IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

EnviroGRIDS Special Issue on "Building a Regional Observation System in the Black Sea Catchment"

ISSN 2156-5570(Online)
ISSN 2158-107X(Print)

www.ijacsa.thesai.org

Editorial Board

Prof. Dorian Gorgan, Technical University of Cluj-Napoca

Dorian Gorgan is professor and PhD supervisor in Computer Science and Information Technology at the Technical University of Cluj-Napoca, and the chair of the CGIS (Computer Graphics and Interactive Systems) Laboratory. His fields of interest concern with parallel and distributed processing over Grid and Cloud infrastructures, high performance graphical processing and visualization, distributed interactive applications, and Earth Science oriented applications and tools development. He was ERCIM fellow at the Rutherford Appleton Laboratory, UK, in the field of graphical interaction and multimedia systems. He was the initiator and the director of the MedioGrid project, a national Grid computing infrastructure for academic education and scientific research. He was involved as scientific coordinator and WP leader in international research projects such as enviroGRIDS, SEE-GRID-SCI, GiSHEO, mEducator, IASON, and I-TRACE. He is member of scientific and reviewing committees of many journals and international conferences in Computer Science and Information Technology.
Dorian Gorgan's homepage: http://users.utcluj.ro/~gorgan/

Dr. Gregory Giuliani, University of Geneva

Dr. Gregory Giuliani, Scientific collaborator at University of Geneva is a geologist and environment scientist who specialized in Geographical Information Systems (GIS) analyses and Spatial Data Infrastructures (SDI). After obtaining a degree in Earth Sciences, he went on to complete a master and a PhD in Environmental Sciences, specializing in remote sensing, GIS, and SDI. He previously worked as a GIS Consultant for the World Health Organization, as a University tutor in remote sensing and GIS and as a GIS Developer in a local Swiss GIS company. He also works at UNEP/GRID-Geneva since 2001 and is the focal point for SDI. He is Work package leader in the FP7 enviroGRIDS project and the FP7 AFROMAISON project where he coordinates SDI development and implementation. Project manager of the FP7 EOPOWER project, he also participated in the FP7 ACQWA project. At GRID-Geneva, he is the lead developer of the PREVIEW global risk data platform (http://preview.grid.unep.ch). Participate and contribute actively to various Global Earth Observation System of Systems (GEOSS) activities.
Gregory Giuliani's homepage: http://www.unige.ch/envirospace/People/giuliani.html

Special Issue Reviewers

- Serena Coetzee (Uni. Pretoria)
- Carlos Granell (JRC)
- Eric Grosso (IGN)
- Joan Maso (UAB)
- Ioan Lucian Muntean (UTCN)
- Antonio Parodi (Uni. Genoa)
- Dana Petcu (UVT)
- Hans van der Kwast (UNESCO-IHE)
- Florin Pop (UPB)

CONTENTS

Editorial

EnviroGRIDS Special Issue on "Building a Regional Observation System in the Black Sea Catchment"

Gregory Giuliani[1,2]
[1]Institute for Environmental Sciences, enviroSPACE
University of Geneva
1227 Carouge, Switzerland
[2]United Nations Environment Programme
Global Resource Information Database
1211 Châtelaine, Switzerland
gregory.giuliani@unige.ch

Dorian Gorgan
Computer Science Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
dorian.gorgan@cs.utcluj.ro

***Abstract*—The Black Sea Catchment Observation System has been developed in the frame of the EU/FP7 enviroGRIDS project to inform about crucial regional environmental issues. This system is now making resources accessible to a large community of users for data management and publishing, for hydrological models calibration and execution, for satellite image processing, for report generation and visualization, and for decision support. In this special issue, we present the different components that were developed as well as the encountered challenges in order to bring innovative contributions into the Global Earth Observation System of Systems. One of the major issues was to enable data exchange across different heterogeneous components and infrastructures, more specifically Spatial Data and Grid infrastructures. The interoperability standards proposed by the Open Geospatial Consortium (OGC) support the scalability and the efficient combination of the complex specialized functionalities and the computation potential of these platforms. Another important issue was to build the human, institutional and infrastructure capacities to contribute and use this new observation system.**

Keywords—enviroGRIDS; Observation System; Spatial Data Infrastructure; Grid computing; Black Sea; Remote sensing; Hydrological modeling; GEOSS

I. INTRODUCTION

Earth Observation specialists recognize that the lack of systematic monitoring and access to reliable time-series on environmental, statistical, and socio-economical data are a major barrier to effective and efficient informed policy- and decision-making [1]. This problem has been recently reinforced by several EU funded projects related to water. They are all highlighting discrepancies between the objectives of guiding policy, and limited access and availability of data [2]. Policy-relevant researchers and end users are still facing the problem of timely access and exchange of needed data.

Supported by the latest technological advances in Earth Observation and Web technologies, Spatial Data Infrastructures (SDIs) have been developed and implemented at accelerated pace at regional and national levels, with the long-term vision of creating global and regional SDIs. The benefits of SDI have been analyzed and reported extensively, as they allow for trans-sectorial and trans-national sharing of- and access to geospatial data, and their assimilation (consumption) in novel and inventive software applications that can provide wide range of social, economic and environmental benefits.

For achieving these purposes, SDIs provide a suite of services for data publishing, discovery, gathering and integration, which enable interoperability of the different components involved [3]. Therefore, the concept of SDI was developed to facilitate and coordinate the exchange and sharing of geospatial data, encompassing data sources, systems, network linkages, standards and institutional issues involved in delivering geospatial and information from many different sources to the widest possible group of potential users. The vision of an SDI incorporates different databases, ranging from the local to the national, into an integrated information highway and constitutes a framework, needed by a community, in order to make effective use of geospatial data.

Climate change is a worldwide concern that is affecting many areas of human activities. The last report of the Intergovernmental Panel on Climate Change [4-6] predicts important changes in the coming decades that will not only modify climate patterns in terms of temperature and rainfall, but will also drastically change freshwater resources qualitatively and quantitatively, leading to more floods or droughts in different regions, lower drinking water quality, increased risk of water-borne diseases, or irrigation problems. These changes may trigger socio-economic crises across the globe that need to be addressed well in advance of the events in order to reduce the associated risks. Consequently water resources are particularly sensitive to climate change and human activities.

Water is a fundamental natural resource and critical for the well being of individuals in terms of health, agriculture, energy production, ecosystem services and economic development. However, water resources are increasingly under pressure causing a shift in balance between demand and supply, and having a negative influence on its quality [7]. Effective and efficient water management requires coordination of actions, one of them being access and provision of reliable data and information (e.g., state of the resources, changes, pressures) and the capacities to interpret correctly and meaningfully these information [8, 9].

Water management and hydrological modeling, due to their interdisciplinary nature and being complex and dynamic systems, intrinsically ask for better integration of data, information and models [10-12]. The aim is to bring to policy/decision-makers suitable and reliable information through efficient scientific tools and models. .

Beniston et al. [2] have reported that researchers in climate and water sciences are regularly facing the problem of searching, finding, and accessing data. These authors have highlighted several barriers that are impending a timely and efficient usage of water-related data. In particular, incomplete and non-standardized time series data are an important issue obliging scientists to spend a lot of their time in data gathering and harmonization [13]. Moreover, these data are often redundant because of the lack of coordination between providers. This situation leads to the fragmentation of repositories [14], making them difficult to find even if they are available. Furthermore, data are not or only poorly documented by their metadata and users cannot evaluate if they fit their purposes. Searching and downloading interfaces are often complex and difficult to understand for non-experts. Therefore, facilitating the exchange and access to water-related data is essential to easily integrate them with other distributed data sources [2]. This requires implementing commonly agreed standards, in particular, documenting data with standardized metadata, and making them searchable through catalogs.

Interoperability is needed to develop an open science framework allowing scientists and researchers to publish, discover, evaluate and access data. Current technologies are suitable to match these requirements only if open software interfaces and standards are developed allowing these technologies to interoperate on a global scale [15]. The Open Geospatial Consortium (OGC) aims at providing such standards enabling communication and exchange of information between systems operated with different software. Indeed, a non-interoperable system cannot share data and computing resources, inducing scientists to spend much more time than necessary on data discovery and transformations. One of the major benefits of interoperability is to enable locally managed and distributed heterogeneous systems (e.g., different operating systems, databases, data formats) to exchange data and provide services [16].

Moreover, with the emergence of technologies (e.g., Web Services, Web 2.0) and the greater affordability of digital devices, we are currently seeing a deluge of data in quantity and diversity (e.g., real-time data, archived data, crowd-sourced data, high-resolution data) [17]. This poses new challenges and offers new opportunities to turn this huge amount of data into understandable information. Consequently, efficient processing solutions are required, and distributed high performance computing infrastructures appear as promising solutions [18-20]. Indeed, there is an increasing need for large computational power to answer the demand for high-resolution modeling. The associated activities of uncertainty and sensitivity analyses bring forward the integration requirements of different sources of geospatial data, which are provided by SDIs via diverse web services, together with other data within Grid or Cloud computing environments. Consequently, an important effort is currently made to improve hydrological modeling [21] on a shared SDI and Distributed Computing platform [7, 22-25].

To tackle all the previously mentioned issues, the intergovernmental Group on Earth Observations (GEO) is coordinating a worldwide effort the development of the Global Earth Observation System of Systems (GEOSS) on the basis of a 10-Year Implementation Plan until 2015 [26]. GEOSS is aiming at connecting already existing SDIs and Earth Observations infrastructures and thus will not create and/or store its own data. The GEOSS portal is foreseen to act as a gateway between producers of environmental data and end-users. The aim is to enhance the relevance of Earth observations for global issues, and to offer a public access to comprehensive, near-real time data, information and analyses on the environment on following nine Societal Benefits Areas: Disasters, Health, Energy, Climate, Water, Weather, Ecosystems, Agriculture, Biodiversity. The mechanisms for data and information sharing and dissemination are described in the 10-Year Implementation Plan Reference Document [26]. Participating members must endorse data sharing principles [27]: (1) There will be full and open exchange of data, metadata, and products shared within GEOSS, recognizing relevant international instruments and national policies and legislation. (2) All shared data, metadata, and products will be made available with minimum time delay and at minimum cost. (3) All shared data, metadata, and products being free of charge or no more than cost of reproduction will be encouraged for research and education. GEOSS is also advocating for an increased sharing of methods for modeling to transform data into useful information.

The Black Sea Catchment (2.2 mio. km^2, 24 countries, 160 million inhabitants) is affected by severe environmental degradations. In 1995, the sea itself was rated with the highest concerns in five out of seven environmental categories, making it the worst of any of the European seas [28]. The Danube River, the major Black Sea tributary, was described as following an *"ecologically unsustainable development and inadequate water resources management"* [29]. The problems are caused by different factors, such as: inadequate management of wastewater/solid waste, ecological unsustainable industrial activities, inadequate land management and improper agricultural practices. These are generating several direct consequences: pollution of surface/groundwater, eutrophication, and accelerated runoff /erosion. These consequences have, on the other hand, the following main effects: decline in quality of life, human health risks, degradation of biodiversity, economic decline, and reduced availability of water. Therefore, the Black Sea hydrological catchment represents a very interesting case study to test the capacity of integrating large data sets to assess vulnerability and sustainability issues related to freshwater resources as various scales.

The EU FP7 enviroGRIDS research project [1] aims at providing approaches for achieving data integration by developing a SDI for the whole Black Sea catchment that can

[1] http://indico.cern.ch/getFile.py/access?resId=0&materialId=0&confId=45555

be utilized by a SWAT (Soil Moisture Assessment Tool) hydrological model [30]. The goal of the integration is to enable the analysis of the impacts of future climate, development-induced land use and demographic changes on selected social benefit areas, such as water, agriculture, energy, health, disasters, ecosystems and biodiversity. The results are made available though the Black Sea Catchment Observation System[2] (BSCOS, fig.1). This system is a shared information system that operates on the boundary of scientific/technical partners, stakeholders and the public. It allows to discover, gather, store, distribute analyze, visualize and disseminate data on the environment with the aim of increasing the capacity of decision-makers and other interested stakeholders to use it for selecting the most relevant management options on a 50-year time horizon. In summary, enviroGRIDS aims[3] at building the capacity of scientist to assemble such a system in the Black Sea catchment, the capacity of decision-makers to use it, and the capacity of the general public to understand the important environmental, social and economic issues at stake.

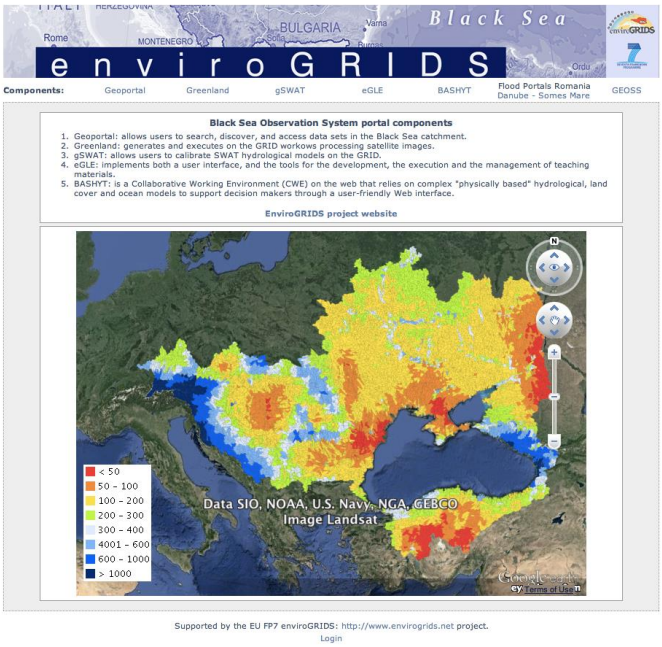


Fig. 1. The Black Sea Catchment Observation System

The objective of this Special Issue is to highlight the main contributions and to present the technical progresses made in building the Black Sea catchment regional observation system. The focus is on putting SDI into practice by improving data and metadata interoperability, by using new geoprocessing tools, by developing innovative geoportal solutions and by building capacity.

The first contribution [31] highlights issues and solutions used for the implementation of the BSCOS portal through heterogeneous technologies, typically SDI and distributed computing, the aim is to create and control the flow, processing, and visualization of spatial data for both the community of Earth Science specialists and Web users. The OGC standards support the scalability and the efficient combination of the complex specialized functionalities, as well as the computation potential of these platforms. These standards act as glue between the different components of the BSCOS portal that are presented in the different papers.

One of the main challenges currently faced is to convince and help regional data holders (like the World Data Center[4]) to make available their data and metadata in order to improve our capacity to assess the sustainability and vulnerability of the environment.. Giuliani et al. [32] present experiences and lessons learnt in the enviroGRIDS project for raising awareness and creating commitments on the benefits of data sharing using interoperable services.

The first component of the BSCOS, the Geoportal, allows to search, discover, view, and access data in the Black Sea catchment. However, setting up services is not sufficent, maintaining them, ensuring they are working correctly and they are offering good and reliable performances are also important. Charvat et al. [33] introduce an innovative approach through a quality check to ensure efficient discovery and access to data services based on OGC standards.

Remote sensing is an important source of Earth Observations data for understanding environmental issues. In a large area such as the Black Sea catchment, it is almost impossible to analyze high to medium resolution remotely sensed data on a single computer. Consequently, distributed computing appears as a promising solution to efficiently process an increasing volume of data. Moreover, a standardized access to these data is required in order to integrate raw and already-processed data in complex models and workflows. Three papers [34,35,36] are discussing these different issues by presenting solutions developed with the Greenland application. This web-based component allows to process large amount of remote sensing images over a Grid infrastructure and implementing OGC standards to access, process, and publish data. Additionally, a contributing paper from Balcik et al. [37] demonstrates the applicability and usefulness of the Greenland component in different case studies. In complement, an e-learning platform was developed to allow non-specialists to easily use computing resources, remote-sensing data in order to develop teaching and learning materials.

Assessing water sustainability and vulnerability of the Black Sea catchment in a global change framework requires to first develop spatially explicit scenarios of climatic, demographic and land cover changes that can serve as inputs for hydrological modeling. One of the software developed in enviroGRIDS is gSWAT for the calibration of SWAT hydrological models in a flexible environment that uses distributed computational infrastructures to speed-up the simulations [38]. SWAT models produce several useful outputs (e.g., evapotranspiration, soil moisture, aquifer recharge, river discharge) as text files. However, visualizing and publishing SWAT outputs as geospatial data is time consuming and repetitive. Moreover, data used and produced are often not interoperable and restricted to dedicated software impending an

[2] http://portal.envirogrids.net

[3] http://indico.cern.ch/getFile.py/access?resId=0&materialId=0&confId=45555

[4] http://wdc.org.ua

efficient use and integration of SWAT outputs with other sources and/or models. To tackle this issue, Giuliani et al. [39] are proposing the OSW4SWAT framework to facilitate SWAT outputs publishing and exchanging with other sources using OGC standards. In addition, Almoradie and Jonoski [40] present a first use-case in Romania using the recently adopted OGC standard WaterML2.0 to publish hydro-metrological time series to monitor and forecast floods.

Finally, to help decision-makers to take sound and informed decisions, the BASHYT component offers a set of web-based components to predict the effect of management decisions on water, sediment, nutrient and pesticide yields on large river basins. This allows users to quantify at different scales (e.g., time, space) the independencies between natural and anthropogenic pressures and states of water bodies [41].

We hope that the readers of this special issue will share the enthusiasm and interest that the enviroGRIDS consortium put into the development of this innovative regional Earth Observation system, on the border between spatial data and Grid infrastructure, and on the edge between computing and environmental sciences. While serving the needs of the Black Sea region, it is clear that all the developed piece of software and solutions can be implemented elsewhere in the World.

ACKNOWLEDGMENT

The authors would like to acknowledge the European Commission ''Seventh Framework Program'' that funded the enviroGRIDS project (Grant Agreement no. 227640). The views expressed in the paper are those of the authors and do not necessarily reflect the views of the institutions they belong to.

REFERENCES

[1] UNEP, Global Environment Outlook (GEO) - 5: Environment for the future we want, 2012. p. 550.

[2] Beniston, M., et al., Obstacles to data access for research related to climate and water: Implications for science and EU policy-making. Environmental Science & Policy, 2012. **17**(0): p. 41-48.

[3] Nebert, D.D., Developing Spatial Data Infrastructure: The SDI Cookbook2005. 171.

[4] IPCC, Climate Change 2007 - Impacts, Adaptation and Vulnerability - Contribution of Working Group II to the Fourth Assessment Report of the IPCC, 2007.

[5] IPCC, Climate Change 2007 - Mitigation of Climate Change - Contribution of Working Group III to the Fourth Assessment Report of the IPCC, 2007.

[6] IPCC, Climate Change 2007 - The Physical Science Basis - Contribution of Working Group I to the Fourth Assessment Report of the IPCC, 2007.

[7] Lecca, G., et al., Grid computing technology for hydrological applications. Journal of Hydrology, 2011. **403**(1-2): p. 186-199.

[8] Gerlak, A.K., J. Lautze, and M. Giordano, Water resources data and information exchange in transboundary water treaties. International Environmental Agreements-Politics Law and Economics, 2011. **11**(2): p. 179-199.

[9] Roehring, J., Information Interoperability for River Basin Management, in Technology Resource Management and Development2002. p. 127-134.

[10] Argent, R.M., An overview of model integration for environmental applications,Äîcomponents, frameworks and semantics. Environmental Modelling & Software, 2004. **19**(3): p. 219-234.

[11] Buytaert, W., et al., Web-Based Environmental Simulation: Bridging the Gap between Scientific Modeling and Decision-Making. Environmental Science & Technology, 2012. **46**(4): p. 1971-1976.

[12] Papajorgji, P., A plug and play approach for developing environmental models. Environmental Modelling & Software, 2005. **20**(10): p. 1353-1357.

[13] Hannah, D.M., et al., Large-scale river flow archives: importance, current status and future needs. Hydrological Processes, 2011. **25**(7): p. 1191-1200.

[14] Reed, C., Integrating Geospatial Standards and Standards Strategies into Business Process, 2004, OGC. p. 1-7.

[15] McKee, L., 18 reasons for open publication of geoscience data, 2010, Earthzine. p. 1-8.

[16] Open Geospatial Consortium, The Havoc of Non-Interoperability, 2004. p. 7.

[17] UN Global Pulse, Big Data for Development: Challenges & Opportunities, 2012: New York. p. 47.

[18] Bosin, A., N. Dessi, and B. Pes, Extending the SOA paradigm to e-Science environments. Future Generation Computer Systems-the International Journal of Grid Computing-Theory Methods and Applications, 2011. **27**(1): p. 20-31.

[19] Fraser, R., T. Rankine, and R. Woodcock, Service oriented grid architecture for geosciences community, in Proceedings of the fifth Australasian symposium on ACSW frontiers - Volume 682007, Australian Computer Society, Inc.: Ballarat, Australia.

[20] Giuliani, G., N. Ray, and A. Lehmann, Grid-enabled Spatial Data Infrastructure for environmental sciences: Challenges and opportunities. Future Generation Computer Systems, 2011. **27**(3): p. 292-303.

[21] Ray, N., et al., Distributed Geocomputation for Modeling the Hydrology of the Black Sea Watershed, in Environmental Security in Watersheds: The Sea of Azov, V. Lagutov, Editor 2012, Springer. p. 141-157.

[22] Diaz, L., C. Granell, and M. Gould, Case study: Geospatial processing services for web-based hydrological application, in Geospatial Services and Applications for the Internet2008. p. 31-47.

[23] Goodall, J.L., B.F. Robinson, and A.M. Castronova, Modeling water resource systems using a service-oriented computing paradigm. Environmental Modelling & Software, 2011. **26**(5): p. 573-582.

[24] Paudyal, D.R. and K. McDougall, Buliding Spatial Data Infrastructure to support sustainable catchment management, in Queensland Spatial Conference2008: Gold Coast. p. 1-9.

[25] Tarboton, D.G., et al., Development of a Community Hydrologic Information System. 18th World Imacs Congress and Modsim09 International Congress on Modelling and Simulation, 2009: p. 988-994.

[26] GEO secretariat, Global Earth Observation System of Systems 10-Year Implementation Plan Reference Document, 2005. p. 209.

[27] GEO secretariat, White Paper on the GEOSS Data Sharing Principles, 2008. p. 93.

[28] Stanners, D.A. and P. Bourdeau, Europe's environment: The Dobris assessment, E.E. Agency, Editor 1995, European Environment Agency: Copenhagen.

[29] Danube Pollution Reduction Programme, Strategic action plan for the Danube river basin 1995-2005, 1999.

[30] Arnold, J., et al., Large area hydrologic modeling and assessment - Part 1: Model development. Water resources bulletin, 1998. **34**(1): p. 73-89.

[31] Gorgan D., Giuliani G., Ray N., Cau P., Abbaspour K., Charvat K., Jonoski A., Lehmann A., "Black Sea Catchment Observation System as a Portal for GEOSS Community". International Journal of Advanced Computer Science and Applications, this issue.

[32] Giuliani G., Ray N., Lehmann A., "Building Regional Capacities for GEOSS and INSPIRE: a journey in the Black Sea Catchment". International Journal of Advanced Computer Science and Applications, this issue.

[33] Charvat K., Vohnout P., Sredl M., Kafka S., Milsdorf T.,De Bono A., Giuliani G., "Enabling Efficient Discovery and Access to Data Services". International Journal of Advanced Computer Science and Applications, this issue.

[34] Mihon D., Colceriu V., Bacu V., Allenbach K., Rodila D., Giuliani G., Gorgan D., "OGC Compliant Services for Remote Sensing Processing over the Grid Infrastructure". International Journal of Advanced Computer Science and Applications, this issue.

[35] Mihon D., Colceriu V., Bacu V., Gorgan D., "Grid based Processing of Satellite Images in GreenLand Platform". International Journal of Advanced Computer Science and Applications, this issue.

[36] Colceriu V., Mihon D., Minculescu A., Bacu V., Rodila D., Gorgan D., "Workflow Based Description and Distributed Processing of Satellite Images ". International Journal of Advanced Computer Science and Applications, this issue.

[37] Balcik F.B., Mihon D., Colceriu V., Allenbach K., Goksel C., Dogru A.Z., Gvilava M., Giuliani G., Gorgan D., "Remotely Sensed Data Processing on Grids by using GreenLand Web-based Platform". International Journal of Advanced Computer Science and Applications, this issue.

[38] Bacu V., Mihon D., Stefanut T., Rodila D., Abbaspour K., Rouholahnejad E., Gorgan D., "Calibration of SWAT Hydrological Models in a Distributed Environment Using the gSWAT Application". International Journal of Advanced Computer Science and Applications, this issue.

[39] Giuliani G., Rahman K., Ray N., Lehmann A., "OWS4SWAT: Publishing and Sharing SWAT Outputs with OGC standards". International Journal of Advanced Computer Science and Applications, this issue.

[40] Almoradie A., Jonoski A., "Web-based Access to Water-Related Data Using OGC WaterML 2.0". International Journal of Advanced Computer Science and Applications, this issue.

[41] Cau P., Manca S., Muroni D., Gorgan D., Bacu V., Lehmann A., Ray N., Giuliani G., "An Interoperable OGC-Compliant Decision Support System for Water Resources Management". International Journal of Advanced Computer Science and Applications, this issue.

Black Sea Catchment Observation System as a Portal for GEOSS Community

Dorian Gorgan[1], Gregory Giuliani[2,7], Nicolas Ray[2,7], Anthony Lehmann[2], Pierluigi Cau[3],
Karim Abbaspour[4], Karel Charvat[5], Andreja Jonoski[6]
[1]Computer Science Department, Technical University of Cluj-Napoca, Cluj-Napoca, Romania, dorian.gorgan@cs.utcluj.ro
[2]EnviroSPACE Laboratory, University of Geneva, Geneva, Switzerland
gregory.giuliani@unepgrid.ch, nicolas.ray@unige.ch, anthony.lehmann@unige.ch
[3]CRS4 - Center for Advanced Studies Research, Sardinia, Italy, pierluigi.cau@gmail.com
[4]EAWAG Institute of Aquatic Science and Technology, Zurich, Switzerland, karim.abbaspour@eawag.ch
[5]CCSS - Czech Centre for Science and Society, Prague, Czech Republic, charvat@ccss.cz
[6]UNESCO-IHE Institute for Water Education, Delft, The Netherlands, a.jonoski@unesco-ihe.org
[7]United Nations Environment Programme Global Resource Information Database, 1211 Châtelaine, Switzerland

***Abstract*—The resources of the enviroGRIDS system are accessible to the large community of users through the BSC-OS Portal that provides Web applications for data management, hydrological model calibration and execution, satellite image processing, report generation and visualization, citizens oriented applications, and virtual training center. The portal publishes through Internet both the geospatial functionality provided by Web technologies, and the high power computation resources supported by the Grid technologies. The paper highlights the issues on the implementation of the portal by heterogeneous technologies, in order to support control flow, processing, and visualization of spatial data for GEOSS community, Earth Science specialists, and generally for Web users.**

***Keywords*—Grids computing; geospatial; SWAT hydrological model; satellite image processing; spatial data processing; distributed computing.**

I. Introduction

To study and to search solutions for improvement of the sustainable development of environment and adequate resource management in the Black Sea catchment region, are ones of the main objectives of the enviroGRIDS (Black Sea Catchment Observation and Assessment System supporting Sustainable Development) project [1]. Moreover the evolution of the complex environmental systems is analyzed in context of land cover, demographic, industrial, and climate changes. One of the main goals of this project is to simulate environmental scenarios concerning the quantity and quality of waters over the coming decades. The enviroGRIDS project aims to develop, calibrate, and provide for execution hydrological models of the Black Sea catchment region. There are four main tasks carried out by the project consortium:

1) Collect environmental sets of data regarding the Black Sea catchment region;
2) Develop a dedicated Spatial Data Infrastructure (SDI) in order to support data sharing and distributed processing;
3) Calibrate and execute high-resolution and large area hydrological models on distributed infrastructures such as Grid;
4) Provide tools and applications to specialists, decision makers, and citizens in order to access data processing and visualization, and develop and run environmental scenarios.

The enviroGRIDS project aims to put together different software technologies and heterogeneous computing resources. One of the main issues of the project is to develop solutions based on interoperability between different technologies, platforms, and applications. For instance, such a case is the interoperability between the Geospatial and Grid infrastructures, in order to get benefits by using in a collaborative manner the both technologies. Each of them comes with important features. While the Geospatial platforms provides very specialized functionalities for Earth Science oriented applications, the Grid infrastructures support scalable, distributed, and parallel high performance computation.

The tools, applications, platforms, and resources of the enviroGRIDS system are available to wide communities of users through its Web portal, called BSC-OS (Black Sea Catchment - Observation System). The system provides graphical user interfaces to Web applications for data management, hydrological model calibration and execution, satellite image processing, report generation and visualization, environmental scenarios, and virtual training center. In order to simplify the access to all these tools and applications the portal has implemented the single sign-on authentication mechanism, through which the user has to authenticate just one times, and then gets authorization to all resources during the same working session.

This paper highlights the availability of the BSC-OS portal, by its resources, tools, applications, and platforms to GEOSS community, Earth Science specialists, and generally to internet users. Meanwhile the presentation focuses on the main challenges and issues regarding the development and using of the BSC-OS portal.

The presentation is structured as follows. Section II presents the works and achievements related with the enviroGRIDS project. Section III describes GEOSS components and services. Section IV sketches the BSC-OS portal architecture and the set of tool and application categories. Each of the next six sections V-X describes a tool and application category such as data management, SWAT model calibration and scenario execution by gSWAT application, satellite image

processing by GreenLand application, geospatial data visualization, two flood scenarios addressing citizens and decision makers, and training material development and execution. The last section XI concludes on the portal development and future work.

II. Related Works

Black Sea catchment region is a huge geographical area and a very complex environment. The watershed related hydrological model involves highly interconnected and continuously evolving interactions at many spatial and temporal scales, and requires to gather and integrate different sets of environmental data such as physical, chemical, and biological [2]. The enviroGRIDS project managed to develop and calibrate the SWAT model [3] as a high-resolution water balance model to the entire Black Sea catchment region, by sub-catchment spatial and daily temporal resolution. The model has been calibrated and validated by using river discharge data, river water quality data, and crop yield data by the approach described in [4]. There are many other projects that are used the SWAT model for limited hydrological areas. The enviroGRIDS project is the first attempt that has accomplished to build and calibrate a such huge hydrological model for the Black Sea catchment region and the Danube River.

There are two other hydrological models such as HEC-HMS and SOBEK that are also used to develop use cases for limited regions. HEC-HMS (Hydrologic Engineering Center - Hydrological Modeling System) [5] is a generic modeling system for simulating precipitation-runoff processes in dendritic catchments. The catchment under study is usually divided in a number of sub-catchments with spatially varying parameters and meteorological inputs. Runoff generation is computed for each sub-catchment and subsequently the generated runoff is routed downstream to the catchment outlet. SOBEK modeling system [6] is used for setting-up the flooding and sediment transport model. SOBEK is a software tool used for flow modeling in many areas such as irrigation systems, drainage systems, and natural streams. The SOBEK 1D/2D model combines one-dimensional (1D) hydraulic modeling of the river channel to a two-dimensional (2D) representation of the floodplains. The enviroGRIDS project has used the HEC-HMS and SOBEK models to study two flood forecasting scenarios in Romania.

Other European projects deal with environmental related subjects [7]. Each project is focused on a specific domain and user community of Earth Science. DRIHM (Distributed Research Infrastructure for Hydro-Meteorology) project [8] attempts to improve the use of Grid and HPC (High Performance Computing) just for HMR Hydro-Meteorological Research) modeling and observational databases. The project supports the study of severe hydrometeorological events, the execution and analysis of high-end simulations, and the dissemination of predictive models as decision analysis tools. ENES and ENSEMBLES projects worked on understanding and prediction of future climate change based on the high resolution, global and regional Earth System models developed in Europe. METAFOR and GENESI-DEC projects aim to provide effective access of users to a variety of data repositories, facilities, tools and services. Other European projects such as SAW-GEO, CYCLOPS, GDI-Grid, GEO-Grid, DEEGREE, DORII, and GENESI-DR address the management of spatial data and environmental tools and applications. EGEE, SEEGRID-SCI, and C3Grid projects have developed Grid based solutions for sharing complex spatial and environmental data sets. Other EU projects such as OBSERVE, EGIDA, Balkan GEONET, BalkanGEONet, and GEONetCab have significant contribution to the development of the environmental network and observation capacity in the South East Europe.

Now, other projects are working on software technologies for developing tools, services, and infrastructures, that can be used by the Earth Science and Environmental community. ENVRI project [9] is a collaboration in the ESFRI (European Strategy Forum on Research Infrastructures) Environmental Cluster, to develop common e-science software components and services. The results will speed up the construction of environmental infrastructures and will allow scientists to use the data and software from each facility to enable multi-disciplinary science. LifeWatch [10] is a European research infrastructure aiming the biodiversity and ecosystem research. Users benefit from integrated access to a variety of data, analytical and modelling tools provided by a variety of collaborating initiatives. LifeWatch offers data and tools in selected workflows for specific scientific communities, and provides as well possibility to construct personalized virtual laboratories, including new data and analytical tools.

The enviroGRIDS project [11] has concerned with Earth Science and Grid based solutions approaching the particularities of the Black Sea catchment region in order to:

- Collect and build a dedicated SDI that is able to support data sharing and distributed processing;
- Process over the Grid infrastructure the huge spatial data such as hydrological models, satellite images, and maps;
- Support scalability in terms of high number of users, applications, and models, high model resolution, large areas, and big dimension of data models;
- Develop interactive applications which hide the complexity of the computation infrastructure and the huge data management. Provide all these applications by the BSC-OS portal to the Earth Science community;
- Calibrate the huge SWAT models of the Black Sea Catchment region and Danube River.
- Process great number of satellite images over the Grid infrastructure;
- Develop training materials by including Earth Science dynamical content and Grid based processing.
- Support interoperability between the Geospatial and Grid platforms, and compatibility of software platforms like URM (Uniform Resource Management), gSWAT, ESIP, GreenLand, gProcess, CWE (Collaborative Work Environment), and eGLE.

III. GEOSS Components and Services

The main tools and applications from the enviroGRIDS portal are registered into the GEO Portal [12] and available as GEOSS services to the large community of Web users.

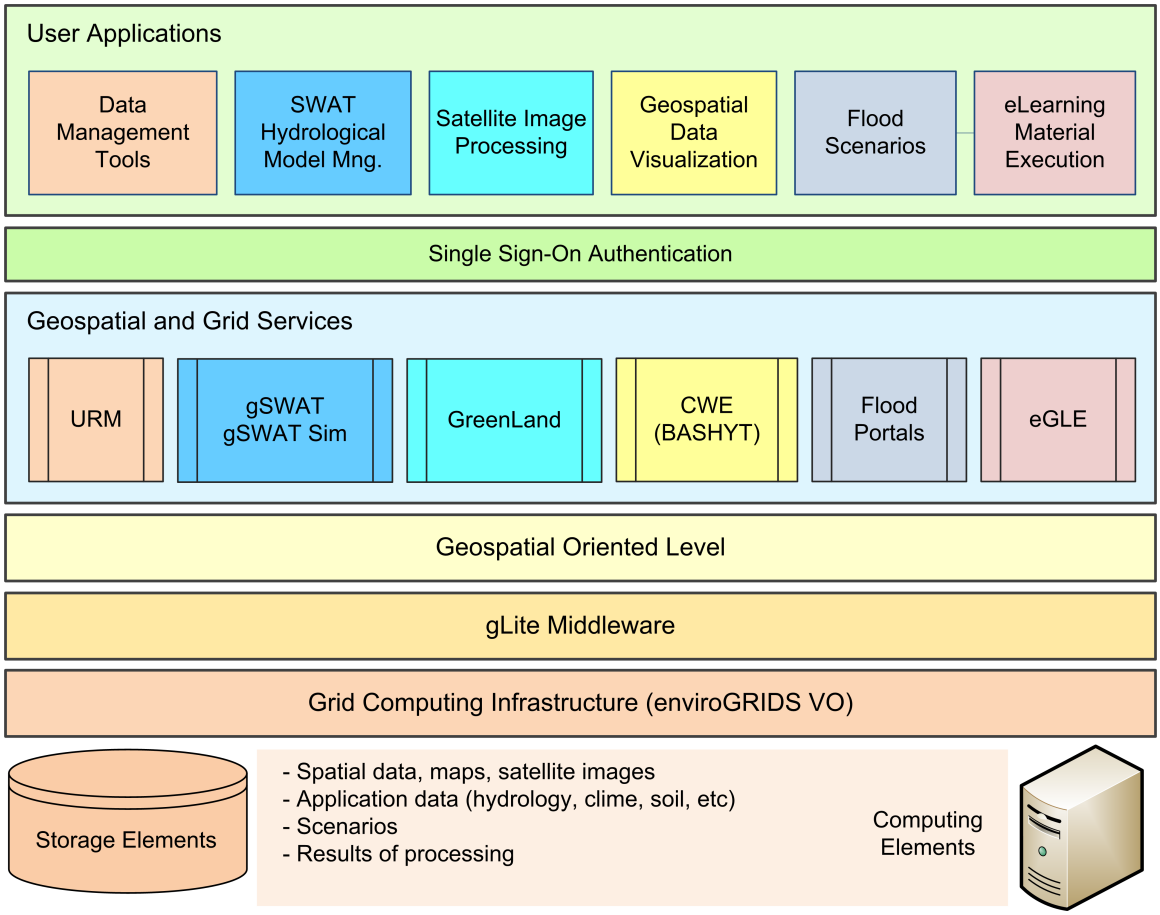


Figure 1. BSC-OS portal architecture

GEOSS (Global Earth Observation System of Systems) [13] is an achievement of the Group on Earth Observations (GEO). The intention of GEOSS is to build Web based global and flexible network compose of content providers. The main idea is to guarantee access of decision makers and a large community of users to range of information. GEOSS is intended to be a system of systems, which will proactively link together existing and planned observing systems around the world and support the development of new systems where gaps currently exist.

GEOSS is simultaneously addressing nine areas of critical importance to people and society. It aims to empower the international community to protect itself against natural and human-induced disasters, understand the environmental sources of health hazards, manage energy resources, respond to climate change and its impacts, safeguard water resources, improve weather forecasts, manage ecosystems, promote sustainable agriculture and conserve biodiversity. The GEOSS solutions are based on SOA (Service Oriented Architecture), which dynamically discover and combine the services on a global scale to support decision-making. The main principles are:

- Services which are based on standard interfaces, utilize common data types, and are well described by standard metadata;
- Distributed computing services may be based on many interaction and transport protocols. Web services based on the HTTP protocol have so far proved to be the usable and interchangeable means of providing access to data and processing resources in a globally federated and diverse environment.

Since 2007 GEOSS has developed the Architecture Implementation Pilots (AIP) [14], which concerns with the development and pilot experiment of new process and infrastructure components for the GEOSS Common Infrastructure (GCI) and the broader GEOSS architecture through an evolutionary development process. The tools and applications developed through various research projects may be registered as components and services into the GEO Portal. The GEOSS Components and Services Registry provides a formal listing and description of all the Earth observation systems, data sets, models and other services and tools that together constitute the Global Earth Observation System of Systems. These various components are being interlinked using standards and protocols that allow data and information from different sources to be integrated. The components and services listed on the Registry can be searched and explored by decision-makers, managers and other users of Earth observations via the GEO Portal.

The GEOSS Components and Services Registry is the main GEOSS catalogue. The GEOSS Standards and Interoperability Registry enable contributors to GEOSS to configure their systems so that they can share information with other systems. One of the key components for interoperability in the GEOSS

architecture is the clearinghouse. Clearinghouse catalogue client specifies that, in order to perform search and discovery of external (to GEOSS) resources, the GEOSS catalogue(s) should operate as discovery brokers. GEOSS Clearinghouse utilizes the OGC Catalogue Service protocol to access the Component and Service Registry and external community catalogues. On the base of analysis, GEOSS defines multiple Web Services interface Implementation Specifications based on OGC Web Services, OGC Web Map Service (WMS), Web Feature Service (WFS), and Web Coverage Service (WCS) [15]. The goal of GEOSS is to develop a dynamically inter-operable infrastructure.

The enviroGRIDS project has developed and registered into GEOSS the tools and applications such as gSWAT, GreenLand, BASHYT, eGLE, and URM related services.

IV. BSC-OS Portal

The BSC-OS portal is the main way of users to access the resources of the enviroGRIDS projects such as environmental data, geospatial services, hydrological models, environmental scenarios, tools and applications, distributed processing tools, satellite image processing applications, geospatial data visualization tools, environmental reports, and training materials (Figure 1). All these functional resources are implemented and provided by the Geospatial and Grid Services level.

The user accesses the portal by Web applications and published by the End User Application level [16]. Each application provides a graphical user interface with high usability. There are five categories of users with different accessing rights according with their professional qualification and assigned role. They are data providers, earth science specialists, decision makers, citizens, and system administrators. Each user may authenticate either locally for a particular application, or globally for all tools and applications within the BSC-OS portal.

The main objective on developing the end user applications was to provide the user the possibility to access from a low performance computation computer, such as his laptop, the processing of huge data on high performance computing resources (Figure 2). The application by its graphical user interface hides the complexity of managing huge distributed spatial data and computing infrastructures. The user accesses the system similarly to a simple local application. Another aim was to access distributed data repositories through a standard manner such as OGC Web Services.

The main categories of end user applications and platforms implemented and published by the portal are the followings:

- *Data Management* – provides the user with spatial data management and operations. The user may enter data and metadata, visualize, modify, update, and remove spatial content from data repositories. The URM platform supports the main functionality required by the end user tools and applications;
- *SWAT Hydrological Model Management* – provides the Earth Science specialists with hydrologic model configuration, model calibration, and hydrologic scenario running. One of the water quality models used in the enviroGRIDS project is SWAT. This model is designed to estimate impacts of land management practices on water quantity and quality in complex watersheds. The SWAT model requires specific information about weather, soil properties, topography, vegetation, and land management practices of the watershed. The gSWAT platform provides the main functionality required on the Web server. The gSWAT Sim platform is a set of services publishing the granular functionality of the gSWAT platform. The platform is able to manage as well the HEC-HMS hydrological models;
- *Satellite Image Processing application* and *GreenLand platform* – the specialists may process satellite data and images in order to search for relevant information (e.g., land cover, vegetation, water, land use, soil composition, etc);
- *Geospatial Data Visualization* and *BASHYT platform* – the specialists visualize various spatial data in different formats and views and compose environmental reports for decision makers and citizens;
- *Flood Scenarios* and *Flood Portals* – provide the decision makers with interactive and graphical tools to access data, maps, reports, and scenarios regarding the floods;
- *eLearning Material Execution application* and *eGLE platform* – supports the specialists to develop Earth Science oriented training materials and the regular users to execute the lessons and Earth Science related processing.

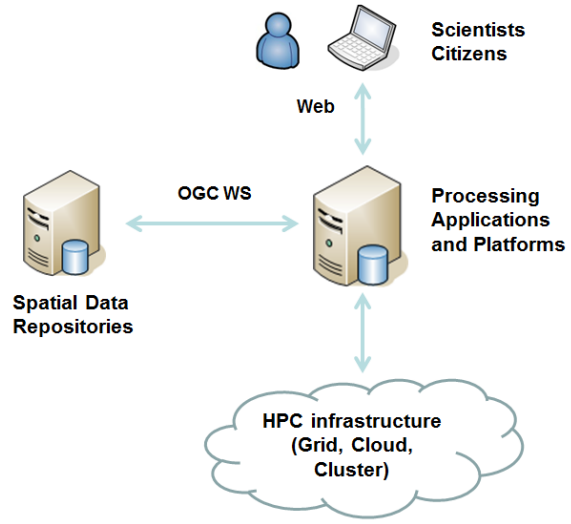


Figure 2. Web applications use remote high performance computation

The regular users visualize the reports generated by the specialists as results of executing environmental scenarios. The input data for the reports are built up by the specialists by running hydrological models of the Black Sea catchment area and by processing related satellite data. All data sets required for building up the hydrological models, environmental scenarios, and spatial models are provided and entered into the system by data providers.

The interoperability between various platforms is supported

by the standard OGC services [15]. The distributed processing is supported by the Grid Computing Infrastructure, based on the gLite middleware, and involving Storage Elements and Computing Elements throughout the Grid [17]. The gLite is lightweight middleware for Grid computing developed through the EGEE project as the foundation of its globally distributed computing infrastructure. Now, the middleware components in gLite became part of the EMI (European Middleware Initiative) distribution [18] and are managed as independent projects providing software to Grid infrastructures such as EGI (European Grid Infrastructure) [19].

V. Data Management Tools

The URM (Uniform Resource Management System) platform [20] supports the sharing, searching and fetching of spatial and non-spatial data, and establishes a network that promotes the GEOSS concept of data sharing for a more sustainable environment. URM Geoportal is a set of modules and services, which are able to communicate through interoperable services defined by OGC (Open Geospatial Consortium), and W3C (World Wide Web Consortium). The URM Geoportal consists of four basic modules interconnected through metadata:

1) *Metadata management* – supported by the MicKa toolset for editing and management of metadata for spatial information, Web services, and other sources;
2) *Data management* – supported by the DataMan application. It provides the import, export, and management of spatial data in files or databases for both raster (IFF/GeoTIFF, JPEG, GIF, PNG, BMP, and ECW) and vector (ESRI Shapefile, DGN, DWG, and GML) data types;
3) *Data visualization* – provided through the MapMan software tool. It supports publication of compositions from locally stored geospatial data with external WMS (Web Map Service) and WFS (Web Feature Service) data services;
4) *Content management* – is supported by the SimpleCMS toolset for publishing in context and connections with social networks.

VI. Hydrological Model Management

One of the main tasks of the enviroGRIDS project is to study environmental scenarios by experimenting hydrological models for the Black Sea catchment region. The Black Sea catchment region needs a high resolution model and involved huge quantity of geospatial data sets. Therefore, the execution on a standalone computer is not efficient at all, especially for the calibration phase. The basic solution proposed through the project is to use the distributed and parallel processing over the Grid infrastructure. The project has experimented on the Grid two model types: SWAT models for Danube River and for Black Sea catchment area, and HEC-HMS for Somes Mare basin.

A. SWAT Hydrological Models

The enviroGRIDS project aims to build up, calibrate and execute huge SWAT models [3] for the Black Sea catchment region [21]. The model allows specialists to develop and study different scenarios, and to make predictions on the impact of management decisions on water, sediment, nutrient and pesticide yields with high accuracy on large river basins.

A good hydrological model is achieved by three steps: development, calibration and evaluation. The model calibration aims to select the best values for model parameters so that the real hydrological behaviour can be simulated. Most hydrological models have two types of model parameters, called physical parameters that represent physical properties of the catchment, which can be measured, and process parameters that represent characteristics which cannot be measured. The calibration process aims to minimize an objective function, which measures the difference between the simulated output of the hydrological model and the measured output. The calibration is a very expensive process requiring high performance computation resources. To reduce the costs, the enviroGRIDS project develops and experiments the calibration and execution of the SWAT model over the Grid infrastructure and evaluate the efficiency of such a solution.

One of the first models built in enviroGRIDS is the Danube model that covers an area of 801,093 km^2, for a river flow distance of 2,826 km. The region has been divided in 1,224 smaller sub-basins. The model has 1.6 GB (compressed) and more than 327,000 files. The calibration process requires running a high number of iterations, each iteration consisting in a high number of simulations. Since the great number of simulations are executed in parallel and distributed over the Grid, the overall execution time of one iteration is dramatically reduced compared with a sequential execution. For instance, the execution over Grid takes 21 hours for 24 simulations, with execution time per simulation of 2,586 sec. It takes 26 hours for 100 simulations, and 939 sec per simulation, and 30 hours for 500 simulations, and that means 215 sec for each simulation. When the system executes a high number of simulations, through distributed and parallel processing over Grid, the average execution time for one simulation is extremely short. Therefore the execution over the Grid becomes efficient for huge models that require very extensive computation scalable to a great number of users, execution processes, and data models.

Another very extensive SWAT model is that of the Black Sea catchment region covering 2.3 million km^2, with rivers from 23 countries, and 160 million inhabitants. The catchment region has been divided in 12,982 sub-basins, for river length of 20,343,825 km. The model has 1,300,000 input output files. The calibration through 200 simulations by a sequential execution on a standalone machine could take 8,059 hours, while the Grid based distributed and parallel calibration manages to reduce the time to 40 hours.

B. gSWAT Application

The gSWAT application has been developed in enviroGRIDS project and is available through the BSC-OS portal in order to support the calibration and execution of the SWAT models [22], [23]. The Grid infrastructure is the basic solution for parallel and distributed processing of the hydrological model by the gSWAT application [21]. It is developed as a Web application that hides to the user the complexity of the Grid infrastructure (Figure 3). The application provides support

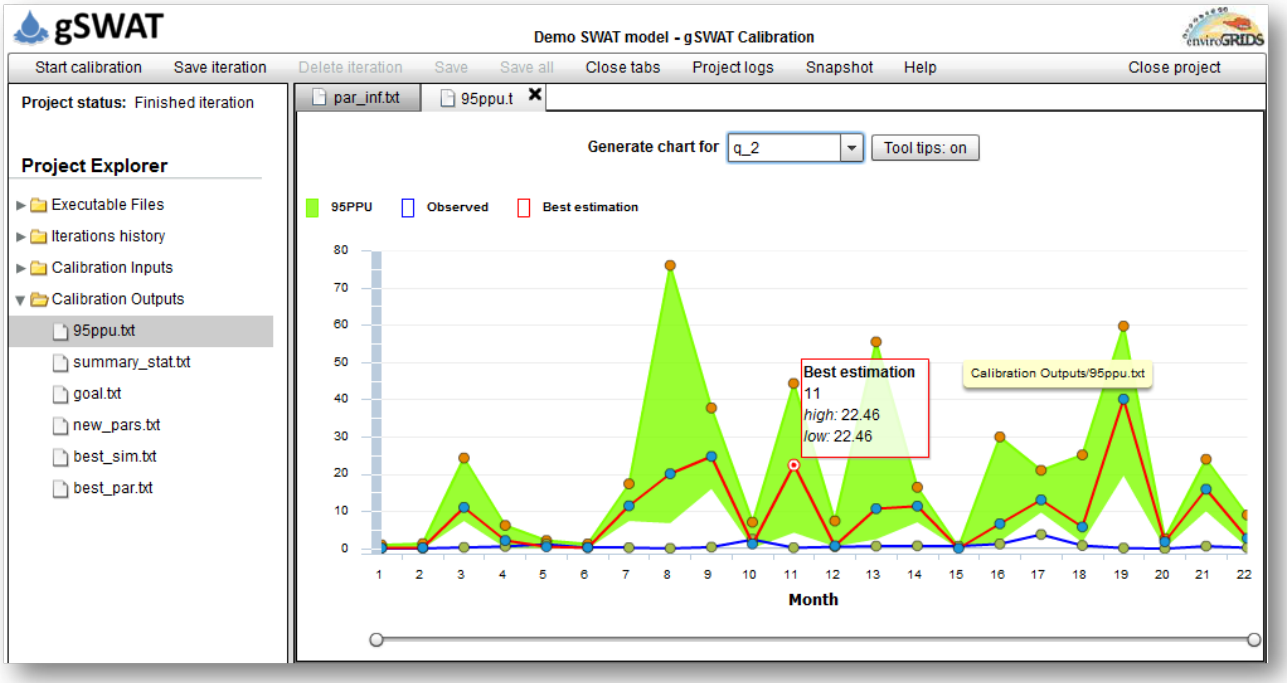


Figure 3. Graphical visualization of the SWAT model calibration results

for scalable models in terms of geographical area, modeling resolution, number of models simultaneously running, and number of users. Cloud [24], Multicore architecture [25], and GPU cluster based solutions are explored as well in order to speed up and optimize the hydrological model processing.

C. SWAT Oriented Services

gSWATSim is a server side extension of the gSWAT platform that is exposed as a collection of REST Web Services supporting the user to create new projects and scenarios; run environmental scenarios; modify some project related information such as name, description, etc.; upload results to visualization module like BASHYT; and get the execution status of scenarios.

D. SWAT Model Development and Running

The hydrological model could be developed, calibrated and run through various approaches based on the gSWAT, gSWATSim, and BASHYT platforms. The specialist could choose one of the following solutions:

1) gSWAT Application: The environmental specialist develops the SWAT model by using ArcSWAT and ArcView tools on his computer. By using the gSWAT application the user uploads the model onto the gSWAT server and executes interactively the calibration of the model [22]. The user controls the steps to reach the optimal calibration by setting up the parameters, the simulations, and the iterations, through user interactive techniques provided by the Web graphical user interface. For instance, in Figure 3 the user visualizes graphically the set of executed simulations and the best one. Finally the user can download the resulted calibrated model.

2) gSWAT and BASHYT Tools: The gSWAT and BASHYT applications collaborate through different working sessions that are connected just at the data level. The main advantage of this solution is the portability of the data model between independent tools. The user carries out the following steps:

1) Develops the SWAT model just in BASHYT;
2) Downloads the archived SWAT files and metadata onto the Storage Element accessed by the gSWAT platform;
3) Performs the calibration by gSWAT as in the first solution;
4) Uploads the results into BASHYT and visualizes the environmental information.

3) gSWATSim Services: The applications work together through a common Storage Element and dedicated Web Services. The control flow of the processing is in BASHYT through which the user develops the model and defines the scenario. The user exports model data onto the Storage Element by gSWATSim services. Then through dedicated Web Service the user customizes the execution environment, and performs the execution of the scenario. The progressing of the scenario execution can be sampled from BASHYT. Finally, after the execution is completed, the results are available automatically into BASHYT for visualization. By this solution the user does not need to switch between the applications. BASHYT accesses a new functionality available through gSWATSim services, which allows both the execution and the monitoring of running scenarios.

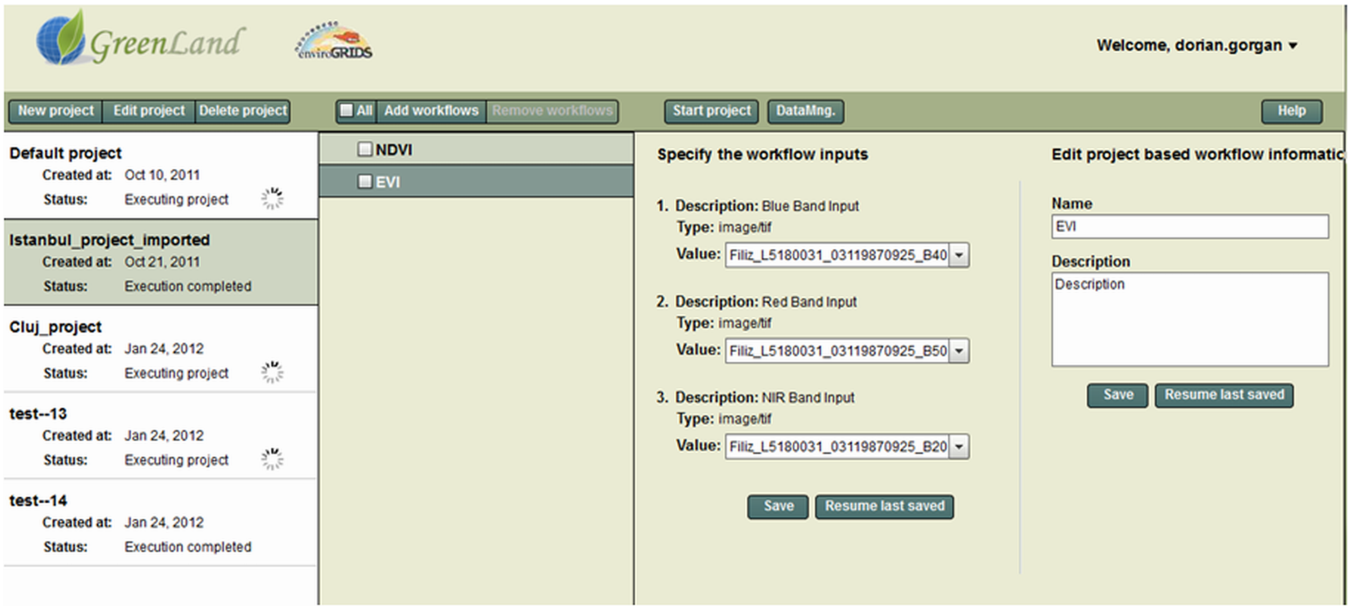


Figure 4. GreenLand application available in the BSC-OS portal for satellite image processing over the Grid infrastructure

VII. Satellite Image Processing

The satellite image processing is supported in the BSC-OS portal by the GreenLand platform and application. Satellite images such as MODIS and Landsat, could reveal information on land cover, precipitations, soil composition, moisture, pollution, and various natural phenomena. Spatial and environment related data could be obtained by imagery classification and processing of the multispectral bands. The image classification is a multivariable process that requires flexible and powerful tools and applications to support an optimal search for the appropriate solutions.

The GreenLand platform supports the development of Grid based applications for satellite image processing, and layers the ESIP (Environment oriented Satellite Data Processing Platform) and gProcess platforms [26]. ESIP supports a workflow based flexible description of the satellite images complex processing over the Grid. the ESIP platform includes the gridified GRASS library [27]. The gProcess platform supports the management and execution of workflows, task distribution, and management of parallel and sequential tasks across the Grid infrastructure.

The BSC-OS portal publishes the GreenLand end user application that is accessible by Web browsers (Figure 4). The GreenLand application offers the following satellite image processing related functionalities:

- Describe the image processing by acyclic graphs. There are two types of graphs: (a) pattern description called PDG (Process Description Graph), and (b) instantiated description called iPDG (Instantiated PDG). PDG describes the processes by basic operators, services, subgraphs, data types, and their connectivity throughout the graph. iPDG completes the PDG description by real data that have to be processed;
- Describe the basic functionality by a set of basic operators. The complex functionality is described by workflows, and remote Web services;
- Complex processing is executed in parallel and distributed over the Grid infrastructure. The simple processing is executed locally if it requires low performance computation resources;
- gProcess platform maps the workflow description onto the physical resources of the Grid infrastructure;
- Supporting the scalability, in terms of number of users, number of projects, number of workflows;
- GreenLand uses OGC Web services to search, visualize, fetch, and store the satellite images;
- Interoperability between GreenLand and URM is supported by standard OGC services (e.g., WMS, WCS, and WFS);
- GreenLand publishes satellite data by OGC services provided by GeoServer, and registered on the URM server;
- GreenLand functionality and operators are published through a standard WPS (Web Processing Service) interface (e.g., NDVI, EVI, and Accuracy Assessment);
- Two graphical editors support the development of basic operators and workflows. The first editor includes into the GreenLand platform the basic operators, which are used later to develop complex functionalities as workflows. The workflow editor supports the diagrammatic description of complex processing to be executed over the Grid.

Through the enviroGRIDS project the features of the GreenLand have been extended to cover the requirements of three main use cases: (a) land cover monitoring for the Istanbul area in Turkey, (b) Rioni River in Georgia, and (c) Mosaic scenario related with the Black Sea catchment region.

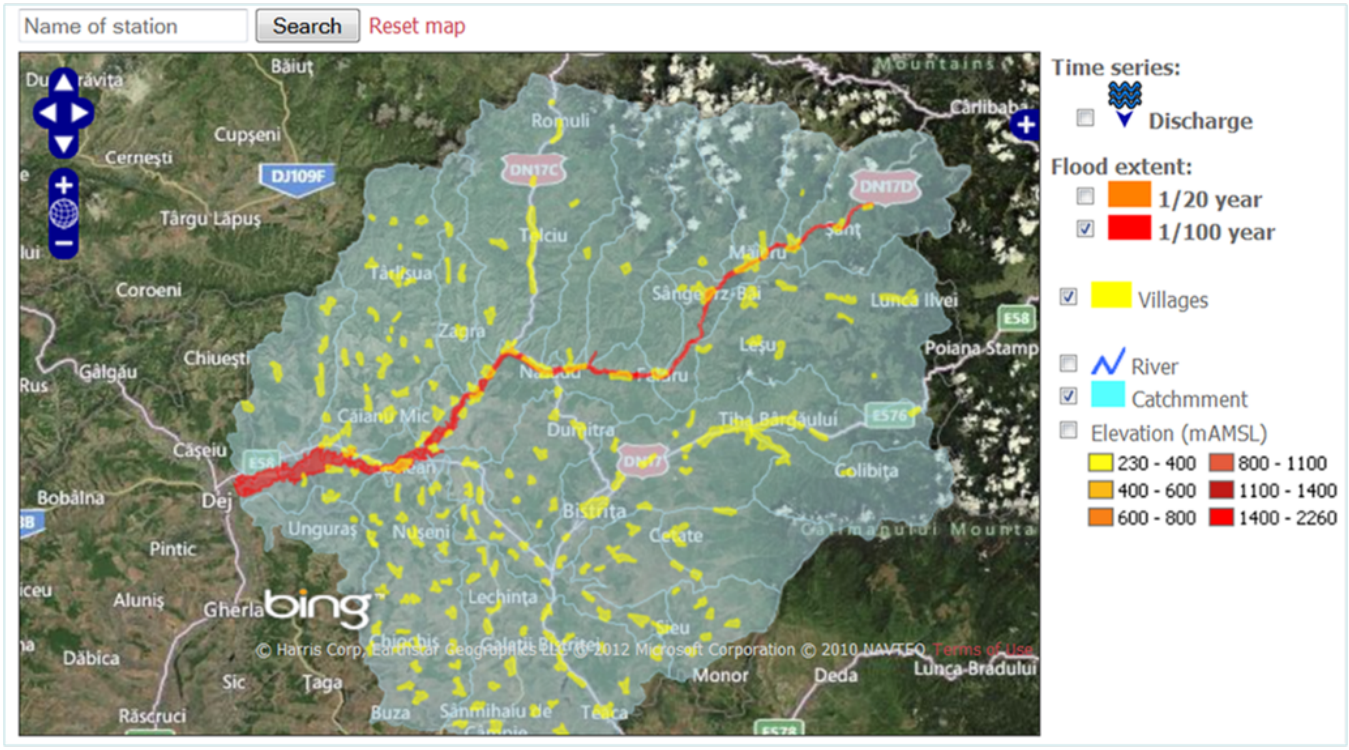


Figure 5. Flood Portal for the Somes Mare catchment area

VIII. Geospatial Data Visualization

BASHYT (The Basin Scale Hydrological Tool) is a Web based interface to SWAT that works together with ArcSWAT and AvSWAT [28]. It can be used to manage many watersheds/scenarios at once and exposes on the Web a template to produce environmental applications. BASHYT supports adaptive strategies for water and soil resource vulnerability. The tools assist decision makers in the field of sustainable water resources management. Such a decision support system is designed to meet the needs of administrations involved in integrating environmental reporting procedures (based primarily on GIS, tables, graphs) and analysis tools. BASHYT supports a Web based, live programming environment, making the programming features available to developers with almost no learning curve. This increases the productivity of the software development by reducing scaffolding code of end user applications.

In BASHYT the SWAT models are stored into a relational database, and a preprocessing step is required to import raw data (vector, raster and tabular data) into the system. After importing SWAT models BASHYT could offer tables, charts, and maps in a transparent way to the end users.

IX. Flood Scenarios

Two demonstrator Web applications, available through the BSC-OS portal, have been developed for citizens within the enviroGRIDS project. The first application, which is related to near real time dissemination of environmental data to citizens, a flood forecasting demonstrator is applied on the Somes Mare catchment in north-western Romania (Figure 5). For the second application, related to long term planning in river basins a demonstrator for long term planning of remediation strategies regarding flooding, sediment and ecosystem problems along the Danube River section between the towns of Braila and Isaccea has been selected.

The first application is supported by the HEC-HMS [5] hydrological model calibrated over the Grid infrastructure. For the calibration the computation system executes a large number of iterations of HEC-HMS model with randomly generated parameters. The model is developed for the Somes Mare catchment area of 5,078 km^2, covering 27 sub-basins, along a length of 136 km. Grid calibration consists of 1,000 iterations executed on 6.72 hours, over 286 worker nodes, and generating 1 TB of data results.

The second use case is supported by the SOBEK 1D/2D [6] hydrodynamic model of flow and sediment transport. Geospatial data are available through the enviroGRIDS URM Portal by standard OGC services, while for water-related time series data the emerging WaterML standard is used.

On the client side, for both applications the main interfaces are map-based such as OpenLayers, Google maps, and Google Earth platforms, over which the additional data are overlaid as spatially distributed data, or point data containing time series of modeled results.

X. Training Material Development and Execution

The BSC-OS portal provides access to the virtual training center based on eGLE (GiSHEO eLearning Environment) [29]. Both the authoring and the execution of the training materials are supported by eGLE. The user plays two roles called generically teacher and student.

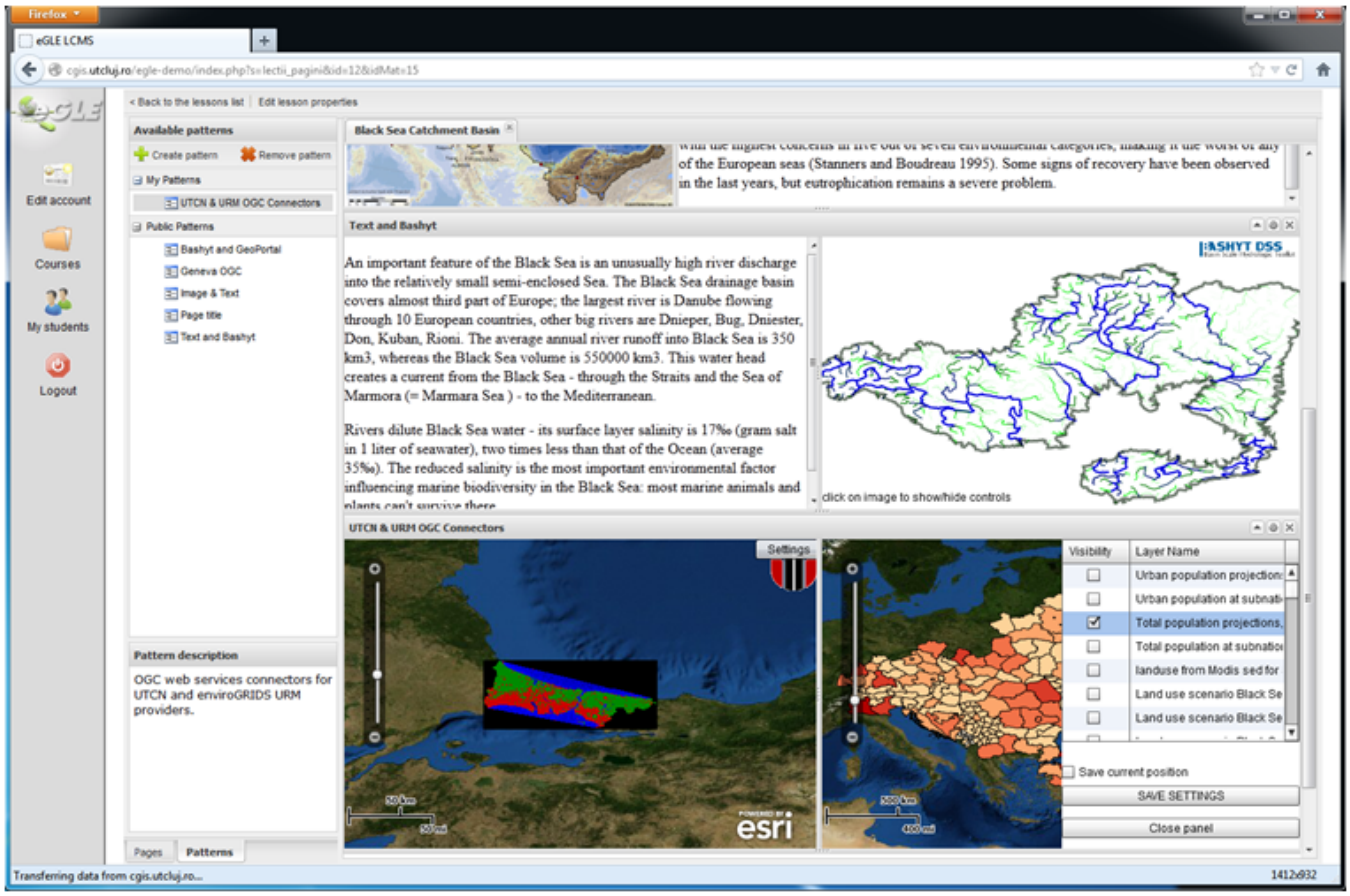


Figure 6. eGLE eLearning platform for teaching materials development and execution

The teacher is the Earth Science specialist who authors teaching materials and coordinates the training sessions. The student is the trainee who accesses the teaching objects organized within lessons in order to get presentations, experiment algorithms on spatial data, process satellite images, execute environmental scenarios, and visualize reports already prepared by the specialists.

The teaching material is organized by lessons in terms of structural templates, patterns, and tools. The Earth Science related content of the lessons may be static or dynamically fetched from data repositories by standard OGC services such as WMS and WCS (Figure 6).

The Grid based processing provided by gSWATSim and GreenLand platforms through Web services can be included and called from the lesson content. The teacher may use the Grid based execution to process satellite images, to execute specific algorithms through workflow descriptions or to visualize previously created teaching resources such as already processed satellite images, geographical maps, diagrams, algorithm workflow descriptions, etc. The students have only the right to execute the lessons according to the constraints established by the teacher. Depending on the specified level of interaction, the students could be allowed to describe and experiment new workflows (i.e. algorithms, scenarios) or choose different input data (e.g., satellite images, parameters) for the current workflow.

XI. Conclusions

The execution of the enviroGRIDS project in general, and the development of the BSC-OS portal in particular, have faced with many challenges and issues regarding the large spectrum of concepts, methodologies, standards, technologies, and practical solutions as well.

One of the main issues is the development of the dedicated SDI, by gathering data from different sources, countries, formats, resolution, consistency, and correctness in order to fit them for the same purpose. The unitary management of huge geospatial data sets involved in the development of hydrological models and environmental scenarios (e.g., Danube, Mosaic, Black Sea Catchment, Istanbul use case) over the Grid has been a challenge indeed.

Another challenge has been the interoperability between Geospatial and Grid infrastructures, which are conceptually and technologically different in terms of security and user access rights and policy, service and tasks based granularity of the software components, flow control of processes and processing, management of data repository and resulted data. An important issue is the connectivity through standard OGC services, and interoperability between different platforms developed by the project partners (e.g., URM, gSWAT, ESIP, gProcess, GreenLand, gLite, BASHYT, and eGLE).

The portal development has to keep compatibility with new technologies and functional requirements. One main concern is

the compatibility with the new European Middleware Initiative (EMI), which aims to improve and standardize the dominant existing middlewares in order to produce one simplified and interoperable middleware [18]. EMI attempts to unify a few Grid platforms such as ARC, gLite, Unicore and dCache. The EMI and Globus platforms will empower the EGI (European Grid Infrastructure) with more stable, useable and manageable software.

The service oriented architecture, multicore, GPGPU based systems, Cloud processing are other technologies that are to be explored in order to extend the scalability, interoperability, standard connectivity, functionality, usability of end user applications, system efficiency, and to improve the performance of data processing.

The BSC-OS portal has to be able to work or to move the applications and data onto new high performance computing infrastructure such as Cloud [24]. Such attempt is the Helix Nebula project [30]. The project aims to prepare the way for the development and exploitation of a Cloud Computing Infrastructure, initially based on the needs of European IT-intense scientific research organizations, while also allowing the inclusion of the needs of other stakeholders such as governments, businesses and citizens. The Cloud Computing Infrastructure as a partnership across academia and industry is working to establish a sustainable European cloud computing infrastructure, supported by industrial partners, which will provide stable computing capacities and services that elastically meet demand.

The enviroGRIDS project provides through its BSC-OS portal and the services published through the GEO Portal the main its achievements consisting of a huge repository of spatial data regarding the Black Sea catchment region, two calibrated SWAT hydrological models of Danube River and Black Sea catchment, the gSWAT application and gSWATSim services for calibrating new SWAT models, the GreenLand application to process satellite images, the BASHYT platform for data visualization, and the eGLE platform for eLearning purposes.

ACKNOWLEDGMENT

This research has been supported by the FP7 enviroGRIDS Project (Black Sea Catchment Observation and Assessment System supporting Sustainable Development), funded by the European Commission, between 2009-2013, through the Contract 226740.

REFERENCES

[1] enviroGRIDS project- Black Sea Catchment Observation and Assessment System supporting Sustainable Development, http://www.envirogrids.net/

[2] The Full Picture, GEO Group on Earth Observation, Geneva, Switzerland, pp. 278, 2007. http://www.earthobservations.org/documents/the_full_picture.pdf

[3] Soil and Water Assessment tool - SWAT, http://swatmodel.tamu.edu

[4] K.C. Abbaspour, J. Yang, I. Maximov, R. Siber, K. Bogner, J. Mieleitner, J. Zobrist, R. Srinivasan, Spatially-distributed modelling of hydrology and water quality in the pre-alpine/alpine Thur watershed using SWAT. Journal of Hydrology, 333, pp. 413-430, 2007.

[5] Hydrologic Engineering Center - Hydrological Model System, HEC-HMS, http://www.wrc-hec.usace.army.mil

[6] Deltares Systems - SOBEK Suite, http://www.deltaressystems.com/hydro/product/108282/sobek-suite

[7] European Commission, Research and Innovation Environment, 2012. http://ec.europa.eu/research/environment/

[8] DRIHM - Distributed Research Infrastructure for Hydro-Meteorology. http://www.drihm.eu/

[9] ENVRI - Common Operations of Environmental Research Infrastructures, http://envri.eu/

[10] LifeWatch European research infrastructure, http://www.lifewatch.eu/

[11] D. Gorgan, V. Bacu, D. Mihon, T. Stefanut, , D. Rodila, P. Cau, K. Abbaspour, G. Giuliani, N. Ray, A. Lehmann, Software platform interoperability throughout enviroGRIDS portal, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS), 5(6), pp. 16171627, 2012.

[12] GEO Portal, http://www.geoportal.org

[13] Global Earth Observation System of Systems, http://www.earthobservations.org/geoss.shtml

[14] GEOSS AIP Architecture, pp. 49, Feb. 2013 http://www.earthobservations.org/documents/cfp/201302_geoss_cfp_aip6_architecture.pdf

[15] Open Geospatial Consortium, OpenGIS Web Service Common Implementation Specification, pp.153, 2007.

[16] D. Gorgan, V. Bacu, D. Mihon, D. Rodila, T. Stefanut, K. Abbaspour, P. Cau, G. Giuliani, N. Ray, A. Lehmann, Spatial Data Processing Tools and Applications for Black Sea Catchment Region. International Journal of Computing - IJC, Vol.11 (4), pp. 327-335, 2012.

[17] gLite - Lightweight Middleware for Grid Computing, http://glite.cern.ch/

[18] EMI - European Middleware Initiative, http://www.eu-emi.eu/

[19] EGI - European Grid Infrastructure, http://www.egi.eu/

[20] K. Charvat, S. Kafka, M. Splichal, M. Alberts, URM for agriculture, environmental education, and knowledge sharing. WCCA 2008 - 6th World Congress on Computers in Agriculture, Tokyo, Japan, (24 - 27 August, 2008), pp. 455-460, 2008.

[21] D. Gorgan, V. Bacu, D. Mihon, D. Rodila, K. Abbaspour, and E. Rouholahnejad, Grid based calibration of SWAT hydrological models, Journal of Nat. Hazards Earth Syst. Sci., Vol. 12/7, pp. 2411-2423, 2012.

[22] V. Bacu, D. Mihon, D. Rodila, T. Stefanut, D. Gorgan, Grid Based Architectural Components for SWAT Model Calibration. HPCS 2011 - International Conference on High Performance Computing and Simulation, pp. 193-198, 2011.

[23] D. Mihon, V. Bacu, D. Rodila, T. Stefanut, K. Abbaspour, E. Rouholahnejad, D. Gorgan, Grid Based Hydrologic Model Calibration and Execution. Chapter in the book: Advanced in Intelligent Control Systems and Computer Science, Dumitrache I. (Ed.), Springer-Verlag, Vol. 187, pp 279-293, 2012.

[24] L. Biro, V. Bacu, D. Rodila, L. Barabas, D. Gorgan, Grid to cloud migration of scientific applications, using dynamically created cloud clusters, IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), pp.335-340, 2012.

[25] D. Rodila, V. Bacu, D. Gorgan, Comparative Parallel Execution of SWAT Hydrological Model on Multicore and Grid Architecture, International Journal of Web and Grid Services, Vol. 8, No. 3, 2012, pp. 304-320, 2012.

[26] D. Gorgan, V. Bacu, D. Rodila, F. Pop, D. Petcu, Experiments on ESIP - environment oriented satellite data processing platform, Earth Science Informatics Journal, Springer, Vol.3/4, pp. 297-308, 2010.

[27] GRASS GIS - Geographic Resources Analysis Support System project, 2011. http://grass.osgeo.org/

[28] S. Manca, C. Soru, P. Cau, G. C. Meloni, M. Fiori, Facing issues of water and soil resources vulnerability: A multimodel and multiscale, GIS - oriented Web framework approach based on the SWAT model, Proceedings of the 2009 SWAT International Conference, Boulder, USA, pp. 91-100, 2009.

[29] D. Gorgan, T. Stefanut, V. Bacu, Grid based Training Environment for Earth Observation. Advances in Grid and Pervasive Computing, LNCS Vol. 5529, pp 98-109, 2009.

[30] Helix Nebula - the Science Cloud, http://www.helix-nebula.eu/

Building Regional Capacities for GEOSS and INSPIRE: a Journey in the Black SeaCatchment

Gregory Giuliani, Nicolas Ray, Anthony Lehmann
Institute for Environmental Sciences, enviroSPACE
University of Geneva
1227 Carouge, Switzerland
gregory.giuliani@unige.ch

Gregory Giuliani, Nicolas Ray
United Nations Environment Programme
Global Resource Information Database
1211 Châtelaine, Switzerland

***Abstract*—To understand environmental systems like the Black Sea catchment, it is required to gather and integrate different datasets. However, data discoverability, accessibility and integration are among the most frequent difficulties that scientists are regularly facing. To tackle these issues, capacity building (at human, institutional, and technical levels) is recognized as a key enabler to raise awareness and create commitments on the benefits of data sharing and publication using interoperable services. In this paper, we present experiences and lessons learnt in the frame of the EU FP7 project enviroGRIDS in developing a network of GEO partners and an efficient strategy to build capacities of scientists from different countries in the Black Sea region. As a result, 27 services, providing access to more than 300 (local or regional) environmental datasets corresponding to around 300'000 layers, are currently registered into the Global Earth Observation System of Systems (GEOSS). Finally, we discuss the added value for stakeholders in the region to participate into GEOSS and the European directive on data sharing INSPIRE, and how to improve its visibility and credibility in the research community, among potential end users.**

Keywords—enviroGRIDS; Capacity Building; GEOSS; Black Sea; Spatial Data Infrastrcuture; Grid computing

I. INTRODUCTION

The Black Sea catchment is a particularly interesting and complex region that is under several environmental pressures from global changes (e.g. climate, demography, land cover) that are influenced by its geophysical and geopolitical situation [1]. First, the Black Sea is almost a closed sea, with reduced exchanges with the Mediterranean Sea through the Bosporus, which led to anoxic conditions in deeper water layers. Second, the Black Sea catchment is very large, covering 2.2 million km^2 and draining more than 150 million inhabitants. Third, by joining the European Union, Bulgaria and Romania brought back the Black Sea on the shores of Europe. Forth, the main tributaries (e.g., Danube, Dnieper, Dniester, Don, Rioni, Kizilirmak) drain large agricultural regions and pass trough numerous dams that both modify significantly the water and sediment quantity and quality reaching the Black Sea. These issues are of particular interest for two important environmental regional commissions, namely the Commission on the Protection of the Black Sea Against Pollution (BSC[1]) and the International Commission for the Protection of the Danube River (ICPDR[2]). The main challenge for the BSC is to fix targets to reduce nutrient loads into the Black Sea from the different catchments/countries. For the ICPDR, the efforts made under the European Water Framework Directive to improve the condition of the Danube river need to have some impacts in the Black Sea as well. The know-how of the ICPDR could be very beneficial to the rest of the catchment to improve the implementation of integrated water resource management in transboundary catchments.

The enviroGRIDS project has explored several scenarios of development for the future of this region [2] with the aim to provide the key spatially explicit information on the past, the present and the future to set the scene for improved decision making. In order to respond to some of the questions related to the water societal benefit areas as defined by the Group on Earth Observations (GEO[3]), the enviroGRIDS project developed for the first time a full catchment hydrological model to predict water quality and quantity according to these different scenarios. To reach its objectives, enviroGRIDS needed to gather, share and process a huge amount of Earth Observation data. In collaboration with other related European projects such as PEGASO[3], BlackSeaScene[4], OBSERVE[5], BalkanGEONet[6], IASON[7], EOPOWER[8], enviroGRIDS is bringing a completely new solution to explore the environment of the Black Sea catchment.

One of the challenges authorities are facing worldwide is the coordination and effective use of the vast amount of geospatial data that is generated continuously [3, 4]. The majority of these data is stored in "electronic silos" at different locations, managed by different organizations [5]. Often, available data are only partly accessible and if they are, often incompatible with one another because of different data formats and standards, data policy, protocols of measurement or analysis, different geographical projection, spatial resolution, lateral overlaps or gaps. Inevitably this can lead to

[1] http://www.blacksea-commission.org

[2] http://www.icpdr.org

[3] http://www.pegasoproject.eu

[4] http://www.blackseascene.net

[5] http://www.observe-fp7.eu

[6] http://www.balkangeo.net

[7] http://iason-fp7.eu

[8] http://www.eopower.eu

inefficiencies and duplication efforts. Moreover the increasing resolution and volume of data require more and more computing resources [6] and consequently limit the possibilities to use them in complex analysis workflow on single desktop computers.

To improve the capacity of scientists to assess the sustainability and vulnerability of the environment and to provide understandable and usable information to decision makers, an essential prerequisite is to convince and help regional data holders to make available their data and metadata to a larger audience in order to facilitate data discovery, access, and analysis.

To address the need of improved environmental data sharing and processing, an interdisciplinary approach can be appropriate. Indeed, Spatial Data Infrastructure (SDI) concept propose a framework to encompass data sources, systems, network linkages, standards and institutional issues in delivering geospatial data and information from many different sources to the widest possible group of potential users [7]. To enable efficient and effective data publication, discovery, evaluation, and access, SDIs mostly rely on interoperability, the capacity to exchange data between two or more systems and to use it. The Open Geospatial Consortium (OGC[9]) is an international voluntary consensus standards organization that promotes and develops open standards for geospatial data and information [8-13]. However, currently SDIs are lacking of computational resources to process the vast amount of data [14]. Therefore, distributed computing paradigm can offer capabilities to complement SDIs. OGC standards can enable an efficient and scalable solution to link these two heterogeneous technologies. This leverages wide and effective exchanges of data, maximizing the value and reuse of data. The capacity to exchange with other systems may also enable new knowledge to emerge from relationships that were not anticipated previously.

Several initiatives at the regional and global scales are promoting the creation of SDIs. These initiatives coordinate actions to promote awareness and implementation of policies, common standards and effective mechanisms for the development and availability of interoperable geospatial data and technologies to support decision making at all levels and for various purposes [15]. At the European level, The Infrastructure for Spatial Information in the European Community (INSPIRE[10]) is a legal directive that is aiming to enable sharing of environmental information to support formulation, implementation, monitoring and evaluation of European policies [16]. At the global scale, the Global Earth Observation System of Systems (GEOSS) is a voluntary effort coordinated by the Group on Earth Observation (GEO) to connect existing SDIs and Earth Observation infrastructures and to act as a gateway between data producers and users [17]. The primary objective is to enhance the relevance of Earth observations for the global problems and to offer public access to comprehensive information and analyses on the environment. To support the nine defined Societal Benefit Areas (SBAs) on disasters, health, energy, climate, water, weather, ecosystems, agriculture, biodiversity, data sharing principles and interoperability arrangements are presented in a 10-year Implementation Plan Reference Document [18] that any participant must endorse.

To reach large adoption, acceptation and commitment on data sharing principles and to increase ability to access and use Earth Observations (EOs) and environmental data, GEO has developed a Capacity Building (CB) Strategy [19]. GEO's definition is based on the one provided by the United Nations Conference on Environment and Development (UNCED) encompassing *"human, scientific, technological, organizational and institutional resources and capabilities"* to *"enhance the abilities of stakeholders to evaluate and address crucial questions related to policy choices and different options for development"* [19]. Three levels are of particular relevance for GEO:

- *Human:* education and training of people to be aware of and able to access, use, and develop EO data and services.
- *Institutional:* development of a working environment (e.g., data policies, organizational and decision structures) for the use of EO to enhance decision-making.
- *Infrastructure:* hardware, software and technology needed to access, use and develop EO services for decision-making.

Particular attention must be devoted to demonstrate the benefits of sharing data through appropriate examples, best practices and guidelines. This helps to strengthen (1) existing observation systems, (2) capacities of decision-makers to use it, and (3) capacities of the general public to understand important environmental, social and economical issues at stake in the region. Additionally, capacity building efforts should aim at convincing a maximum of data owners/providers that they have an opportunity to become more visible nationally and internationally by joining the effort of GEOSS [20].

GEO's survey has revealed several issues related to capacity building, particularly in developing countries [19]:

- Limited access to CB resources;
- Lack of e-science infrastructure for EO education and training;
- Need for criteria and standards for EO CB,;
- Gaps between EO research and operational application;
- Inefficient connectivity between providers and users of EO systems;
- Need for cooperation within and between developed and developing countries and regions;
- Lack of awareness about the value of EO among decision makers; and
- Duplication of EO CB efforts.

9 http://www.opengeospatial.org

10 http://inspire.jrc.ec.europa.eu

Consequently, there are many opportunities to improve this situation [21-23]. GEO is seeking to coordinate and build synergies upon existing efforts and best practices to enhance efficient use of CB resources by:

1) responding and focusing to users needs;
2) fostering collaboration and partnership;
3) concentrating on end-to-end EO needs;
4) enhancing the sustainability of existing and future EO capacity building efforts by raising awareness amongst decision makers, and
5) facilitating the development of comprehensive, sustainable CB efforts to address the needs for infrastructure, education and training, and to build local institutional capacity.

GEO has a dedicated committee on Capacity Building[11] to support the countries to use and benefits from EO products and services and to contribute to GEOSS. There is also a Capacity Building section on the GEO portal, the entry point to discover content in GEOSS, and to access capacity building resources[12]. In complement, there is also a Best Practices Wiki maintained by IEEE to compile and review best practices in all fields of EO[13]. Finally the task ID-02 "Developing Institutional and Individual Capacity" of the GEO workplan is seeking to promote and coordinate actions related to capacity building in GEOSS like the Architecture Implementation Pilots (AIP) activities, the Data Sharing Principles implementation, or the contributions from EU FP7 projects.

Recognizing these opportunities, enviroGRIDS built the capacity of scientists to publish data on the Black Sea catchment using OGC standards, the capacity of decision-makers to use them, and the capacity of the general public to understand the important environmental, social and economic issues in the region. The main objective remains bridging the gap between science and policy by targeting the needs of the Black Sea Commission (BSC) and the International Commission for the Protection of the Danube River (ICPDR). Based on these considerations the aim of this paper is to explore (1) Why does the Black Sea catchment need EO?, (2) Is the Black Sea region ready for EO at the human, institutional and infrastructure levels?, and (3) What is still needed to further improve these capacities?

II. GAP ANALYSIS IN THE BLACK SEA CATCHMENT

To better understand the status of EO in the Black Sea Catchment region a gap analysis was carried out during the first two years of the project (1) to identify the list of existing datasets and observation systems (OS) within the Black Sea catchment, (2) to assess their level of compatibility with the international standards of interoperability, and (3) to identify areas where further efforts are needed to reinforce existing observation systems in this region.

To gather this information an online questionnaire was developed and sent to all project partners, who were requested to provide information about used and available data, observation systems and information networks within their areas of activity from local, national, regional, and global scales. In addition, they were also requested to provide lists of end-users and data needs. To complement the information provided by project partners an intensive Internet search for available data and OS was performed. In total, information about 162 datasets and 30 observations systems covering the Black Sea catchment were identified. The analysis of the identified datasets and observation systems against the project requirements revealed spatial and temporal gaps in data coverage, gaps in observation systems, problems with data accessibility, compatibility and interoperability.

The datasets reported by project partners' cover all 9 GEOSS Societal Benefits Areas (SBAs). The initial statistic of relevance of the reported datasets to SBAs is presented in Table 1. The GEOSS SBAs in this table are sorted according to their relevance frequency. Statistic shows that most of the datasets are related to the Water, Ecosystems, and Climate SBAs, while least covered SBAs are Energy, Weather and Health SBAs. Considering the importance of weather data to build the SWAT hydrological model [24], the limited amount of data for this SBA was an important gap.

TABLE I. RELEVANCE OF PARTNER'S DATASETS TO GEOSS SBAS

Datasets	***GEOSS SBAs***								
	Water	***Ecosystems***	***Climate***	***Agriculture***	***Disasters***	***Biodiversity***	***Energy***	***Weather***	***Health***
Collected / operated by partners	61	57	50	50	47	41	33	21	21
Used by partners	15	13	12	10	8	10	7	3	6

Even if large amount of data sets relevant for the project and end-users was available at different scales (e.g., national, regional, global), data access was often limited or restricted, particularly at the national level. Project partners reported national datasets only for four countries: Georgia, Hungary, Romania and Ukraine, whereas Black Sea catchment is situated on the territory of 23 countries. Thus, there is a large spatial gap in data coverage at country scale. This gap is partly covered by available regional and European scale datasets containing data from Danube basin countries, however for the rest of the Black Sea catchment the problem persists.

With respect to the river basins of the Black Sea Catchment:

- The Danube river catchment has the best data coverage. Data are available at all scales: global, European, regional and national;
- The large river basins of Ukraine (Dnieper, Dniester, Bug) seem to have rather acceptable data coverage,

11 http://www.earthobservations.org/ag_cbc.shtml
12 http://www.geoportal.org/web/guest/geo_capacitybuilding
13 http://wiki.ieee-earth.org

however due to lack of access to data it is difficult to assess their completeness;

- For the large river basins of Russia (Don, Kuban) and Turkey (Kizilirmak, Yesilirmak) project partners did not report any dataset. This is identified as a significant gap in data, particularly taking into account that these river basins are important for the project end-users and decision-makers: they cover large territories populated by millions of people and have important socio-economic value for these countries.

The methods of access to data are various: direct Internet links, FTP, e-mail, CD and USB devices. The datasets of country scale are usually not accessible online and have to be requested via e-mail from data holders. The variety of formats for data storage, as well as the absence of online access to the data hamper the data exchange and appear to be a significant gap for the datasets at country scale. Consequently, data accessibility was the main problem for an effective and efficient use of data. Finally, this analysis highlighted the problem of data compatibility while integrating data from different sources and scales. This require users a lot of efforts to make these data compatible before starting to analyze them.

In term of observation systems, satellite-based platforms are the most important. The available observations systems were analyzed regarding their ability to satisfy the project and end-users data requirement. Based on the fact, that all required data types exist in the Black Sea catchment, it can be concluded that respective observation systems, networks and services also exist. The identified gaps in data may result from different factors such as imperfection of respective observation systems, scarcity of monitoring networks, weakness of data exchange mechanisms and services.

However, the results of the gap analysis of the available datasets clearly indicate that in most cases the real problem is the limited or restricted access to data produced by observation systems rather than gaps in observation systems. The relevant problems are also not developed ownership of datasets and lost datasets after projects are completed. With respect to the most problematic data categories identified, they result from the gaps in observation systems, (i.e. the capacity of monitoring networks/services) for (1) pollutants deposition from atmosphere, (2) oceanography (e.g., in situ measurements), (3) sea water quality, and (4) marine biology and biodiversity.

The issue of data accessibility and availability is of primary importance. Even access to the project partners data in many cases is limited or restricted It is recommended to elaborate appropriate data policy, which envisages different types of data access licenses and encourages open data access and exchange for non-commercial purposes. Then projects partners – data-holders have to share their data for the project under the data policy, further encouraging other stakeholders to do the same.

All these gaps reveal the necessity to enable interoperability among project partners' and raise awareness about the benefits of using EO products and services. In particular, this requires building capacities on Earth Observation in the Black Sea catchment through improved data collection, management, storage, analyses and dissemination.

III. CAPACITY BUILDING BY ENVIROGRIDS

To enable wide data sharing in the Black Sea catchment, the enviroGRIDS capacity building strategy was articulated around 6 components (fig.1), corresponding to those identified to implement an SDI [7, 25-28]. Following the definition of an SDI, it can be thought as a framework of governance, infrastructures, data, and skills that when associated with funding can achieve geospatial data discovery, access and use.

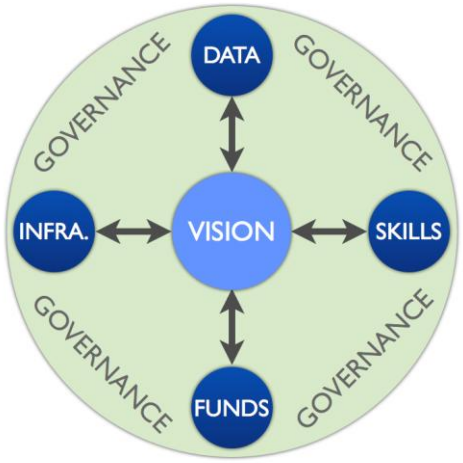


Fig. 1. Components of the enviroGRIDS capacity building strategy.

The central element represents the *vision*, which should define the objectives that enabling data sharing must target. In the case of enviorGRIDS it was (1) supporting the needs of main end-users (e.g., BSC, ICPDR), (2) facilitating discovery and access to existing data, (3) creating and making available new datasets, (4) contributing to data sharing initiatives like GEOSS and INSPIRE.

To support this vision, *funds* must be available to have people working on certain number of activities. This will allow developing also the *skills* of these people through dedicated capacity building activities at the three levels defined by GEO:

- At *institutional* level, the project has created a network of 30 partners in 15 countries targeting the needs of main end-users: BSC and ICPDR. A gap analysis was completed to give a first overview of the EO capacities in the region. Different factsheets, newsletters and policy briefs were written and translated in regional languages to raise awareness about GEO/GEOSS. International organizations (e.g., UNEP, UNESCO, CERN) were also involved as partners. The project was integrated officially in the work plan of UNEP and GEO. Institutional connections were also enabled with other EU FP7 projects in the region to foster data sharing. Finally, an active collaboration with IEEE and OGC allowed developing and sharing teaching material. EnviroGRIDS also strongly promoted the new membership of countries such as Georgia, Bulgaria and Armenia in GEO, as well as the creation of national GEO nodes.

- At *human* level, enviroGRIDS has essentially organized a series of workshops on "Bringing GEOSS into practice"[14] to teach policy and decision makers about GEOSS and INSPIRE, to teach technicians how to install required software to share data and metadata using OGC standards, and finally to teach partners how to become the future trainers. This series of workshops was developed to demonstrate the benefits of data sharing and to show the potential of GEOSS. To disseminate as much as possible this content, all the teaching material and courses are available on the enviroGRIDS [15], GEOSS, and OGC websites. In complement, a Virtual Training Center[16] was developed for providing various learning resources to the project partners, stakeholders from the Black Sea Catchment involved in environmental management at different levels and anyone who is interested in the research topics covered by enviroGRIDS. A network of scientists working in the region was established through the LinkedIn social network. Finally, an enviroGRIDS channel was created on YouTube [17]to broadcast several important videos and presentations on the project outputs. For instance, an animation entitled "the Story of Data on the Environment" as well as a documentary prepared by Euronews "Coloring the Black Sea" are clearly promoting data sharing for a more sustainable future.
- At *infrastructure* level, a distributed grid-enabled spatial data infrastructure shared between several partners of the project was developed to gather, store, discover, access and process key environmental data on the region. Along with the development of the enviroGRIDS SDI, initiatives like GEOSS, INSPIRE or UNSDI were promoted together with the use of OGC and ISO interoperability standards. This enabled partners to develop different tools, build pilot observation systems. In particular, the ICPDR decided to develop its own SDI based on enviroGRIDS recommendations to have more efficient data sharing mechanisms and to improve their environmental assessment processes. Finally, all the components developed in the frame of the project were registered into GEOSS.

As a result of the different capacity building activities we taught more than 300 participants how to share and use data and metadata using OGC and ISO standards, and how to benefit from GEOSS. Based on interoperable services, partners became able to develop different tools to discover, access, process and evaluate data in the Black Sea catchment as well as developing dedicated portals to raise awareness about flooding issues in Romania. All these tools are available in the enviroGRIDS portal [18] that integrates different components supported by different types of infrastructures, enabling communication and data exchange between them.

Additionally, the project created new datasets to explore different scenarios of climate, land cover and demographic changes in the Black Sea catchment, and their impacts on water resources. Several pilot studies were also implemented in different countries on the other GEO Societal Benefit Areas. All the data created by enviroGRIDS is made freely available through web services on the enviroGRIDS portal, where all the available data covering the Black Sea countries are exposed. At the end of the project this has resulted in a set of 27 resources registered into GEOSS (fig.2) corresponding approximately to 300 datasets, giving access to more that 300'000 layers. The effort will continue in different ways. Therefore, this list should increase in the following years.

GROUP ON EARTH OBSERVATIONS

Back

My Previous Registration

My registration list (✔ indicates Approved, ⚠ indicates Pending)

1	Black Sea Catchment DEM - WMS
2	Land Use, Climate, and Demographic Changes Scenarios in the Black Sea Region - WFS
3	Land Use, Climate, and Demographic Changes Scenarios in the Black Sea Region - WMS
4	EnviroGRIDS Map toolkit - WMS
5	EnviroGRIDS.ch metadata catalog - CSW
6	EnviroGRIDS Hydrological modeling (Wp4) - WMS
7	ITU GeoServer
8	Odessa National I.I. Mechnikov University - WMS
9	Black Sea Commission
10	SWAT model outputs - WFS
11	SWAT model inputs - WFS
12	SWAT model outputs - WMS
13	Land Use, Climate, and Demographic Changes Scenarios in the Black Sea Region - WCS
14	Taurida National University Geoserver - WMS
15	EnviroGRIDS URM Portal - Catalog
16	SWAT model outputs - WCS
17	SWAT model inputs - WMS
18	Land Use, Climate, and Demographic Changes Scenarios in the Black Sea Region - WFS
19	Flood Portal Romania - Danube
20	enviroGRIDS
21	Basin Scale Hydrological Tool
22	Grid Based Sattelite Image Processing Platform
23	Grid Based SWAT Hydrological Models Calibration
24	GiSHEO eLearning Environment
25	Flood Portal Romania - Somes Mare
26	EnviroGRIDS Climate scenarios
27	enviroGRIDS - Bringing GEOSS services into practice

Last updated: Wed Jul 03 2013

Fig. 2. EnviroGRIDS resources registered in GEOSS.

IV. Discussion/Lessons Learnt

In its 4-year time frame, the enviroGRIDS project members gained some experience and learned some lessons on developing capacities of different user groups.

A. Success stories

The proposed approach for building capacities in the Black Sea catchment had an impact on several project partners, countries and institutions. Indeed, different partners from Turkey, Ukraine and Romania decided to implement their own

[14] http://bit.ly/15H2SVy

[15] http://bit.ly/14ThgJe

[16] http://bit.ly/JtlEb3

[17] http://www.youtube.com/envirogrids

[18] http://portal.envirogrids.net

SDI to share local datasets they are custodians. They all recognize that having participated to the "Bringing GEOSS services into practice" workshops convinced them about necessity to share data and to use interoperable standards. In complement, for the project partners that were not able to develop their own SDI solution, but wanted to make available their data, one of the project partners (e.g., Czech Centre for Science and Society) offered the possibility to publish their data directly on the enviroGRIDS portal.

At the institutional level, the ICPDR, one of the main end-users of the project, found out that data sharing using OGC and ISO standards could bring several benefits for their assessments and reporting processes. They are currently upgrading their system to enable data exchange among the 14 countries covered by the Danube catchment. It will be entirely based on open source software and open standards promoted by enviroGRIDS allowing them to efficiently fulfill the requirements of the INSPIRE directive as well as the Water Framework Directive (WFD).

At the country level, enviroGRIDS was able to raise awareness about GEO/GEOSS. Actually, Georgia and Armenia have contacted the GEO secretariat to become officially new participating members and have already endorsed the GEOSS 10-Year Implementation Plan [29] and its Data Sharing Principles [30]. Bulgaria is also seriously considering its membership to GEO, which would fill the last gap in the countries within the Black Sea catchment.

Finally, the wide adoption among project partners of OGC standards has permitted the development of several components based on different software and computing infrastructures to discover, visualize, access, integrate and analyze environmental data of the Black Sea catchment. In particular, it enabled the communication between geospatial data repositories (e.g., SDIs) and the Grid computing infrastructure to analyze remote sensing high-resolution images and hydrological modeling for the entire catchment (i.e., 2.2 millions square kilometers).

B. Benefits

At the end of the enviroGRIDS project it is probably too early to highlight major benefits in term of data sharing through GEOSS in the Black Sea Catchment region. However, after 4 years a lot of services were registered facilitating discovery of hundreds of datasets. This can be already considered as a positive result and the number of services and datasets will certainly increase in the forthcoming years.

A relevant impact of GEOSS is the fact that it has enabled networking activities between different contributing projects, creating synergies and fostering information exchange and knowledge development. Moreover, it has also permitted different scientific communities to come together and start talking to each other. In particular, participating to GEOSS allows taking part to activities like Architecture Implementation Pilots (AIP) and other meetings that stimulate and coordinate efforts like efficient data sharing, models integration, user engagement, or capacity building.

Lastly, the use of Free and Open Source Software (FOSS) was truly beneficial in term of capacity building and implementation of data sharing solutions. They are especially attractive for students, GIS professionals, small and medium enterprises, companies and institutions in emerging countries and international organizations. The zero-license cost is obviously an advantage but more important the promoted solutions (e.g., GeoServer, GeoNetwork, PostGIS, OpenLayers, PyWPS, THREDDS) have proven to be efficient [31]. Additionally the fact that all the teaching material including software can be freely disseminated allowed trainees to become trainers.

This contributes to lower entry barriers for both resource users and providers, facilitate development of technical skillsand empower local people.

C. Limitations

Several obstacles were encountered while trying to promote data sharing in the Black Sea catchment. Besides technological aspects main issues identified are related to (1) political/cultural context, (2) policies, (3), organization, (4) people and (5) resources. Same issues were also reported by different authors in various assessments and consultations in Europe [32-38].

The main obstacle faced during the project was the lack of institutional and political wills to publish and share. Indeed, data providers tend to hide data mostly for confidentiality, national security or "misuse prevention" reasons. Additionally, lack of awareness and insufficient staff skills induce shortcomings in standardization (e.g., data, metadata, procedures) and documentation. This results in an incoherent, inconsistent and unshared vision and creates (1) difficulties in finding/accessing dataand (2) lack of knowledge from data providers about the value of what they have.

D. Recommendations

Based on the experience acquired, the success stories and both benefits and limitations encountered, several recommendations can be formulated for data providers and data users for (1) continuing and improving the development of capacities in the region and (2) raising awareness about the benefits of sharing data. For data providers we recommend:

- Asking the UN, EU and national institutions to show the example by making all their data available.
- Improving the GEOSS and INSPIRE geoportal interface to transform the experience of data searching into something more efficient.
- Enhancing data policies to facilitate provision and publication of data. For Arzberger et al. [39] publicly funded data are a public good, produced in the public interest and thus should be freely available to the maximum extent possible. Ideally this should be a guiding principle for every institution.
- Strengthening the sustainability of observation systems especially if capacities are developed in the frame of projects financed for a dedicated period. Memorandums of Understanding (MoUs) can be useful means to ensure the maintenance of essential components of an infrastructure.

- Raising financial resources and engage donors in capacity building activities. To reach this objective it should be demonstrated that EO products and services could offer social and economic impacts. Some reports are already highlighting positive financial impacts and associated costs of non-actions [32, 40-42]. The coordinated approach of GEO should facilitate the engagement of donors by matching identified development needs and their priorities and by developing networks of donors [19].
- Enlarging EO network should be also a priority. It is recognized by GEO as an cost-effective mean of coordination for capacity building efforts [19]. It facilitates exchange of ideas and best practices, creates opportunities of collaboration, encourage exchange for training purposes, promote an open and sharing spirit. Encouraging people and institutions to participate to GEO events like the GEO European Projects Workshop or the GEO & OGC AIP activities are good opportunities to collaborate and exchange with others. Moreover web 2.0 technologies and e-learning platforms were coined as promising solutions in developing capacity building networks [43-45].
- Keeping it simple and let users experience the benefits of interoperability.
- Making the data services directly discoverable by web browser, while reconnecting the metadata with the data itself with a unique identifier for each dataset.
- Developing network of sensors and means to acquire new data, particularly time-series, on identified data gaps.
- Moving away from data formats suitable for 2 dimensions towards multidimensional formats such as NetCDF [46].
- Developing further transparent solutions for large data sharing and processing on distributed computing solutions.
- Developing local/regional node to support GEO. This can help to leverage human, technical and institutional capacities and knowledge.
- Keeping some independence from dedicated solution. Making data available with interoperable services will allow disseminating data to the maximum extent possible and ensuring participating to different initiatives.
- Sharing and documenting data is part of the elementary scientific approach, enhancing scientific accountability and credibility.
- Publishing data and making them discoverable using interoperable standards. There are a lot of different end-users communities that are willing to use EO products and services and that may add value to those products and services.
- Encouraging scientists to share their datasets by allowing them to publish a short description of their datasets on a referenced journal like traditional articles.
- Improving automatic data and services quality checking on all geoportals.
- Developing tailored tools to match the requirements and needs of end-users. If they perceive a benefit then it will facilitate reaching commitments and endorsements. In particular, dedicated thematic and regional portals can be beneficial.
- For end users, we recommend:
- Participating in events of targeted end-users. This facilitates exchange, discussions and ensures that the capacity building activities are answering a specific need of end-users. This is also an opportunity to raise awareness about the benefits of data sharing initiatives like GEOSS and to understand what are the needs of end-users.
- Promoting the use of open source software and the development of freely available education and teaching material. This will help to reach and disseminate resources to the widest audience possible. Additionally, this will ensure a sustainable technology transfer by making accessible cost effective and end-user friendly solutions.
- Promoting regional and thematic geoportals that can more easily implement added values to the shared data.
- Investing in massive learning solutions (Massive Open Online Courses - MOOC) to better promote data sharing needs and solutions among all potential end users.
- Enhancing an "open and sharing spirit" through participative approach. Capacity building activities should demonstrate the benefits of data sharing through appropriate examples communicate best practices and develop guidelines and policies. Altogether this will help to reach agreement and endorsement on the use of new standards.
- Getting involved early in the decision processes and discussions of targeted end users in order to favor the uptake of the promoted solutions.
- Enabling institutions and people to work together and share a common vision.

All these recommendations are aiming to positively influence both data providers and end-users to endorse standards and to commit to data sharing. However, it remains that currently SDI concepts and methods are still strongly related to the geospatial community. In our view, it is required establishing interdisciplinary networks to cross-fertilize disciplines and to promote integration. GEO/GEOSS and INSPIRE represent promising arenas to face this challenge.

Capacity building is a key element to gain acceptance and adoption about data sharing [22]. However, it is a long-term process and the best solution is to establish a long-term

commitment to education and research [47]. Like in any new technologies, the old generations are often more reluctant to adopt them, while often still occupying the positions where decisions are taken.

To improve support and commitment to data sharing, Rajabifard et al. propose [26]: (1) increasing the level of awareness about the nature and value of EO products and services (e.g., capacity building), (2) assessing and understanding the dynamic nature of collaboration and partnership in order to sustain a culture of sharing, (3) improve SDI models to better match the needs of various communities, (4) improve SDI definition to give a clearer vision of its potential benefits and (5) identifying the key factors (in a given context) that can facilitate interactions between social, economical and political issues.

V. CONCLUSIONS

Without sharing data: (1) doing science can be difficult, (2) taking sound decisions can be problematic and (3) envisioning a sustainable development can be complicated. There are a lot of enablers that can influence data sharing. From a technological point of view, all the building blocks are available but the most important component to reach endorsement is not technology but it relies on people (e.g., collaboration, cooperation, social relation, willingness to share and to learn). Indeed, developing the technological component is rather simple but building and maintaining the social one is much more difficult requiring important human and financial resources as well as collaboration, partnership, commitment and trust. Consequently, SDIs can be thought of as social networks of people and organizations supported by data and technology [48].

The answers to the three initial questions can be found below:

1) The Black Sea catchment clearly needs improved EO solutions, like any region of the world, because this catchment represents a federating transboundary unit that is feeding with water the most emblematic geographic feature of the region, the Black Sea itself. Only a well-organized EO system will allow for the important institutions such as the ICPDR and the BSC to address the complex environmental issues influencing water resource sustainability and its vulnerability towards global changes in climate, land cover and demography.

2) The scientists, especially the younger ones, are ready to implement largely the directives and principles of the data sharing promoted by INSPIRE and GEO. From an institutional point of view, there are still too many barriers to encourage a change in paradigm around the true value of EO data. The potential direct commercial value is still dominating the decisions, slowing down significantly the development of an entire economical sector dedicated to geospatial services. The adoption of the INSPIRE directive in European countries and beyond, certainly represent a very promising prospect. From an infrastructure perspective, the main problem resides in the costs for maintaining and developing proper EO systems combining remote sensing with networks of field stations and sensors. Then, the data sharing solutions are becoming really easy to implement with efficient open source and commercial solutions. Data sharing and distributed geoprocessing solutions are largely dependent on fast and reliable Internet infrastructures. The dissemination of EO will become more and more oriented towards mobile devices that are themselves dependent on good cell phone and Wi-Fi coverages.

3) We gave above a long list of recommendations to improve EO in the Black Sea region for data providers and users. These recommendations will only be transformed into actions if there is a strong political understanding and support that data sharing of EO data is essential for guiding the region into a more sustainable future. In a time of important financial and economical difficulties, the very reasonable additional cost of making existing or newly collected data available should be perceived as a high-return strategy for the society.

ACKNOWLEDGMENT

The authors would like to acknowledge the European Commission ''Seventh Framework Program'' that funded the enviroGRIDS project (Grant Agreement no. 227640).

The views expressed in the paper are those of the authors and do not necessarily reflect the views of the institutions they belong to.

REFERENCES

[1] Stanners, D.A. and P. Bourdeau, Europe's environment: The Dobrisassessment, E.E. Agency, Editor 1995, European Environment Agency: Copenhagen.

[2] Mancuso, E., et al., Future land use change scenarios for Black Sea Bastin.

[3] Gerlak, A.K., J. Lautze, and M. Giordano, Water resources data and information exchange in transboundary water treaties. International Environmental Agreements-Politics Law and Economics, 2011. **11**(2): p. 179-199.

[4] Roehring, J., Information Interoperability for River Basin Management, in Technology Resource Management and Development2002. p. 127-134.

[5] Gore, A., The Digital Earth: Understanding our planet in the 21st Century, 1998. p. 4.

[6] Lecca, G., et al., Grid computing technology for hydrological applications. Journal of Hydrology, 2011. **403**(1-2): p. 186-199.

[7] Nebert, D.D., Developing Spatial Data Infrastructure: The SDI Cookbook2005. 171.

[8] Ames, D.P., et al., Introducing the Open Source CUAHSI Hydrologic Information System Desktop Application (HIS Desktop). 18th World Imacs Congress and Modsim09 International Congress on Modelling and Simulation, 2009: p. 4353-4359.

[9] Open Geospatial Consortium, The Havoc of Non-Interoperability, 2004. p. 7.

[10] Open Geospatial Consortium, Geospatial Portal Reference Architecture, 2004. p. 23.

[11] Open Geospatial Consortium, OGC Reference Model, 2008. p. 35.

[12] Open Geospatial Consortium, OGC Market Report: Open Standards and INSPIRE, 2012. p. 32.

[13] Open Geospatial Consortium, The Importance of Going "Open", 2005.

[14] Giuliani, G., N. Ray, and A. Lehmann, Grid-enabled Spatial Data Infrastructure for environmental sciences: Challenges and opportunities. Future Generation Computer Systems, 2011. **27**(3): p. 292-303.

[15] Ray, N., et al., Distributed Geocomputation for Modeling the Hydrology of the Black Sea Watershed, in Environmental Security in Watersheds: The Sea of Azov, V. Lagutov, Editor 2012, Springer. p. 141-157.

[16] European Commission, Priorities for a new strategy for European Information Society, 2010.

[17] GEO secretariat, Building a Global Earth Observation System of Systems, 2008. p. 7.

[18] GEO secretariat, Global Earth Observation System of Systems 10-Year Implementation Plan Reference Document, 2005. p. 209.

[19] GEO secretariat, GEO Capacity building strategy, 2006. p. 13.

[20] Noort, M., GeoNetCab: Marketing Earth Observation Products and Services, Part #2, 2013.

[21] Noort, M., The GEONetCab approach to Capacitiy Building, 2012.

[22] Rajabifard, A. and I.P. Williamson. SDI Development and Capacity Building. in GSDI-7. 2004. Bangalore, India.

[23] Williamson, I.P., A. Rajabifard, and S. Enemark, Capacity Building for SDIs, in 16th United Nations Regional Cartographic2003: Okinawa, Japan. p. 1-14.

[24] Arnold, J., et al., Large area hydrologic modeling and assessment - Part 1: Model development. Water resources bulletin, 1998. **34**(1): p. 73-89.

[25] Craglia, M. and A. Annoni, Building a Spatial Data Infrastructure for Europe: the many research questions for which we need answers, in GISRUK2007: NUI Maynooth, Ireland. p. 1.

[26] Rajabifard, A., M.-E. Feeney, and I. Williamson, The Cultural Aspects of Sharing and Dynamic Partnerships within an SDI Hierarchy. Cartography Journal, 2002. **31**(1).

[27] Rajabifard, A. and I.P. Williamson. Spatial Data Infrastructures: Concept, SDI Hierarchy and Future directions. in Geomatics'80. 2001. Tehran, Iran.

[28] Williamson, I., A. Rajabifard, and A. Binns, Challenges and Issues for SDI development. International Journal of Spatial Data Infrastructures Research, 2006. **1**: p. 24-35.

[29] GEO secretariat, The Global Earth Observation System of Systems (GEOSS) 10-Year Implementation Plan, 2005. p. 11.

[30] GEO secretariat, White Paper on the GEOSS Data Sharing Principles, 2008. p. 93.

[31] Giuliani, G., A. Dubois, and P. Lacroix, OGC Web Feature and Coverage Services performance testing: towards an efficient delivery of geospatial data. Journal of Spatial Information Science, 2013.

[32] Craglia, M. and M. Campagna, Advanced Regional Spatial Data Infrastructures in Europe, in JRC Scientific and Technical Reports2009. p. 132.

[33] Bregt, A.K., et al., Changing demands for Spatial Data Infrastructure assessment: experience from The Netherlands, in A Multi-View Framework to Assess Spatial Data Infrastructures2008. p. 357-369.

[34] Eelderink, L., J. Crompvoets, and W.H. Erik De Man, Towards key variables to assess National Spatial Data Infastructures (NSDIs) in developing countries, in A Multi-View Framework to Assess Spatial Data Infrastructures2008. p. 307-325.

[35] Giff, G., et al., Geoportals in Selected European States: A Non-Technical Comparative Analysis. International Journal of Spatial Data Infrastructures Research, 2008.

[36] Grus, L., J. Crompvoets, and A.K. Bregt, Multi-view SDI Assessment Framework. International Journal of Spatial Data Infrastructures Research, 2007. **2**: p. 33-53.

[37] Henricksen, B., UNSDI Compendium: A UNSDI Vision, Implementation Strategy, and Reference Architecture, 2007, UNGIWG. p. 150.

[38] Vandenbroucke, D., Spatial Data Infrastructures in Europe: State of play sptring 2010, 2010. p. 1-72.

[39] Arzberger, P., et al., Promoting Access to Public Research Data for Scientific, Economic and Social Development. Data Science Journal, 2004. **3**: p. 17.

[40] Booz, Allen, and Hamilton, Geospatial Interoperability Return on Investment, 2005, NASA. p. 47.

[41] European Commission, Assessing the impacts of Spatial Data Infrastrucutres, J.R. Center, Editor 2006: Ispra. p. 1-61.

[42] Garcia Almirall, P., M. Moix Bergada, and P. Queralto Ros, The Socio-Economic impact of the Spatial Data Infrastructure of Catalonia, M. Craglia, Editor 2008. p. 62.

[43] Alvarez, M., T. Delgado Fernandez, and R. Cruz Iglesias, Social networks and Web 2.0 tools as a good complement to the local SDI's, in GSDI-122010: Singapore. p. 1-14.

[44] Alvarez, M. and D. Gallego Gil, Training in Web 2.0 tools: a way to bring Spatial Data Infrastructures to people, in GSDI-122010: Singapore. p. 1-10.

[45] Gonzalez, M.E. and M.A. Bernabe, E-learning training for Spanish Compulsory Secondary Education Teachers to Use SDI as an ICT Educational Resource, in GSDI-122010: Singapore. p. 1-15.

[46] Signell, R.P., et al., Collaboration tools and techniques for large model datasets. Journal of Marine Systems, 2008. **69**(1‚Äì2): p. 154-161.

[47] Gonzalez, M.E., Spatial Data Infrastructures as an Educational Resource in Secondary Education in Spain and Argentina. International Journal of Spatial Data Infrastructures Research, 2008.

[48] Craglia, M. and M. Campagna, Advanced Regional SDIs in Europe: comparative cost-benefit evaluation and impact assessment perspectives. International Journal of Spatial Data Infrastructures Research, 2009.

Enabling Efficient Discovery of and Access to Spatial Data Services

Karel Charvat, Premysl Vohnout, Michal Sredl,
Stepan Kafka, Tomas Mildorf
Czech Centre for Science and Society
Prague, Czech Republic
ccss@ccss.cz

Andrea De Bono[1,2], Gregory Giuliani[1,2]
[1]Institute for Environmental Sciences, enviroSPACE
University of Geneva 1227 Carouge, Switzerland,
[2]United Nations Environment Programme
Global Resource Information Database
1211 Châtelaine, Switzerland
{debono, gregory.giuliani}@unepgrid.ch

***Abstract*—Spatial data represent valuable information and a basis for decision making processes in society. The number of specialisms that use spatial data for such purposes is increasing. Increasing is also the number of services enabling to search, access, process, analyse or visualise spatial data. Standardisation activities of the Open Geospatial Consortium (OGC) support standardised sharing of services through the Web. However, many services declared as OGC compliant do not respond or they are not available. The paper introduces an innovative solution for efficient discovery of and access to spatial data services compliant with OGC specifications. The research was performed in the context of the EnviroGRIDS geoportal. Several thousands of harvested services were quality checked and the summary of the testing including the identified problems are presented.**

***Keywords*—EnviroGRIDS; web services; discovery; metadata; geoportal; SDI; INSPIRE; OGC; SuperCAT**

I. Spatial Data and Interoperability

Spatial data, sometimes referred to as geographic data, geodata or geospatial data, are defined by INSPIRE (Infrastructure for Spatial Information in the European Community) as "data with a direct or indirect reference to a specific location or geographic area." [1] It has been estimated that over 80% of all data have a spatial component. Spatial references enable to locate objects, processes and other phenomena; to model their shape and to analyse their relation to other data [2].

Spatial data are collected by various organisations all over the world, from local to global level. Data are collected using different techniques. The purposes of data collection also vary. Then there are issues for example of data storage, processing, analysing and visualisation. All of these aspects and many others contribute to heterogeneity of spatial data. Due to these aspects, it is not easy to combine data from various resources. In order to make spatial data usable in cross border activities, interoperability framework must be agreed.

Interoperability is defined by the International Organisation for Standardization (ISO) as "capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units." [3]

Recent activity of the European Commission paid due attention to data interoperability in a document describing the European Interoperability Framework (EIF) for European public services. EIF distinguishes four levels of interoperability including legal, organisational, semantic and technical levels. As shown in Figure 1, the political context underlines all the interoperability levels and creates the environment for successful and meaningful cooperation.

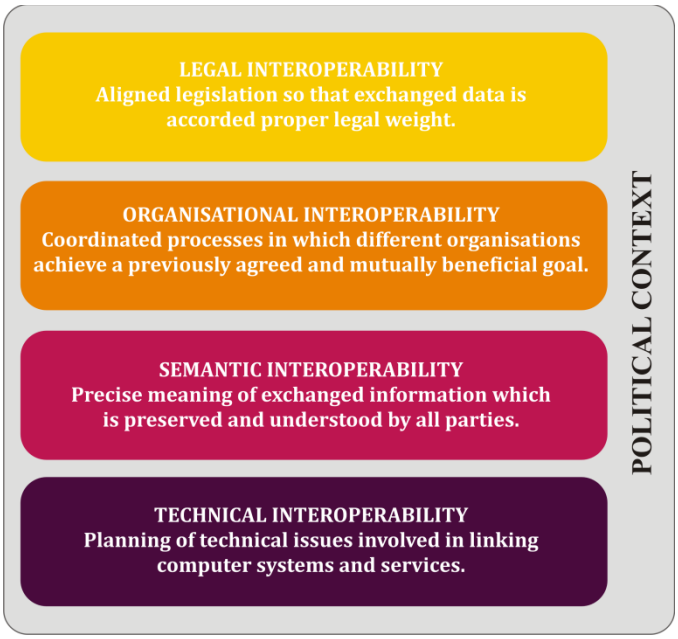


Fig. 1. Levels of interoperability (adapted from [4])

Interoperability on all levels can be achieved through adoption of common standards, specifications and other agreements. The most important international and well respected standards in the field of spatial data are those created by the Technical Committee 211 of the International Organization for Standardization (ISO/TC 211) and by the Open Geospatial Consortium (OGC). Together with the INSPIRE specifications and national standards they create the core of the European spatial data infrastructure (SDI), mainly from the technical and semantic points of view.

Spatial data represent one of the key components of SDI. Spatial data infrastructure, sometimes referred to as spatial information infrastructure, is generally understood as "a computerised environment for handling data that relate to a position on or near the surface of the earth." [5] There are many definitions of SDI. INSPIRE defines SDI as "the metadata, spatial data sets and spatial data services; network services and technologies; agreements on sharing, access and use; and coordination and monitoring mechanisms, processes and procedures, established, operated or made available in an interoperable manner." [6]

This paper presents an innovative solution for efficient discovery of and access to spatial data services and the OGC interoperability standards. "Spatial data services means the operations which may be performed, by invoking a computer application, on the spatial data contained in spatial data sets or on the related metadata." [6] Chapter II describes the rationale for the research in the context of the EnviroGRIDS project. Chapter III describes the innovative solution SuperCAT and Chapter IV provides the results of the spatial data service testing using SuperCAT.

II. EnviroGRIDS Geoportal

EnviroGRIDS is an FP7 project that aims at building capacities in the Black Sea region to use new international standards to gather, store, distribute, analyse, visualise and disseminate crucial information on past, present and future states of the Black Sea region in order to assess its sustainability and vulnerability. To achieve its objectives, EnviroGRIDS built a modern SDI that became one of the components of the Global Earth Observation System of Systems (GEOSS), compatible with the INSPIRE Directive.

GEOSS is being built by the Group on Earth Observations (GEO). GEOSS is focused on user needs and support better utilisation of environmental data and decision-support tools by users. GEOSS is focused on global infrastructure, supplying near-real-time environmental data, information and analyses. GEOSS supports utilisation of information by wide range of users. There are nine areas of interests in GEOSS: disasters, health, energy, climate, water, weather, ecosystems, agriculture and biodiversity. Potential user groups include decision makers in the public and private sectors, resource managers, planners, emergency responders, and scientists. [7]

An important part of the EnviroGRIDS SDI is a geoportal. The geoportal was designed and implemented as a virtual database. It uses the principles of web services, Uniform Resource Management (URM) [8], social media, Geoportal4everybody [9] and semantic web. The geoportal integrates social networking tools supporting social assessment. These services are not implemented directly on the EnviroGRIDS geoportal but as virtual services on different places all over Europe.

The design of the EnviroGRIDS geoportal is based on the analysis of the INSPIRE and GEOSS principles and the principles of the Service Oriented Architecture (SOA) that is INSPIRE compliant. The INSPIRE requirements give to the overall system architecture a loosely coupled integration based on OGC standards, which allows to use any OGC compliant software components.

The EnviroGRIDS geoportal allows management of spatial and non-spatial data across the Black Sea catchment and integration of different existing resources in this area. The geoportal is not only a set of client applications but also a gate to all data and services registered on the geoportal and interconnected servers. The interconnection with other servers is achieved by using the OGC specifications for data and service interoperability. The geoportal enables for example to connect Web Map Services (WMS), Web Feature Services (WFS), Web Coverage Services (WCS) as well as Catalogue Services for the Web (CSW) from other servers.

III. Metadata Catalogue SuperCAT

The central part of the EnviroGRIDS geoportal is a metadata catalogue. The catalogue enables harvesting of external catalogues published by other servers using the OGC Catalogue Service for the Web (CSW). User can then search other interconnected servers and discover and access available services through metadata records (Figure 2).

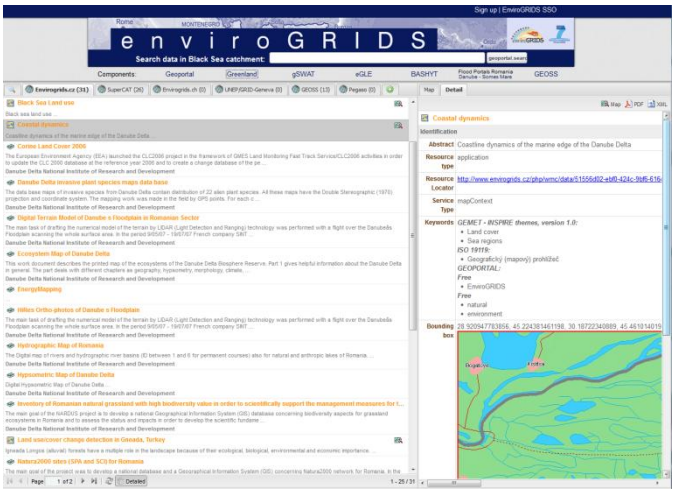


Fig. 2. Metadata search of the EnviroGRIDS geoportal showing the metadata records

There are many OGC web services available worldwide which may be widely used in different applications. In order to make SDI or other applications interoperable, it is crucial to have the services available and reliable at any time.

In real life, many problems occur with regard to operability and access to services. Several examples based on the authors experience can be mentioned:

- Some services do not have a public metadata record that can be used to search the services. The services are not catalogued.
- Services are usually registered in catalogues in order to discover them. However, the services cannot be simply searched by common search engines like Google or Bing. The services can be then searched only through catalogues in which the services are registered or by knowing the Uniform Resource Locator (URL) of the service.

- In many cases, metadata records are not up-to-date and links to services are invalid or not provided at all. Sometimes, the provided URL links to a web page or a viewer rather than to the service itself.
- Some catalogues do not respond.
- Service metadata can be catalogued in different ways. There are no common "global" rules on how to uniquely code certain elements, e.g. serviceType. Querying a catalogue may then result with en error.

The implementation of the EnviroGRIDS geoportal revealed that many OGC services from external servers show the above mentioned errors and this leads to user dissatisfaction. This experience led to the design and implementation of SuperCAT - a metadata catalogue for discovery of reliable services that are OGC compliant. In the first phase, the OGC Web Map Service (WMS) compliance testing was implemented and is further described in the next chapter.

SuperCAT is an independent and extended installation of the Micka metadata catalogue. Micka is a complex system for metadata management used for building Spatial Data Infrastructure (SDI) and geoportal solutions. It contains tools for editing and management of metadata for spatial information, web services and other sources (documents, web sites, etc.). It includes online metadata search engine, portrayal of spatial information and download of spatial data to local computer. Micka is compatible with obligatory standards for European SDI building (INSPIRE). Therefore it is ready to be connected with other nodes of prepared network of metadata catalogues.

SuperCAT is independent on the EnviroGRIDS geoportal and enables verification of the services which are accessed within the catalogue and other catalogues connected through the OGC Catalogue Services for the Web (CSW) 2.0.2.

IV. Service Testing Using SuperCAT

A. Catalogue Interoperability Problems

As shown in the previous chapter, there are many problems with accessing web services. The authors performed an analysis of selected catalogues with regard to their functionality and compliance to the OGC CSW 2.0.2 and the INSPIRE specifications. The following problems were identified:

- The catalogue is not functioning. The catalogue was moved to another address, is temporarily unavailable or is password protected.
- The catalogue is not properly implemented according to the CSW 2.0.2 specification (older version, errors, etc.) or it is based on another standard.
- Many aspects in the current version of the CSW 2.0.2 specification are unclear or missing. As a result, service vendors implement them in different ways (e.g. harvesting of catalogues, behavior of different typenames, queryables).
- CSW should support Dublin Core (csw:Record) and other profiles are optional. There are implementations of ebRIM, ISO, FGDC and others. For INSPIRE the ISO AP 1.0 standard is mandatory. However, not all catalogues support it.
- CSW should support GET, POST and optionally SOAP protocol. In most implementations, POST and SOAP are not used. Not all catalogues implement these protocols for the GetRecord and GetRecordById operations.
- Query languages. CQL and OGC Filter should be supported. Some catalogues have errors in implementation or do not support full language properties.
- There is a mandatory set of queryables. INSPIRE requires additional ones which are usually not implemented by the service vendors.

B. Central Catalogue Implementation and Testing

In order to provide users easy access to services that are correctly implemented, a service metadata repository was built on our server. The repository:

- is CSW 2.0.2 ISO AP 1.0 compliant;
- supports the INSPIRE metadata profile and queryables;
- enables to register remote catalogues (CSW) and harvest them periodically;
- enables to harvest other services (WMS, WFS, WCS) and individual metadata files;
- enables verification of registered services.

The service metadata repository was tested using the following set of metadata catalogues. The most important catalogues of INSPIRE and GEOSS we complemented by other catalogues which are of knowledge by the authors:

- INSPIRE national catalogues of Austria, Belgium, the Czech Republic, Finland, France, Germany, Luxembourg, Poland, Portugal, Slovakia and the United Kingdom;
- GEOSS
- EuroGEOSS
- European Environmental Agency SDI
- EnviroGRIDS
- Habitats (http://www.habitats.cz/)
- One Geology Europe
- Plan4all (http://www.plan4all.eu/)
- World Health Organization

These resources are daily harvested and services are filtered (type=service) using the SuperCAT catalogue. A harvesting protocol form is generated for checking the availability of the

catalogues. Mail notifications are sent to corresponding users to ensure feedback.

The test of service availability is performed on daily basis for all services (only WMS at this phase). If a service is not running, the corresponding metadata record is not deleted but only hidden. This can ensure that temporarily unavailable services can be tested in the future.

C. Testing Results

2222 services were harvested from the registered catalogues. WMS services were analysed in detail. See the results in Table I. 87% of all the services were responding.

TABLE I. TESTING RESULTS

Service type code	Number	Responding (number)	Responding (%)
WMS	96	88	92
OGC:WMS	1418	1351	95
view	343	190	55
View	1	1	100
VIEW	2	2	100
View Service	73	73	100
Total	1860	1632	87

The following problems were identified during the testing:

- serviceType is in many cases ambiguous (see service type codes in Table 1). The INSPIRE Directive brings more confusion into service classification.
- There is no thematic classification for services and metadata are of poor quality (missing elements such as abstract, wrong bounding boxes, etc.). As a result, catalogue queries are not efficient. At least the INSPIRE theme keywords or other commonly used code lists would be a good step to introduce thematic classification.
- Unique service URL is not stated and in many cases it is coded in different ways.

V. CONCLUSIONS

The EnviroGRIDS portal is now part of the GEOSS infrastructure and is an important tool for capacity building in the Black Sea region. Currently, it offers a list of basic services for data integration, supports harvesting of available data, metadata and services and is an integrated access point for data in the region. The architecture enables to combine spatial data, metadata and services from different sources.

The core component of the system is represented by the metadata system management allowing to manage any type of information contained in the geoportal and to use catalogue services for sharing this information with other portals and social network sites.

The paper presented the innovative solution for efficient discovery of and access to spatial data services which are performed using SuperCAT - a metadata catalogue for discovery of reliable services that are OGC compliant.

The test demonstrates that many metadata catalogues including the GEOSS registry do not guarantee that registered services are operational. A better situation is with the INSPIRE national catalogues where the services are mostly guaranteed. But the test of accessibility and operability of services is still needed.

The future work includes implementation of an OGC Web Processing Service (WPS) client to execute external services and also for implementation of Sensor Observation Services (SOS).

ACKNOWLEDGMENT

The paper was prepared on the basis of the outputs of the EnviroGRIDS project - the solution was achieved with the financial co-funding of the European Commission within the Seventh Framework Programme with registration number 226740 and the name "Building Capacity for a Black Sea Catchment Observation and Assessment System supporting Sustainable Development".

REFERENCES

[1] European Commission, 2010b. INSPIRE Glossary. Available at: http://inspire-registry.jrc.ec.europa.eu/registers/GLOSSARY [Accessed February 17, 2012].

[2] Charvát, K. et al., 2013. SDI, INSPIRE and other initiatives. In INSPIRE and Social Empowerment for Environmental Sustainability, Results from the HABITATS project. Spain: TRAGSA.

[3] International Organization for Standardization, 1993. ISO/IEC 2382-1 Information technology -- Vocabulary -- Part 1: Fundamental terms. Available at: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=7229 [Accessed June 15, 2011].

[4] European Commission, 2010a. European Interoperability Framework (EIF) for European public services.

[5] International Organization for Standardization, 2011. CEN/TR 15449 Geographic information - Standards, specifications, technical reports and guidelines, required to implement Spatial Data Infrastructures. Available at: http://esearch.cen.eu/esearch/Details.aspx?id=6880372 [Accessed August 31, 2012].

[6] European Parliament, 2007. DIRECTIVE 2007/2/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Available at: http://eurlex.europa.eu/JOHtml.do?uri=OJ:L:2007:108:SOM:EN:HTML [Accessed May 31, 2012].

[7] http://www.earthobservations.org/geoss.shtml

[8] Janečka, K., Raitis, B. & Charvát, K., 2011. URM GeoPortal4Everybody: the modern platform for vocational education. In E-Learning, Distance Education or ... the Education of 21st Century. Sofia, Bulgaria.

[9] Charvát, K. et al., 2011. Social Space for Geospatial Information. In IST-Africa 2011 Conference Proceedings. IST-Africa 2011. Gaborne, Botswana: IIMC International Information Management Corporation.

OGC Compliant Services for Remote Sensing Processing over the Grid Infrastructure

Danut Mihon, Vlad Colceriu, Victor Bacu,
Denisa Rodila, Dorian Gorgan
Computer Science Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
{vasile.mihon, vlad.colceriu, victor.bacu,
denisa.rodila, dorian.gorgan}@cs.utcluj.ro

Karin Allenbach[1, 2], Gregory Giuliani[1, 2]
[1]Institute for Environmental Sciences, enviroSPACE
University of Geneva 1227 Carouge, Switzerland
[2]United Nations Environment Programme,
Global Resource Information Database
1211 Châtelaine, Switzerland
allenbach@unepgrid.ch, gregory.giuliani@unige.ch

***Abstract*—The latest issues in simulating and analyzing different Earth Science phenomena require the development of complex algorithms, based on satellite images in different formats. The main goal of this paper is to process these data in a standard manner and to automatically publish the execution results by using the latest Web Processing Services (WPS). The development of these services needs to be slightly different when involving large volume of data processed over the Grid infrastructure opposed to standalone machines. This paper provides an implementation solution of the WPS standard within the GreenLand platform, and exemplifies it on the Black Sea catchment hydrologic modeling use case.**

I. Introduction

This paper concerns with the integration of the Web Processing Services (WPS) into the GreenLand application [1] that is built over the Grid infrastructure. The research is part of the FP7 enviroGRIDS project [2], funded by the European Commission through contract no. 226740.

GreenLand is a Grid based software application that operates in the Geographic Information System (GIS) domain, and it is used for geo-spatial data management and analysis, satellite image processing, graphics/maps generation, spatial modeling and visualization. In particular this application offers support for three major case studies: Black Sea catchment hydrologic modeling, land cover/land use analysis of the Istanbul geographic area, and the Rioni river hydrologic analysis [3].

By geo-spatial data we mean raster inputs such as satellite images in different formats (e.g. Moderate Resolution Imaging Spectro-Radiometer (MODIS), Landsat, Aster, etc.) and vector inputs (e.g. ESRI shapefiles). These data are retrieved from remote repositories or from the user's local machine, and converted into a GeoTIFF internal representation.

There are several objectives that are highlighted throughout this paper: OGC WPS integration within the GreenLand platform, the access mechanisms to the GreenLand WPS services, and the interactive implementation of new geospatial algorithms based on the WPS guidelines.

The first one proposes the integration of the Open Geospatial Consortium (OGC) standards - WPS in particular [4] - with the Grid infrastructure, where the GreenLand application is the intermediate communication layer. The advantages of performing this integration are: standardized data access and processing, interoperability with external platforms, flexible data models for storing and exposing the spatial information.

The WPS standard was successfully used in small scale processing, where the computation complexity was not that high. Different issues were encountered in cases of large scale scenarios, where the standard does not provide enough details about the geo-spatial data execution, monitoring, and results presentation.

The three GreenLand case studies, described earlier, require an intensive processing of a large volume of spatial data. All these aspects motivate the usage of the Grid infrastructure and the necessity of extending the WPS standard towards the Grid platform.

The WPS package consists of a set of three services (GetCapabilities, DescribeProcess, and Execute) that are accessible over the Web. Another objective of this paper is related with the usage of these services: directly from the GreenLand application, API access from external platforms, and browser based access.

Usually the Execute service requires a list of mandatory parameters, such as the path to the inputs data set that are going to be used at runtime. In some cases this attribute is hard to be written manually (e.g. in case of accessing the WPS as a link given in the Web browser address bar). Instead an automated generation of the path could be achieved from the GreenLand application.

This is the main reason why the GreenLand-WPS integration is the most easy to use solution (against the three available ones, described earlier), regardless of the users' experience in computer science domain. More details about these concepts are going to be presented in the following sections.

The last objective of this paper is related to the possibility of adding new WPS processes or updating the existing ones. This is possible due to the PyWPS [6] tool that allows the specification of each process as a Python script. The flexible model of the extended WPS standard has an important role in achieving this objective that aims to improve the platforms interoperability, by re-using these algorithms (implemented on external platforms) instead of developing them from scratch.

II. Related work

Different implementations of the WPS standard demonstrated the benefits of this approach [5]. Small scale processing is the common characteristic for all these solutions. When using this standard for complex use cases, extra computation resources need to be added. The most appropriate methods are related to the distributed infrastructures, Grid and Cloud in particular that offer high power computation and storage resources.

There are several research studies of large scale applications development over the Grid [7], [8], [9] and Cloud infrastructures [10]. None of them experiments the WPS integration, and use instead non-standardized executions.

The GreenLand platform also uses the Grid infrastructure for parallel and distributed computations. What differentiates it from the previous mentioned applications is the implementation of OGC-based mechanisms that allow a standardized execution of spatial data. This means that the GreenLand platform is able to interoperate with external systems for sharing, processing, and visualizing of spatial data.

The majority of Earth Observation researches recommend standard access to geo-spatial data. Article [11] provides a proof of concept solution for executing the spatial data on certain backend infrastructures, by using a mediation approach. The paper [12] proposes an integration of the WPS standard with the Cloud infrastructure by using the Hadoop Map-Reduce model.

This paper describes the integration of the WPS standard directly within the GreenLand platform that allows the possibility of developing complex use cases in an interactive manner, closely related to the human natural language. This approach differentiates from the previous mentioned solutions, and allows user access to these features, without the need to have background knowledge related to computer science domains.

There are multiple frameworks that implement the WPS standard. The 52n WPS version could be used for deploying services on standalone machines. It provides a standardized access to data, and allows the creation of new processes through the Development Kit that was integrated in the last release. It uses the R language for implementing the geo-spatial algorithms [13].

Degree WPS [14] is another framework that implements this standard. Currently it is integrated with some of the most known spatial data processing applications, such as Sextante [15]. The Geographic Resources Analysis Support System (GRASS) library [16] is used for implementing different geo-spatial algorithms, and it is used especially on standalone machines.

The PyWPS [6] is a Python based framework that uses the GRASS library for describing the geo-spatial features. On the other hand it offers the possibility to access remote services that provide the same types of algorithms. Because the GreenLand platform uses the GRASS library for the operators' development, the PyWPS implementation was adopted as support for accessing and executing these operators by following the WPS standards.

III. OGC standard general overview

The rules and guidelines of data sharing represent the basis for several standards organizations that put them into practice in the GIS domain fields. According to these issues, they manage to increase the interoperability between systems and geo-spatial data.

There are several standards that are related with the GIS system (e.g. Open Geospatial Consortium OGC [4], International Organization for Standardization ISO [17], Spatial Data Transfer Standard SDTS [18], Organization for the Advancement of Structured Information Standards OASIS [19]), but for spatial data management the most important ones are OGC and ISO.

The Open Geospatial Consortium (OGC) is a non-profit organization that provides guidelines for service oriented spatial data processing and visualization. Based on these standards, the developers are able to create interoperable services for information access and information exchange, but also complex data structures that could be accessible to a large number of applications. The goal of the OGC organization is to provide standards for developers and users in order to produce services for accessing spatial data, and to assure that geo-spatial interoperability:

- Allows the creation of standards (integrated into daily processes) regarding spatial data computing;
- Facilitates interoperability between GIS applications;
- Facilitates the implementation of open architectures.

Without interoperability and standardization, data access and data exchange is really difficult among organizations. In general, any Web service must have the ability to describe its own capabilities, enabling this way other services and products to interoperate based on its standard functionalities. The OGC usage allows this process flow between different GIS software applications without the need to develop new translation mechanisms to assure their interconnection.

There are several OGC standards that are commonly used for accessing, visualizing, analyzing, and processing data throughout Web services:

- Web Map Service (WMS): defines a Web interface that allows geo-referenced data retrieval as map layers. For security reasons the WMS service does not give access to the real data, but instead it creates different layers representation (e.g. JPEG, PNG, TIFF) of this data.
  The WMS interface is a three-steps process that consists in the following operations: GetCapabilities (exposes the service functionalities and lists all its available layers), GetMap (produces a map based on the selected layers set), and GetFeatureInfo (returns information about the generated map content).
  For each WMS request there are lists of mandatory and optional parameters. The response is an XML file that contains information about the service and the data availability;
- Web Feature Service (WFS): it allows features retrieval and management using the GML format. It is intended to be used only for vector datasets, while the

WCS works mainly on raster images. The WFS interface is a thee-steps process: GetCapabilities (similar to the one described for WMS), DescribeFeatureType (defines the structure of the feature), and GetFeature (returns the response encoded with GML schema).
There are two implementations of this standard: basic WFS and transactional WFS. The first one is used to query and to retrieve features, while the WFS-T provides services for features creation, deletion, and update;

- Web Coverage Service (WCS): provides standardized access to raster datasets. Like the rest of the OGC services, the WCS interface consists in three types of requests: GetCapabilities, DescribeCoverage, and GetCoverage. Based on the XML result, generated by the GetCapabilities, the user is able to select and download data for a specific area of interest. The area's geographic coordinates, image format, width, height, and projection are a few of the mandatory fields required by this service;
- Web Processing Service (WPS): besides data accessibility, the OGC standard also provides geo-spatial data processing services through the WPS interface. It allows users to: know what processes are available, to select the proper input data, to create and run different models, to perform management operations to the output results, etc.
As any OGC standard implementation, it includes three types of operations: GetCapabilities (returns a list of the service capabilities together with all its available processes), DescribeProcess (for each process it provides a general description of the parameters and their types), and Execute (performs the process execution. It does not offer tools for monitoring the execution progress, and once the result is available it is send to the user);
- Simple Features SQL (SFS): It is an open standard that offers rules and guidelines for storing, querying, updating, and retrieving geo-spatial features from SQL databases. SFS establishes an architectural framework for features representation, provides syntaxes for defining geometric attributes attached to those features, and describes a set of geometry types in order to ease the data exchange processes.

The GreenLand platform implements the majority of the previous mentioned services (WMS, WCS, and WPS), but this paper offers details only about the Web Processing Service.

IV. System related architecture

This section describes the concepts, the solutions, and the technologies involved throughout the experiments of integrating the WPS standard within the GreenLand platform. The high power computing resources are crucial in optimizing the processing of large volume of geo-spatial data [20].

In order to fulfil the three objectives proposed in this paper, a combination of multiple tools and technologies (Figure 1) was performed, such as: the Grid infrastructure for geo-spatial data processing, the OGC services for standardized data access and specification, PyWPS that acts like a middleware between the WPS Execute operation and the hardware platform, the GreenLand together with the gProcess and GRASS libraries that provides a framework for developing and using the WPS related services.

The Grid infrastructure could be defined as a worldwide computer network that offers parallel and distributed support for storing and processing large volume of data [21].

A. WPS general overview

The Open Geospatial Consortium (OGC) is a non-profit organization that provides guidelines, rules, and software API, recognizable throughout the entire geo-spatial community. Based on these standards, the developers are able to create interoperable services for information access and information exchange, but also complex data structures that could be processed by a large number of applications [4].

The OGC Web Processing Service allows standardized access to data and geo-spatial algorithms, by using the Web technologies. There is a list of mandatory parameters that must be attached to each such service. The inputs and outputs lists, type of request, version of the WPS, and the unique identifier for the algorithm are the most frequently used.

The WPS standard includes three types of operations, accessible as URLs: GetCapabilities, DescribeProcess, and Execute. The GetCapabilities is used for obtaining certain information about all the available services. This information is packed within an XML metadata document that specifies their identifiers, name and description, the provider, the type of projection, etc. The URL http://<server_domain>/wps/wps.py?service=wps&version=1.0.0&request=GetCapabilities can be used to exemplify this operation.

The <server_domain> represent the URL location of the PyWPS server. The service, version and the request parameters are mandatory for all WPS operations and are used to identify the OGC standard (WPS in this case) and its implementation version.

For detailing a particular service, the DescribeProcess operation should be used. Based on the unique identifier, the inputs and outputs lists for the algorithm could be retrieved. The URL associated with this operation has the following structure: http://<server_domain>/wps/wps.py?service=wps&version=1.0.0&identifier=NDVI&request=DescribeProcess.

The identifier parameter is needed only for the DescribeProcess and the Execute operations that are closely related with a specific process.

The last operation allows the execution of a certain process, based on a well-defined list of inputs data set. One such example is given bellow: http://<server_domain>/wps/wps.py?service=wps&version=1.0.0&identifier=NDVI&dataInputs=[input1=value1; input2=value2; ...;inputn=valuen]&request=Execute. For this operation, the URL structure is more complex, because it involves the specification of a valid path to the input data that is stored locally or on remote repositories.

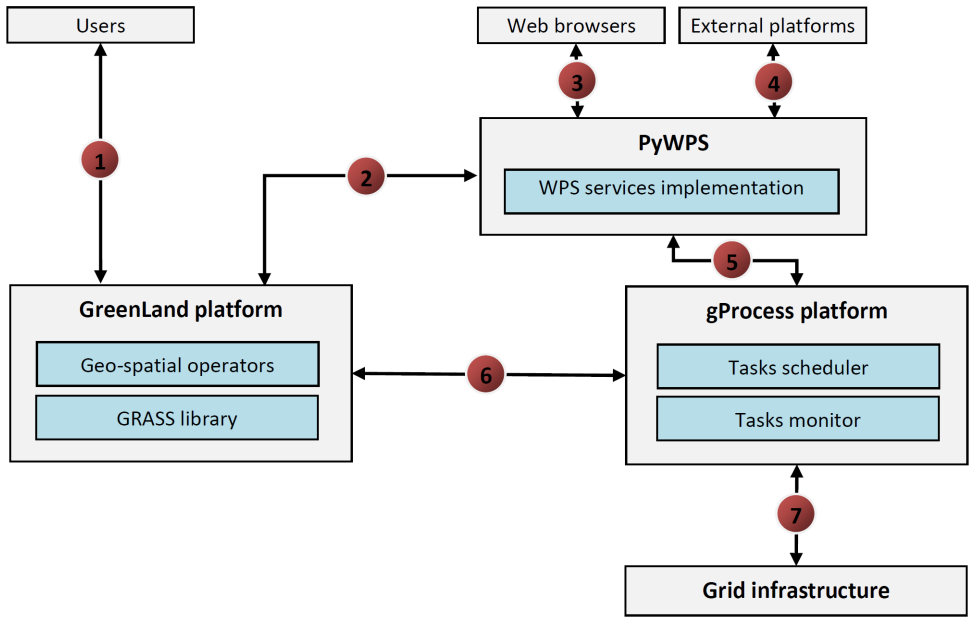


Fig. 1. System related architecture

B. Modules interconnection and characteristics

The PyWPS is a software application, written in Python programming language that implements the WPS standard. It offers the possibility to access HTTP services that contain different geo-processing algorithms. On the other hand, the PyWPS also allows the direct access of GRASS functions.

In cases of executing large scale scenarios, the direct usage of the GRASS features is not enough. The GreenLand platform addresses such complex use cases that need tens of minutes for completing their execution. This is the main reason why a new algorithm development method was implemented that uses the PyWPS for accessing them over the HTTP protocol.

The gProcess tool [22] could be defined as an intermediate layer between the GreenLand platform and the Grid infrastructure. All the geo-spatial algorithms within this platform are transcoded at runtime as nodes of a workflow (graph). Using this data model, it is easier to execute and monitor the processes launched over the Grid infrastructure.

The gProcess platform is used for scheduling the tasks over the worker nodes of the Grid. In this case, each node of the workflow is considered to be a task. This process involves the placement of all tasks within a queue and dynamically deploys them on the Grid CPUs. When the number of tasks is greater than the number of available hardware resources, the gProcess scheduler waits until one of the worker node finishes, and assigns it with another task. The execution of the entire workflow is completed when all its inner nodes (tasks) are successfully processed.

The gProcess also offers a monitor mechanism that provides to the users information about the status of the execution. Based on this feature the users are able to remotely control the Grid executions.

The geo-spatial algorithms within the GreenLand platform were implemented based on the functionalities provided by the GRASS library. Currently, this library consists of more than 400 data processing modules that are organized in several categories: vector, raster, image processing, database management, etc.

The GRASS product uses the GDAL and OGR libraries [23] for data conversion between multiple types. GRASS modules could be accessed based on command line instructions. This way it is easier to integrate them with other GIS applications [16].

Figure 1 highlights how the previous mentioned components are related to each other. The WPS standard is implemented through the PyWPS server that could be accessed in three ways: directly from browsers by using the URL structure, from external platforms based on the API provided by the server, and from the GreenLand application.

Integrating the WPS with the Grid infrastructure generates issues that are not encountered in the traditional usage of the standard:

- Create a flexible data model that allows the development and the maintenance of the WPS processes. In order to deal with this issue, the PyWPS tool was chosen. It provides the possibility to access (via HTTP) remote algorithms that describe complex use case scenarios. The Python script is simple, and includes definition of the inputs and the output of the process.
  The remote algorithm (resident on the GreenLand platform) is the core of this flexible data model. The Python script is not affected when this algorithm changes its inner functionality, but only in cases in which the inputs or the output parameters are updated;
- In traditional usage of the WPS services, the data security is not that important. On the other hand, the Grid infrastructure filters its users, and offers

the execution rights only to the ones that have valid certificates, emitted by a Certification Authority (CA). This security issue is solved at the GreenLand platform level, where only users with valid credentials are allowed to use its functionality;

- Usually the time required to execute a WPS process is not that high. This gives the possibility of presenting the output result to the user almost instantaneously (in the same request-response loop).
  Processes that use the Grid infrastructure, demand powerful computing resources, and their execution take a long time to complete (e.g. tens of minutes). As seen, there is no possibility to provide an immediate final result to the user. Instead the GreenLand application periodically interrogates the execution status, and when completed, it offers the user the possibility to download the result, based on an URL address.

V. Implementation

This section describes the relationships between the components from Figure 1, and tries to provide the best solutions to different issues that arose on the way. The implementation of the WPS standard, within the GreenLand platform, is a 7-steps process that symbolizes each phase with a filled circle, like in the previous figure.

The first step represents the user authentication within the GreenLand platform, by using the username and password credentials. Once the user is authenticated, it is assigned with a valid Grid certificate that could be used for later executions.

The GreenLand platform, among other functionalities, provides a list of geo-spatial operators that were developed by using the GRASS library. These algorithms are from different Earth Science domains, and prove to be useful for: land cover/land use analysis (e.g. NDVI, Accuracy assessment, Density slicing, etc.), hydrologic modeling (e.g. Black Sea catchment use case study), atmospheric pollution, etc.

By default, the execution of these operators does not involve the WPS standard. Instead it is performed on a regular basis, by submitting Grid jobs through the gProcess platform in a non-standardized manner. This type of execution corresponds to the following sequence of steps: 1, 6, and 7 (see Figure 1).

One of the objectives of this paper is to prove that the WPS standard could be successfully used within the Grid parallel and distributed executions. Steps sequence: 1, 2, 5, and 7 represent an alternative WPS scenario for non-standard execution of the GreenLand processes. The main difference from the previous execution method is the involvement of the PyWPS server that allows the implementation of the WPS standard.

A. Creation of a new WPS process

The PyWPS is used for describing a standardized access method to the GreenLand operators. It uses Python scripts that allow HTTP invocation of remote services, and it consists of four sections (Algorithm 1):

- Header (lines 1-8): contains the information about the process, such as the unique identifier, the title and description, an attribute that indicates the usage of GRASS code within the script, etc.;
- Specification of inputs (lines 10-17): the name, description, and the type represent the minimal set of parameters that need to be specified for each input. There are two available types: *literalData* and *complexData*. The first one is used for sending numerical values, strings, or Boolean operators. The second type is used when the input requires the byte array of a local or a remote data source;
- Specification of the output (lines 18-21): has the same data structure as the input section;
- The body section (lines 23-29): depending on the value of the grassLocation parameter, this section contains GRASS functions, or a method that performs a remote service call to the GreenLand operators.

Algorithm 1 PyWPS exemplification for the NDVI algorithm

```
1:  WPSProcess.__init__(self,
2:      identifier = "NDVI",
3:      title="NDVI process",
4:      abstract="Computes the NDVI index.",
5:      version = "1.0",
6:      storeSupported = True,
7:      statusSupported = True,
8:      grassLocation = False)
9:  self.NIR = self.addLiteralInput(
10:     identifier="nirBand",
11:     type=type(""),
12:     title="NIR band input;image/tif")
13: self.Red = self.addLiteralInput(
14:     identifier="redBand",
15:     type=type(""),
16:     title="Red band input;image/tif")
17: self.Result = self.addLiteralOutput(
18:     identifier = "result",
19:     title="Output result",
20:     type=type(""))
21: data=urllib.urlopen('http://cgis2ui.mediogrid.utcluj.ro/GreenL
22:     andv2/executeChain&process="NDVI"&inputs=[{
23:     "type":"tif", "value":"'+self.NIR.getValue()+'"},
24:     {"type":"tif","value":+ "'+self.Red.getValue()+'"}]).
25:     read()
26:
27: self.Result.setValue(data)
```

The Normalized Difference Vegetation Index (NDVI) is a common algorithm used in the classification of the land use/-land cover from different geographic areas. Its implementation is given in Algorithm 1, and it uses the following formula in order to perform the classification process:

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{1}$$

As can be seen there is a perfect match between this formula and the Python script defined in previous algorithm. The two inputs of the algorithm represent different types of

satellite images, converted to GeoTIFF [24] by the GreenLand platform.

Usually, a satellite image contains several spectral bands. The NDVI algorithm uses the Red and NIR bands in order to create the classes of the land cover. The result of this operator generates a single-band satellite image that is available for the user to download and to perform further analysis.

After specifying the inputs and output, the Python script invokes a remote service (*executeChain*) that is hosted on the GreenLand platform. This service contains the actual implementation of the NDVI formula, while Algorithm 1 provides only a standardized access to this service. The invocation is done by using the HTTP protocol, and attaches the two parameters as additional arguments. It is worth mentioning that the *nirBand* and *redBand* identifiers store the full path to the physical resources that are going to be used at execution time.

B. WPS access types

The processes provided through the PyWPS server have three access modes: directly from the GreenLand application, API access from external platforms, and browser based access. The first mode assumes that the user is authenticated at the GreenLand level, and has access to all its geo-spatial operators. Within this platform, the user has the possibility to select one of these algorithms and to specify its inputs. After that it can establish a bidirectional communication channel with the PyWPS server (step 2 of Figure 1):

- PyWPS $\rightarrow$ GreenLand platform: the user is able to interrogate the list of existing processes (GetCapabilities) and to visualize the detailed description for each of them (DescribeProcess);
- GreenLand platform $\rightarrow$ PyWPS: in order to start a new processing over the Grid infrastructure, the user is able to invoke the Execute WPS operation. The inputs selected by the user have to be sent as additional parameters to the Python script that further passes them to the *executeChain* service that starts the Grid processing.

In some cases the inputs path to the physical resources involves a complex combination of folders and files. In such situations there is a high probability of introducing syntactical errors while specifying these paths manually.

The main advantage of accessing the WPS processes from the GreenLand platform is that it leaves no space for such errors. The user is provided with a list of aliases, instead of the real names of the satellite images. After selecting one input, the system generates in the background the entire path to that resource, and sends it to the PyWPS server.

The WPS processes could also be accessed directly from the browser (step 3 of Figure 1), by using its URL address. The main disadvantage is that the user has the full control in building the URL, including the paths to the input resources. Errors may occur in these cases, and the system is not able to provide specific support. This is the main reason why this paper recommends the usage of the first access method of the WPS services.

Using external platforms for invoking the PyWPS processes (step 4 of Figure 1) represent the third accessibility mode. It is similar with the GreenLand-WPS mechanism, but it lacks of some important features: the customized version of the metadata retrieved by using the GetCapabilities and DescribeProcess operations, the access to the GreenLand geo-spatial data repository, and the automatic generation of the input paths.

All accessing modes from Figure 1, are bidirectional. The connection to the PyWPS is used when invoking the Execute WPS operation, while the connection from the PyWPS is required to expose information about the available services (as metadata) by using the GetCapabilities and DescribeProcess operations.

C. WPS execution over the Grid infrastructure

After the PyWPS receives an Execute request with the proper inputs data set, it identifies the process and invokes the *executeChain* service that in his turn calls one of the algorithms (e.g. NDVI) available in the GreenLand repository. At this stage (step 5 from Figure 1) the next action is to use the gProcess platform in order to partition the geo-spatial algorithm into atomic tasks, and to schedule them to different Grid worker nodes.

At runtime, the process enters several stages: submitted, running, completed, and cancelled. The first stage allocates a specific number of Grid computing nodes, and sends the tasks together with their additional dependencies to these physical machines.

The running stage is identified as the actual Grid execution, where each node processes a task or a group of tasks. If the tasks are related between them, when an intermediate result is generated, it is used as input for other tasks.

When all intermediate results are available, the entire execution of the algorithm is completed. At this point, the user is able to access the results in a standardized manner, by using the WPS operations.

In cases in which errors occurred in the Grid based processing of the geo-spatial algorithms, the cancelled status is displayed. For such situations, the recommendation is to restart the execution.

In traditional WPS usage most of the algorithms need a relatively small amount of time to finish their executions (e.g. a few seconds). This allows the system to provide the WPS execution result almost instantaneously. On the other hand, the GreenLand platform addresses large scale use cases that require tens of minutes of Grid processing. Because of this aspect, a monitoring module was implemented at the gProcess level that periodically notifies the PyWPS server about the status of the execution.

The monitor module displays a "Not executed" message or a valid URL from where the user is able to download the results. If the WPS in integrated within the GreenLand platform, this link is masked under the shape of a download button. In case of direct browser access the URL in display as plain text. When accessing the PyWPS services from external platforms, the metaphor of downloading the result should be customized by certain criteria.

VI. EXPERIMENTS

This section describes the standardized execution of the Mosaic algorithm, exemplifying at each step the implementation details, from the GreenLand perspective. This proof of concept experiment is related to the validation of integrating the WPS standard within the Grid based execution processes, by using the GreenLand platform as an intermediate level. In conclusion, the experiment does not aim to measure the performance of the Grid processing, but only to validate the proposed solution.

A. Experiment description

The Mosaic is the core algorithm of the Black Sea catchment hydrologic modeling use case, which generates the shape of the entire geographic area of this catchment, by merging multiple tiles one to another. These tiles represent single-band satellite images of MODIS products (MOD15 and MOD16 in particular) that are extracted at runtime, based on the users' requests [3].

These products are multi-layers stacks of 1 km resolution provided on 8-days basis, where the maximum file size is 5.8 Mb for each. The spatial data is stored in large repositories, such as Numerical Terradynamic Simulation Group (NTSG) that offers remote access to the geo-spatial information, via File Transfer Protocol (FTP) protocol. New measurements are added periodically by means of satellite sensors that provide raw data that is pre-processed and then inserted into these repositories.

For an easier identification of these data, the NTSG consortium partitioned the Earth surface into horizontal and vertical tiles. For this experiment only the tiles that cover the Black Sea catchment are needed. The MODIS data is organized in years, starting from 2000 until the current time. Because it is an 8-days temporal resolution product, each year contains day-folders (e.g. D001, D009, D0017) indexed from 1 and increasing up to 361.

Taking into account all these aspects, the Mosaic implementation requires an automatic data retrieval script that transfers at runtime the relevant information from remote repositories to the Grid worker nodes.

B. Selecting the inputs data set

Before executing a WPS process, the user has to query the list of available services. This is possible by performing the GetCapabilities request that will return a XML metadata file. Due to the lack of space, Figure 2 highlights only a fraction of this document that contains the description of the processes. Among other details, the metadata contains information about the provider of these services and the Web addresses to the WPS operations.

In order to perform the experiment, the Mosaic process is going to be selected. The next step is to get a detailed description of its inputs and output, by using the DescribeProcess request. The ows:Identifier attribute is used as a unique identifier of the process.

The product type (MOD15 or MOD16), a list of bands for each product (e.g. Evapo-transpiration - ET, Leaf Area Index - LAI), and the processing year (e.g. 2000, 2001, ..., until current year) are the input parameters that are requested by the WPS Mosaic process. For a better understanding of this action, Figure 3 highlights the structure only for the year input of the process.

```
<wps:ProcessOfferings>
    <wps:Process wps:processVersion="1.0">
        <ows:Identifier>NDVI</ows:Identifier>
        <ows:Title>NDVI process</ows:Title>
        <ows:Abstract>Computes the NDVI </ows:Abstract>
    </wps:Process>
    <wps:Process wps:processVersion="1.0">
        <ows:Identifier>Mosaic</ows:Identifier>
        <ows:Title>Mosaic process</ows:Title>
        <ows:Abstract>Computes the Mosaic</ows:Abstract>
    </wps:Process>
                        . . .
</wps:ProcessOfferings>
```

Fig. 2. GetCapabilities metadata file

```
<DataInputs>
    <Input minOccurs="1" maxOccurs="1">
        <ows:Identifier>year</ows:Identifier>
        <ows:Title>Year</ows:Title>
        <LiteralData>
            <ows:DataType ows:reference= "http://www.w3.
org/TR/xmlschema-2/#string"> string </ows:DataType>
            <ows:AnyValue/>
        </LiteralData>
    </Input>
                        . . .
</DataInputs>
```

Fig. 3. DescribeProcess metadata file for the Mosaic workflow

The GreenLand platform is able to parse the DescribeProcess metadata, and to present it to the user in a more user friendly manner, by means of combo boxes, check boxes, buttons, dynamic text, etc.

The next action is to initiate the Grid execution, through the PyWPS server. This is possible by using the WPS Execute operation. The URL of this action is automatically generated in the background, based on the values selected by the user at the graphical level of the GreenLand application.

Let's assume that we want to process the LAI band from the MOD15 product, ET band of the MOD16 product, and set the processing year to 2010. The URL for the Execute operation has the following structure: http://<server_domain>/wps/wps.py?service=wps&version=1.0.0&identifier=Mosaic&request=Execute&datainputs=[year=2010;mod15=LAI;mod16=ET]. The <server_domain> represents the URL location of the PyWPS server, while the datainputs field contains all the parameters specified by the user.

C. Grid based execution

The Execute request is received by the PyWPS sever that interprets it, and assigns each parameter from the *datainputs*

to local variables in the Python script. Such an example was described in Algorithm 1 for the NDVI process.

The Mosaic is a GRASS based algorithm that is resident on the GreenLand platform. Once the WPS execution request reaches the PyWPS server, it is redirected to this geo-spatial algorithm. Among the inputs list, additional information needs to be sent, such as: a unique identifier, a short description, the type required for each input together with its entry from the GreenLand database.

Until this stage, the Mosaic experiment was performed based on the rules and guidelines provided by the WPS standard. The next steps represent the actual Grid execution that is outside of the standard scope. This execution starts at the gProcess level that is responsible for partitioning the Mosaic algorithm into atomic tasks.

The one year processing contains multiple executions of data with a time delay of 8 days (the NTSG data repository updates every 8 days with new satellite images). So, in one year we have 365 days that divided by 8, results 45 executions. Because the user selected two MODIS products (MOD15 and MOD16) there will be 90 independent executions.

Allocating one Grid worker node for each task is not efficient, because its execution is not that time consuming. In order to optimize the entire processing flow, the gProcess platform creates groups of nine tasks, where each group is going to be executed on a single worker node.

After scheduling the entire Mosaic process, transferring the data to each Grid machine is the next step. Initially, data is resident on remote repositories and needs to be copied at runtime to the worker nodes. These repositories contain horizontal and vertical data tiles that cover the entire Earth surface. For this experiment only 12 tiles (the ones from the Black Sea catchment area) are needed.

Each of the 90 executions will process all 12 tiles, where their results will provide a good indicator for hydrologic prediction in the Black Sea catchment region.

Usually, the Mosaic process requires two hours to complete its execution. During this time, the user needs to know the status of the Grid execution. The gProcess platform provides a monitor mechanism that periodically interrogates the execution state, and sends the feedback to the GreenLand platform that in its turn displays it to the user in an easy to understand manner.

When the Grid execution completes, the user is able to download the results or to perform further operations. Figure 4 presents one of the results generated by processing the Mosaic algorithm that is partitioned into 12 tiles that cover the Black Sea catchment geographic area.

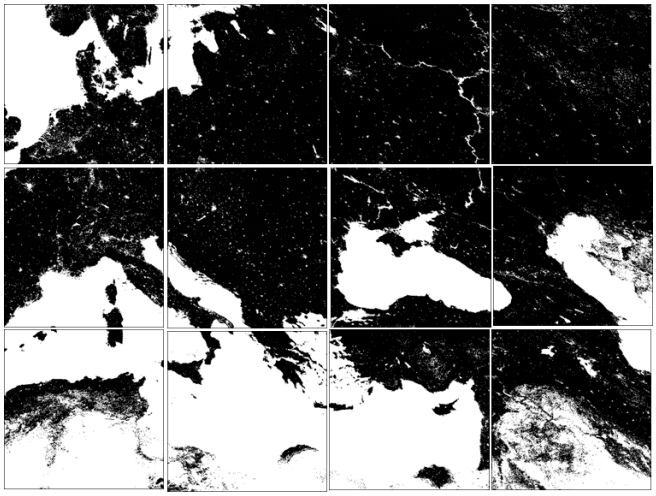


Fig. 4. The result generated by the Mosaic operator

VII. Conclusions

This paper aims to integrate the WPS standard for executing the GreenLand algorithms over the Grid infrastructure. The Mosaic experiment proofs that this approach is valid and could be extended for other types of algorithms.

The other two objectives (different WPS access modes and the flexible implementation of WPS processes) were also described, by highlighting the possible implementation solutions.

The OGC consortium does not offer enough implementation details for all situations. For example, applying the WPS standard in Grid computing is difficult due to the fact that the Grid infrastructure allows only authenticated users to perform different operations. One recommendation is to extend the current standard to also enhance these use cases.

Acknowledgment

This research is supported by the enviroGRIDS Project funded by the European Commission, through the Contract 226740.

References

[1] D. Mihon, V. Colceriu, F.B. Balcik, K. Allenbach, M. Gvilava, D. Gorgan, "Spatial Data Exploring by Satellite Image Distributed Processing", Geophysical Research Abstracts, EGU General Assembly 2012, vol.14, pp.13278, 2012.

[2] enviroGRIDS project, http://envirogrids.net/

[3] B.F. Bektas, C. Goksel, S. Sozen, K. Allenbach, M. Gvilava, K. Rahman, D. Gorgan, D. Mihon, "Remote Sensing Services - ESIP Platform and Hot Spot Inventory Case Studies", enviroGRIDS Deliverable D2.11, 2012. Available at: http://envirogrids.net/index.php?option=com_jdownloads&Itemid=13 &view=finish&cid=139&catid=11

[4] Open Geospatial Consortium, (2007), "OpenGIS Web Service Common Implementation Specification", pp.1-153.

[5] L. Diaz, C. Granell, M. Gould, "Case study: Geospatial Processing Services for Web- based Hydrological Application", Book chapter: Geospatial Services and Applications for the Internet, pp.31-47, 2008.

[6] J. Cepicky, "PyWPS 2.0.0: The Presence and the Future", Geoinformatics, 2007, http://geoinformatics.fsv.cvut.cz/gwiki/PyWPS_2.0.0:_The_presence_and _the_future

[7] G. Aloisio, M. Cafaro, "A Dynamic Earth Observation System", Parallel Computing, vol.29, pp.1357-1362, 2003.

[8] D. Gorgan, V. Bacu,T. Stefanut, D. Rodila, D. Mihon, "Grid Based Satellite Image Processing Platform for Earth Observation Application Development", Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, pp.247-252, 2009.

[9] D. Gorgan, V. Bacu, D. Mihon, D. Rodila, T. Stefanut, K. Abbaspour, P. Cau, G. Giuliani, N. Ray, A. Lehmann, "Spatial Data Processing Tools and Applications for Black Sea Catchment Region", International Journal of Computing, vol.11 (4), pp. 327-335, 2012.

[10] Y. Wang, S. Wang, D. Zhou, "Retrieving and Indexing Spatial Data in the Cloud Computing Environment", 1st International Conference on Cloud Computing, pp.322-331, 2009.

[11] G. Giuliani, S. Nativi, A. Lehmann, N. Ray, "WPS Mediation: an Approach to Process Geospatial Data on Different Computing Backends", Computers and Geosciences, vol.47, pp.20-33, 2011.

[12] Z. Chen, N. Chen, C. Yang, L. Di, "Cloud Computing Enabled Web Processing Service for Earth Observation Data Processing", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol.5, pp.1637-1649, 2012.

[13] A. Wytzisk, M. Gould, F. Holsmuller, "Research and Instruction Mixing Commercial and Open Source Tools", 14th Workshop AGILE International Conference on Geographic Information Science, 2011.

[14] Degree Webservices documentation, 2013, http://download.deegree.org/documentation/3.2-pre13/deegreeWebservices.pdf

[15] Sextante documentation, 2011, http://gvsigce.sourceforge.net/sextante_web/pdf/SextantePaper.pdf

[16] M. Neteler, H. Mitasova, "Open Source GIS: A GRASS GIS Approach", Kluwer Academic Publishers/Springer, Boston, Dordrecht, and London, 2004.

[17] D. Wesloh, J. Sturley, "Implementing ISO Data Quality Standards Using ESRI's GIS Data ReViewer", National Geospatial-Intelligence Agency, 2004.

[18] FGDC Facilities Working Group, "Spatial Data Transfer Standard (SDTS). Part 7: CADD Profile", Federal Geographic Data Committee, 2000.

[19] OASIS standard, http://www.oasisopen.org

[20] J. Brauner, T. Foerster, S. Bastian, B. Bastian, "Towards a Research Agenda for Geoprocessing Services", 12th AGILE International Conference on Geographic Information Science, Hannover, Germany, pp.2-12, 2009.

[21] Z. Young, I. Raicu, S. Lu, "Cloud Computing and Grid Computing 360-Degree Compared", Grid Computing Environments Workshop, pp.1-10, 2008.

[22] V. Bacu, T. Stefanut, D. Rodila, D. Gorgan, "Process Description Graph Composition by gProcess Platform", HiPerGRID - 3rd International Workshop on High Performance Grid Middleware, Proceedings of CSCS-17 Conference, vol.2., pp. 423-430, 2009.

[23] T. Mitchell, "Web Mapping Illustrated", O'Reilly, First edition, ISBN 978-0-596-00865-9, pp.349, 2005.

[24] J. Rhodes, C. Bailey, P. Brown, "FalconView GeoTIFF Profile", Georgia Teach Research Institute, pp.1-25, 2006.

Grid Based Processing of Satellite Images in GreenLand Platform

Danut Mihon, Vlad Colceriu, Victor Bacu, Dorian Gorgan
Computer Science Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
{vasile.mihon, vlad.colceriu, victor.bacu, dorian.gorgan}@cs.utcluj.ro

***Abstract*—Geographical Information System (GIS) applications that process large amount of data require intensive usage of hardware capabilities provided by distributed platforms, such as the Grid infrastructure. Due to the constant demand of data availability and data sharing, without concerning its format and size, a new software solution is needed. GreenLand is a system capable to provide such a solution, based on its constituent modules: GreenLandGUI, gProcess, ESIP, WorkflowEditor, and OperatorEditor. This paper highlights each of them and how they interact in order to create a platform capable of fetching, processing, and visualizing large amount of data exposed in a uniform and standardized manner.**

I. Introduction

The description and processing of natural phenomena and experiments, from different domain fields, is a complex process that usually involves: a solid understanding of the background context, the collection of the adequate input data set, the syntactic and semantic description of the adopted solutions, the execution over distributed environments in order to speed up the entire process, optimized tools for partial results integration, and some special interaction techniques for visualizing and analyzing the final outputs.

This paper describes the theoretical concepts and practical solutions involved in solving the previous mentioned issues, through the perspective of the GreenLand platform [1]. This system was developed within the enviroGRIDS project [2], and its functionality was validated through three case studies: Black Sea catchment hydrologic modeling, land cover/land use analysis of the Istanbul geographic area, and the Rioni river hydrologic analysis [3].

Modeling large scale environmental use case scenarios is most of the times a challenging task, due to the multitude of conditions, restrictions, and algorithms that need to interconnect in order to provide the desired output. Regarding this aspect, the new Geographic Information System (GIS) applications try to provide advanced interaction techniques that facilitate the end-user work and increase the usability of the entire platform.

In order to overcome these issues, the adopted solution was to represent the entire use case as a workflow, where each node identifies one of the algorithms (function) of the main process. The uni-directional edges of the graph specify the interaction between the algorithms and how they communicate in order to generate the output results.

This type of approach is useful in many cases, but when the user is required to manually specify all the connections, errors may occur. This is the main reason why the GreenLand platform provides the WorkflowEditor tool [4] for an easy and flexible description of the workflows.

Executing such large use cases on standalone machines is not a feasible solution. On the other hand, the correct approach is to use the storage and computation benefits of the distributed infrastructures (e.g. Grid, Cloud, clusters, multi-core machines). This way an execution speed up will be obtained, by partitioning the main process into smaller tasks and execute them in parallel.

The GreenLand platform uses the Grid infrastructure [5] in order to improve the execution time, where each node (or a group of nodes) of the workflow is processed onto a different physical machine. The gProcess platform [6] connects the two environments and acts like a middleware between them. The input data set and the expanded structure of the workflow are the only information required by this platform.

Based on the process complexity, the gProcess is able to group the tasks and to discover the optimal execution schema. Monitoring the Grid-based processing and sending the feedback to the GreenLand system is another feature offered by this platform.

In order to provide useful results that could be reused by external applications (without further processing) the GreenLand platform implements the WMS, WCS, and WPS OGC standard services [7]. They allow the satellite images retrieval and exposure in a standardized manner, and facilitate the user actions regarding this types of tasks.

II. Related work

The availability of high performance applications, broadband Internet access, high storage and processing capability devices, and the Web technologies accelerate the usage of geographic information into our daily lives. GIS applications are widely spread across Earth science domains, such as: hydrology, meteorology, agriculture, air and water pollution, urban planning, etc. They offer standard services for storing, processing, analyzing, and visualizing spatial data of different types and formats.

Some of the most known such platforms work either on standalone or distributed infrastructures. In the first category we can include Sextante [8], uDig [9], and GRASS [10]. As

for the distributed environments class, the QuantumGIS [11] tool could be mentioned.

The Sistema EXTremeno de ANalisis TEritorrial (Sextante) is an open source spatial data analysis library that contains more than 300 geospatial algorithms that handle raster and vector data types, and provides rich common functionalities, useful for the entire geospatial communities. It allows the creation of complex workflows, in an interactive manner, but it does not support the sub-workflow concept (nodes imbrications within other nodes) as the GreenLand platform does.

The main goal of the uDig is to fill the functional gaps between the geospatial standards and the open source communities. It provides integration support with the latest Open Geospatial Consortium (OGC) standards, and it is mostly used to represent database geospatial information in a simple and interactive manner. Similar with this tool, the GreenLand platform adheres to the latest OGC standards, by offering support in data retrieval, execution, and visualization.

The Geographic Resources Analysis Support System (GRASS) is based on GDAL and OGC libraries, and provides features for reading and writing various raster and vector data formats. It offers more than 400 geospatial algorithms and it can be easily implemented in other platforms (this is the case of the GreenLand system that integrates its functions directly within the Web services, consumed by the end-user).

The Quantum GIS system is useful for spatial data processing, displaying data layers over interactive maps, performing distance measurements, creating map symbologies, data reprojection, etc. Another important aspect is the support it offers for distributed and parallel computations, in case of large experiments. The workflow-based description of the scenarios is the main advantage of the GreenLand platform, and proves useful especially when dealing with a large set of algorithms that need to be connected by certain rules.

The GreenLand platform allows the parallel and distributed execution of the tasks, and benefits from the computing and storage characteristics of the Grid infrastructure. One of the important advantages of this solution (compared with the previous mentioned environmental applications) is the execution speedup, obtained for large scale use cases (experiments). Because the entire process is partitioned into multiple tasks, the system is able to schedule them onto different physical Grid nodes. This means that the total processing time is significantly reduced, the only overhead appears when transferring input data set and combining the partial results in order to generate the final output.

The ability of executing the use cases over the Grid infrastructure is the main feature that differentiates the GreenLand platform from the previous mentioned standalone applications, and makes it suitable for implementing large environmental scenarios from different Earth Science domains.

The flexible and interactive use cases description is the main advantage of the GreenLand platform compared with the QuantumGIS application. Instead of independent execution of all the inner algorithms, this solution allows the relationships definition between them and the possibility to create a single execution thread for the entire workflow.

The gProcess platform is used as a middleware between the Grid infrastructure and the user requests, and provides support for: workflows partition into tasks, scheduling mechanisms, and execution and monitor features. The GANGA [12] and DIANE [13] tools represent two of the alternatives to this approach. The first one is a job management tool, capable of scheduling the entire execution process. On the other hand, the DIANE is mostly used for monitoring the processing, and gives periodic feedback about its status (e.g. the number of executed jobs, on what Grid nodes the tasks are resident, etc).

The main advantage of the gProcess platform (compared with the features provided by these two alternative applications) is the ability to interpret the workflow-based data structures, and to create groups of nodes, similar in complexity. This way a balanced Grid execution is obtained.

III. Theoretical concepts and implementation

This section highlights the main concepts related to the possibility of describing the spatial data execution process as complex workflows that encapsulate within their nodes an abstract representation of an algorithm, function, or experiment.

A. Spatial data classification

Based on the data structure and on the collection mechanisms, the spatial data are grouped into: satellite images, airborne images, and ground data measurements.

The satellite images are obtained onboard the artificial satellites that orbit around the Earth, collecting information about its surface (e.g. temperature, humidity) by scanning it in multiple frequency levels. The collected data are organized in bands that contain on each layer one of the measured characteristic. The GreenLand platform supports various satellite images, regardless of their number of bands: Landsat (organized on 7 layers), ASTER (15 bands), 36 levels MODIS images, etc.

The airborne data are useful in applications that require high accuracy results, because these images scan the Earth's surface in more detail. Some of the most known products (e.g. SPOT and QuickBird) are also supported in the GreenLand framework.

The information obtained from ground based measurements has the best accuracy and penetrate in dense areas where the artificial sensors are not able to record the data. They are used especially for calibrating different experimental models, related to a small geographic region (due to the limited measurement capacity).

The GreenLand platform offers support for all these data categories, but in this paper only the satellite images are presented in more detail, due to the requirements of the three case studies highlighted in the introduction section.

The GreenLand platform uses the workflow concept for use cases development. The physical execution of such graphs can be defined as a multi-variable function P that produces, in a finite amount of time, a valid result, based on a specific input data set. It also contains sub-processes (represented as the nodes of the graph) combined in a specific order that corresponds to the use case description flow. Two types of

processes were identified in the context of the GreenLand platform: basic operators and complex workflows.

B. Basic operators

The operator is the smallest unit that can be processed, without the possibility to divide it into atomic modules. It integrates the representation of an algorithm (e.g.: vegetation indices, atmospheric correction functions, statistics computation, distance measurements, etc.) under the form of an executable file that is further used at runtime over the computing infrastructures.

A formal description of the basic operator is given bellow,

$$O(IN, OUT, DATA) \tag{1}$$

where:

- $IN = \{in_1, in_2, \ldots, in_n\}$: all the available inputs data set;
- OUT: the output of the operator;
- $DATA = \{d_1, d_2, \ldots, d_m\}$: all the available data resources that can be used for inputs instantiation.

Each input ($in_k, k = \overline{1,n}$) and the output is a triplet <name, value, type> that has a name, a value (or resource from the *m* possible entries), and an associated type. The order in which the inputs are specified has a major impact on the final result of the core process. In this case, there should be a perfect match between the *n* arguments and the variables of the algorithm described by the operator

C. Complex workflows

The description of natural phenomena (experiments, use cases) that belongs to different Earth Science domains can be modeled as workflows (graphs) that contain a collection of operators, interconnected by uni-directional edges. Using this approach, we can achieve the goal of optimal representation and data model organization of the natural phenomena.

From mathematical point of view, the workflows can be described as in (2)

$$W(IN, OUT, DATA, N, C) \tag{2}$$

The first three arguments of the function *W* have the same significance as in the case of the basic operators. The only difference is the fact that the workflows allow the possibility to specify multiple outputs ($out_1, out_2, \ldots, out_s$), compared to a single operator's output. Information about the inner layout of the graph is stored in the last two arguments of the function:

- $N = \{n_1, n_2, \ldots, n_u\}$: a finite list of nodes that, in the basic form are identified as operators. A more complex node is called sub-workflow that has the ability of storing other graphs within;
- $C = \{c_1, c_2, \ldots, c_v\}$: a list of uni-directional edges that describe the execution flow inside the graph.

The conceptual representation of a complex workflow is described in Figure 1. As can be seen, it contains *u* operators

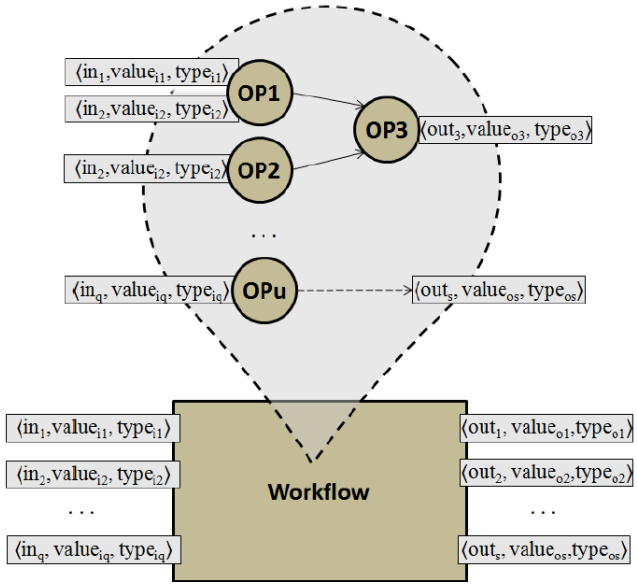


Fig. 1. The abstract representation of the workflow

(marked with $OP1, OP2, \ldots, OPu$). The inputs list of the workflow is distributed to its inner basic operators. This part is extremely important because it influences the final output result. For example, if we switch the inputs of the $OP1$ and OPu than the obtained results are different from the original ones. This change propagates to the next description levels ($OP3$ in this case) and affects the out_3 and *outs* of the core workflow.

The example from Figure 1 highlights only the mathematical significance of the concepts, but at runtime, these inputs are instantiated with the data specified by the end-user. In the GreenLand framework multiple data types are supported, such as: generic satellite images (e.g. Landsat, MODIS, Aster, etc.), vector shape files, projection files, integers, strings, etc.

Another important aspect is the connection establishment between the operators, because it describes the entire execution flow of the use case (scenario). Even though the GreenLand platform offers support for multiple users' categories (e.g. data providers, decision makers, specialists in Earth Sciences, regular users), this step is recommended to be realized by a domain field specialist.

To exemplify the basic operator and workflow concepts, the Istanbul case study is very useful. Shortly, this experiment consists in classifying the vegetation, water, and urban areas around the geographic region of Istanbul, by implementing multiple algorithms that interconnect at four stages: spatial data pre-processing, vegetation index computation, satellite image classification, and the accuracy statistics generation.

The algorithms that are used in each stage can be defined as basic operators (e.g. geometric correction of satellite images, NDVI, EVI, Density slicing, etc.). The entire Istanbul scenario can be described as a workflow, where the nodes are identified as operators and the data flow process is described through uni-directional edges.

One important aspect is the fact the GreenLand platform limits to the acyclic graph structure. This means that the system

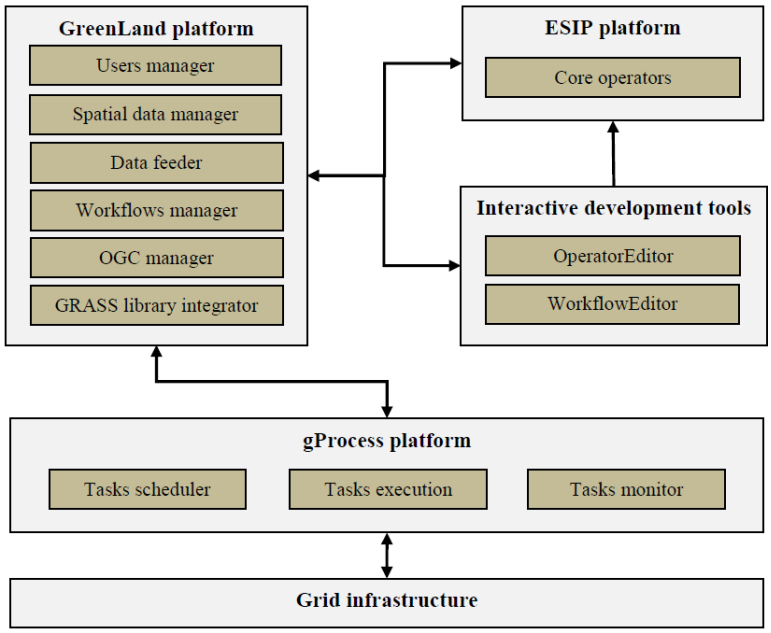


Fig. 2. System related architecture

is protected from infinite looping cycles. One future research direction is to offer support for repetitive structures (e.g. for, while) and the conditional ones.

D. System related architecture

The GreenLand platform is built upon other modules that interconnect in order to fulfill the satellite images processing requirements (Figure 2). The main entry point is the GreenLand system that listens for the user-actions at the graphical interface level. Once a new event is triggered, it is automatically interpreted and converted into an internal representation. In cases that involve satellite images processing, the system also stores the operators and the workflows specified by the user together with their input data set.

The Environment oriented Satellite Image Processing (ESIP) is responsible for providing the core algorithms (operators) to the GreenLand platform. It contains only the main operators that are useful in different Earth Science domains, while the GreenLand allows the development of customized operators and workflows (based on the OperatorEditor and WorkflowEditor tools).

This delineation is also useful when installing the platform on other environments, because only the ESIP is exported together with the data schema, while the GreenLand is ported with an empty data repository.

Once the use case (scenario) development process is complete, it can be executed over the Grid infrastructure. This is possible based on the gProcess platform that acts like a middleware, translating the client requests into commands recognized by the Grid environment. It is also used for sending the execution feedback to the application's graphical user interface.

When the execution is finished the final output can be analyzed by using the online specific tools or it can be downloaded onto the user's local machine. The GreenLand platform offers support for visualizing these results in an interactive manner and promotes data sharing with external systems and applications.

1) GreenLand general overview: It is a GIS platform that provides services for geospatial data retrieval, processing, and visualization. The frontend of this platform acts like a gateway that masks all the complex mechanisms that are implemented within the system, such as: workflow partition into tasks, the scheduling process, Grid based execution, data interoperability with external platforms, standards implementation, etc. [1].

The GreenLand is an open platform that allows data import from three main sources: directly from the user local machine (regular upload), from File Transfer Protocol (FTP) data repositories, and by OGC means. Depending on the requirements, the users are able to utilize one of these methods, or to combine them as desired.

In case of near real time processing algorithms the idea of automatic data fetching from different remote repositories is very useful. This feature is especially used in prediction experiments that require a large data set for the calibration process. The GreenLand platform allows the automatic data extraction, based on the FTP protocol.

The Mosaic Black Sea catchment workflow is a perfect example to offer insights about this solution. The user is requested to specify the remote repository that stores these data, the processing time period, and the satellite image bands that he is interested in. Once these steps are completed the system processes the user's request and automatically starts collecting data, by applying the filters selected at the GreenLand graphical interface level.

Because various spatial data types are used in the GreenLand framework, new scripts were needed in order to interpret and process these data. GRASS library proved to be flexible

enough to allow its functionality extension to fulfill the GreenLand requests.

On the other hand the GRASS library fits perfectly on standalone platforms, but needs several adjustments in order to run properly on distributed environments.

The adopted solution was to identify the GRASS version that comply the best with the Grid infrastructure, and to describe all the operators based on that type of library. At runtime, the GRASS files were packed together with the inputs specified by the user and transferred to the Grid machines. Locally, on each Grid node, the operators (algorithms) are executed similar to standalone machines.

One of the GIS fundamental ideas is to develop an open platform that is able to interoperate with external systems in terms of data sharing. Based on the OGC standard that was implemented within the GreenLand framework, the system is able to: query and download remote data directly in the GreenLand repository (through WCS operation), interactively visualize spatial information (based on the WMS service), execute the scenarios in a standardized manner (by using the WPS service), and to publish the Grid processing results to external remote storages.

The flexibility characteristic is another main aspect that was taken into account when developing the GreenLand system. First of all it can be used as a Web-based platform. In this case the users are able to perform different actions directly from the application frontend, in an interactive and user friendly manner. The complexity of the internal mechanisms is hidden from the users, and only light weighted operations are exposed.

Extending the GreenLand functionalities to other activity domains (e.g. archeology, physics, etc.) represents the second utilization mode of this platform. This feature can be achieved by integrating the constituent services directly into the backend architecture of other systems. Because of the GreenLand flexibility, the modules described in Figure 2 do not necessarily need to work in their original schema. Instead their installation can be extended to different physical machines (e.g. the operators repository can be resident on other servers).

Platform interoperability is the third way of using the services exposed by the GreenLand system. Because it implements the OGC standard, external applications are able to connect to the GreenLand (by means of standard services) in order to: query, visualize, and download the satellite images made available by their owners, and to process the GreenLand workflows exposed as WPS items

2) ESIP platform: The Environment oriented Satellite Image Processing (ESIP) [14], [15] can be defined as a set of basic operators (e.g. radiometric correction, vegetation index computation, histogram generation, mathematic computations, etc.) that handle various types of data, such as: satellite images, vector data, ground based measurements, etc.

The GreenLand platform provides services for the GIS domain. By default, when a fresh copy of the platform is installed, it contains a predefined basic operators set, resident in the ESIP platform. As the system develops, new operators can be added to the ESIP repository.

The content of this platform is rarely updated, and once an operator is implemented it is recommended to maintain its functions (because it may be already used in other workflows and the change of its internal structure will affect the entire data flow). Adding new operators to this repository can be done interactively, through the OperatorEditor tool (Figure 2). More details about this application are presented in the next section.

In other words the ESIP platform is recommended to be used as a repository of operators that provides information to different instances of the GreenLand system.

3) Interactive development tools: There are two interactive applications (OperatorEditor and WorkflowEditor) [4] which are integrated within the GreenLand platform and used for basic operators and workflows development. They are called interactive because they facilitate the entire implementation process, allowing the users to easily describe the inner functionalities of the algorithms (as operators) and complex use cases (as workflows).

There are several important characteristics about the operator concept: a list on inputs, an output, and the inner algorithm (function or formula) that describes the operator's behavior. Each input/output has the triplet form (<name, value, type>) that makes it easier to distinguish among other items, and to map various data formats.

The OperatorEditor tool allows the user to describe the inner functionality of the operator by extending a specific Java API. The resulting algorithm has several input variables and one output that stores the result generated when instantiating the algorithm with the input data resources.

Once the operator is implemented its owner has the possibility to make it available to other users. These users do not have access to the kernel of the algorithm and do not know what inputs it expects and what its functions are.

For this reason the OperatorEditor tool provides to the owner of the operator an interactive mapping technique for specifying all these features. As can be seen in Figure 3 the user is able to specify the operator's name and a short description. It is also recommended to give a full description of its functionality and to attach it to the operator by means of external files (PDF in this case). Once the operator is completed, it can be shared with the entire users' communities (by making it public).

The user is also able to map the algorithm's inputs and output to the operator's ones, by using the same interaction technique as the one presented in the bottom side of Figure 3. The order of the inputs must completely match the order in which they were specified within the Java algorithm, otherwise the final output result will be altered.

As mentioned in previous sections, the GreenLand use cases are described as workflows. When this process is done manually (e.g. by means of XML tags) it is most likely that errors may occur. The WorklfowEditor tool was developed in order to avoid such issues and to facilitate the workflows implementation by providing: several interactive techniques, validation mechanisms, layout algorithms, and proper adjustments performed automatically by the system itself.

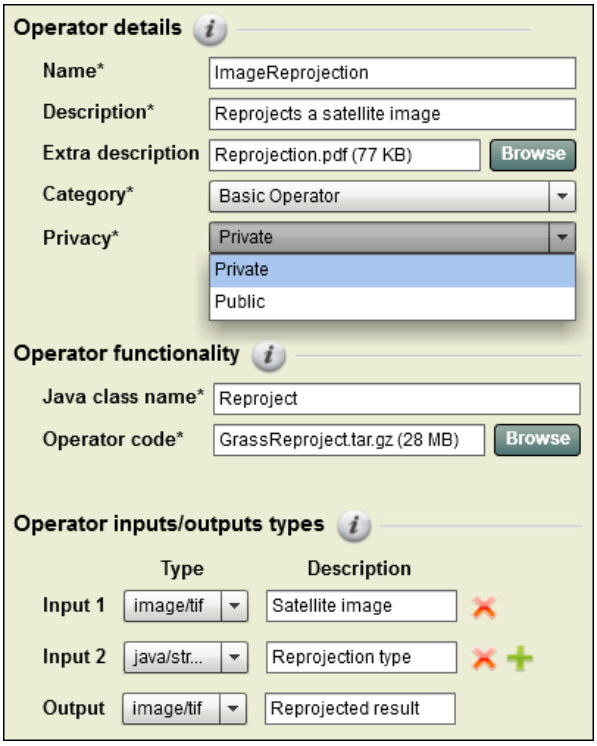


Fig. 3. Interactive description of the ImageReprojection operator

Usually each workflow (Figure 4) contains a list of operators (marked with the circle graphical symbol) and sub-workflows (highlighted as rectangles) that integrate other inner nodes. The imbrications can extend to multiple levels, while the user has the possibility to navigate through these structures by using the mouse device.

On the other hand the workflow development process itself is highly interactive, and includes:

- Placing the graph nodes with the drag and drop actions;
- Connecting the items by tracing a uni-directional edge with the mouse (once a node is selected the system enables only the inputs that have the same data type). At this stage it is worth mentioning that only nodes that have the same type are allowed to be connected;
- Nodes reposition on the canvas surface, while the corresponding edges update automatically;
- Navigating through the sub-workflows hierarchy;
- Auto-arrangement of the nodes by using one of the automatic workflow layout algorithms that minimize the surface on which the workflow is represented and reduces as much as possible the number of intersections between the edges. Creating edges of similar lengths and preserving their angular resolution are also taken into account when using the automatic layout algorithms.

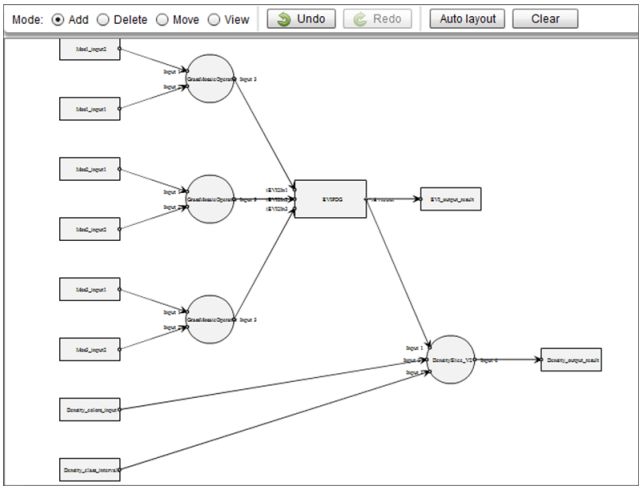


Fig. 4. The workflow development process based the WorkflowEditor tool

Once the workflow development process is complete, it becomes available in the GreenLand platform, and can be used in further Grid processing. The inputs of the workflow represent the inputs collection of all its internal nodes, excepting the case when one input is connected to the output of another item.

4) gProcess platform: Once the workflows are instantiated with the inputs specified by the user at the graphical interface level, the Grid execution begins. The gProcess acts like a middleware that interprets the processing user-requests, converts them into an internal representation structure, and forwards them to be executed over the Grid infrastructure [6].

The communication with the gProcess platform is accomplished by using a customize XML format that contains the description of the entire GreenLand workflow. The XML structure is slightly different from the one generated by the WorkflowEditor tool, meaning that the sub-workflow concept is not included. Instead, each node of the graph is represented on the same hierarchy level.

The main advantage of this representation approach consists in the possibility of creating execution groups (that run on different Grid machines) in order to balance the processing of the entire workflow.

Taking into account this aspect, we can say that the gProcess platform is able to partition an execution graph into smaller tasks that are interconnected upon the relations specified within the XML structure. Each task is then submitted to a specific Grid machine, together with all its input data and additional dependencies.

The gProcess platform is also responsible for monitoring the entire Grid execution process. When interconnected tasks are executed on multiple machines, it is most likely that one node has to wait for the other one to finish. In this case the gProcess is involved in managing the data transfer between the two entities, and to generate the final output of the workflow based on these partial results.

Each gProcess task is considered to have one of the following statuses: submitting, running, completed, canceled, and failed. The input data transfer to the Grid nodes takes place in the submitting stage. At this moment the gProcess is also partitioning the workflow into tasks, and schedule their execution by mapping each task on a specific Grid node.

It is worth mentioning that initially a list of all available Grid machines is retrieved and tasks are assigned to consecutive worker nodes from this queue. If the number of tasks is greater than the length of the list, the remaining processes are marked with the pending status. Once a Grid machine completes, it receives a new task to process. The running stage consists in executing the workflow modules over the Grid infrastructure. When a task finishes, its output is automatically transferred to other nodes that require this information as input.

The workflow execution completes when all its tasks are processed correctly. The final result of the workflow is generated by combining the partial outputs of each task, based on the XML representation file.

The failed status identifies an error that was encountered during the execution process. The user is also able to stop the Grid based workflow processing. In this case the gProcess will mark this execution as cancelled.

5) Grid infrastructure: It can be described as a worldwide computer network that offers support for storing and processing large volume of data. The storage nodes are called Storage Elements (SEs), while the computational stations are referred as Computing Elements (CEs) [5].

The motivation behind using the Grid, as a processing infrastructure for the GreenLand platform, is that in case of complex scenarios the standalone machines do not provide enough computation power to execute them in reasonable amount of time. In order to speed up the entire execution process, this platform benefits from the Grid parallel and distributed capabilities regarding the large data processing.

On the other hand the GreenLand platform can also be used for executing the basic operators that are simpler algorithms that take a few seconds to compute. In these cases the Grid infrastructure is not needed, because it will slow down the entire computation process, taking into account:

- The time required to partition the workflow into tasks and to schedule them onto the available Grid nodes;
- The time required for transferring the input data sets, together with the additional dependencies;
- The actual Grid execution of all tasks and the final output generation, based on the partial results of each task;
- The time required to transfer the workflow result from the SE node to the GreenLand server and to make it available to the user.

In order to avoid using the Grid for unnecessary executions, one of the research directions for extending the GreenLand platform is to implement a decision module that is able to redirect the processing (to Grid or multi-core infrastructures) based on the complexity of the workflow. This research is only at the beginning, but it proves to be useful in increasing the platform flexibility and scalability.

IV. Experiments

This chapter exemplifies the theoretical concepts, described in the previous sections of this paper. The goal of the conducted experiment was to analyze the water quality/quantity for the Black Sea catchment in the last 10 years. The MODIS satellite images were used as input data sets for this use case. In order to simplify the entire execution process, one additional request was to automatically collect the data from remote repositories, by keeping the user graphical interface as simple as possible [3].

A. General description

The MODIS satellite produces data by scanning the Earth's surface on an 8-days time basis. This sensor partitions the entire Black Sea catchment into 12 adjacent tiles, represented as satellite images. Regarding all these aspects a new GreenLand workflow was needed in order to:

- Recompose the Black Sea catchment area from the 12 adjacent tiles, and apply the analysis algorithms on the extended model;
- Automatically collect MODIS satellite images from remote repositories, over a specific time period;
- Handle both MOD15 and MOD16 products. The differences between them are the internal bands organization and the data contained within each frequency interval;
- Extract information relevant to the use case requirements. Because the MODIS data are organized in multiple bands, only specific information is required for this particular experiment (e.g. the Evapo-Transpiration, the Photosynthetically Active Radiation, etc.);
- Optimize the entire execution process by performing parallel computations over the Grid infrastructure;
- Expose the results to external platforms, by using the OGC standard.

B. GreenLand workflow development

A new workflow was implemented (called BlackSeaMosaicPDG or Mosaic12) that based on 12 MODIS satellite image input generates a single model for the entire Black Sea catchment. The internal algorithm is based on the classical Mosaic operator that combines 2 bands in order to generate a single satellite image, containing the extended area.

On the left side (in Figure 5) there are 6 Mosaic operators that receive the 12 input images. Each of the next levels reduces the number of the operators, until the final result image is generated. The inputs order is important and has to match the horizontal or vertical position of the adjacent tiles. The Mosaic12 workflow can be created directly from the WorkflowEditor tool that allows the interactive placement of the operators and the specification of the inter-nodes relations by using the mouse device.

C. Input data specification process

The main goal of this experiment is to model the Black Sea catchment area, based on information dated from 2000 up to 2010. In order to optimize the entire execution, the workflow was implemented to process one year at a time (Figure 6

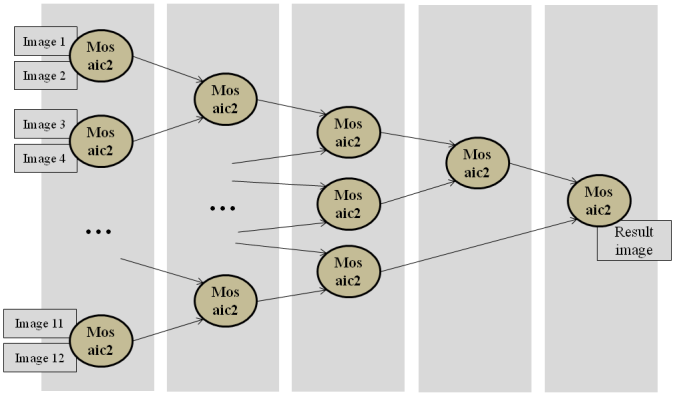


Fig. 5. The internal representation of the Mosaic12 workfow

highlights how the user is able to specify the processing time period).

One of the requirements of this experiment was related to the workflow capability of being able to handle MOD15 and MOD16 products. This is the reason why the graphical interface (Figure 6) allows the user to select bands from both products. By default the Evapo-Transpiration and the Photosynthetically Active Radiation items are already selected.

Until now the user specified only the metadata for the workflow, meaning the processing time and the relevant satellite image bands. But there were no specifications about the actual data. This process is done internally by the GreenLand platform that query at runtime the entire content of the NTSG (ftp://ftp.ntsg.umt.edu/pub/MODIS/Mirror/) and USGS (ftp://e4ftl01.cr.usgs.gov/MOTA) data repositories.

If the content information from the two storages match the metadata specified by the user, then they are automatically downloaded to the machines that process the Mosaic12 workflow and mapped to the corresponding inputs of the graph.

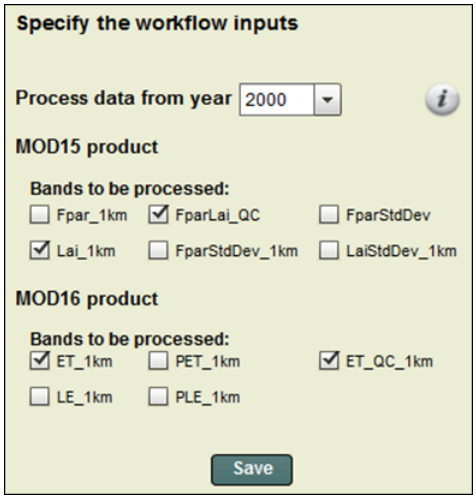


Fig. 6. The inputs specification for the Mosaic12 workflow

D. Optimizing the execution process

When using the workflow for a large time interval (e.g. 10 years) the entire execution process will take a long time to complete. Based on the Grid parallel computing capabilities the GreenLand is able to complete the entire process in approximately two hours.

The MODIS sensor generates data for the same geographic area every 8 days. This means that in a year we have 45 samples for the same tile for each product, and 90 data samples for both MOD15 and MOD16.

This experiment requires that a result is generated for each data sample, meaning that for one processing year the Mosaic12 workflow will generate 90 independent results (if taking into account both MODIS products).

In order to optimize the Grid execution, the gProcess platform partitions the use case into multiple tasks and schedules them to be executed on a different Grid node. Each task contains a group of 9 data samples. Using this approach, we've obtain a parallel execution that significantly improves the total workflow execution time.

E. Results visualization

Another important aspect about the Mosaic workflow is the ability of sharing the results among different scientific communities or between regular users that are not necessarily registered within the GreenLand system.

The implementation of the OCG services proved to be the best solution, regarding the fulfillment of data level interoperability between multiple platforms. The GreenLand offers support for the majority of the OGC products, such as: Web Map Service (WMS) for spatial data visualization, Web Coverage Service (WCS) for remote data retrieval, and Web Processing Service (WPS) useful for standardized execution of the workflows.

The results visualization is managed by the WMS service that provides a standardized method of accessing spatial data, regardless of the location of the remote repository. This service does not offer access to the original information, instead it generates at runtime a graphical representation of the data (under the form of JPEG, TIF, or PNG files). Such a representation is known as layer and can be identified as a frequency band of the satellite images.

The results visualization using the WMS service is an open feature that can be used by any platforms, regardless of its location. The only requirements are the availability of the results (resident on a GeoServer or MapServer) and the Internet connection of all the systems that are implementing the visualization feature.

The WMS service can be accessed directly as a Web based resource (http://<server_domain>/service=WMS&request=GetCapabilities&version=1.1.1) with multiple parameters that specify the results that need to be visualized, the image type that is used for exporting the result (e.g. JPEG, PNG, etc.), the projection type, etc.

Figure 7 highlights the results visualization when using the WMS service from different GIS platforms. As can be

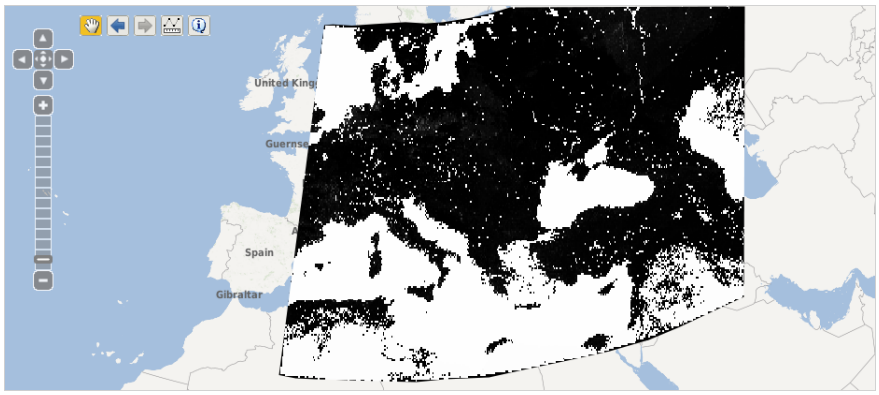


Fig. 7. Mosaic workflow results visualization on different platforms

seen there is the possibility to overlap the WMS result over interactive maps, increasing this way the user satisfaction and the usability of the system.

V. Conclusions

The natural language description of large use case studies is a complex process that requires a good understanding of the context of the problem. Modeling and implementing software architectures based on these use cases is even harder and usually involves specialists from both computer science and Earth science domains.

This paper describes the GreenLand platform that implements the previous mentioned features and exposed them in a user friendly Web based application. The complexity of the inner mechanisms is hidden from the user. Special interaction techniques were developed in order to ease the use cases description in an interactive and intuitive manner.

The system related architecture highlights all the modules of the GreenLand platform and exemplifies their contributions by modeling the Black Sea catchment scenario as a GreenLand workflow.

OGC standard implementation provides the advantage of achieving data interoperability with other external platforms. This feature is useful especially when retrieving, processing, and visualizing spatial data from different remote repositories.

Acknowledgment

This research is supported by the enviroGRIDS Project funded by the European Commission, through the Contract 226740.

References

[1] D. Mihon, V. Colceriu, F. Bektas, K. Allenbach, M. Gvilava, D. Gorgan, "Spatial Data Exploring by Satellite Image Distributed Processing", Geospatial Research Abstracts, EGU General Assembly, vol.14, 2012.

[2] enviroGRIDS project, http://envirogrids.net/

[3] B.F. Bektas, C. Goksel, S. Sozen, K. Allenbach, M. Gvilava, K. Rahman, D. Gorgan, D. Mihon, "Remote Sensing Services - ESIP Platform and Hot Spot Inventory Case Studies", enviroGRIDS Deliverable D2.11, 2012. Available at: http://envirogrids.net/index.php?option=com_jdownloads&Itemid=13 &view=finish&cid=139&catid=11

[4] D. Mihon, A. Minculescu, V. Colceriu, D. Gorgan, "Diagramatic description of distributed spatial data processing", Romanian Journal of Human-Computer Interaction, pp.59-80, 2012.

[5] K. Czajkowski, S. Fitzgerald, I. Foster, and C. Kesselman, "Grid Information Services for Distributed Resource Sharing", In IEEE International Symposium on High Performance Distributed Computing, (2001), IEEE Press

[6] V. Bacu, T. Stefanut, D. Rodila, D. Gorgan, "Process Description Graph Composition by gProcess Platform", 3rd International Workshop on High Performance Grid Middleware, CSCS-17 Conference, vol.2, pp.423-430, 2009.

[7] Open Geospatial Consortium, (2007), "OpenGIS Web Service Common Implementation Specification", pp.1-153.

[8] Sextante, "A Versatile Open-Source Library for Spatial Data Analysis", 2011.

[9] P. Ramsey, "User Friendly Desktop Internet GIS (uDIG) for OpenGIS Spatial Data Infrastructures", Refractions Research Inc., 2004.

[10] M. Landa, "New GUI for GRASS GIS Based on wxPython", Departament of Geodesy and Cartography, pp.1-17, 2008.

[11] T. Athan, O. Dassau, A. Ghisla, "Quantum GIS 1.7.0 Geographic Information System User Guid", Open Source Geospatial Foundation, 2011.

[12] A. Maier, "A Job Management and Optimising Tool", International Conference on Computing in High Energy and Nuclear Physics, Journal of Physics: Conference Series, vol. 119, pp.2-9, 2008.

[13] J.T. Moscicki, "DIANE Distributed Analysis Environment for Grid-enabled Simulation and Analysis of Physics Data", Nuclear Science Symposium, vol. 3, pp.1617-1620, 2004.

[14] V. Bacu, "Error Prevention and Recovery Mechanisms in the ESIP Platform", IEEE 6th International Conference on Intelligent Computer Communication and Processing, pp.411-417, 2010.

[15] D. Gorgan, V. Bacu, T. Stefanut, D. Rodila, D. Mihon, "Earth Observation Application Development based on the Grid Oriented ESIP Satellite Image Processing Platform", Journal of Computer Standards & Interfaces, vol. 34/6, pp. 541548, 2012.

Mathematical Modeling of Distributed Image Processing Algorithms

Vlad Colceriu, Danut Mihon,
Angela Minculescu
{vlad.colceriu, vasile.mihon}@cs.utcluj.ro
angela.minculescu@gmail.com
Technical University of Cluj-Napoca
Cluj-Napoca, Romania

Victor Bacu, Denisa Rodila
Dorian Gorgan
{victor.bacu,denisa.rodila}@cs.utcluj.ro
dorian.gorgan@cs.utcluj.ro
Technical University of Cluj-Napoca
Cluj-Napoca, Romania

***Abstract*—Satellite images play an important role in developing Geographical Information System software applications that prove to be useful for different Earth Science phenomena analysis. Accurate results are obtained from high resolution images, or by applying the same algorithm multiple times over a specific input data set. In both cases the data volume that needs to be processed is large, and usually involves distributed infrastructures. In order for non-technical users to use these algorithms, they should be described in a flexible manner, using workflow structure models. This paper highlights the main achievements within the GreenLand platform, regarding scheduling, executing, and monitoring the Grid processes. Its development is based on simple, but powerful, notion of mathematical directed acyclic graphs that are used in parallel and distributed executions over the Grid infrastructure.**

I. Introduction

This paper highlights the parallel and distributed satellite image processing over the Grid infrastructure, as implemented within the GreenLand platform. GreenLand is a free GIS (Geographical Information System) software used in the geospatial data management and visualization domain, which was integrated as part of the BSC-OS(Black Sea Catchment-Observation System) portal[1][2] alongside other software platforms, designed for calibration of SWAT models, such as gSwat[3] and BASHYT[4] and other general purpose GIS web applications, such as GeoServer[5] and GEOSS[6]. The following sections present some of the main goals of this system: provide a flexible description of spatial data processing, schedule, execute and monitor Grid processes, GRASS (Geographic Resources Analysis Support System) [7] library integration, and interoperability with other software platforms.

All the executable processes implement a specific functionality, related to the Earth Science domains: satellite images data extraction, thematic map creation, arithmetic operations on spatial data, raster and vector data conversion, etc. All these processes are represented within the GreenLand platform as acyclic graphs, composed from basic operators, Web services and sub-graphs [15].

The operators are identified as atomic components and represent the smallest unit of work that can be executed without further decomposition. The workflow is another GreenLand concept, used to fulfill the user needs. It could be defined as a collection of basic operators, adopting a graph-style representation. Each node implements a particular function, while the entire workflow can be used to simulate specific dataflow scenarios.

The availability of the GreenLand system for non-technical persons was the main reason for workflow based data representation. Otherwise they should have been familiar with the XML standard and with developing Linux based scripts. In order to ease the user actions, two editor tools were implemented for operator and workflow description. Another advantage of using this approach is the portability within other platforms, as described in section System related architecture.

The Grid infrastructure processing capabilities are needed due to the large volume of satellite data that could reach a few GB is size. Executing such data is a complex process and should be optimized even when executed over the Grid worker nodes. Some workflows executions are light weight, while other might take hours to complete. This way it is up to the gProcess platform [8] to apply the best scheduling techniques. Currently no solutions exist to overcome this shortcoming, but several research directions have already analyzed and put into practice[9].

The gProcess platform is used for Grid process schedule, execution and monitoring. More information about the operations performed by this platform can be found in section entitled Grid based execution.

II. Related works

The Grid processes are described using the mathematical graph concept that seems to fulfill the GreenLand requirements of extensibility and simplicity . The major disadvantage in using such a method is represented by the cyclic workflows that handle looping execution. This is a restrictive case in the GreenLand workflows editor, and the user has no possibility to define such kinds of structures. There are several applications that could be used to create workflows: Pegasus [10], Taverna [11], GridFlow [12], etc. All of these are working only with acyclic graphs, called DAG (Direct Acyclic Graph). The main difference between these tools and the OperatorEditor and WorkflowEditor, developed within the GreenLand platform, is related to the flexibility in managing the data structure, the possibility of creating hyper-graphs, depth workflow navigation, or ease in creating new basic operators by attaching a specific functionality (described throughout an executable file, script file, Web service, etc.).

Most of the GreenLand operators encapsulate GRASS functionalities that operate with raster or vector data formats. The GRASS library allows the usage of more than 300 operators, supports over 2500 different CRS (Coordinate Reference System) and handles the most common used spatial data types: Landsat, MODIS, GeoTIFF, ESRI shapefiles, etc. Due to its popularity, there are several geospatial applications that integrate this library: Sextante [13] and QGIS (Quantum GIS) [14].

The main goal of Sextante is to provide an easy method for implementing rich geo-processing algorithms, and it integrates tools like Java GIS, OpenJUMP, ArcGIS, etc. QGIS allows the user the possibility to execute geospatial data, to analyze the results, edit raster and vector data, data type conversion, etc.

One of the main goals of the GreenLand platform is to provide workflows that could be reused in other applications, such as Pegasus, Taverna, PGRADE [15], etc. This could be achieved by using the SHIWA (SHaring Interoperable Workflows for large-scale scientific simulations on Available DCIs) [16] platform that offers interoperability services in order to standardize the workflow development and portability.

Workflow interoperability enables their execution over different infrastructures, allows data sharing among scientific communities around the world, facilitates workflows migration between applications, and offers the usage of the most appropriate system or infrastructure in order to execute one specific workflow.

In order to access GRASS functions, the user has to write its own Linux bash script, in the Sextante and QGIS frameworks. On the other hand the GreenLand offers the user the possibility to do the same operations but in a more intuitive manner, by using the workflow editor. This approach allows the non-technical users to develop and process their own scenarios, without the uncertainty of introducing semantic or syntactic errors.

The GreenLand uses the gProcess platform in order to schedule, execute and monitor processes over the Grid infrastructure. Other approaches that share the same experience regard the GANGA [17] and Diane (Distributed Analysis Environment) [18] tools. Grid process configuration and monitoring is based on the GANGA tool, while the execution scheduling and task submission is related to the Diane application

III. System related architecture

GreenLand is a client-server application, available over the Web. The client-side represents the graphical user interface that fulfills user requests for a extensible, parallel running and internet accessible GIS platform. The server-side is Java based and implements functionalities for users, projects and data management. Data exchange between these two modules is based on Web services.

The only way for the user to access the backend functionality of the GreenLand application is through its graphical interface (Figure 1). A username and password authentication is required for system access.

The second architectural level consists of a set of services exposed by the GreenLand platform: users management, workflows development, execution and management, data retrieval,

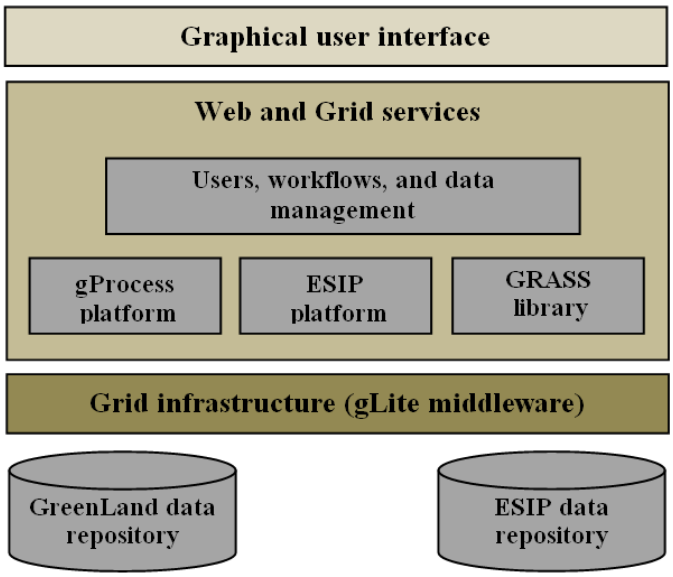


Figure 1. System related architecture

data storage, data conversion, etc. These services are available by integrating the gProcess and ESIP (Environment oriented Satellite Data Processing Platform) [19] platforms. The Web services provided by the gProcess fulfil the user requirements regarding the process scheduling, execution and monitoring.

The workflows developed by the users have two internal standard representations, both of them using the XML description. The first one is called PDG (Process Description Graph) and it is a pattern that describes only the workflow nodes types and position, and the relationship between them, but it has no knowledge about its physical inputs and outputs. This pattern is only used to store the workflow representation, and it expands during the Grid execution into a so called iPDG (instantiated PDG). This second representation shares the same XML structure as the PDG, and allows the gProcess to gather all the inputs information specified by the user (e.g. spatial data files, numerical constants, external dependencies, etc).

Based on the iPDG format, the gProcess platform performs the Grid scheduling operation. In most cases a single node in the workflow will be processed on a single CPU, but there are situations in which groups must be created in order to improve the execution efficiency. Currently this is not an automated process, because it requires a complexity analysis of the entire workflow. Several research studies were conducted in this direction, and the bases for such a module were already adopted.

The gProcess platform establishes the connection with Grid infrastructure, by implementing a subset of the gLite middleware. These services allows the data transfer (i.e. input data specified by the user) to SE (Storing Element), tasks execution over CE (Computing Element), proxy creation and delegation, Grid execution information retrieval, etc.

The ESIP platform are a set of Web services that provide the following functionalities: basic operators and workflows development, workflow representation based on DAG (Direct Acyclic Graph) patterns, spatial data management, etc. Internal

representations of the basic operators are also part of the ESIP platform, exposed as: vegetation indices (e.g. NDVI, EVI), spatial data processes (e.g. mosaic, density slicing, and data extraction), statistics (e.g. histogram generation, standard deviation computation), etc.

Other services provided by the GreenLand platform are related to users management (i.e. create new account, update profile, etc.), data retrieval using the local upload mechanism, FTP data transfer and OGC services [20].

Finally it is worth mentioning that the current application stack is enrolled within the envirogrids.vo-eu.egee.org Virtual Organization, of which for testing purposes we used the sites or computing elements: RO-09-UTCN and AM-02-SEUA.

IV. Data model

This section describes the basic operator, workflow and project concepts, their development using the GreenLand editor tools as well as their internal representation within the ESIP data repository (Figure 1).

A. Project and workflow relationship

GreenLand projects are defined as virtual containers that allow workflow organization and instantiation. Each project has a unique name in the user workspace, and supports workflows attachment. A workflow can be added as multiple instances within the same project. At graphical user interface level, the project content is displayed as a forest of trees, where each tree root represents the workflow name, and leafs consists of the workflow instances. Each item inside the project, stores information about its name, description, author who developed it, inputs and outputs, etc.

From the graphical interface the user is able to specify the physical inputs for this item (workflow). For each input, only the available values are displayed to the user (e.g. if the inputs type requires a spatial data attribute, only the list of available satellite images are shown). All these information are retrieved based on ESIP services.

Executing a project consists of processing its entire list of workflows. This operation is achieved by using the gProcess services. After the Grid process begins, a monitoring mechanism gives feedback about execution progress.

B. Basic Operator Concept

Operators lie at the center of the gProcess execution environment and GreenLand management system. They represent the basic units of work, the only constructs which can get executed.

The GreenLand application allows users to create, alter and delete these structures. By doing so, it allows full customization of the Grid execution processes, from its most coarse grained constructs represented by iPDGs to its most simple, atomically executed statements.

Operators represent the most fine grained execution units; they are the only constructs that get executed on the nodes of the Grid. These units must have their respective program or executable script defined as well as any dependencies they might require, since environment in which they run is heterogeneous and offers no guarantees on shared library or version.

Table I. Operator editing conditions

	Operator is owned		Operator is used		Operator is validated	
	True	False	True	False	True	False
Insert	N/A	N/A	N/A	N/A	N/A	N/A
Update	Yes	No	Partial	Yes	Yes	Partial
Delete	Yes	No	Partial	Yes	Yes	Yes

The insertion of operators is supported via a visual editor which takes the users program and annotations and inserts it in the gProcess and GreenLand databases.

When creating an operator one has to provide besides the executable code of the program, certain additional information, which allows the GreenLand application to track the visibility, unique name, description and category of the operator.

There are two types of visibility properties defined:

- Public means that all the users may view and use the operator.
- Private means that only the owner of the operator may view or use it.

The public operators, to which a user is not owner to, but uses within its private or public workflows, can still be accessed even if the visibility of the program in question is changed. However creation of new graphs containing that element is prohibited.

The category allows the user to create its own hierarchy of operators, facilitating a quicker lookup when browsing for them.

An Application Programming Interface (API) has been created to allow the user to create operators. The problem with it is that if reuse is desired, the implementer would have to create a new program form scratch or call its desired application from within the provided ESIP (Environmental Satellite Image Processing) API.

Entering, updating and deleting Operators is not a straight forward operation, since there are some constraints involved in it as expressed in (Table I).

The first of these limitations refers to ownership of the operator, since there is a strict traceability of Grid execution which needs to be maintained. The idea is that each user should be responsible for its own distributed application. Additionally, before such an operator is made visible, it is tested locally for compliance, so that any malicious or unintended effects of the program may be detected.

The second limitation refers to whether the operator to be removed or updated is already in use. If it is used, removal and updating is done only at a formal level; else it is removed entirely from the database of operators. Deleting or altering an operator at a formal level means that any of the existing workflows which use it, can do so without becoming invalid or having their functionality changed.

Finally before moving on to workflows and hypergraphs the programming interface is discussed. It is implemented in using only the Java programming language, which brings up certain constraints regarding the generality of the platform. Of course one can call a program or script implemented in any programming or scripting language from the required Java wrapper, as long as it is supported by the operating system on the worker nodes of the Grid. Where the current worker node distributions include a CentOS version of the Linux kernel.

Also to be noted is the fact that all operators must be implemented in such a way so as to be able to parse Linux type paths, end of line characters and call executables which were compiled in Linux, preferably having all their library dependencies packaged alongside themselves.

Further constraints on the program include aspects of code structure such as [21]:

- Including the Operator class in a certain package "gPOperators"
- Extending a certain class, which includes the code for launching the operator on the Grid node "OperatorExec"
- Overriding a certain method included in the "OperatorExec" class

All these limitations exist due to the fact that these operators need to be integrated inside the gProcess platform, which was not designed to support such rich and powerful interaction as exposed by the GreenLand application.

This programming interface also includes all the dependencies and prerequisites needed for generating GRASS and GDAL based programs as described in section V-B. In order to do this a different class needs to be extended "GenericOperator" and a different method overridden "grassExecute".

C. Workflow and Hypergraphs Concept

gProcess and GreenLand give users the opportunity to develop their own parallel and distributed programs. These are implemented with the help of Process Description Graphs (PDG), which plainly put are directed acyclic graphs.

Describing programs with the help of graphs is not a new concept; it has been extensively studied within [16] which presents a general solution to integrate already existing platforms together. It is also present in other well established frameworks for Grid execution such as [22] and [23].

PDG's cannot be executed on the gProcess platform since they represent only the program definition; they lack the input data necessary to perform useful actions. For execution we use another construct called Instantiated Process Description Graphs (iPDG).

iPDG's are morphologically similar with their counterparts but they give the possibility to specify user input to the defined program.

Both PDG's and iPDG's may also be referred to as workflows, since they present the flow of data, from node to node, in a Grid program.

```
<?xml version="1.0" encoding="UTF-8" ?>
<Workflow>
<Nodes>
    <Resource …></Resource>
    <Operator id="6" name="NDVI" idDB="64">
        <Preconditions>
            <Input id="1"/>
            <Input id="2"/>
            <Input id="3"/>
            <Input id="4"/>
            <Input id="5"/>
        </Preconditions>
    </Operator>
</Nodes>
<Groups></Groups>
</Workflow>
```

Figure 2. Simple PDG representing an NDVI program

The internal structure of a PDG is represented by nodes and directed edges. The nodes can be matched to operators or other PDG's. These particular types of entities, which do not make the scope of the top level structure are called sub-workflows and are similar to the idea of functions in programming languages. A structure which has multiple levels of imbrication is called a hypergraph.

Recursive structures are not supported within workflows since there is no control structures currently implemented within workflows. The reason they are not supported is due to the fact that no control structures have been implemented.

Control statements would allow the distributed program to test for termination conditions, otherwise not encountered in the current solution.

The arcs described inside a PDG and iPDG represent the flow of data. All information passed from a source node to destination passes trough gProcess file system, where it is forwarded to the corresponding execution, as specified in the workflow.

The constraints and operations presented for nodes also apply here. The major difference is that workflows are automatically created once such a request is submitted and require no additional validation of their behavior. One may assume that their behavior is implicitly safe since all their individual parts function correctly. We can make this assumption because it is only the operators that get directly executed.

gProcess and GreenLand have different representations of these two notions. gProcess uses a lightweight XML representation (Figure 2) of the directed acyclic graph.

The XML format is disadvantageous in allowing for an editable and extensible program structure mainly because of the fact that the user must specify the inputs and be able to validate the program structure manually. This means that it would need intimate knowledge about application structure. Such a solution would be impractical and furthermore unsafe since it would give the user direct access to resources, without any possibility to restrict or refute its actions.

On the other hand GreenLand allows for a database representation of the model, which gives the user the possibility

to dynamically create and modify workflows, without having to know anything about internal representation. The model described was created so as to serve to the purpose of categorizing, extending and validating the workflows and their subcomponents.

The basic concepts behind the GreenLand application is the gProcess workflow, which is represented by a directed acyclic graph also called a PDG. In this graph each node represents the executable code submitted on a worker node, an operator. On the other hand an arc represents a communication path between two operators. They are not explicitly modelled since they can be inferred from the connection between two node(Figures 2 and 3).

The GreenLand data model supports ranking of operators according to categories in order make searching for a given functionality easier. Atop of this each category element offers the possibility to generate other subcategories(Figure 3), thus generating a infinitely extendible structure.

Each node of a workflow can be either a operator or another workflow, generating a multi-layered structure, inside which no cycles or self-calling elements can exist.

Additionally resources in the form of inputs and outputs are attached to a node. The amount of inputs or outputs a node may contain is unlimited, except for the case of operators, which may contain at most a single output. This constraint is imposed by gProcess functionality, which requires this in order to be able to detect operator output and communicate results between the nodes of the program graph.

Each resource supports either a string value or a file type. In order to assure that these elements are matched correctly, two types of validations need to be performed.

First a syntactic validation assuring that the file is of the required type. This validation is not done by filtering the file through a extension sieve, but by pre-emptively inspecting the file type at import time.

The second type of validation is done at the semantic level, where each file is checked so that the meta-data attached does not have conflicting values. An example of this would be the projection of the files, which according to GRASS and GDAL operators would have to be the same in order to obtain a successful execution.

Additionally it is worth mentioning that Greenland is accompanied by an interface application, which allows the user to interactively manipulate workflows, as easily as one would create, update and delete an operator [24].

V. Grid based execution

This Section presents the gProcess and GreenLand in intimate detail, highlighting their interfaces and communication protocols, which help the user to submit, create and manage distributed Grid programs.

A. GreenLand and gProcess Compatibility

GreenLand and gProcess are a pair of symbiotic applications designed to complement each other and in some cases of degraded functionality even work independently. The current implementation however requires that both applications be housed by the same machine.

GreenLand is a workflow, operator and file manager which allows the user to generate, edit and categorize Grid programs. On the other hand gProcess is a Grid execution manager , which allows the submission and cancelling of complex execution workflows.The task scheduler implemented in gProcess was also studied in [25].

Although they were thought with the idea of separability in mind, they still have to communicate with each other, to pass programs created in GreenLand to gProcess and to synchronize GreenLand data to gProcess executions.

As mentioned in Sect. IV-C, these two applications have different representations of PDG's and iPDG's. Where GreenLand has a recursive database hierarchy of operators and workflows, which contains additional information such as categories, descriptions and ownership information. Also the arcs and nodes of the graphs are represented as separate entities within the storage space. On the other hand gProcess has a lighter representation, where the entire program is contained within an executable file.

In order for things to work GreenLand must know the internal implementation of gProcess programs. This means that the GreenLand application must be able to create gProcess execution files. To do this it interrogates the gProcess database for all available operators and input types, which it uses to generate and validate its own programs.

gProcess offers services for uploading operators, workflows and required input files. These services are then called by GreenLand, so that the data edited within can become available to the Grid execution environment.

Execution and monitoring of workflows is the most important part of the GreenLand/gProcess communication and is divided in 4 distinct steps.

The first operation is the transfer of the iPDG file from the GreenLand application to gProcess. Even though both applications are housed by the same machine, they were designed to operate remotely. This is done by calling the "importXML" service of the gProcess application.

The second step requires that the file be registered as a PDG by calling "insertPDG" and then as a iPDG by calling "insertIPDG". This step is done on the same file, due to the similarities between the two file types.

After uploading the program, it is executed by calling the "execute" service, which returns information about monitoring identification number. This is then later used to single out the workflow, from within the set of monitored executions.

Monitoring is done at 2 different levels:

- Top level, which polls the execution in order to discover the state of the workflow
- Operator level, which inquires about the state of each node execution separately and extracts the output

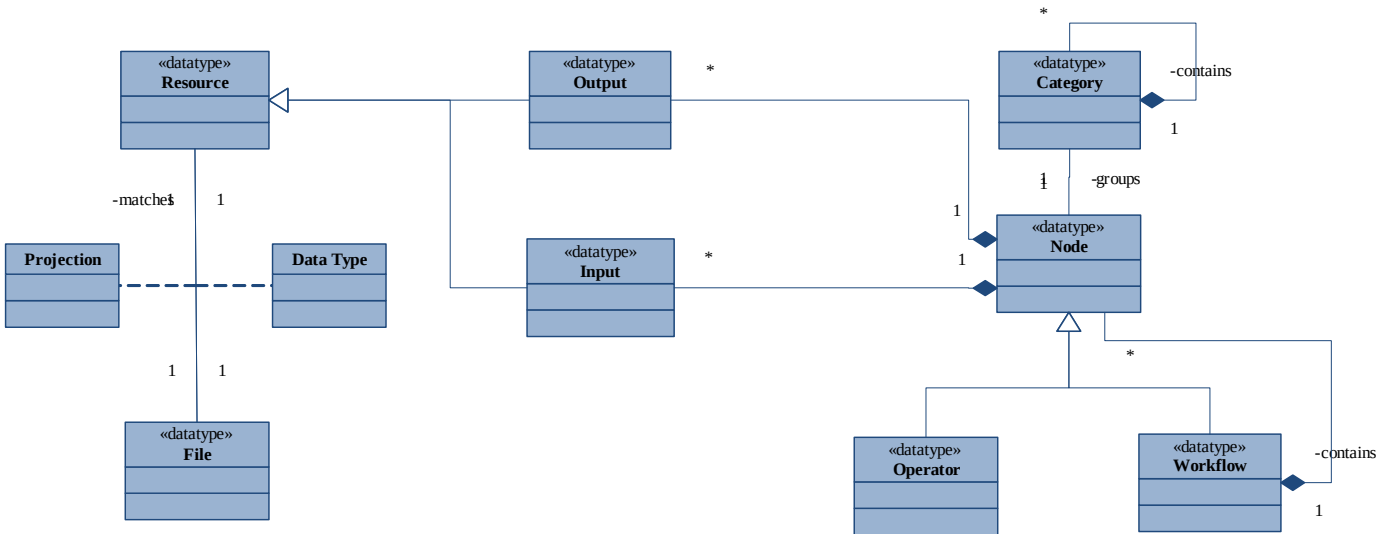


Figure 3. GreenLand Data Model

B. GRASS Integration

The operators developed for gProcess can be configured to run already existing applications. GRASS is one such case of a fully fledged desktop application, running on the Grid.

All programs which run on the Grid must not require any interactive user input. They must be applications which have non-interactive interfaces, meaning that all input must be known in advance.

The Grid platform exposes to its users a heterogeneous environment, where program versions and installed shared libraries can differ from system to system. The only constant we can count on is that the background operating system is running a Linux kernel. In such a case we cannot make any assumptions about whether an application which runs perfectly on a desktop environment will run in the same manner on all of the nodes.This means that in order to use GRASS there are certain steps which have to be performed before one can be sure of its functionality.

The primary condition that must be satisfied is that all executable and configuration files used by the operator be packaged with it as described in (Figure 4).

GRASS has a binary folder which contains all functions, which must be included in the operator dependencies. Also a configuration file specifying some of the parameters of the application, ex. *DATABASE*, *LOCATION_NAME* and *MAPSET*. More on this topic can be found in [7].

On a desktop solution the operating system will satisfy all needed shared libraries at install time. On the Grid platform an executing operator has limited privileges when writing files, accessing system state and installing programs. To compensate for this drawback all needed shared libraries were packaged with the operator.

Finally a script must be created, which generates the above mentioned configuration file and appends all executables to the *$PATH* system variable and prepends the shared libraries to the *$LD_LIBRARY_PATH* variable. The class that implements this functionality within the GreenLand programming interface is "GrassGeneric".

C. Grid Execution and Monitoring

Each program created by GreenLand is later executed, monitored and managed by gProcess. Once the former mentioned application is done creating and launching the workflow, the second jumps into action.

The executor service processes the iPDG description in order to accomplish workflow execution on the Grid [26], where it parses the XML file and generates the appropriate internal representation. It then tries to check the file for consistency by matching input and output types. The input data of one operator, service or resource must match the output of the node on the other end of the arc which links them.

The executor service also checks for consistency relating to the availability of the individual operators instantiated within the internal representation. If any of the operators are missing or unavailable the system tries to find an operator or service capable of substituting it, while also checking for cycles and recursive declarations. Doing so, it creates a planar structure, which is the expanded structure of the program.

When an internal representation has been created the backend application then submits each individual node of the workflow to a CE (Computing Element) of the Grid.

Once a workflow has been launched into execution, the hierarchies which existed within it are no longer visible. The user can only see the flattened, instantiated graph. This means that from the moment the workflow was launched, the monitoring can follow only the state of the entire structure and of individual operators, but not of intermediate structures.

Also canceling an entire workflow is supported, but not a singular node, since operators downstream might suffer from unsatisfied input constraints. This would require the system to cancel all dependent nodes, but since this would lead to results which would be hard to predict without having advanced knowledge of internal structure.

VI. Practical Use Case Scenarios

This section is divided into 2 subsections each detailing a different type of program generation mechanisms for gProcess corresponding to different levels of program abstraction.

The first use case will detail an operator, which was designed for merging a series of satellite images from a FTP repository into a single large image. Thus giving the user the possibility to select year, month and day of the given image and the region which required combining, without any prior knowledge of how the data had been organized on that particular repository, in order to obtain a single image of the entire Black Sea catchment area.

The reason for generating a new operator instead of a workflow was chosen due to the very particular functionalities of this use case, which could not be satisfied by other more general operators.

The second use case will detail a complex workflow generated, from a series of predefined operators. Where the requirement to be satisfied was the generation of a thematic map highlighting land use in the Istanbul metropolitan area.

More information can be about the particularities of both these use cases can be found in document [27]

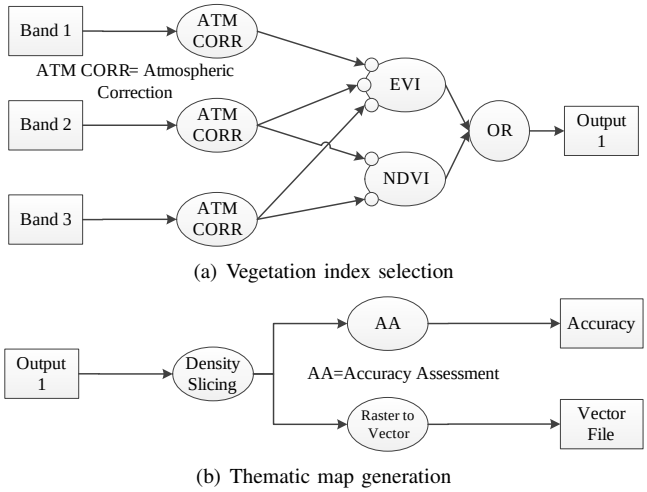
A. Mosaic Operator Use Case

This section presents the usage of a complex atomic structure within this framework. It gives an idea of how powerful and general the interface for Grid program generation really is.

The atomic operator is divided into several steps. The idea of atomicity is implemented under the paradigm of all or nothing execution. Meaning that if the operator fails, at one of the steps, no partial result will be available to the workflow.

Inside the workflow there exist a list of operators allowing the user to generate a sequence of images representing a given time interval.

Figure 4. GRASS Operator Setup Script



(a) Vegetation index selection

(b) Thematic map generation

Figure 5. Workflows representing the Istanbul Thematic Map Use Case

The "Special Mosaic" operator takes multiple multiband images of various formats and glues them together according to certain metadata embedded within their corpus, which may refer to the projection of the individual bands, as well as the geographic region which they occupy. Such information provide the operator a way to combine the images.

The operator receives as its arguments the following: a link to an ftp server, a directory of that server plus username and password if necessary. The operator then decides which files to download given a specified algorithm.

The steps of the operator are divided as follows:

1) Download the images via ftp.
2) Split the images in their respective bands.
3) Combine each band from its parts.
4) Merge all results into a single image.

B. Istanbul Thematic Map Generation Use Case

In order to generate a thematic map for the Istanbul area from a given set of Landsat satellite images a series of operations needed to be performed.

Since the thematic maps are of land use in urban areas, the main operators of the workflows are those exposing vegetation indices, of which the current implementations opted for EVI and NDVI. Therefore the bands of the Landsat image being used are 1,3 and 4 corresponding to blue, red and infrared bands. Bands 3 and 4 are required for NDVI and 1,3 and 4 for EVI. Both algorithms return an image with values between -1 and 1, where values from -1 to 0 represent water bodies and 0 to 1 increasing values of vegetation.

Before the vegetation index operations can be performed, there is the need for atmospheric correction, which is based on metadata attached to the multi-band image and a series of mosaic and cropping operations, which are required due to the fact that the location of Istanbul is spread across 2 distinct Landsat images. Cropping and mosaicking are removed from figure 5 due to them not bringing any added value to the use case outside of solving a technical issue.

After applying one of the 2 vegetation indexes a density slicing algorithm is applied reducing the number of possible values of the resulting image from 256 floating point intervals to just 3 classes representing water, urban and wooded areas.

The last step of this algorithm is composed of an accuracy assessment operator and a Raster to Vector image converter, which guarantee that a sufficiently accurate thematic map represented by vector file is generated. If the accuracy is below a given threshold the workflow is executed again using different intervals for the 3 classes of the density slicing operator.

It is because of this fact that the implementation of this logical workflow has been divided into 2 parts so as to remove redundant work regarding atmospheric correction, mosaicking, cropping and vegetation index calculation (Figure 5).

VII. Conclusion

Due to the high complexity and size of input data satellite image processing requires high computing power. In order to be able to meet these requirements gProcess uses the Grid execution platform.

GreenLand extends the functionalities of gProcess by giving the user an interface with which he can customize his own programs from the coarse grained constructs represented by top level workflows to the most fine grained represented by operators.

Additionally to submission and management gProcess offers optimized execution and scheduling of multiple workflows so as to obtain the highest possible throughput.

Acknowledgment

This research is supported by the enviroGRIDS Project funded by the European Commission, through the Contract 226740.

References

[1] D. Gorgan, V. Bacu, D. Mihon, T. Stefanut, D. Rodila, P. Cau, K. Abbaspour, G. Giuliani, N. Ray, and A. Lehmann, "Software platform interoperability throughout envirogrids portal," *International Journal of Selected Topics in Applied Earth Observations and Remote Sensing –" JSTARS*, vol. 5, no. 6, pp. 1617–1627, 2012.

[2] D. Gorgan, V. Bacu, D. Mihon, D. Rodila, T. Stefanut, A. K., P. Cau, G. Giuliani, N. Ray, and A. Lehmann, "Spatial data processing tools and applications for black sea catchment region," *International Journal of Computing*, vol. 11, no. 4, pp. 327–335, 2012.

[3] D. Gorgan, V. Bacu, D. Mihon, D. Rodila, K. Abbaspour, and E. Rouholahnejad, "Grid based calibration of swat hydrological models," *Journal of Nat. Hazards Earth Syst. Sci.*, vol. 12, no. 7, pp. 2411–2423, 2012.

[4] P. Cau, C. Meloni, S. Manca, D. Soru, and D. Muroni, "A java based framework optimized for scientific modeling and analysis," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1, 2011.

[5] J. Deoliveira, "Geoserver: uniting the geoweb and spatial data infrastructures," in *Proceedings of the 10th International Conference for Spatial Data Infrastructure, St. Augustine, Trinidad*, 2008.

[6] M. L. Butterfield, J. S. Pearlman, and S. C. Vickroy, "A system-of-systems engineering GEOSS: Architectural approach," *Systems Journal, IEEE*, vol. 2, no. 3, pp. 321–332, 2008.

[7] M. Neteler, M.H. Bowman, M. Landa, and M. Metz, *GRASS GIS: a multi-purpose Open Source GIS*, Environmental Modelling and Software, vol.31, pp.124-130, 2012.

[8] V. Bacu, T. Stefanut, D. Rodila, D. Mihon, and D. Gorgan, *Process Description Graph Composition by gProcess Platform*, HiPerGRID, May 28, Bucharest, vol.2, pp.423-430, 2009.

[9] V. Colceriu and D. Gorgan, "Execution time estimating framework on distributed platforms,", 2013, Unpublished.

[10] J.S. Vockler, G. Juve, E. Deelman, M. Rynge, and G.B. Berriman, *Experiences Using Cloud Computing for a Scientific Workflow Application*, ScienceCloud'11, pp.15-24, 2011.

[11] W. Tan, P. Missier, I. Foster, R. Madduri, D. De Roure, and C. Goble, *A comparison of using Taverna and BPEL in building scientific workflows: the case of caGrid*, Concurrency and Computation: Practice and Experience, vol. 22, pp.1098-1117, 2010.

[12] J. Cao, S.A. Jarvis, S. Saini, and G.R. Nudd, *GridFlow: Workflow Management for Grid Computing*, In 3rd International Symposium on Cluster Computing and the Grid (CCGrid), IEEE CS Press, May 12-15, Tokyo, Japan, pp.198-205, 2003.

[13] V. Olaya, *Sextante User's Manual*, 2011.

[14] T. Sutton, O. Dassau, and M. Sutton, *Geographical Information System User Guide*, Open Source Geospatial Foundation Project, 2011.

[15] P. Kacsuk, *P-GRADE Portal Family for Grid Infrastructures*, Concurrency and Computation: Practice and Experience, vol.23, pp.235-245, 2011.

[16] N. Cerezo, and J. Montagnat, *Scientifc Workflows Reuse through Conceptual Workflows on the Virtual Imaging Platform*, Proceedings of 6th WORKS2011, Seattle, pp.1-10, 2011.

[17] J.T. Moscicki, F. Brochu, J. Ebke, U. Egede, J. Elmsheuser, K. Harrison, R.W.L. Jones, H.C. Lee, D. Liko, A. Maier, A. Muraru, G.N. Patrick, K. Pajchel, W. Reece, B.H. Samset, M.W. Slater, A. Soroko, C.L. Tan, D.C. van der Ster, and M. Williams, *Ganga: A tool for Computational-task Management and Easy Access to Grid Resources*, Computer Physics Communications, vol. 180, pp.2303-2316, 2009.

[18] J.T. Moscicki, *DIANE - Distributed Analysis Environment for GRID-enabled Simulation and Analysis of Physics Data*, Nuclear Science Symposium, vol. 3, pp.1617-1620, 2004.

[19] V. Bacu, D. Rodila, D. Mihon, T. Stefanut, and D. Gorgan, *Error prevention and recovery mechanisms in the ESIP platform*, IEEE 6th International Conference on Intelligent Computer Communication and Processing, ICCP2010, pp.411-417, 2010.

[20] A. Padberg, and K. Greve, *Gridification of the OGC Web Processing Service: Challenges and Potential*, AGILE Workshop, pp.5-11, 2009.

[21] V. Colceriu and D. Mihon, *Operator Editor*, 2012. [Online]. Available: http://cgis.utcluj.ro/documents/OperatorEditor_user_manual.pdf

[22] P. Kacsuk, T. Fahringer, Z. Nemeth. *Distributed and Parallel Systems. Cluster and Grid Computing* , 2nd edition, 223 pages, Springer Verlag, ISBN: 0387698574 (2007)

[23] E. Deelman, G. Singh, M.H. Su, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, G. B. Berriman, J. Good, A. Laity, J. C. Jacob, D. S. Katz, *Pegasus: a Framework for Mapping Complex Scientific Workflows onto Distributed Systems*, Scientific Programming Journal, Vol 13(3), 2005, Pages 219-237

[24] D. Mihon, A. Minculescu, V. Colceriu, and D. Gorgan, "Diagramatic description of distributed spatial data processing," *Romanian Journal of Human - Computer Interaction*,pp. 129-134, 2013.

[25] Pop F., *A Fault Tolerant Decentralized Scheduling in Large Scale Distributed Systems, chapter in Handbook of Research on P2P and Grid Systems for Service-Oriented Computing: Models, Methodologies, and Applications*, N. Antonoupoulos, G. Exarchakos, M. Li, A. Liotta (Eds.), Ed. Information Science Reference (IGI Global), ISBN: 978-161-520-686-5, pp. 566-588, February 2010

[26] Gorgan D., Bacu V., Stefanut T., Rodila D., Mihon D., *Grid based Satellite Image Processing Platform for Earth Observation Applications Development.* IDAACS'2009 - IEEE Fifth International Workshop on "Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications", 21-23 September, Cosenza, Italy, IEEE, Computer Press, ISBN: 978-1-4244-4901-9, 247-252 (2009).

[27] F. B. Balcik, C. Goksel, K. Allenbach, M. Gvilava, K. Rahman, D. Gorgan, and V. Mihon, *Building Capacity for a Black Sea Catchment Observation and Assessment supporting Sustainable Development*, 2012.

Remotely Sensed Data Processing on Grids by using GreenLand Web Based Platform

Filiz Bektas Balcik[1], Danut Mihon[2], Vlad Colceriu[2], Karin Allenbach[3,4], Cigdem Goksel[1], A. Ozgur Dogru[1], Gregory Giuliani[3,4], Dorian Gorgan[2]

[1]Istanbul Technical University, Istanbul, Turkey
[2]Technical University of Cluj-Napoca, Cluj-Napoca, Romania
[3]Institute for Environmental Sciences, enviroSPACE
University of Geneva 1227 Carouge, Switzerland
[4]United Nations Environment Programme
Global Resource Information Database 1211 Châtelaine, Switzerland
bektasfi@itu.edu.tr, danut.mihon@cs.utcluj.ro, vlad.colceriu@cs.utcluj.ro, karin.allenbach@unige.ch, goksel@itu.edu.tr, ozgur.dogru@itu.edu.tr, gregory.giuliani@unepgrid.ch, dorian.gorgan@cs.utcluj.ro

***Abstract*—Developing applications for analyzing and processing different remotely sensed data is very important for environmental predictions and management strategies. Applications focusing on environmental and natural resource monitoring need large data sets to be processed and fast response to actions. These requirements mostly imply high computing power that can be achieved through the parallel and distributed capabilities provided by the Grid infrastructure. This paper presents the GreenLand application as a user friendly web based platform for the use of environmental specialists engaging remote sensing applications using Grid computing technology. Theoretical concepts and basic functionalities of GreenLand platform were tested in two detailed case studies: a land cover/use determination analysis in Istanbul (Turkey) by conducting vegetation indices and density slice classification on Landsat 5 Thematic Mapper (TM) imagery, and the retrieval of large remote sensing products datasets (The Moderate Resolution Imaging Spectroradiometer (MODIS)) for the entire Black Sea Catchment. All the results of different image processing scenarios used in the reported experiments have been developed through the enviroGRIDS project, targeting the Black Sea Catchment (BSC) area.**

Keywords*—GreenLand; Landsat; MODIS; image processing; Grid computing

I. INTRODUCTION

Environmental applications require large volume input data sets that mainly consist of remotely sensed images (some of them up to 1 GB in size). Another important aspect regards the fact that most of the applications in the Earth Science domain use algorithms based on big sets of parameters that have to be combined in a certain way to obtain the accurate results [1]. Remote sensing image processing is a very demanding procedure in terms of data manipulation and computing power. As a result, it is mostly impossible to obtain reasonable processing times in environmental applications by using a stand alone machine. Grid infrastructure provides the solution of this problem, by providing parallel and distributed computation methods.

The Grid infrastructure [2] is the execution environment where all the data processing takes place. This emerging technology provides access to computing power and data storage capacity distributed over the globe [3].

Grid computing is the use of multiple computers to solve a single problem at the same time usually a scientific problem that requires a great number of computer processing cycles or access to large amounts of data [4].

The approach presented in this paper is to use the Grid infrastructure that offers high power computation machines that allow parallel and distributed execution of tasks for satellite image processing. The main use case partition into smaller tasks is done automatically at runtime, by the GreenLand platform. The GreenLand application is conceived as free Geographic Information System (GIS) software for geospatial data management and analysis, image processing, graphics/maps production, spatial modelling, and visualization [1]. Grid-based GreenLand platform produced within the enviroGRIDS project [5], and available through the Black Sea Catchment Observation System (BSC-OS) Portal [6], offers scalability when dealing with a large number of users and/or a large processing data volume.

The ability of the Grid based platform tested based on two different case studies using remotely sensed data as Landsat 5 TM (Thematic Mapper) and MODIS (Moderate Resolution Imaging Spectroradiometer). In the first case study, land use/land cover categories of Istanbul, were derived by using remote sensing vegetation indices such as Normalized Difference Vegetation Indices (NDVI) and Enhanced Vegetation Indices (EVI) and density slice classification. In the second case study, a workflow was developed to retrieve two MODIS products MOD15-Leaf Area Index (LAI) and Fraction of Photosynthetically Active Radiation (FPAR) and MOD16 - Surface Resistance and Evapotranspiration (ET) at the scale of the BSC. GRASS (Geographic Resources Analysis Support System) library was used in the study as a code source that necessary for image processing [7].

GreenLand web-based application is able to provide smart solution by automating repetitive processes and using

distributed or Grid computing technology when needed. Due to the computing and storage capabilities offered by the Grid infrastructure, the workflow execution times are significantly reduced in comparison with standalone/cluster processing. Therefore this powerful tool will certainly be very useful for sustainable management of the Black Sea catchment by using remote sensing technology.

II. Study Area And Data

A. Study Area

Case Study I: Determination of Land Cover/Land Use of Istanbul, Turkey

Istanbul is located in north-western Turkey within the Marmara Region on a total area of 5,343 square kilometers. There are several reasons why Istanbul is considered as the test case study for deriving land cover categories phenomena. Humans are increasingly disturbing natural resources, ecosystems and the environments in the city.

As a result, the city is facing serious water quality problems, deforestation, desertification, soil erosion, degradation of land productivity, and the disappearance of biodiversity and sensitive regions. There is an urgent need to determine and monitor the land cover types of the mega city. It is very important to derive land cover/land use information by using freely available remotely sensed data and freely available image processing platforms [1].

Case Study II: MODIS Mosaic at Black Sea Catchment

The Black Sea Catchment area covers more than 2 million square kilometers, overspreading entirely or partially 24 countries. Approximately a hundred and sixty million inhabitants live in this area which is annually frequented by millions of tourists.

One of the aims of the enviroGRIDS project was to assess water resources in the past, the present and the future using the Soil and Water Assessment Tool (SWAT) [8] for the entire catchment. Combined to in-situ data, remote sensing products could be a valuable source of information to improve modeling such spacious and complex environment by providing homogenous datasets over broad area with high temporal resolution.

B. Data

In this paper, two case studies were highlighted. In case study I, 2009 dated Landsat 5 TM data were used to derive land cover/land use categories of Istanbul by using vegetation indices and density slicing classification in GreenLand platform. Landsat 5 TM sensor acquires data in seven spectral bands that cover a wavelength range from 450 nm-2350 nm with a spatial resolution of 30 m. The remotely sensed data were obtained from NASA, by the Warehouse Inventory Search Tool (WIST) [9].

In the second case study, two MODIS level 4 products were selected to develop a workflow which facilitates their retrieval at the scale of the Black Sea catchment; MOD15 and MOD16 [10]. These products are multilayers stacks of 1 km resolution issue from EOS (Earth Observation Services) instrument and freely provided by NASA on 8-day basis in *.hdf* format.

These high level processed products are specially used for monitoring wildfire danger and crop/range drought, and to describe the canopy structure. MOD15 Leaf Area Index (LAI) defines the one-sided leaf area per unit ground area (value between 0 and 8) when Fraction of Photosynthetically Active Radiation (FPAR) measures the proportion of available radiation (400 to 700 nm) that a canopy can absorbs (value between 0 and 1). MOD16 consists of surface resistance and evapotranspiration.

III. GreenLand Overview

The GreenLand [11] main goal is to provide support for geospatial algorithms development in different Earth Science domains. At a general level they can be classified as: vegetation index operators (e.g. NDVI, EVI), correction operators (Dark Object Subtraction for atmospheric correction), satellite images bands manipulation (Mosaic and Extraction of bands), statistical and arithmetic operators (e.g. add/substract/multiply/dividethe pixels from a satellite image by a given constant value), etc.

Currently the platform is used in two major case studies concerning with the Istanbul geographic area, and the Black Sea catchment region.

When studying complex use cases (like the ones presented in this paper) it is hard to model and simulate them as a whole. Instead the domain field specialists need to divide the use cases into smaller modules and to analyze them separately, and only after that they are able to create the global results.

The solution related to these issues, which was implemented in the GreenLand platform, represents the complex use cases based on mathematical notions from the graph theory. This means that each node of the graph represents one of the algorithms within the use case (e.g. NDVI, EVI, Density slicing), while the edges specify the relations between these algorithms.

Usually the complex use cases take a long time to execute, and for this reason the Grid-based data processing solution was adopted. Because of the workflow-like description of the scenarios, the GreenLand can easily optimize the entire execution by creating group of nodes, similar in complexity, that are processed in parallel on different Grid machines.

The nodes of the workflow are not independent one from another; instead their inputs and outputs are connected through uni-directional edges. This means that at runtime some of the nodes will wait until the corresponding ones will complete their execution. Only after that they can start to process the data.

Based on these aspects, the execution of a workflow will always generate multiple partial results, and it is up to the GreenLand platform to combine them and to create the final outputs that corresponds to the main workflow.

The complexity of processing data over the Grid infrastructure is hidden from the user, by implementing special interactive techniques in the graphical user interface, which

Funded by the European Commission seventh research framework through the enviroGRIDS project (Grant Agreement n° 226740).

Fig. 1. Workflows organization in the GreenLand graphical user interface.

allow the creation, instantiation, and execution of the use cases, represented as workflows.

Among other features provided by the GreenLand platform, the following ones are the most important:

- Automatic data retrieval from remote repositories, by using the Open Geospatial Consortium (OGC) [12] standard and the File Transfer Protocol (FTP);
- Manual upload of spatial data from the user local machine;
- Parallel and distributed execution of satellite images over the Grid infrastructure. This involves also the partition of the use case into smaller processes and the tasks schedule over the physical machines;
- The execution optimization by creating groups of workflow nodes that have similar complexities. In this way a balanced Grid processing is achieved;
- The standardized data processing, by using the OGC WPS service, meaning that other external systems are able to access the exposed workflows and to execute them remotely;
- GRASS [13] support for developing new geospatial algorithms that can be further used as nodes within the workflows;
- Data level interoperability with other platforms through the Web Map Service (WMS) and Web Coverage Service (WCS) that are part of the OGC standard;
- Dynamic data visualization, by overlapping the execution results directly over interactive maps.

From the graphical user interface level the specialist is able to execute simultaneously several workflows during the same working session. The GreenLand uses the project concept that can be defined as a virtual container that stores groups of workflows, defined by the user.

Fig. 1 highlights a project example that contains four distinct workflows that were selected from the right side list. Before starting the Grid execution the user must instantiate all these items with specific data inputs.

IV. Implementation And Experiments In Greenland

A. Case Study I: Determination of Land Cover/Land Use of Istanbul, Turkey

Istanbul case study reports an application of remote sensing image processing steps to derive land cover/land use categories especially vegetation areas in Istanbul, Turkey by using GreenLand platform.

As it is depicted in the Fig.2 case study starts with pre-processing of data and then vegetation indices are calculated as the following step. Then, after classification, the accuracy assessment steps are executed and finally the results are presented as thematic maps.

Satellite data pre-processing comprise of radiometric calibrations (atmospheric corrections) for 2009 dated Landsat TM data. The objective of radiometric correction is to recover the "true" radiance and/or reflectance of the target of interest [14]. Conversion from Digital Number (DN) to radiance (analogue signal) was conducted by using calibration parameters such as gain and offset. These are available in published sources and image header files [15].

Equation (1) is used for the calculation of radiance values from DN values:

$$L_\lambda = C_0 + C_1 * DN \quad (1)$$

where L is top of atmosphere (TOA) upwelling radiance, C_0 and C_1 ($mWcm^{-2}sr^{-1}\mu m^{-1}$) are Offset and Gain values, and DN is digital number.

L was converted to TOA reflectance, R (without unit) using the (2).

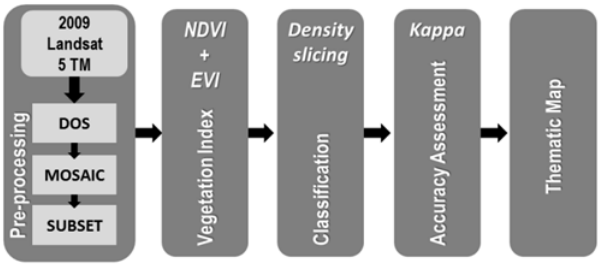


Fig. 2. Flowchart of the Case Study I

$$R = (\pi * L_\lambda * d^2) / (ESUN_\lambda * Z) \quad (2)$$

where R is planetary reflectance, d is Earth Sun distance, L_λ is at-sensor radiance, Z is the solar zenith angle in degrees, and $ESUN_\lambda$ is Mean solar Exoatmospheric irradiances on the top of the atmosphere.

The Dark Object Subtraction (DOS) method was used to correct for atmospheric scattering in the path [16]. DOS is an image-based approach that assumes dark objects exist within an image and these objects should have values very close to zero (such as water bodies), and that radiance values greater than zero over these areas can be attributed to atmospheric scattering and thereby subtracted from all pixel values in an image. The correction is applied by subtracting the minimum

observed value, determined for each specific band, from all pixel values in each respective band.

The DOS method was implemented as a workflow within the GreenLand platform, and has the effect of correcting the satellite images affected by the atmospheric conditions. This workflow takes a single input, representing one band of the satellite image and generates an atmospheric corrected one that maintains its original size, projection, and location.

The DOS algorithm is implemented based on a combination of GRASS functions. The advantage of using these functions inside the GreenLand platform is that they can be involved in the parallel and distributed executions over the Grid infrastructure. This means that at runtime the platform transfers to the Grid machines the input band specified by the user together with the GRASS scripts.

A vegetation index is a number that is generated by some combination of remote sensing image bands and may have some relationship to the amount of vegetation in a given image pixel. Description of vegetation indices tested in this study such as Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) is shown in Table. 1.

TABLE I. REMOTE SENSING INDICES

Index	Reference	Formula
Normalized DifferenceVegetation Index (NDVI)	[17]	(NIR-RED)/(NIR+RED)[a] (3)
Enhanced Vegetation Index (EVI)	[18]	2.5*(NIR-RED)/(NIR+C1*RED-C2*BLUE+L)[b] (4)

[a, b] NIR (Near Infrared), RED, BLUE denotes reflectance values derived from Landsat 5 TM bands of B4, B3, and B1, respectively and C1 =6, C2 =7.5 and L=1 for EVI index.

The Normalized Difference Vegetation Index (NDVI) is one of the oldest, most well-known, and most frequently used Vis (Vegetation Indices). The combination of its normalized difference formulation and use of the highest absorption and reflectance regions of chlorophyll make it robust over a wide range of conditions. It can, however, saturate in dense vegetation conditions when LAI becomes high. The value of this index ranges from -1 to 1. The common range for green vegetation is 0.2 to 0.8 [17].

The enhanced vegetation index (EVI) was developed as an alternative vegetation index to address some of the limitations of the NDVI. The EVI was specifically developed to be more sensitive to changes in areas having high biomass (a serious shortcoming of NDVI), reduce the influence of atmospheric conditions on vegetation index values, and correct for canopy background signals. EVI tends to be more sensitive to plant canopy differences like LAI, canopy structure, and plant phenology and stress than does NDVI which generally responds just to the amount of chlorophyll present. The value of this index ranges from -1 to 1. The common range EVI value for green vegetation is 0.2 to 0.8.

Two alternative indices were taken into account (NDVI and EVI) and implemented within the GreenLand platform as independent workflows. The development of these algorithms is based on the formulas described in Table 1.

For this experiment the 2009 dated Landsat 5 TM satellite image bands were used. The NDVI requires the Red and NIR bands as input, while the EVI workflow expects the usage of valid inputs for the Blue, Red, and NIR layers (Table 2).

TABLE II. VEGETATION INDEX OPERATORS

NDVI Uses two single band images and creates a NDVI image	**Inputs:** 1.Image representing the Red band (Geotiff) 2.Image representing the NIR band (Geotiff)
	Outputs: 1. NDVI image band (Geotiff). Each pixel is in the [0,1] range
EVI Uses three single band images and creates an EVI image, highlighting areas of increased vegetation	**Inputs:** 1.Image representing the Blue band (Geotiff) 2.Image representing the Red band (Geotiff) 3.Image representing the NIR band (Geotiff)
	Outputs: 1. EVI image band (Geotiff). Each pixel is in the [0, 1] range

It is worth mentioning that the order in which the inputs are specified by the user is very important, and should be identical with the one that is used inside the algorithms (see the two computation functions in Table 1). If the inputs are switched, the workflow will not fail at runtime, but will generate an erroneous result.

Fig.3 highlights how these concepts are mapped for the NDVI workflow.Without going into further details; we can say that the inputs specification process is similar for the rest of the existing resources.

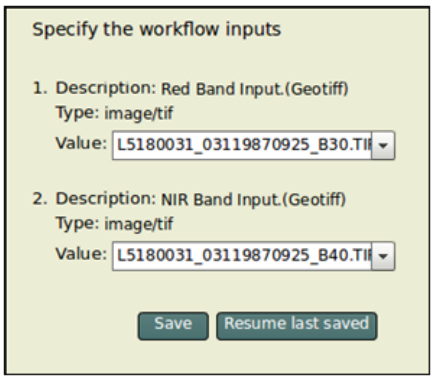


Fig. 3. The inputs specification for the NDVI workflow.

When executing the NDVI workflow, the system automatically generates its internal representation that is used

for the parallel and distributed execution over the Grid infrastructure.

Based on the NDVI formula, a graphic representation is highlighted in Fig.4, together with the XML-based structure.

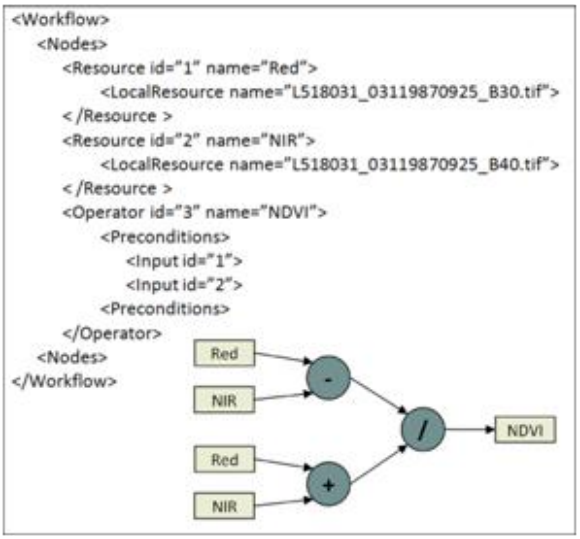


Fig. 4. The XML and graphical representations of the NDVI

Digital image classification uses the spectral information represented by the DN in one or more spectral bands, and attempts to classify each individual pixel based on this spectral information. The resulting classified image is comprised of a mosaic of pixels, each of which belong to a particular theme, and is essentially a thematic "map" of the original image [19].

Density slicing, also known as double threshold, is a classification technique using computer processing of digital data [1].

Density slicing allows the user to define sub-intervals for characterizing the data. The advantage of density slicing is that it allows one to gain a greater degree of variability of brightness within the remotely sensed image compared to the original image (e.g. black and white imagery). The method works best if the range of brightness values covers a single band of frequencies. Each interval is then assigned a color value. The intervals may be defined based on the application. The range of input pixel values is assigned a single output pixel value in a density sliced image.

The range of pixel values may be defined by the user. Density slicing is most effective when the value of particular pixels have significance to a physical variable.

The density slicing workflow acts like a pseudo-color algorithm, used in the creation process of the thematic maps that can be shared and analyzed by the scientific communities.

If in the case of the DOS, NDVI, and EVI workflows the input data type was identified as satellite image, in the density slicing algorithm a new type had to be created that represents the classes range chosen by the user at the graphical interface level.

Once this information was specified, internally the algorithm loops through the entire pixels structure of the satellite image and assigns to each item a specific color. In the end a thematic map is generated.

If more than one domain field specialists are involved in the scenarios development process (e.g. Istanbul case study) the GreenLand platform provides a collaborative environment for developing new algorithms and workflows, but also for visualizing and analyzing the results.

Not in all cases the scientific community members are using the same applications for local development of the scenarios. For such situations it is suitable to create standard services that can be used by all these tools. This is the case of the visualization and interpretation of the GreenLand results that can be access through the OGC standard.

For sharing the density slicing results among the scientific community members, the GreenLand platform provides the Web Map Service (WMS) that is able to expose the output in a standardized format. Its two operations (GetCapabilities and GetMap) allow the user to periodically query the data repository and to retrieve, as a static image, the data they are interested in.

In order to create a dynamic visualization environment, the image returned by the GetMap operation is overlapped onto an interactive map. Once the image is displayed, the user is able to extract relevant information from a specific area within the image boundaries. The area selection is also interactive, and can be performed directly with the mouse. In case of higher accuracy a set of input fields are provided where the user can specify a more detailed area.

Fig. 5 exemplifies the visualization of the NDVI result, after applying the density slicing algorithm, based on the WMS service.

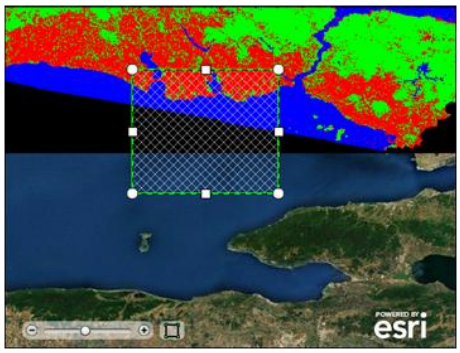


Fig. 5. The standardized visualization of the NDVI density sliced result.

Classification accuracy is the main measure of the quality of thematic maps produced and required by users, typically to help to evaluate the fitness of a map for a particular purpose [20]. Ground truth and classified classes are compared to assess classification accuracy. Error matrix is constructed for this comparison [21]. Each row of the matrix is reserved for one of the information classes used by the classification algorithm. Each column displays the corresponding ground truth classes in an identical order. The diagonal elements of the error matrix show the number of pixels classified correctly in each class.

In this research, to assess the accuracy of classification, the error matrix and some common measures derived from this matrix namely, overall accuracy, user's accuracy, producer's accuracy and kappa coefficient are used. The confusion matrix is a simple cross-tabulation of the mapped class label against that observed in the ground or reference data for a sample of cases at specified locations [21].

The accuracy assessment workflow, implemented in the GreenLand platform, determines the quality of the Istanbul land cover/land use classification, by comparing the obtained results with the ground based measurements.

The accuracy assessment workflow implements a set of GRASS functions, in order to obtain the required statistics. It expects two inputs: the classified image and a vector file that contains the ground based measurements.

B. *Case Study II: MODIS Mosaic at Black Sea Catchment*

Large amount of existing remote sensing products are freely available through the web. In this paper, a challenge was to find freely-available products of scientific quality, covering the entire Black Sea catchment, with large temporal availability, necessary especially for improving crop's type monitoring and for helping SWAT's results validation. Two MODIS products MOD15 - Leaf Area Index (LAI) and Fraction of Photosynthetically Active Radiation (FPAR) and MOD16 - Surface Resistance and Evapotranspiration (ET) presented great interest in validating SWAT result in the Vit river basin in Bulgaria [22]. Therefore, developed flowchart has the ability for the validating the results of SWAT models applied in BSC at local and regional scale.

A specific workflow "BlackSeaMosaicPDG" was developed in the GreenLand platform which permits to retrieve directly these products at the scale of the Black Sea catchment. Twelve tiles are necessary to cover the entire area of interest. The flowchart (Fig. 6.) consists in downloading one-year time series from an FTP server, then extracting the bands separately and mosaicking adjacent tiles together in a single operation.

There are several disadvantages in retrieving large datasets without any automation help. On top of requiring specific software for the analysis, repetitive processes are very unexciting, time consuming and require powerful computing and storage capabilities. Moreover processing made on a stand-alone computer without publishing results on web-based application are not made visible to others, and therefore the reuse of retrieved datasets remains unlikely.

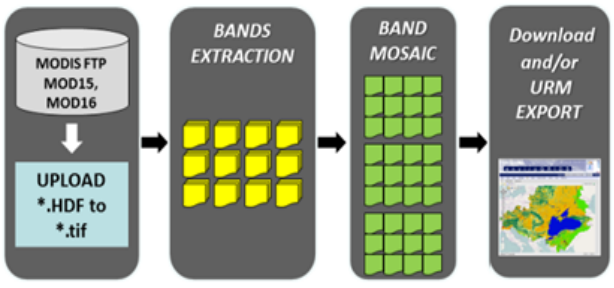


Fig. 7. Flowchart of Case Study II.

The MOD15 and MOD16 products are used in the implementation of this case study. The difference between them is in terms of bands organization, different repositories where these products are stored, and different structure inside the repositories.When implementing it within the GreenLand platform, the main goal was to provide an easy to use graphic interface that hides from the user the entire data retrieval and Grid execution processes.

At the graphical user interface level of the application the user is required to specify only the processing year and the bands of the two MODIS products that he is interested in. In the background the application automatically retrieves the related satellite images from the corresponding remote data repository and transfers them onto the Grid machines, where the tasks are going to be processed.

In order to optimize the execution process, the GreenLand platform partitions the use case into groups of tasks, where each group integrates five BlackSeaMosaicPDG workflows, instantiated with different input data sets. The content of each workflow consists in processing the selected bands of one MODIS product, for an entire year time period.

V. RESULTS

A. *Case Study I: Determination of Land Cover/Land Use of Istanbul, Turkey*

The selected region used in this study contains diverse land cover types, including vegetative area, high and medium density built up spaces (artificial surfaces-other), and water surfaces. Fig. 7 shows that the NDVI and EVI values for the area are consistent with the theoretical values.

Fig. 7 indicates the NDVI values fluctuated from – 0.40 to +0.80. In the figure, the values between 0.4 and 0.8 indicate the green areas in Istanbul. The positive values (bright pixels) less

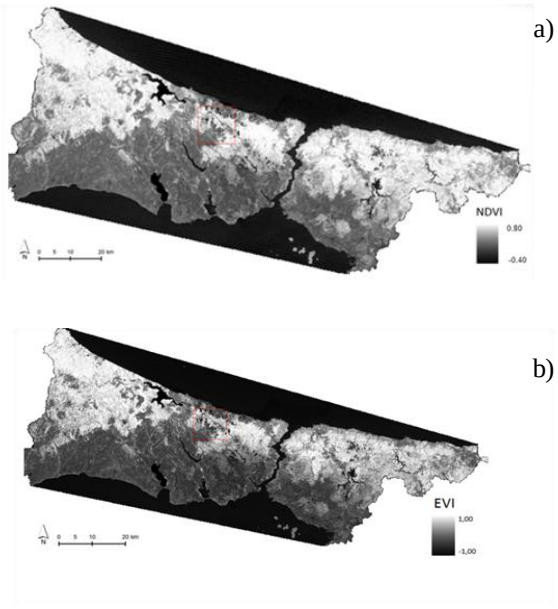


Fig. 6. Results of GreenLand a)NDVI b)EVI.

than '0.4' indicate the other artificial surfaces in Istanbul. Therefore, the negative values (dark pixels) indicate water surfaces in the selected region.

Density slicing classification result is given for NDVI image in Fig. 8.

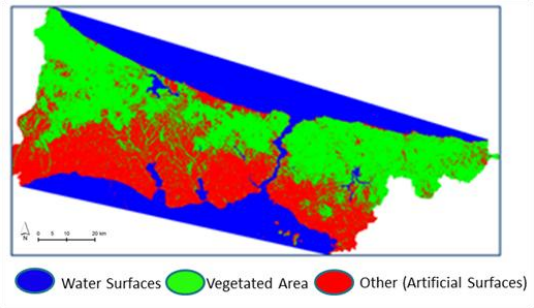


Fig. 8. Classified image.

Accuracy assessments were applied based on Foody, 2002 [21]. 25 random points were selected for accuracy assessment. By using error matrix, the overall accuracy (OA) was calculated as 0.92 and Kappa was calculated as 0.87 for classified NDVI and EVI image. Although errors and confusion exist because of mixing problem, these two indices (NDVI and EVI) showed satisfying classification results (OA and Kappa > 0.80).

B. Case Study II: MODIS mosaic at Black Sea catchment

After a couple of hours of processing time on powerful servers, 45 dates (one year) of four MODIS mosaicked products are available in geotiff format as input for the application, for download or for direct export into the enviroGRIDS The Unified Resource Management (URM) portal [23] using OGC standards.

Such development simplifies the access to LAI (Fig. 9.), ET and FPAR MODIS product collections at the scale of the Black Sea catchment, by considerably reducing time for data processing without needing any particular remote sensing skills neither specific software, while benefiting of GRID technology to process and to store voluminous datasets.

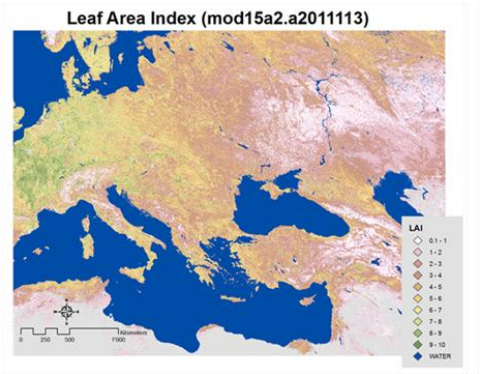


Fig. 9. Mosaic of 12 tiles of LAI band (MOD15) on 23 April 2011.

VI. CONCLUSION

Grid technologies provide powerful tools for huge volume of remotely sensed data sharing and high performance processing. After an overview of the recent initiatives of 'gridifying' satellite image processing, two specific usage scenarios in which the Grid is conceived as a powerful computing resource were analyzed. GreenLand web-based platform and application are able to provide smart solutions by automating repetitive processes and using distributed or Grid computing technology when needed. Moreover this application is linked to the EnviroGRIDS URM geo portal where processed results could directly be exported according OGC standards, increasing visibility of existing datasets and encouraging the reuse of processed and available data. Studies to extend the capabilities of the GreenLand application are in progress. Above stated case studies prove that GreenLand is a useful and flexiable platform to implement open web based remote sensing applications.

REFERENCES

[1] enviroGRIDS D2.11 Report, Remote Sensing Services, 2012 http://www.envirogrids.net/

[2] D. Mihon, V. Bâcu V, T. Ştefănuţ, and D. Gorgan, "Considerations on the Grid Oriented Environmental Application Development Framework", IEEE 6th International Conference on Intelligent Computer Communication and Processing, p. 419-426, 2010.

[3] B. Jacob, M. Brown, K. Fukui, N.Trivedi, "Introduction to GridComputing", SG24-6778-00. http://ibm.com/redbooks, 2005.

[4] C. G. Serban, C. Maftei, C. Filip, "Assessment of Multi-spectral Vegetation Indices using Remote Sensing and Grid Computing",International Journal of Computers, i.4, v.5, p 469 -475, 2011.

[5] D. Gorgan, V. Bâcu, T. Ştefănuţ, D. Rodilă and D. Mihon, "Earth Observation application development based on the Grid oriented ESIP satellite image processing platform", Journal of Computer Standards & Interfaces, Vol. 34, November, pp. 541–548,2012.

[6] D. Gorgan, V. Bâcu, D. Mihon, T. Ştefănuţ, D. Rodilă, P. Cau, K. Abbaspour, G. Giuliani, N. Ray, A. Lehmann, "Software platform interoperability throughout enviroGRIDS portal", in International Journal of Selected Topics in Applied Earth Observations and Remote Sensing – JSTARS, Vol. 5/6, pp. 1617-1627, 2012.

[7] Neteler M., Mitasova. H., 2004: Open Source GIS: A GRASS GIS Approach, Kluwer Academic Publishers/Springer, Boston, Dordrecht, London.

[8] D. Gorgan, V. Bacu, D. Mihon, D. Rodila, K. Abbaspour, and E. Rouholahnejad E, "Grid based calibration of SWAT hydrological models", Journal of Nat. Hazards Earth Syst. Sci., Vol. 12/7, ISSN: 1561-8633, pp. 2411-2423, NASA, Warehouse Inventory Search Tool (WIST): https://wist-ops.echo.nasa.gov/api/

[9] Moderate Resolution Imaging Spectroradiometer, http://modis.gsfc.nasa.gov/

[10] D. Mihon, V. Colceriu, F. Bektaş, K. Allenbach, M. Gvilava, D. Gorgan, "Spatial Data Exploring by Satellite Image Distributed Processing", Geospatial Research Abstracts, EGU General Assembly, vol.14, 2012.

[11] OGC standard, "OpenGIS Web Service Common Implementation Specification",http://www.opengeospatial.org/standards/common

[12] M. Landa, "New GUI for GRASS GIS Based on wxPython", Departament of Geodesy and Cartography, pp.1-17, 2008.

[13] R. G. Lathrop,"The integration of remote sensing and geographic information systems for Great Lakes water quality monitoring", PhD Thesis, University of Wisconsin, USA, 1988.

[14] G. Chander, &B. Markham, "Revised Landsat-5 TM Radiometric Calibration Procedures and Post calibration Dynamic Ranges", IEEE Transactions on Geoscience and Remote Sensing 41, 2674-2677, 2003.

[15] M. S. Moran, R. D. Jackson, P. N. Slater, and P. M. Teillet, "Evaluation of simplified procedures for retrieval of land surface reflectance factors from satellite sensor output", Remote Sensing of Environment, v.41, p. 169– 184, 1992.

[16] J. W. Rouse, R. H. Haas, J. A. Schell, and D. W. Deering,**"**Monitoring Vegetation Systems in the Great Plains with ERTS", Third ERTS Symposium, NASA SP-351 I,p. 309-317,1973.

[17] A. Huete, C. Justice, and H. Liu, "Development of vegetation and soil indices for MODIS-EOS", Remote Sensing of Environment, v.49, p. 224– 234, 1999.

[18] P.S. Roy, S.A. Ravan, "Biomass estimation using satellite remote-sensing data – an investigation on possible approaches for natural forest", J. Bioscience, v.21, p. 535–561,1996.

[19] G. M. Foody,"Harshness in image classification accuracy assessment", International Journal of Remote Sensing, V.29, no.11, p. 3137-3158, 2008.

[20] G. M. Foody, "Status of land cover classification accuracy assessment", Remote Sensing of Environment, v.80, p. 185-201, 2002.

[21] A. Stefanova, " The use of remotely sensed data for validating a SWAT model: Application on the Vit River Basin, Bulgaria", Master thesis, UNESCO-IHE Institute for Water Education, p. 92, 2011.

[22] enviroGRIDS URM data portal (http://www.envirogrids.cz/view)

Calibration of SWAT Hydrological Models in a Distributed Environment Using the gSWAT Application

Victor Bacu, Danut Mihon, Teodor Stefanut, Denisa Rodila, Dorian Gorgan
Computer Science Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
{victor.bacu, vasile.mihon, teodor.stefanut, denisa.rodila, dorian.gorgan}@cs.utcluj.ro

***Abstract*—Topics such as the sustainability and vulnerability of land management practices on water quality and quantity are very important in these days both for decision makers and for citizens. The enviroGRIDS FP7 project addresses some of these topics in the Black Sea Catchment area. One of the software tools developed in this project is gSWAT. It allows the calibration of SWAT hydrological models in a flexible development environment and uses distributed computational infrastructures to speed-up the simulations. The development of SWAT (Soil Water Assessment Tool) hydrological models is a well-known procedure for the hydrological specialists and this paper highlights, from the end-users point of view, the scenarios related with the calibration procedures available in the gSWAT application.**

I. Introduction

Currently a lot of effort is put into topics such as sustainability and vulnerability of land management practices on water quality and quantity. Both decision makers and citizens are interested in these aspects. The enviroGRIDS [1] project, funded by the European Commission (EC) through 7th Framework Programme (FP7) aimed at building capacity in the Black Sea region providing specialists, decision makers and citizens with tools and applications specialized in processing spatial data, processing and visualization of satellite images, calibration and simulation of hydrologic models, etc.

One of the software applications developed in this project is gSWAT, targeting the calibration of SWAT [2] hydrologic models. In the frame of the project a very complex SWAT model of the Black Sea catchment basin has been developed, which required a complex calibration process. For small and medium scale models the calibration process can be easily performed on a desktop computer and in a reasonable amount of time. But for complex models this process is very difficult to be made in this way, mainly because of the size of the model (and the space that is needed in order to store all the results) and in executing all the required simulations in a reasonable amount of time. The gSWAT application addresses this issues and allows a flexible calibration process of complex (but not only) SWAT hydrologic models in a Web based environment. The user has access to high power computational resources and storage space. The execution of simulations is performed in a distributed environment, Grid.

A distributed infrastructure offers high power computation and storage resources, but the access to them is difficult for many users mainly because the interaction with this kind of infrastructure is not made in a graphical manner. For this reason the gSWAT application is developed as a Web application to allow users to access and use the computational resources provided by the Grid infrastructure in the process of hydrologic model calibration. Management of processes, data distribution, task parallelization, monitoring, load balancing, authentication and authorization, scalability represents topics that are solved transparently by the gSWAT application from the user point of view.

In this paper we are presenting the scenarios related with the calibration process available in the gSWAT application. In section 2 are presented notions related to hydrologic models, calibration process, and execution. Several working sessions are presented in section 3. Section 4 presents the architecture and the module of the gSWAT application. Section 5 presents the interoperability aspect of the application by using services and the way in which this is implemented on some particular case.. The performance evaluation is discussed in section 6, section 7 presenting the conclusions.

II. Hydrologic models

Hydrological models are widely used for water resource planning, flood prediction, water quality, etc. They represent, in a simplified manner, the hydrological cycle which can be used for hydrological prediction. Three phases are required in order to provide a good hydrological model: development, calibration and evaluation. Model calibration aims at selecting the best values for model parameters so that the real hydrological behavior can be simulated [3]. Most hydrological models have two types of model parameters, namely physical parameters (represents physical properties of the catchment, which can be measured) and process parameters (represents characteristics which cannot be measured). The objective function measures the difference between the simulated output of the hydrological model and the measured output and in general is based on least squares or maximum likelihood methods.

A classification of hydrological models based on their model structure, spatial distribution, stochasticity, and spatial-temporal application is presented in [4]. Metric models such as Data Based Mechanic (DBM) [5] and Artificial Neural Networks (ANN) [6] are based on observations. ANN uses measured rainfall and runoff data to map the behavior of the rainfall-runoff processes. Physic-based models are using the equations of motion in order to represent hydrological processes. The hybrid physically-based-conceptual models aim at simplifying the model structure.

In the enviroGRIDS project the Soil and Water Assessment

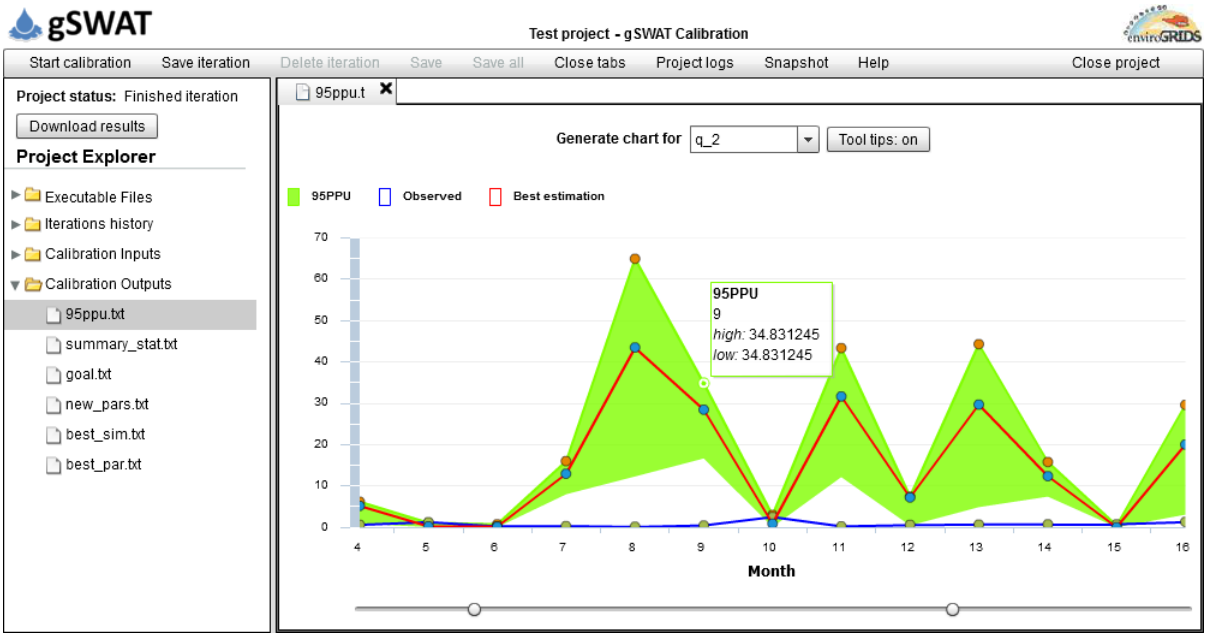


Fig. 1. Output visualization in gSWAT.

Tool (SWAT) has been used to model and simulate the Black Sea catchment basin. SWAT is a continuous simulation model that operates on a daily time step and quantifies the impact of land management practices on water quality and vegetation growth. The calibration and uncertainty analysis is a very important step in the flow of creating a SWAT model. In the enviroGRIDS project the Sequential Uncertainty Fitting program SUFI-2 [7] was used. One advantage of using this algorithm is that the simulations are independent one from another, meaning that we can achieve a high level of parallelism. It allows analyzing a large number of parameters which can be specified by the users.

A recent research paper [12] showed that the Grid technology is suitable for hydrology domain mainly for reducing the processing time. Different studies, such as in [13] and [14] prove that a Grid infrastructure, using efficient planning mechanisms, can lead to an increase of system performances. In [15] the authors present a parallelization framework for hydrological models calibration, but at a reduced scale, using a 24 CPUs cluster. A method involving Message-Passing Interface (MPI) is presented in [16]. A comparative analysis of three method of parallelization of 2D hydraulic models is presented in [17]. The usage of GPUs for processing a 2D flood simulation model is presented in [18] and [19]. Other methods of parallelization are described in [20] and [21].

III. Working Sessions

A. Projects

gSWAT is a Web based application supporting the calibration of complex SWAT hydrological models. It offers both computational resources to minimize the time needed to calibrate the models and storage resources to access remotely the SWAT models and also the results of the calibration process. This application is exposed to the users similar to the Software as a Service (SaaS) level from Cloud. The complexity of the underlying computational infrastructure is hidden and the users can focus on the calibration process rather than aspects related to Grid computing.

A project in gSWAT represents a SWAT model together with other information related to it. The first step to create a new project is to define the project name and description. After this step the user specifies the SWAT model that will be uploaded to the gSWAT server. At the server side the SWAT model is remapped to the structure needed for the calibration process, meaning a new directory structure. The new structure is after that archived and stored on the Storage Element (SE), the LFN (Logical File Name) for the archive being updated in the database. A feedback with the status of this process is provided to the user.

A calibrated model is obtained after a set of iteration steps, each iteration step consisting in executing a variable number of simulations. For each calibration project only one iteration step is the active one, meaning that the user can start only one execution at a time for a calibration process. When starting a new iteration process the user has the possibility to save the previous one, and has access to all iterations that are already executed.

B. Process execution and monitoring

Only one iteration step can be active (in execution) for each calibration project at a time. From the user's point of view the complexity of the calibration process execution over the Grid infrastructure is transparent. From the graphical interface the user selects the start calibration button which will trigger the execution of the steps already detailed in a previous section. Before starting the execution the user should modify all the input parameters that will have an impact on the results. The gSWAT database is periodically interrogated in order to provide users with feedback about the execution (in terms of total execution time and number of completed simulations).

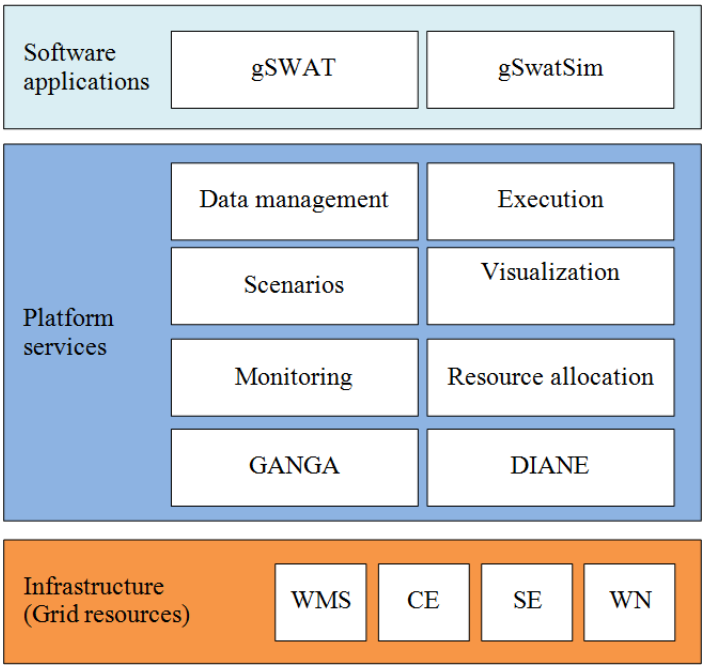


Fig. 2. gSWAT architectural modules.

The user has the possibility to stop the execution of the current iteration step by clicking the stop calibration button. This will trigger the cancelling of all the Grid jobs and cleaning of the current iteration step intermediary files.

C. Input and output data visualization

After the internal structure of the SWAT model is created the user has the possibility to explore it by using the graphical user interface. The text files can be edited directly in the text editor which supports opening multiple files at the same time, and basic operations such as save file, save files, redo, undo, copy and paste, etc.

The output results can be visualized as text or as charts. The chart module parses the 95ppu.txt file and output this data in a graphical manner. The chart presents the best estimated parameters values together with the observed values (see Figure 1). The user has the possibility to adjust the horizontal axis which represents the temporal scale. All the output data can be downloaded, as an archive, by the user. This archive is created on the fly when the user tries to download it.

IV. GSWAT APPLICATION

The gSWAT application [22], [23] is based on the client-server architectural model and uses Web 2.0 technologies in order to provide a flexible calibration interface for different categories of users, such as hydrology specialists or students.

By exposing an intuitive graphical interface, the gSWAT application overcomes the command line based interface exposed by gLite [8]. GANGA [9] offers a flexible programming interface and facilitates the accessibility to Grid infrastructures. DIANE (DIstributes ANalysis Environment) [10] provides an efficient usage of Grid infrastructures and it is based on the master-slave paradigm. The gSWAT application is using both GANGA and DIANE to provide a flexible environment and to minimize the execution time.

A. General architectures

The architecture is composed of three layers, where each layer provides different functionalities (presented in Figure 2). The distributed infrastructure that is used to minimize the calibration time is the Grid infrastructure. The services layer offers services both for the graphical user interface and for other applications that are interconnected with it, such as BASHYT. The graphical user interface is built in Adobe Flex and being a web based interface it can be used from different devices (such as desktops, laptops or even tablets). The layers are similar to the ones in Cloud computing, the infrastructure level can be mapped to the Infrastructure as a Service (Iaas), the Platform services can be mapped to the Platform as a Service (PaaS) and the software applications can be mapped to the Software as a Service (SaaS). An experimental study of migration of scientific applications (where the experiment was made on the gSWAT application) from Grid to Cloud Cluster infrastructure was presented in [11].

B. gSWAT Modules

1) Data management: In gLite a Storage Element (SE) offers a uniform access to various data storage resources (such as disk or tape) and allows users and applications to store/retrieve data in a very simple manner. From the users point of view the file location is hidden, he has access to files based on a logical file name. The data could be replicated to several SEs in order to minimize the transfer cost or to

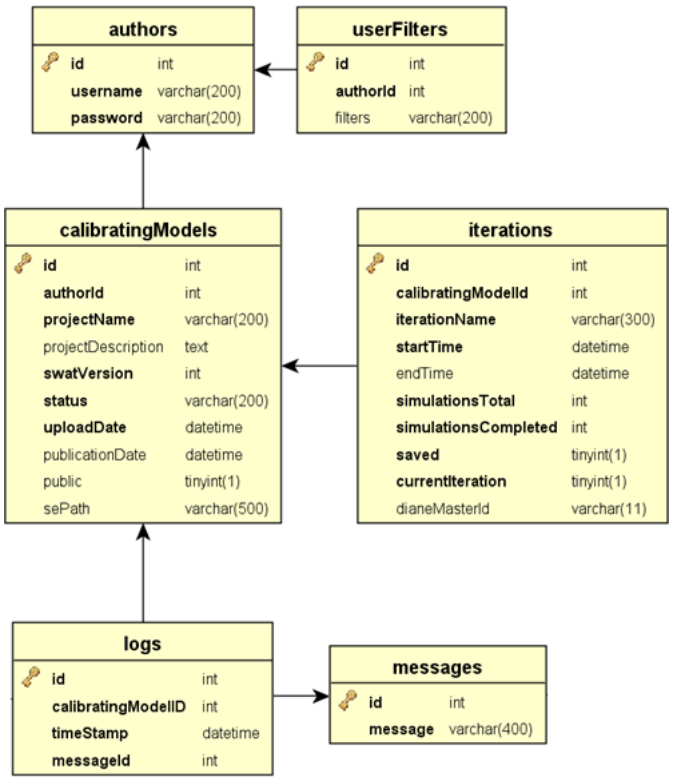


Fig. 3. Database structure.

increase the availability of data. Files are shared by the users in a Virtual Organization (VO) and are protected by security mechanisms. In such environment the files (data) are written once and they cannot be modified, the only solution for doing this is by removing and replacing the files. The protocol that is used by the SEs is GSIFTP which offers a high-speed, reliable and secure data transfer.

This module is responsible mainly for exchanging data to and from the Grids Storage Element. It offers services related to this functionality which provides a transparent access to data resources for the users. A specific data structure is needed by the calibration process and this module creates the necessary directory structure and store the SWAT model to the SE. Another service provides the results from the execution of calibration.

Figure 3 presents the database that stores all the information related to projects, iterations, etc. In gSWAT, each hydrological model is represented in the database as a calibrating model. The most important information about calibrating model are the SWAT version (is used to know which executable is needed in order to execute the model simulations), the logical file path (is used to retrieve the SWAT model from the SE after each job is started on a WN), status (is modified by the execution module to update the state of the calibration process and is used by the graphical user interface to inform the users about the current state).

The status of the calibrating model could be one of the followings:

1) Empty the project doesnt have a valid SWAT model attached to it;
2) Uploading the SWAT model archive is fetched to the gSWAT server, validated and transformed to the structure needed by the calibration process and finally uploaded to the SE;
3) Incomplete uploading the SWAT model is not valid, or another problem occurred when storing the model on the Grid repository (missing Grid certificates, problem in communicating with LFC server, transfer error, etc.);
4) Loaded the project contains a valid SWAT model stored on the SE and on which the calibration process can start;
5) Finished - the current iteration execution is completed and the model can be used to define and execute scenarios;
6) Running a iteration execution is currently ongoing;
7) Incomplete iteration some errors occurred during the execution (bad SWAT model, missing files, etc.).

For each calibrating model there can be zero to many iterations steps, but only one is currently active (meaning is in running). The users have the option to visualize all the input and output data related to one iteration. The start and finish time of the execution is stored in the database, the execution time for each individual simulations can be retrieved from the output files. The number of simulations that are completed is updated by the monitoring module and is reflected in the graphical user interface.

In a dynamic environment, such as Grid, errors can occur at different level, data or execution. In order to minimize the possible errors due to data the data management module tries to detect and recover the execution.

2) Execution: In order to validate a SWAT model a complex calibration process is being conducted, this process being completed when a calibration criteria is satisfied. By performing a variable number of iteration steps we try to accomplish this goal. In each iteration step several simulations of the SWAT model are executed (independently on the other ones) by performing 3 phases (presented in Figure 4): pre-processing, actual execution and post-processing.

Because the complexity of the pre-processing phase is not very high this phase is performed at the server side, once for each iteration step. The user has the possibility to modify some parameters of the SWAT model by defining intervals from which, by using the Latin hypercube sampling method, new parameters values are generated. The outcome of this phase is a list of new parameters values (one list for each simulation needed) which will be propagated in the next step in the SWAT model.

The most complex step is the actual execution of simulations. The execution module uses DIANE and GANGA to interact with Grid jobs. DIANE is used to start and manage the execution of simulations. Each simulation is mapped in DIANE as a task which will be executed on a Grid WN. GANGA starts the Grid jobs and connects to DIANE master

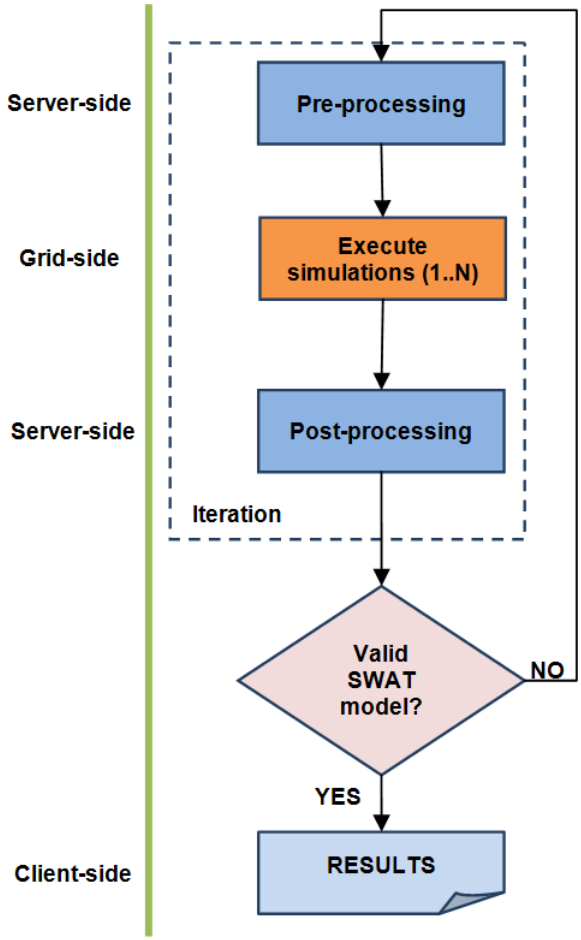


Fig. 4. Calibration steps.

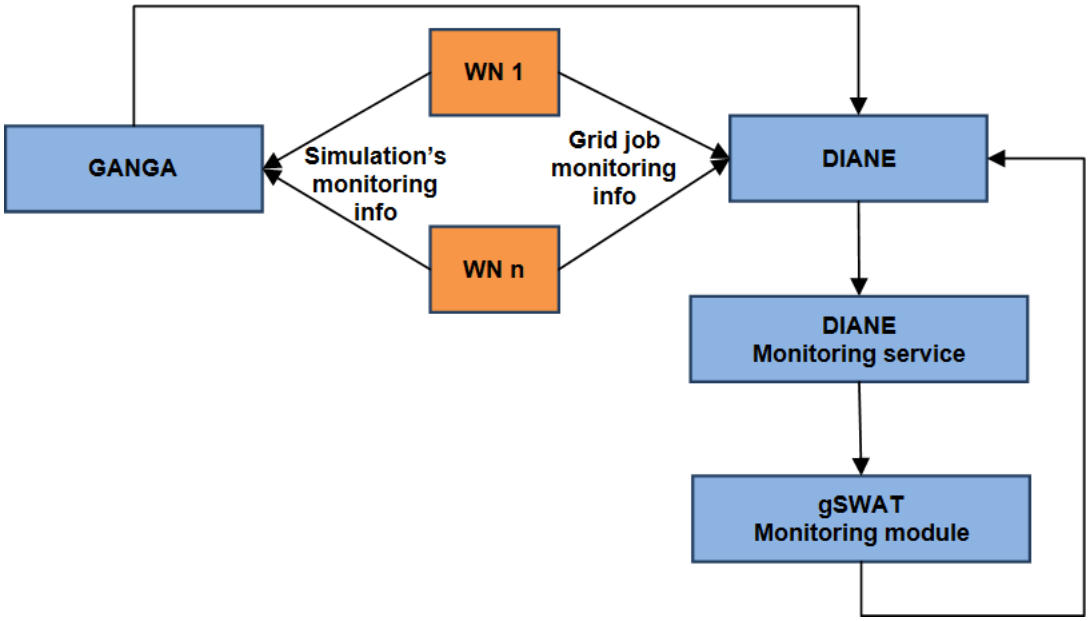


Fig. 5. Monitoring module flow of actions.

in order to receive tasks to be executed. Because the number of simulations that should be performed is very high (varying in general from 200 to 1000) and in order to minimize the number of Grid resources that are used each Grid job will execute one or more simulations.

The following steps are performed by the execution module:

- Define the script that will be executed by the Grid job. This script will copy locally the SWAT model archive (stored on the SE), extract the files, modify the model parameters accordingly to the new values generated in the pre-processing phase, execute the SWAT simulation and in the end archive and send back the SWAT outputs;
- Define the DIANE script that maps the simulations to be performed to tasks;
- Start a new DIANE master for each iteration step on a different port to which the Grid jobs (DIANE workers) will connect;

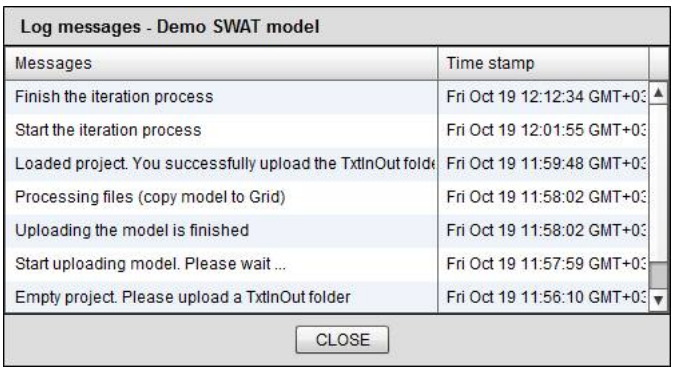


Fig. 6. Detailed log messages.

- Start Grid jobs using GANGA. After this each job will connect to the DIANE master and will receive tasks (simulations) to be executed;
- Monitor the execution of the simulations and store this info to the gSWAT database from which the graphical user interface will provide feedback to the users;
- Download the output results from each simulation at the server side.

The final phase is the post-processing which is also executed at the server side and creates the output for the current iteration step based on the output results provided by each simulation. In the graphical user interface the users have the possibility to visualize, in a graphical manner, the results or to download the files.

3) Scenarios: Scenarios can be defined starting from a calibrated SWAT model to highlight different aspects regarding the modeled catchment basin. The gSWATSim module allows the execution of basic scenarios which are created by modifying some of the model parameters. Similar to the execution module, it uses the Grid infrastructure to run the scenarios. It offers a complex execution and management solution and also the possibility to integrate some of the functionalities in other applications.

The database related to scenarios stores information such as: scenario name, scenario description, scenario fingerprint, SWAT version, status, scenario location on the Storage Element and scenario execution output location on the Storage Element. Scenario execution means the execution of only one simulation. The output could be fetched to other applications, in order to visualize to results in a graphical manner. The following steps are performed by this module in order to run scenarios:

- Start the DIANE master;

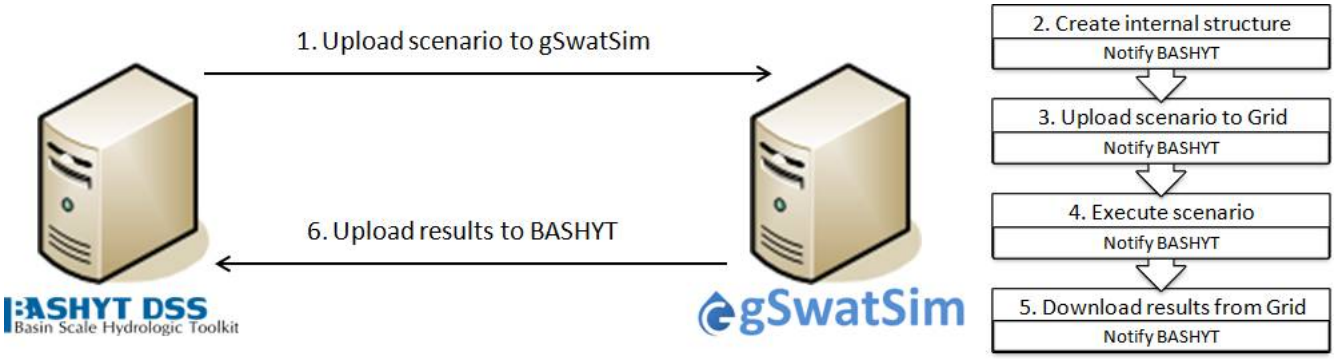


Fig. 7. gSwatSim and BASHYT service based interoperability.

- Start one GANGA worker;
- Execute the job on the Worker Node (WN) (meaning copy locally the SWAT scenario archive from the SE, execute SWAT simulation, archive the output results and upload them to SE).

4) Monitoring: The monitoring module is used both to check the execution flow and to provide feedback to users. Each calibration project has attached a status field (in the database) that characterizes the current situation. This status is updated based on commands initiated by users (uploading a new project, change the project information, etc.) or based on the execution of the iterations. Every job that is executed in the Grid environment has one of the following states: submitted, waiting, ready, scheduled, running, done, cleared. These states are reached with the successful execution; other states are reached with the failure of execution. The status information that is provided by the gSWAT application is different that the states reachable by Grid jobs. These states were presented in a previous section.

The DIANE monitoring system is used by the gSWAT application to gather information regarding the execution flow (meaning the number of simulations that were successfully executed). The flow of interaction between the different components of the system is presented in Figure 5. The DIANE master connects to the monitoring server (which is gridmsg101.cern.ch) and the monitoring messages are sent automatically to it. The DIANE master updates the status based on the information received from the Grid WNs. Two levels of execution status is available, one from DIANE which is responsible for providing info at a higher level, (meaning at simulation level) and the other one from GANGA which is used to monitor the Grid jobs and provide info at a lower level (which is important mainly to recover the execution if some error occurred). The Diane Dashboard application makes available all the monitoring messages in JSON format. The JSON format is used to transfer structured data between a server and a web application. The monitoring module has incorporated a JSON parser that update the gSWAT database with relevant information, such as start time, end time (if it is available), total simulations, completed simulations. At predefined time interval, the information is updated also in the graphical user interface.

The status of the calibration projects offers only limited information about it. Beside this, the user has access to more detailed information about the progress of the calibration process in the form of system logs. Every time a calibration project changes his status, much more detailed message info is stored in the database together with a timestamp (used to be able to order the messages). The user can visualize system logs related to a single calibration project or for all of his calibration projects (Figure 7).

5) Resource allocation: The calibration process for large scale SWAT models is quite complex (mainly because of the size of the model and the number of simulations that are needed to be performed) and in order to minimize the execution time and also to improve the usage of Grid resources the resource allocation module [25] selects the optimum number of resources that are needed. The model complexity is defined based on the number of files, model size and an estimated complexity provided by the specialist in hydrology. Other important aspects are the availability of the Grid resources (free WN) and also the number of users that are using the application. The steps followed by this module are the following: gathering requirements (specified complexity, number of files, model size, etc.), discovering Grid resources (available WNs, waiting jobs, etc.) and determining the necessary resources (based on the requirements and the available resources).

The actual execution time has the following mathematical expression:

$$TotalExecutionTime = T(resourceAllocation) + T(copyModel)$$

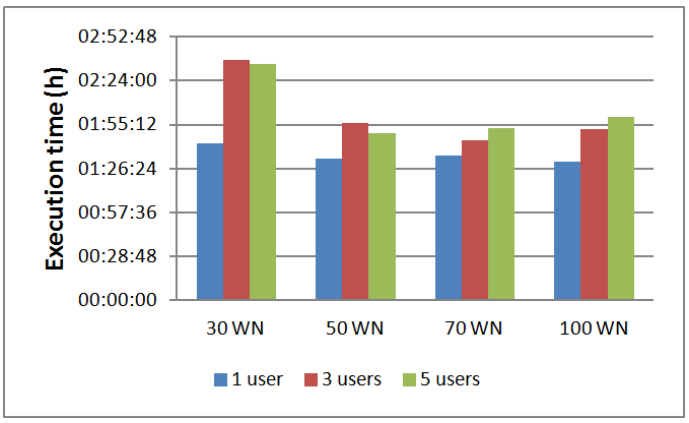


Fig. 8. The variation of execution time with the number of users.

+ T(propagateModifications) + T(actualExecutionOfSWAT) + T(getOutputData)

V. gSwatSim Interoperability

The gSWATSim [26] exposes a collection of REST Web Services [27] that allows the user to create new projects (scenarios), to modify some information about the projects (such as project name, description, etc.), to run scenarios, to upload output results to BASHYT, etc.

BASHYT [28] offers in a Web based interface the possibility to produce reports for SWAT models in a flexible manner. The interoperability between gSWATSim and BASHYT brings some advantages:

- scenarios are developed in a flexible environment by using BASHYT functionalities;
- by using GRID capabilities, gSwatSim speeds up the processing (simulation) of large scenarios;
- the results can be visualized by using BASHYT dedicated tools and modules.

The interoperability between gSwatSim and BASHYT is presented in Figure 6. The first step is to upload scenario to gSwatSim. At server side the internal structure is created and BASHYT is notified about it. After that the scenario is archived and uploaded to SE from where it will be available. The next step is to execute the scenario and store the results to SE. The results are downloaded by gSwatSim from the Grid and uploaded to BASHYT. Notification messages are sent to BASHYT each time the status of the scenario execution changes. In the end the output results can be visualized in BASHYT.

VI. Performance evaluation

A. Distributed infrastructure

By using a distributed infrastructure we gain computational power, efficient storage solution and flexibility. From the user point of view the access to the distributed infrastructure is made automatically. The computational resourced needed by the gSWAT application are provided by the enviroGRIDS project VO. Currently three CEs are providing resources for it but the main CE is ce01.mosigrid.utcluj.ro providing 128 physical CPUs, with a 1024 logical CPUs. This VO is using one SE (se01.mosigrid.utcluj.ro) with a storage capacity of 13 TB. Being a production site, and not just a test site, the availability of resources is not constant (the resources are shared with other VOs), this being reflected on the experiments that were made on it. A comparative analysis of parallel execution of SWAT hydrological model on multicore and Grid architectures is presented in [24].

B. Black Sea catchment basin calibration results

The gSWAT application is addressed to specialists in hydrology to help them to calibrate complex SWAT model. It can also be used as a teaching tool in workshops related to SWAT and calibration. The total area of the Black Sea Basin is around 2.3 million km2 with rivers from 23 countries. A complex SWAT model consists of a very high number of files (at least 1.000.000 files).

For the first experiments we have used a small scale model. The size of the SWAT model archive stored on SE is 256 MB and the size of the extracted archive is 327 MB. The number of input files, without the ones from the backup directory, is 17,990 files and the number of the hydrological sub-basins is 1,629. The number of input parameters for this model was 14. The variables for this experiments are the number of simulations (100, 500 and 1000 simulations), and the number of allocated WNs (30, 50, 80 and 100 WNs).

1) gSWAT scalability with the number of user: A first experiment targets the scalability of the application in terms of number of users that are performing calibrations. In Figure 8 is represented the influence of the number of users on the overall execution time of the calibration process. A first remark is that the calibration time when only one user is running the application is lower than when 3 or 5 users are also performing a calibration. This is obvious because only one user is using the Grid resources. It is also important to notice that even though the execution time increases with the number of users it is not a linear increasing. The overall execution time is higher mainly because the number of Grid resources is not scaled with the number of users and the Grid services have to manage more jobs. The number of Grid resources was fixed and the other VOs could use them as well, reducing in this way the number of possible computational resources for gSWAT. In all cases the overall execution time decreases when adding more computational resources even though more users are performing calibrations.

2) gSWAT scalability with the number of computational resources: Another experiment aims to show what is the influence of the number of computational resources used (WNs) on the overall execution time. When adding more resources the execution time should decrease. The improvement is not in all cases proportional with the additional computational resources that are used. In Figure 9 are presented the results. The execution time decreases when adding more resources, the decrease is accentuated better when the number of simulations is higher. The trend is the same even if the number of simulations is 100, 500 or 1000, proving in this way the scalability of the application with the number of simulations and with the number of computational resources. In some cases even though we add more resources the speedup is small and it shows that is not always a good idea to add more resources.

Table 1 presents the speedup (by percentage values) gained by increasing the computational resources from 30 to 50 WNs, from 50 to 80 WNs and from 80 to 100 WNs. If we increase the number of WNs from 30 to 50 we gain 44% for 100 simulations. However, the number of resources needed is with 67% more. The speedup gained by increasing the computational resources from 80 to 100 (and any number of simulations),

TABLE I. Speedup percentages

Sn=T1/Tn, n=#WN	100 Sims	500 Sims	1000 Sims
50 WNs / 30 WNs	44%	83%	40%
80 WNs / 50 WNs	14%	31%	57%
100 WNs / 80 WNs	3%	12%	13%

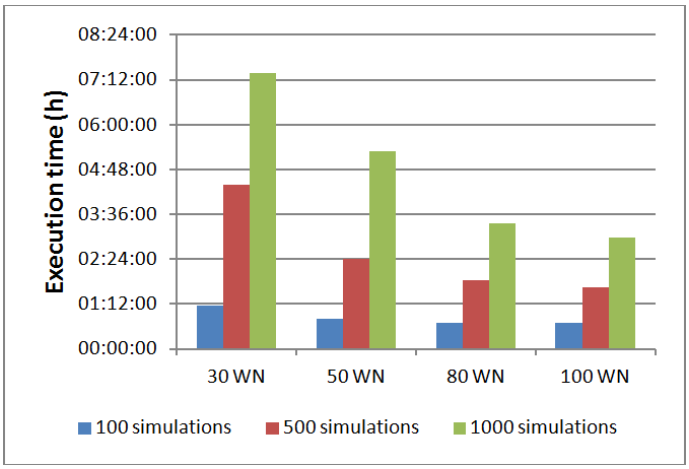


Fig. 9. Total execution time.

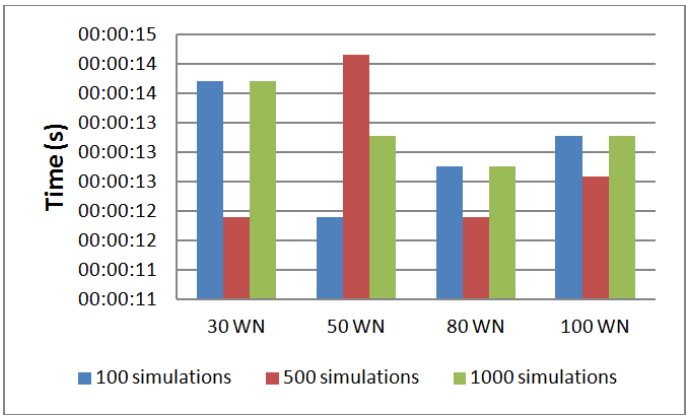


Fig. 11. Submission time for one job.

is not very high compared with the 25% additional resources that are needed. Figure 10 presents the execution time per simulation. In all cases (variation of the number of simulations) the execution time for one simulation decreases (keeping the same trend) when we use more computational resources.

Figure 11 presents the submission time, which is constant (around 13 seconds) and does not depend on the number of computational resources used or on the number of simulations that were executed. The submission process consists in all the steps performed by the gSWAT application before the execution of simulations can begin. Even though the submission time is constant the impact on the total execution time is different.

For the complex SWAT model we have executed 8 iteration steps, each iteration step requiring 200 simulations. Because of the complexity of the model we split the execution of each iteration step in 4 blocks of 50 simulations. The average execution time for one iteration step was around 170 hours, meaning a virtual execution time per simulation of around 50 minutes. The actual execution time for one simulation was around 40 hours. The increase of performance is in this case a significant one, execution of all the simulations on only one computer is impossible in this case in a reasonable amount of time. The execution times for each simulation are different but there are no significant differences regarding the total execution time (see Figure 12 where results from three iteration steps are presented). The minimum and maximum execution time for each simulation block varies mainly because of the availability of Grid resources.

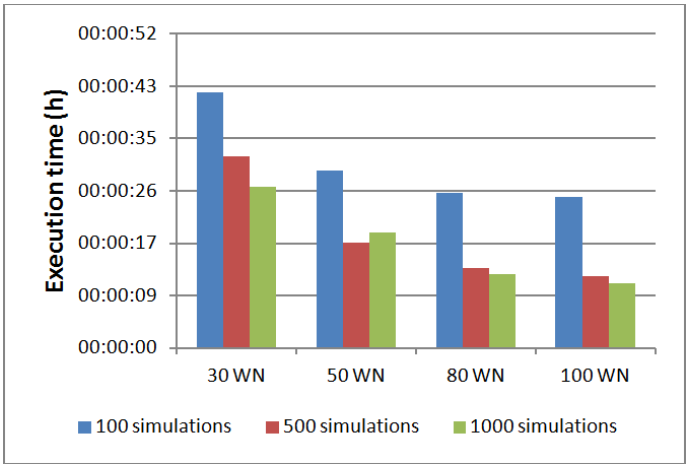


Fig. 10. Execution time per simulation.

For a complex model where the number of files is very high (more than 1.000.000 files) we have to start fewer jobs on the same physical machine. The execution of the simulation needs to read and write in this case many files, and if multiple jobs are executed on the same physical machine, they will make concurrently access to the hard-disk and the execution time will grow excessively. In some cases the execution of one or more simulations takes longer than the execution of the other ones (as is the case of the second simulation block for the second iteration steps presented in Figure 12). The availability of the Grid resources is the cause for this higher execution time but as can be seen the impact is not significant. This experiment proves that in this case (calibration of complex models) the Grid offers a very good solution, decreasing very much the time needed to execute all the simulations required by the calibration process.

VII. Conclusions

Complex SWAT hydrologic models are used to assess the sustainability and vulnerability of land management practices on water quality and quantity. The gSWAT application offers a flexible environment to calibrate SWAT models over distributed infrastructures such as Grid. The execution time could be minimized by running several simulations in parallel, on different WNs. In some cases (according to the number of simulations or the model complexity) the speedup obtained by increasing the number of computational resources is quite small. The experiments proved that the calibration process can benefit by the scalability offered by the Grid infrastructure.

Acknowledgment

This research is supported by the FP7 enviroGRIDS Project funded by the European Commission, through the Contract 226740.

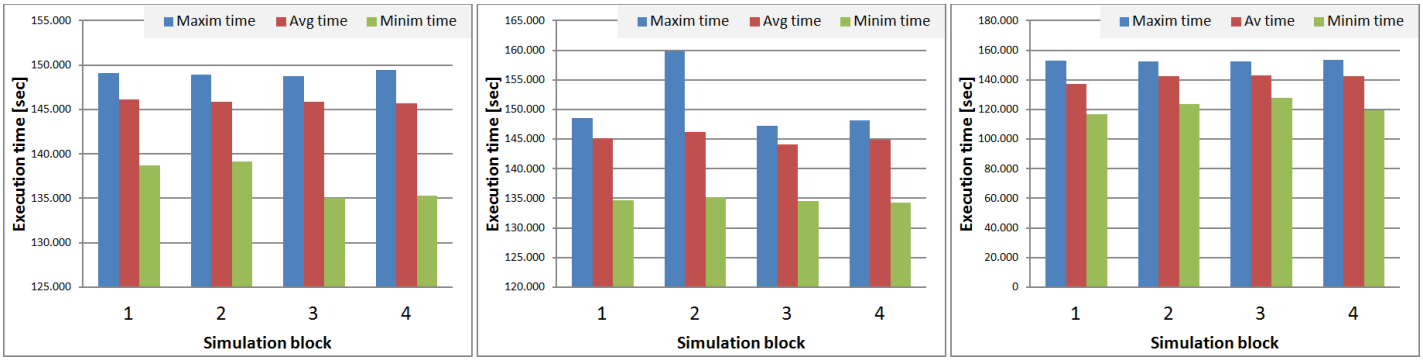


Fig. 12. Execution times for complex SWAT models (results for three iteration steps using 4 simulation blocks for each one).

REFERENCES

[1] enviroGRIDS project, http://www.envirogrids.net/

[2] SWAT: Soil and Water Assessment Tool, available at: http://swatmodel.tamu.edu/

[3] C. Moore, and J. Doherty, "Role of the calibration process in reducing model predictive error", Water Resour. Res., 41, 2005.

[4] I. G. Pechlivanidis, B. Jackson, N. McIntyre and H.S. Wheater, "Catchment scale hydrological modelling: A review of model types, calibration approaches and uncertainty analysis methods in the context of recent developments in technology and applications", Global NEST Journal, 13(3), pp. 193-214, 2011.

[5] P. C. Young, A. J. Jakeman and D. A. Post, "Recent advances in the data-based modelling and analysis of hydrological systems", Water Science and Technology, 36(1), pp. 99-116, 1997.

[6] C. W. Dawson, R. J. Abrahart, A. Y. Shamseldin and R. L. Wilby, "Flood estimation at ungauged sites using artificial neural networks", Journal of Hydrology, 319(1-4), pp. 391-409, 2006.

[7] K. C. Abbaspour, C. A. Johnson, and M. Th. van Genuchten, "Estimating Uncertain Flow and Transport Parameters Using a Sequential Uncertainty Fitting Procedure", J. Vadouse Zone, 3 (4), pp. 1340-1352, 2004.

[8] gLite, http://glite.cern.ch/

[9] J.T. Moscicki, F. Brochu, J. Ebke, U. Egede, J. Elmsheuser, K. Harrison, R.W.L. Jones, H.C. Lee, D. Liko, A. Maier, A. Muraru, G.N. Patrick, K. Pajchel, W. Reece, B.H. Samset, M.W. Slater, A. Soroko, C.L. Tan, D.C. van der Ster, M. Williams, "Ganga: A tool for computationaltask management and easy access to Grid resources", Computer Physics Communications, Volume 180, Issue 11, November 2009, pp. 2303-2316, ISSN 0010-4655, 2009.

[10] J.T. Moscicki, DIANE - Distributed Analysis Environment for GRID enabled Simulation and Analysis of Physics Data, Nuclear Science Symposium Conference Record, pp. 1617 - 1620, Vol.3, ISSN 1082-3654, 2003.

[11] L. Biro, V. Bacu, D. Rodila, L. Barabas, D. Gorgan, "Grid to cloud migration of scientific applications, using dynamically created cloud clusters", Intelligent Computer Communication and Processing (ICCP), 2012 IEEE International Conference on , pp.335,340, 2012.

[12] G. Lecca, M. Petitdidier, L. Hluchy, M. Ivanovic, N. Kussul, N. Ray, and V. Thieron, "Grid computing technology for hydrological applications", Journal of Hydrology, 403(1-2), pp. 186 199, 2011.

[13] F. Pop, "A fault tolerant decentralized scheduling in large scale distributed systems", Chapter in: Handbook of Research on P2P and Grid Systems for Service-Oriented Computing: Models, Methodologies, and Applications, edited by: Antonoupoulos, N., Exarchakos, G., Li, M., and Liotta, A.,Information Science Reference (IGI Global), pp. 566-588, 2010.

[14] B. Simion, C. Leordeanu, F. Pop, and V. Cristea, "A hybrid algorithm for scheduling workflow applications in grid environments (icpdp)", In Proceedings of the 2007 OTM confederated international conference on On the move to meaningful internet systems: CoopIS, DOA, ODBASE, GADA, and IS - Volume Part II, OTM07, pp. 13311348, Berlin, Heidelberg, 2007.

[15] E. Rouholahnejad, K. Abbaspour, M. Vejdani, R. Srinivasan, R. Schulin, and A. Lehmann, "A parallelization framework for calibration of hydrological models", Environ. Model. Softw., 31, pp. 2836, 2012.

[16] T. Li, G. Wang, J. Chen, and H. Wang, "Dynamic parallelization of hydrological model simulations", Environ. Model. Softw., 26(12), pp. 17361746, 2011.

[17] J. Neal, T. Fewtrell, P. Bates, and N. Wright, "A comparison of three parallelisation methods for 2d flood inundation models", Environ. Model. Softw., 25(4), pp. 398411, 2010.

[18] A. J. Kalyanapu, S. Shankar, E. R. Pardyjak, D. R. Judi, and S. J. Burian, "Assessment of gpu computational enhancement to a 2d flood model", Environ. Model. Softw., 26(8), pp. 10091016, 2011.

[19] B. Singh, E. R. Pardyjak, A. Norgren, and P. Willemsen, "Accelerating urban fast response lagrangian dispersion simulations using inexpensive graphics processor parallelism", Environ. Model. Softw., 26(6), pp. 739750, 2011.

[20] Y. Cui, B. E. Vieux, H. Neeman, and F. Moreda, "Parallelisation of a distributed hydrologic model", Int. J. Comput. Appl. Technol., 22(1), pp. 4252, 2005.

[21] V. Sharma, D. A. Swayne, D. Lam, and W. Schertzer, "Parallel shuffled complex evolution algorithm for calibration of hydrological models", In Proceedings of the 20th International Symposium on High-Performance Computing in an Advanced Collaborative Environment, HPCS 06, pages 30, Washington, DC, USA, 2006.

[22] D. Gorgan, V. Bacu, D. Mihon, T. Stefanut, , D. Rodila, P. Cau, K. Abbaspour, G. Giuliani, N. Ray, A. Lehmann, "Software platform interoperability throughout enviroGRIDS portal", Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS) , IEEE Journal of, 5(6), pp. 16171627, 2012.

[23] D. Gorgan, V. Bacu, D. Mihon, D. Rodila, K. Abbaspour, E. Rouholahnejad, "Grid based calibration of SWAT hydrological models", Natural Hazards and Earth System Science (NHESS), 12(7), pp. 24112423, 2012.

[24] D. Rodila, V. Bacu, D. Gorgan, "Comparative parallel execution of SWAT hydrological model on multicore and grid architectures", International Journal of Web and Grid Services (JWGS) 8, 3, pp. 304-320, 2012.

[25] V. Bacu, D. Gorgan, "Grid application oriented computational resource allocation strategy", High Performance Computing and Simulation (HPCS), 2012 International Conference on, pp.581-587, 2012.

[26] V. Bacu, D. Mihon, T. Stefanut, , D. Rodila, D. Gorgan, P. Cau, S. Manca, "Grid based services and tools for hydrological model processing and visualization", In Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2011 13th International Symposium on, pp. 291298, 2011.

[27] R. T. Fielding, "Architectural Styles and the Design of Network-Based Software Architectures", Ph.D. Dissertation, University of California, Irvine, 2000.

[28] S. Manca, C. Soru, P. Cau, G. Meloni, M. Fiori, A multi model and multiscale, GIS oriented Web framework based on the SWAT model to face issues of water and soil resource vulnerability. Presentation at the 5th International SWAT Conference, August 3-7, 2009

An Interoperable GIS Oriented Information and Support System for Water Resources Management

Pierluigi Cau
Center for Advance Research and Studies in Sardinia (CRS4), Pula (CA) Italy. plcau@crs4.it

Simone Manca
Center for Advance Research and Studies in Sardinia (CRS4), Pula (CA) Italy.

Costantino Soru,
Center for Advance Research and Studies in Sardinia (CRS4), Pula (CA) Italy.

Davide Muroni
Center for Advance Research and Studies in Sardinia (CRS4), Pula (CA) Italy.

Dorian Gorgan
Technical University of Cluj-Napoca, Cluj-Napoca, Romania

Victor Bacu
Technical University of Cluj-Napoca, Cluj-Napoca, Romania

Anthony Lehman
Institute for Environmental Sciences, enviroSPACE
University of Geneva
1227 Carouge, Switzerland

Nicolas Ray[1,2]
[1]Institute for Environmental Sciences, enviroSPACE
University of Geneva
1227 Carouge, Switzerland
[2]United Nations Environment Programme Global Resource Information Database 1211 Châtelaine, Switzerland

Gregory Giuliani[1,2]
[1]Institute for Environmental Sciences, enviroSPACE
University of Geneva
1227 Carouge, Switzerland
[2]United Nations Environment Programme
Global Resource Information Database
1211 Châtelaine, Switzerland

Abstract*—Important objectives of the four-year enviroGRIDS project encompass the improvement of transnational cooperation, the use of state of the art Information and Communication Technologies for data analysis and sharing and the application of environmental models for monitoring present and predicting future states of the environment for the Black Sea region. In such a transnational context, there is a dire need for the environmental sciences to evolve from a simple, local-scale vision toward a complex, multi-user, multilayered holistic approach. BASHYT (http://swat.crs4.it/) is a Web based, GIS oriented, information and support tool, part of the Black Sea Catchment – Observation System (BSC-OS). It exposes a set of applications for data management, analysis and visualization and a complete server and client side development framework (wiki like) to create Web contents. The core of the portal relies on the hydrological semi distributed SWAT code to model the water cycle and predict the effect of management decisions on water, sediment, nutrient and pesticide yields on large river basins. Furthermore, BASHYT aims at quantifying the interconnectedness between (human and natural) pressures and states of water body receptors at different space and time scales. The aim is to enhance environmental management capacity to assess water resource and to share and process large amounts of key environmental information. Within an experimental and innovative programming environment, modules have been developed to run near real-time applications based on numerical solvers (SWAT is just one example), run pre- and post-processing codes, query and map results through the Web browser. A set of web OGC services and a complete Application Programming Interface (API) are also exposed by the portal. We expect to improve the ways in which land management systems can operate and improve model usability to aid in making management decisions and watershed-scale modeling.

Keywords—enviroGRIDS; GIS; SWAT; DSS; Argilla; Black Sea Catchment; Mapserver; BASHYT; Hydrology; Interoperability; OGC; Portal

I. INTRODUCTION

Prediction, prevention, or minimization of point and diffuse pollution is an open issue for the Black Sea region. The Black Sea Catchment is going through an ecologically unsustainable development and inadequate resource management, which is leading to severe environmental, social and economic problems. The Black Sea ecosystems are endangered by eutrophication, pollution, and irresponsible exploitation of natural resources which resulted in an steadily decline of biological diversity in ecosystems and in a degradation of

landscapes. As a matter of fact point and diffuse pollution from priority sources such as oil spills, or insufficiently treated waters, mismanagement of agricultural lands needs decreasing. The complexity of water resources management in such a complex basin represents an increasing challenge to policy makers of the region, where an interdisciplinary approach is needed to design effective management strategies.

By one side, the use of ICT, such as High Performance Computing (HPC) Infrastructures, Geographical Information System (GIS), numerical models, and web-based applications involves major investments in terms of acquisition of quality data and the development of an interdisciplinary approach to the study. By the other side, such technologies can provide a significant contribution in the description of environmental dynamics, simplifying the management, access, share, and analysis of data and providing efficient report production mechanisms. Web portals [1, 2] are becoming strategic gateways where scientists, citizens, stakeholders, and end users can securely use applications, storage and computational infrastructures and services. Analysis and management tools for the environmental sciences need evolving from a local and single-user oriented approach toward a complex, multi-user, and multilayered global vision.

The experience gathered from many EU initiatives, such as, CLIMB (http://www.climb-fp7.eu/) or DRIHM, (http://www.drihm.eu), that ultimately aim at contributing to Global Earth Observation System of Systems (GEOSS) and Copernicus goals, highlights the need of increasing the interoperability abilities for the sharing of information and knowledge between data repositories and service providers from different sources across Europe. So far, many open standards and interoperability services are being considered, such as Web Map Service (WMS) and Web Feature Service (WFS) proposed by the Open Geospatial Consortium (OGC - http://www.opengeospatial.org/), although their use is still limited. Management and analysis of very large and growing volumes of geo-data is challenging the scientific community without clear long term solutions. The main open issues, tightly bound to technological development are: scalability and flexibility of the application level; web data accessibility and security; limitation use of web services. Important objectives of the four-year enviroGRIDS project (http://envirogrids.net/) include the enhancement of transnational cooperation, the use of web-based technologies for data analysis and sharing and the application of environmental models for monitoring present and predicting future states of the environment for the Black Sea region. One particular objective is to contribute to the achievement of the goals of the intergovernmental Group on Earth Observations (GEO), that is leading a worldwide effort to build a Global Earth Observation System of Systems (GEOSS). Two specific goals are recognized to be of paramount importance: to raise awareness of Societal Benefit Issues of the general public and to build regional capacities on Earth Observations and INSPIRE standards and approaches. Such needs have been addressed by developing a Black Sea catchment Observation System (BSC-OS) [3], that integrates several web-based information technologies to exploit complex models, quality data and the EnviroGRIDS storage and Spatial Data Infrastructure (SDI). The BSC-OS portal is composed of a set of loosely coupled components that aim at addressing specific needs: data and catalogs management is provided by the URM portal; model calibration and execution of scenarios is accomplished by gSWAT; data visualization, scenarios development, and report generation is based on BASHYT; eGLE provides a web-based access to training material and lesson execution.

BASHYT is one important effort to develop and promote a innovative environmental management system particularly targeting observational (e.g., agricultural droughts and water quality measurements) and technological gaps for the water domain. During the EnviroGRIDS project, several hydrological models at the catchment's scale and one large implementation for the whole Black Sea Basin have been set up using the SWAT numerical code. The evaluation of historical changes to support environmental monitoring and reporting has been also carried out, leading to the evaluation of the impact on water resources as a result of natural and/or man-made change. BASHYT aims at using such models to assess the state of the water basin and to identify the reasons affecting the conditions. Furthermore, it aims at fostering integration of expertise from various fields to create a lively system where end-users and scientists can cooperatively work and create applications to assess water quality and quantity status.

In this work, BASHYT is examined, providing a detailed description of the architecture and technologies used, and how it has been applied to real case studies.

II. Description of the Framework

BASHYT is web-based software, which relies on environmental models and state of the art Information and Communication Technologies (ICT) to support decision makers in the field of sustainable water resources management. Initially it was designed to expose hydrological applications based on the Soil and Water Assessment Tool (SWAT) [4]. Currently also other models have been deployed on the system such as the General Estuarine Transport Model / General Ocean Turbulence Model (GETM/GOTM) [5, 6] and the OilSpill Module [7]. In this paper, we will describe the information system particularly focusing on the hydrological applications.

BASHYT is a web-based operational tool to share SWAT model applications on the web and to standardize as much as possible the report production mechanism. The system, for the general user, exposes analysis based on the semi distributed "physically based" hydrological SWAT model and on geo-processing tools that make use of large volumes of geographical data. Free software and in-house technologies are combined to transparently and automatically access to and process SWAT data repositories, run and manage the model, and expose web-based user-friendly environmental applications. The system does not require additional software or plugins, but works directly on any web browser, improving the potential for its utilization by water management administrations, being programmable and assessable directly on the Internet/Intranet through any WEB browser (Internet Explorer, Mozilla Firefox, Opera, ecc.).

A. SWAT – the Soil and Water Assessment Tool

SWAT as described by Neitsch et al. (2005) [8] is a watershed-scale hydrological model, developed by the U.S. Department of Agriculture USDA-ARS and Texas A & M University, which allows to simulate the integrated water cycle and to assess the impact of point and diffuse pollution in the medium/long term. The model has been tested successfully worldwide and is supported and farther developed by a very active community [9]. Its application requires specific information on weather, soil characteristics, topography, vegetation and land use. It is computationally efficient and uses readily available inputs, enabling users to study long-term impacts. The model works on two levels: land and routing phase.

Hydrological processes are first simulated in SWAT for the land phase at the HRU spatial unit. HRUs are Hydrologic Response Unit that represent areas with a unique combination of land cover, soil type and management practice. This yields the water, sediment, nutrient, and pesticide loadings to the main channel in each subbasin. The division of a watershed enables the model to reflect differences in evapotranspiration, runoff, movement and transformation of chemicals, etc., for various crops and soils. This improves accuracy of model predictions and gives a much better physical description of the catchment's water balance and water quality.

In the second phase, the water, sediments, nutrients, etc. are routed through the channel network of the watershed to the outlet. The system incorporates a variety of physical, chemical, and biological (Nitrogen, Phosphorus, Pesticide and sediment fate) processes that control the transport and transformation of pollutants within the water body. The water quality module of the SWAT model, based on Qualk 2E as described by Brown et al. (1987) [10], is driven by hydrodynamics, point and non-point source loadings, and key environmental forcing functions, such as temperature, precipitation, solar radiation, wind speed, and light attenuation coefficients.

For each subbasin there is one reach, one outlet, and many HRUs. Water quality and quantity state variables are computed for each subbasin at a daily temporal scale. BASHYT interconnects directly to the ARCGIS SWAT [11] and AV SWATX [12] desktop processing software. Such desktop tools are particularly aimed at creating the input files for the SWAT model. A possible data workflow for the system consists of the following phases:

1) Users upload SWAT project (input / output data) to BASHYT. Earth scientists create their model on their local resources, exploiting the GIS functionalities of the desktop AvSWAT or ArcSWAT programs. When the project has been uploaded, users can utilize the BASHYT web interface to analyze each simulation, to compare simulations or to run new scenarios.

2) Once the project is loaded to BASHYT, it can be calibrated over the enviroGRIDS infrastructure exploiting gSWAT [13, 14]. After calibration it can be loaded back to BASHYT. Calibration is a highly computational consuming process and desktop programs may not be efficient.

3) Scenarios can be run on BASHYT exploiting transparently the computational and storage resources granted by the enviroGRIDS SDI: scenarios execution can be also a highly computational consuming processes as well as time consuming.

The gSWAT system allows the calibration of SWAT models in a flexible manner. gSWAT is built as a distributed system composed of a graphical user interface and a service related component. The BASHYT framework enables earth scientists to exploit transparently the whole EnviroGRIDS computational and data storage resources as well as the components of the BSC-OS portal. Interoperability standards and a single sign on authentication mechanism are used to let users have a single entry point.

B. The information and support system

BASHYT is based on the Driving forces-Pressures-State-(DPS) paradigm, introduced by the European Environmental Agency (similar to the PSR model developed by the Organization for Economic Co-operation and Development - OECD [15]) and also adopted in the EU Water Framework Directive (WFD). Whenever a new SWAT project/simulation is loaded to the BASHYT interface, data are automatically digested by the software and a database is created from a fixed schema. The following sections will expose geodata and simulations on the web through thematic applications:

- Driving forces (D) and Pressures (P): this section exposes two main categories where D and P are grouped in: point and diffuse pollution;
- State of the environment (S): water balance and water quality states are analyzed. Results are viewed on various space and temporal scales (e.g., monthly, yearly, subbasin, basin).

The DPS methodological approach is useful to demonstrate the interconnectedness and estimate the effectiveness of the actions aimed (responses) at solving problems at hand. Driving forces stand for processes underlying to environmental changes such as land use or demographic development. Pressure indicators measure the level of environmental impairment (e.g., total quantity of phosphorous in chemical or biological fertilizers applied per hectare of agricultural land). State indicators are the conditions of the environment (e.g., average concentration of phosphorous in surface waters). States have impacts on different receptors causing damages such as loss of biodiversity, eutrophication of surface water or water becoming unsuitable for drinking. Responses (policies and measures) can be manually designed to solve problems and then simulated within BASHYT.

The environmental applications are exposed through a "user friendly" interface, which supports a coherent management of the Driver, Pressure, State indicators as distributed (in space and time) catchment's variables. This approach encourages the user to increase the awareness of the effects of subjective judgments or misjudgments on the final result. Through its thematic sections, users are guided to analyze pressures on the environment from natural (e.g., climate) and anthropogenic (e.g., land use) sources and improve the understanding of the

complex watershed system. The DPS "Conceptual Model" represents the causal links between current human activities (D), the pressures they exert (P), and the state (S) of the environment. Responses to the problems at hand or more generally Scenarios of interest can be run on BASHYT and compared to any other scenarios. Water managers can design management strategies to solve environmental problems and evaluate the performance of the choices through the "compare scenario section". The SWAT code is employed to evaluate the performance of the response, on the basis of the chosen indicators. BASHYT reads the SWAT input/output (IO) and dynamically produce standardized reports, making the analysis of the complex SWAT IO considerably easier.

The GIS analysis and visualization tools help identifying critical areas (e.g., the major contributors to nutrient losses or affected by desertification processes) and prioritize critical sub-areas in order to develop a multi-year management analysis. This analysis can be essential, for instance, to reduce the nutrient impact from point and non-point source pollution to downstream water bodies or to design and evaluate sub-regional and regional remediation strategies through the DPS conceptual framework.

BASHYT also exposes an interface to run the SWAT model directly from the Web, where climate change scenarios can be created in an easier way. In the back end, server side procedures:

- Process climate data to produce new input file
- Run the model on the EnviroGRIDS infrastructure (gSWAT) or in the "in house" HPC environment
- Post process results to be viewed in BASHYT

We aim at improving model usability at all levels to aid in making management decisions and watershed-scale modeling.

III. Architecture and Technologies

BASHYT comprises computing and storage resources, data, and a complex software system composed of a relational database management system, visualization software, numerical applications and geo-processing tools. Interoperability is of paramount importance because, in such a system, heterogeneous data sources (databases and filesystem), computing resources and services are shared among the different components of the BSC-OS Portal. Users, logged on BASHYT, transparently access to the shared physical resources, such as a HPC cluster environment to submit jobs, and use catalogs of Geo data or applications found in different domains and organizations. Within BASHYT, the in-house HPC infrastructure is exploited to run the SWAT model as well as all geo-processing phases required by the reporting production mechanism and in particular by the hydrological applications. The data flows, the data storage, and the application workload have been designed in a non-conventional fashion to hide the user the complexity of the infrastructure. The system is more than a simple sum of modules: it is a software to consume and expose Web services for data mapping, querying and sharing, processing and distributing, with a high degree of freedom.

One important component of the system is the Argilla engine [16], a Java development framework to construct web pages and applications. This can be thought of as an open interoperable, and extensible development framework to build spatially enabled web-based applications. It is based on the Model View Controller (MVC) architecture. The MVC architectural pattern has been used to isolate business logic from input and presentation, enabling, for each component, independent development, testing and maintenance [17, 18]. Our MVC software implements the web template system, which is a fast and flexible processing system for web content management and application development, making the programming features available to developers with almost-zero learning curve. This increases developer productivity by reducing scaffolding code when developing web, GUI, database/GIS or any web-based application.

A. The data infrastructure

We have designed a new prototype of a distributed spatial infrastructure based on SpatiaLite (http://www.gaia-gis.it/), particularly useful for large distributed data-intensive applications based on the SWAT model. While sharing many of the same goals as other distributed systems, our design has been driven by observation of our application workload and technological environment, both present and expected. This has led us to reconsider the traditional choice of one comprehensive PostGIS database (which still keeps its validity when dealing with a limited number of watersheds) and explore radically different design points. PostgreSQL/PostGIS system could not be flexible enough to meet the requirements of scalability in a regional/continental context where virtually hundreds and even thousands of basins needed to be simulated.

Given the amount of spatial data required for cases such as the Black Sea watershed scale model, we decided to experiment a solution based on the SQLite technology with spatial extension (SpatiaLite). SQLite is an embedded database engine distributed as a common library; it is widely used on many popular applications like Mozilla Firefox, Apple Mac OS X, Google Apps and many more. Spatialite provides a large set of spatial functions and data structures like what PostGIS does for PostgreSQL.

Inputs and outputs of the SWAT model are stored in SpatiaLite database files. Each DB file contains one model set up (one watershed and many simulations) that is accessed by BASHYT when the user activates the watershed in the portal. This choice guarantees good performance when a large number of simulations/watersheds are accessed from several users at the same time. Within the EnviroGRIDS project, we have set up a test environment of 5 servers with 8 cores each. The application can be scaled and enlarged on a dedicated computing/storage environment (typically using virtualization mechanism) to meet user workload. Each node of the system contains many SpatiaLite DB files that are controlled by dedicated instances of the application framework. BASHYT acts as a workflow manager posting requests and getting results. In this configuration the computing and storing tasks are resolved outside the web framework.

The nature of the SQLite engine (without dependencies) assures high scalable scenarios, since all operations work as in

common read/write filesystem. Using this approach, we are gaining the power of a complete transactional RDBMS, without the need of external server process to query and with a useful portability freedom. SQLite offers the capability to load personal or third party extensions (shared libraries), written in C or other languages. This mechanism can be used to straighten the SQL functionalities of the engine or override its functions. This structure, combined with the absence of a dedicated DBMS process, reduces lags and resources needed by network communications. BASHYT commands also the map and graph rendering and other applications for the report production mechanism. The application includes the libsqlite library to manage the whole database repository. When one (or more) SWAT simulation is uploaded or executed in the server, an ETL (Extract, Transform, Load) procedure is run to format the model input/output (IO) and import it into one or more SQLite DB file (internal flow). The SQLite architecture does not impose restrictions on distribution, size or number of files. Each DB file is completely independent from the others. The main issue to consider when using SQLite is its strict dependence on the filesystem. SQLite inherit any fault coming from the layer below without chance for recovery. Distributed network filesystems suffer often from file locking bugs. In general this can cause SQLite data corruption or inconsistency in high traffic volume contexts. In SQLite, one reading operation locks all write requests on files and vice versa. In high concurrency conditions, when read/write actions alternate themselves with high-frequency, this could represent a performance bottleneck. Although our system aims at working in high volume data and traffic situations, the above issues are minor, because end-user operations are read only operations. As a matter of fact all write operations are done by the ETL procedure to import SWAT IO. During this task, the simulation is not available to users for reading.

The SQLite and its SpatiaLite extension engine have been carefully and positively tested in real situations with a large number of competitive access to the web environment. On one hand, this technology can be considered still young and does not have the reliability level or spatial functions of other engines like PostGIS. On the other hand for a limited controlled use, SpatiaLite meets our needs, although some changes on JDBC SQLite driver for Java were needed to let it work on our distributed system.

B. The GIS visualization

The GIS rendering is optimized using the Open Source MapServer (http://mapserver.org/) technology. This is accomplished, exploiting the scripting languages capabilities to access the MapServer CGI and OGC (WMS, WFS) interfaces. MapServer works as a map engine providing a spatial context where it is required. On the client side, AJAX (web 2.0) technologies, such as the msCross [19] cross-browser interface, is customized to allow users dynamically display and browse the geographical information layers. Our system inherits all the Geographic Information System (GIS) capabilities granted by these technologies. The system aspires to become desktop like for the geospatial data management and analysis, image processing, graphics/maps, spatial modeling, and visualization productions. Complex spatial and alphanumeric query capabilities have been implemented to meet requirements and specifications of the SWAT data structure. GIS functionalities have been also developed from scratch and/or adapted to serve sophisticated applications to query and analyze spatial data produced by the models. In this way users can easily display on maps complex analysis and queries. It is possible, for instance, to execute spatial queries on any simulation map and get a report on the different SWAT output ready to use.

BASHYT supports a multitude of raster and vector data formats (e.g., ESRI Shapefiles, PostGIS, Oracle Spatial, MySQL, OGC web specifications WMS and WFS) exploiting the functionalities of the Geospatial Data Abstraction Library (GDAL - http://www.gdal.org/) and the OGR Simple Feature Library (http://www.gdal.org/ogr/).

C. Interoperability and the web development environment

BASHYT exposes a fully programmable environment accessible directly from the web, wiki like, and an Application Programming Interface (API), that specifies how software components interact with each other (http://swat.crs4.it/Documentation/).

The API we developed is a particular set of rules and specifications that an external software program can follow to access and make use of the services and resources provided by BASHYT. The API defines the "vocabulary" and resources request conventions (e.g., function: getFile()). It includes general specifications for data structures, object classes, and protocols that are to be used to communicate with the framework. The API enables not just to access data but also allows writing and creating new contents exploiting the server side report production mechanism of BASHYT.

The API offers a uniform way of identifying and accessing resources, and thus increasing the interoperability between applications. Web applications are mostly data-driven, and it is easy predictable that they will benefit from the increasing interoperability of our framework. Other web applications of the BSC-OS portal, such as eGLE (http://cgis.utcluj.ro/egle-demo/), exploit our system and its capabilities to merge in new ways information, model outputs, or simply territorial data. The eGLE e-Learning environment is used to support the development and the execution of lessons in Earth Observation domain.

In addition, BASHYT exposes a fast and flexible processing system on the Web (Wiki like) for web content management and application development. On the web, client and server side code can be edited to create complex web applications. Earth scientists, through a dedicated web editor, write their own GUI's and applications. The development process (e.g., layout, connection and query to db, etc.) can be controlled on the fly by switching from edit to view mode. No compilation is required: this increases development or maintenance productivity of web based applications. When ready and validated, applications can be made public by the administrator. Hydrologists, scientists, web designers, and developers are asked to concentrate on generating web contents without getting bogged down in programming matters, making the whole process of developing, updating and maintaining portals significantly easier.

BASHYT development capabilities enable to write services merging server side and client side codes within a uniform web based interface. Dedicated sections of the development framework exposes modules that enable to shape XML objects for the production of graphs, maps, tables, PDF reports, and forms. These modules permit the massive use of preset schemas stored in the database (virtual file system) in a structured form (XML). Each object refers to its schema and describes parameters (e.g., to control layout) and data sources. The development framework exposes GUIs to produce in a easier manner these objects.

IV. The Black Sea Case Study

The Black Sea is suffering from poor water resource management, partially due to the lack of effective transnational cooperation, limited data sharing and the lack of scientific tools to bring together scientists, administrations, social partners, and environment protection agencies. Exchange of information, sharing of good practices, and working together towards common solutions in a multicultural environment are still open challenges. BASHYT aims at contributing to some of these issues by bringing several new emerging information technologies to build a data-driven vision of the planet that is feeding into models and scenarios to explore the past, the present and the future of the Earth and, in the specific context of the enviroGRIDS project, particularly targeting the Black Sea regions. A double objective achieved during the project is the set up of a complex modeling system for inland water analysis and protection exposed on the web by the BSC-OS portal. Within BASHYT, applications are grouped in different thematic sections:

- Data Manager: user can choose which model and watershed to analyze.
- SWAT: users access to the "Watershed" and "Scenarios" sections.

Watersheds in different regions around the word have been deployed on BASHYT, as shown on figure 1, to obtain the needed information and analysis.

In the "SWAT" section, the output of the model is used to produce reports for the watershed organized in the DPS structure, while in the "Basin" section, a physical description of the territory is produced, taking into account topography, land use, soil and climate. In figure 2, the Digital Elevation model for the Black Sea is shown. Reports are based on the SWAT IO data and default parameterization datasets. In figure 3, under current land use and climate condition, the BASHYT exposes the distribution of the water balance components at the subbasin spatial scale for the Black Sea watershed. Over a 38-year simulation period (1970-2008) the main hydrological components assessed on a monthly basis reads as follow (figure 4): average (standard deviation) precipitation is 59.02 mm (16.14). Average evapotranspiration (standard deviation) is 33.9 (20.35). Average water yield (standard deviation) is 23.89 (6.84). Within the same environment, users can assess analysis for the other watersheds in the same standardized fashion.

The components of the hydrological balance as well as the other SWAT output variables, computed on a daily time step for each HRU, subbasin or river reach are integrated and assessed by BASHYT back-end procedures to expose on the web spatial and temporal analysis (outputs are presented with time series graph, tables and spatial representations by means of dedicated interactive web GIS within the portal).

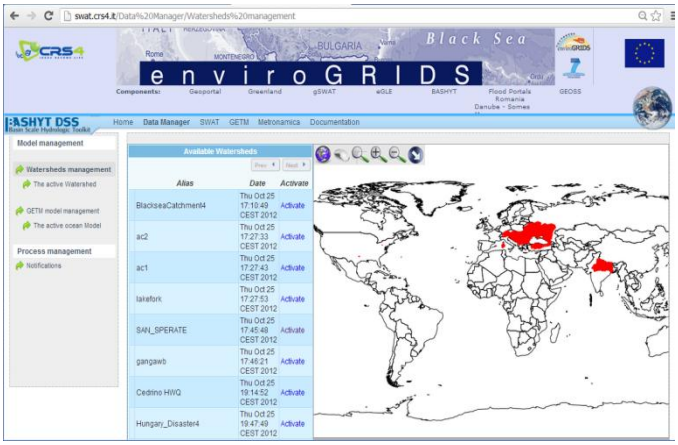


Fig. 1. The data Manager section. The interface shows all watershed that have been deployed on BASHYT. The Black Sea watershed is clearly visible in the center.

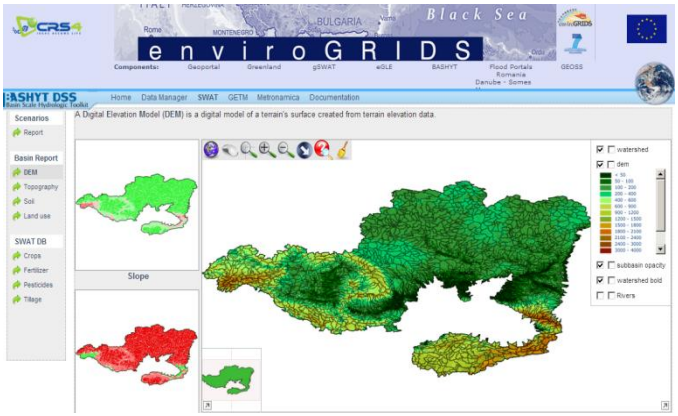


Fig. 2. Digital elevation model viewed on the BASHYT interface for the Black Sea watershed..

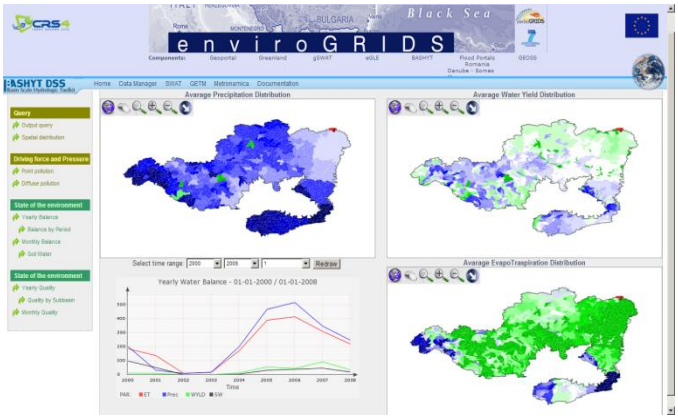


Fig. 3. Water Balance. The water balance is mapped on the Web GIS. Automatic procedures read the SWAT results and produce reports in the form of maps, charts or tables.

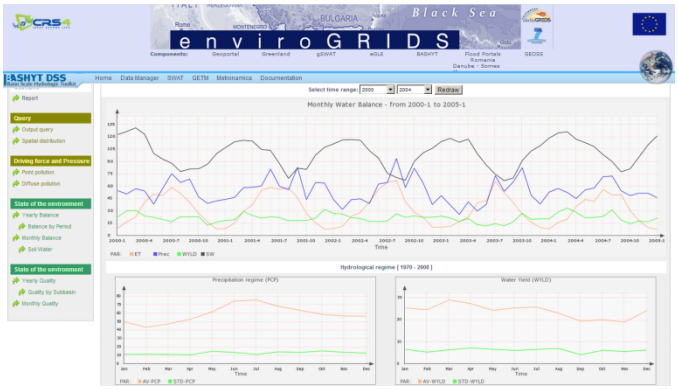


Fig. 4. Water balance for the whole Black Sea basin. Water balance is computed at the subbasin scale and daily time step. Processing algoritm in the back-end integrates these values to assess watershed avarage values.

A methodology to evaluate agricultural drought conditions, applied also to the Black Sea basin, has been set up during the project. Drought is a temporary condition of relative scarcity of water resource compared to values that can be considered normal for a period of time and on a region [20]. We may distinguish between meteorological, agricultural, hydrological and operational drought [21, 22, 23]. While the meteorological drought is identified on the basis of a deficit of precipitation, the agricultural drought depends on the soil moisture deficit, which is dependent on many factors such as the precipitation regime and weather, the soil characteristics and the evapotranspiration rate. The persistence of agricultural drought condition produces negative effects both on natural vegetation and agriculture. Drought periods have an important impact on water supply system causing water shortage, negatively affecting the economic and social system.

The Soil Moistures Deficit (SMD) agricultural drought index implemented in BASHYT is a variation of the approach proposed by Narasimhan [24]. SMD is calculated on a monthly basis as proposed in the formula (1) and at the subbasin spatial scale. For the given month the index expresses the ratio between the anomaly of the monthly value compared to the average multi-annual data, and the difference between the maximum and minimum values for the entire time series available (for the Black Sea: 1970-2008).

The SMD index reads as follows:

$$SMD_i = \frac{SW_i - SW_i^{mean}}{SW_i^{max} - SW_i^{min}} \qquad (1)$$

where SMDi is the deficit of soil water content of months i, SW_i the monthly average soil water content of month i, SW_i^{mean} the long-term average of the soil water content of month i, SW_i^{min} and SW_i^{max} respectively the minimum and the maximum soil water content of month i for the entire simulation.

The index can be positive or negative, signifying for a given month a surplus and a deficit of water content respectively for a given soil. BASHYT automatically quantifies the anomaly magnitude of the SMD drought index, mediated on each month and on a subbasin spatial scale. In figure 5, for July 2002, the spatial distribution of the monthly SMD index is provided. Yellow/orange colors represent area under water stress while green colors show high water content values.

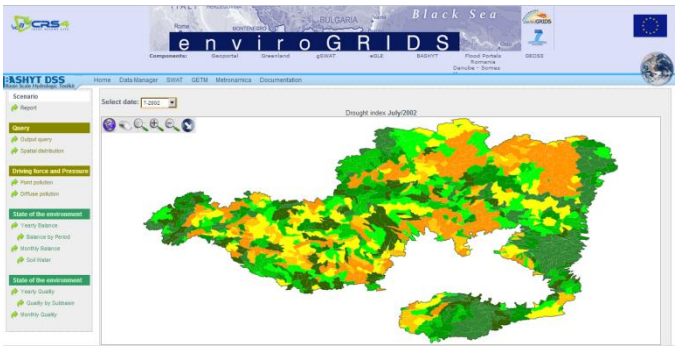


Fig. 5. Spatial distribution of the monthly SMD index (July 2002). BASHYT automatically quantify the SMD drought index. Yellow/orange areas show drought conditions while green colors show high water content values.

The correct characterization of the spatial and temporal distribution of rainfall, of the land use, soil and anthropogenic pressures are strategic to represent the complex dynamic of surface and ground water resources and to design its sustainable use.

V. Conclusions and Future Work

The BASHYT provides a framework for analyzing management scenarios based on valuable data and computing resources over the Web. The system is based on client/server architecture and can be used within the Internet/Intranet cyberspace, offering to the community services to extract meaningful information about the environment. The DPS conceptual model can be used as a base for environmental management allowing the linkage between pressures and state-indicators. The application of this causality model and the use of GIS capabilities in combination with the SWAT hydrological model have the advantage of allowing the spatial visualization and complex analysis and better integration/exploitation of the different indicators on which water and territorial management is based.

In general, the web interoperability is of paramount importance to control the redundancy of replicated datasets, and it allows users to retrieve updated certified information, avoiding the latency due to administrative and technological barriers.

As a matter of fact, environmental analysis will benefit from near real-time data processing, making territorial management and planning more efficient. The BASHYT system can contribute to the development and the exchange of information relative to the environment, offering administrations standardized procedures to manage, control and study water resources. The development of a web-based framework such as BASHYT offers an infrastructure for optimizing data-sharing and solving application development problems in a multi-user environment. It improves model(s) usability by simplifying data I/O flow management and application development to aid in making management decisions.

The current version of the software has been used to expose on the web various SWAT model implementations for different regions of the world. The Black Sea case study is just one example. The system address the subjects related to SWAT data archiving, distribution and interpretation on a web-based environment through the use of interoperability standards and automated procedures. Improved capabilities for coordinating, accessing, sharing, and using environmental and geo-data have been also implemented in the system. Building web applications cooperatively through the web development environment contributes to the creation of enlarged multi-cultural working groups to improve public consciousness for environmental problems and strategic remediation strategies on a transnational scale.

ACKNOWLEDGMENT

This work was co-sponsored by the Regione Autonoma della Sardegna, Italy (RAS -. http://www.regione.sardegna.it/), the "Argilla Group" project funded by the LR 37/98, the FP VII CLIMB Project (http://www.climb-fp7.eu/) and the FP VII EnviroGRIDS Project (http://envirogrids.net/).

REFERENCES

[1] Berners-Lee, T.: Weaving the Web – The Past, Present and Future of the World Wide Web by its Inventor. Texere (2000)

[2] Berners-Lee, T., Hall, W., Hendler, J.A.: A Framework for Web Science (Foundations and Trends(R) in Web Science). Now Publishers Inc., (2006)

[3] D. Gorgan, P. Cau, K. Charvat, D. Rodila, V. Bacu, A. Jonoski, A. Van Griensven, P. Horak, K. Abbaspour, S. Manca, G. Giuliani, N. Ray, A. Lehmann, 2010. Requirements and specifications for the development of BSC-OS Portal. Technical report (enviroGRIDS_D61) enviroGRIDS – FP7 European project.

[4] Neitsch, S. l., J. G. Arnold, J. R. Kiniry and J. R. Williams. 2005. Soil and Water Assessment Tool. Theoretical documentation. Version 2005. Blackland research center - Texas agricultural experiment station. Grassland, soil and water research laboratory - USDA agricultural research service.

[5] H.Burchard, K.Bolding, L.Umlauf. GETM, source code and test case documentation. Version pre 1.8, http://getm.eu.

[6] L.Umlauf,H.Burchard,K.Bolding.GOTM, sourcecode and test case documentation. Version 4.0. , http://www.gotm.net.

[7] A. Vargiu, E. Peneva, S. Manca, M. G. Mulas, F. Murgia, M. Pintus, C. Soru,R. Biella, P. Cau, 2010. A web based interface for coastal zones modelling: a test case for the Orosei Gulf in Sardinia (Italy). Procedings of the Phisics and Estuaries and Coastal Seas (PECS) Conference, Sri Lanka – Colombo, 09-2010.

[8] S.L. Neitsch, J.G. Arnold, J.R. Kiniry, R. Srinivasan, J.R. Williams, "Soil and Water Assessment Tool, User's Manual". Published 2002 by Texas Water Resources Institute, College Station, Texas

[9] Jayakrishnan, R., R. Srinivasan, C. Santhi, and J.G. Arnold. 2005. Advances in the application of the SWAT model for water resources management. Hydrol. Process. 19(3):749-762.

[10] Brown, L.C. and T.O. Barnwell, Jr.. 1987. The enhanced water quality models QUAL2E and QUAL2E-UNCAS documentation and user manual. EPA document. EPA/600/3-87/007. USEPA, Athens, GA.

[11] Francisco O., V. Milver, R. Srinivasan, and C. Janghwoan, ARCGIS-SWAT: a geodata model and GIS interface for SWAT. 2006. Journal of the American Water Resources Association.

[12] Di Luzio, M., R. Srinivasan, and J. G. Arnold. 2002. Integration of watershed tools and SWAT model into BASINS. J. of the American Water Resources Association. 38(4):1127-1141.

[13] Mihon D., Stefanut T., Bacu V., Rodila D., Abbaspour K., Kokoszkiewicz L., Rouholahnejad E., van Griensven A., and Gorgan D., Grid Based Hydrologic Model Calibration and Execution. CSCS-18, International Conference on Control Systems and Computer Science, Bucharest, Romania, 24-27 May, 2011.

[14] Bacu V., Mihon D., Rodila D., Stefanut T., Gorgan D., gSWAT Platform for Grid based Hydrological Model Calibration and Execution. ISPDC 2011 Conference, Cluj-Napoca, July 2011.

[15] OECD. OECD –Environmental Indicators- Development, Measurement and Use. Organization for Economic Co-operation and Development. http://www.oecd.org/dataoecd/7/47/24993546.pdf.

[16] P. Cau, G. C. Meloni, S. Manca, C. Soru, D. Muroni. A Java Based Framework Optimized for Scientific Modeling and Analysis, International MultiConference of Engineers and Computer Scientists 2011 Vol. 1 (IMECS), Hong Kong (2011), ISBN: 978-988-18210-3-4.

[17] Burbeck, S.: Applications Programming in Smalltalk-80: How to use Model–View–Controller (1987)

[18] Reenskaug, T.: Models, views, controllers. Tech. rep., Xerox PARC (1979)

[19] S. Manca, P. Cau, E. Bonomi, A. Mazzella. 2006. The Datacrossing DSS: a Data-GRID Based Decision Support System for Groundwater Management. Second IEEE International Conference on e-Science, Amsterdam, December 4-6. ISBN: 0-7695-2734-5.

[20] Rossi G.. 2000. Drought Mitigation Measures: a Comprehensive Framework. In Vogt J.V. e Somma F. (eds.), Drought and Drought Mitigation, Kluwer Academic Publishers, pp. 233-246.

[21] Palmer, W.C., (1965). Meteorological Drought. Research Paper No. 45, U.S. Department of Commerce Weather Bureau, Washington, D.C., 58 pp.

[22] Gibbs, W.J., Maher J.V., (1967). Rainfull deciles as drought indicators. Bureau of Meteorology Bulletin No. 48, Commonwealth of Australia, Melbourne.

[23] McKee, T.B., N.J. Doesken and J. Kleist. 1993. The relationship of drought frequency and duration to time scales. Preprints, 8th Conference on Applied Climatology, 17-22 January, Anaheim, CA, Amer. Meteor. Soc., 179-184.

[24] Narasimhan, B., and R. Srinivasan. 2002. Development of a Soil Moisture Index for Agricultural Drought Monitoring Using a Hydrologic Model (SWAT), GIS and Remote Sensing. Texas Water Monitoring Congress. September 9-11, 2002. Austin, TX.

Web Based Access to Water Related Data Using OGC WaterML 2.0

Adrian Almoradie
Dept. of Integrated Water Systems and Governance
UNESCO-IHE Institute for Water Education
Delft, the Netherlands
a.almoradie@unesco-ihe.org

Andreja Jonoski
Dept. of Integrated Water Systems and Governance
UNESCO-IHE Institute for Water Education
Delft, the Netherlands
a.jonoski@unesco-ihe.org

Ioana Popescu
Dept. of Integrated Water Systems and Governance
UNESCO-IHE Institute for Water Education
Delft, the Netherlands
i.popescu@unesco-ihe.org

Dimitri Solomatine
Dept. of Integrated Water Systems and Governance
UNESCO-IHE Institute for Water Education
Delft, the Netherlands
d.solomatine@unesco-ihe.org

***Abstract*—As the Open Geospatial Consortium (OGC) has adopted WaterML 2.0 as encoding standard for representing hydro-meteorological time series data, the water community is in need of tools and methods for delivering such data over the web.**

This article presents experiences with one approach for publishing water-related data over the web based on the OGC WaterML 2.0-GeoServer framework and methods developed by Australia's Commonwealth Scientific and Industrial Research Organisation (CSIRO).

This implementation is one component of a web based flood information system for Somes Mare basin in Romania, which has been developed within the enviroGRIDS EU FP7 research project.

Keywords—WaterML 2.0; GeoServer; Flood information dissemination; web-based; Somes Mare

I. INTRODUCTION

Environmental disasters such as floods are expected to increase because of rapid urban development, population growth and climate change. Such example is the flood-related disaster that happened in Pakistan in the year 2010. The flood affected over twenty million people, more than the combined major disaster of the 2005 Pakistan Earthquake, 2005 USA Katrina Cyclone, 2008 Myanmar Nargis Cyclone, 2004 Indian Ocean Basin Tsunami and the 2010 Haiti Earthquake.

In Europe, floods are the most common cause of natural disasters. During the summer of 2013 some parts of Germany and Czech Republic experienced severe flooding. Record breaking water levels, were observed in 2013, on the Elbe and Danube rivers. Estimated damages in Germany alone may have reached $12 billion.

As flood related disaster increases, it is likely that decision makers and authorities will increase their hydrological monitoring for the purpose of better management and planning of flood risk.

Increased monitoring will produce large amounts of data that needs to be properly managed and utilized in different modeling and decision support applications. Further, to better make use of available data coming from different sources (i.e. institutions, organizations and flood management authorities) there is a need for web-based systems for sharing and accessing these data.

There are several existing data publication methods [1] used by different organizations. In general these organizations have different systems in collecting, formatting, archiving and publishing data.

Since environmental data may come from different sources, these data are likely to be syntactically and semantically heterogeneous. Syntactic heterogeneity means different data structure or format [1]. Semantic heterogeneity is defined as the differences in the objects and attributes that define the data, leading to disagreement on the meaning, interpretation and use of the same data [2]. Semantic heterogeneity can further be divided in two types, structural and contextual, as further elaborated in [1].

To overcome data heterogeneity and to better manage, share and analyze these large amounts of data there is a need to make use of standards (recognized by scientific community) that help these data to be organized and published [1]. None of the existing data publication methods has been widely embraced by the scientific community as a standard for publishing data. However there has been relevant progress when the Open Geospatial Consortium (OGC) community started the collaborative effort on developing such standards for geospatial data, many of which are of high relevant to environmental applications. In the area of water, the OGC recently accepted the Water Mark-up Language (WaterML) 2.0 as a standardized marked up language for publishing water-related time series data.

Flood risk management (FRM) will greatly benefit in using standards for publishing, sharing and accessing data. However,

how to properly make use and present these data to users currently presents a major challenge.

Users can be decision makers, experts, flood authorities, stakeholders and citizens. An essential component for flood risk management is raising awareness of potentially-affected citizens. However for the past two decades environmental managers and authorities have seen the growing need for people to more actively participate in the environmental management [3]. Incorporating stakeholder's beliefs, values and their local knowledge of the environment in FRM will lead to more sustainable measures and decisions [4-5].

Web-based systems are increasingly seen as potential tools for flood information sharing, dissemination and participation, which may greatly enhance flood risk management. Moreover, the use of standards for publishing data on these web-based systems will enable a more efficient data transmission, subsequent analyses and decision making. These tasks are performed by different users and require different applications for meeting their needs. The integration of such diverse application can greatly benefit from the use of established standards for data publishing and sharing.

This article presents experiences for publishing water-related data over the web based on the OGC WaterML 2.0-GeoServer framework and methods developed by Australia's Commonwealth Scientific and Industrial Research Organization (CSIRO).

This implementation is one component of a web-based flood information system (FIS) developed for Somes Mare basin in Romania. The FIS was developed to be used by flood management authorities and potentially-affected citizens. The latest technologies for collection, archiving and sharing of environmental data, using web-based Spatial Data Infrastructure (SDI) were implemented.

This system was developed within a research project entitled enviroGRIDS at the Black Sea Catchment, funded by the EU FP7 Research Framework.

The article is organized as follows: Section II presents WaterML 2.0 as a standardized markup language. WaterML2.0-GeoServer architecture is presented in section III. Overview of the Somes Mare FIS development and implementation of the WaterML 2.0-GeoServer is presented in section IV. The final section discusses the advantages of using this approach, experiences and future development.

II. WaterML 2.0 as a Standardized Markup Language

In the past decades there have been initiatives from scientific communities to make use of standardized markup languages to address data heterogeneity from different sources. Such examples are the Earth science Markup Language (ESML) [6], Ecological Metadata Language (EML) [13], Observations and Measurements (O&M) by OGC [7] and Water Markup Language (WaterML) [8].These standardized markup languages are presented in Extensible Mark-up Language (XML) format.

Their differences are the vocabularies used (e.g. discharge-streamflow, precipitation-rainfall) and how the objects and their attributes that describe the data are structured.

Different versions of WaterML have already been used by several main international organizations, such as Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) from USA, Australia's Commonwealth Scientific and Industrial Research Organization (CSIRO) and Bureau of Meteorology (BOM), United States Geological Surveys (USGS) and National Oceanic and Atmospheric Administration (NOAA) from USA. Although these organizations, institutes and scientific communities have already used this language, one issue that remains is its interoperability with other systems. To address this issue, an international group of organizations that encourage development of open standards (OGC) accepted and endorsed WaterML 2.0 schema [9] as an encoding standard for publishing time series of hydrological observation data. WaterML 2.0 is an updated version of WaterML that incorporates the OGC O&M standards.

In Europe, scientific communities, organizations and institutes are encouraged to make use of this standard (WaterML 2.0) and establish a system for publishing hydrological observation data in WaterML 2.0 format.

III. WaterML 2.0-GeoServer Architecture

One known system that publishes time series in WaterML via web services is the CUAHSI-Hydrologic Information System (HIS). CUAHSI-HIS web services system called WaterOneFlow has been tested and is in use by several US environmental agencies such as USGS, NOAA and NASA.

Although CUAHSI-HIS is free to install, it requires several commercial products for it to be functional, such as Microsoft Windows Server, Microsoft ASP.NET 2.0, Microsoft SQL Server 2008, ESRI ArcGIS 9.3.1 Desktop and Server for .Net (Enterprise Advanced).

As these commercial software packages are expensive to acquire and maintain, scientific communities, organizations and institutes are searching for alternatives in publishing data over the web. Such alternative is to make use of an open source or freely available systems for publishing data over the web.

A team from Commonwealth Scientific and Industrial Research Organization (CSIRO) in Australia developed a framework and methods that implements WaterML 2.0 schema using the GeoServer Web Feature Services (WFS). CSIRO developed the WaterML 2.0-GeoServer to publish water storage time series information (from lakes and reservoirs) from the Australian Bureau of Meteorology (BoM) [10].

GeoServer and its components are an open source - general public licensed (GPL) software/technology (e.g. PostgeSQL DB, Tomcat server, etc.). GeoServer implements OGC standards for publishing spatial data. Such standards are the Web Mapping Services (WMS), Web Feature Services (WFS) and Web Mapping Content (WMC).

Fig. 1 presents the WaterML 2.0-GeoServer architecture. Aspects of data storage, discovery and access, and those related to GeoServer configuration are briefly presented below.

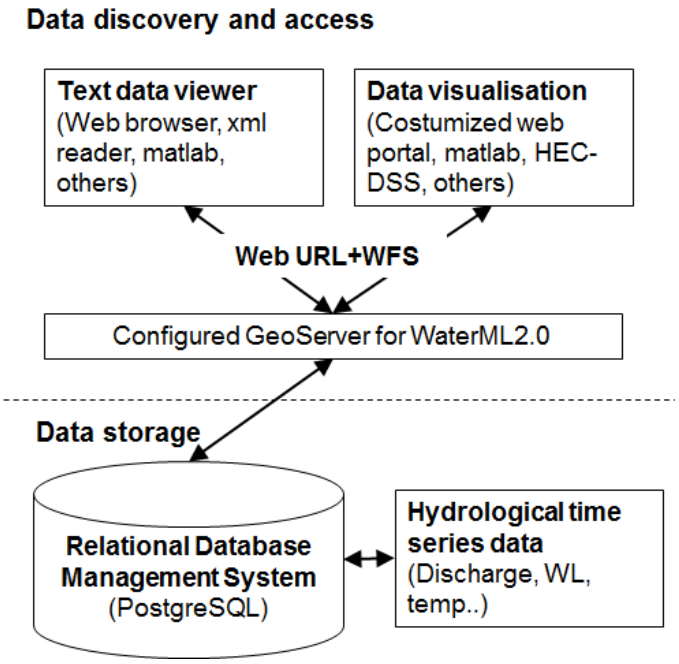


Fig. 1. WaterML 2.0-GeoServer Architecture

A. Data storage, discovery and access

Discovery and access to data is made possible through the use of Universal Resource Locator (URL) that host the GeoServer web services accompanied with WFS operators. With WFS it is possible for clients to query the data structure and the actual data. WFS can perform several operations to query the data. Such operators are the *GetCapabilities* - retrieves a list of the server's data; *DescribeFeatureType* - retrieves information and attributes about a particular dataset; and *GetFeature* - retrieves the actual data.

For WaterML 2.0 the WFS GetFearture is the main operator key to retrieve and query the actual data. For details on querying keys for WFS please see http://docs.geoserver.org/stable/en/user/services/wfs/.

Calling the URL with WFS operators can be done via web browsers or other tools that interpret an XML format. There are also some software tools that can read directly WaterML 2.0 schema and present the data in a table or charts, such as the HEC-DSS software tool by the US Army Corps of Engineers. HEC-DSS is a database management system for HEC modeling softwares (e.g. HEC-RAS and HEC-HMS).

For data storage and management the framework made use of an open source - GPL relational database management system technology (postgreSQL) compatible with GeoServer.

B. WaterML 2.0 - GeoServer configuration

The GeoServer software was re-configured by installing plug-ins for database connection, updating the schema files and GeoServer properties (e.g. workspaces). The updated Schema files contain the WaterML 2.0 structure.

The WaterML 2.0-GeoServer framework structured the information about the data in four main parts: (1) time series data, (2) geographical information and geometry, (3) data provider details and (4) details about the stations.

In reference to CSIRO BoM's implementation of WaterML 2.0-GeoServer for water storage observation [10], Table I presents the schema files description and its corresponding database and Fig. 2 presents the schematic of database table relationships. As will be shown later this structure has been adapted for the implementation of the application presented in this article.

TABLE I. CSIRO BOM IMPLEMENTATION OF WATERML 2.0-GEOSERVER: SCHEMA XML WITH ITS CORRESPONDING DATABASE TABLE

Workspace	Xml schema	Slake database table	Description
om	OM_Observation.xml	mv_om_observation	time series data
slake	SurfaceReservoir.xml	mv_surface_reservoir	lake information and geometry
	ProviderDetails.xml	mv_provider_details	data provider details
	StorageDetails.xml	mv_storage_details	lake storage details

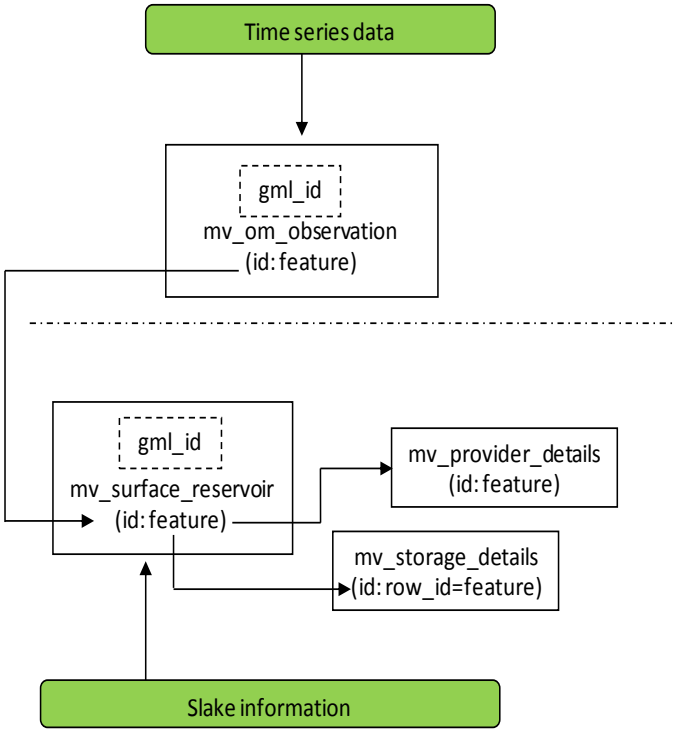


Fig. 2. Database table relationships

Reference [10] provides a development guide (designed for BoM water storage observation) to configure GeoServer for WaterML 2.0.

IV. FIS IMPLEMENTATION OF WATERML 2.0-GEOSERVER

This section first presents an overview of the web-based FIS followed by the implementation of WaterML2.0-GeoServer.

A. Overview of the Somes Mare Flood Information System

The web-based FIS was designed and developed for flood information dissemination and stakeholders-citizen participation. Latest web development technology freely available was used in its development. The following briefly presents the case study area and the FIS conceptual design.

1) Case study area

Somes Mare is a catchment of the larger Somes basin. The basin is located in the northern part of Romania. Somes Mare has an area of about 5078 km^2. The catchment is vulnerable to flooding especially during the spring season when snow from the mountains melts. Flood is most devastating when combination of rainfall with snowmelt occurs. In the past 50 years the most important flood was the one of 1970, corresponding to the 100 years return flood.

Lately, on Somes Mare River there are many occurrences of flash floods. The most devastating one was the one from 2009. This situation demonstrates the need for further studies in order to build and implement a better flood risk management strategy in the Somes Mare catchment.

Flood risk awareness of the citizens and information sharing is one of the approaches in management of floods. An innovative solution to reach the citizens and share information is through a web-based flood information system.

2) Conceptual design

The web-based FIS was designed to be simple, informative, interactive, customizable and flexible. Map based applications were extensively used for publishing geospatial data and accessing information. Furthermore, web infrastructure services and data standards were used.

The FIS has three main components: (1) FRM awareness, (2) Flood information access and (3) Citizens participation. Fig. 3 presents the conceptual design of the FIS portal.

The component FRM awareness is intended to raise citizens' awareness on the catchments flooding problems and its management plans. Flood information access is intended to raise the citizens' and stakeholders awareness on local flooding. Historical floods, observed data and model results (time series and flood maps) are presented. Moreover, flood information access has a sub-component for data access. This may be of less interest to users such as citizens and non-expert stakeholders, but more for professional users. Through this sub-component users can access the actual data (hydro-meteorological, time series and spatial data). Flood information access is provided by using the OGC standards such as WMS, WFS and WaterML 2.0. Details on the implementation of the OGC WaterML 2.0-GeoServer standards in FIS are presented in the later section.

The citizens' participation component provides the citizens and stakeholders with opportunities to discuss flood related issues, share information and timely report on local flooding.

The FIS portal can be accessed through the following URL:

- http://hikm.ihe.nl/envirogrids/Platform/Somes/

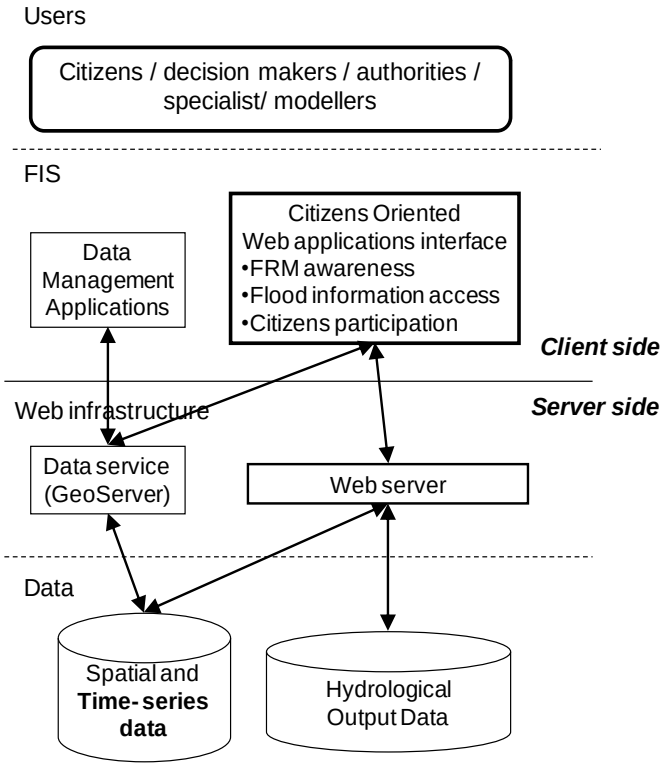


Fig. 3. Web-based FIS generic conceptual design

Reference [11-12] presents more details on the FIS design, implementation and its evaluation.

B. Implementation of WaterML 2.0- GeoServer

The web-based Flood Information System (FIS) for Somes Mare basin of Romania implemented the WaterML 2.0-GeoServer framework and methods developed by CSIRO to publish hydro meteorological time series.

The original version of GeoServer was re-configured by installing plug-ins and updating the schema files and properties (updated schema files were provided by CSIRO). To have the correct database structure the database "Slake" from CSIRO was uploaded and updated with the id's, data and related information of the Somes Mare basin. Table II presents the data tables that were updated and renamed. Fig. 4 is a sample snapshot of the modified Id.

TABLE II. DATA TABLES MODIFICATION

Slake database table	Somes Mare database table
mv_om_observation	mv_om_observation
mv_surface_reservoir	*mv_hydromet*
mv_om_observation	mv_om_observation
mv_storage_details	*mv_hydrometdetails*

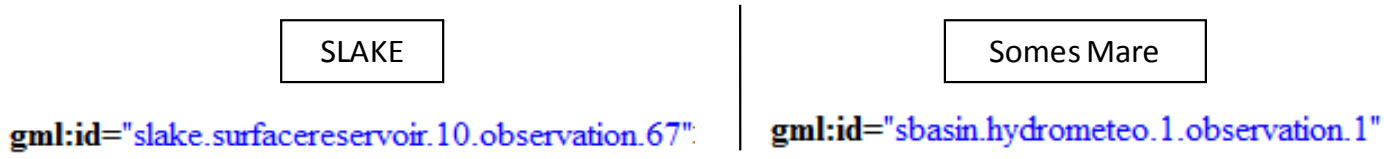


Fig. 4. Schema-database modification of id.

As previously mentioned, one sub-component of the Somes Mare FIS portal is the "Data access" (Fig. 5). The data access component allows users to view and download available spatial and time series data. The Water ML 2.0 formatted time series data are accessible through this sub-component.

Fig. 5. Somes Mare FIS portal: Data Acess.

The sub section "Hydrometeorological Data" of the data access provides a map based interface for accessing and downloading time series data of precipitation, discharge and temperature for the year 2007 in WaterML 2.0 format. The map based interface (using google maps) presents markers representing monitoring stations that provides the three types of data. The markers change as user switches from precipitation to discharge or temperature data.

The selection of stations is made possible by clicking on a station marker, using a drop-down list, or by searching for a station nearest to a given address. Once the type of data of the station has been selected, this selection is then displayed in a list box displayed to the right side of the map. Users then need to select in the list box a button to download or view the data that is in WaterML 2.0 format.

Presented in Fig.6 is an example where 4 data items have been selected (two precipitations, one discharge, and one temperature). When "Download" is selected a file type in xml is made available for download. The pre-set file name contains the type of data and name of station (e.g.

Discharge_CHIRALES.xml). When "View" is selected the data are presented in a separate web browser window, as shown in Fig. 7.

With this implementation the available precipitation, discharge and temperature data from the monitoring system of Somes Mare are accessible as web services in WaterML 2.0 format. Since this implementation is for demonstration and testing purposes, such data are provided only for the year 2007. Nevertheless, the implementation allows for other applications to access these web services and test the usage of the data for other purposes.

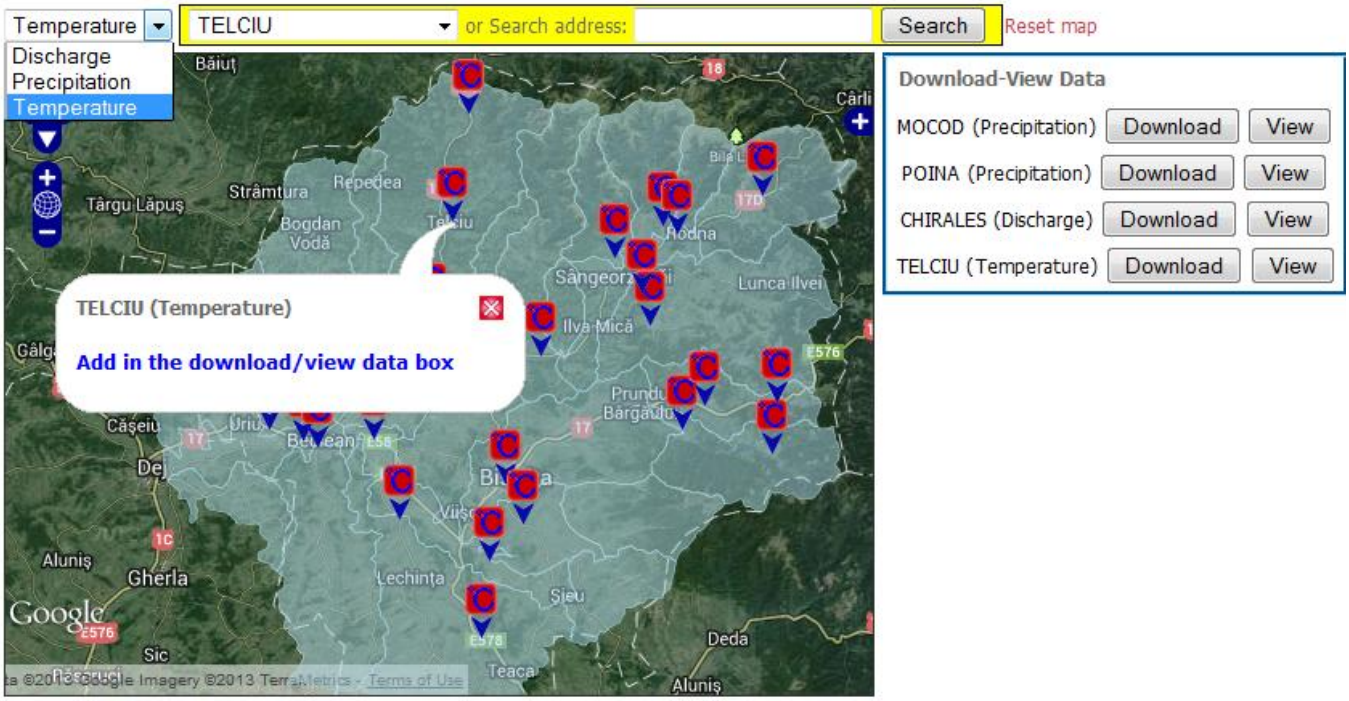


Fig. 6. Data selection

```
        2007-12-27T06:00:00 2007-12-27T18:00:00 2007-12-28T06:00:00 2007-12-28T18:00:00
        2007-12-29T06:00:00 2007-12-29T18:00:00 2007-12-30T06:00:00 2007-12-30T18:00:00
        2007-12-31T06:00:00 2007-12-31T18:00:00
      </wml2dr:timePositionList>
    </wml2dr:TimePositionList>
  </gml:domainSet>
- <gml:rangeSet>
  - <gml:QuantityList uom="http://vocabs.auscope.org/classifier/bom/sbasin/0.1/uom/celcius">
      -1.0 1.0 2.0 2.5 0.0 2.0 1.5 2.0 0.5 2.5 2.5 2.0 1.5 2.0 2.0 3.5 3.0 4.5 4.0 5.5 4.0 5.5 4.0 4.0 0.0 2.5 2.5 5.0 -0.5
      1.5 -5.0 1.0 -4.5 1.5 1.2 3.0 3.5 4.0 4.0 4.5 3.0 5.5 2.5 4.0 2.6 4.5 4.5 5.0 4.0 2.0 -1.5 -0.5 -2.0 -2.0 -4.0 -3.0
      -2.0 1.0 -9.0 -5.0 -7.5 -0.5 -1.5 2.0 -2.0 0.0 -9.0 -2.0 -3.5 2.0 -6.0 -0.5 -2.5 -0.5 0.5 1.5 1.5 3.0 2.0 2.5 1.5 3.0
```

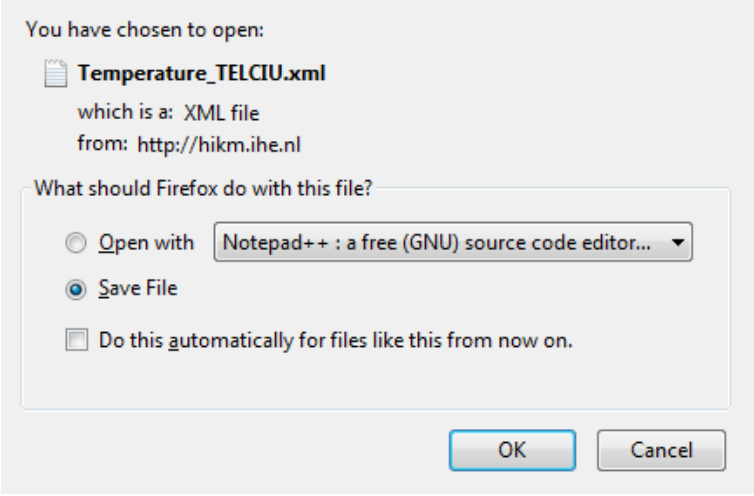


Fig. 7. Data web browser viewing and download.

The following is an example of accessing time series data using URL with WFS request.

http://sditest.unesco-ihe.org:8080/GeoserverWaterML/wfs?service=WFS&version=1.1.0&request=getFeature&typeName=om:OM_Observation&outputFormat=gml32&featureID=sbasin.hydrometeo.6.observation.2

V. CONCLUSIONS

The WaterML 2.0-GeoServer method has been successfully tested with data from the Somes Mare case study. A web-based front-end interface was developed for FIS to access these time series data. It needs to be noted that the implementation of web services using GeoServer technologies for publishing time series data in WaterML 2.0 format was not straight forward. It required several back-end configurations for achieving the intended functionalities. In general this method for publishing time series data over the web (in WaterML 2.0 format) is practical and promising. However for this method to become more usable for other case study applications there is a need for customized GeoServer package that will set-up a WaterML 2.0 web service (accompanied with a database structure) without any back-end configuration.

ACKNOWLEDGMENT

This implementation was developed within the enviroGRIDS research project for Black Sea Catchment, funded by the 7th Framework Programme (FP) of the European Union (EU). All data was provided by the Romanian Water Authority, Somes-Tisa Water Branch. CSIRO provided a sample BoM data base structure.

REFERENCES

[1] J. Horsburgh, D. Tarboton, M. Piasecki, D. Maidment, I. Zaslavsky, D. Valentine, and T. Whitenack, "An integrated system for publishing environmental observations data," Journal of Environmental Modelling & Software, vol. 24, pp. 879-888, January 2009.

[2] J. A. Sheth and J.Larson, "Federated database systems for managing distributed, heterogeneous, and autonomous databases," Journal of ACM Computing Surveys, vol. 22, pp. 183-236, September 1990.

[3] S. Hickey, and G. Mohan, Participation: from tyranny to transformation? Exploring new approaches to participation in development., Zed Books, 2004.

[4] M.B. Abbott, and A. Jonoski, "The democratisation of decision-making process in the water sector II," Journal of Hydroinformatics,vol. 3(1), pp. 35-47, 2001.

[5] M.S. Reed, "Stakeholder participation for environmental management: A literature review," Journal of Biological Conservation, vol. 141, pp. 2417-2431, October 2008.

[6] R. Ramachandran, S.A. Christopher, S. Movva, X. Li, H.T. Conover, K.R. Keiser, S.J. Graves, R.T. McNider, "Earth Science Markup Language: a solution to address data format heterogeneity problems in atmospheric sciences," Journal of American Meteorological Society, vol. 86 (6), 791–794, June 2005.

[7] S. Cox, Observations and Measurements, Open Geospatial Consortium Best Practices Document, 2006.

[8] I. Zaslavsky, D. Valentine, T.Whiteaker, CUASHI WaterML, Open Geospatial Consortium Discussion paper 07-041r1, 2007.

[9] P. Taylor, OGC WaterML 2.0: Part 1- Timeseries, Open Geospatial Consortium Implemented Standard 10-126r3, 2012.

[10] CSIRO, Publish BOM Observational Data (online), https://www.seegrid.csiro.au/wiki/ASRDC/PublishBomObservationalData, 2012, last accessed July 5, 2013.

[11] A. Almoradie, A. Jonoski, F. Stoica, D. Solomatine, and I. Popescu, " Web-based Flood Information System-: Case study of Somes Mare, Romania," Environmental Engineering and Management Journal, 2013, in press.

[12] A. Jonoski, I. Popescu, A. Almoradie, F. Stoica, S. Teodor, I. Jelev, and D. Gorgan, Functional prototypes of BSC-OS Flood Portals for citizens: Implementation and Evaluation, Final report enviroGRIDS Deliverable D6.11, 2013.

[13] EML Project, Ecological Metadata Language (EML) (online), http://knb.ecoinformatics.org/software/eml/, 2008, last accessed July 5, 2013.

OWS4SWAT: Publishing and Sharing SWAT Outputs with OGC standards

Gregory Giuliani[1,2],Kazi Rahman[1], Nicolas Ray[1,2], Anthony Lehmann[1]
[1]Institute for Environmental Sciences, enviroSPACE University of Geneva 1227 Carouge, Switzerland
[2]United Nations Environment Programme Global Resource Information Database, 1211 Châtelaine, Switzerland
gregory.giuliani@unige.ch

***Abstract*—The Soil and Water Assessment Tool (SWAT) is a widely used hydrological model that produces several useful outputs (e.g. evapotranspiration, soil moisture, aquifer recharge, river discharge) as text files. Currently, visualizing and publishing SWAT outputs as geospatial data requires a lot of time and repetitive processing steps. Moreover, data used and produced are often not interoperable and restricted to software like ArcGIS or MapWindow. Consequently, integrating SWAT outputs with other datasets and/or models is difficult. To solve these issues, we propose an innovative, scalable and interoperable framework allowing (1) the automatization of post-processing tasks using orchestrated Web Processing Services (WPS) and (2) the publishing of SWAT outputs using interoperable data services (e.g Web Feature Service, Web Map Service). The proposed framework simplifies map/data production and facilitates exchange/integration of hydrological data with other sources.**

Keywords—enviroGRIDS; SWAT; Water; Interoperability; WPS; OGC; GEOSS; Black Sea

I. INTRODUCTION

Humans are exerting significant impacts on the global water system [1] through activities such as the accelerated melting of snow and ice in alpine zones, the removal of trees that lead to increased runoff, reduced transpiration and impacts on the water table and its salinity, the draining of wetlands, the irrigation for agriculture, the alteration of flow through dams, the transfer of water between catchments, and finally the pollutions from industrial, agricultural and domestic sources. To better understand these modifications and impacts, water science research needs to follow a holistic research approach in order to effectively inform policy for sustainable water management about the dynamics of water in the context of global needs.

However, it is recognized that many policy-relevant research areas are still facing the problem of readily and timely access and exchange of data. Access and availability of reliable time-series on environmental, statistical, and socio-economical data is still a major barrier to effective and efficient informed policy-making [1]. Additionally, there are currently also gaps in term of knowledge when analyzing different water cycles: water scarcity (e.g. droughts), water abundance (e.g. floods), water quality (including sediment loads evaluation), water use and renewability, interactions between extremes (e.g. interconnections between drought and flood distribution), and ecosystem services maintenance.

Effective and efficient water management requires coordination of actions, the first one being the access and provision of reliable data and information (e.g., state of the resources, changes, pressures) and second the capacities to interpret correctly and meaningfully these information [2, 3]. Hydrological modeling, being interdisciplinary, complex and dynamic by nature, intrinsically asks for better integration of data, information and models [4-6]. The objective is to bring to policy/decision-makers efficient tools as well as suitable and reliable information, both supported by scientific knowledge and models.

Hydrological models are simplified representations of parts of the water cycle. They are primarily used for predictions and for understanding hydrological processes. The Soil and Water Assessment Tool [1] (SWAT) [7, 8] is a semi-distributed, continuous watershed simulator operating on a daily time step. It is developed by USDA Agricultural Research Service (USDA-ARS) and Texas A&M AgriLife Research, for assessing the impact of management options as well as global changes on water supplies, sediment transportation and agricultural chemical yields in watersheds and larger river basins. This model performs simulations that integrate various processes such as hydrology, climate, chemical transport, soil erosion, pesticide dynamics and agricultural management. SWAT accounts for soil and land cover conditions by subdividing the simulated catchment into homogeneous Hydrological Response Unit (HRU). The model uses a daily to sub-hourly time step, and can perform continuous simulation for a 1- to 100-year period. SWAT simulations are typically prepared from ArcGIS [2] (ArcSWAT) or MapWindow [3] interfaces (MWSWAT). For the simulation, SWAT requires data on elevation, soil, land cover, reservoirs, agricultural practices and weather for model setup. River discharges, water quality and crop yield (as available) are needed for calibration and uncertainty analyses. Once this data is gathered and formatted, SWAT is able to model the water cycle inland and in-stream components.

SWAT was used to simulate the continent of Africa [9] and in the "Hydrologic Unit Model for the United States" (HUMUS) [7], where the entire U.S. was simulated with good results for river discharges at around 6000 gauging stations.

[1] http://swat.tamu.edu

[2] http://swat.tamu.edu/software/arcswat/

[3] http://swat.tamu.edu/software/mwswat/

This study is now extended within the national assessment of the USDA Conservation Effects Assessment Project (CEAP 21). Other large scale SWAT application included the work of Gosain et al. [10] where twelve large river catchments in India were modeled with the purpose of quantifying the climate change impact on hydrology. SWAT is recognized by the U.S. Environmental Protection Agency (EPA) and has been incorporated into the EPA's BASINS system (Better Assessment Science Integrating Point and Non-point Sources) [11].

In every SWAT simulation, several important hydrological variables are estimated (e.g., precipitation, snow or ice melting, evapotranspiration, water yield, groundwater recharge/transfer, suspended/dissolved load, pollutants) and stored in different text files[4]. They provide temporal resolution of yearly, monthly and daily time steps based on user interest for subbasin (output.sub), main river reach (output.rch) and HRU (output.hru). Potentially, SWAT provides many useful outputs for both scientists and decision-makers, but in a not very accessible format.

Preparing, calibrating, executing and publishing outputs of a SWAT model are often time-consuming and repetitive tasks. In particular, while gathering the required data to set up a SWAT model, users are regularly facing the problem of data accessibility and data heterogeneity (e.g., different data formats). Additionally, results of a SWAT simulation can not be visualized directly on a map. Different processing steps in various software are required for generating geospatial data from the output text files. Finally, these results are often prepared using closed/proprietary formats and therefore limit their usability and making them difficult to integrate with other data sources and/or models.

Recognizing the need for efficient and effective data accessibility and considering that SWAT outputs can be valuable for different community of users (e.g, hydrologists, environmentalists, biologists, decision-makers), a scalable and interoperable framework simplifying and automatizing repetitive tasks like data gathering and map production can be beneficial. Based on these considerations the aim of this paper is to present a proof of concept (1) to facilitate gathering and harmonization of SWAT data inputs, (2) to facilitate the publishing of SWAT simulation outputs, (3) to expose these results in a standardized way using interoperable services, and (4) to facilitate the exchange of data, and integration with other resources.

II. BACKGROUND

A. SWAT models preparation and outputs visualization

SWAT models preparation and outputs visualization are generally realized in ArcGIS [12] using ArcSWAT and VizSWAT[5] extensions for building the model and respectively visualizing results as dynamic graphs or maps. There is also an open source alternative to ArcSWAT providing the same functionalities for model preparation based on the MapWindow[6] GIS system and the MWSWAT plugin [13]. In term of output visualization two other solutions are currently available. Field_SWAT [7] [14] is a graphical user interface developed in MatLab for preparing maps and SWATShare[8] is a web-based tool for uploading, executing and visualizing SWAT simulations.

The first task required to users while setting up a SWAT model is to gather the necessary data. Traditionally, users must:

1) Identify the relevant data sources,
2) Download these data on their computer,
3) Harmonize data formats, resolution, and projections.

The data needed to prepare a SWAT model are:

- *Geospatial data (raster and vector):* Digital Elevation Model (DEM), Land Use (LU), Soils, and river network.
- *Weather data (tables):* (daily) precipitation and temperature.

Nutrient and sediment loads can be used to predict water quality. Additionally, climatic data from global or regional climate models can be used to predict the impacts of climate changes on the hydrological model. Once all these data have been gathered, downloaded and harmonized, then users can prepare, calibrate and execute their SWAT models.

Once a model has been executed then users want to visualize the results of their simulations. A typical workflow for processing outputs to prepare maps involves the following steps:

1) Open a text editor to filter and remove unnecessary columns in the output file (e.g., output.sub, output.rch)

2) In a spreadsheet editor open the cleaned output file, separate independent variables using tab delimited option, use pivot table to calculate average the values, and finally save independent value (each variable) in a text file for each variable (in csv format).

3) In a GIS software, join the shapefile of the watershed delineation or the river reaches with the table values, classify data according to the values of the variable, and finally save the map.

These tasks involve different proprietary software that use closed formats [12]. Consequently, a standardized approach for a rapid collection of required datasets for a given geograpical area is needed. It should automatically structure the data into the input format of SWAT. An automatic procedure to publish maps-results based on non-proprietary formats can be very convenient as well to improve the interoperability of SWAT outputs.

[4] http://swat.tamu.edu/media/69395/ch32_output.pdf

[5] http://swat.tamu.edu/software/vizswat/

[6] http://www.mapwindow.org

[7] http://baegrisk.ddns.uark.edu/SWAT_Model_Tools/Field_SWAT/

[8] https://water-hub.org/swat-tool

B. Interoperability in the water domain

Presently, hydrological and meteorological data still remain difficult to find, access, and integrate because of various incompatibilities (e.g., data formats, models specifications, quality needs), missing documentation (e.g., metadata), data fragmentation and replication, data policies, and these systems are operating in isolation [15]. Interoperability, the ability to exchange and use information between two or more systems/components, is therefore an essential condition to enable efficient data publishing, discovery, evaluation and access to environmental data [16].

Current technologies are suitable to match these requirements only if open software interfaces and standards are developed allowing these technologies to interoperate at a large scale [17]. The Open Geospatial Consortium (OGC) aims to develop and provide such standards enabling communication and exchange of information between systems of different types operated with distinctive software. Indeed, a non-interoperable system cannot share data and computing resources, inducing scientists to spend much more time than necessary on data discovery and transformations. One of the major benefits of interoperability is to enable locally managed and distributed heterogeneous systems (e.g., different operating systems, databases, data formats) to exchange data and provide a service [18]. A good example is the distributed information system developed by the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI). It is a web-based system for storing, publishing, sharing, processing, and analyzing hydrological data through a full suite of software and standardized/interoperable services [19].

The OGC, completed by ISO standards, is providing a suite of standard specifications to search, discover, and access of heterogeneous geospatial resources. These resources can be maps served with Web Map Service (WMS) [20], vectors and raster data published respectively as Web Feature Service (WFS) [21] and Web Coverage Service (WCS) [22], or processing algorithms exposed as Web Processing Service (WPS) [23]. Data and services can be documented through International Organization for Standardization (ISO) 19115 (resource metadata), 19139 (metadata encoding) and 19119 (service metadata). ISO standards are complemented by the OGC Catalog Service for the Web (CSW) specification [24] defining an interoperable interface to publish, discover, search and query metadata.

Currently, the OGC has several projects underway related to water resources [9]. In particular, it has a Hydrology Domain Working Group [10] that is seeking to develop and provide solutions for describing and exchanging data related to water resources. For example, WaterML 2.0 has been recently accepted as a standard [11]. WaterML is an XML-based specification used to formally describe hydrological data and act as an interchange format via the Internet through web services. It contains specifications for both point and spatial coverage data (via XML elements) as well as a set of generic vocabularies. Additionally, the OGC is also conducting Interoperability Experiments on Surface Water, Ground Water, and Hydrologic Forecasting as well as developing pilot studies on Hydro-climatology Information Sharing.

Finally, several initiatives are catalyzing data sharing by promoting interoperability to maximize the (re)use of data and supporting easy access to and utilization of geospatial data. At the global level, the Global Earth Observation System of Systems (GEOSS) [25] has a dedicated Societal Benefit Area on Water [12] and related activities like the Water Cycle Integrator or Interoperability Experiments on Weather, Ocean, and Water. Similarly, Eye on Earth has recently launched a special initiative on Water Security[13]. At the European level the Infrastructure for Spatial Information in the European Community (INSPIRE) [26] has a "Cross-Border Water Management" Initiative[14] to contribute to the implementation of the Water Framework Directive.

Accordingly, making SWAT outputs interoperable will simplify their sharing/exchange, expend their potential applicability and facilitate their contribution initiatives like GEOSS or INSPIRE.

III. OWS4SWAT FRAMEWORK

The proposed framework called OGC Web Services for SWAT (OWS4SWAT) is entirely based on OGC standards to help: (1) data gathering and harmonization, while setting up a model, and (2) map preparation and publication, when results of simulations are available. Currently, OWS4SWAT is not influencing model preparation, calibration, and execution. These operations are still manually executed on desktop computers (Table 1).

TABLE I. COMPARISON BETWEEN "TRADTIONAL" AND OSW4SWAT WORKLOWS

	Traditional	**OWS4SWAT**
Data gathering and harmonization	Manual download, processing/harmonization, repetitive tasks, heterogeneous and not interoperable data.	Automatic download and processing accomplished by webservices on the server, based on OGC standards
Model preparation, calibration, and execution	Manual on desktop computer	Manual on desktop computer
Outputs preparation and publication	Manual preparation and upload on server, repetitive tasks, results often not interoperable	Automatic preparation and publication done by webservices, OGC-compliant

Thanks to the use of interoperable services, this will not be dedicated to specific software, consent the use of different clients (e.g., desktop, web), and facilitate data access, exchange and integration.

9 http://www.opengeospatial.org/node/1535

10 http://www.opengeospatial.org/projects/groups/hydrologydwg

11 http://www.opengeospatial.org/standards/waterml

12 http://www.earthobservations.org/geoss_wa.shtml

13 http://www.ogcnetwork.net/system/files/EoE%20SI%20Water%20Security%20-12-for%20Summit.pdf

14 http://inspire.jrc.ec.europa.eu/index.cfm/pageid/42/list/7/id/2688

A. Architecture and design

The two main functions of OWS4SWAT are built with different software packages:

- Data processing/handling: is programmed using PyWPS[15], an OGC WPS 1.0.0 implementation written in Python. PyWPS does not process data by itself but wraps different backends to both access geospatial (GRASS) and statistical (R) functionalities.
- Data publishing: is based on the OpenGeo Suite Community Edition[16]. It is an integrated software package made of different components to store (e.g., PostgresSQL/PostGIS), publish (e.g., GeoServer), and develop web-mapping applications (e.g., OpenLayers, GeoExt) based on WMS, WFS, and WCS OGC standards.

The advantages of using these software solutions is that they fully implement OGC standards, are Free and Open Source, have an important community of both users and developers, and can be installed on different Operating Systems (e.g., Linux, Windows, Mac).

Within this architecture, different WPS services are developed to answer the requirements of data download/harmonization and SWAT results publishing. Outputs data are made available using WMS for visualization and WFS for data access. Once published data can be accessed, manipulated, styled, and integrated in desktop (e.g., QGIS) or web-based (e.g., GeoExplorer) clients (fig.1).

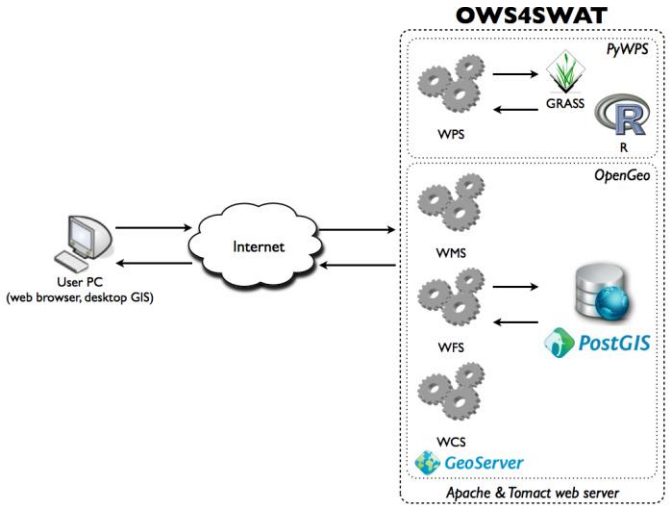


Fig. 1. OWS4SWAT architecture for managing SWAT data inputs and publishing its outputs as OGC web services.

B. Web Processing Service (WPS)

To enable interoperability and automatization of tasks, OWS4SWAT framework mostly relies OGC Web Processing Service (WPS) specification. WPS processes are flexible and remotely accessible algorithms available through web services and can be reused in different workflows [27, 28]. The core element of a WPS is the process, a calculation with defined inputs and outputs [29, 30]. It allows users to know which processes are available, to select the required input data and their formats, to create a model and run it, to manage processes (status, storage for the output ...) and to return the output once computation is completed.

As any web service, a WPS instance must expose different operations that are accessible through standardized web communication (e.g., XML). In particular, descriptions of algorithms through metadata that are usable and understandable both by humans and other web services [31] are essential elements to develop chains of services [32]. WPS specification includes a set of three operations that can be called using HTTP-GET, HTTP-POST or SOAP/WSDL:

- *GetCapabilites:* answer to a client describing its capabilities in an XML document. It tells the client which kinds of process are available.

```
http://localhost/cgi-bin/wps?
service=WPS&request=getcapabilities
```

- *DescribeProcess:* describe the parameters of a selected process also through an XML document (e.g., input and output).

```
http://localhost/cgi-
bin/wps?service=WPS&version=1.0.
0&identifier=buffer&request=describeproc
ess
```

- *Execute:* execute a selected process.

```
http://localhost/cgi-
bin/wps?service=WPS&version=1.0.0&identi
fier=buffer&request=execute&datainputs=[
data=http://foo.bar/cities.gml; width=3]
```

The output of a process can be obtained either by a direct download (e.g., result sent immediately to the client after the end of the execution) or as resource stored on the server and accessible through the web using URLs [33]. In such a case, the Execute response will be an XML document providing the URLs to access each stored output.

In a web service environment, it is possible to seamlessly couple and reuse services to perform various (complex) tasks organized in sequences or chains of processes. Service chaining can be defined as a mechanism to build flexible, coherent, and efficient workflows by combining individual web services to create customized web applications based on distributed services [34-36]. This offer significant potential to modularize, reuse, and share software components [37]. The International Organization for Standardization (ISO) through its ISO 19119 standard defines three types of chaining mechanisms: (1) *transparent* where the workflow is defined and managed by users, (2) *translucent* where users are familiar with atomic services that compose the chain and invoke a service to manage the chain, (3) *opaque* where users invoke an aggregated service that execute the chain but do not have knowledge about the atomic service constituting the chain. These chains of services can be coordinated by orchestration engines generally based on Simple Object Access Protocol (SOAP) and Web Service Description Language (WSDL) to exchange structured information [38].

C. Implementation

The general workflow when working with SWAT is the following:

1) A user needs to download and harmonize required data to build a SWAT model.

2) Once he has all data he prepares, calibrates and executes the model on a desktop computer.

3) The user retrieves the results of a simulation from output.sub, output.rch, output.hru files and prepare maps and/or graphs to visualize and share its results.

This general workflow has been transposed in a web service environment (fig.2) assuming that the step 2 (e.g., model preparation, calibration, execution) is done on a Desktop computer and therefore can be considered as a standalone task that do not require interactions with web services.

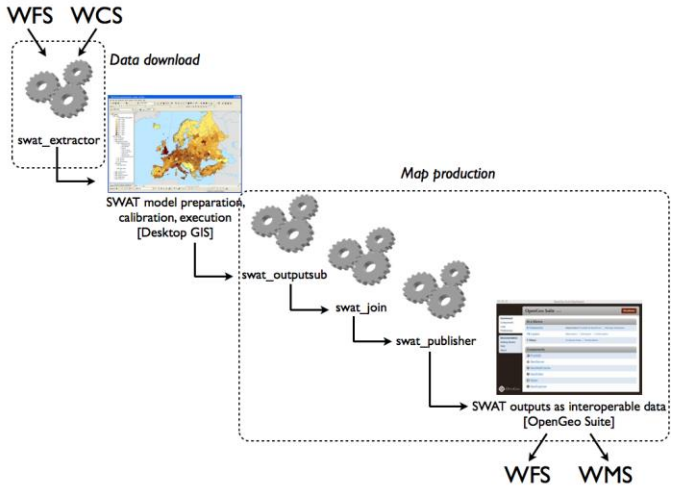


Fig. 2. General workflow in a web service environment for downloading SWAT input data (step1), modeling (step2) and publishing of the outputs (step3)..

To differentiate between data download (step 1) and map production tasks (step 3), the general workflow has been sub-divided into two independent workflows. This also reflects the fact that step 1 (fig.3) is accomplished before the modeling exercise, while step 3 (fig.4) is completed after successful execution of the model.

The tasks mentioned in section II.A have been disaggregated in atomic functionalities and implemented as dedicated WPS processes that can be chained to achieve the objective of automating of processing tasks for data download and maps production. The different processes are written as PyWPS scripts that, depending on the required functionalities, will interact with GRASS for geoprocessing algorithms and R for statistical functions

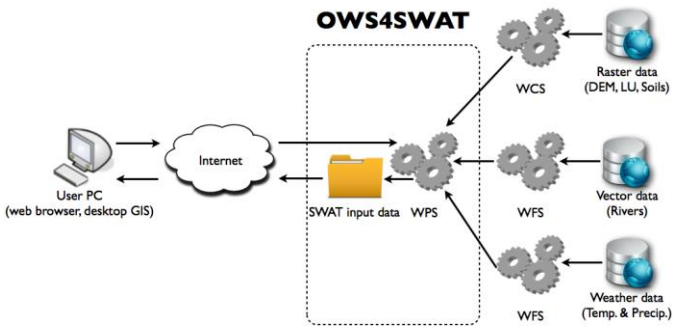


Fig. 3. Download workflow implemented in OWS4SWAT with the swat_extractor WPS process (step1).

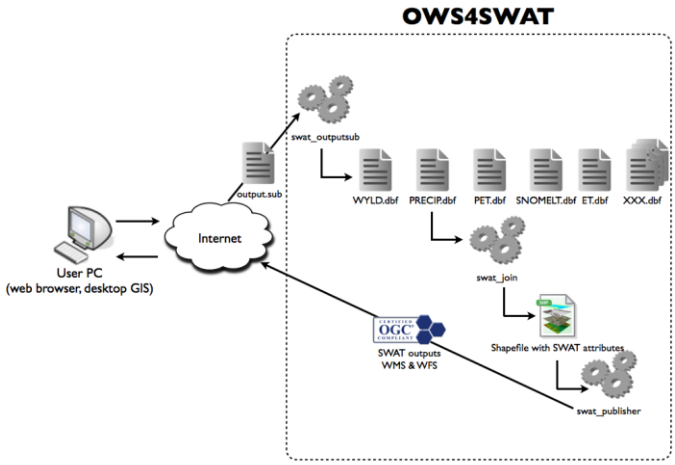


Fig. 4. Maps production workflow in OWS4SWAT with a chain of three WPS processes (swat_outputsub, swat_join, swat_publisher) (step3).

- *swat_extractor:* extract required SWAT input data from different WFS/WCS endpoints for a dedicated area (e.g, bounding box) specified by the user and store them in zip file.
- *swat_outputsub:* generates from the output.sub (e.g., .rch and .hru files are not handled) file a set of DBF files, one for each SWAT simulated variables.
- *swat_join:* join all the DBF files with the a geospatial file of the computed subbasin.
- *swat_publisher:* store in a PostGIS database the geospatial file with all SWAT variables as attributes and publish this file with GeoServer using OGC WMS and WFS standards.

Processes with their identifier, computing backends, data input and output variables are summarized in Table 2.

TABLE II. OWS4SWAT PROCESSES DESCRIPTION

identifier	backend	inputs	outputs
swat_extractor	GRASS	riversInUrl demInUrl soilsInUrl landuseInUrl tempInUrl precipInUrl northBound southBound eastBound westBound	swatOut
swat_outputsub	R	outputSub	swatDbf
swat_join	GRASS	dbfFile shpFile	shpOut
swat_publisher	GRASS	shpIn wkspIn dstoreIn pgIn	Shapefile stored in PostGIS and published in GeoServer with WMS and WFS

All these WPS processes are individually available and they can be executed with various desktop or web-based WPS clients. For example, ArcGIS 10.1 or QGIS 1.8 (with WPS plugin) are able to consume the proposed services and to execute the different tasks of data download and map production/publication. More specifically, with the map production/publication workflow, swat_outputsub, swat_join, and swat_publisher processes can be chained because output of a process can be used as input for the following process [39]. Since version 3.2, PyWPS implements a SOAP/WSDL interface enabling chaining WPS services in WSDL-based Workflow Management Systems. [39].

The WSDL file is generated dynamically by making a WSDL request to the WPS instance: `http://localhost/cgi-bin/wps?wsdl`. The description is created applying a XSLT template to a *DescribeProcess* operation output and contains all the processes request/responses provided by the WPS instance allowing exposing them as WSDL-based services. These different processes have been successfully chained with Taverna[15] and SEXTANTE[16] modeler. The former is a scientific workflow management system while the latter is a geospatial data processing framework available for different Desktop GIS clients like QGIS and providing different tools such as a graphical modeler. Interestingly, Taverna, that is not meant to be a geospatial workflow manager, has been able to efficiently handle geospatial data.

Consequently, executing this workflow enables users to automatically process and publish the different simulated SWAT variables with OGC standards using both specialized and non-specialized software. Users only have to provide their output.sub file and the chain of processing services will execute all the tasks and greatly simplified the publication of their results.

D. Use-case: publishing results of the enviroGRIDS Black Sea catchment SWAT model.

The prototype OWS4SWAT framework has been developed in the context of the enviroGRIDS project, funded by the European Commission (EC) Seventh Framework Program. This project concentrates on the unsustainable development and the inadequate resource management that is affecting the Black Sea catchment region. A large catalog of environmental data sets (e.g., land use, hydrology, and climate) has been gathered, published and used to carry out distributed spatially explicit simulations to build scenarios of key environmental changes.

Advances in distributed computing in conjunction with data availability from interoperable web services have made high-resolution modeling of distributed hydrologic processes possible [40, 41]. In the frame of enviroGRIDS, a high-resolution SWAT (sub-catchment spatial and daily temporal resolution) model of the entire Black Sea catchment has been developed. This model, divided into 12982 subbasins, which were further divided into 89202 HRUs, will be used to predict water quality and quantity according to the different scenarios in the region (e.g, Land Use, Climate, and Demographic changes). Subsequent analyses of land use change, agricultural management change, and/or climate change can then predict the consequence of various scenarios. Finally, all results of the Black Sea SWAT model and scenarios will be registered and made available as OGC services to feed the Global Earth Observation System of Systems (GEOSS) and contribute to the Water Societal Benefit Area (SBA). Consequently, the OWS4SWAT framework has been used to test the validity of the proposed approach for facilitating the publication of the Black Sea SWAT model results. This first experiment have shown that it facilitates and accelerates the publication of SWAT outputs, allow to process larger files that where difficult to handle on desktop computer, and simplify the access and integration with other data sources in different clients (fig.5&6).

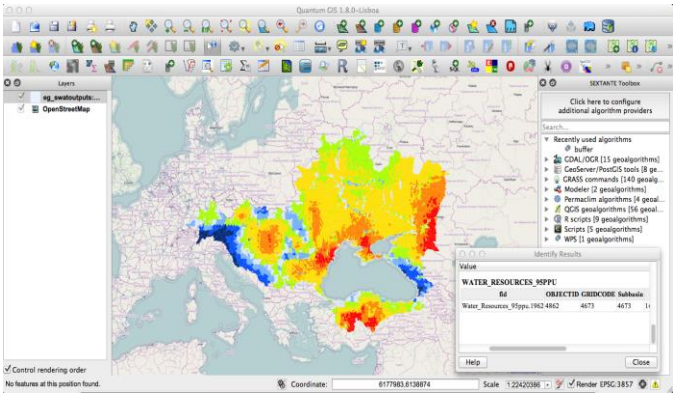


Fig. 5. Soil Water content (e.g., SW variable) modeled with SWAT, accessed in WFS in QGIS Desktop GIS, and integrated with OpenStreetMap background.

[15] http://www.taverna.org.uk

[16] http://www.sextantegis.com

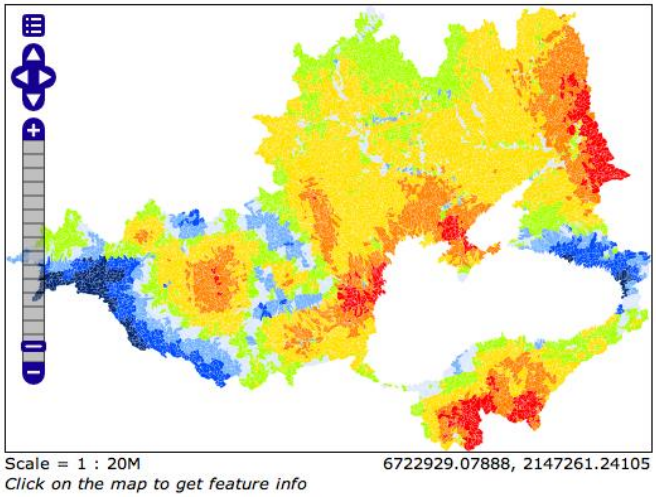


Fig. 6. Soil Water content (e.g., SW variable) in the Black Sea catchment visualized with WMS in a web-based OpenLayers client.

IV. Discussion

The OWS4SWAT framework is, to our knowledge, among the first attempt to bring interoperability based on OGC specifications around SWAT software. The only study we have found concerns the derivation of HRUs based on a WPS implementation [42]. The proposed approach was developed as a proof-of-concept and the implementation was successful. The first results have highlighted both benefits and limitations but we are convinced that such approach can bring relevant benefits for the numerous SWAT users (e.g., hydrologists) as well as users of SWAT results (e.g., scientists of other communities).

A. Benefits

The major benefit of this approach is to enable interoperability around SWAT applications. Indeed, using OGC standards helps to solve the problem of seamlessly integrating multiple heterogeneous data source [43]. From a scientific perspective, having data published in a comparable form considerably facilitate data acquisition, interpretation and comprehension [43]. Several studies have already demonstrated the benefits of interoperability and web services in the water domain for data visualization [44], data publication [45], data distribution [46], data discovery and retrieval [47] and modeling [48]. All these authors stress the fact that interoperability offers new and promising opportunities to the water research community for systematic data management, publication and analysis [49].

Our approach helps overcoming the problems of poor data accessibility and of different formats for specialized scientific models. It enables a standardized approach for rapid data collection on a given geographic area and automatically formats data into the input format of the targeted model. It makes environmental modeling much more efficient by making better use of existing data sources and by reducing the time for finding, gathering and preparing environmental input data. It also simplifies and accelerates the publication of model results by automatizing repetitive tasks through a chain of processing services. Moreover, due to the fact that the processing is executed on a server and not on a desktop computer, it allows faster processing of data of a given size and also allows processing larger data sets (i.e., higher resolution). Finally, it increases model results availability and discovery by publishing them according to OGC standards. This facilitates sharing, exchange, and integration of SWAT output data with other data sources. Data can be interactively visualized with WMS, accessed with WFS for subsequent geospatial or graph analysis [50], and processed with WPS. Consequently, data are no more restricted to dedicated software or formats but instead can be consumed by a wide variety of clients. It can be light web-based client like the enviroGRIDS portal[17], desktop GIS clients like QGIS, or specialized hydrological software like HydroDesktop [51].

The scalability of the OWS4SWAT is another advantage thanks to the use of interoperable services. Indeed, it is not difficult to incorporate different (data and/or processing) services implemented by other providers or create new services to build new and more complex workflows. Therefore, composability, extension, and integration are further enhanced and allow envisioning interaction with different scientific disciplines and coupling other models. Integration through web service of heterogeneous data and modeling resources have been demonstrated in hydrology and climatology [52], surface dynamics [53], biodiversity [54], and ecosystem services [55]. Despite the fact that many challenges need to be overcome (e.g., discovery, semantic, ontologies, performances) [56], all authors recognize that there is a need to improve collaboration among scientific disciplines and the diversity of environmental models requires effective and efficient solutions of using and reusing functionalities or services provided by others [57]. Promising solutions have been developed to expose models either using WPS [58] or Open Modeling Interface (OpenMI) [59] standards. Model integration based on interoperable services can significantly reduce time, efforts, and technical challenges for scientists who only have to concentrate on their expertise while developing components [4]. In the coming years, the Group on Earth Observation (GEO) Model Web initiative will certainly catalyze and improve models access, sharing, and integration [56] and ultimately enable an holistic vision/understanding of the Earth system to better address decision-making processes.

B. Limitations.

The current implementation of OSW4SWAT framework has different technical limitations that may be overcome with further developments.

At the moment, this is still a prototype and has only been tested for publishing some intermediate SWAT outputs. Moreover, only output.sub files at daily time step can be handled and two other processes need to be developed to process .rch and .hru files. The next step is therefore to move to a production scenario ingesting and publishing all results of a SWAT model.

SWAT outputs visualization not only concerns maps but should also graph generation. Two solutions exist to tackle this issue: (1) develop a dedicated process to generate graphs

[17] http://portal.envirogrids.net

directly from raw results (e.g., output.sub file) or (2) due to the fact that SWAT results can already be published in WFS, it is possible to access the attribute table and consequently developing a process to access directly data in WFS and generate graphs.

The use of interoperable services can also be perceived as a limitation because as of today not all data providers are publishing their data using standards. Moreover, even if data providers use standards, they may differ from one scientific community to another (e.g., netCDF in climatology, WaterML in hydrology). Consequently, the current implementation of OWS4SWAT is limited to resources exposed as WMS, WFS, WCS, and WPS. A possible solution to access resources based on standard and non-standard capacities is to rely on the brokering approach to access heteregeneous resources in a consistent and uniform manner [60].

C. Perspectives

Thanks to its scalability the OWS4SWAT framework will be improved with forthcoming developments:

- Process to publish SWAT results using the recently adopted OGC WaterML2.0. This will increase interoperability between SWAT and other hydrological systems/applications.
- Implementation of a process to visualize SWAT outputs as graphs.
- Develop a new workflow to compute vulnerability maps of water resources based on SWAT outputs.

V. Conclusions

The proposed OWS4SWAT framework enables SWAT users to share more easily and efficiently their model results using OGC standards. By making these results interoperable, it facilitates the exchange of data, and integration with other resources. SWAT model outputs accessibility and visualization are consequently no more restricted to dedicated software but can be available for various types of (desktop or web-based) clients. This can greatly expand the use of SWAT results and their applications.

The use of chained WPS services has simplified the processing of SWAT results and the automating of repetitive tasks of data handling, map preparation, and data publication. Typically, these tasks are executed more rapidly compared to the "traditional" way of manipulating SWAT model outputs for preparing maps. Furthermore, it enhances the scalability of the framework allowing one to easily and seamlessly incorporate new services, to extend existing workflows or create new ones.

Interoperable data and processing services not only help scientists to share their data or computational algorithms but also enhance their reusability and therefore can facilitate the development of complex scientific workflows to solve complex problems. Indeed, by linking distributed and heterogeneous data and processing services, it offers new opportunities to scientists for processing data and for communicating scientific results and hence contributing of better resource management and help decision-makers. It makes possible to benefit from the abundant scientific resources to more efficiently explore and better understand complex interactions between the different components of the Earth system. Ultimately, it offers a promising potential for coupling different models (e.g., the output of one model serves as input in the other, the communication is based on interoperable services) and therefore facilitate the development of integrated models. Finally, sharing data, processes and models, is also part of the elementary scientific approach and thus enhance scientific accountability, credibility, and facilitate the replication and comparison of workflows and methodologies.

Acknowledgment

The authors would like to acknowledge the European Commission ''Seventh Framework Program'' that funded the enviroGRIDS (Grant Agreement no. 227640) and ACQWA projects (Grant Agreement no. 212250). We thank Jorge Mendes de Jesus, Jachmy Cepicky, and PyWPS community for their valuable help, support and developments. The views expressed in the paper are those of the authors and do not necessarily reflect the views of the institutions they belong to.

References

[1] UNEP, Global Environment Outlook (GEO) - 5: Environment for the future we want, 2012. p. 550.

[2] Gerlak, A.K., J. Lautze, and M. Giordano, Water resources data and information exchange in transboundary water treaties. International Environmental Agreements-Politics Law and Economics, 2011. **11**(2): p. 179-199.

[3] Roehring, J., Information Interoperability for River Basin Management, in Technology Resource Management and Development2002. p. 127-134.

[4] Argent, R.M., An overview of model integration for environmental applications,Äîcomponents, frameworks and semantics. Environmental Modelling & Software, 2004. **19**(3): p. 219-234.

[5] Buytaert, W., et al., Web-Based Environmental Simulation: Bridging the Gap between Scientific Modeling and Decision-Making. Environmental Science & Technology, 2012. **46**(4): p. 1971-1976.

[6] Papajorgji, P., H.W. Beck, and J.L. Braga, An architecture for developing service-oriented and component-based environmental models. Ecological Modelling, 2004. **179**(1): p. 61-76.

[7] Arnold, J., et al., Large area hydrologic modeling and assessment - Part 1: Model development. Water resources bulletin, 1998. **34**(1): p. 73-89.

[8] Srinivasan, R., et al., Large area hydrologic modeling and assessment - Part II: Model application. Water resources bulletin, 1998. **34**(1): p. 91-101.

[9] Schuol, J., et al., Modelling Blue and Green Water Availability in Africa at Monthly Intervals and Subbasin Level. Water Resources Research, 2008. **44**: p. 1-18.

[10] Gosain, A.K., S. Rao, and D. Basuray, Climate change impact assessment on hydrology of Indian river basins. Current Science, 2006. **90**(3): p. 346-353.

[11] Gassman, P.W., et al., THE SOIL AND WATER ASSESSMENT TOOL: HISTORICAL DEVELOPMENT, APPLICATIONS, AND FUTURE RESEARCH DIRECTIONS. Transactions of the ASABE, 2007. **50**(4): p. 1211-1250.

[12] Winchell, M., et al., AcSWAT interface for SWAT2005 - User's guide, 2007. p. 436.

[13] George, C. and L.F. Leon, WaterBase: SWAT in an Open Source GIS. The Open Hydrology Journal, 2008. **2**: p. 1-6.

[14] Pai, N., D. Saraswat, and R. Srinivasan, Field_SWAT: A tool for mapping SWAT output to field boundaries. Computers & Geosciences, 2012. **40**(0): p. 175-184.

[15] Beniston, M., et al., Obstacles to data access for research related to climate and water: Implications for science and EU policy-making. Environmental Science & Policy, 2012. **17**(0): p. 41-48.

[16] Giuliani, G., et al., Sharing Environmental Data through GEOSS. International Journal of Applied Geospatial Research, 2011. **2**(1): p. 1-17.

[17] McKee, L., 18 reasons for open publication of geoscience data, 2010, Earthzine. p. 1-8.

[18] Open Geospatial Consortium, The Havoc of Non-Interoperability, 2004. p. 7.

[19] Ames, D.P., et al., Introducing the Open Source CUAHSI Hydrologic Information System Desktop Application (HIS Desktop). 18th World Imacs Congress and Modsim09 International Congress on Modelling and Simulation, 2009: p. 4353-4359.

[20] Open Geospatial Consortium, OpenGIS Web Map Server Implementation Specification, 2006. p. 85.

[21] Open Geospatial Consortium, Web Feature Service Implementation Specification, 2005. p. 131.

[22] Open Geospatial Consortium, Web Coverage Service (WCS) Implementation Specification, 2006, 143.

[23] Open Geospatial Consortium, OpenGIS Web Processing Service, 2007. p. 87.

[24] Open Geospatial Consortium, OpenGIS Catalogue Services Specification, 2007. p. 218.

[25] GEO secretariat, Global Earth Observation System of Systems 10-Year Implementation Plan Reference Document, 2005. p. 209.

[26] European Commission, Directive 2007/2/EC of the European Parliament and the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), 2007: Brussels. p. 14.

[27] Kiehle, C., K. Greve, and C. Heier, Standardized geoprocessing - Taking Spatial Data Infrastructures one Step Further, in 9th AGILE Conference on Geographic Information Science2006: Visegrad. p. 273-282.

[28] Michaelis, C.D. and D.P. Ames, Evaluation and Implementation of the OGC Web Processing Service for Use in Client-Side GIS. Geoinformatica, 2009. **13**(1): p. 109-120.

[29] Granell, C., et al., Interlinking Geoprocessing Services. Second International Conference on Advanced Geographic Information Systems, Applications, and Services: Geoprocessing 2010, Proceedings, 2010: p. 99-104.

[30] Granell, C., M. Gould, and M.A. Esbri, Geospatial Web Service Chaining, in Handbook of Research on Geoinformatics, H.A. Karimi, Editor 2009. p. 9.

[31] Kiehle, C., Business logic for geoprocessing of distributed geodata. Computers & Geosciences, 2006. **32**(10): p. 1746-1757.

[32] Schaffer, B. and T. Foerster, A client for distributed geo-processing and workflow design. Journal of Location Based Services, 2008: p. 194-210.

[33] Schaeffer, B., Towards a Transactional Web Processing Service (WPS-T), in GI Days2008: Münster, Germany. p. 27.

[34] Diaz, L., C. Granell, and M. Gould, Case study: Geospatial processing services for web-based hydrological application, in Geospatial Services and Applications for the Internet2008. p. 31-47.

[35] Friis-Christensen, A., et al., Designing Service Architectures for Distributed Geoprocessing: Challenges and Future Directions. Transactions in GIS, 2007. **11**(6): p. 799-818.

[36] Granell, C., L. Diaz, and M. Gould, Distributed Geospatial Processing Services, in Encyclopedia of Information Science and Technology2009, Information Science Reference. p. 1186-1193.

[37] Guru, S., M., et al., Challenges in using scientific workflow tools in the hydrology domain, in World IMACS/MODSIM Confress2009: Cairns, Australia.

[38] Fleuren, T. and P. Muller. BPEL workflows combining standard OGC Web services and grid-enabled OGC Web services. in 2008 34th Euromicro Conference Software Engineering and Advanced Applications (SEAA). 2008. Parma, Italy: Ieee.

[39] de Jesus, J., et al., WPS orchestration using the Taverna workbench: The eScience approach. Computers & Geosciences, 2012. **47**(0): p. 75-86.

[40] Lecca, G., et al., Grid computing technology for hydrological applications. Journal of Hydrology, 2011. **403**(1-2): p. 186-199.

[41] Ray, N., et al., Distributed Geocomputation for Modeling the Hydrology of the Black Sea Watershed, in Environmental Security in Watersheds: The Sea of Azov, V. Lagutov, Editor 2012, Springer. p. 141-157.

[42] Cepek, A., Proceedings of the Workshop Geoinformatics FCE CTU. 2008. **3**.

[43] Horsburgh, J.S., et al., Components of an environmental observatory information system. Computers & Geosciences, 2011. **37**(2): p. 207-218.

[44] Kao, S., et al., Visualisation of hydrological observations in the water data transfer format. Environmental Modelling & Software, 2011. **26**: p. 1767-1769.

[45] Horsburgh, J.S., et al., An integrated system for publishing environmental observations data. Environmental Modelling & Software, 2009. **24**(8): p. 879-888.

[46] Kanwar, R., U. Narayan, and V. Lakshmi, Web service based hydrologic data distribution system. Computers & Geosciences, 2010. **36**(7): p. 819-826.

[47] Huang, M.T., D.R. Maidment, and Y. Tian, Using SOA and RIAs for water data discovery and retrieval. Environmental Modelling & Software, 2011. **26**(11): p. 1309-1324.

[48] Goodall, J.L., B.F. Robinson, and A.M. Castronova, Modeling water resource systems using a service-oriented computing paradigm. Environmental Modelling & Software, 2011. **26**(5): p. 573-582.

[49] Tarboton, D.G., et al., Development of a Community Hydrologic Information System. 18th World Imacs Congress and Modsim09 International Congress on Modelling and Simulation, 2009: p. 988-994.

[50] Jiang, J., et al., The research and application of WFS based GML, in XXIst ISPRS Congress2008: Beijing, China.

[51] Ames, D.P., et al., HydroDesktop: Web services-based software for hydrologic data discovery, download, visualization, and analysis. Environmental Modelling & Software, 2012. **37**(0): p. 146-156.

[52] Salas, F.R., et al., Crossing the digital divide: an interoperable solution for sharing time series and coverages in Earth sciences. Nat. Hazards Earth Syst. Sci., 2012. **12**(10): p. 3013-3029.

[53] Peckham, S.D. and J.L. Goodall, Driving plug-and-play models with data from web services: A demonstration of interoperability between CSDMS and CUAHSI-HIS. Computers & Geosciences, (0).

[54] Nativi, S., et al., Biodiversity and climate change use scenarios framework for the GEOSS interoperability pilot process. Ecological Informatics, 2009. **4**(1): p. 23-33.

[55] Feng, M., et al., Prototyping an online wetland ecosystem services model using open model sharing standards. Environmental Modelling & Software, 2011. **26**(4): p. 458-468.

[56] Nativi, S., P. Mazzetti, and G.N. Geller, Environmental model access and interoperability: The GEO Model Web initiative. Environmental Modelling and Software, 2013. **39**: p. 214-228.

[57] Papajorgji, P., A plug and play approach for developing environmental models. Environmental Modelling & Software, 2005. **20**(10): p. 1353-1357.

[58] Castronova, A.M., J.L. Goodall, and M.M. Elag, Models as web services using the Open Geospatial Consortium (OGC) Web Processing Service (WPS) standard. Environmental Modelling & Software, 2013. **41**(0): p. 72-83.

[59] Castronova, A.M., J.L. Goodall, and M.B. Ercan, Integrated modeling within a Hydrologic Information System: An OpenMI based approach. Environmental Modelling and Software, 2013. **39**: p. 263-273.

[60] Nativi, S., M. Craglia, and J. Pearlman, The brokering approach for multidisciplinary interoperability: a position paper. International Journal of Spatial Data Infrastructures Research, 2012. **7**: p. 1-15