A Science Gateway for Biodiversity and Climate Change Research

Donatello Elia*, Alessandra Nuzzo*, Paola Nassisi*, Sandro Fiore*, Ignacio Blanquer†, Francisco V. Brasileiro‡, Iana A. A. Rufino‡, Arie C. Seijmonsbergen§, Niels S. Anders§, Carlos de O. Galvão‡, John E. de B. L. Cunha‡, Mariane de Sousa-Baena¶, Vanderlei P. Canhos¶ and Giovanni Aloisio*‖
*Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici, Lecce, Italy
†Universitat Politecnica de Valencia, Valencia, Spain
‡Universidade Federal de Campina Grande, Campina Grande, PB, Brasil
§IBED, University of Amsterdam, Amsterdam, Netherlands
¶Centro de Referência em Informação Ambiental, Campinas, SP, Brasil
‖University of Salento, Lecce, Italy

***Abstract*—Climate and biodiversity systems are closely interlaced 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, tools and a large variety of heterogeneous, distributed data sources. In this regard, the EUBrazilCloudConnect project provides a user-centric research environment built on top of a federated cloud infrastructure across Europe and Brazil to serve scientific needs. One of the test cases implemented in this project focuses on climate change and biodiversity research. The BioClimate is the Science Gateway of the use case. It aims at providing end-users with a highly integrated environment, addressing mainly data analytics requirements. This paper presents a complete overview about BioClimate and the scientific environment delivered to the user community at the end of the project.**

***Keywords*—Science Gateways, Scientific Data Management and Analytics, Environmental Sciences.**

I. Introduction

Climate and biodiversity systems are closely interlaced across a wide range of scales. In order 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.

Direct measurements of climate and biodiversity are often difficult and time-consuming to obtain, instead it is common practice to use climate and biodiversity indicators. These interactions can be studied at various scales, ranging from microscopic scales, and at (genomic, taxonomic, ecosystem) scales of individual plant and animal species. A multi-scale and integrated approach is required to investigate the climate-biodiversity system as a whole. Presently, in this scenario, researchers and professionals are burdened by scattered data sources, wealth of analysis tools to master and implement, and computational limitations to upscale their analysis.

EUBrazilCloudConnect [1] is a project from the third coordinated EU-Brazil call. It is a preliminary step towards providing a user-centric environment for the scientific research communities to test the execution of challenging applications exploiting a federated cloud infrastructure. The project addresses the scientific challenges of three multidisciplinary and highly complementary scenarios, among which the one on biodiversity, natural resources and climate change represents the most challenging one from the scientific data management standpoint. 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 visualisation.

The Biodiversity and Climate Change use case (BioClimate) involves multiple heterogeneous data sources (e.g. SEBAL, LiDAR, CRU, CMIP5, speciesLink, GBIF, etc.) and several processing pipelines, integrated through the BioClimate Scientific Gateway. The gateway sits on top of the databases and enables near-real-time analysis of large volume datasets (from multi-GBs to multi-TBs scale depending on the specific data source) through the Parallel Data Analysis Service (PDAS). PDAS clusters are deployed on the site where the databases are stored providing the end-user with a high-level, parallel, and server-side interface for scientific data analysis.

The design of the software infrastructure and the BioClimate Scientific Gateway for end-users facilitates joint research using data that is otherwise difficult to access or for which availability is fragmented and/or too large to process using traditional computational means. With regard to existing approaches and tools that are mainly client-side/desktop based, the use case delivers a well-integrated environment for climate change and biodiversity research with cloud-based infrastructure and server-side capabilities.

This work presents the BioClimate Scientific Gateway, the scientific challenges addressed and the implementation details. The remainder of this work is organised as it follows. Section II provides an overview of the BioClimate use case and its main goals. Section III provides a general description of the BioClimate Scientific Gateway architecture, whereas Section IV and Section V give, respectively, a detailed description of the graphic interface and the back-end. Finally, Section VI draws the main conclusions and describes the future activities.

II. A Biodiversity & Climate Change Use Case

The EUBrazilCloudConnect (EUBrazilCC) use case on climate change and biodiversity is a data-driven use case, aiming at better understanding the interactions between the biodiversity system and the climate system. This use case focuses on bringing together a wide variety of climate and biodiversity data and analysis tools into a user-friendly and web-based Science Gateway to provide an integrated approach of investigating climate and biodiversity across different temporal and spatial scales.

To address all these scientific challenges, the use case joins together heterogeneous data sources, on-premises cloud infrastructures, multiple data services, and a Science Gateway into a single, federated trans-Atlantic environment.

The Science Gateway provides access to historical temperature and precipitation records, different climate model scenarios with predictions of future temperature and precipitation, Landsat [2] satellite imagery for climate and biodiversity indicators, LiDAR 3D forest metrics and biodiversity indicators at a very high resolution, and plant occurrences data for ecological niche models for the prediction of future plant distribution based on different climate scenarios. The proposed pipelines/ workflows combine the analysis of data acquired from these different technologies 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 analysis of remote sensing images provides 3D information concerning the structure of the vegetation, which improves biodiversity indicators such as the energy balance and evapotranspiration.

The EUBrazilCC infrastructure provides the computing power needed to support data processing and analysis, the management of metadata to enable search and discovery as well as provenance management to address re-usability and reproducibility, both strongly relevant for scientific data environments. The BioClimate Scientific Gateway integrates in a web-based environment the data sources and the processing and analysis capabilities exploiting the project infrastructure. More specifically, the gateway has been designed to fulfil some key requirements:

- *Integration of heterogenous data sources*. The gateway provides a unified interface to access and process satellite images (from Landsat), environmental data, future climate scenarios, biodiversity data like species distributions and LiDAR datasets related to some target areas. Furthermore, the gateway provides also metadata information describing these data sources.
- *Implementation of processing tools*. To support data analysis, several tools are integrated in the gateway to allow: computation of 3D vegetation products based on LiDAR data [3] (e.g. Digital Surface Model (DSM), Digital Terrain Model (DTM), Canopy Height Model (CHM), Relative Height at 50% (RH50)), execution of Ecological Niche Modeling over species data and processing of datasets from climate models and the SEBAL algorithm.

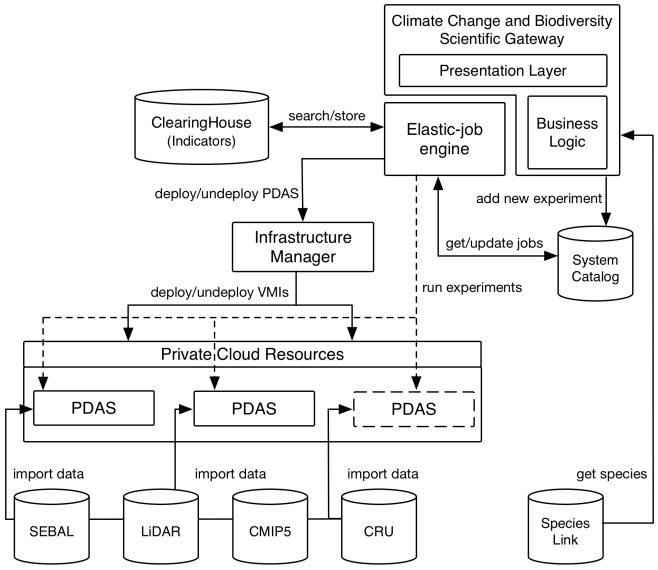

Fig. 1. BioClimate high-level use case architecture

- *Usability*. The interface is designed to: (i) facilitate the end-user to select the target data source, an area of interest and the temporal scale; (ii) submit an experiment computation; (iii) visualise the processed results in terms of maps, graphs, tables and comparative charts; and (iv) download the aggregated results and products regarding satellite images and 3D vegetation products (CSV, Raster, GeoTIFF and PNG formats).

III. Gateway Architecture

The software architecture of the use case is shown in Figure 1. The BioClimate Scientific Gateway represents the high-level user interface provided by the use case. It allows data access, analysis and visualisation over multiple, heterogeneous data sources, by exposing an integrated view of the data level. It supports several features, such as time-series and statistical analysis, data inspection, intercomparison and subsetting.

The elastic-job engine takes care of the execution of the requests submitted through the gateway interface by translating the requests in PDAS tasks and then properly scheduling the jobs on the available resources. To guarantee scalability, it elastically adapts to the analytics workload exploiting the underlying cloud resources. The engine interacts with the Infrastructure Manager (IM) [4] to deploy and un-deploy PDAS cluster instances on-demand. A detailed description of the implementation and the main features of both the Science Gateway interface and the engine is provided in the next sections.

A system catalog is used by both the front-end and the back-end to store useful information regarding user management, experiment execution requests and results, PDAS cluster usage history and it also serves as a centralised data repository.

The PDAS, a core component of the Ophidia project [5], [6], provides support in terms of data analytics applied to large scientific datasets. It includes functionalities to deal with different

scientific data formats, such as NetCDF (Network Common Data Form) [7] and satellite data, and allow mathematical and statistical operations on this data. Python scripts, integrated in the PDAS, provide additional functionalities to process LiDAR products and interact with external tools (e.g GDAL [8]) and services (e.g OpenModeller [9]).

The gateway also provides access to the BioClimate Clearing House, a database where the user can persistently store the results of the experiment run during a session and retrieve them through the search functionalities.

The lowest layer of the diagram comprises the several private clouds, running OpenNebula or OpenStack at the Infrastructure as a Service (IaaS) level, and the data sources, made available by the project partners or already available from national and international agencies, which are part of the infrastructure with a more static setup.

The data sources integrated through the gateway are reported in the following:

- *SEBAL datasets*. These are an output of satellite images series (Landsat) processed by the SEBAL [10], [11] algorithm to produce estimates of energy balance and evapotranspiration of water to the atmosphere. 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 Semiarid region.
- *LiDAR data*. For the areas near Manaus in Brazil, where hyper-spectral imagery is apparently absent, EUBrazil Cloud Connect will leverage of the available LiDAR data provided by EMBRAPA [12] (Brazilian Agricultural and Livestock Research Corporation). Vegetation and terrain metrics represent the key indicators that can be inferred from these datasets.
- *Biodiversity data sources*. The speciesLink datasets [13], provided by CRIA, the Reference Center on Environmental Information, are an output of networking activities to provide free and open access to 7.3 million primary research-grade data, derived from the federation of 350 Brazilian biodiversity datasets, gathered from 150 institutions in Brazil and abroad. They represent valuable biodiversity data sources.
- *Climate data from the CMIP5 Federated Data Archive (ESGF)* [14]. 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, NetCDF format, CF conventions. Starting from these datasets, multiple climate indicators can be computed.
- *Climate data from observed data*. These high-resolution gridded datasets (CRU TS v.3.23 [15]) provide monthly values for several variables, such as temperature and precipitation, for an historical time period and are made available under the Open Database License by Climatic Research Unit, University of East Anglia.

Finally, security cuts across the whole architecture and is taken into account at several levels. With regard to the front-end, the 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, based on a Java implementation of a Cryptographically Secure Pseudo-Random Number Generator, called Password-Based Key Derivation Function 2 (PBKDF2) [16]. 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 can exploit the X509v3 digital certificates-based authentication and the VOMS-based authorisation. 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 GSI/VOMS enabled interface, supporting both X.509 certificates and VOMS-based authorisation and addressing the interoperability with the EGI Fed Cloud environment [17], has been defined.

IV. User interface insights

In order to address portability of the system and the separation of concerns between the presentation layer and the business logic, the gateway has been implemented according to the Model-View-Controller pattern.

The presentation layer, running on the client side (i.e. a browser), provides a rich user interface to submit the data analysis tasks and visualise their results. It is implemented as a JavaScript web application based on the ExtJS library [18], which offers a number of gadgets such as panels, charts and grids, and Google Maps API [19] for the visualisation of georeferenced 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 is based on Java and Apache Struts2 framework [20].

To increase the performance and make the output visualisation faster, it has been decided to perform the heavier tasks, related to the post-processing of the outputs, on the server side and to present the ready-to-use result to the JavaScript library on the presentation layer.

Usability has been addressed by defining and implementing a set of pre-defined experiments regarding the different data sources and type of analysis. Each experiment defines a customisable template to perform data analytics tasks on climate and biodiversity data and requires a specific pipeline of operations, including subsetting, data reduction and mathematical/statistical functions.

The following subsections provide a description of the main views and interfaces made available by the gateway.

A. Interactive analysis

The ”Interactive analysis” panel allows a real-time, exploratory analysis of time series from the climate data available

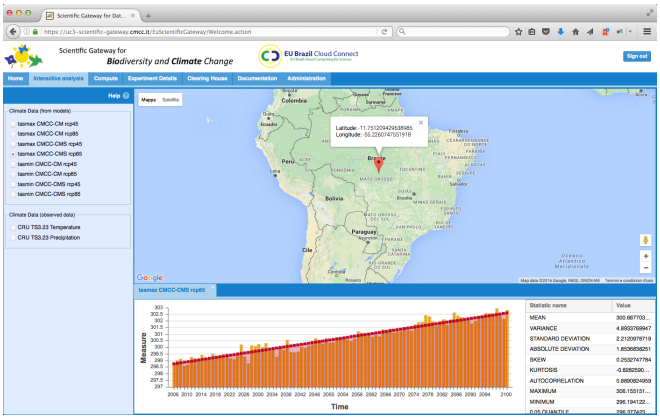

Fig. 2. Interactive analysis

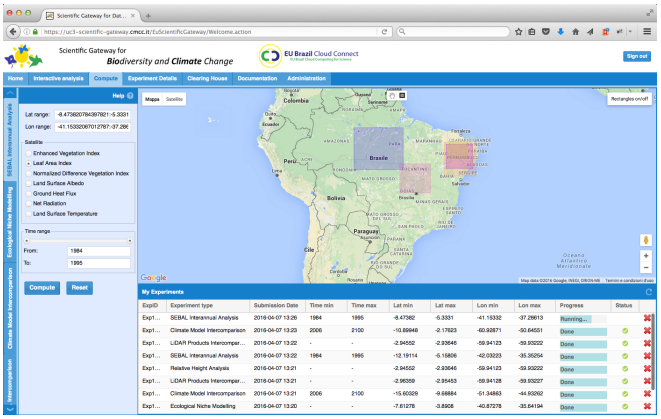

Fig. 3. SEBAL Interannual analysis compute interface

in the use case. In particular, it provides access to 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 Figure 2, the interface allows 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 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 statistics.

B. Batch analysis

The ”Compute” panel provides the features to define and submit complex experiments regarding the available data sources. For each experiment, a map for spatial selection and a form to set the input parameters is provided. The following experiments are defined:

- *Interannual analysis of SEBAL output* (see Figure 3) provides information about interannual trends and statistical information of a specific SEBAL variable. The Science Gateway integrates data processed by the SEBAL algorithm and provides functionalities to analyse several variables produced by this algorithm (e.g. Enhanced Vegetation Index, Leaf Area Index, Normalized Difference Vegetation index, etc.). The interface allows both spatial and temporal selection.
- *Climate and SEBAL variables intercomparison* allows the comparison of the behaviour of climate and SEBAL variables. In particular it supports analysis over the variables produced by the SEBAL algorithm and variables (precipitation and temperature) from historical climate data. From a scientific point of view, this experiment provides useful information about the relationship between climate and vegetation indices.
- *Climate indices intercomparison* allows comparison of indicators computed on CMIP5 datasets belonging to different climate models and future emission scenarios (RCP4.5 and RCP8.5 [21]). Four well-known indicators based on maximum and minimum temperature are available for comparison (i.e. TXx, TNx, TXn, TNn [22]).
- *Ecological Niche Modelling* (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 and future pessimistic). The models are created with the maximum entropy algorithm [23] and are based on the species occurrences selected by the user.
- *LiDAR products intercomparison* allows comparison and evaluation of the statistical relationship between LiDAR products available through the gateway (e.g. DSM, DTM, CHM). In this case, a LiDAR tile can be selected from the map.
- *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.

Fig. 4. SEBAL Interannual analysis details interface

C. Experiment visualisation & download

Once the computation of the experiment is completed, details about the experiment are available through the ”Experiment Details” section. Figure 4 displays the output produced

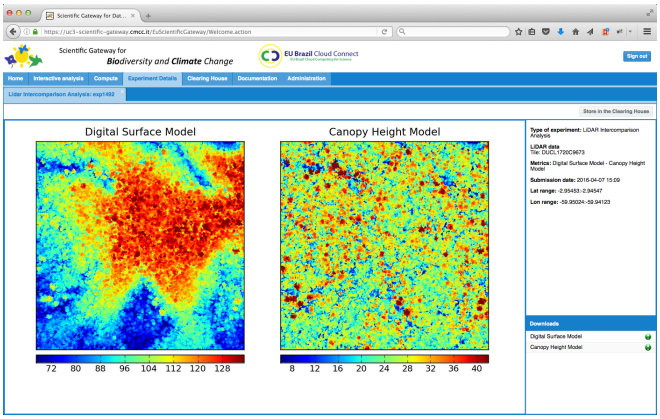

Fig. 5. LiDAR intercomparison details interface

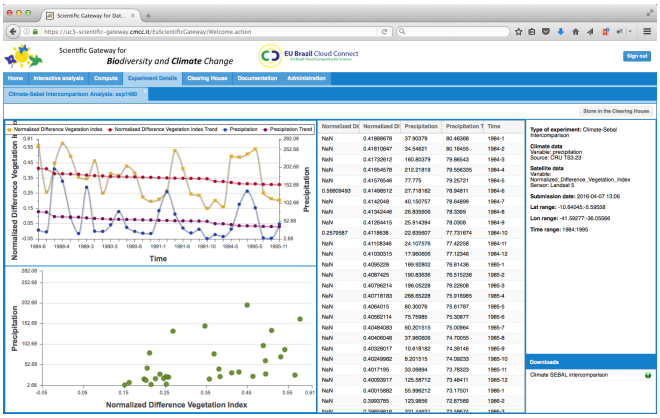

Fig. 6. Climate-SEBAL intercomparison details interface

by a SEBAL interannual experiment, whereas Figure 5 and Figure 6 display the output produced by a LiDAR intercomparison experiment and Climate-SEBAL intercomparison experiment respectively.

In particular, to better suit the experiment peculiarities, a specific detail view is provided for each experiment defined above. Hence, various gadgets organised 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 histograms of the point distribution.

Most of the information provided through the gadgets is also available for download in CSV, raster, GeoTIFF or PNG format, depending on the type of experiment run. Furthermore, metadata regarding the experiment is available in the same view.

D. BioClimate Clearing House

The BioClimate Clearing House system allows users to store a relevant experiment, run during a session, for future analysis. A smart search feature is available to filter out the experiments saved into the Clearing House, based on: (i) spatial domain used for the experiment, (ii) experiment type and (iii) submission date.

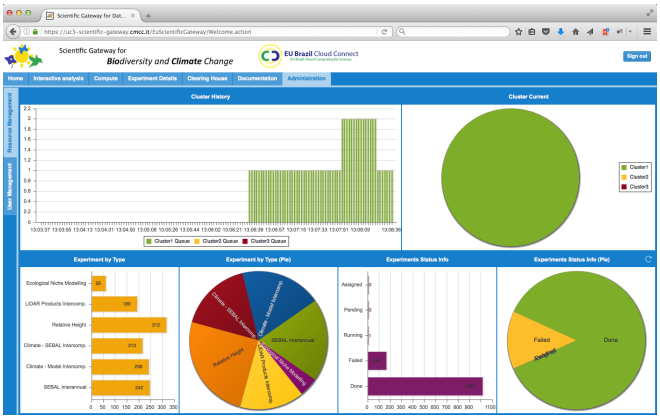

Fig. 7. Monitoring Dashboard

E. Infrastructure Monitoring

The BioClimate Scientific Gateway includes two administrative interfaces that (i) allow managing users and their privileges and (ii) provide some information about the resources exploited dynamically by the gateway (i.e. PDAS cluster instances) as well as some statistics regarding the number of experiments executed in terms of their type and status. Through this dashboard (see Figure 7) it is possible to get some insights about the use of the system by the end-users. The charts mainly provide real-time monitoring information regarding the number of experiments running/pending and the status of the resources. In particular, a histogram shows the set of experiments and their distribution across the active PDAS instances for the last couple of minutes, whereas a pie chart shows the set of clusters currently running the experiments.

V. Elastic-job engine

The elastic-job engine is designed to guarantee fast processing of the user requests by exploiting dynamically and elastically the federated cloud infrastructure. To meet scalability and performance requirements, the engine is implemented as multi-threaded daemon, based on GNU C libraries, that exploits the PDAS capabilities to perform pipelines of analytics tasks.

Data-driven processing pipelines, based on PDAS operators, have been defined integrating different tools, services and data formats.

Management of the workload is performed exploiting a smart scheduling algorithm, which provides dynamic job scheduling over a set of queues. A job queue is associated to each PDAS cluster running on the infrastructure. To horizontally scale on the workload, a new PDAS instance is deployed automatically on the private cloud resources when the number of pending jobs on all the queues exceed a configurable threshold. A more detailed description of the automated cloud deployment (through the elastic-job engine) of the PDAS, as well as of the queue policy adopted and its rationale, are out of the scope of this paper and can be found in [24].

A. PDAS

As mentioned before, the PDAS provides the capabilities to perform data analytics on large scientific datasets and includes a set of libraries able to deal with different data formats. In the EUBrazilCC project, the PDAS addresses scientific challenges related to the BioClimate use case and it is used for, both batch and interactive data analysis on NetCDF, LiDAR and remote sensing data.

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

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. In particular, regarding the interactive analysis, an operator that allows data inspection and on-the-fly exploration of time series has been implemented, whereas to run the batch experiments, processing pipelines made up of several new operators and functions have been defined. To integrate external tools, an operator to run scripts has also been developed. Besides the previous extensions, the import process has also been optimised to reduce the time required to import large-scale datasets such as SEBAL output data. Finally, to automate the deployment of PDAS instances in the EUBrazilCC federated infrastructure, some cloud-based scenarios, based on RADL files [4], have been implemented as reported in detail in [24].

VI. Conclusion

During the final validation phase of the EUBrazilCC project, the BioClimate use case was highly appreciated by the end-users, due to its ability to provide and deliver in the same environment tools, pipelines, analysis/visualisation features, and several data sources in an integrated manner.

User experience was good and the change of paradigm (process the data on the server-side) was evaluated as the key added value. Despite it requires a learning process, the BioClimate Scientific Gateway provides multiple views and analyses of the retrospective data gathered. High level user experience and usability have been two key requirements considered in the implementation phase.

A lot of interest was also raised by governmental & environmental agencies (both research & education) especially in Brazil. A set of follow-up actions will be put in place from the different partners even beyond the project lifetime (that was part of the project sustainability plan).

Finally, the impact on the user community was very high. The gateway was evaluated as seamlessly, flexibly and efficiently able to integrate a comprehensive and useful set of scientific data management tools to increase the mutual understanding between climate change and biodiversity.

Acknowledgment

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

References

[1] EUBrazilCC. [Online]. Available: http://eubrazilcloudconnect.eu
[2] The landsat program. [Online]. Available: http://landsat.gsfc.nasa.gov/
[3] M. A. Lefsky, W. B. Cohen, G. G. Parker, and D. J. Harding, "Lidar remote sensing for ecosystem studies," *BioScience*, vol. 52, no. 1, pp. 19–30, 2002. [Online]. Available: http://bioscience.oxfordjournals.org/content/52/1/19.short
[4] M. Caballer, I. Blanquer, G. Moltó, and C. Alfonso, "Dynamic management of virtual infrastructures," *Journal of Grid Computing*, vol. 13, no. 1, pp. 53–70, 2014.
[5] S. Fiore, A. D'Anca, C. Palazzo, I. T. Foster, D. N. Williams, and 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.
[6] S. Fiore, C. Palazzo, A. D'Anca, I. Foster, D. N. Williams, and G. Aloisio, "A big data analytics framework for scientific data management," in *Big Data, 2013 IEEE International Conference on*, Oct 2013, pp. 1–8.
[7] R. K. Rew and 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.
[8] Gdal library. [Online]. Available: http://www.gdal.org/
[9] M. E. Souza Muñoz, R. Giovanni, M. F. Siqueira, T. Sutton, P. Brewer, R. S. Pereira, D. A. L. Canhos, and V. P. Canhos, "openmodeller: a generic approach to species' potential distribution modelling," *GeoInformatica*, vol. 15, no. 1, pp. 111–135, 2009.
[10] W. Bastiaanssen, M. Menenti, R. Feddes, and A. Holtslag, "A remote sensing surface energy balance algorithm for land (sebal). 1. formulation," *Journal of hydrology*, vol. 212, pp. 198–212, 1998.
[11] W. Bastiaanssen, H. Pelgrum, J. Wang, Y. Ma, J. Moreno, G. Roerink, and T. Van der Wal, "A remote sensing surface energy balance algorithm for land (sebal).: Part 2: Validation," *Journal of hydrology*, vol. 212, pp. 213–229, 1998.
[12] Embrapa. [Online]. Available: https://www.embrapa.br/
[13] C. Centro de Referencia em Informacao Ambiental. Specieslink service. [Online]. Available: http://splink.cria.org.br/
[14] K. E. Taylor, R. J. Stouffer, and G. A. Meehl, "An overview of cmip5 and the experiment design," *Bulletin of the American Meteorological Society*, vol. 93, no. 4, pp. 485–498, 2012.
[15] I. Harris, P. Jones, T. Osborn, and D. Lister, "Updated high-resolution grids of monthly climatic observations - the cru ts3.10 dataset," *International Journal of Climatology*, vol. 34, no. 3, pp. 623–642, 2014.
[16] B. Kaliski, "Pkcs #5: Password-based cryptography specification version 2.0," RFC 2898, Sep. 2000. [Online]. Available: http://tools.ietf.org/html/rfc2898
[17] Egi fedcloud. [Online]. Available: http://www.egi.eu/infrastructure/cloud/
[18] Extjs library. [Online]. Available: http://docs.sencha.com/extjs/
[19] Google maps api. [Online]. Available: https://developers.google.com/maps/
[20] Apache struts2 framework. [Online]. Available: https://struts.apache.org/
[21] Rcp emission scenarios. [Online]. Available: http://www.wmo.int/pages/themes/climate/emission_scenarios.php
[22] Climate change indices. definitions of the 27 core indices. [Online]. Available: http://etccdi.pacificclimate.org/list_27_indices.shtml
[23] Maximum entropy algorithm. [Online]. Available: http://openmodeller.sourceforge.net/algorithms/maxent.html
[24] S. Fiore et al., "Big data analytics for climate change and biodiversity in the eubrazilcc 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.