices Guidelines for the Operation and Management of Biological Resources Centers. Leader of early biodiversity data infrastructure projects, including the speciesLink network and openModeller, his current work is focused on strategies for the sustainable consolidation of biodiverity data infrastructures. Former member of the Board of Directors of Species2000 /Catalogue of Life, and Board of Directors of ETI Bioinformatics (University of Amsterdam) and the Advisory Board of Global Research Data Infrastructures (GRDI 2020), he has been involved in various European projects from FP7 and Horizon 2020 programme.

Giovanni Aloisio is Full professor of Information Processing Systems at the Dept. of Innovation Engineering of the University of Salento, Lecce, Italy, where he leads the HPC laboratory. Former director of the ''Advanced Scientific Computing'' (ASC) Division at the Euro-Mediterranean Center on Climate Change (CMCC), he is now a member of the CMCC Strategic Council and Director of the CMCC Supercomputing Center. His expertise concerns high performance computing, grid & cloud computing and distributed data management. He has been involved into several EU projects such as GridLab, EGEE, IS-ENES, EESI, EXDCI, EU-BrazilCC, INDIGO-DataCloud, OFIDIA, EUBra-BIGSEA. He has also contributed to the IESP (International Exascale Software Project) exascale roadmap. He has been the chair of the European panel of experts on WCES that has contributed to the PRACE strategic document ''The Scientific Case for HPC in Europe 2015-2020''. He is a member of the ENES HPC Task Force. He is the author of more than 250 papers in referred journals on high performance computing, grid & cloud computing and distributed data management.