Scalable Computing: Practice and Experience

Scientific International Journal for Parallel and Distributed Computing

ISSN: 1895-1767

Universitatea de Vest
din Timişoara

Volume 19(2) June 2018

SUBSCRIPTION INFORMATION: please visit http://www.scpe.org

Scalable Computing: Practice and Experience

Volume 19, Number 2, June 2018

TABLE OF CONTENTS

Special Issue on E-Infrastructures for Excellent Science: Advances in Life Sciences, Digital Cultural Heritage and Climatology:

DOI 10.12694/scpe.v19i2.1401
Scalable Computing: Practice and Experience ISSN 1895-1767
Volume 19, Number 2, pp. iii–iv. http://www.scpe.org

INTRODUCTION TO THE SPECIAL ISSUE ON E-INFRASTRUCTURES FOR EXCELLENT SCIENCE: ADVANCES IN LIFE SCIENCES, DIGITAL CULTURAL HERITAGE AND CLIMATOLOGY

It is our pleasure to present this special issue of the scientific journal Scalable Computing: Practice and Experience. In this issue (Volume 19, No 2 June 2018), we have selected 14 papers which have gone through peer review and represent novel results in the fields of Life Sciences, Digital Cultural Heritage and Climatology, using state-of-the-art e-infrastructures, regionally integrated under the framework of the VI-SEEM project, and the related Virtual Research Environment (VRE). e-Infrastructures offer state-of-the-art IT resources which are the foundation that supports the scientific excellence in contemporary research. The VI-SEEM VRE integrates resources across all layers of the e-Infrastructure (networking, computing, data, software, user interfaces) to foster scientific excellence in selected fields and encourage cross-disciplinarity as well.

In the content of this special issue the papers are ordered thematically in 4 groups: Climatology (5 papers), Life Sciences (3 papers), Digital Cultural Heritage (2 papers) and Tools and Services (4 papers).

Climatology. The first paper presents an online interactive platform that aims to provide weather information about Armenia by integrating observations, model and satellite data. The topic is interesting from the practical point of view and might be very useful, especially for meteorologists.

The second papers studies the effect of the dust on climate in the Caucasus region, with a specific focus on Georgia, using the Regional Climate Model RegCM interactively coupled to a dust model. The simulations cover the period 1985-2014 encompassing most of the Sahara, the Middle East, the Great Caucasus with adjacent regions.

The third paper provides insight in the performances of wind simulations for high resolution models of the terrain. The presented results rationalize the possibility to run in reasonable time high resolution models, while showing that the impact of turbulence does not have significantly increases the computing requirement.

The fourth paper presents adaptation and tuning of the RegCM model for the Balkan Peninsula and Bulgaria and development of a methodology able to predict possible changes of the regional climate for different global climate change scenarios and their impact on spatial/temporal distribution of precipitation, hence the global water budgets, to changes of the characteristics and spatial/temporal distribution of extreme, unfavorable and catastrophic events.

The fifth paper presents comparison of two approaches (static and dynamical) used to compute the vibrational spectra of two conformers of the free formic acid molecule. The topic is interesting within the context of the atmospheric chemistry research field and it is of sufficient importance regarding the vibrational spectroscopic data and induced temperature effects of intramolecular motions.

Life Sciences. The manuscript from Astsatryan et al. describes a platform, which consists of data repository and workflow management services for Molecular Dynamics simulations. The platform focuses on an interactive data visualization workflow service as a key to perform more in-depth analyzing of research data outputs.

The manuscript from Bigovic et al. describes the organic synthesis of three enol carbonate derivatives and the analysis of their interactions with T4 lysozyme L99A/M102Q using molecular dynamics (MD) simulations. The results obtained by different software packages are discussed.

The manuscript from Koteska et al. describes a semi-empirical Molecular Dynamics study of irinotecan, a colon cancer drug, using the atom-centered density matrix propagation approach. The described methodology was used to study the structure, dynamics, and rovibrational spectrum of irinotecan.

Digital Cultural Heritage. The paper of Charalambous and Artopoulos presents the deployment of the Clowder CMS system and the development of extraction services to handle, manage and automatically process Digital Cultural Heritage data in order to enable virtual collaboration for research in the South East and Eastern Mediterranean region. Technical descriptions of the system are given and some results are provided.

In the paper of Elfarargy and Rizq a software system called Virtual Museum Framework (VirMuF), which is a set of tools that can be used by non-developers to easily create and publish 3D virtual museums in a very short time is presented. VirMuF is an open-source and teams including software developers can further extend VirMuF to fit their needs.

Software Tools and Services. Dimitrov and Stoyanov present the Data Discovery Service supporting the VI-SEEM project Virtual Research Environment - VRE. The solution is based on an open source platform with special customization regarding the data harvesting methods from diverse data sources and updating the available content so that the users will seamlessly access all the data from a single point.

The paper of Golubev et al. addresses the problems of optimization of medical image storing and secure access, using the DICOM system. Based on the Moldova DICOM Network architecture, the system enables distributed search, and transportation of DICOM images. Additionally, several optimization problems are addressed by the authors, along with the integration challenges within the VI-SEEM VRE.

In the paper of Mishev et al. the design, requirements and implementation of a federated virtual research environment, based on the service orientation paradigm, offering anything as a service solutions, have been considered. The challenges of the service management implementation focusing on interoperability by design and service management standards have been discussed.

The manuscript of Vudragovic et al. gives an extensive insight of the development and implementation of the DREAM dust model (DREAMCLIMATE service). Additionally, a use-case study of the premature mortality due to the desert dust in the North Africa - Europe - Middle East region for the 2005 obtained by the application of this model is presented, justifying the model and the applicability of the service itself.

We would like to thank all those who kindly contributed to this Special Issue: the authors who submitted their papers, reviewers for their help and proposed improvements, especially to Dr. Zoe Cournia, Dr. Theodoros Christoudias and Dr. George Artopoulos for their valuable remarks and suggestions. Our special gratitude is for the Editor-in-Chief, Professor Dana Petcu, for her constant support.

Aneta Karaivanova, IICT-BAS, Bulgaria
Anastas Mishev, UKIM, FYR of Macedonia

DOI 10.12694/scpe.v19i2.1351
Scalable Computing: Practice and Experience ISSN 1895-1767
Volume 19, Number 2, pp. 79–86. http://www.scpe.org © 2018 SCPE

WEATHER DATA VISUALIZATION AND ANALYTICAL PLATFORM

HRACHYA ASTSATRYAN, HAYK GRIGORYAN, ELIZA GYULGYULYAN, ANUSH HAKOBYAN, ARAM KOCHARYAN, WAHI NARSISIAN, VLADIMIR SAHAKYAN, YURI SHOUKOURIAN, ARTUR MKOYAN,[*] RITA ABRAHAMYAN, ZARMANDUKHT PETROSYAN [†] AND JULIEN ALIGON [‡]

Abstract. This article aims to present a web-based interactive visualization and analytical platform for weather data in Armenia by integrating the three existing infrastructures for observational data, numerical weather prediction, and satellite image processing. The weather data used in the platform consists of near-surface atmospheric elements including air temperature, pressure, relative humidity, wind and precipitation. The visualization and analytical platform has been implemented for 2-m surface temperature. The platform gives Armenian State Hydrometeorological and Monitoring Service analytical capabilities to analyze the in-situ observations, model and satellite image data per station and region for a given period.

Key words: Weather data, OLAP, web-based visualization, observational data, numerical weather prediction, satellite image processing.

AMS subject classifications. 68M14, 76M25

1. Introduction.

Armenia occupies the north-eastern part of Armenian plateau and central part of Lesser Caucasus range (latitude 38.51' to 41.18' North, longitude 43.29' to 46.37' East), with the area of about 30 000 sq.km. The geographical location of Armenia and complex mountainous relief has led to the diversity of natural conditions across the country. Armenia is on the northern edge of the sub-tropical zone, in latitudes characterized by an arid and continental climate. Due to a mountainous relief, different climatic zones exist and the weather may have high spatial gradients. High fluctuations in annual and daily temperatures are typical for the Armenian climate. The presence of six climatic zones from dry subtropical to rigorous high mountainous and from everlasting snowcaps to warm humid subtropical forests and humid semi-desert steppes make additional challenges on weather forecasting and climate prediction for the Armenian State Hydrometeorological and Monitoring Service (AHMS).

The meteorological data, received from 47 meteorological stations, serves as an input for the global atmospheric models to produce weather forecasts at the global scale. Only four stations provide historical data and monthly updates to the Global Climate Observing System Surface Network and three meteorological stations provide synoptic data to the gridded analysis dataset.

The observation data received from the meteorological stations and data received from a global model used as inputs and outputs to the high-resolution numerical weather prediction models to produce outputs of temperature, precipitation, and other meteorological elements from the ground to the top of the atmosphere [1, 2].

The Three-Dimensional Variational (3DVAR) data assimilation method is used to combine all available information on the atmospheric state in a given time-window to generate an estimate of atmospheric conditions valid at a prescribed analysis time [3]. Sources of information used to produce the analysis include observations, previous forecasts (the background or first-guess state) and satellite images. Currently, satellite imagery is used for the future experiments, as the availability of reasonable data over regions, where observations are scarce, is crucial to increase the accuracy of numerical prediction. The high-performance computational (HPC) resources of the Armenian e-infrastructure are used to resolve mesoscale weather events better and hence to give reasonably accurate forecasts in a short range [4, 5].

This article aims to present the weather data interactive web-based visualization and analytical platform[1]. The platform has been developed for the weather data in Armenia by integrating the three existing platforms

[*] Institute for Informatics and Automation Problems of the National Academy of Sciences of the Republic of Armenia, 1, P. Sevak str., 0014 Yerevan, Armenia (hrach@sci.am).

[†] Armenian State Hydrometeorological and Monitoring Service, 09/8 A. Mikoyan Str. 4th Block of Davitashen, 0054 Yerevan, Armenia.

[‡] University of Toulouse, 41 Allés Jules Guesde - CS 61321, Toulouse, France.

[1] Weather data interactive web-based visualization and analytical platform: http://meteo.grid.am

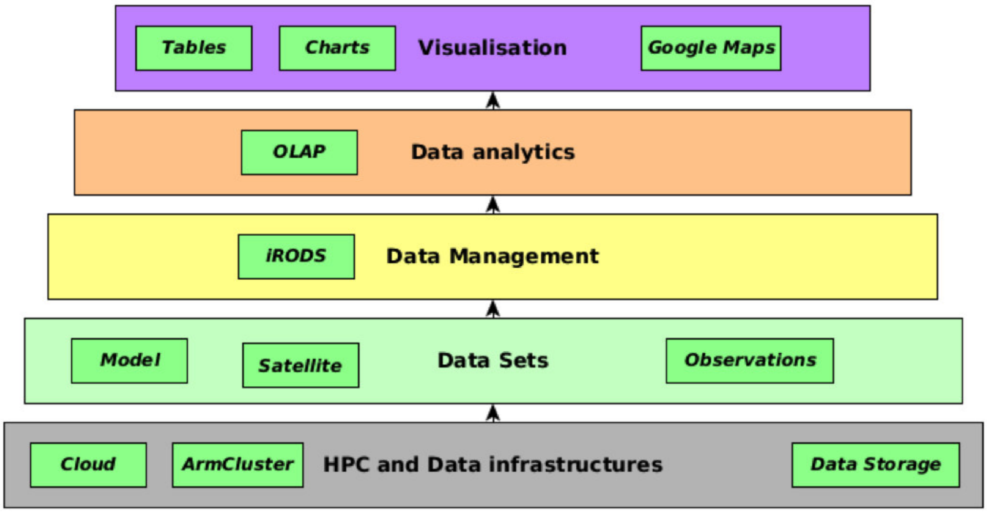


Fig. 2.1. *The framework of the platform.*

for observational data, numerical weather prediction and satellite image processing. The platform provides a way to compare the output of forecasting model with the observation data gathered from different stations for a chosen frame of time. The suggested platform is essential for a wide range of applications, such as urban area management, sustainable development and nature protection, regional and local planning, agriculture, forestry and fisheries, health, civil protection, infrastructure, transport and mobility or tourism.

The remainder of this paper is divided into the following sections: section 2 introduces the infrastructure, section 3 represents the discussions and analyzes and finally, section 4 is the conclusion.

2. Infrastructure. The suggested interactive web-based visualization and analytical platform consist of 5 main layers (see Fig. 2.1). The bottom layer provides HPC and data resources, which is especially important for the digital models and satellite image processing [6]. The resources of the Armenian e-infrastructure are used, which is a complex national IT infrastructure consisting of both communication and distributed computing infrastructures.

The datasets layer combines three types of data platforms for further analysis:

- Model output: outputs of weather prediction models;
- Satellite images: multispectral satellite images covering the territory of Armenia;
- In-situ data: meteorological stations observations, as a base to analyze the deviation values with other model outputs and satellite images.

The Data management layer provides intelligent tools to transfer raw data to data analytics layer. The Integrated Rule-Oriented Data System (iRODS) provides a middleware between the physical data storage systems and the user interface [7].

As soon as data reaches data analytics layer, it is processed and only several indexes are left from huge amount of initial raw data. Finally, the top layer and final destination of already processed data is visualization layer, where the outcome indexes are transformed to more user-friendly graphs or tables. Moreover, the advantages of Google Maps are used to map these indexes with their real location on the map.

2.1. Datasets. Observational datasets provide from various weather stations obtained with codes SYNOP (surface station reports observations). SYNOP reports are typically sent every three hours, which consists of groups of numbers describing general weather information, such as the temperature, sea level pressure, visibility, wind direction and speed, etc.

The numerical weather prediction models are initialized using NCEP (National Centers for Environmental Prediction) Global Forecast System analysis and forecasts at 0.5 deg horizontal resolution [8]. Data produced

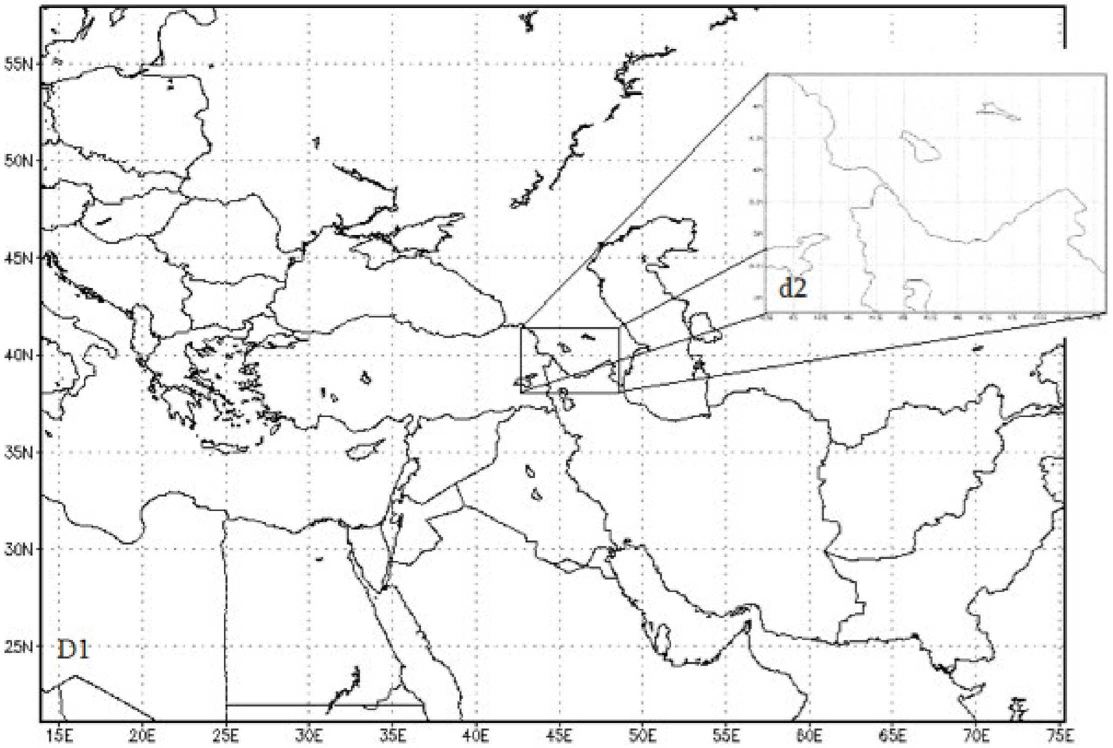


Fig. 2.2. *Parent (D1) and nested (d2) domains using in the models.*

during pre-processing and simulations of the models are in the Lambert conformal projection, which is well-suited for mid-latitude domains.

The model's setup (see Fig. 2.2) consists of a parent D1 domain (with a common center located at longitude 44.7, latitude 40.0) covering partly of Europe and all the Caucasus and parts of Central Asia and the Middle East and the nest domain d2 covering the whole territory of Armenia. The model uses 1-way nesting strategy and vertical 31 eta_levels.

Earth surface temperature, including land surface temperature (LST), is an important parameter reflecting earth surface environment and is widely used for climate change and weather forecasting. Landsat imagery is used, which supplies high-resolution visible and infrared imagery, with thermal imagery and a panchromatic image also available from the ETM+ (Enhanced Thematic Mapper Plus) sensor [9]. The single-channel method is used for LST retrieval. This method employs only the single thermal band of satellite imagery. This method is suitable for sensors that have only one thermal band such as Landsat TM/ETM+. The data acquisition dates had highly clear atmospheric conditions, and the images were acquired through the United States Geological Survey Earth Explorer Data Center, which has corrected the radiometric and geometrical distortions of the images to a quality level of 1G before delivery. The images are available in the GeoTiff format allowing to reformat, re-project and easily perform operations. Two images per day for daytime and nighttime is available covering the territory of Armenia.

2.2. Tools and models. As a geographic information system (GIS) software suite, the Geographic Resources Analysis Support System (commonly termed GRASS GIS) is used for satellite image processing, producing graphics and maps, spatial and temporal modeling, and visualization [10]. GRASS GIS is currently used in academic and commercial contexts around the world, as well as in many governmental agencies and environmental consulting companies. It can handle raster, topological vector, image processing, and graphic data. GRASS GIS contains over 350 modules to render maps and images on monitor and paper; manipulate raster and vector data including vector networks; process multispectral image data; and create, manage and store spatial data.

In the beginning, the workspace containing the following information is set up for every GRASS GIS session:

- Location - defines the projection, default spatial extent, and resolution for all data in the project;
- Mapset - defines collections of data within a given location. Mapsets can be used to organize data of similar types or categories;
- Database - defines where in the file system the files for the GIS will be stored for the given location and mapset.

As soon as all the images are the same GeoTiff type the created workspaces are identical and the same ESPG geodetic parameter dataset is used. After the workspace is created, "r.in.gdal" is used to import raster image to the workspace. Than by using r.mapcalc the LST index is calculated for every point on the map to allow quickly retrieve LST for a given coordinates by using "r.what" module. By implementing these steps we get the LST as an output.

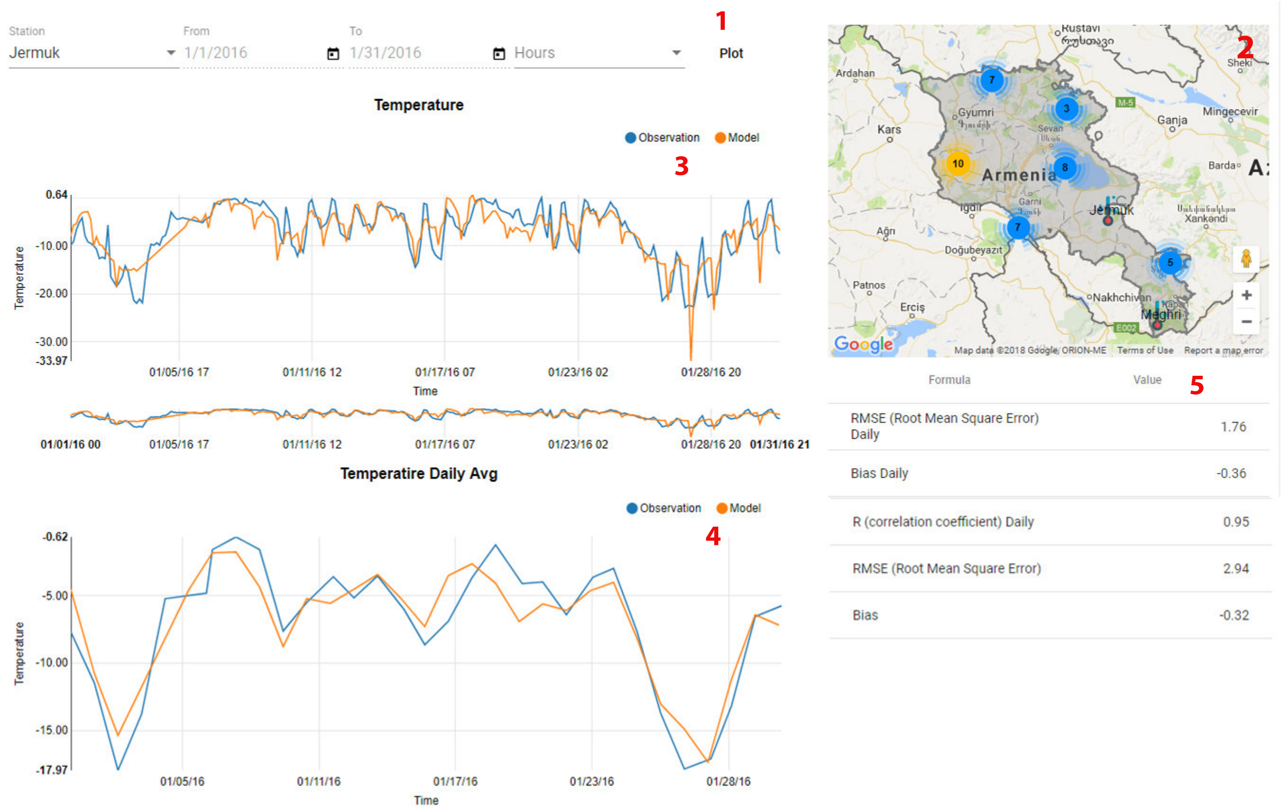
The mesoscale Weather Research and Forecasting (WRF) model [11, 12], which is adapted for the territory of Armenia, is used for operational weather forecasting (WRF-ARW version 3.6). The WRF model has become one of the world's most widely used numerical weather prediction models. Designed to serve both research and operational needs, it has been grown to offer a spectrum of options and capabilities for a wide range of applications. The WRF model is initialized using NCEP (National Centers for Environmental Prediction) Global Forecast System analysis and forecasts at 0.5 deg horizontal resolution. USGS (US Geological Survey) 30 arc-second digital topography database is used to interpolate the topography and land use. The Lambert conformal projection is used, as it is well suited for the mid-latitude domains. The model operates on two grids, the first parent domain covers the major part of Europe and all the Caucasus and some parts of the Central Asia and the Middle East (40.0_ N, 44.7_ E) with 202x202 grid points at 18-km, the second nest domain covers the whole territory of Armenia with 6-km horizontal resolution and 97x70 grid points. The following physical parameterization schemes are used in the model:

- microphysics; WRF Single-Moment 6-class (WSM6);
- radioactive processes: RRTM/ Dudhia;
- the surface layer: Eta similarity based on Monin-Obukhov with Zilitinkevich;
- the processes on the underlying surface and in the soil: Noah Land Surface Model;
- the planetary boundary layer: Mellor-Yamada-Janjic;
- cloudiness parameterizations: Kain-Fritsch.

Due to the limited computational resources, the inner domain with 6 km resolution has been implemented. As a range of every run, 24hours always starting at 0000 UTC of each day has been considered. Observable 2-meters long temperatures from 42 operational stations in Armenia are used for data assimilation and to study the accuracy of forecasting air temperature by the model.

2.3. Data analytics. The everyday rapid growth of data and need in analyze of this data pushed the development of analytical processing tools. OLAP (Online analytical processing) is a model for accessing multidimensional data in data warehouses [13]. Data cubes and OLAP session are central concepts in OLAP. A data cube is a collection of facts and dimensions organizing the data of a data warehouse according to different analysis axes and aggregation measures. OLAP provides a set of operations (such as drill-down and slice-and-dice) that transform one multidimensional query into another, which provide high querying. OLAP queries are formulated as sequences called OLAP sessions. For analyzing data with the spatial and georeferenced components the Spatial OLAP (SOLAP) technology is used, which allows rapid and easy navigation within spatial databases and that offers many levels of information granularity, many themes, many epochs and many display modes synchronized or not: maps, tables and diagrams [14]. It allows tight integration of GIS and OLAP systems. A SOLAP system supports three types of spatial dimensions: the non-geometric spatial dimensions, the geometric spatial dimensions and the mixed spatial dimensions. During an OLAP session, the user analyzes the results of a query and, depending on the specific data, applies an operation to determine a new query that will give a better understanding of information.

2.4. Visualization. The implemented tool consists of 3 main logical blocks and the data circulates between these blocks. The first block is the Database. PostgreSQL with its PostGIS extension is used to store datasets with geometrical data. The next block is implementing data analytic logic. The special type of OLAP approach



Station: Jermuk
From: 1/1/2016
To: 1/31/2016
Hours
Plot

Formula	Value
RMSE (Root Mean Square Error) Daily	1.76
Bias Daily	-0.36
R (correlation coefficient) Daily	0.95
RMSE (Root Mean Square Error)	2.94
Bias	-0.32

Fig. 2.3. *Web-based visualization and analytical platform consists of the following sections: 1- query form, 2 - map, 3 - temperature chart, 4 - daily average temperature chart, 5 - coefficients table.*

is used (Spatial OLAP) to process required data. As a query language for OLAP, the MultiDimensional eXpressions (MDX) is used to interact and perform tasks with multidimensional databases (OLAP Cubes). Afterwards, the processed data is transferred to User Interface where user can create various graphs, tables and see the output on the map (see Fig. 2.3).

The main screen includes many parts which are marked with red numbers.

- Query form - user can select the station, start and end dates and the period (by default it takes all hours from 0 to 21) for requesting the needed plot information. After plot action the charts and the table will be updated with corresponding values;
- Map - highlights the territory of Armenia and with markers displayed the physical locations of the stations in the Earth coordinate system;
- Temperature and daily average temperature - displays the observation and model temperature data lines correspondingly for each period and the daily average;
- Coefficients table - shows the RMSE, BIAS and R correlation coefficients.

The platform provides the ability to upload new data to the database by using API endpoints. Currently only the administrator of the platform has access for this actions. For observation data SYNOP files (the records are in ascii format) must be used which will be parsed using script written in JavaScript language. For WRF model outputs which are generated in netCDF file format the Python script was created for finding corresponding values based on the stations information stored in the database (see Fig. 2.4).

For observation data the parser script was made with JavaScript for parsing SYNOP files (the records are in ASCII format).

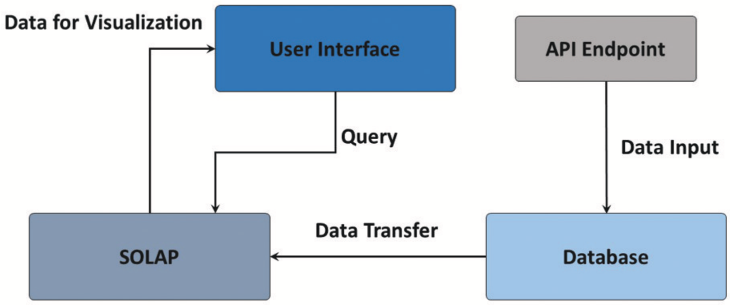


Fig. 2.4. *Data workflow consists of API endpoint, database, SOLAP and user interface steps.*

Table 3.1
Mean estimates of verification of temperature forecasts for several meteo stations per region.

Region	Station	Height above sea-level, m	RMSE	BIAS	R
4*Ararat	Armavir	870	3.97	-3.53	0.82
	Artashat	829	5.69	-5.10	0.73
	Ararat	818	5.65	-4.83	0.62
	Merdzavan	942	4.71	-4.11	0.79
2*Yerevan	Zvartnots	853	4.44	-3.91	0.82
	Arabkir	1113	3.24	-1.92	0.77
2*Syunik	Meghri	627	9.89	-7.37	-0.70
	Kapan	705	9.69	-7.81	-0.43
2*Tavush	Ijevan	732	7.61	-4.47	-0.40
	Bagratashen	453	7.56	-6.02	-0.31
2*Shirak	Ashotsk	2012	2.52	-1.10	0.91
	Gyumri	1513	3.91	-3.04	0.82
2*Gegharkunik	Gavar	1960	2.74	-1.61	0.89
	Lake Sevan	1917	2.59	1.74	0.90

3. Discussions and analyzes. Various statistical and object-oriented methods are used in the suggested web-based analytical platform to investigate the characteristics of model-forecast and satellite image errors, which is important for providing useful guidance to end-users.

As a case study, the 2m temperature has been investigated using the observational and regional high-resolution WRF model data for the January, 2016. For the studied period, the RMSE, BIAS and R correlation coefficients are calculated for the observational data and model-forecast data for 42 observation points distributed in the territory of Armenia (see Table 3.1). The table shows that the average difference between the forecast data and observations is about 4-5^{0}C. For all the stations studied, an acceptable correlation coefficient has been obtained, which allows one to judge the adequacy of the model.

The analyzes, which have been carried out for 42 stations (some high-altitude stations are not considered), show that the model data for almost all stations are overstated. RMSE values are about 1.6-2.5^{0}C for the stations located at altitudes above 1500m-2000m, 4.6-9.8^{0}C for the stations below 1000m. It means that the model predicts well the temperature values at a height of 2m for stations located above 1500m and gives unsatisfactory results for stations below 1500m. The worst results are obtained for stations below 1000m, such as stations located in Meghri, Kapan, Ijevan, Bagratashen, and the valley of Syunik and Tavush region.

The January 2016 temperature for Ararat valley and Yerevan is also projected unsatisfactory. In Ararat valley RMSE 5.6-6.4^{0}C, the worst result was obtained for Ararat station. In Yerevan, a good result was obtained for Yerevan-Arabkir station, which is located at an altitude of 1113m, and the worst result for Yerevan-Zvartnots. The correlation coefficient varies from 0.80-0.97 for mountain and foothill areas, as for valleys it is from 0.7 to 0.25 and from -0.7 to -0.31.

From all that has been described above, it can be concluded that the WRF model with the described configuration, for a cold period of time, gives a positive forecast for a variable air temperature of 2 m for the mountain and foothill regions of the republic (Shirak, Kotayk, Gegharkunik, Lori, mountain and foothill areas of Aragatsotn) and at the same time an unsatisfactory forecast for the Valley of Syunik, Tavush, Ararat valley and Yerevan.

Analysing all the factors that make up the temperature, we see that there were fogs in the valley during the selected period, which was the reason for low temperatures. Therefore we can conclude that the model WRF, is not predicted by low temperatures of occurrence of the lowland, effect of the fog at surface inversion. Such a result does not satisfy, therefore, it is necessary to improve the accuracy of the model data, through a proper tuning, it would be useful to test the impact of resolution given, because the Armenian terrain has a rather complex orography and is characterized by several land-category types.

4. Conclusion. The suggested platform enables to integrate already available observational, model-forecast and multispectral satellite images and use these data sources for studies and analyzes in a web-based visualization environment. The interactive comparison charts for 2m air temperature allows to visually analyze and gather the information about model accuracy. It enables to adjust the forecasting results with additional methods by implementing statistical analyzes and provides a fairly high result in cases where the model's sensitivity is low.

It is planned to improve the functionality of the platform by adding new visualization tools of various formats, such as to analyze and compare other near-surface atmospheric elements. Different nowcasting methodologies based on artificial intelligence and the utilization of satellite imagery will be implemented for the development of a hazardous hydro-meteorological phenomena alarm system. The extended platform will be integrated with the available cloud services [15, 16, 17] by providing access to the required specialized climatic data.

The ultimate goal is to develop an integrated web-based service, which can be used by AHMS for operational weather forecasting and for data analytics for scientific studies.

Acknowledgments. The research leading to these results has been co-funded by the European Commission under the H2020 Research Infrastructures contract no. 675121 (project VI-SEEM) and the State Target Program "Creating a cloud computing environment in order to solve scientific and applied problems".

REFERENCES

[1] A. Gevorgyan, *Summertime wind climate in Yerevan: valley wind systems*, Climate Dynamics, 48:5–6 (2017), pp. 1827–1840.

[2] Artur Gevorgyan, Hamlet Melkonyan, Rita Abrahamyan, Zarmandukht Petrosyan, Anna Shachnazaryan, Hrachya Astsatryan, Vladimir Sahakyan, Yuri Shoukourian, *A Persistent Surface Inversion Event in Armenia as Simulated by WRF Model*, in IEEE Proceedings of the International Conference on Computer Science and Information Technologies, CSIT'2015, pp. 105–110.

[3] C. Faccini, D. Cimini, R. Ferretti, F.S. Marzano, A.C. Taramasso, *3DVAR assimilation of SSM/I data over the sea for the IOP2b MAP case*, Adv Geosci, 2 (2005), pp. 229–235.

[4] Hrachya Astsatryan, Vladimir Sahakyan, Yuri Shoukourian, Pierre-Henri Cros, Michel Dayde, Jack Dongarra, Per Oster, *Strengthening Compute and Data intensive Capacities of Armenia*, in IEEE Proceedings of 14th RoEduNet International Conference - Networking in Education and Research, NER'2015, pp. 28–33.

[5] Yuri Shoukourian, Vladimir Sahakyan, Hrachya Astsatryan, *E-Infrastructures in Armenia: Virtual research environments* , in IEEE Proceedings CSIT 2013 - 9th International Conference on Computer Science and Information Technologies, Revised Selected Papers, CSIT'2013, pp. 1–7.

[6] H. Astsatryan, Yu. Shoukourian, V. Sahakyan, *The ArmCluster Project: Brief Introduction*, in Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications, PDPTA ?2004, pp. 1291–1295.

[7] M. Hedges, T. Blanke, A. Hasan, *Rule-based curation and preservation of data: A data Grid approach using iRODS.*, Future Gener. Comput. Syst, 25:4 (2009) , pp. 446–452.

[8] W. S. Jefrey , T. M. Hamill, X. S. Yucheng, Z. Toth, *Ensemble data assimilation with the ncep global forecast system*, Monthly Weather Review, 136 (2008), pp. 463–482.

[9] J. A. Sobrino, J. C. Jimenez-Munoz, L. Paolini, *Land surface temperature retrieval from Landsat TM 5*, Remote Sensing of Environment, 90(2004), pp. 434–440.

[10] M. Neteler, M. Bowman M. Landa M. Metz, *Grass gis: A multi-purpose open source gis*, Environmental Modelling & Software, 31(2011), pp. 124–130.

[11] William C. Skamarock, Joseph B. Klem, *A time-split non-hydrostatic atmospheric model for weather research and forecasting applications*, Computational Physics, 227–7(2008), pp. 3465–3485.

[12] J.G. Powers, J.B. Klemp,et. al, *The Weather Research and Forecasting Model: Overview, System Efforts, and Future Directions*, Bulletin of the American Meteorological Society, 98–8(2017), pp. 1717–1737.

[13] Surajit Chaudhuri, Umesh Dayal, *An Overview of Data Warehousing and OLAP Technology*, SIGMOD Record, 26–1(1996).

[14] S. Aissi, M.S. Gouider, T. Sboui, L.B. Said, *Enhancing spatial data warehouse exploitation: a solap recommendation approach*, In: Computer and Information Science, Springer (2016), pp. 131–147.

[15] Hayk Grigoryan, Hrachya Astsatryan, Tigran Gevorgyan, Vahe Manukyan, *Cloud Service for Numerical Calculations and Visualizations of Photonic Dissipative Systems*, Cybernetics and Information Technologies, 15–5(2017), pp. 89–100.

[16] H. Astsatryan, W. Narsisian, Sh. Asmaryan, *SWAT Hydrological Model as a DaaS Cloud Service*, Springer Earth Science Informatics, 9–3(2016), pp. 401–407.

[17] Hrachya Astsatryan, Vladimir Sahakyan, Yuri Shoukouryan, Michel Dayd, Aurelie Hurault, Ronan Guivarch, Arutyun Terzyan, Levon Hovhannisyan, *Services Enabling Large-Scale Linear Systems of Equations and Algorithms based on Integrated P-Grade Portlal*, Springer Journal of Grid Computing, 11–2(2013), pp. 239–248.

Edited by: Aneta Karaivanova
Received: Dec 31, 2017
Accepted: Mar 28, 2018

DOI 10.12694/scpe.v19i2.1398
Scalable Computing: Practice and Experience ISSN 1895-1767
Volume 19, Number 2, pp. 87–96. http://www.scpe.org © 2018 SCPE

EFFECT OF DUST AEROSOLS IN FORMING THE REGIONAL CLIMATE OF GEORGIA

TEIMURAZ DAVITASHVILI[*] NATO KUTALADZE[†] RAMAZ KVATADZE[‡] AND GEORGE MIKUCHADZE[§]

Abstract. The effect of the dust on the climate in the Caucasus region, with a specific focus on Georgia, was investigated with a Regional Climate Model RegCM interactively coupled with a dust model. For this purpose we have executed sets of 30 years simulations (1985-2014) with and without dust effects by RegCM4.7 model with 16.7 km resolution over the Caucasus domain and with 50 km resolution encompassing most of the Sahara, the Middle East, and the Great Caucasus with adjacent regions. Results of calculations have shown that the dust aerosol is an active player in the climate system of Georgia. Mineral dust aerosol influences on temperature and aerosol optical depth spatial and temporally inhomogeneous distribution on the territory of Georgia and generally has been agreed with MODIS satellite data. Results of numerical calculations have shown that dust radiative forcing inclusion has improved simulated summer temperature. The mean annual temperature increased across the whole territory of Georgia in simulations when dust direct effect was considered.

Key words: Climate change; Caucasus; Georgia; Dust; RegCM4.7.

AMS subject classifications. 68M14, 65C48

1. Introduction. Dust aerosols in the form of fine particles are scatter and absorb solar and terrestrial radiation and therefore are affect climate [1]. Besides, lifted up from the soils, rocks, plants, volcanic eruptions and anthropogenic pollutants into the atmosphere, dust aerosols can reduce evaporation and as a consequence precipitation processes by reducing the earth surface temperature [2]. Investigations have shown that anthropogenic activities on the average lead to 30 percent of the dust load whereas the storms are the major sources of mineral loading in the environment [3]. The world's biggest desert (Sahara and Sahel in Africa, the Gobi, Kyzylkum, Karakum, Taklamakan in central Asia) storms usually represent the primary sources for the mineral dust aerosols transfer in the atmosphere, its sediment on the earth surface and spreading across the Europe, Asia and Africa continents [1,4]. For instance, desert dust is the principal aerosol component over the Mediterranean basin and they strongly influence the Mediterranean climate [3]. For short or long time periods the dust storms are significantly affecting the earth's atmosphere quality, modifying the clouds microphysics, their optical properties and have a strong influence on both regional and global climate systems. Indeed, dust storms influence the atmospheric radiation budget, the ground surface albedo, the air quality and consequently the human health and the entire biota [5-7]. Numerical modeling of the past, present and future climate processes represents a good means to study the main factors affecting the modern climate change. Several attempts were made to examine the global or regional climate models ability and to clarify the dynamical and physical mechanisms responsible for climate change over the particular regions [3-11]. For instance, for the purpose of assessing the ability of regional climate model to simulate surface solar radiation patterns over Europe the RegCM4.4 model was used [11]. The results of calculations from 2000 to 2009 and their comparisons against the satellite-based observations have shown that the model slightly has overestimated surface solar radiation patterns in Europe.

At present, the simulations of dust cycle in the global and regional climatic models continue to be an important research area. In order to better understand the origin conditions of dust storms and their migration trends, a worldwide effort has been undertaken by numerous researchers to detect the dust aerosols sources into the main and accessory regions, based on the meteorological monitoring networks and satellite observations, in the last decades [12]. As atmospheric aerosols have substantial impacts on the Earth's climate through their direct, semi-direct and indirect effects, the inclusion of aerosol processes (evolution during transportation, deposition, chemical, physical and optical properties) is essential in global and regional climate models. Thus, the study of the dust aerosols life cycle, migration and dust-climate interaction processes by numerical climate models with dust modules currently are widely appreciated in numerous studies [13-17]. Currently climatic

[*] Faculty of Exact and Natural Sciences, I.Vekua Institute of Applied Mathematics of Iv. Javakhishvili Tbilisi State University (teimuraz.davitashvili@tsu.ge)

[†] Hydrometeorological Department/ National Environmental Agency of Georgia

[‡] Georgian Research and Educational Networking Association GRENA

[§] Hydrometeorological Department/ National Environmental Agency of Georgia impacts of aerosols at the global scale are relatively well understood. However, a number of uncertainties still exist in understanding these effects at regional scales [13-15] [16]. It's known that the finest desert dust particles at the near-surface wind threshold conditions are lifted up to high altitudes of troposphere and then are transported thousands of kilometers from the source regions [18]. That is why the effects of the desert dust on climate can be felt not only locally but also in regions far from the sources [14]. Dust aerosol influences and responds to climate change mainly due to its extinction in shortwave radiation, that leads to surface cooling, especially over the arid and semi-arid areas [17].

In order to understand aerosol impacts on climate and environment, a new dust aerosol scheme including emission, transport, gravitational settling and optical property calculations were implemented and tested by the Regional Climate Model (RegCM) [14]. The RegCM/Dust model was run from episodic (few days) to seasonal (climate mode) periods and the model was able to simulate the occurrence of strong dust outbreaks in different regions and to capture the main dust load areas over the Sahel [14]. The coupled chemistry aerosol regional climate model RegCM was used to investigate the dust emission size distribution impact on aerosol budget and its radiative forcing over the Mediterranean region [13]. RegCM-Dust model, with 30 km horizontal resolution, had been used for estimation of the direct radiative forcing by mineral dust aerosols over the Indian subcontinent throughout 2009 [19]. The results of calculations have shown that the model was able to capture the seasonality of dust emissions, they were reasonably well transported and distributed from the major sources across the Indo-Gangetic Basin to Himalayan foothills and have 700-850 hPa. The simulations' results of 38 summer monsoon seasons (1969-2006) has shown reduction of average precipitation over the Sahel region executed by the RegCM model with and without dust effects over the African continent. Numerical calculations also have shown that inclusion of dust module into the model has improved the West African monsoon simulation quality [20].

Several attempts were made to examine the RegCM/Dust coupled models' ability for the particular regions [21-23]. For example, the RegCM/Dust coupled model was capable to simulating dust seasonal transport from Sahara towards the South and Central America appropriately [21]. Topography-Modulated dust aerosol distribution and its effects on the atmospheric heat source over the Tibetan Plateau, East Asian summer monsoon onset and prediction of precipitation under different terrain settings in East Asia were successfully studied by RegCM4/Dust coupled model [22]. The results of calculations have shown that the dust greatly increments in the Taklamakan desert (accompanied with the uplift at the northern Tibetan Plateau) and greatly suppresses precipitation in East Asia. Seasonal mean air temperature (C°) and precipitation (mm/day) for the three periods of 2011-2040, 2041-2070 and 2071-2100, with respect to the control period of 1971-2000 above the Central Asia domain was studied by RegCM4.3 model [23]. The results of calculations have shown high rate of warming in the warm season with a decrease in precipitation in almost all parts of the domain and warming trend especially for the northern part of the domain during the cold season [23].

So far, no attempts have been made to study the effect of dust on climate change in the Caucasus region, with particular attention to the territory of Georgia, by regional climate models including dust module. Although, there are several publications on this topic [24-27]. The impact of dust deposition on the Caucasus glacial environment has recently attracted attention of scientists due to the accelerated melting of glaciers in the Caucasus region [24-26]. Generally, dust deposited on glaciers originates from the products of decay of biogenic material, locally-produced mineral dust and long-travelled desert dust [25-26]. For example, a significant desert dust deposition event occurred on Mt. Elbrus (Caucasus Mountains, Russia) on 5th of May 2009, where the deposited dust later appeared as a brown layer in the snow pack which originated in the foothills of the Djebel Akhdar in eastern Libya. The dust sources were activated by the intrusion of cold air from the Mediterranean Sea and Saharan low pressure system and transported to the Caucasus along the eastern Mediterranean coast, Syria and Turkey [27].

It's known that the dust aerosols have an indirect effect on the radiation through effecting cloud microphysics [28]. The most dust aerosols are settled in the atmospheric surface and planetary boundary layers where the atmosphere-surface exchanges energy and water takes place [29]. The dust particles influence microphysical and optical properties of the cloud by scattering and absorbing the short and long wave radiation [30]. The observations have shown that the dust aerosols effect cloud thermodynamic (temperature, relative humidity) and microphysics [31]. The dust aerosols modify cloud properties, the amount, size of cloud droplets and ice crystals [32-34]. Mineral dust aerosols are important components of the Earth's system and have influence on the cloud system [28,35] by reflecting solar energy and resulting in surface cooling [36].

The problem of the forthcoming global climate change resulting from natural and growing anthropogenic factors (economical and technological development, overexploitation of land, water, oil and gas resources) gain a particular importance for the territory of Georgia because of its location and compound orography. In Georgia there are 11 types of climate zones from semi-desert to subtropics including mountainous zone of Caucasus with constant snow and glacier. Climate temperature data for the last 100 years have shown climate cooling process in the western and climate warming in the eastern Georgia and also permanence in some micro regions of Georgia. It is necessary to find constantly acting thermal and advective-dynamic sources being responsible for this change. Especially interesting is the impact of increasing concentration of radiation gases, aerosols and dust (dust is one of the main pollutants of the territory of Georgia) on the regional climate, as their accumulation in the lower atmospheric layer plays the role of the scum, which intensifies solar warming of the atmosphere and considerably decreases long-wave flow directed from the Earth to the outer space.

In this study, the RegCM4.7 model coupled with a dust module is configured with a relatively high horizontal resolution (16,7 km) and used to simulate dust aerosol distribution and its effects on the climate in the Caucasus region. This article examines the role of dust (mineral aerosols) in the regional climate of Georgia by comparing two 30 year simulations executed with and without fully coupled radiatively interactive dust emissions. It was found that RegCM4.7-BATS and its dust model had simulated well the temporal and spatial distributions of mineral aerosols over the Georgia.

2. Model description and data. This study is based on the fourth generation regional climate modeling system RegCM4 [37], where the mineral dust particles' emission, transport and deposition are included [14]. The RegCM4 is a sigma regional climate model with a dynamical core based on the hydrostatic version of the PSU/NCAR Mesoscale MM5 Model [38]. The model has been widely used and tested for the study of regional climatic change and especially for simulations of the effect of dust aerosols in forming the regional climate [6,14,37]. The coupled dust module includes dust emission, transport (wet and dry removal), gravitational settling and optical properties' calculations. The dust emission scheme completely depends on the simulated surface wind threshold friction velocity value, boundary layer atmosphere processes and land surface characteristics (surface roughness and soil moisture) which are provided by the surface biosphere-atmosphere transfer scheme BATS. We examine the role of mineral dust effect in forming the regional climate of Georgia by RegCM4/dust model as dust represents the main pollutant for the territory of Georgia [25,39]. During the last decades, there has been a significant improvement in understanding of the dust sources, its transportation, properties and in modeling capabilities [39]. In our study the dust particles are divided into four size bins: fine (0.01-1.0 μm), accumulation (1.0-2.5 μm), coarse (2.5-5 μm), giant (5.0-20.0 μm) and we used four steps in dust parameterization [39]. The dust transport, deposition and removal processes have been described in detail in articles [40] [41] and were used in this study.

2.1. Dust Parameters Satellite Measurements. The modeled dust Aerosol Optical Depth (AOD) data have been examined against the Moderate Resolution Imaging Spectroradiometer's (MODIS) data with a 1° × 1° resolution. There are two MODIS sensors which are observing Earth from polar orbit of NASA's Terra (since February 2000) and Aqua (since June 2002) satellites [42-44]. For this reason, some simulations relating to AOD were executed over 15-year (2000-2014). The retrieved AOD outputs from both Terra and Aqua satellites were used in our study.

The contours of dust load volume concentration were retrieved from the CALIPSO monthly mean gridded (2° × 5°) outputs. The CALIPSO products gave us the aerosol extinction coefficient at 532 nm, column aerosol optical depth and aerosol layer properties in the global grid cells.

2.2. Meteorological data. The results of modeled climate characteristics (among them precipitations) executed by the RegCM4 model are significantly impacted by the boundary conditions (BCs). In our study the BCs for the RegCM4 model domain have been created from ERA-Interim the state-of-the-art global atmospheric reanalysis data which were developed by the European Centre for Medium Range Weather Forecasts [45]. It should be mentioned that observational data have been combined with modeled information from the previous time step in order to construct the global atmospheric conditions. The results of calculations have been validated against Climate Research Unit (CRU) data which present the gridded global climate database of

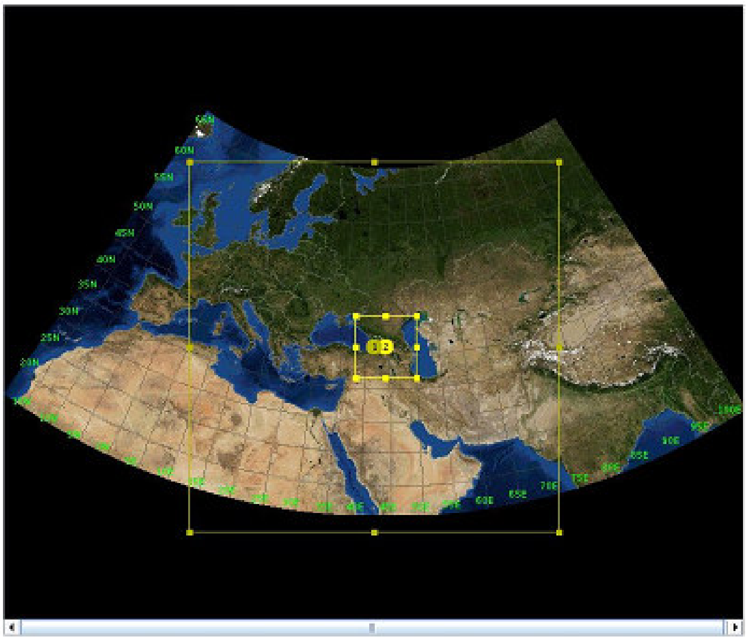


Fig. 3.1. *Location of the study area. Coarse and nested domains.*

monthly meteorological measurements from ground-based stations [46]. The surface measurements of temperature, among of six meteorological variables datasets, were interpolated into a 0.75× 0.75 grid that have covered the entire land surface of the planet.

3. Experiments design and method. Our investigation was concentrated on modeling the regional climate change of the territory of Georgia based on the latest version of RegCM4.7 using BATS (Biosphere-Atmosphere Transfer Scheme) surface code. The study is focused on the impact of locally-produced mineral dust aerosols and long-travelled desert mineral dust on Georgia's regional climate change. As mentioned above, the study over Georgia's territory, similarly to the whole Caucasus region, has a very fragmented character. These studies mostly have focused on strong dust events, when dust particles were observed in Abastumani (Georgia) and in Mt. Elbrus (Russia) [27,47]. The study of regular and long term dust impact on the regional climate has not been carried out for the territory of Georgia yet.

For simulation the period from 1985 to 2014 with boundary conditions from ECMWF ERA-Interim data (with 0.75 degree resolution) was selected. Our model has coarse domain with 50 km resolution, (it covers all of the regions, which mainly take part in the formation of the atmospheric processes over the Caucasus region, namely: the most of south and east Europe, Ural and Siberian Region, Middle East and Central Asia) and one nested domain (fully covering Caucasus region) with 16.7 km resolution – see Fig. 3.1.

The impact of dust on the regional climate was evaluated by comparing two numerical experiments in which the first was executed without dust and the second one was simulated through interactive dust inclusion. In order to explore the resolution effect each run was downscaled with nested simulation. For the coarse domains we use time step 100 sec. and for the nested ones - 30 sec. The coarse domains contain 128 grid points in each of the horizontal directions and 18 vertical levels and the nested domains - 64 and 18 correspondingly.

The same physical schemes were used for Dust and NoDust experiments:

– Holtslag PBL Boundary layer scheme;

– Tiedtke Cumulus convection scheme over the land and the ocean;

– Explicit moisture scheme; In the experiment with Dust only dust tracers - 4 dust bins scheme was activated and aerosol direct effects on radiation and dynamics of atmosphere were considered.

Simulations of the model without chemistry were performed on GRENA's (Georgian Research and Ed-

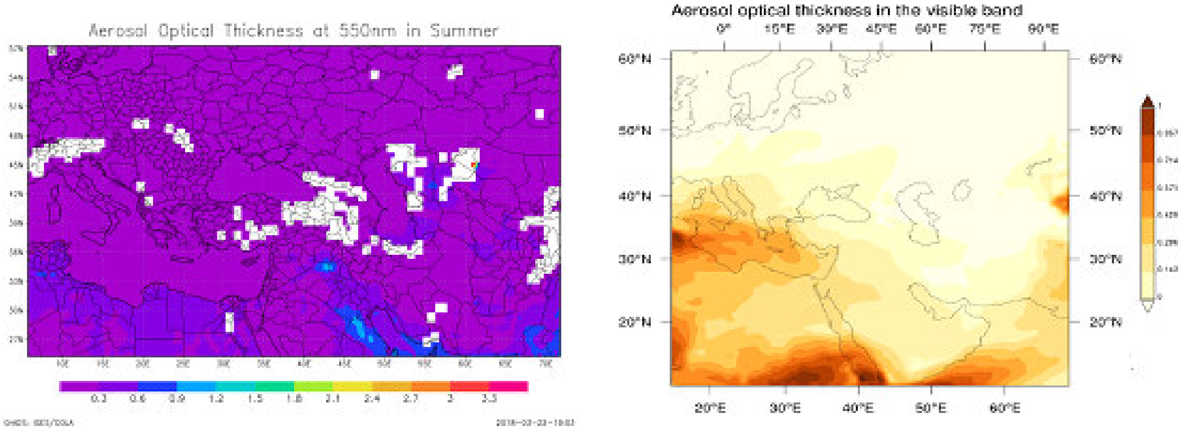


FIG. 4.1. *Observed (seawifs) and simulated AOD of summer season for 2000-2014 period.*

ucational Networking Association) cluster (one computing node with 15 cores of Intel Xeon CPU E5-2670 @2.60GHz.) It took approximately 135 hours for the coarse domain and 70 hours for the nested one. For the simulations with chemistry more powerful computing resource - high performance computing cluster ARIS of GRNET (Greek Research and Technology Network) was used. The model run was performed on the 60 cores (it corresponds to the 3 nodes) of Intel Xeon CPU E5-2680 v2 @ 2.80GHz. It took approximately 84 hours for the coarse and 60 hours for the nested domains. In order to investigate the usage effect of two different computational resources on model results the same simulations were performed on GRENA and ARIS clusters, the results of calculations were the same.

4. Results and discussions. On the first stage, simulation with dust was validated. As in South Caucasus region we have no ground base stratosphere aerosols observations the only source to examine dust concentration in the air is satellite derived data. The simulated AOD results (aerosol optical depth) were compared to MODIS (MISR, seawifs) data for the four seasons during the period 2000-2014. From the analyses of MODIS data value of the AOD is small in winter as this period of year is characterized by heavy snows that prevent dust accumulation and consequently its values over Caucasus reach only 0.1-0.2 aerosol optical depths observed at characteristic wavelength of 550 nm. The maximum values of the AOD occur in spring and at the beginning of summer season (March-June), when dust is uplifted and transported from the Sahara and Middle East across Mediterranean to the Black Sea's east coast and reaches the Caucasus regions [27]. Indeed, in spring due to activation of dust transportation from Libyan and Egyptian deserts, the circulation process values of AOD increases and it reaches over Caucasus 0.5-0.6 at 550 nm. Also, the observations have shown that the aerosol episodes are frequent during the dry period too, from June to October [42]. In contrast, AOD is minimal in winter. From June to October the subtropical Atlantic high (Azores) prevails over the black sea basin, enhances and causes subsidence. Thus, it results in an extremely stable atmosphere and in absence of rainfall, conditions that favor the aerosol accumulation in the atmosphere. These variations related to synoptic meteorological conditions were investigated previously by authors [23,27,42].

The seasonal distribution pattern of simulated AOD's agrees well with MODIS data. It should be noted that in spring and summer sessions the simulated dust concentrations and corresponding AOD in the dust storm generation regions are overestimated, but for the Caucasus region calculated AOD values are very close to satellite data. For comparison the observed and simulated AOD are presented for summer period on the Fig. 4.1.

To evaluate the impact of dust on the simulated 2 m temperature all of the 4 runs have been interpolated on nested domain's grid and compared to the 0.50-resolution Climatic Research Unit (CRU) surface temperature (for land only) for annual and seasonal scale. On Fig. 4.2 observed and modeled summer and winter mean

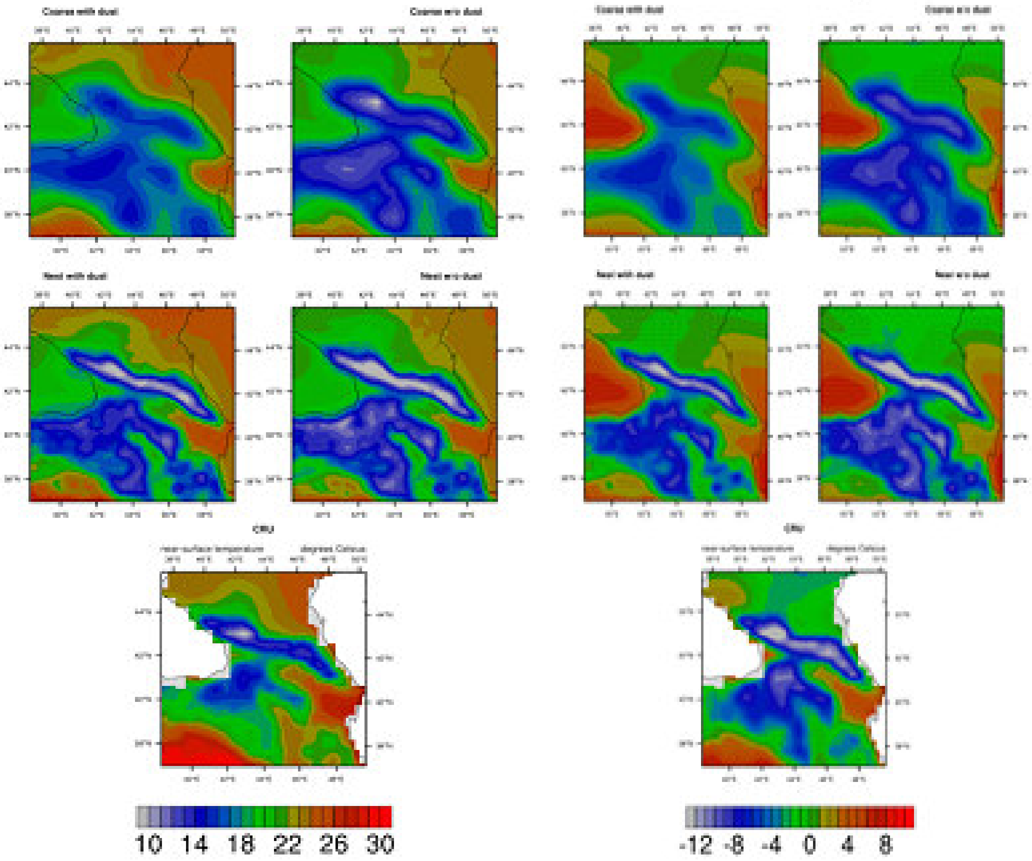


Fig. 4.2. *Summer and winter mean surface temperatures observed (CRU) and simulated (with Dust and NoDust on different resolutions).*

temperature plots are presented. From these plots difference between experiments with Dust (left on the Fig.) and NoDust is evident, as well as difference between spatial resolutions. It depends on different sub-regions and is in agreement with CRU data.

To examine the simulation performance across the experiments on different sub-regions of the South Caucasus nested domain was divided in 8 sub-regions. These regions mostly cover Georgia's territory but also include some other parts, according to the factors of local climate formation. On Fig. 4.3 location and names of sub-regions are presented, where SCC is Central part of South Caucasus, SWC - Western part of South Caucasus, SEC - Eastern part of South Caucasus, EP - Eastern plane territory, KP - Kolkhety Down land, AJ - South mountainous part of Ajara, JP - Javakhety Plato, CP - Central part of Georgia including Likhi range.

The mean annual, as well as summer and winter temperatures bias, standard deviation and correlation coefficient in comparison with CRU data have been calculated from all of the 4 simulations and mean values across mentioned 8 sub-regions evaluated (Table 4.1).

According to the Table 4.1, annual negative bias appears in all sub-regions for the NoDust simulation, the biggest one is in EP sub-region, bias from Nested NoDust run is even bigger for the most of sub-regions. Dust simulation has evident benefit in reduction of mentioned cold bias, and in the western sub-regions produces warm bias. Nested Dust run also improves performance. Namely for EP sub-region it continues the reduction of negative bias and smoothes bias in other regions produced from the course domain.

On the seasonal scale, Dust experiment improves winter mean temperature simulation for all sub-regions,

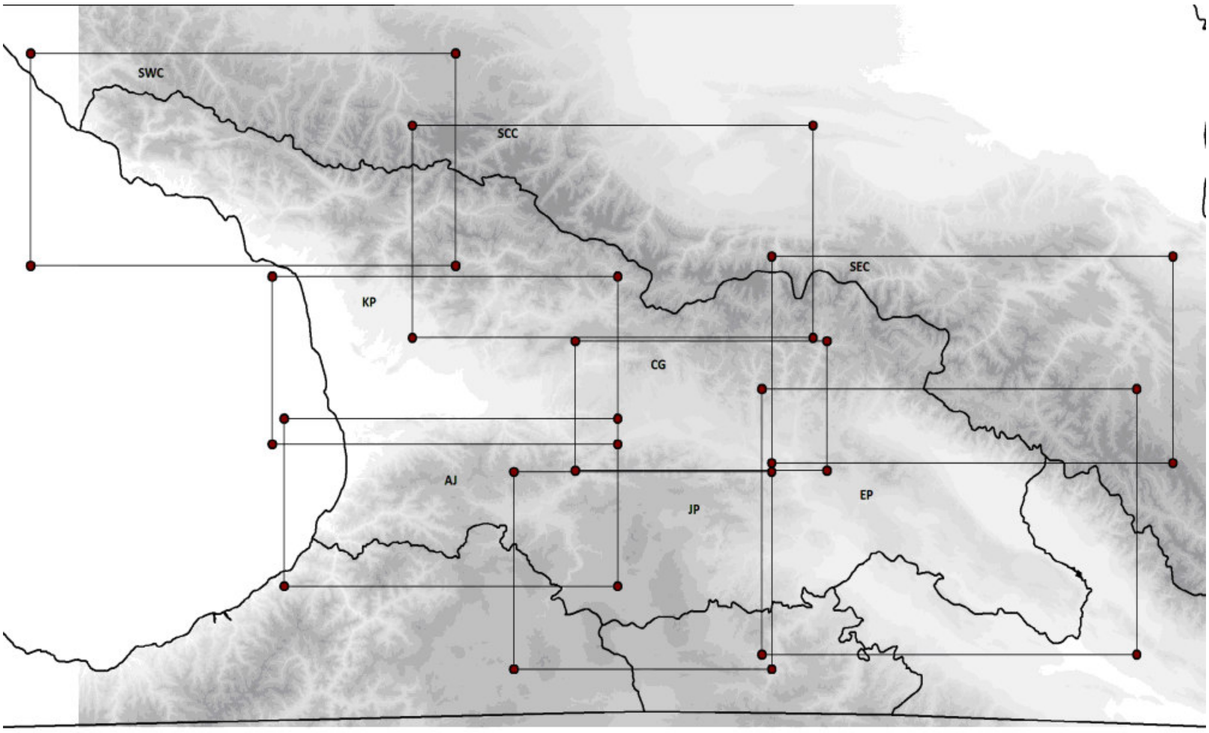


Fig. 4.3. *The contours represent the terrain elevation (m). The boxes indicate the 8 sub-regions.*

and Nested Dust run performance is better. But summer Dust experiment produces relatively large warm bias for south west sub-regions. The Nested Dust run for summer has better results for some sub-regions than course run. But for others - mostly central regions Nested Dust run simulation increases warm bias. These results are obtained after comparison with 0.5^o resolution observations gridded dataset. Comparing them with finer spatial resolution's observations will be useful to avoid mistakes raised from smoothing local effects.

5. Conclusion. We have investigated the dust's direct effect and its influence on the Georgia climate. This is first attempt to study this problem using RegCM model with dust module taking into account the aerosols radiative forcing in Georgia. In this paper, one climate parameter 2 m temperature was examined and the difference between Dust and NoDust simulations was found. The mean annual temperature warmed across the whole territory of Georgia in simulations where dust's direct effect was considered. The temperature performance was improved, as it had negative bias in simulation where dust effect wasn't taken into account. The finer resolution temperature results on annual scale have been improved more. On seasonal scale nested run has inhomogeneous results. It differs from sub-region to sub-region and from season to season, and it's especially diverse in summer. This result should be verified by means of rigorous comparison with observations from different sources and resolution. We begin our study from near surface temperature characterization, as it is a well-known and observed variable, but changes in temperature due to the changes in the radiation budget and cloud microphysical processes and thermodynamic state considering the effect of the dust and aerosol. To conclude how our results are relevant for the temperature, it is necessary to examine parameters of the cloud and radiation and continue this study by evaluating these variables.

Acknowledgment. The research in the paper was supported by the EU H2020 project "VRE for regional Interdisciplinary communities in Southeast Europe and the Eastern Mediterranean" VI-SEEM, grant ' 675121.

Table 4.1

Mean Annual, Temperatures Bias, Standard Deviation and Correlation Coefficient for 8 sub-regions from 4 simulations.

	Dust			NoDust		
	BIAS	STDV	CORREL	BIAS	STDV	CORREL
SCC	1.91	1.79	0.987	-0.53	1.94	0.984
CG	-1.03	1.62	0.988	-2.79	1.66	0.986
SEC	2.00	1.90	0.990	-0.57	1.77	0.988
EP	-1.99	1.49	0.990	-3.15	1.47	0.989
JP	-0.81	1.57	0.987	-2.61	1.70	0.984
KP	0.57	2.11	0.984	-1.67	2.21	0.981
AJ	0.41	1.79	0.983	-0.99	1.93	0.981
SWC	0.60	1.48	0.987	-1.41	1.54	0.985

	Nested Dust			Nested NoDust		
	BIAS	STDV	CORREL	BIAS	STDV	CORREL
SCC	-0.57	1.95	0.984	-2.11	2.00	0.981
CG	-0.81	1.72	0.987	-1.99	1.71	0.985
SEC	-0.44	2.08	0.987	-1.72	1.89	0.987
EP	-1.21	1.61	0.988	-2.15	1.46	0.989
JP	-0.73	1.70	0.985	-2.07	1.77	0.983
KP	- 1.03	2.16	0.983	-2.43	2.10	0.982
AJ	-0.41	1.93	0.980	-1.90	1.93	0.979
SWC	-0.52	1.57	0.984	-1.83	1.64	0.982

REFERENCES

[1] P. Forster, et al., *Changes in atmospheric constituents and in radiative forcing, in Climate Change 2007*, The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by S. Solomon et al., pp. 129- 234, Cambridge Univ. Press, Cambridge, U. K.

[2] R.L. Miller, I. Tegen and J.P. Perlwitz, *Surface radiative forcing by soil dust aerosols and the hydrologic cycle*, J. Geophys. Res., 109, D04203, doi:10.1029/2003JD004085, 2004.

[3] F. Barnaba and G. P Gobbi, *Aerosol seasonal variability over the Mediterranean region and relative impact of maritime, continental and Saharan dust particles over the basin from MODIS data in the year 2001*, Atmos. Chem. Phys., 4, 2367-2391, doi:10.5194/acp-4-2367-2004, 2004.

[4] C. Cavazos, M. C. Todd and K. Schepanski, *Numerical model simulation of the Saharan dust event of 6-11 March 2006 using the Regional Climate Model version 3 (RegCM3)*, Journal of Geophysical Research, Vol. 114, pp.1-24, D12109, doi:10.1029/2008JD011078, 2009.

[5] C. Zhao, et al., *The spatial distribution of mineral dust and its shortwave radiative forcing over North Africa: modeling sensitivities to dust emissions and aerosol size treatments*, Atmos. Chem. Phys. 2010, 10, 8821-8838. http://dx.doi.org/10.5194/acp-10-8821-2010.

[6] C. Zhao, et al., *Radiative impact of mineral dust on monsoon precipitation variability over West Africa*, Atmos. Chem. Phys. 11, 1879-1893. http://dx.doi.org/10.5194/acp-11-1879-2011.

[7] J. Huang, et al., *Dust and black carbon in seasonal snow Across Northern China*, Bull. Am. Meteorol. Soc. 92, 175-181. http://dx.doi.org/10.1175/2010BAMS3064.1, 2011.

[8] K.W. Kim, Y.J. Kim and S.J. Oh, S.J., *Visibility impairment during Yellow Sand periods in the urban atmosphere of Kwangju, Korea*, Atmos. Environ. 35 (30), 5157-5167, 2001.

[9] Y.S. Chen, et al., *Effects of Asian dust storm events on daily mortality in Taipei? Taiwan*, Environ. Res. 95, 151-155, doi: 10.1016/j.envres.2003.08.008, 2004.

[10] M. Aloysius, et al., *Role of dynamics in the advection of aerosols over the Arabian Sea along the west coast of peninsular India during pre-monsoon season: A case study based on satellite data and regional climate model*, J. Earth Syst. Sci. 120, No. 2, 269-279, 2011.

[11] G. Alexandri, et al., *On the ability of RegCM4 regional climate model to simulate surface solar radiation patterns over Europe: an assessment using satellite-based observations*, Atmos. Chem. Phys., 15, 13195-13216, 2015, www.atmos-chem-phys.net/15/13195/2015/ doi:10.5194/acp-15-13195-2015.

[12] D. G. Kaskaoutis, et al., *Desert Dust Properties, Modelling, and Monitoring*, Advances in Meteorology, vol. 2012, Article ID 483632, 2, doi:10.1155/2012/483632, 2012.

[13] P. Nabat, et al., *Dust emission size distribution impact on aerosol budget and radiative forcing over the Mediterranean region: a regional climate model approach*, Atmos. Chem. Phys., 12, 10545-10567, doi:10.5194/acp-12-10545-2012.

[14] A.S. Zakey, F. Solmon, and F. Giorgi, *Implementation and testing of a desert dust module in a regional climate model*, Atmos. Chem. Phys., 6, 4687-4704, 2006.

[15] Y. P. Shao, et al., *Northeast Asian dust storms: real-time numerical prediction and validation*, Journal of Geophysical Research, vol. 108, article 4691, 2003.

[16] D. F. Zhang, et al., *Simulation of dust aerosol and its regional feedbacks over East Asia using a regional climate model*, Atmospheric Chemistry and Physics, vol. 9, no. 4, 1095-1110, 2009.

[17] S. Zhao et al., *Simulating direct effects of dust aerosol on arid and semi-arid regions using an aerosol-climate coupled system*, International Journal of Climatology, Published online in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/joc.4093, 2014.

[18] N. Mahowald, et al., *Understanding the 30-year Barbados desert dust record*, J. Geophys. Res., 107(D21), doi:10.1029/2002JD002097, 2002.

[19] S. Das, et al., *Examining mineral dust transport over the Indian subcontinent using the regional climate model, RegCM4.1*, Atmospheric Research 134 (2013) 64-76, 2013.

[20] A. Konare, et al., *A regional climate modeling study of the effect of desert dust on the West African monsoon*, Journal of Geophysical Research, Vol. 113, d12206, pp.1-15, doi:10.1029/2007jd009322, 2008.

[21] A. Tsikerdekis et al., *Modeling the trans-Atlantic transportation of Saharan dust*, Bulletin of the Geological Society of Greece, vol. L, p. 1052-1061, Proceedings of the 14th International Congress, Thessaloniki, May 2016.

[22] H. Sun, and X. Liu, *Numerical Modeling of Topography-Modulated Dust Aerosol Distribution and Its Influence on the Onset of East Asian Summer Monsoon*, Hindawi Publishing Corporation Advances in Meteorology, Vol. 2016, Article ID 6951942, 15 pages, http://dx.doi.org/10.1155/2016/6951942, 2016.

[23] T. Ozturk, et al., *Projected changes in temperature and precipitation climatology of Central Asia CORDEX Region 8 by using RegCM4.3.5*, Atmospheric Research, 183 296-307, 2017.

[24] M. Shahgedanova, et al., *Climate Change, Glacier Retreat, and Water Availability in the Caucasus Region*, In Book Threats to Global Water Security, Editors: J. Anthony A. Jones, Trahel G. Vardanian, Christina Hakopian ,Publisher Springer Netherlands, 131-140, 2009.

[25] L. Shengelia et al., *Possibilities of the use of remote sensing technologies for the estimation of modern climate change impact on the Caucasus glaciers*, Georgian National Academy of Sciences, Monthly Scientific-Reviewed Magazine, Science and Technologies, Vol. 4-6, 25-30, 2012.

[26] C.R. Stokes, et al., *Late 20th century changes in glacier extent in the Caucasus Mountains, Russia/Georgia, J. Glaciol.*, 52, 99-109, 2006.

[27] M. Shahgedanova, et al., *Using the significant dust deposition event on the glaciers of Mt. Elbrus, Caucasus Mountains, Russia on 5 May 2009 to develop a method for dating and "provenancing" of desert dust events recorded in snow pack*, Atmos. Chem. Phys., 13, 1797-1808, 2013.

[28] P. Knippertz and M. C. Todd. *Mineral dust aerosols over the sahara: meteorological controls on emission and transport and implications for modeling*, Reviews of Geophysics, 50, 1-28, RG1007 / 2012

[29] Z. Li. *Influence of absorbing aerosols on the inference of solar surface radiation budget and cloud absorption.* Journal of

Climate, 11, 5-17, 1998

[30] M.Bangert, et al. *Saharan dust event impacts on cloud formation and radiation over Western Europe*, Atmos. Chem. Phys., 12, 4045-4063, doi:10.5194/acp-12-4045-2012, 2012.

[31] Z. Li, et al. *Aerosols and Their Impact on Radiation, Clouds, Precipitation, and Severe Weather Events*, Environments, Environmental Processes and Systems, Online Publication Date: Sep 2017, DOI:10.1093/acrefore/9780199389414.013.126, 2017

[32] W.-K.Tao, et al. *Impact of aerosols on convective clouds and precipitation.* Reviews of Geophysics, 50, 2012.

[33] N. M. Mahowald, Kiehl, L. M.. *Mineral aerosol and cloud interactions.* Geophysical Research Letters, 30, 1475-1478, 2003.

[34] G Myhre, et al. *Aerosol-cloud interaction inferred from MODIS satellite data and global aerosol models.* Atmospheric Chemistry and Physics, 7, 3081-3101, 2007.

[35] C. Denjean, et al. *Size distribution and optical properties of mineral dust aerosols transported in the western Mediterranean*, Atmos. Chem. Phys., 16, 1081-1104, https://doi.org/10.5194/acp-16-1081-2016, 2016.

[36] S. Twomey. *The influence of pollution on the shortwave albedo of clouds.* Journal of the Atmospheric Sciences, 34, 1149-1152, 1977.

[37] F. Giorgi, et al., *RegCM4: Model description and preliminary tests over multiple CORDEX domains*, Climate Research, 52, 7-29, doi: 10.3354/cr01018, 2012.

[38] G. A. Grell, J. Dudhia and D. R. Stauffer, *A description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5)*, NCAR Tech. Note NCAR/TN- 398+STR, 121 pp, 1994.

[39] H. Sun, Z. Pan and X. Liu, *Numerical simulation of spatialtemporal distribution of dust aerosol and its direct radiative effects on East Asian climate*, Journal of Geophysical Research, Atmospheres, vol. 117, no. 13, Article IDD13206, 2012.

[40] Y. Giorgi, et al., *Regional simulation of anthropogenic sulfur over East Asia 30 and its sensitivity to model parameters*, Tellus B, 53, 171-191, doi:10.1034/j.1600-0889.2001.d01-14.x, 2001.

[41] Y. Qian and F Giorgi, *Interactive coupling of regional climate and sulfate aerosol models over East Asia*, J. Geophys. Res., 104, 6477-6499, doi:10.1029/98JD02347, 1999.

[42] A. Gkikas, et al., T*he regime of intense desert dust episodes in the Mediterranean based on contemporary satellite observations and ground measurements*, Atmos. Chem. Phys., 13, 12135-12154, doi:10.5194/acp-13- 12135-2013, 2013.

[43] A. Gkikas, et al., *Atmospheric circulation evolution related to desert-dust episodes over the Mediterranean*, Q. J. Roy. Meteor. Soc., 141, 1634-1645, doi:10.1002/qj.2466, 2015.

[44] A. Gkikas, et al., *Characterization of aerosols episodes in the greater Mediterranean Sea area from satellite observations (2000-2007)*, Atmos. Environ., 128, 286-304, doi:10.1016/j.atmosenv.2015.11.056, 2016

[45] D. P. Dee, et al., *The ERA-Interim reanalysis: configuration and performance of the data assimilation system*, Q. J. Roy. Meteorol. Soc., 137, 553-597, doi:10.1002/qj.828, 2011.

[46] I. Harris, et al., *Updated high-resolution grids of monthly climatic observations- the CRU TS3.10 Dataset*, Int. J. Climatol., 34, 623-642, doi:10.1002/joc.3711, 2014.

[47] P. Kokkalis et al., *Ground-, satellite- and simulation-based analysis of a strong dust event over Abastumani, Georgia, during May 2009*, International Journal of Remote Sensing, 33:16, 4886-4901, DOI: 10.1080/01431161.2011.644593, 2012.

Edited by: Aneta Karaivanova
Received: Jan 5, 2018
Accepted: Apr 12, 2018

DOI 10.12694/scpe.v19i2.1342
Scalable Computing: Practice and Experience ISSN 1895-1767
Volume 19, Number 2, pp. 97–105. http://www.scpe.org © 2018 SCPE

AN ANALYSIS FOR PARALLEL WIND SIMULATION SPEEDUP USING OPENFOAM [*]

NEKI FRASHERI[†] AND EMANOUIL ATANASSOV [‡]

Abstract. An analysis of speedup for parallel execution of OpenFOAM software for wind simulation over rugged terrain is presented in the paper. Runtime speedup is analyzed using small and medium resolution DEM models for icoFoam and pisoFoam solvers, the latter due to consideration of turbulence, running in the parallel system Avitohol of Institute of Information and Communication Technologies of the Bulgarian Academy of Sciences. The results gave a clearer view about the possibility to run in reasonable time medium and high resolution models in regional scale, while indicating the weight of turbulence calculations for computing runtime requirements.

Key words: OpenFOAM; wind simulation; HPC scalability

AMS subject classifications. 68U20, 68W10, 68W40

1. Introduction.

The H2020 project VI-SEEM aims at creating a unique Virtual Research Environment (VRE) in Southeast Europe and the Eastern Mediterranean (SEEM), with special focus on the scientific communities of Life Sciences, Climatology and Digital Cultural Heritage [https://vi-seem.eu/]. One of the tasks in the project was to realize in VI-SEEM VRE the wind simulation over rugged terrain in regional scales, important for environmental studies and green energy production, which had to be undertaken by Polytechnic University of Tirana (UPT), Albania.

Involved for the first time in such simulations at UPT, we selected for this purpose the OpenFOAM software [https://www.openfoam.com/], which is one of available open source packages to solve Navier-Stokes equations for fluid dynamics, applicable for air flows.

Tanasescu has presented OpenFOAM as a leading Open Source CFD (Computational Fluid Dynamics) having qualities of accessibility, transparency, customization, and extensible [1]. The package offers a multitude of solvers, parallelized with MPI, for different Navier-Stokes problems.

A number of factors that impact the volume of requested computing capacities were identified during the literature review, related with the scalability dependence from the nature of concrete models, model size, interprocess communication, etc. In particular, air flow around sails solutions are presented in [2] by Lombardi et al, and the impact of data interchange between cores is evaluated. Ravelli et al analyzed air flows in turbines [3]. Lysenko et al presented several cases of turbulent flows in [4], considering both weak and strong scalability for the OpenFOAM. Rivera et al studied large eddy flows and related scalability [5].

Dagna and Hertzer studied the scalability of OpenFOAM with hybrid MPI and OpenMP parallelization with up to 4096 cores in BlueGene/Q system [6]. Lysenko et al simulated aerodynamic sound from a circular cylinder using up to 256 cores with an average efficiency of 70% [7]. Karasek et al compared different benchmarks for evaluation of OpenFOAM performance using up to 1024 cores [8].

Kornyei studied the speedup and scalability of simulating combustible gas flows with up to 256 cores in [9]. Ponweiser et al presented fluid-structure simulations for aircraft design in [10], dealing with memory limitations for limited number of 128-256 cores. Sidlof and Ridky studied the scalability of simulated flows past airfoil in [11], experiencing 50% speedup reduction when over 36 cores were used.

Atmospheric simulations were done by Flores et al for turbulent flows over complex urban geometry [12]. Vermier et al analysed the atmospheric boundary layer in complex terrain [13]. Garcia et al studied the buoyancy effect of temperature over terrain, using SRTM Digital Elevation Model and Landsat infrared imagery for the ground temperature [14].

Garcia and Boulanger simulated the altitude winds over mount Saint Helens using SRTM Digital Elevation Model [15]. Hardin offered a solution how to build OpenFOAM terrain mesh using Digital Elevation Models

[*] The work presented in this paper has been supported by EU under the H2020 programme Grant agreement no. 675121 VRE for regional Interdisciplinary communities in Southeast Europe and the Eastern Mediterranean (VI-SEEM).

[†] Faculty of Information Technology, Polytechnic University of Tirana Tirana, Albania (nfrasheri@fti.edu.al).

[‡] Institute of Information and Communication Technologies, Bulgarian Academy Of Sciences, Sofia, Bulgaria (emanouil@parallel.bas.bg).

from GRAS GIS system [16]. Tapia made a synthesis of theoretical issues of OpenFOAM with cases of wind farms and simple hill terrain [17].

Most of the reviewed literature presented for concrete problems some runtime analysis, but without details on memory usage, which resulted to be critical in our case and confirmed the remark of Culpo that *the size of problems that can be handled on a HPC cluster lies beyond the limitations imposed by smaller in-house clusters* [18].

Our study was focused on the evaluation of OpenFOAM scalability when it runs for 3D regional wind simulations in HPC systems in our disposal. Simulations were planned for the geographical region that includes Albania, two thirds of which is mountainous. In order to include rugged terrain in OpenFOAM data, we used the Digital Elevation Model (DEM) fragment of mountainous area including Albania from the NASA Shuttle Radar Topography Model (SRTM) data obtained from USGS archive [https://earthexplorer.usgs.gov/]. The selection of SRTM DEM was based on the facts that it can be obtained freely from the Internet and we had previous experiences with its usage.

The models were designed to take into account air turbulence, which requires iterative solution of equations for a sequence of time intervals and periodical storage of related temporal results in disk space. While increasing the model resolution through decreasing of spatial mesh steps, another problem emerged related with the balance between spatial and temporal discretization steps [19], leading to the need of decreasing time steps proportionally with spatial steps, and increasing the number of iterations in order to keep the same time span.

First experiments were carried out in local workstations and the small multiprocessor system of UPT, obtaining a first evaluation of computing capacities, as a preparatory phase for running it in the VI-SEEM VRE [20], [21]. In the actual paper results from new experiments in Avitohol are presented.

2. Experimental Setup. Atmospheric winds can be simulated solving Navier-Stokes equations applied for incompressible laminar and turbulent flows in spatial 3D volumes. Equations include the temporal dimension as well, especially necessary for turbulence problems. Software OpenFOAM solves Navier-Stokes equations using finite volume methods based on digitized spatial 3D mesh, through a suite of iterations in temporal dimension. In order to include the effect of buildings and mountains in air flows, 3D mesh generated for a rectangular volume should be deformed following a digital elevation model of the relief.

Mountainous ranges in Western Balkans are cut by a network of narrow deep valleys, which impact the air flows due to meteorological conditions. Valleys less than 1km wide require wind simulation models of high resolution. NASA SRTM DEM with 3 arcsec per pixel, which corresponds with a rectangular metric resolution of 100x100m in equator and 70x100m in latitudes 40 degrees (meridians are nearer each other farther from equator), and it would be suitable for wind simulations in narrow valleys. We selected the DEM section covering Albania with the highest resolution available 3 arcsec per pixel with size 3600x4800 pixel (Fig. 2.1.a), defined with corner coordinates in degrees, minutes and seconds:

Upper Left (18d59'58.50"E, 43d 0' 1.50"N)
Lower Left (18d59'58.50"E, 38d59'58.50"N)
Upper Right (22d 0' 1.50"E, 43d 0' 1.50"N)
Lower Right (22d 0' 1.50"E, 38d59'58.50"N)

The 3D model size of lower atmosphere volume in kilometers was 270x480x10, where the variation of arcsec metric distances by latitude was considered. Our experiments were based on several models of the same volume with different resolutions in the range from DEM 36x48 pixels up to 500x667 pixels obtained reducing proportionally the Fig. 2.1.a image, with a spatial meridian step varying respectively from 10km to 720m.

The characteristics of used model resolutions are presented in Table 2.1, where respective numbers of digitized finite volume elements in three spatial dimensions (X, Y, and Z), their total number, and requested virtual memory (RAM) for each case are presented. The factor dedicated for each of models, proportional with related one-dimensional size in pixels, is used to simplify the graphical presentation of results. It was not possible to run models with higher resolutions, for which the extrapolation was used to evaluate requested computing capacities.

The Reynolds number for the air was taken 1e-5, situated in the file *transportProperties.* Boundary conditions were defined in files *0/p* and *0/U*, applied for boundary faces of 3D rectangle in the file *blockMeshDict* as

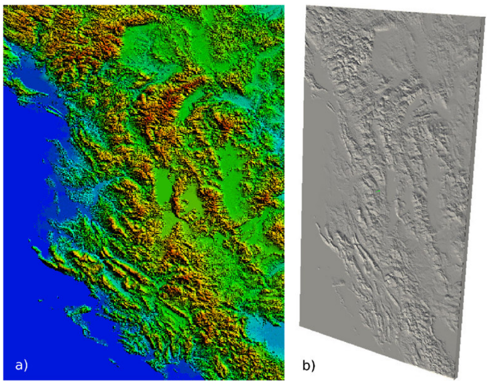


Fig. 2.1. *a) 3600x4800 DEM image for Albania and surrounding area; b) bottom face of 3D volume deformed based on DEM.*

Table 2.1
Model sizes used for experiments

Factor	X pixels	Y pixels	Z pixels	Elements	RAM KB
10	36	48	10	1.73E+04	1.20E+05
43	154	206	43	1.36E+06	1.25E+06
60	216	288	60	3.73E+06	3.07E+06
100	360	480	100	1.73E+07	1.31E+07
139	500	667	139	4.64E+07	3.34E+07
1000	3600	4800	1000	1.73E+10	1.43E+09

follows (orientation of front-end and left-right faces were defined following the direction of the wind north-south):

- front end vertical faces: fixed walls with uniform values +1 and -1 and zero gradients
- left and right vertical faces: fixed zero gradient walls
- top horizontal face (upper atmosphere): fixed zero gradient wall
- bottom horizontal face (ground surface): fixed zero values wall

The model meshes were generated with the OpenFOAM module *blockMesh*. Before running OpenFOAM solvers, the generated spatial coordinates of mesh nodes situated in the file *./constant/polyMesh/points* were modified to include the relief, modifying the altitudes of nodes starting from the bottom of mesh (ground surface) and decreasing linearly to zero until the top.

Similar with other files used by OpenFOAM, the file *points* contains in editable format ASCII the data: a header and the suite of coordinates (x,y,z) of mesh nodes. DEM files were downloaded from USGS repository in binary format. GDAL software tools [http://www.gdal.org/] were used to process binary DEM data, using *gdalinfo* to obtain related metadata, and *gdal_translate* to convert data into Surfer ASCII grid format [http://www.tifton.uga.edu/sewrl/flownet/flownet.htm]. The already known and editable grid text format contains the header with number of pixels and the 2D array of heights in meters, which were used to modify vertical coordinates in *points* file using an in-house written software.

Splitting of input data for parallel processing with the solver module and recombination of partial results were done with modules *decmposePar* dhe *reconstructPar*. Used OpenFOAM MPI prallelized solvers were *icoFoam* for incompressible laminar flows and *pisooFoam* for turbulent flows.

The execution of OpenFOAM suite in parallel requires running of four modules: the mesh generator *blockMesh* (necessary even for simple modification of boundary conditions), followed by the *decomposePar* module, the solver (*icoFoam* and *pisoFoam* in our case), terminating with *reconstructPar* module.

The OpenFOAM solution gives the distribution of scalar potential and vectorial velocity fields, for the latter magnitude and three vector components are given for axes X (west to east), Y (south to north), and Z (bottom-up).

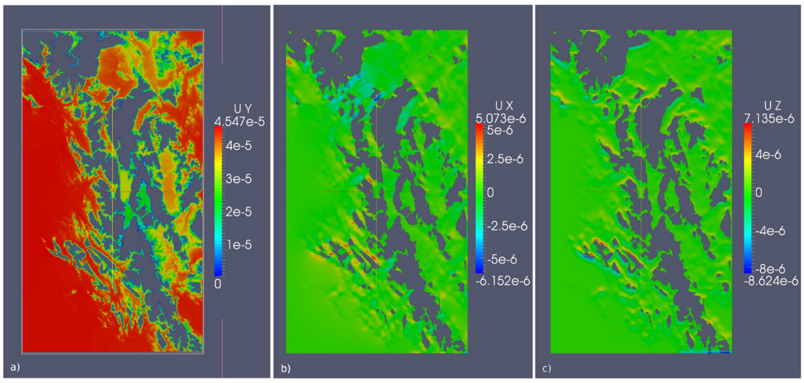


Fig. 3.1. *a) North-south wind flow in altitude 1000m; b) induced west-east wind flow in altitude 1000m; c) induced vertical wind flow in altitude 1000m.*

Recent experiments were carried out running OpenFOAM solver for 10,000 time steps of length 0.001 second, storing results in external storage at the end of each second for duration of 10 seconds. The same was done for all model sizes, taking into account the time step length required for higher resolution models. Use of small spatial steps with long time steps in previous experiments has led to miss-balance between spatial and temporal discretization, increasing courant numbers greater that one (the case when fluid particles *jump over* one mesh cell while moving from one time instance to the next one), and the divergence of iterative process [19].

The used computer system was the supercomputer Avitohol, with 150 servers with dual Intel Xeon E5-2650v2 8C @2.6GHz running Scientific Linux 6 and Intel compilers, in the Institute of Information and Communication Technologies, Bulgarian Academy of Sciences (IICT-BAS).

3. Preliminary Simulation Results. The used SRTM DEM data (Fig. 2.1) cover territories of Albania and small parts of surrounding countries. The area is characterized by a group of high Alps Mountains in North, and ranges of mountains that extend in direction north-north-west to south-south-east cut by narrow river valleys flowing from east to west. In the western part of this area there is the PreAdriatic Depression, lowlands bordered by Adriatic Sea.

Experiments presented here were done with boundary conditions for a constant regional wind flowing north to south, using potentials +1 and -1 in respective northern and southern faces of the 3D volume. We used zero conditions for the bottom ground face, and gradient zero on the rest of faces (east, west and top).

In the Fig. 3.1 there are presented the distribution of magnitude and components X and Z of the velocity field in a plan at altitude 1,000m, where tops of mountains are shown in gray. A near to surface reduction of wind magnitude is visible around mountain peaks. Low values of velocity are characteristic for protected *closed* mountainous valleys (image 'a'), with a decrease of magnitude down to 25%.

The effect of relief is more visible in images 'b' and 'c' presenting west - east and vertical wind flows. Due to oblique extension of mountain ranges, in valleys there is generated a component of west - east wind flow with magnitude 50%, flowing to the east or west depending to the orientation of open valleys. In the image 'c' strong vertical wind flow component is visible in high mountain slopes facing north. This phenomenon is significant in two areas - northern boundaries of Mirdita and mountainous areas east of Vlora city.

The results show the importance of such simulations for planning of wind energy farms, and for the air transport. A more detailed analysis is necessary for such purpose, considering the complexity of relief.

4. OpenFOAM Scalability In Avitohol. New results presented in this paper consist of runtime analysis for the solver icoFoam and comparison of both runtime and disk space required by both solvers icoFoam and pisoFoam, for models which sizes are described in Table 2.1, except for the case of highest resolution for which extrapolation was used in few cases.

The volume of virtual memory resulted similar for all modules depending on the quantity of 3D mesh nodes (Fig. 4.1; while volume of external storage correlates with the virtual memory size multiplied with the number of stored temporal results (in the figure case of 100 results is presented).

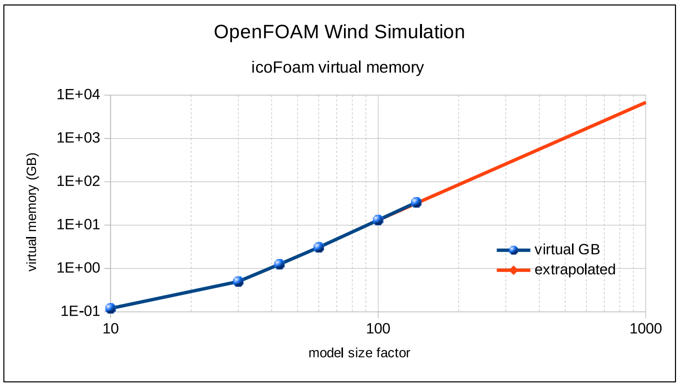


FIG. 4.1. *Memory usage of OpenFOAM - single process run.*

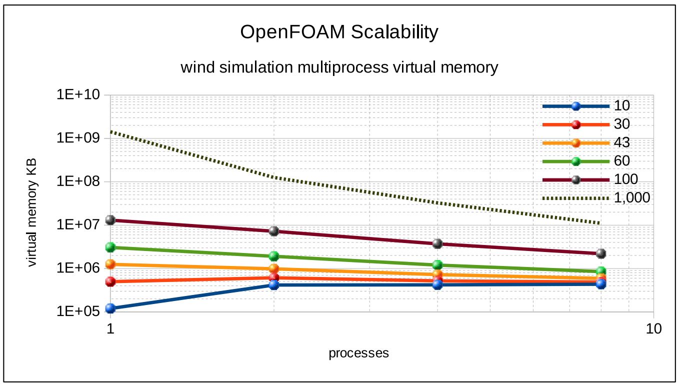


FIG. 4.2. *Memory usage of OpenFOAM - multi process run.*

Only the solver was possible to run in parallel, with this case due to the splitting of 3D mesh between processes, for each of them the required virtual memory is reduced converging towards around 0.5 GigaBytes per process with the increase of parallelism (Fig. 4.2).

Other pre- and post- processing modules required the total of virtual memory, decomposing and reconstructing modules duplicate the volume of data in external storage (each solver process stores its partial results and during post-processing there are stored combined files). Central memory requirements, time steps, and runtime conditioned the size of models executed in Avitohol.

Increase of time steps in order to have the requested temporal span of dynamic solutions when increasing mesh resolution was necessary to avoid the iterations divergence when spatial and temporal discretization steps were not balanced leading to the increase of courant numbers, a case of such divergence is presented in Fig. 4.3.

Runtime of different models per number of processes is shown in Fig. 4.4. For the model size 100 a trend line is calculated extrapolating the runtime for 1,000 processes.

The same runtime plotted versus the size of models is shown in Fig. 4.5. The trend line is calculated for a single process run of different models, extrapolating the runtime for single process run of the highest model size.

The speedup of parallel execution for different model sizes is given in Fig. 4.6. The lowest resolution model speedup degenerates quickly while for other models the trend seems constant for the used range 1:64 of the number of processes.

Efficiency of parallelization is given in Fig. 4.7. Even for higher resolution models the trend of reduction of efficiency is visible. The reduction trend of efficiency is expected to reach values of 50% when running models with up to 1,000 parallel processes. Further degradation of efficiency due to virtual memory requirements was not considered in these extrapolations.

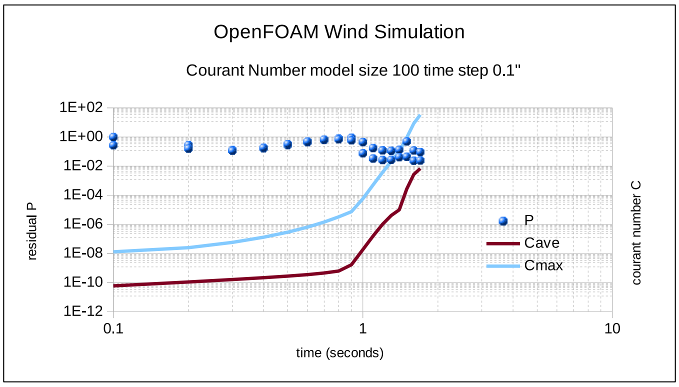


Fig. 4.3. *Divergence of solution processes: error of potential field P, average Cave and maximal courant Cmax numbers for time steps 0.1 seconds.*

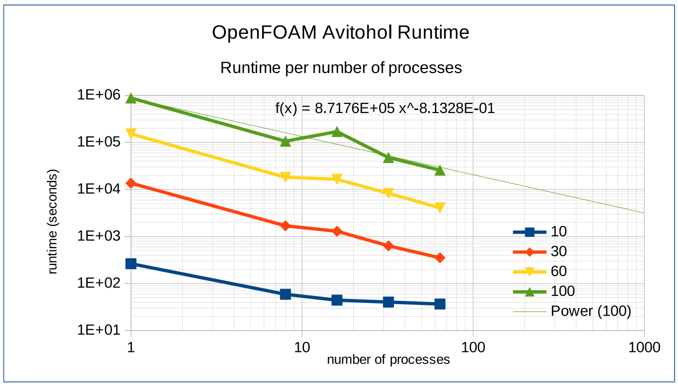


Fig. 4.4. *Runtime of models per number of parallel processes.*

Comparing the trends of runtime for different number of parallel processes in Fig. 4.5, it was possible to make an simple approximate evaluation and extrapolation of runtime for the highest resolution model (of factor 1,000) for up to 640 parallel processes. The runtime trends for different process numbers are parallel with the sequential run. Extrapolating the trend of 64 processes for the model factor 1,000 the approximate value 1.00E+8 was obtained. Supposing 100% efficiency, evaluated runtime for the highest model size for 64 and 640 processes given in Table 4.1.

Comparison of solver *icoFoam* and *pisoFoam* was done in two planes - runtime and external storage requirements. Comparison of the runtime is given in Fig. 4.8. Runtime trend in case of *pisoFoam* resulted at least 35% more compared with *icoFoam*.

Comparison of *icoFoam* and *pisoFoam* external storage requirements is given in Fig. 4.9. Requirements for both solvers are quite similar with each other, indicating that the use of *pisoFoam* does not require excessive extra external storage compared with *icoFoam*.

The runtime problem with *pisoFoam* is related with long time span of solution sequences necessary to indicate turbulence and eddies, compared with laminar solutions.

5. Conclusions. Usage of OpenFOAM for wind simulation models of medium resolution (spatial discretization steps 10km-1km) over rugged mountainous regional area of Albania and surroundings was possible in the Avitohol using reasonable computational resources. Beside the runtime, another limitation was the virtual memory in levels of 10-30 GB for medium sized models, requested by preparatory modules *blockMesh*, *decomposePar*, and *reconstructPar*.

An optimistic extrapolation of the runtime for models with resolution 100m of spatial discretization step

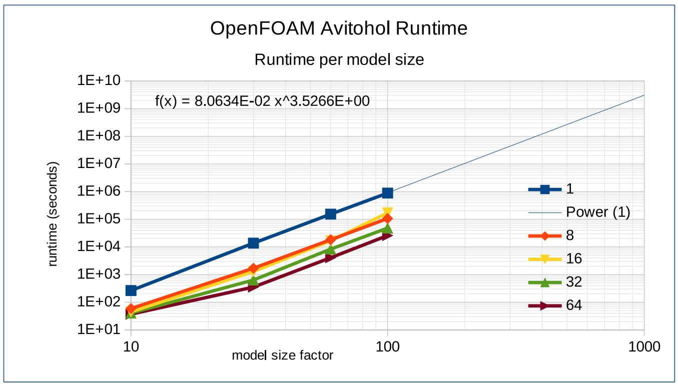


Fig. 4.5. *Runtime of parallel processes per model size.*

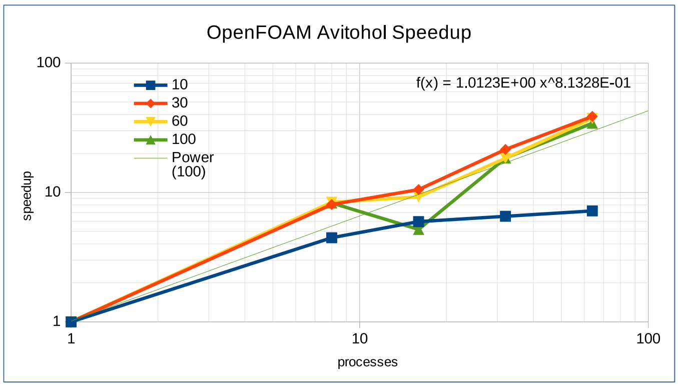


Fig. 4.6. *Speedup of parallel execution in Avitohol*

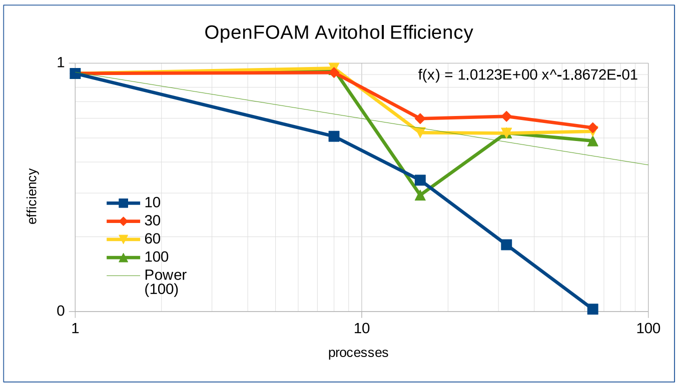


Fig. 4.7. *Efficiency of parallel execution in Avitohol*

Table 4.1
Model sizes used for experiments

processes	seconds	hours	days	months	years
64	1.00E+08	27,778	1,157	37.95	3.16
640	1.00E+07	2,778	116	3.79	0.32

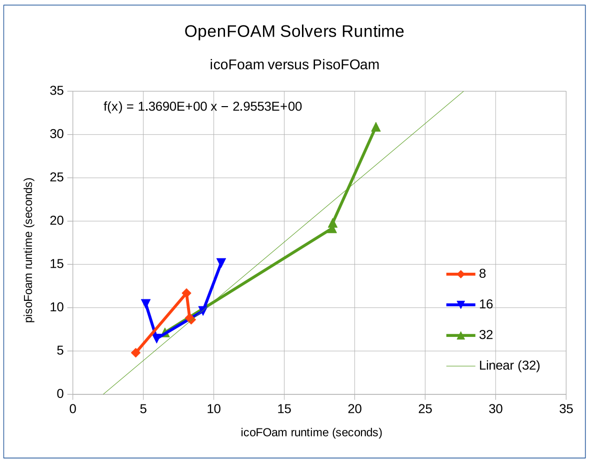


FIG. 4.8. *icoFoam versus pisoFoam runtime*

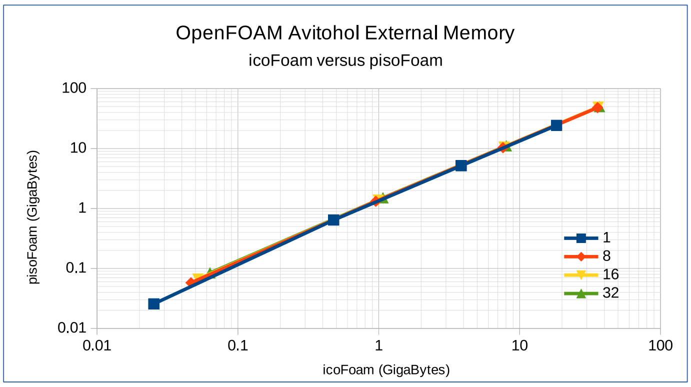


FIG. 4.9. *icoFoam versus pisoFoam external storage requirements*

using 640 parallel processes resulted at least 4 months, not considering limitations due to virtual memory and degeneration of efficiency from inter-process and external storage communication. Comparison of two OpenFOAM solvers, *icoFoam* for laminar flows and *pisoFoam* for turbulence flows showed that the latter requires about 35% runtime more than the former, while memory requirements are similar. It is also necessary to consider studies of turbulence and eddies would require the storage in disk of a greater number of temporal results, compared with experiment setup presented in this paper.

Acknowledgments. The research leading to these results has been co-funded by the European Commission under the H2020 Research Infrastructures contract no. 675121 (project VI-SEEM). Authors are grateful for all the help of the staff of IICT-BAS involved with the preparation of the OpenFOAM environment. Collaboration in framework of COST Action IC1305 NESUS promoted the speedup of this work.

REFERENCES

[1] C. Tanasescu, OpenFOAM *and* SGI *designed to work together*, http://docplayer.net/3118206 -Openfoam-and-sgi-designed-to-work-together-christian-tanasescu-vice-president-software- engineering.html

[2] M. Lombardi, N. Parolini, A. Quarteroni, and G Rozza, *Numerical simulation of sailing boats : dynamics,* FSI*, and shape optimization*, MATHICSE Technical Report No. 03.2011, April 2011.

[3] S. Ravelli, G. Barigozzi, F. Pasqua, R. Pieri, and R. Ponzini, *Numerical and experimental study for the prediction of the steady, three dimensional flow in a turbine nozzle vane cascade using* OpenFOAM, in International CAE Conference 2014, Verona, Italy, October 2014, pp. 27-28.

[4] D. Lysenko, I. Ertesvag, and K. Rian, *Testing of* OpenFOAM CFD *code for plane turbulent bluff body flows within conventional* URANS *approach*, in International Conference on Computational Fluid Dynamics in the Oil Gas, Metallurgical and Process Industries - CFD11, Trondheim, Norway, June 2011.

[5] O. Rivera, K. Furlinger, and D. Kranzlmuller, *Investigating the scalability of* OpenFOAM *for the solution of transport equations and large eddy simulations*, in ICA3PP11 Proceedings of the 11th International Conference on Algorithms and Architectures for Parallel Processing Volume Part II, LNCS 7017, 2011, pp. 121-130.

[6] P. Dagna and J. Hertzer, *Evaluation of multi-threaded* OpenFOAM *hybridization for massively parallel architectures*, Partnership for Advanced Computing in Europe (PRACE) white paper, 2017, http://www.prace-ri.eu/IMG/pdf/wp98.pdf

[7] D.A. Lysenko, I.S. Ertesvag, and K.E.Rian, *Towards simulation of far-field aerodynamic sound from a circular cylinder using* OpenFOAM, Aeroacoustics, vol. 13, no. 1, 2014.

[8] T. Karasek, D. Horak, V. Hapla, A. Markopoulos, L. Riha, V. Vondrak, and T. Brzobohaty, *Application of* CFD *and* CSM *open source codes for solving multiscale multiphysics problems*, IT4Innovations, VSB - Technical University of Ostrava, www.prace-ri.eu

[9] L. Kornyei, *Simulation of gas flow in a combustion chamber using high performance computing hardware*, in Workshop on the Occasion of the 60th Birthday of Ferenc Igloi, Budapest, Hungary, October 3, 2012.

[10] T. Ponweiser, P. Stadelmeyer, and T. Karasek, *Fluid-structure simulations with* OpenFOAM *for aircraft designs*, Partnership for Advanced Computing in Europe (PRACE) Report, www.prace-ri.eu/IMG/pdf/wp172.pdf

[11] P. Sidlof and V. Ridky, *Scalability of the parallel* CFD *simulations of flow past a fluttering airfoil in* OpenFOAM, in EPJ Web of Conferences 92, (2015) DOI: 10.1051/ epjconf / 201 5 9 2 020 8 0.

[12] F. Flores, R. Garreaud and R. C. Munoz, CFD *simulations of turbulent buoyant atmospheric flows over complex geometry: Solver development in* OpenFOAM, Elsevier Computers & Fluids 82 (2013) 113.

[13] J. Vermeir, M. Runacres and T. De Troyer, CFD *modelling of the boundary layer in complex terrain validated by field measurements*, Proceedings of the EWEA Annual Event, 19 Apr 2012, Copenhagen.

[14] M. Garcia, P. Boulanger and D. Giraldo, CFD *analysis effect buoyancy due terrain temperature based integrated* DEM *landsat infrared imagery*, Ingenieria y Ciencia, ISSN 17949165, Vol. 4, no. 8, December 2008, pp. 6584

[15] M. Garcia and P. Boulanger, *Low altitude wind simulation over mount Saint Helens using* NASA SRTM *Digital Terrain Model*, Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06) 0-7695-2825-2/06, IEEE 2006.

[16] E. Hardin, *Simulating wind over terrain: how to huild an* OpenFOAM *case from* GRASS GIS *Digital Elevation Models*, 2013, http://www.cybermanual.com/simulating- wind-over-terrain-how-to-build-an-openfoam-case-from-grass-gis-digital-elevation-models /download.html

[17] X. P. Tapia, *Modelling of wind flow over complex terrain using* OpenFoam, Masters Thesis in Energy Systems, June 2009, http://www.diva-portal.org/smash/get/diva2

[18] M. Culpo, *Current bottlenecks in the scalability of* OpenFOAM *on massively parallel clusters*, Partnership for Advanced Computing in Europe (PRACE), white paper 2010, http://www.prace-ri.eu/IMG/pdf

[19] ESI-OpenCFD, OpenFOAM *lid-driven cavity flow*, http://www.openfoam.com/documentation /tutorial-guide/tutorialse1.php

[20] N. Frasheri and E. Atanassov, *Scalability Issues for Wind Simulation using* OpenFOAM, Third Nesus Action Workshop NESUS 2016, IICT-BAS, Sofia, Bulgaria. October, 6-7, 2016.

[21] N. Frasheri and E. Atanassov, *Scalability Issues for Wind Simulation Using* OpenFOAM, Bulgarian Academy of Sciences, Journal Cybernetics and Information Technologies, 2017 (submitted).

Edited by: Aneta Karaivanova
Received: Dec 19, 2017
Accepted: Mar 27, 2018

DOI 10.12694/scpe.v19i2.1347
Scalable Computing: Practice and Experience ISSN 1895-1767
Volume 19, Number 2, pp. 107–118. http://www.scpe.org © 2018 SCPE

CLIMATE APPLICATIONS IN A VIRTUAL RESEARCH ENVIRONMENT PLATFORM

GEORGI GADZHEV[*], IVELINA GEORGIEVA[*], KOSTADIN GANEV[*], VLADIMIR IVANOV[*], NIKOLAY MILOSHEV[*], HRISTO CHERVENKOV[†], AND DIMITER SYRAKOV[†]

Abstract. Previous atmospheric composition studies were based on extensive computer simulations carried out with good resolution using up-to-date modelling tools and detailed and reliable input data.

The oncoming climate changes will exert influence on the ecosystems, on the all branches of the international economy, and on the quality of life. Regional climate models (RCMs) are important instruments used for downscaling climate simulations from Global circulation models (GCMs).

The air quality (AQ) impact on human health and quality of life is an issue of great social significance. Evaluating this impact will give scientifically robust basis for elaborating efficient short term measures and long term strategies for mitigation of the harmful effects of air pollution. The AQ impact is evaluated in the terms of Air Quality Indices (AQI). Some extensive numerical simulations of the atmospheric composition fields in Bulgaria and Sofia have been recently performed. A quite extensive data base was created from simulations which were used for different studies of the atmospheric composition, including the AQ climate.

Main aims of the numerical experiment presented in this paper are: (1) Adaptation and tuning of the RegCM model for the Balkan Peninsula and Bulgaria and thus development of a methodology able to predict possible changes of the regional climate for different global climate change scenarios and their impact on spatial/temporal distribution of precipitation, hence the global water budgets, to changes of the characteristics and spatial/temporal distribution of extreme, unfavorable and catastrophic events (drought, storms, hail, floods, fires, sea waves, soil erosion, etc.). (2) Development of a methodology and performing reliable, comprehensive and detailed studies of the impact of lower atmosphere parameters and characteristics on the quality of life (QL) and health risks (HR) for the population.

Key words: Virtual Research Environment, Regional climate models, RegCM, Air Quality Indices

AMS subject classifications. 86A10, 65Y05

1. Introduction.

The climate modelling community has very strong computational needs. In particular, the integration of various computational resources such as High-performance computing (HPC) and Grid jointly with data infrastructure. VI-SEEM is a project that aims at creating a unique Virtual Research Environment (VRE) in Southeast Europe and the Eastern Mediterranean (SEEM), in order to facilitate regional interdisciplinary collaboration, with special focus on the scientific communities of Life Sciences, Climatology and Digital Cultural Heritage. In the frame of the VI-SEEM project, the existing e-Infrastructures are being unify into an integrated platform to better utilize synergies, for an improved service provision within a unified Virtual Research Environment to be provided to scientific communities of high impact in the combined South East Europe and Eastern Mediterranean region. Perhaps the largest focus is on regional climate modelling and weather forecasting, where local weather and regional climate phenomena are investigated. This is complemented by global climate modelling where the impact of global phenomena on the regional climate is the focus. These results are crucial to predict extreme weather in the region and understand the future trends of the regional climate. Another strong field of related research is the study of air pollution that includes the influence on the climate and human health. These activities jointly enable the assessment of the impact on regional climate due to climate change. Climate impact studies provide the analysis of the upcoming change on humans, the environment and society that is so crucial for policy makers.

In this paper we will be present the results from two applications – ACIQLife (Atmospheric Composition Impact on Quality of Life and Human Health) and TVRegCM (Tuning and Validation of the RegCM) in the frame of VI-SEEM project, climate section.

The ACIQLife application is focused on development of a methodology and performing reliable, comprehensive and detailed studies of the impact of lower atmosphere parameters and characteristics on the quality of life (QL) and health risks (HR) for the population in our country. The TVRegCM reached to adaptation

[*] National Institute of Geophysics, Geodesy and Geography–Bulgarian Academy of Sciences, Acad. G. Bonchev str., bl. 3 1113 Sofia, Bulgaria (ggadjev@geophys.bas.bg),

[†] National Institute in Meteorology and Hydrology–Bulgarian Academy of Sciences, 66, Tsarigradsko Shose blvd 1784 Sofia, Bulgaria

Table 2.1
Computer resource requirements on 16 CPU-s for 1 Day simulation for ACIQLife

	WRF	CMAQ and SMOKE	Total
Time (h)	3	2	5
HDD (GB)	0.5	1	1.5

Table 2.2
Computer resource requirements on 16 CPU-s for TVRegCM

	1 Month Simulation	x120 Months	x20 Cases
Time (h)	6	720	14400
HDD (GB)	6	720	14400

and tuning of the RegCM model for the Balkan Peninsula and Bulgaria and thus development of a methodology able to predict possible changes of the regional climate for different global climate change scenarios and their impact on spatial/temporal distribution of precipitation, hence the global water budgets, to changes of the characteristics and spatial/temporal distribution of extreme, unfavorable and catastrophic events (drought, storms, hail, floods, fires, sea waves, soil erosion, etc.). All these changes will have influence on the ecosystems and on practically all sectors of the economy and human activity and consequently on the quality of life.

2. HPC computing. The model simulations were performed day by day for two periods.

The computer resource requirements for the (WRF) Weather Research and Forecasting Model, (SMOKE) Sparse Matrix Operator Kernel Emissions Modeling System, (CMAQ) The Community Multiscale Air Quality Modeling System and RegCM, simulations are rather big [16] (Tables 2.1 and 2.2) and that is why the numerical experiments were organized in effective High-performance computing (HPC) environment. The simulations were organized in two separate jobs: one job for WRF simulations and one job for SMOKE, CMAQ and post-processing procedures. This makes the jobs run time for 6 days real time fairly reasonable for ACIQLife application and 3 months for TVRegCM application.

The calculations were implemented on the Supercomputer System Avitohol at IICT–BAS (Institute of Information and Communication Technologies–Bulgarian Academy of Sciences). The supercomputer consists of 150 HP Cluster Platform SL250S GEN8 servers, each one equipped with 2 Intel Xeon E5-2650 v2 8C 2600 GHz CPUs and 64GB RAM per server. The storage system is HP MSA 2040 SAN with a total of 96 TB of raw disk storage capacity. All the servers are interconnected with fully non-blocking FDR Infiniband, using a fat-tree topology [1] and [2]. The needed libraries and programs were installed on supercomputer for proper functioning and working of models used in this study. The Avitohol system is a part of the Virtual Research Environment platform (VRE platform) built in the framework of the VI-SEEM project [3]. The both applications - ACIQLife and TVRegCM use not only HPC resources provided by the VRE platform, but they use also other services like VI-SEEM Simple Storage (VSS) and VI-SEEM Archival Service (VAS) to save the obtained data. The training materials about both applications are available in the VI-SEEM Training portal [4] and [5]. According the VI-SEEM accounting system [6], 730 ACIQLife jobs and 810 TVRegCM jobs were run to receive some of the current scientific results. The needed CPU time and storage per job is shown in the Tables 2.1 and 2.2.

The models output from ACIQLife and TVRegCM applications are uploaded on VRE repository website. The results are free and can be use by the scientific communities in the region. The workflows wre also created and uploaded for both applications.

The ACIQLife ouput is a NetCDF file with surface concentrations on an hourly basis of the most important pollutants (which are used for calculation of AQI) and annually/seasonally averaged hourly values of the different AQI value for the selected area.

The TVRegCM ouput is also NetCDF file, but for each month of the period and consist of daily and hourly averaged values of the meteorological parameters for the area of interest.

3. ACIQLife application. Some extensive numerical simulations of the atmospheric composition fields in Bulgaria and Sofia have been recently performed. Quite extensive data base has been created from the

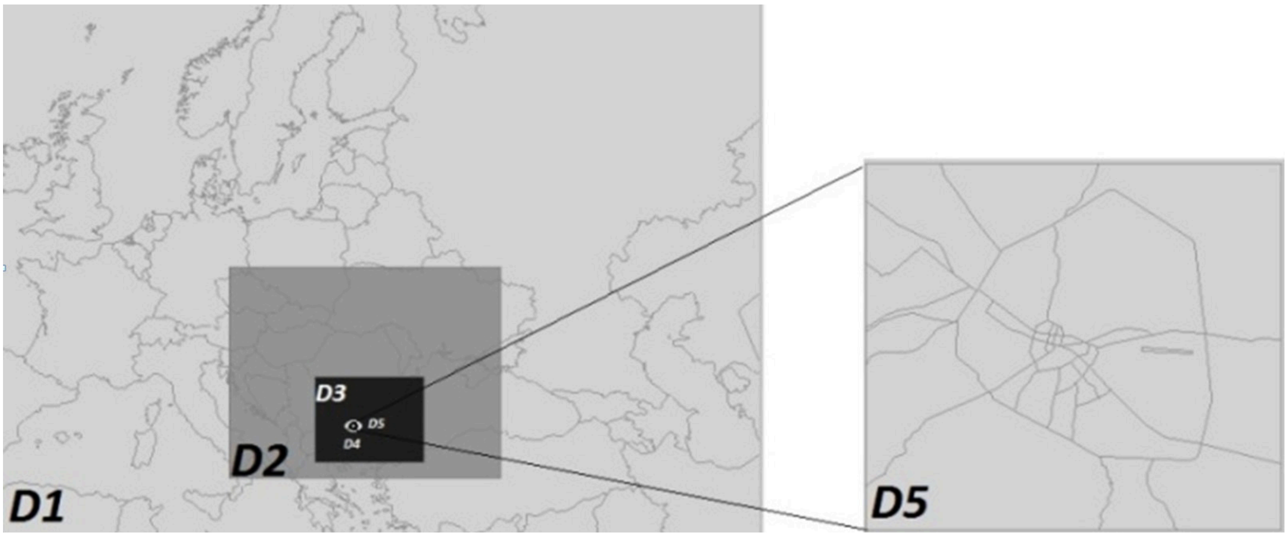


Fig. 3.1. *Model domains - D1 81×81 km (Europe), D2 27×27 km (Balkan Peninsula), D3 9×9 km (Bulgaria), D4 3×3 km (Sofia region) and D5 1×1 km (Sofia city).*

simulations which is used for different studies of the atmospheric composition, including the AQ climate.

The atmospheric composition studies were based on extensive computer simulations carried out with good resolution using up-to-date modelling tools and detailed and reliable input data. All the simulations were based on the United States Environmental Protection Agency (US EPA) Model-3 system, which consists of 3 models: WRF [7] used as meteorological pre-processor; CMAQ [8, 9] and [10] – the Community Multiscale Air Quality System, being the Chemical Transport Model (CTM) and SMOKE [11, 12] – the Sparse Matrix Operator Kernel Emissions Modelling System – the emission pre-processor. The simulations were performed for 7 years period (2008 to 2014) with Two-Way Nesting mod on.

The large scale (background) meteorological fields, used by the application were taken from the National Centers for Environmental Prediction (NCEP) Global Analysis Data with 1°×1° resolution. The WRF and CMAQ nesting capabilities were used to downscale the simulations to a 9 km for domain D3 – Bulgaria and to a 1 km horizontal resolution for the innermost domain – Sofia. The simulations were carried out for 5 nested domains Figure 3.1. The used WRF model parametrizations and schemes are as follows: micro physics – WRF single moment 6-class , cumulus physics – Kain-Fritsch, boundary layer scheme – ACM2Pleim, surface physics – Pleim-Xiu Land Surface Model and the model vertical levels are 27.

The Bulgarian emission inventory was used as an emission input for Bulgaria, while outside the country the high resolution inventory of the the Netherlands Organization for Applied Scientific Research (TNO, see `https://www.tno.nl/en/`) with resolution 20×15 km (0.25°×0.125°) was exploited. The latest one is produced by proper disaggregation of the European Monitoring and Evaluation Program (EMEP) 50-km data base [13, 14]. In both inventories the emissions are distributed over 10 Selected Nomenclature for Sources of Air Pollution (SNAP) categories [15].

The Air Quality is a key element for the well-being and quality of life of the European citizens and that is why the AQ impact on human health and quality of life is an issue of great social significance. The AQ impact on human health and quality of life is evaluated in the terms of Air Quality Indices (AQI), which give an integrated assessment of the impact of pollutants and directly measuring the effects of AQ on the human health. The evaluations are based on extensive computer simulations of the AQ for Bulgaria and Sofia city carried out with good resolution using up-to-date modelling tools and detailed and reliable input data [16, 17, 18]. All the AQI evaluations are on the basis of air pollutant concentrations obtained from the numerical modelling and make it possible to reveal the climate of AQI spatial/temporal distribution and behavior.The AQI is defined as a measure of air pollution and provides an integrated assessment of the impact of the pollutants on human health. The index is defined in several segments, each of which is a linear function of the concentration of each pollutant

TABLE 3.1
Air Pollution Bandings and Index Impact on Human Health

Banding	Value	Health Descriptor
Low	1–3	Effects are unlikely to be noticed even by individuals who know they are sensitive to air pollutants
Moderate	4–6	Mild effects, unlikely to require action, may be noticed among sensitive individuals.
High	7–9	Significant effects may be noticed by sensitive individuals and action to avoid or reduce these effects may be needed. Asthmatics will find that their 'reliever' inhaler is likely to reverse the effects on the lung.
Very High	10	The effects on sensitive individuals described for 'High' levels of pollution may

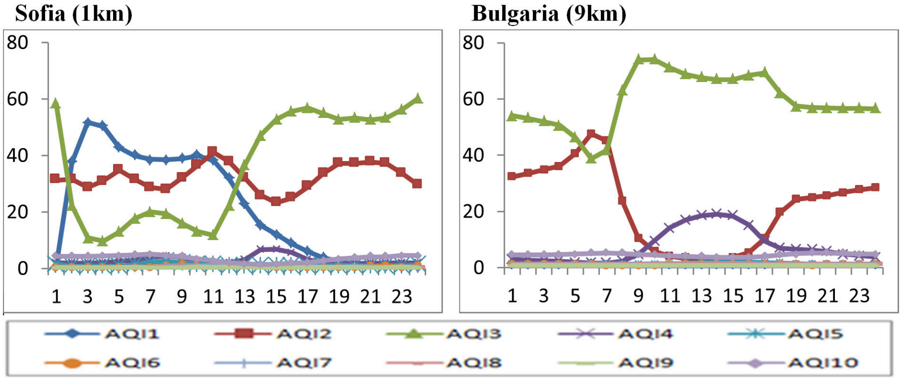


FIG. 3.2. *Annual Diurnal variations [%] of the different AQI (1 to 10) integrated over territory of Bulgaria and Sofia*

considered [19]. The index falls in different ranges of the dimensionless scale. In each range the index values are associated with an intuitive color code ((from green to red), a linguistic description (e.g. from very good to very poor) and a health description. In order to evaluate the air quality situation in Europe, all measurements are transformed into a single relative figure: the Common Air Quality Index (CAQI) which has 5 levels using a scale from 0 (very low) to > 100 (very high). The index is based on 3 pollutants of major concern in Europe: Particulate matter, with diameter <10µm (PM10), Nitrogen Dioxide NO_2, Ozone O_3 and will be able to take into account to 3 additional pollutants Carbon Oxide (CO), Particulate matter with diameter <2.5µm (PM2.5) and Sulphur Dioxide SO_2. In different countries use different AQI on basis of different monitor pollutants.

The index, calculated in Bulgaria in the frame of Bulgarian Chemical Weather Forecast System [20, 21, 22], follows the United Kingdom (UK) Daily Air Quality Index [23]. This index has ten grades, which are further grouped into 4 bands: low, moderate, high and very high and is based on the concentrations of 5 pollutants - NO_2, O_3, SO_2, CO and PM10 (Table 3.1). Different averaging periods are used for different pollutants. The reference levels and Health Descriptor used in the tables are based on health-protection related limit, target or guideline values set by the European regulations, at national or local level or by the World Health Organization [24, 25].

Annually averaged hourly values of the AQI for Sofia and Bulgaria with different horizontal grid resolution are presented in Figure 3.2. The graphs represent the daily percent recurrence of the AQI (1 to 10). This results, allow to follow highest recurrence of the indices during the day (during the seasons), and to analyze the possible reason for high values in the High and Very High bands. The meteorological conditions from one hand and the pollutant emissions from other one could be the cause for different possible AQI statuses. That representation of the index makes it possible to evaluate the atmospheric composition in the context of impacts on human health and quality of life.

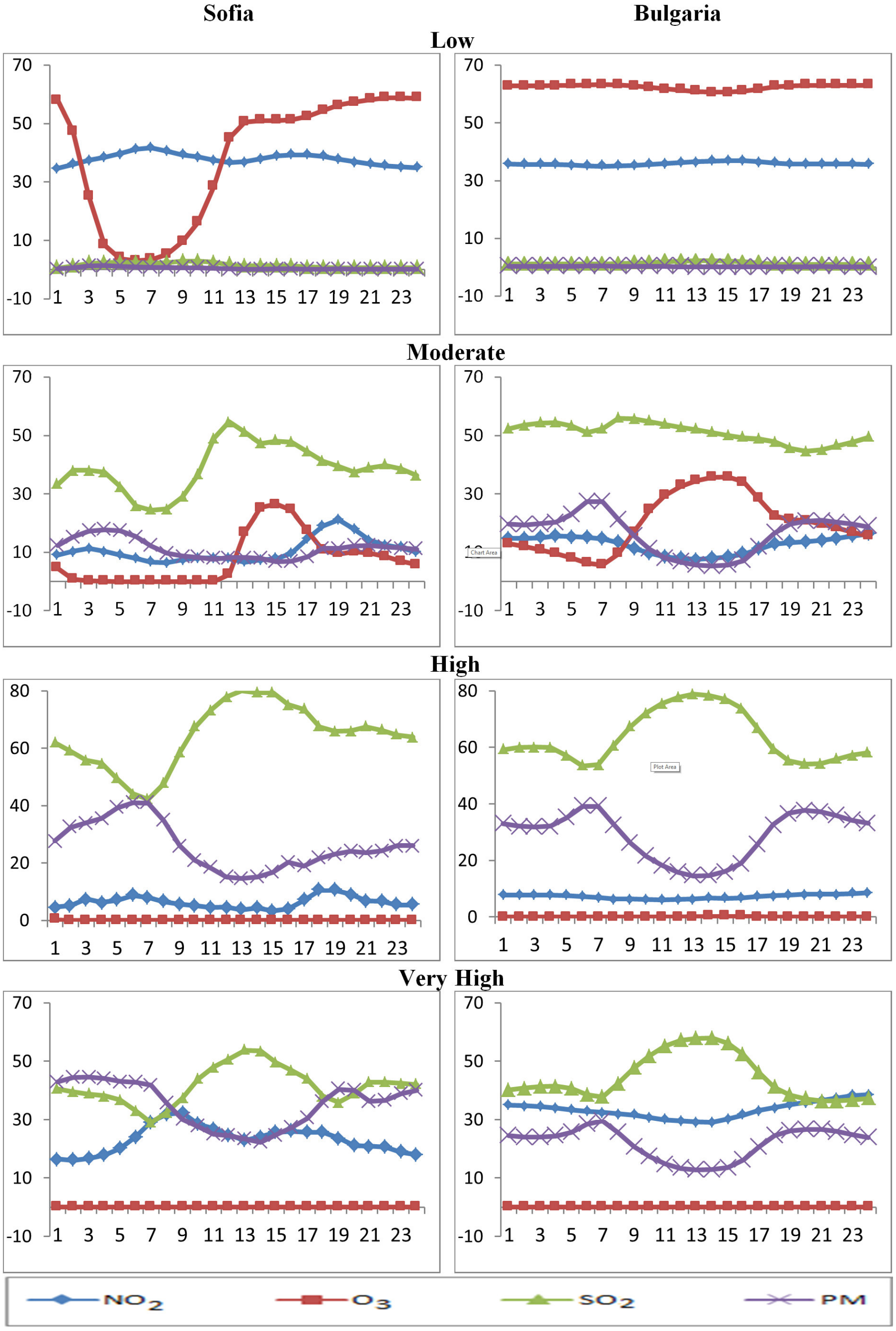


FIG. 3.3. *Diurnal variations of the annually averaged recurrence [%] of the dominant pollutant.*

The graphics on Figure 3.3 demonstrate the annual recurrence of the pollutant with highest AQI, which determines the overall AQI for the 4 bands (the dominant pollutant). The pollutants involved in the calculation of AQI – NO_2, O_3, SO_2 and PM are presented in different colours. The seasonal cases are not present here, but they differ from the annually averaged graphics. The dominant pollutants are different for each band with well displayed diurnal course.

The air pollution pattern is formed as a result of interaction of different processes, so knowing the contribu-

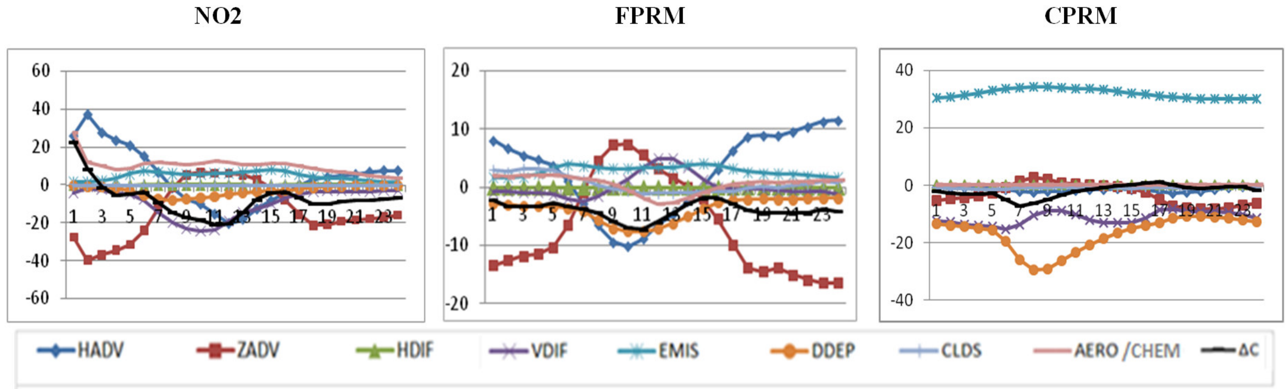


FIG. 3.4. *Annually averaged contribution of the different processes to the formation of NO_2 [$\mu g/m^3$] and FPRM, CPRM [pPMv/h] for Sofia city.*

tion of each one of these processes for different meteorological conditions and given emission spatial configuration and temporal behavior could be helpful for understanding the atmospheric composition formation and air pollutants behavior. Therefore the CMAQ "Integrated Process Rate Analysis" option was applied to discriminate the role of different dynamic and chemical processes for the air pollution formation. The procedure allows the concentration change for each compound for an hour ΔC to be presented as a sum of the contribution of the processes, which determine the concentration. The results were averaged over the whole ensemble and so the "typical" seasonal and annual evaluations were obtained.

The diurnal/annual behavior of the processes contribution to the surface concentrations change of pollutant NO_2, fine- and coarse particulate matter ($FPRM$ and $CPRM$), averaged for the territory of Sofia, is given in Figure 3.4.The considered processes are advection (horizontal – HADV – and vertical – VADV), diffusion (horizontal – HDIF – and vertical – VDIF), mass adjustment, emissions (EMIS), dry deposition (DDEP), chemistry (CHEM), aerosol processes (AERO) and cloud processes/aqueous chemistry (CLDS) and they are present in different colors.

The total concentration change (ΔC), leading to a change in a concentration is determined mainly by a small number of dominating processes which have large values, and could be with opposite sign and phases. The ΔC is different for each pollutant with well displayed seasonal and diurnal course. The sign of the contributions of some of the processes is obvious, but some of them may have different sign and it depends on the type of emissions, weather conditions and local atmospheric dynamics.

4. Conclusion related to ACIQLife. A very small part of the obtained results is presented in the present paper, just to demonstrate the opportunity HPC platforms give for detailed and extensive study of the atmospheric composition its behavior, origin and health impact. Due to volume limitations the spatial variability of the air pollution characteristics is not demonstrated at all.

The generated ensembles of atmospheric composition characteristics have still to be carefully and extensively treated and analyzed, which will be objective of the future work of the authors.

5. TVRegCM application. The simulations with the RCM RegCM version 4.4 [26] were made for the SE Europe covering ten years period from 01.12.1999 to 30.11.2009 and are driven by the ERA-Interim reanalysis [27], providing the required atmospheric Initial and Boundary Conditions (ICBC) as well as sea surface temperatures. The ERA-Interim boundary conditions can be considered to be of very high quality [27], particularly in the Northern Hemisphere extratropical areas where reanalysis uncertainty is negligible [28]. The simulation domain covers entirely the Balkan Peninsula, a minor part of Italy and a part of Asia Minor Peninsula. The model grid is in Lambert Conformal Conic projection with spatial resolution of 10 km. Hence the previous experiments reveals that time step equal to 25 seconds, and 27 vertical levels are optimal, they are selected for the model integration. The default land surface parameterization method in RegCM4 is the BATS scheme [29]. In the current study, we have used it without the subgridding option. The considered planetary boundary layer (PBL) schemes are the one proposed by Holstlag [30, 31] and the University of Washington (UW) [32, 33]. One of the most significant novelties in RegCM4.4 is the incorporation of the new cloud microphysics scheme (for brevity: M-scheme), proposed by Nogherotto and Tompkins (NT) [34]. EURO-CORDEX (`http://www.euro-cordex.net/`) is the European branch of the international CORDEX initiative, which is a program sponsored by the World Climate Research Program (WRCP) to organize an internationally coordinated framework to produce improved regional climate change projections for all land regions world-wide. Med-CORDEX (`https://www.medcordex.eu/`) project has been proposed by the Mediterranean climate research community within EURO-CORDEX as a follow-up of previous and existing initiatives. The NT-scheme was released after MedCORDEX experiments started. The cumulus convection (CC) parameterizations include Grell scheme [35] with Arakawa-Schubert (AS) [36] and Fritsch-Chappell (FC) closure assumption [37], Emanuel scheme [38, 39], Tiedtke scheme [40] and Kain-Fritsch scheme [41, 42]. The simulations with Kuo [43] convective parameterization scheme have shown instability and interruptions of the model simulations at some periods, so was not used in the present research.

Thus, the number of the possible combinations, which means RegCM4.4 model set-ups, between 2 PBL schemes, 2 M-schemes and 5 CC ones, is 20 and the performance of all of them have been investigated.

The well-known and freely available for the research community data-base E-OBS version 12.0 of the European Climate Assessment & data-set (ECA&D) project [44] is used as reference in the model validation. E-OBS is based only on observations, covers entire Europe and the surroundings, and the version with 0.25°×0.25° regular grid spacing is implemented. It is worthy to emphasize that E-OBS is the standard validation data-base for the EURO-CORDEX.

Hence the multi-annual seasonal mean temperature (referred further for brevity only temperature) and the multi-annual seasonal mean precipitation sum (precipitation) are probably the most important quantities from climatological point of view, the validation study thus far is focussed on them. The E-OBS is on daily basis and RegCM is set to produce output on every 6 hours. Thus, the climatological quantities are calculated after every successive model run with the CDO operators [45]. The detailed results from the validation are presented in [46]. Only the most relevant conclusion will be listed briefly here.

According to the simulated temperature behavior, the models can be divided in two groups those with prevailing warm bias and those with prevailing cold bias. Generally, the biases are more remarkable in the summer than in the winter and are in the interval from about -3.5°C to 3.5°C, but over the bigger part of the domain typically from about -2°C to 2°C.

The simulation outcome from all 20 model set-ups produces almost identical picture for the precipitation distribution in winter: The biases are nearly equally distributed and are positive (i.e. the model overestimates the precipitation), with some minor exceptions. The summer biases however, show significant distinction in their distribution and magnitude. They are positive, with some minor exception in Greece. Generally, the precipitation biases however, are very big. Their values vary from below -100% to above 160%.

The main conclusions are, first, that the relative weight of the CC-schemes is the biggest and, second, the simulations with the smallest biases are with Grell one with the both closures. The sensitivity of RegCM4.4 to the PBL- and M-scheme seems is significantly weaker. Thus, there are not clear evidences for clear distinction between the model skill with Holstlag or UW PBL parameterization from one side or for over performance of the NT M-scheme in comparison with the default SUBEX. As overall, 7 from 20 model setups show recognizable better performance. They are listed in Table 5.1.

Main aim of the current, second stage of TVRegCM is to "narrow" the selection, i.e. to perform further examination of these 7 model configurations.

Our previous results [46, 47] indicate that the biases are bigger in summer. Thus, we will use another approach to assess the model performance in that season, called Taylor diagram [48]. The observation data base is E-OBS version 13.1, but the differences with the version 12.0 are insignificant for our purpose. We will consider the temporal and spatial Taylor diagrams of the normalized (in such a way that the observations standard deviation is equal to the model results with respect to the temporal and spatial variability respectively, for the mean summer daily 2m temperature and the mean summer daily rainfall from 2001 to 2008 years. The spatial diagram is constructed from the season average for each location, and the temporal diagram from the spatial average by whole domain on daily basis. The correlation is shown by radial dashed lines, and the

Table 5.1

List of the model set-ups with better (in comparison with the others) performance. The original index and notation is preserved from the first stage of TVRegCM experiment

Index	Notation	ICBC	PBL-scheme	M-scheme	CC-scheme
1	r11111	EIN15	Holtslag	SUBEX	Grell/FC
2	r11112	EIN15	Holtslag	SUBEX	Grell/AS
5	r11155	EIN15	Holtslag	SUBEX	Kain-Fritsch
11	r12121	EIN15	UW	SUBEX	Grell/AS
12	r12122	EIN15	UW	SUBEX	Grell/FC
15	r12155	EIN15	UW	SUBEX	Kain-Fritsch
16	r12221	EIN15	UW	Nogherotto/Tompkins	Grell/AS

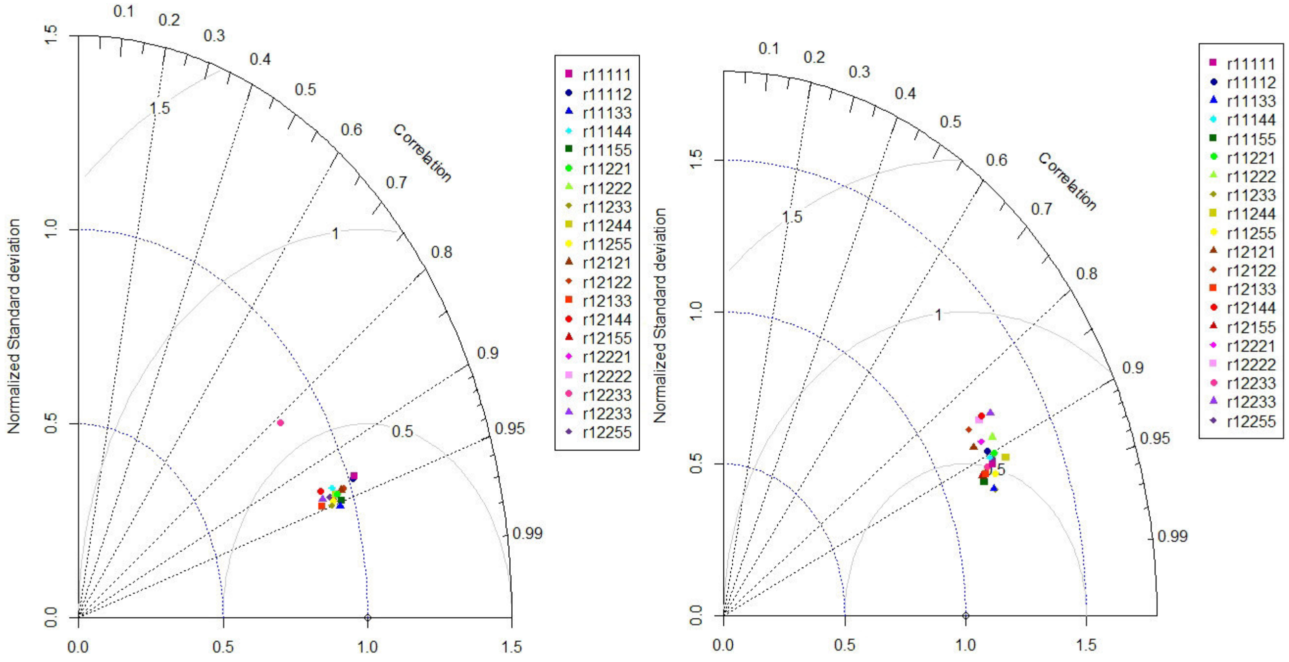


Fig. 5.1. *Taylor diagrams for the mean summer daily temperature concerning the temporal (on left) and spatial (on right) variability*

normalized standard deviation on the horizontal and vertical axes. The normalization is made in a way that the reference standard deviation is 1. The lines of centered root mean square difference (RMSD) values are also given.

The Taylor diagram of the mean summer daily temperature for temporal variability is given on the left pane of Figure 5.1. The simulations are depicted by solid color points, and the reference E-OBS data by an empty circle. The normalized standard deviations are below 1.0, except for the cases r11111 and r11112. All simulations except r12222 have RMSD below 0.5. Although relatively small differences, the simulation cases with the slightly better than other cases performances are r11133, r11233, r12133. The Taylor diagram of the mean summer daily temperature concerning the spatial variability is shown on the right pane of Figure 5.1. The correlation coefficient is between about 0.85 and 0.95. The normalized standard deviation is between 1.0 and 1.5, and the RMSD in most cases is above 0.5. The simulations with the best performance are r11233, r11133 and r11155.

The Taylor diagram with respect of the temporal variability of the mean summer daily precipitation is shown on the left pane of Figure 5.2. The normalized standard deviations are between 0.75 and 1.5. The RMSD

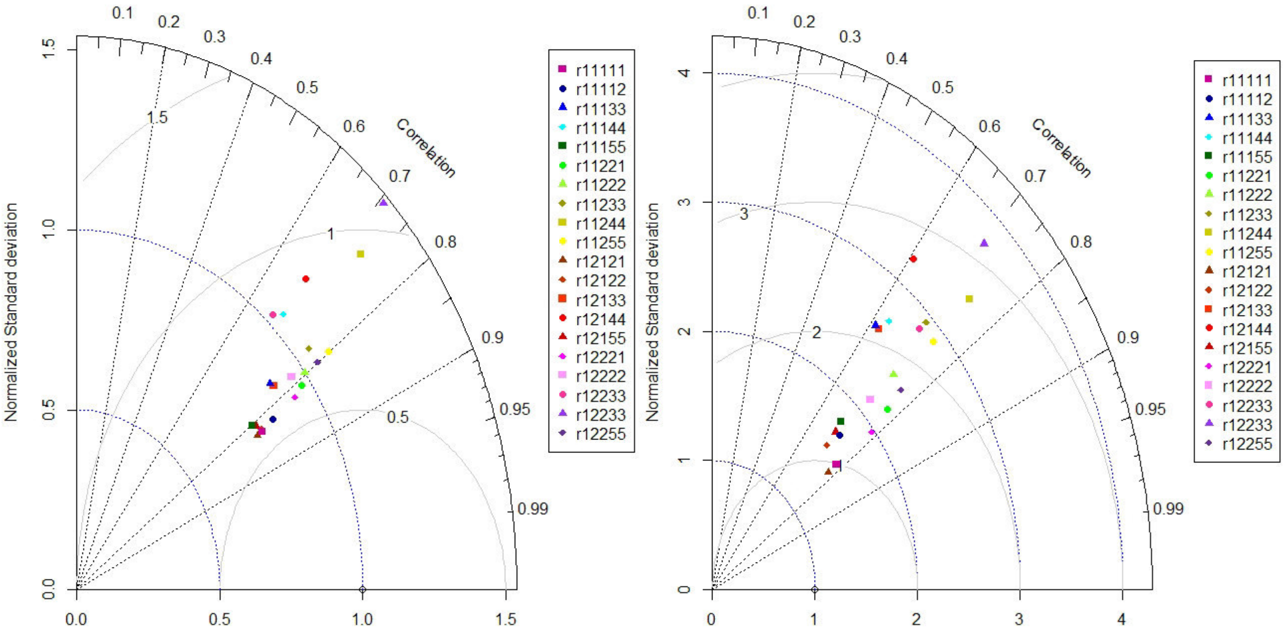


FIG. 5.2. *Same as Figure 5.1, but for the precipitation*

are between 0.5 and 1.1, and the correlation is between 0.65 and 0.85. The simulations results are much more scattered than the ones for the mean daily temperature. The performance of the cases is more distinguishable, and the best ones are r12121, r12122, r11111, r12155 and r11155.

The Taylor diagram of the the mean summer daily precipitation with respect to the spatial variability is shown on the right pane of Figure 5.2. The scattering of the simulation points are bigger than for the mean daily temperature ones, as in the Taylor diagram with respect to the temporal variability. The correlation coefficient is between 0.6 and 0.8. The normalized standard deviation is from about 1.2 to 4.0, and the RMSD is between 1.5 and 2.5. The cases with the best performance are r12121 and r11111, although r12122, r11112, r12155 and r11155 form a cloud of points with slightly worse performance but with normalized standard deviation below 2.

6. Conclusion related to TVRegCM. The main conclusion of the presented part of the RegCM numerical experiment is that our new test does not reveal single one model set-up that definitely over performs the other considered ones. Nevertheless this exercise was a necessary step forward in the authors' evaluation strategy.

The results of the model temperature field lead us to the following conclusions. The spatial variability is bigger than the temporal one. It is worth to note that the choice of the boundary layer scheme also has some meaning in a spatial variation meaning. The cases with Holstag boundary layer scheme show more resemblance with the observations, than the simulations with UW boundary layer scheme. We can note that the best convective scheme concerning the temperature field is of the Tiedtke.

The spatial variability of the precipitation field is bigger than the temporal one, and from the spatial one for the temperature. The results for the model mean daily precipitation variability on the other hand is as much as about 2 times the reference one and the correlation is weaker than the one for the temperature. Therefore, the model performance is worse for the precipitation than for the temperature. The results suggest that the Grell scheme with FC or AS closures is the best scheme for the precipitation simulation. Although, the Kain-Fritsch cumulus convective scheme with SUBEX moisture scheme and Holstag boundary layer parameterization scheme is also a good case. These results confirm our previous ones [46, 47] that the results for temperature are more spatially homogeneous and the correlation for the temperature field is higher than the one for precipitation. The bigger, in comparison with the temperature, spread of the results on the Taylor diagram for the precipitation demonstrates the bigger sensitivity of this output variable from the parameterization selection and combination from different schemes. The results, together with these from the previous stage, are in general agreement with the outcomes in [49] and [50]. In particular, we confirm the outlined in [50] primary importance of the convection scheme. Obviously, many other factors have to be investigated, including:

- It is relevant to investigate the model option to switch the CC-scheme by transition from land to sea and vice versa. It is worth to emphasize that the default setting (and it is explicitly recommended from the RegCM authors'), which is confirmed in [50], is Grell over land and Emanuel over sea.
- It is necessary to perform sensitivity tests over shorter periods, including case studies for warm/cold/ wet/dry years.
- Other output quantities, which are more or less also relevant for many practical applications, such as cloud cover, soil moisture, radiation fluxes, etc should be also considered. Although the availability of independent data-sets, which can be used as reference, seems limited, this further step seems is reasonable.
- The computational efficiency of the selected model set-ups should be estimated.

7. Conclusion. The virtual research environment platform allows to different scientific communities to make research which require big computational and storage resources. A small part of the obtained results from both applications are presented in the present paper, just to demonstrate the opportunity of HPC platforms.

Acknowledgments. Deep gratitude to the organizations and institutes (ICTP, ECMWF, NCEP-NCAR, ECA&D, Unidata, MPI-M, EMEP, TNO, and all others), which provides free of charge software and data. Without their innovative data services and tools this study would not be possible.

Special thanks to:

- VRE for regional Interdisciplinary communities in Southeast Europe and the Eastern Mediterranean (VI-SEEM), Horizon 2020 project 675121
- The EC-FP7 grant PIRSES-GA-2013-612671 (project REQUA)
- Bulgarian National Science Fund (grant DN-04/2/13.12.2016)
- Program for career development of young scientists, BAS.

REFERENCES

[1] E. Atanassov, T. Gurov, A. Karaivanova, S. Ivanovska, M. Durchova and D. Dimitrov *On the Parallelization Approaches for Intel MIC Architecture*, AIP Conf. Proc. 1773, 070001 2016 http://dx.doi.org/10.1063/1.4964983.

[2] A. Radenski, T. Gurov, et al *Big Data Techniques, Systems, Applications, and Platforms: Case Studies from Academia*, in Annals of Computer Science and Information Systems, Volume 8, Proceedings of the 2016 Federated Conference on Computer Science and Information Systems" FedCSIS'16, 2016, pp. 883–888 DOI:http://dx.doi.org/10.15439/978-83-60810-90-3.

[3] https://vi-seem.eu/

[4] https://training.vi-seem.eu/index.php/domain-specific-software-and-tools/climate-software-and-tools#tuning-and-validation-of-the-regcm-tvregcm

[5] https://training.vi-seem.eu/index.php/domain-specific-software-and-tools/climate-software-and-tools#atmospheric-composition-impact-on-quality-of-life-and-human-health-wrf-cmaq-aciqlife

[6] https://accounting.vi-seem.eu

[7] W. Shamarock, J. Klemp, J. Dudhia, D. Gill, D. Barker, W. Wang, J. Powers *A description of the Advanced Research WRF Version 2* 2007 http://www.mmm.ucar.edu/wrf/users/docs/arw_v2.pdf

[8] D. Byun, J. Young, G. Gipson, J. Godowitch, F.S. Binkowski, S. Roselle, B. Benjey, J. Pleim, J. Ching, J. Novak, C. Coats, T. Odman, A. Hanna, K. Alapaty, R. Mathur, J. McHenry, U. Shankar, S. Fine, A. Xiu, and C. Jang *Description of the Models-3 Community Multiscale Air Quality (CMAQ) Modeling System*, 10th Joint Conference on the Applications of Air Pollution Meteorology with the A&WMA, 11–16 January 1998, Phoenix, Arizona, 264–268.

[9] D. Byun and J. Ching *Science Algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) Modeling System*. EPA Report 600/R-99/030, Washington DC, 1999 http://www.epa.gov/asmdnerl/models3/doc/science/science.html

[10] *CMAQ user guide* 2006. [Online] Available from: https://www.cmascenter.org/help/documentation.cfm?model=cmaq&version=4.6

[11] CEP *Sparse Matrix Operator Kernel Emission (SMOKE) Modeling System*, University of Carolina, Carolina Environmental Programs, Research Triangle Park, North Carolina, 2003.

[12] D. Schwede, G. Pouliot, and T. Pierce *Changes to the Biogenic Emissions Inventory System Version 3 (BEIS3)*, Proc. of 4th Annual CMAS Models-3 Users's Conference, September 26-28, 2005, Chapel Hill, NC.

[13] V. Vestreng *Emission data reported to UNECE/EMEP: Evaluation of the spatial distribution of emissions* Meteorological Synthesizing Centre – West, The Norwegian Meteorological Institute, Oslo, Norway, Research Note 56, EMEP/MSC-W Note 1/2001, 2001.

[14] V. Vestreng, K. Breivik, M. Adams, A. Wagner, J. Goodwin, O. Rozovskaya, J.M. Pacyna *Inventory Review 2005 (Emission Data reported to LRTAP Convention and NEC Directive)*, Technical Report MSC-W 1/2005, EMEP, 2005

[15] A. Visschedijk, P. Zandveld, H. van der Gon *A high resolution gridded European emission database for the EU integrated project GEMS* TNO report 2007-A-R0233/B, The Netherlands Brunekreef B, Holgate S: Air pollution and health., Lancet 2002, 2007, 360: 1233–1242

[16] Gadzhev, G., Ganev, K., Prodanova, M., Syrakov, D., Atanasov, E., Miloshev, N *Multi-scale Atmospheric Composition Modelling for Bulgaria* NATO Science for Peace and Security Series C: Environmental Security, 137, 2013, 381–385.

[17] Gadzhev G., K. Ganev, N. Miloshev, D. Syrakov, and M. Prodanova *Analysis of the Processes Which Form the Air Pollution Pattern over Bulgaria* in I. Lirkov et al. (Eds.): LSSC 2013, LNCS 8353, Springer-Verlag Berlin Heidelberg, 2014, 390–396.

[18] Gadzhev G., K. Ganev, N. Miloshev, D. Syrakov, and M. Prodanova *Some Basic Facts About the Atmospheric Composition in Bulgaria – Grid Computing Simulations* in I. Lirkov et al. (Eds.): LSSC 2013, LNCS 8353, Springer-Verlag Berlin Heidelberg, 2014, 484–490.

[19] EPA *Technical assistance document for the reporting of daily air quality–the Air Quality Index (AQI)* EPA- 454/B-09-001, US Environmental Protection Agency, Research Triangle Park, North Carolina, Office of Air Quality Planning and Standards, Research Triangle Park, North Carolina 27711, 2009.

[20] Syrakov, D., Etropolska, I., Prodanova, M., Ganev, K., Miloshev, N., Slavov, K. *Operational Pollution Forecast for the Region of Bulgaria*, American Institute of Physics, Conf. Proc. 1487, 2012, 88–94;

[21] Syrakov, D., Etropolska, I., Prodanova, M., Slavov, K., Ganev, K., Miloshev, N., Ljubenov T. *Downscaling of Bulgarian Chemical Weather Forecast from Bulgaria region to Sofia city*, American Institute of Physics, Conf. Proc. 1561, 2013, 120–132.

[22] Syrakov D., M. Prodanova, I. Etropolska, K. Slavov, K. Ganev, N. Miloshev, and T. Ljubenov *A Multy-Domain Operational Chemical Weather Forecast System* in I. Lirkov et al. (Eds.): LSSC 2013, LNCS 8353, Springer-Verlag Berlin Heidelberg 2014, 413–420,

[23] Leeuw, F. de, Mol, W. *Air Quality and Air Quality Indices: a world apart* ETC/ACC Technical Paper 2005/5, 2005, `http://acm.eionet.europa.eu/docs/ETCACC_TechnPaper_2005_5_AQ_Indices.pdf`

[24] World Health Organization (WHO) *Fact Sheet Number 187*, 2000

[25] World Health Organization (WHO) *Health Aspects of Air Pollution. Results from the WHO Project Systematic Review of Health Aspects of Air Pollution in Europe*, 2004.

[26] F. Giorgi and 20 others *RegCM: model description and preliminary tests over multiple CORDEX domains*, Clim. Res., 52, 2012, 7–29

[27] D. P. Dee, S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, H., E. V. Hèlm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J. J. Morcrette, B. K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J. N. Thèpaut and F. Vitart *The ERA-Interim reanalysis: configuration and performance of the data assimilation system.* Q.J.R. Meteorol. Soc., 2011, 137: 553–597. `doi:10.1002/qj.828`

[28] S. Brands, J. M. Gutiérrez, S. Herrera and A. S. Cofiño *On the Use of Reanalysis Data for Downscaling*, J. Climate, 25, 2012, 2517–2526.

[29] R. E. Dickinson, A. Henderson-Sellers, and P. J. Kennedy *Biosphere-atmosphere transfer scheme (BATS) version 1e as coupled to the NCAR community climate model*, Tech. rep., National Center for Atmospheric Research, 1993

[30] A. A. M. Holtslag, E. I. F. de Bruijn and H.-L. Pan *A high resolution air mass transformation model for shortrange weather forecasting*, Mon. Wea. Rev., 118, 1990, 1561–1575.

[31] A. A. M Holtslag and B. A. Boville *Local versus nonlocal boundary-layer diffusion in a global climate model*, J. Climate, 6, 1993, 1825–1842

[32] C. S. Bretherton, J. McCaa, and H. Grenier *A new parameterization for shallow cumulus convection and its application to marine subtropical cloud-topped boundary layers. part I: Description and 1D results*, Monthly Weather Review, 132, 2004, 864–882

[33] H. Grenier and C. S. Bretherton *A moist PBL parameterization for large-scale models and its application to subtropical cloud-topped marine boundary layers*, Monthly Weather Review, 129, 2001, 357–377

[34] N. Elguindi, X. Bi, F. Giorgi, B. Nagarajan, J. Pal, F. Solmon, S. Rauscher, A. Zakey, T. O'Brien, R. Nogherotto and G. Giuliani *Regional Climate Model RegCM User Manual Version 4.4.*, 2014, p.34, ICTP, Trieste

[35] G. Grell *Prognostic evaluation of assumptions used by cumulus parameterizations*, Mon. Wea. Rev., 121, 1993, 764–787

[36] A. Arakawa, W. H. Schubert *Interaction of a cumulus cloud ensemble with the large scale environment. Part I.* J. Atmos. Sci., 31, 1974, 674–701.

[37] J. M. Fritsch and C. F. Chappel *Numerical prediction of convectively driven mesoscale pressure systems. Part I: Convective parameterization.* J. Atmos. Sci., 37, 1980, 1722–1733

[38] K. A. Emanuel *A scheme for representing cumulus convection in large-scale models*, J. Atmos. Sci., 48(21), 1991 ,2313–2335

[39] K. A. Emanuel and M. Zivkovic-Rothman. *Development and evaluation of a convection scheme for use in climate models*, J. Atmos. Sci., 56, 1999, 1766–1782

[40] M. Tiedtke *A Comprehensive Mass Flux Scheme for Cumulus Parameterization in large-scale models.* Bulletin of the American Meteorological Society, 117, 1989, 1779–1800

[41] J. S. Kain *The Kain-Fritsch convective parameterization: an update.* J Appl Meteorol 43, 2004,170–180

[42] J.S. Kain and J. M. Fritsch *A one-dimensional entraining/detraining plume model and its application in convective parameterization.* J Atmos Sci 47, 1990 , 2784–2802

[43] R. A. Anthes *A cumulus parameterization scheme utilizing a one-dimensional cloud model*, Mon. Wea. Rev., 105, 1977, 270–286

[44] M. R. Haylock, N. Hofstra, A. M. G. Klein Tank, E. J. Klok, P.D. Jones, M. New. *A European daily high-resolution gridded dataset of surface temperature and precipitation.* J. Geophys. Res (Atmospheres), 2008, p. 113.

[45] CDO 2015: Climate Data Operators. Available at: `http://www.mpimet.mpg.de/cdo`

[46] Gadzev G., Ivanov, V., Ganev K., Chervenkov H. *TVRegCM Numerical Simulations – Preliminary Results*, In: Lirkov I., Margenov S. (eds) Large-Scale Scientific Computing. LSSC 2017. Lecture Notes in Computer Science, 2018 vol 10665, 266–274, Springer, Cham

[47] Chervenkov H. Ivanov, V., Gadzev G., Ganev K. *Sensitivity Study of Different RegCM4.4 Model Set-Ups–Recent Results from the TVRegCM Experiment* Cybernetics and Information Technologies Vol. 17, No. 5 17–26

[48] Taylor K. E. *Summarizing multiple aspects of model performance in a single diagram* Journal of Geophysical Research, 106, 2001, 7183–7192

[49] S. Kotlarski, K. Keuler, O. B. Christensen, A. Colette, M. Déqué, A. Gobiet, K. Goergen, D. Jacob, D. Lüthi, E. van Meijgaard, G. Nikulin, C. Schär, C. Teichmann, R. Vautard, K. Warrach-Sagi, and Wulfmeyer, V. *Regional climate modeling on European scales: a joint standard evaluation of the EURO-CORDEX RCM ensemble.* Geosci. Model Dev., 7, 2014, 1297–133

[50] I. Pieczka, R. Pongrcz, K. S. Andrè, F. D. Kelemen, J. Bartholy *Sensitivity analysis of different parameterization schemes using RegCM4.3 for the Carpathian region*, Theor Appl Climatol, 130, 2017, 1175–1188

Edited by: Aneta Karaivanova
Received: Dec 21, 2017
Accepted: Mar 30, 2018

DOI 10.12694/scpe.v19i2.1352
Scalable Computing: Practice and Experience ISSN 1895-1767
Volume 19, Number 2, pp. 119–130. http://www.scpe.org © 2018 SCPE

DYNAMIC VERSUS STATIC APPROACH TO THEORETICAL ANHARMONIC VIBRATIONAL SPECTROSCOPY OF MOLECULAR SPECIES RELEVANT TO ATMOSPHERIC CHEMISTRY: A CASE STUDY OF FORMIC ACID

BOJANA KOTESKA,[*] VERČE MANEVSKA,[†] ANASTAS MISHEV,[‡] AND LJUPČO PEJOV[§]

Abstract. Vibrational spectra of the two conformers of the free formic acid molecule are computed by two approaches, with a special emphasis on the region of O-H stretching modes. The first approach (referred to as a static one) is based on sequential computation of anharmonic O-H stretching vibrational potential and numerical solution of the vibrational Schrödinger equation by the Numerov method. The second approach (referred to as a dynamic one) is based on molecular dynamics (MD) simulations performed within the atom-centered density matrix propagation scheme (ADMP) followed by spectral analysis of the velocity-velocity and dipole moment autocorrelation functions computed from the ADMP MD trajectories. All calculations are carried out within the density functional tight binding (DFTB) formalism. The computed properties are compared to the available experimental data and the advantages of the dynamic versus the static approach are outlined and analyzed in the context of detection of individual and non-covalently bonded molecular species relevant to climate science and atmospheric chemistry.

Key words: formic acid, atmospheric chemistry, molecular dynamics, atom-centered density matrix propagation scheme, *anharmonic* vibrational frequencies, statistical physics simulations, theoretical spectroscopy, density functional tight binding.

AMS subject classifications. 70F99, 82B30, 92E10

1. Introduction.

In the last decade, we have evidenced a thorough paradigmatic shift in many scientific areas, with a notable emphasis on fundamental (i.e. molecular-level) understanding of a wide variety of processes. Within the climate science community in particular, it has become clear that for a thorough in-depth understanding of various phenomena and processes taking place in the Earth's atmosphere, it is necessary to view this complex system at molecular level. Thus, the mesoscale-level models are effectively being replaced by more general molecular models. This, on the other hand, has led to the establishment of chemistry-climate models at various level of sophistication. Numerous molecular species are present in the Earth's atmosphere. However, from the viewpoint of both biosphere and atmospheric chemistry, organic acids are especially important. The title compound studied in the present paper formic acid is the simplest organic acid. It is in a sense a prototypical molecular system for this whole class of compounds. At the same time, it is the most abundant (and ubiquitous) organic acid in the atmosphere [1, 2, 3]. It is generated both by anthropogenic as well as biogenic factors and it is widespread in aerosols and also in atmospheric precipitates (acid rain in particular). Aside from the Earth's atmosphere, formic acid is a prototypical "astrophysical molecule" as well. Numerous studies have been devoted to understanding of the physical and chemical processes that govern its formation in interstellar and cometary ices [3, 4, 5, 6, 7, 8, 9, 10].

From a purely fundamental aspect, the formic acid molecule is a prototypical molecular system exhibiting two rotameric forms due to intramolecular hindered rotational motion around the single C-O bond. Existence of two conformational isomers in this simple molecule significantly enriches the spectral appearance in the region where bands due to the O-H stretching modes are expected to appear in the vibrational (IR, Raman, inelastic neutron scattering etc.) spectra of gaseous formic acid, as well as in certain noncovalently bonded dimers thereof [1-10]. This spectral region is rather characteristic for detection of noncovalent interactions in which formic acid takes part. Therefore, understanding of the exact shape of this spectral region in the case of free formic acid is of essential importance for further analysis of its appearance in more complex clusters in which it participates. Vibrational modes that involve motion of hydrogen atoms are known to be strongly

[*] Faculty of Computer Science and Engineering, "Ss. Cyril and Methodious University", Rugjer Boskovikj 16, 1000 Skopje, Republic of Macedonia (bojana.koteska@finki.ukim.mk).

[†] Institute of Chemistry, Faculty of Science, "Ss. Cyril and Methodius University", P.O. Box 162, 1001 Skopje, Republic of Macedonia.

[‡] Faculty of Computer Science and Engineering, "Ss. Cyril and Methodious University", Rugjer Boskovikj 16, 1000 Skopje, Republic of Macedonia (anastas.mishev@finki.ukim.mk).

[§] Institute of Chemistry, Faculty of Science, "Ss. Cyril and Methodius University", P.O. Box 162, 1001 Skopje, Republic of Macedonia(ljupcop@pmf.ukim.mk).

anharmonic. For accurate prediction of the X-H stretching vibrational frequencies, therefore, it is crucial to go beyond the often employed harmonic approximation. Further, all the routine computations with widely used quantum chemical codes (either in harmonic approximation, or using an anharmonic approach e.g. by perturbation theoretical treatment [11]) give results that refer to 0 K. Most of the experimental data, however, have been collected at temperatures significantly higher than absolute zero. Having in mind that intramolecular torsional motions can be thermally activated, it is interesting to establish a computationally feasible approach in which the temperature effects could be explicitly accounted for.

Continuing our previous work in the field [12, 13], in the present study we tackle both of the previously mentioned issues. We compare the performances of static and dynamics methodologies for computation of anharmonic O-H stretching frequencies of the two rotameric forms of free formic acid molecule. The static approach is based on computation of anharmonic O-H stretching vibrational potentials of both conformers followed by numerical solution of the vibrational Schrödinger equation. The dynamic approach, on the other hand, is based on sequential molecular dynamic simulation employing the atom-centered density matrix propagation scheme (ADMP [14]) followed by analysis of several autocorrelation functions computed from the MD trajectories [15].

2. Computational details.

2.1. General theoretical methodology and calculation of anharmonic O-H stretching vibrational frequencies. All calculations carried out for the purpose of the present study were performed with the Density functional tight binding methodology (DFTB [16, 17]). The so-called DFTB-A variant of this theoretical approach was used, which is based on usage of analytic form of the relevant matrix elements. The potential energy surface (PES) of free formic acid was thoroughly investigated, employing Schlegel's gradient optimization algorithm [18]. The two minima on the PES, corresponding to the cis- and trans-rotameric forms of this molecular system were located and further characterized. The true-minimum character of the located stationary points on the studied PES was proven by harmonic vibrational analysis performed subsequently to geometry optimization. Absence of negative eigenvalues of the Hessian matrices confirmed that true minima are in question.

To compute the anharmonic O-H stretching vibrational frequencies of the two rotamers of formic acid molecule, we have relied on an implementation of a localized mode approach. We start from the equilibrium geometries (corresponding to the minima on the considered PESs) and generate a series of configurations by moving simultaneously the oxygen and hydrogen atoms in a manner that resembles the realistic motion taking place during the excitation of the O-H stretching vibration (i.e. keeping the center-of-mass of the O-H bond fixed). In the course of these movements, the O-H distances were varied from 0.85 to 1.55 Å. To generate these configurations, we have used our in-house developed FORTRAN code. A series of 15 single-point energy calculations were further carried out for the series of geometries generated this way to compute the vibrational potentials $V(r_{\mathrm{OH}})$. Subsequently, the resulting vibrational Schrödinger equation was solved by the finite-difference Numerov method [12, 13]. Frequencies of the fundamental $|0> \rightarrow |1>$ vibrational transitions were calculated from the energies of the ground ($|0>$) and first excited ($|1>$) vibrational energy levels.

2.2. Density functional tight binding (DFTB) molecular dynamics simulations with the atom-centered density matrix propagation (ADMP) scheme. To account explicitly for the finite-temperature effects on molecular structure and dynamics, we have carried out molecular dynamics simulations of free formic acid molecule by the atom-centered density matrix propagation (ADMP) scheme. These simulations were carried out on the basis of forces and energies computed with the DFTB approach, as described in the previous section. The particular MD method that has been implemented in the present study belongs to the extended Lagrangian approaches to molecular dynamics. It is actually based on propagation of the density matrix, using Gaussian-type basis functions [14, 19, 20]. The extended Lagrangian of the studied system is written in the form:

$$L = \frac{1}{2}Tr(V^T MV) + \frac{1}{2}\mu Tr(WW) - E(R,P) - Tr[\Lambda(PP-P)] \tag{2.1}$$

In (2.1), M, R and V are the nuclear masses, positions and velocities, respectively, while P, W and μ denote the density matrix, density matrix velocity and the fictitious mass for the electronic degrees of freedom, correspondingly. Λ is a Lagrangian multiplier matrix, and is here used to impose the constraints on the total number of electrons in the system and on the condition of idempotency of the density matrix. If one applies the principle of stationary action, the Euler-Lagrange equations for density matrix propagation can be written in the form:

$$\mu \frac{d^2 P}{dt^2} = -\left[\left. \frac{\partial E(R,P)}{\partial P} \right|_R + \Lambda P + P\Lambda - \Lambda \right] \tag{2.2}$$

$$M \frac{d^2 R}{dt^2} = - \left. \frac{\partial E(R,P)}{\partial R} \right|_P \tag{2.3}$$

Throughout the present study, we have used the velocity Verlet algorithm to integrate numerically equations (2.2) and (2.3). The time-evolution of the density matrix (i.e. its propagation) is given by:

$$P_{i+1} = P_i + W_i \Delta t - \frac{\Delta t^2}{2\mu} \left[\left. \frac{\partial E(R_i,P_i)}{\partial P} \right|_R + \Lambda_i P_i + P_i \Lambda_i - \Lambda_i \right] \tag{2.4}$$

$$W_{i+1/2} = W_i - \frac{\Delta t}{2\mu} \left[\left. \frac{\partial E(R_i,P_i)}{\partial P} \right|_R + \Lambda_i P_i + P_i \Lambda_i - \Lambda_i \right] = \frac{P_{i+1} - P_i}{\Delta t} \tag{2.5}$$

$$W_{i+1} = W_{i+1/2} - \frac{\Delta t}{2\mu} \left[\left. \frac{\partial E(R_{i+1},P_{i+1})}{\partial P} \right|_R + \Lambda_{i+1} P_{i+1} + P_{i+1} \Lambda_{i+1} - \Lambda_{i+1} \right] \tag{2.6}$$

Within the approach based on extended Lagrangian molecular dynamics with ab initio or semiempirical quantum mechanical Hamiltonian, the electronic subsystem is not treated by a full solution (e.g. by a self-consistent field procedure). It is propagated along with the nuclear degrees of freedom which are treated classically. To achieve this aim, the time scales of the electronic and nuclear motions are properly adjusted.

Two series of ADMP simulations in the present study were started from the geometries corresponding to the two minima on the DFTB PES of free formic acid, i.e. to the cis- and trans-conformers of this molecular system. Starting from each of the located minima on the DFTB PES, ADMP molecular dynamics simulations have been performed in the microca-nonical (NVE) ensemble. To reach the finally desired temperatures, various amounts of initial nuclear kinetic energies were in the beginning injected into the system and distributed among the atoms. No thermostats were applied to maintain a constant temperature during each of the ADMP simulations. Such approach has led to acceptable temperature fluctuations throughout the simulation. As the main focus of the present study is to compute the spectroscopic properties of the title system within the dynamical approach, i.e. within the time correlation function approach, the dynamics of molecular system has to be sampled properly. In such context, introducing a thermostat to maintain constant temperature, although leading to much smaller temperature fluctuations, would in parallel severely distort the system's dynamics [19, 20]. Series of ADMP simulations were carried out at target temperatures of 10 K, 100 K, 200 K and at 300 K.

Upon initial velocity assignment, the system was allowed to equilibrate for 2 ps. Equilibration phase was followed by production (simulation) phase which was 11 ps long. To integrate the equations of motions, a time step of 0.1 fs was used for productive computations. The fictitious electron mass was set to 0.1 amu and the Cholesky basis for the orthonormal set was used.

All static and dynamics calculations in the present study were performed with the Gaussian09 series of codes [21]. Computation of the autocorrelation functions and their subsequent analyses, as well as generation of series of geometries for the pointwise energy calculations were done with our locally developed FORTRAN codes and LINUX scripts.

2.3. Time-correlation functions approach to spectroscopic properties. In the present study, to compute the finite-temperature vibrational spectra of cis- and trans-conformers of free formic acid from the ADMP molecular dynamics simulations, we have implemented the time correlation functions approach. This approach is actually based on the linear response formalism [15]. A particular autocorrelation function is computed from the collected data throughout the MD trajectory and it is sequentially Fourier-transformed to arrive at a spectrum of a given type. The autocorrelation of a time-dependent function f(t), according to Wiener-Khintchine theorem, is given by:

$$\langle f(\tau)f(t+\tau)\rangle_\tau = \frac{1}{2\pi}\int\left|\int f(t)e^{-i\omega t}dt\right|e^{i\omega t}d\omega \tag{2.7}$$

In accordance with the mathematical properties of Fourier transforms, it follows that the autocorrelation of $f(t)$ can be obtained by first taking the Fourier transform of $f(t)$ and sequentially computing the square of its modulus and taking the inverse Fourier transform.

In the present study, we have relied on two types of autocorrelation functions to compute the vibrational spectra: autocorrelation function of the nuclear velocities (the velocity-velocity autocorrelation function) and the dipole moment autocorrelation function [19, 20].

We have used the following definition of the velocity-velocity autocorrelation function (VV-ACF):

$$\langle\overrightarrow{v}(t)\overrightarrow{v}(0)\rangle = \sum_i\sum_j\int\limits_0^{T_{lag}} v_{i,j}(t')\cdot v_{i,j}(t'+t)dt \tag{2.8}$$

where i ranges from 1 to the total number of atoms, while the index j refers to the three principal Cartesian directions and ranges from 1 to 3. VV-ACF was computed from the data collected from the production (simulation) part of the ADMP trajectory and subsequently normalized with respect to the initial value $\langle\overrightarrow{v}(0)\overrightarrow{v}(0)\rangle$. From the normalized VV-ACF, the rovibrational density of states spectra, which are proportional to the kinetic energy spectra were computed by:

$$I_{vv}(\omega) = \lim_{T\to\infty}\int\limits_0^T \frac{\langle\overrightarrow{v}(0)\overrightarrow{v}(0)\rangle}{\langle\overrightarrow{v}(0)\overrightarrow{v}(0)\rangle}e^{-i\omega t}dt \tag{2.9}$$

Analogously to VV-ACF the dipole moment autocorrelation function (DM-ACF) was also computed from the production phase of the ADMP simulation. Infrared absorption cross-sections were computed by a subsequent Fourier transformation, i.e.:

$$I_{\mu\mu}(\omega) \sim \lim_{T\to\infty}\int\limits_0^T \frac{\langle\overrightarrow{\mu}(0)\overrightarrow{\mu}(0)\rangle}{\langle\overrightarrow{\mu}(0)\overrightarrow{\mu}(0)\rangle}e^{-i\omega t}dt \tag{2.10}$$

To compute the spectra of given types, both types of autocorrelation functions were subjected to Fourier transformation by the fast Fourier transform algorithm (FFT). As the corresponding time series have been obtained from finite length ADMP simulations, we have used the Blackman's window function to account for the fact that $T < \infty$ and cause the integrand to diminish at suitable T values, which (in discrete notation) has the following form [22]:

$$w(n) = 0.42 - 0.5\cdot\cos\left(\frac{2\pi n}{N-1}\right) + 0.08\cdot\cos\left(\frac{4\pi n}{N-1}\right);\quad 0\le n\le N-1 \tag{2.11}$$

3. Results and discussion. The two minima that have been located on the DFTBA PES of free formic acid molecule, corresponding to the cis- and trans-conformers of this molecular system, are shown in Fig. 3.1. These two structures have further on been used as starting points in the course of MD simulations within the ADMP scheme.

The computed structural parameters for the two minima are compared to the corresponding experimental values in Table 3.1. In this table, also the changes in intramolecular parameters upon mutual interconversion of

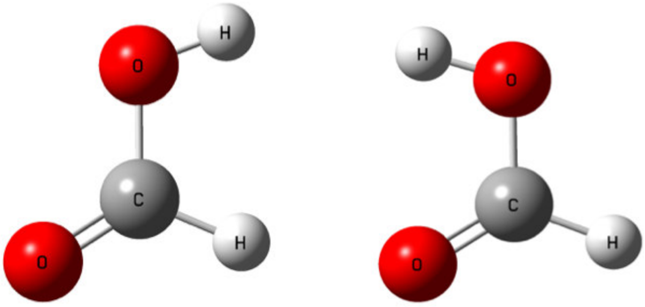


FIG. 3.1. *The minima located on the DFTB PES of formic acid molecule, corresponding to the cis- (left) and trans- (right) conformers.*

TABLE 3.1

The computed structural parameters for the two minima compared to the corresponding experimental values

Parameter	Trans		Cis		Δ(Cis-Trans)	
	Exp.	DFTB	Exp.	DFTB	Exp.	DFTB
$R_{\text{C-H}}$ / Å	1.097	1.135	1.105	1.153	0.008	0.018
$R_{\text{C-O}}$ / Å	1.342	1.377	1.352	1.379	0.010	0.002
$R_{\text{O-H}}$ / Å	0.972	0.982	0.956	0.978	-0.016	-0.004
$R_{\text{C=O}}$ / Å	1.203	1.211	1.195	1.202	-0.008	-0.009
θHCO(-H) / °	112.0	112.6	114.6	112.9	2.6	0.3
θCOH / °	106.3	107.9	109.7	110.4	3.4	2.5
θHCO / °	123.2	125.8	123.2	124.6	0.0	-1.2
θOCO(-H) / °	124.8	121.6	122.1	122.5	-2.7	0.9

TABLE 3.2

The computed anharmonic O-H and O-D stretching frequencies for the two minima compared to the corresponding experimental values

Parameter	Trans		Cis		Δ(Cis-Trans)	
	Exp.	DFTB	Exp.	DFTB	Exp.	DFTB
vO-H / cm-1	3550.5	3506.0	3618.0	3557.4	67.5	51.4
vO-D / cm-1	2631.0	2569.9	2685.0	2605.9	53.0	36.0

the two conformers are given. As can be seen, the overall agreement between theory and experiment is excellent, confirming the appropriateness of the employed theoretical method for the purpose of the present study.

The O-H stretching vibrational potentials, computed at the DFTBA level of theory for the cis-and trans-conformers are shown in Fig. 3.2. Table 3.2, on the other hand, compiles the computed anharmonic O-H and O-D stretching frequencies for the two minima compared to the corresponding experimental values. As can be seen, concerning the fact that these are the "raw" (i.e. unscaled) values, the agreement between theoretically computed and experimentally measured O-H(D) stretching frequencies is rather satisfactory.

In the present study, to carry out the molecular dynamics simulations, we have relied on the atom-centered density matrix propagation scheme. This is an extended Lagrangian molecular dynamics method in which the electronic structure is accounted for by a single-particle density matrix representation of the electronic subsystem. The density matrix is propagated simultaneously with the nuclear degrees of freedom (treated in a classical manner). This is achieved by introducing the fictitious inertia tensor μ in the course of adjusting the nuclear with the electronic time scales. The overall effect of such adjustment is an achievement of a fictitious dynamics that allows controllable oscillations around the Born-Oppenheimer surface. However, within the ADMP scheme the self-consistent field (SCF) convergence is not achieved. Therefore, one has to analyze

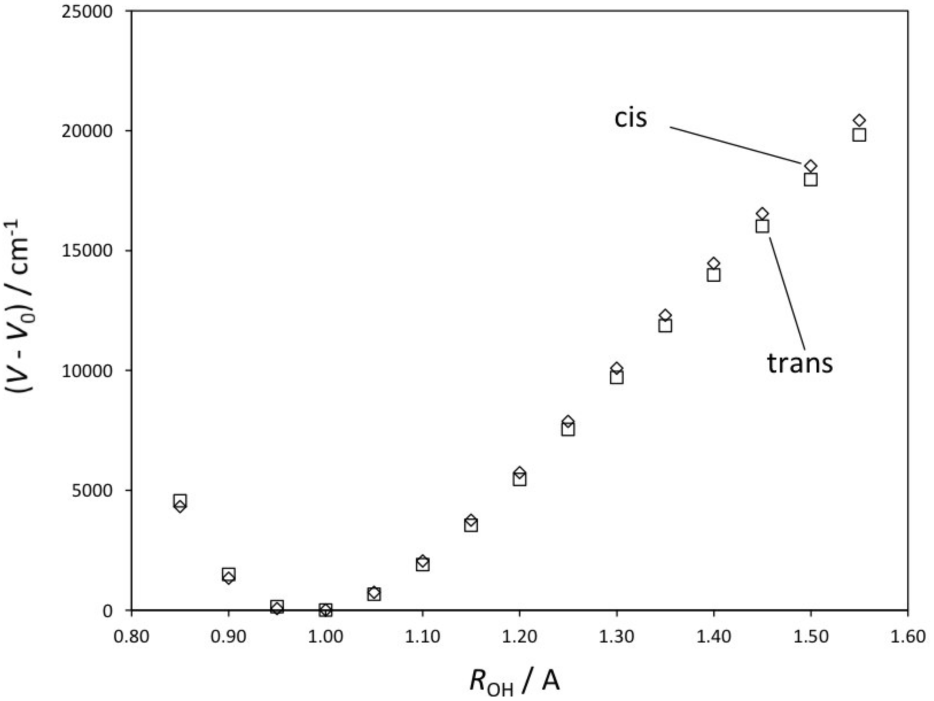


FIG. 3.2. *The O-H stretching vibrational potentials for the cis-and trans-conformers of free formic acid, computed at the DFTBA level of theory.*

carefully the errors in order to be certain of the accuracy of the dynamics as well as of its physical meaningfulness.

To analyze the errors inherent to the ADMP scheme, in our present study, we have thoroughly analyzed the time-evolution of the adiabaticity index, the idempotency of the density matrix [19, 20], as well as time-dependence of the total angular momentum throughout the productive part of the simulation. On Fig. 3.3, the time-dependence of the adiabaticity index in the production phase of the simulations carried out at 10 K, starting from both cis- and trans-conformers of the formic acid molecule is shown. Comparison with the literature-suggested threshold values of this quantity [19, 20] allowed us to conclude the stability of the simulations. As an additional test, to confirm the previous conclusion, we have also checked the idempotency of the density matrix; this parameter was kept within the threshold value of 10^{-12}. At the same time, the total angular momentum value was conserved to $< 10^{-13}\ \hbar$ as well.

In Table 3.3, the target MD temperatures are compared to the actual ones achieved during the productive part of the ADMP simulation runs with cis- and trans-conformers as starting points for the dynamical simulations (i.e. the minima on the DFTB PESs corresponding to these structures). The observed temperature fluctuations around the target and average values presented in Table 3.3 throughout the productive phase of the simulation were acceptable and, at the same time, in line with the statistical physics expectations for a dynamical simulation of molecular system with the current size. Certainly, more precise temperature control (and therefore, further diminishing of the temperature fluctuations) could have easily been achieved by e.g. coupling the system to a suitably chosen thermostat. However, such temperature control has not been applied in the present study, since the main intention of the study was computation of spectroscopic properties from dynamical simulations through the time correlation functions formalism. In order to do this in a physically correct manner, any distortion of the dynamics, which would be introduced by the imposed temperature control [19, 20], has to be avoided.

In Fig. 3.4, the frequency region of the kinetic energy spectra (i.e. the kinetic energy density of states spectra) obtained by Fourier transformation of the velocity-velocity autocorrelation function in which peaks

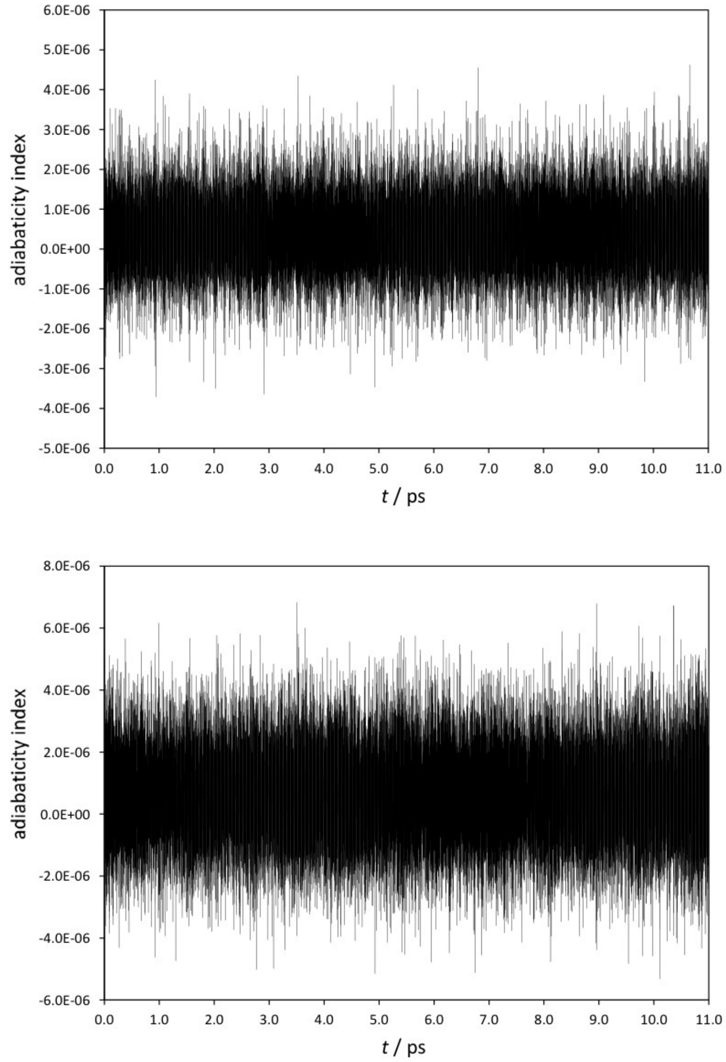


Fig. 3.3. *The time-dependence of the adiabaticity index in the production phase of the simulations carried out at 10 K, starting from both cis- (top) and trans-conformers (bottom) of the formic acid molecule.*

Table 3.3

Target and actual temperatures achieved during the productive part of the ADMP simulation runs with cis- and trans-conformers as starting points for the dynamical simulations (i.e. the minima on the DFTBA PESs corresponding to these structures)

T_{target} / K	$T_{\text{sim.,cis}}$ / K	$T_{\text{sim.,trans}}$ / K
10	10.15	9.97
100	101.91	99.59
200	201.97	199.90
300	299.72	298.45

due to the O-H stretching vibrational motions are expected to appear is shown for the series of simulations carried out at the four different temperatures: 10, 100, 200 and 300 K starting from the equilibrium geometry

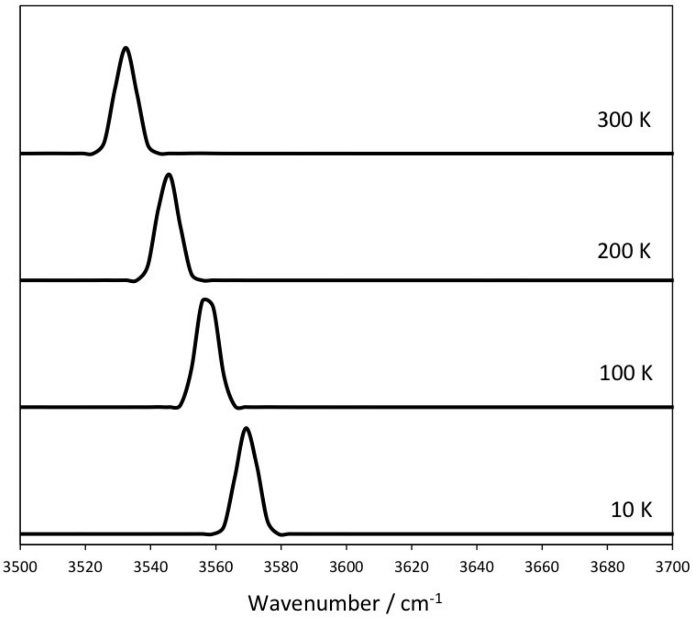


Fig. 3.4. *The kinetic energy spectra (kinetic energy density of states spectra) obtained by Fourier transformation of the velocity-velocity autocorrelation function for the series of simulations carried out at the four different temperatures, starting from the trans-conformer of the free formic acid molecule, in the frequency region where signals due to the O-H stretching vibrations are expected to appear (from 3500 to 3700 cm^{-1}).*

of the trans-conformer of formic acid. In Fig. 3.5, on the other hand, the same spectral region is shown in the kinetic energy spectra computed from the trajectories at the mentioned series of temperatures started from the equilibrium geometry corresponding to the cis-conformer of free formic acid molecule. The intensity pattern in the kinetic energy spectra is, of course, not directly comparable to the infrared spectrum. It is more directly comparable to the deep inelastic neutron scattering spectrum. However, these spectra still contain essential information concerning the spacings between vibrational energy levels of different intramolecular modes. Following, therefore, the kinetic energy spectra, one indirectly follows the temperature evolution of the energy level differences. As will be seen from the subsequent discussion, the particular frequency region analyzed in the present study maps quite clearly the temperature evolution of the molecular conformational flexibility along with the intramolecular vibrational energy redistribution.

In summary, the kinetic energy spectra carry out the essential information concerning molecular rovibrational density of states (or, perhaps even more precisely, about the existence of energy level difference at particular frequency, i.e. wavenumber value). Although for a direct comparison with the experimental infrared spectra, it is better to rely on a straightforwardly comparable quantity derived by a Fourier transformation of the dipole moment autocorrelation function, throughout this present study, we will use the kinetic energy spectra in parallel with those computed from the dipole autocorrelation function. Although this may, at first sight, seem to complicate the analysis, we have chosen such an approach due to the following reasons. Existence of vibrational energy level difference, manifested with an appearance of a band (signal) at a given frequency value, does not tell us anything about the particular mode that is involved in this energy level pattern. As a matter of fact, the intramolecular nuclear motions that correspond to a particular mode may or may not involve a change in the dipole moment. Such change, however, would be responsible for absorption of light quanta upon interaction with the incident radiation from infrared spectral region. By restricting the analysis solely on the basis of spectra obtained from the dipole moment autocorrelation function, therefore, the thermally-induced behavior of modes which are not infrared active would be disregarded.

In Fig. 3.6, the O-H stretching region in the spectra computed by Fourier transformation of the dipole moment vector autocorrelation function for the series of simulations of free formic acid molecule carried out

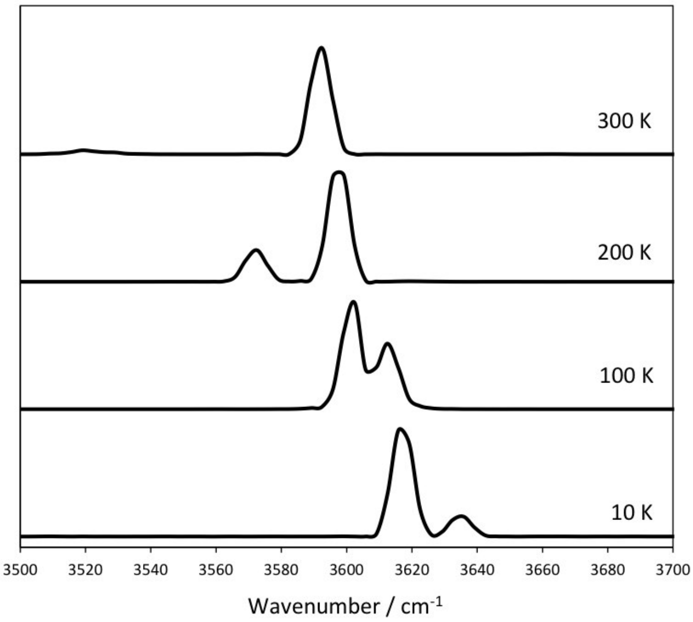


Fig. 3.5. *The kinetic energy spectra (kinetic energy density of states spectra) obtained by Fourier transformation of the velocity-velocity autocorrelation function for the series of simulations carried out at the four different temperatures, starting from the cis-conformer of the free formic acid molecule, in the frequency region where signals due to the O-H stretching vibrations are expected to appear (from 3500 to 3700 cm^{-1}).*

at the four different temperatures (10, 100, 200 and 300 K), started from the equilibrium geometry of the trans-conformer is shown. Fig. 3.7, on the other hand, shows the same spectral region in the dipole moment autocorrelation spectra extracted from series of simulations at the mentioned temperatures, this time started from the equilibrium geometry of the cis-conformer of free formic acid molecule. As implied before, these spectra are actually directly comparable to the experimentally measured frequency dependencies of the infrared absorption cross-sections obtained by experimental infrared spectroscopic techniques, i.e. directly comparable to the temperature-dependent infrared spectra of the studied species.

We start the discussion by analyzing the appearance of the spectral region in which signals due to O-H stretching motions are expected to appear in the kinetic energy spectra computed from the velocity-velocity autocorrelation function (Figs. 3.4 and 3.5), corresponding to the rovibrational density of states. As can be seen from Fig. 3.4, the O-H stretching band exhibits a notable shift towards lower wavenumbers (energies) upon temperature increase. We attribute this behavior to the following. At lower initial kinetic energies, corresponding to achievement of lower effective temperatures, the amplitudes of all intramolecular motions are modest, much lower than in the case of higher energies (and consequently, temperatures). At lower temperatures, therefore, the O-H oscillators sample only the region in the vicinity of the "harmonic part" of the vibrational potential. As the temperature increases, the amplitudes increase, so that the studied oscillators more frequently reside in the "more anharmonic" region of the potential. As the differences between vibrational energy levels are directly dependent on the anharmonicity, the more anharmonic the effective (sampled) potential is, the lower in energy will the corresponding spectral band appear. Since the trans- rotamer of the free formic acid is the more stable one, only a single effective O-H stretching band appears at temperatures up to 300 K in the rovibrational density of states spectra. The situation in the case of simulations started from the cis-form of free formic acid is much more complex, however (Fig. 3.5). Though the general trend in the temperature-dependence of the most intensive band in this spectral region basically follows the one evidenced in the case of trans-conformer, it can be seen from Fig. 3.5 that in the studied spectral region of the cis-conformer, even at temperatures as low as 10 K, two bands appear, as a consequence of enhanced amplitudes of the intramolecular OCOH torsional motion. At temperature of 200 K, a clearly defined band appears at $\sim$ 3572 cm^{-1}, which

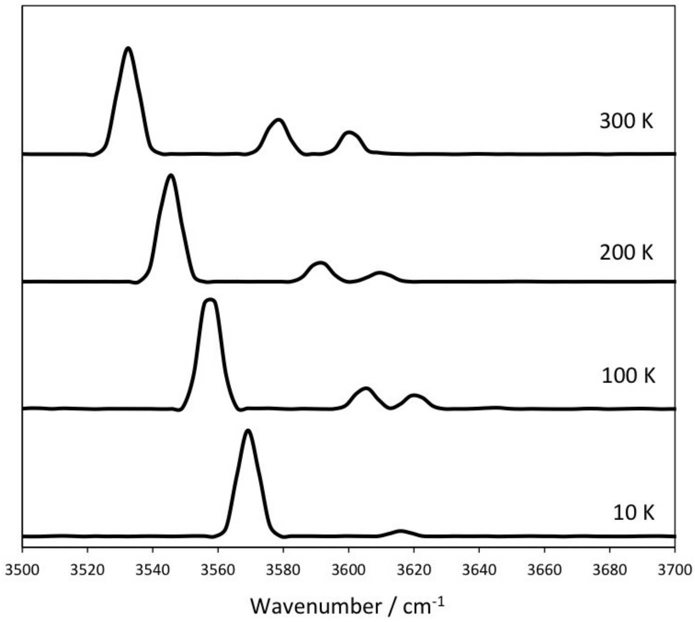


Fig. 3.6. *The dipole moment autocorrelation spectra (∼ infrared absorption spectra) obtained by Fourier transformation of the dipole moment autocorrelation function for the series of simulations carried out at the four different temperatures, starting from the trans-conformer of the free formic acid molecule, in the frequency region where signals due to the O-H stretching vibrations are expected to appear (from 3500 to 3700m^{-1}).*

could correspond to an O-H stretching vibration within the deeper well of the torsional potential. The velocity-velocity autocorrelation spectra computed from the ADMP simulations in the present study are based on the autocorrelation functions of averaged nuclear velocities. Therefore, these are statistical quantities, obtained as a statistical average from numerous different configurations spanned by the MD simulation and the computed resultant spectrum corresponds to a dynamically averaged picture of the studied molecular system. In the course of dynamical simulation, the immediate surroundings of the O-H stretching vibrational mode changes in an anisotropic manner, which, in turn, leads to broadening and flattening of particular spectral regions, in line with previous results reported in the literature [19, 20].

The spectral pictures observed in the O-H stretching regions of the dipole moment autocorrelation spectra (Fig. 3.6 and 3.7) computed from the trajectories started from trans- and cis- minima on the DFTB PES of free formic acid resemble quite much the analogous regions in the velocity-velocity autocorrelation spectra, with some intensity redistributions.

On the basis of all previously outlined theoretical results one can conclude that the thermally-induced dynamical effects in the rovibrational density of states (as well as in the infrared absorption cross-sections) of even rather simple individual molecular systems may lead to substantial differences in comparison to the corresponding "static" properties. When one compares theoretical with experimental spectroscopic data, therefore, to account for all the complexities inherent to the intramolecular motions of the studied system, the temperature effects should be explicitly accounted for. This is especially valid considering the fact that the usually available experimental vibrational spectroscopic data have often been obtained at temperatures far above the absolute 0 (which is the effective temperature at which "static" *ab intio* or semiempirical computations are often carried out).

When the studied molecular system is characterized by a certain degree of intramolecular flexibility, i.e. in cases when several rotamers close in energy may exist, thermally-induced intramolecular motions and intramolecular vibrational energy redistributions can cause effective dynamical transitions between the corresponding wells on the molecular PES. As the statistically averaged spectra computed from statistical physics simulations by the linear response formalism account for such dynamical effects (and for the intramolecular conformational

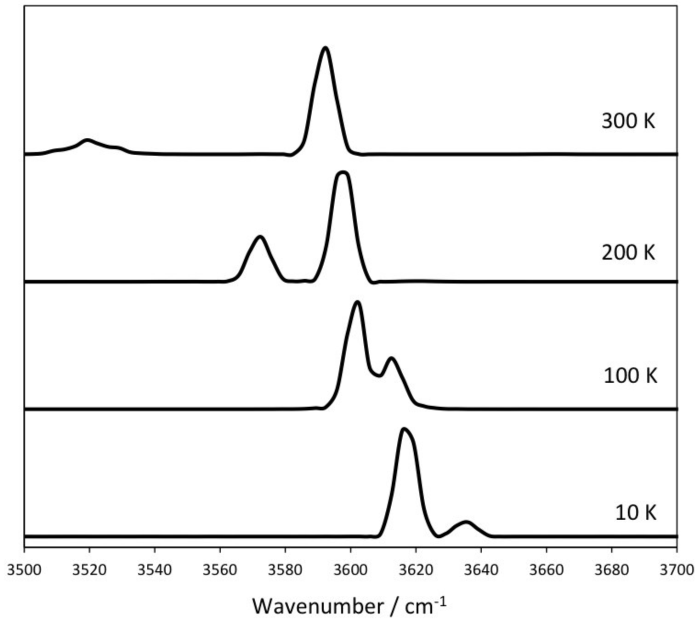


Fig. 3.7. *The dipole moment autocorrelation spectra (∼ infrared absorption spectra) obtained by Fourier transformation of the dipole moment autocorrelation function for the series of simulations carried out at the four different temperatures, starting from the cis-conformer of the free formic acid molecule, in the frequency region where signals due to the O-H stretching vibrations are expected to appear (from 3500 to 3700 m^{-1}).*

flexibility of the studied molecule as well), these should be favored over the static ones.

4. Summary and conclusions. In the present study, we have carried out a density functional tight binding study of the O-H stretching vibrations in the case of two conformers of free formic acid molecule. We have calculated the anharmonic O-H stretching vibrational frequencies implementing a static approach (based on calculation of the O-H stretching vibrational potentials of the two formic acid conformers and subsequent numerical solution of the vibrational Schrödinger equation), and a dynamic approach (based on sequential DFTB MD simulation within the ADMP scheme followed by analysis of several autocorrelation functions calculated from the MD trajectories). Dynamic calculations allowed us to compute the temperature-dependent spectroscopic properties of this molecule, i.e. to explicitly involve the thermal motion effects in single-molecule computational anharmonic vibrational spectroscopy. We have computed various types of spectra at series of temperatures (ranging from 10 up to 300 K), including the rovibrational density of states spectra, as well as the infrared absorp-tion cross section spectra. The thermally induced changes in the single-molecule spectroscopic properties were deduced and the reasons behind them were analyzed and discussed. The advantages of the dynamic approach to computational vibrational spectroscopy at finite temperatures are outlined and discussed, in the context of research devoted to molecular-level understanding of the phenomena and processes relevant to climate science and atmospheric chemistry.

Acknowledgments. This work was supported by the European Unions Horizon 2020 research and innovation programme, project Virtual Research Environment for Regional Interdisciplinary Collaboration in Southeast Europe and Eastern Mediterranean VI-SEEM [675121].

REFERENCES

[1] D.P. Tew and W. Mizukami. *Ab initio vibrational spectroscopy of cis-and trans-formic acid from a global potential energy surface.* The Journal of Physical Chemistry A, 120(49):9815–9828, 2016.

[2] S. Roszak, R.H. Gee, K. Balasubramanian, and L.E. Fried. *New theoretical insight into the interactions and properties of formic acid: Development of a quantum-based pair potential for formic acid.* The Journal of chemical physics, 123(14):144702, 2005.

[3] M. Pettersson, J. Lundell, L. Khriachtchev, and M. Räsänen. *Ir spectrum of the other rotamer of formic acid, cis-hcooh. Journal of the American Chemical Society*, 119(48):11715–11716, 1997.

[4] M. Pettersson, E.M.S. Maçôas, L. Khriachtchev, R. Fausto, and M. Räsänen. *Conformational isomerization of formic acid by vibrational excitation at energies below the torsional barrier.* Journal of the American Chemical Society, 125(14):4058–4059, 2003.

[5] E.M.S. Macôas, L. Khriachtchev, M. Pettersson, J. Lundell, R. Fausto, and M. Räsänen. *Infrared-induced conformational interconversion in carboxylic acids isolated in low-temperature rare-gas matrices.* Vibrational spectroscopy, 34(1):73–82, 2004.

[6] M. Pettersson, E.M.S. Maçôas, L. Khriachtchev, J. Lundell, R. Fausto, and M. Räsänen. *Cis → trans conversion of formic acid by dissipative tunneling in solid rare gases: Influence of environment on the tunneling rate.* The Journal of chemical physics, 117(20):9095–9098, 2002.

[7] E.M.S. Maçôas, L. Khriachtchev, M. Pettersson, J. Juselius, R. Fausto, and M. Räsänen. *Reactive vibrational excitation spectroscopy of formic acid in solid argon: Quantum yield for infrared induced trans → cis isomerization and solid state effects on the vibrational spectrum.* The Journal of chemical physics, 119(22):11765–11772, 2003.

[8] E.M.S. Maçôas, J. Lundell, M. Pettersson, L. Khriachtchev, R. Fausto, and M. Räsänen. *Vibrational spectroscopy of cis-and trans-formic acid in solid argon.* Journal of Molecular Spectroscopy, 219(1):70–80, 2003.

[9] S. Ioppolo, B.A. McGuire, M.A. Allodi, and G.A. Blake. *Thz and mid-ir spectroscopy of interstellar ice analogs: methyl and carboxylic acid groups.* Faraday discussions, 168:461–484, 2014.

[10] G. Buemi. *Dft study of the hydrogen bond strength and ir spectra of formic, oxalic, glyoxylic and pyruvic acids in vacuum, acetone and water solution.* Journal of Physical Organic Chemistry, 22(10):933–947, 2009.

[11] V. Barone. *Anharmonic vibrational properties by a fully automated second-order perturbative approach.* The Journal of chemical physics, 122(1):014108, 2005.

[12] B Koteska, A. Mishev, and L.J. Pejov. *Computational study of intramolecular oh stretching vibrations in the two rotamers of free formic acid.* Romanian Reports in Physic, in press.

[13] B. Koteska, A. Mishev, and L. Pejov. *Computational approach towards vibrational spectroscopic detection of molecular species relevant to atmospheric chemistry and climate science: The formic acid rotamers.* In Smart Technologies, IEEE EUROCON 2017-17th International Conference on, pages 926–931. IEEE, 2017.

[14] H.B. Schlegel, J.M. Millam, S.S. Iyengar, G.A. Voth, A.D. Daniels, G.E. Scuseria, and M.J. Frisch. *Ab initio molecular dynamics: Propagating the density matrix with gaussian orbitals.* The Journal of Chemical Physics, 114(22):9758–9763, 2001.

[15] M. Thomas, M. Brehm, R. Fligg, P. Vöhringer, and B. Kirchner. *Computing vibrational spectra from ab initio molecular dynamics.* Physical Chemistry Chemical Physics, 15(18):6608–6622, 2013.

[16] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert. *Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties.* Physical Review B, 58(11):7260, 1998.

[17] G. Zheng, H.A. Witek, P. Bobadova-Parvanova, S. Irle, D.G. Musaev, R. Prabhakar, K. Morokuma, M. Lundberg, M. Elstner, C. Köhler, et al. *Parameter calibration of transition-metal elements for the spin-polarized self-consistent-charge density-functional tight-binding (dftb) method: Sc, ti, fe, co, and ni.* Journal of chemical theory and computation, 3(4):1349–1367, 2007.

[18] H.B. Schlegel. *Optimization of equilibrium geometries and transition structures.* Journal of Computational Chemistry, 3(2):214–218, 1982.

[19] S.S. Iyengar. *Dynamical effects on vibrational and electronic spectra of hydroperoxyl radical water clusters.* The Journal of chemical physics, 123(8):084310, 2005.

[20] S.S. Iyengar, M.K. Petersen, T.J.F. Day, C.J. Burnham, V.E. Teige, and G.A. Voth. *The properties of ion-water clusters. i. the protonated 21-water cluster.* The Journal of chemical physics, 123(8):084309, 2005.

[21] MJ Frischet et al. al., gaussian 09, revision a. 1, 2009.

[22] F.J. Harris. *On the use of windows for harmonic analysis with the discrete fourier transform.* Proceedings of the IEEE, 66(1):51–83, 1978.

Edited by: Aneta Karaivanova
Received: Dec 22, 2017
Accepted: Apr 10, 2018

DOI 10.12694/scpe.v19i2.1337
Scalable Computing: Practice and Experience ISSN 1895-1767
Volume 19, Number 2, pp. 131–138. http://www.scpe.org © 2018 SCPE

AN INTEGRATED WEB-BASED INTERACTIVE DATA PLATFORM FOR MOLECULAR DYNAMICS SIMULATIONS

HRACHYA ASTSATRYAN, WAHI NARSISIAN, ELIZA GYULGYULYAN, VARDAN BAGHDASARYAN[*] ARMEN POGHOSYAN[†] YEVGENI MAMASAKHLISOV[‡] AND PETER WITTENBURG[§]

Abstract. The article aims to introduce an integrated web-based interactive data platform for molecular dynamic simulations using the datasets generated by different life science communities from Armenia. The suggested platform, consisting of data repository and workflow management services, is vital for current and future scientific discoveries in the life science domain. We focus on interactive data visualization workflow service as a key to perform more in-depth analyzes of research data outputs, helping to understand the problems efficiently and to consolidate the data into one collective illustration platform. The functionalities of the integrated data platform is presented as an advanced integrated environment to capture, analyze, process and visualize the scientific data.

Key words: molecular dynamics simulations, high-performance computing, persistent identifier, Web-based interactive visualization, DNA, biological systems.

AMS subject classifications. 68U20, 65Y05, 78A70

1. Introduction.

In addition to the main pillars of science, which are the theory and experiments, modeling and numerical simulations are the heart of multiple domains, which are not only scientific, but also societal (energy, health, environment), economic, financial, and life ethics [1]. They also appear increasingly as decision-making tools for critical cases such as global warming or natural disasters. Since the modeling and simulations are essential for many scientific advances, the control of all the aspects of high-performance computing (HPC) - as well as the capacity to exploit the masses of data to tackle the solution of these complex models - is inescapable.

The explosion in computational power helps the biologists and life science researchers to conduct more advanced experiments and in its turn lead to a rapid increase in the amount of experimental data, research outputs, etc. [2]. As computational experiments, molecular dynamics (MD) simulations are widely used in the domain of life science to evaluate the equilibrium nature of classical many-body systems [3, 4]. The study of systems with a large number of atoms in long trajectory intervals (from nano to milliseconds) is required to explore a broad range of exciting phenomena, which is undoubtedly unfeasible without using appropriate HPC resources and storage facilities to manage and visualize these data.

Various life science communities from Armenia use HPC resources and generate a significant amount of research outputs by storing them in local repositories. Usually, such local datasets are incomplete, and there is a demand to cluster them into a central repository by managing this data using appropriate metadata and identifiers. Because there is no centralized repository to hold all these data, the data sharing between these communities is a challenge, which also leads to being unable to have a single platform to process and visualize this data. The volume, complexity, and heterogeneity of data originating from these communities have created challenges to have a complete understanding of complex biological processes and systems [5].

This article aims to introduce an integrated web-based interactive data platform for molecular dynamic simulations using the datasets generated by the several Armenian life science communities. The suggested platform, which consists of a data repository and workflow management services, is vital for current and future scientific discoveries. We focus on interactive data visualization workflow service as a key to perform a more in-depth analyzes of research data outputs to help to understand the problems efficiently and to consolidate the data into one collective illustration platform. The integrated data platform is presented as an advanced integrated environment to capture, analyze, process and visualize the scientific data.

[*] Institute for Informatics and Automation Problems of the National Academy of Sciences of the Republic of Armenia, 1, P. Sevak str., 0014 Yerevan, Armenia (hrach@sci.am).

[†] International Scientific Educational Center of the National Academy of Sciences of the Republic of Armenia, 24D, M. Baghramyan ave., 0019 Yerevan, Armenia.

[‡] Yerevan State University, 1 A. Manoogian str., 0015 Yerevan, Armenia.

[§] Max Planck Computing and Data Facility, Gießenbachstraße 2, 85748 Garching, Germany.

The remainder of this paper is divided into the following sections: section 2 introduces the life science communities and applications in Armenia widely using MD simulations, section 3 represents the integrated data platform, section 4 workflow services and finally section 5 is the conclusion.

2. Life science communities and applications in Armenia. The life science communities in Armenia produce a significant amount of data and widely use HPC resources. The scientific outputs and datasets generated by the Bioinformatics Group of the International Scientific and Educational Center of the National Academy of Sciences of the Republic of Armenia (Bioinformatics group) and the Molecular Physics department of the Yerevan State University (MolPhys YSU) are used in the suggested data platform.

The Bioinformatics group studies complex systems, including surfactants, polymers, and proteins, which can be found in everyday life, as well as in many industrial applications, such as in pharmaceuticals, food processing or agrochemicals [6]. The analyzes of interaction between surfactant and protein/polymer plays a critical role in many physical processes and opens up a wide range of commercial applications, including cosmetic formulations [7] or drug delivery systems [8]. Moreover, such kinds of complex systems have been extensively investigated for many years as model systems for biological membranes being of vital importance in studies of cell membranes [9]. The large-scale simulations play a vital role to offer a detailed picture of the structure and dynamics of complex systems by improving our knowledge and understanding of many interesting phenomena, such as slow conformation changes or folding/unfolding. The efficiency of such kind of longtime simulations can be reached via HPC platforms, as many interesting phenomena occur at nano-to-milliseconds time scale, and require massively large systems with many atoms to mimic real systems. Although, it is also essential and necessary to analyze MD simulations with experimental data for validity ensuring. The main aim of various of studies of the group is to use the classical MD simulation method getting a deeper insight into dynamic processes occurring on longtime scale range [10, 11, 12].

The MolPhys YSU studies nucleic acids, e.g. double-stranded DNA and single-stranded RNA molecules, which play an essiential role in the living systems functionality, biomedical research, biological sensors development, etc. [13, 14]. For example, hybridization process is a keystone of many essential processes, including transcription, replication, polymerase chain reaction or DNA sensors functioning. Thermodynamics and kinetics of the nucleic acids hybridization have been extensively investigated both on the surface and in bulk [15, 16, 17]. The all-atom and coarse-grained simulations can give a substantial impact on the understanding of hybridization thermodynamics and kinetics. Such large-scale simulations require HPC platform and massively parallelized MD computations. The main research focus of MolPhys YSU is the single and double-stranded nucleic, which require validation using MD simulations [18, 19, 20]. The combination of experimental and various computational methods can give us a deep understanding of the complex processes, behind the nucleic acids hybridization.

3. Integrated Web-based interactive data platform. The suggested integrated web-based interactive data platform gives users a possibility to have a single entry point to different services[1]. First, it enables to store and manage the output data of diverse simulations and assign meta-data to each output to increase the data reliability. In addition, the platform allows searching for any required molecule type and download it. For the visualization, an interactive web-based solution is provided to use the web browser directly to have a quick look at the molecule structure and based on that decide to use it or not[2]. Finally, the public link is provided to enable data downloading in case the user does not have an account on iRODS (Integrated Rule-Oriented Data System) [21].

The platform consists of the following layers, which are illustrated in Figure 3.1.

3.1. Local data repository. The local data repository is the most common place for preserving digital research data. A repository is an online database service, which archives and stores the data, and also provides a possibility to discover and access the data. The repository offers several features and benefits:

- To preserve the data for future work.
- To assign metadata and persistent identifiers for each data which in its turn increase their reliability if others will use them.

[1] Web-based interactive data platform, http://irods.asnet.am:8080/irods-cloud-backend

[2] Data visualization platform, http://irods.asnet.am/Jmol/jsmol.htm

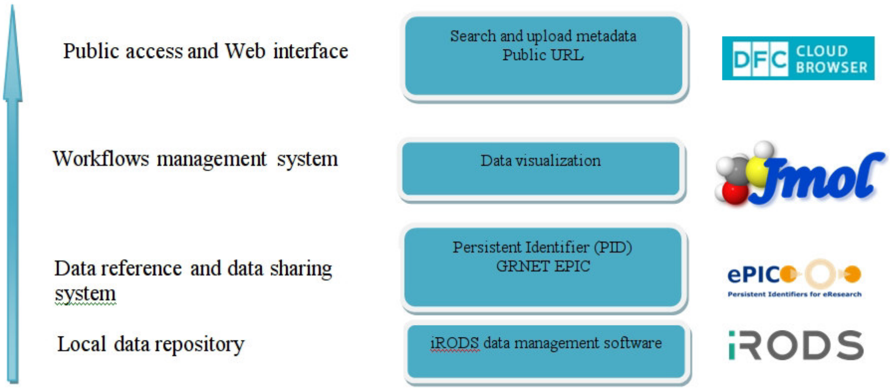

Public access and Web interface
Search and upload metadata
Public URL
DFC CLOUD BROWSER
Workflows management system
Data visualization
Jmol
Data reference and data sharing system
Persistent Identifier (PID)
GRNET EPIC
ePIC
Persistent Identifiers for eResearch
Local data repository
iRODS data management software
iRODS

Fig. 3.1. *Diagram of the integrated web-based interactive data platform.*

- To increase the data discovery over the net.
- To prevent the users from maintaining the data by themselves, because the professional administrators will maintain the data.
- To enable data sharing between different communities.

The most significant benefit of these repositories is sharing the data opening a lot of new opportunities for research, collaboration, increasing research visibility, the use of high-quality data collected by other researchers, etc. Some repositories have restricted constraints and policies about how to use or store the data, and the user needs to register to be able to use the stored data. The data repository platform has been developed using the iRODS, which is an open-source data management software. iRODS provides a rule-based system management approach which makes data replication much easier and provides extra data protection. The metadata system of iRODS is comprehensive and allows users to customize their application level metadata, instead of the metadata supplied by traditional file systems. The types of datasets mainly generated by the Bioinformatics and MolPhys YSU groups include:

- **trr file format** - the trajectory of a simulation including all the coordinates, velocities, forces and energies.
- **gro file format** - a molecular structure in Gromos87 format. gro files, as trajectory by simply concatenating files.
- **xtc file format** - a portable format for trajectories. It uses the xdr routines for writing and reading data which was created for the Unix file system.
- **pdb file format** - molecular structure files in the protein databank file format. The protein databank file format describes the positions of atoms in a molecular structure. Coordinates are read from the ATOM and HETATM records until the file ends or an ENDMDL record is encountered.
- **psf file format** - contain atoms, residues, segment names, residue types, atomic mass and charge, and the bond connectivity.

All in all the metadata are stored in Mysql database, and all related data are stored in the replicated storage to ensure the availability of the resources in case of any damage or failover problem.

3.2. Data reference and data sharing. As the amount of the digital data rises exponentially [22] the relations between them becoming more and more essential and as data repositories are various by their size and format, it can be vital for data in the repositories to change its physical location, as this can cause a loss of the link to that data. It is commonly known that the Uniform Resource Locator (URL) of a data is not a permanent link to the physical location of the data and when the physical location is changed the URL also should be changed everywhere. As the number of places where data URL is cited can be huge the URL changing procedure can be a hard work to do and URLs might be unchanged from several places. This can lead that particular data to be unreachable from the old URL. For this reason, scientific institutions need a long-term preservation of resources with long-term accessibility. Persistent Identifier (PID) [23] is used to enable a long-lasting data reference and sharing, meaning it guarantees reliability in citing sources even if the URL of data changes. It has two components: a unique identifier and a service that locates the resource when its location changes. PID name consists of a prefix that is globally unique within the context of the system providing the PID, and suffix, which is unique within the local organization. The unique prefix "21.15104" is used for the platform and a suffix starting by "ASNET" is automatically generated per dataset. The PIDs are generated and registered by data centers enabled through European Persistent Identifier Consortium (EPIC). EPIC provides PID services for the European Research Community to allocate and to resolve persistent identifiers using handle systems. The handle system of GRNET (Greek Research and Technology Network) is used by the suggested platform to generate PIDs per each simulation with using GRNET Handle Restful web service. As the EPIC API supports the automatic generation of a local name of the PID, the suffix of PID is generated and executed automatically with a POST HTTP request in curl. This gives an opportunity to reach our data even if we change the physical address. We also use EPIC PID for the visualization described in the section below.

3.3. Workflow management system. Research and scientific processes in biology heavily use workflow systems to have all necessary steps to address the complexity of any scientific experiment and to enhance the discovery of new methods and solutions based on the execution of complex algorithms together with the access and analysis of experimental data. All this help to produce more accurate and reliable results and outcomes, which can confirm real experiments or provide a proper and deep understanding of several processes. With the availability of HPC resources and different cloud services [24, 25], the biologists in Armenia can run complex workflows that integrate programs, methods, and data from various resources and run different simulations in a single consolidated platform. Using different scientific computing methods with the support of scientific discovery development, a new area of scientific methods arise with new data analysis strategies enabled which are called e-science paradigm. The ultimate goal of the deployed platform is to provide a complete solution of life science communities to conduct different workflows based on their experimental output. The suggested solution is to have a molecular visualization to be able to check the shape and the construction of the molecule before conducting any experiment. The service enables the users to upload their molecules, or use the sample uploaded on iRODS, or even use any public URL, which refers to some molecule. This will enable users to have more insight and understanding about the molecule and use it in any similar experiment.

3.4. Public access and Web interface. At a top layer, the data visualization layer is used, which is a web interface for molecules visualization (see Fig 3.2). This interface is based on Jmol, which is an open-source browser-based HTML5 viewer and stand-alone Java viewer for chemical structures in 3D [26]. Since it is written in the Java programming language, it is compatible with all major operating systems and, in the applet form, with most modern web browsers. Two beneficial features of this system have been customized, which allows users directly check the molecules stored in the iRODS repository using its public URL link, and upload the molecules from the personal computer and visualize it. The interface enables to colorize the molecules based on their type, animate them, add labels etc. At the top level, there is a web interface, which is a Cloud browser developed by DICE group [27]. The web interface simplifies the researcher's work to not think about where to store the data and corresponding metadata, the researcher needs only to upload the data and then fill the corresponding sections of the metadata. A new template is provided for each community containing the relevant fields in order to gather or input all required information for each uploaded data. There is also a separate service that can be accessed from the above-mentioned web interface. It is a public link to all stored data on iRODS. This link will enable any user to use the data on iRODS repository without the need of registration, which as a result increases the usage of the data.

4. Demonstration and discussion. To perform data analysis in biology workflows are very useful. The workflow management system helps users to create and run different experiments without concerning about the programming and what is going on at the back-end of the system. In bioinformatics, these systems usually concentrate more on a visual representation of molecules using graphical user interface, which enables scientists to run several scientific tasks in parallel and visualize the results in order to get more details and proper information about the experiments outcome. The 3D presentation or visualization of studied systems is an important key to understand the intra-molecular structure better and get insight from MD simulation.

Many complex biological systems have been studied using the platform, for instance, a system, consist of

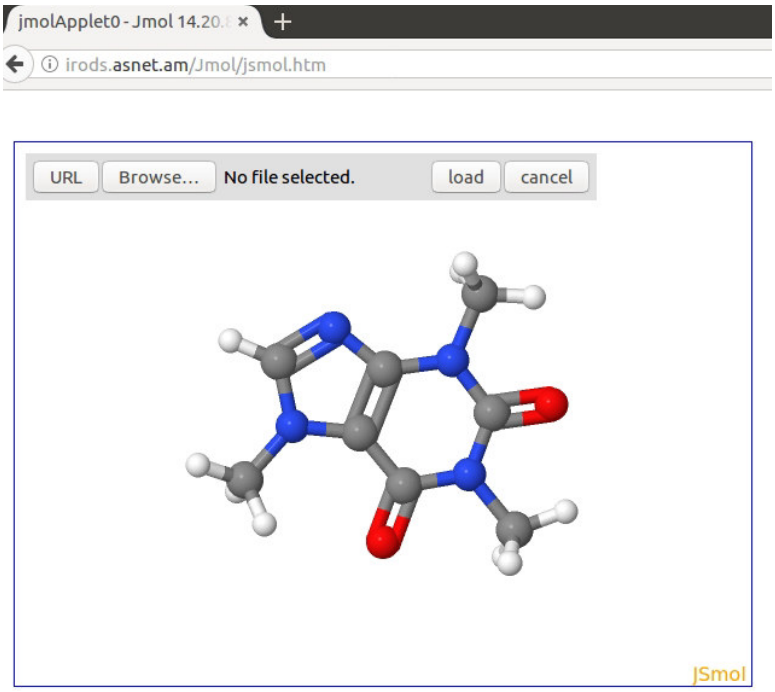


FIG. 3.2. *Data visualization interface.*

a cationic poly (dially ldimethy lammonium chloride) (PDADMAC)/ sodium dodecyl sulfate (SDS)/Decanol and aqueous solution. Such kinds of systems are widely used in different fields, such as pharmaceutics or cosmetics, as well as for the synthesis of nanostructured material [28]. The mentioned system is extracted from the MD simulations. The visualization capabilities of the given system are illustrated in Figure 4.1. The colors are: red - O, silver - H, gray - C, yellow - S, green - CL and purple - Na. The data are the outputs of MD experiments using GROMACS software package [29]. The visual presentation makes it possible to reveal the polymer absorption features on SDS bilayer, as well as, the information coming from the decanol molecules' orientation. One can track that the decanols, which are located between the SDS methyl groups, are mostly in upright position. The vital information coming from this visualization is that the PDADMAC molecules in two layers have different conformations, i.e. a more folded and a more flat conformation. Note that the MD results are in full agreement with our experimental findings, for instance the coexistence of two conformations have been experimentally observed, which argues that MD simulations under the same conditions are consistent to the experiments. Thus, the visualization of the systems gives us information about the coexistence of two lamellar phases in surfactant-based systems induced by polyelectrolyte, as well as, about the lamella features, such as undulations, etc.

Using a present platform, important processes such as force pulling and nucleic acids hybridization have been addressed. Pulling simulation of helical B-DNA with the sequence d(CGCAAATTTCGC)2 has been performed. The system presented in Fig. 4.2 contains 417688 atoms, including dsDNA, water molecules, and 100 mM NaCl. We also considered the same system, containing MgCl2 instead of NaCl.

In Fig. 4.2. d(CGCAAATTTCGC)2 in presence Na+ and Cl- ions, Water molecules are hidden. Ions dissolved in water are indicated in red, the surface of the electronic density of the double - stranded DNA is indicated in macaroon, and the atoms of the double - stranded DNA are indicated in blue, orange, yellow and grey colours. During MD simulation the double-stranded DNA molecule was pulled by an external force and the free energy of double-stranded DNA was measured directly.

The typical final confirmation of the DNA molecule is presented in Fig. 4.3.

The strands of d(CGCAAATTTCGC)2 are separated. Na+ ions are indicated blue, Cl- ions are indicated

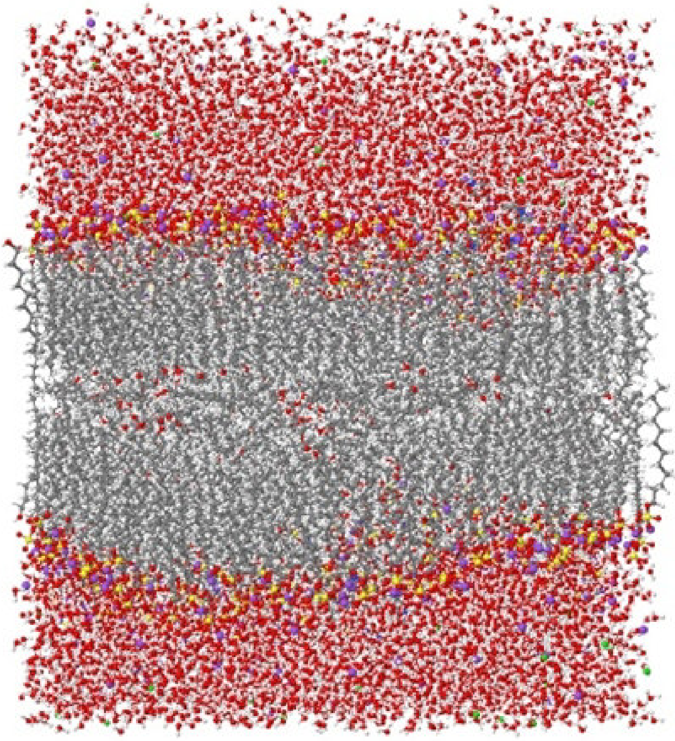


Fig. 4.1. *PDADMAC/SDS/Decanol in water bulk.*

sky, atoms of the two single - strands DNA are indicated using various colors depending on the type of atom.

The visual presentation makes it possible to observe the single strands separation of the double-stranded DNA caused by external pulling.

5. Conclusion. Workflow system is a basic model or pattern that provides support for research and scientific experiments by containing all the all necessary steps of discovery based on the execution of different simulations and having the possibility to visualize the results to get a better understanding of the outputs.

The deployed infrastructure gives the biologists in Armenia and beyond the possibility to increase the visibility of the actual laboratory processes, help them in examining the processes' impact on each other and which activities have more influence on the whole process. The system also gives a possibility to understand the relationship of the small processes in a larger system and how they interact with each other. Having a single point for all distributed data in Armenia enables researchers to collaborate more easily and to share their knowledge.

Using this system helps to unfold the complexity of any scientific problem and their domain, and also to identify the redundancy of the conducted steps to avoid them in the future. There is a plan to expand the system to contain pre-defined workflows for different activities in the domain of biology, and also to enhance the system with more visualization features such as a visual comparison between two elements, display different animation etc.

Acknowledgments. The research leading to these results has been co-funded by the European Commission under the H2020 Research Infrastructures contract no. 675121 (project VI-SEEM) and the State Project entitled "Towards the Green e-Infrastructures" funded by the Government of the Republic of Armenia.

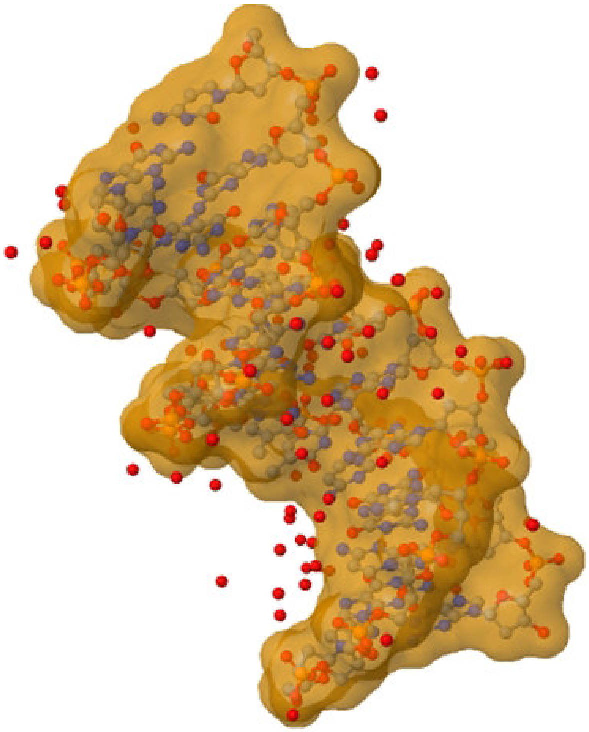


Fig. 4.2. *d(CGCAAATTTCGC)2 in presence Na+ and Cl- ions, Water molecules are hidden.*

REFERENCES

[1] X. Yang, L. Wang,G. von Laszewski, *Recent Research Advances in e-Science, Cluster Compute 12*, (2009), pp. 353–356.

[2] Z. Yin, H. Lan, G. Tan, M. Lu, A. V.Vasilakos, W. Liu, *Computing Platforms for Big Biological Data Analytics*, Perspectives and Challenges, Computational and Structural Biotechnology Journal, Volume 15 (2017), pp. 403–411.

[3] B. Alder, T. Wainwright, *Studies in Molecular Dynamics. I. General Method*, J. Chem. Phys. Vol. 31 (1959), pp. 459.

[4] M. Tuckerman, G. Martyna, *modern molecular dynamics methods: Techniques and Applications*, J. Chem. Phys. Vol. 104 (2000), pp. 159–178.

[5] E. Deelman, D. Gannon, M. Shields, *An overview of workflow system features and capabilities. Future Generation Computer Systems* , Workflows and e-science (2009), pp. 528–540.

[6] J.C.T. Kwak, *Polymer-Surfactant System* , Marcel Dekker; New York, Surfactant Science Series volume 77 (1998)

[7] M.M. Rieger, L.D. Rhein, *In Surfactants in Cosmetics* , Marcel Dekker; New York, Surfactant Science Series volume 68 (1997)

[8] M. Malmsten, *Surfactants and Polymers in Drug Delivery*, Drugs and the Pharmaceutical Sciences, Volume 122 (2002)

[9] V. Luzzati, *X-ray diffraction studies of lipid-water systems*, In Biological Membranes, ed. by D. Chapman. Academic Press, London. 1 (1968), pp. 71–123.

[10] A. Poghosyan, L. Arsenyan, H. Astsatryan, *Dynamic Features of Complex Systems*, A Molecular Simulation Study - Modeling and Optimization in Science and Technologies. Vol. 2 (2014), pp. 117–121.

[11] A. Poghosyan, L. Arsenyan, A. Shahinyan, J. Koetz, *Polyethyleneimine Loaded Inverse SDS Micelle in Pentanol/Toluene Media*, Colloids and Surfaces A: Physicochemical and Engineering Aspects (2016), pp. 402–408.

[12] A. Poghosyan, H. Astsatryan, A. Shahinyan, *Parallel Peculiarities and Performance of GROMACS Package on HPC Platforms*, International Journal of Scientific and Engineering Research (2013), pp. 1755–1761.

[13] D.M. Hinckley, J.P. Lequieu, J.J. de Pablo, *Coarse-grained modeling of DNA oligomer hybridization: length, sequence, and salt effects*, J Chem Phys. (2014) 141(3), doi: 10.1063/1.4886336.

[14] A.W. Peterson, R.J. Heaton, R.M. Georgiadis RM, *The effect of surface probe density on DNA hybridization*, Nucleic Acids Res (2001), 29(24):5163-8.

[15] E. Arslan, J. Laurenzi, *An efficient algorithm for the stochastic simulation of the hybridization of DNA to microarrays*, BMC Bioinformatics 10 (2009), pp. 411–427.

[16] I. Wong, N. Melosh, *An Electrostatic Model for DNA Surface Hybridization*, Biophys. J 98 (2010), pp. 2954–2963.

[17] T. Schmitt, B. Rogers, T. Knotts, *Exploring the mechanisms of DNA hybridization on a surface*, J. Chem. Phys 138

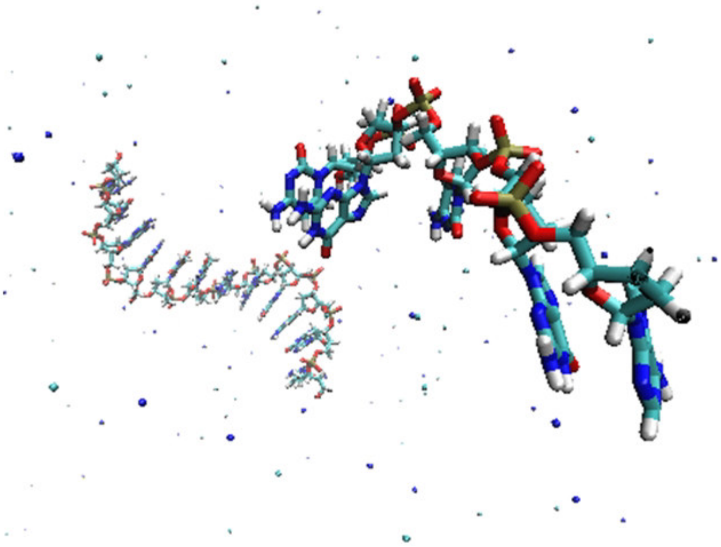


Fig. 4.3. *The strands of d(CGCAAATTTCGC)2 are separated*

(2013), pp. 1755–1761.

[18] A. Karapetian, Z. Grigoryan, Y. Mamasakhlisov, M. Minasyants, P. Vardevanyan, *Theoretical treatment of helixcoil transition of complexes DNA with two different ligands having different binding parameters*, J. Biomol. Struct. Dyn. 34 (2016), pp. 201–205.

[19] G. Hayrapetyan, F. Iannelli, J. Lekscha, V. Morozov, R. Netz, Y. Mamasakhlisov, *Cold melting of RNA with quenched sequence randomness*, Phys. Rev. 113 (2014).

[20] Y. Mamasakhlisov, Sh. Hayryan, V. Morozov, C. Hu, *Kinetics of the long ssRNA: Steady state*, Europhys. Lett. 106 (2014).

[21] Xu, Hao and Russell, Terrell and Coposky, Jason, *iRODS Primer 2: Integrated Rule-Oriented Data System*, Morgan and Claypool (2017).

[22] Matteo Golfarelli and Stefano Rizzi. Data Warehouse Design: Modern Principles and Methodologies. McGraw-Hill, Inc., New York, NY, USA, 1 edition, (2009).

[23] W. Cockshot, M. Atkinson, K. Chisholm, P. Bailey, R. Morrison, *Persistent object management system*, Softw: Pract. Exper., 14 pp. 49–71.

[24] H. Astsatryan, V. Sahakyan, Y. Shoukourian, P. Cros, M. Dayde, J. Dongarra, P. Oster, *Strengthening Compute and Data intensive Capacities of Armenia*, IEEE Proceedings of 14th RoEduNet International Conference - Networking in Education and Research (2015), pp. 28–23.

[25] Y. Shoukourian, V. Sahakyan, H. Astsatryan, *E-Infrastructures in Armenia: Virtual research environments* , in IEEE Proceedings CSIT 2013 - 9th International Conference on Computer Science and Information Technologies, Revised Selected Papers, CSIT'2013, pp. 1–7.

[26] A. Herrez, *Jmol to the rescue, Biochemistry and Molecular Biology Education*, Journal of Colloid and Interface Science(2011), pp. 255–261.

[27] *Irods Cloud Drowser*, https://github.com/DICE-UNC/irods-cloud-browser

[28] A. Poghosyan, L. Arsenyan, J. Koetz, A. Shahinyan, *Molecular dynamics study of poly diallyldimethylammonium chloride (PDADMAC)/sodium dodecyl sulfate (SDS)/decanol/water system*, The Journal of Physical Chemistry B.(2009), pp. 1303–1310.

[29] S. Pronk, et al., *GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit*, Bioinformatics 29, 845–854 (2013)

Edited by: Aneta Karaivanova
Received: Dec 8, 2017
Accepted: Mar 28, 2018

DOI 10.12694/scpe.v19i2.1353
Scalable Computing: Practice and Experience ISSN 1895-1767
Volume 19, Number 2, pp. 139–148. http://www.scpe.org © 2018 SCPE

MODELING AND MOLECULAR DYNAMICS SIMULATIONS STUDY OF ENOL-CARBONATES AND THEIR DERIVATIVES [*]

MILJAN BIGOVIC [†] LUKA FILIPOVIC [‡] ZARKO ZECEVIC [§] AND BOZO KRSTAJIC [¶]

Abstract. In view of the fact that the mechanisms of the interaction of organic molecules with the properties of the drug occur in most cases by their binding to the receptors (proteins), we wanted to examine the interactions of small organic molecules that we synthesized and certain proteins. In this paper the comparison between molecules, modeled by different software packages, and experimental results is performed. The series of new molecules are synthesized using Barbier reaction of allylation of carbonyl compounds with 4-(bromomethyl)-1,3-dioxol-2-one as a highly fictionalized allylic synthon. Their structure was determined by spectroscopic methods (NMR-, IR- and UV/VIS-spectroscopy and MS-spectrometry), while for three individual molecules analysis of x-ray diffraction (X-ray analysis) was also performed, which gave the final confirmation of the exact arrangement of all atoms in the space. The molecules are represented in standard chemical format, visualized and prepared for simulations. In order to obtain datasets that further can be used to examine and analyze interactions with well-known proteins, molecular dynamic simulations are performed. The purpose of this research was to present that using powerful computer infrastructure and appropriate software tools, an accurate molecule models can be created.

Key words: molecule modeling, molecular dynamics simulations, organic molecules, protein, interaction, parallel simulation

AMS subject classifications. 65Y05, 92C40, 92E10

1. Introduction.

Synthesis of biologically active organic molecules in the laboratory is one of the most important methods in the procedure of the creation of new drugs today. In this way, the possibility of synthesizing molecules that are available from natural sources in very small quantities is opened up. Moreover, the synthesis of molecules that are not present in nature at all, but that could have far stronger and better biological effects in relation to natural ones has become possible. All methods of chemical synthesis used in the process of obtaining such molecules are of great importance.

In order to speedup the research of synthetic molecules and their interactions, as well as reduce analysis costs, molecular dynamics (MD) simulations on distributed computing resources are used. They consist of many emerging techniques with potential applications in diverse areas of modern chemistry, pharmacology, pharmaceutical chemistry and biochemistry. Over the past three decades, MD has evolved as an area of importance for understanding the atomic basis of complex phenomena such as molecular recognition, protein folding, and the transport of ions and small molecules across membranes. The application of MD simulations with experimental approaches have provided an increased understanding of protein structure-function relationships and demonstrated capacity in pharmaceutical and medical analysis and drug discovery [1]. Existence of computing model that can simulate and predict molecule interactions with well-known proteins [2] will give enormous contribution in resolving this complex problem. The discovery of a new drug takes 12-15 years, and costs between 600 and 800 million US dollars. The application of these methods will contribute to the selection of candidate molecules that could be biologically active and thus reduce the list of tested molecules. This would certainly contribute to reducing the time spent in the laboratory and the financial expenditures [3]. A variety of publications testifies about growth and successful MD research studies in various fields of life science [4][5][6][7].

Here, models of three newly synthesized molecules are presented. In order to obtain datasets that further can be used to examine and analyze interactions with well-known proteins, molecular dynamic simulations are performed. Since the MD simulations are computationally intensive, they are significantly accelerated by using dozens or hundreds of computing cores, which gave a significant benefit of using of computer models in comparison to the traditional way of synthesis of new molecules and their biological activity. In our analysis we used GROMACS [8], a software for parallel simulations in the molecular dynamics of the given ligand-receptor

[*] The research leading to these results has been co-funded by the European Commission under the H2020 Research Infrastructures contract no. 675121 (project VI-SEEM).

[†] Faculty of natural sciences, University of Montenegro, Podgorica, Montenegro (miljan@ac.me).

[‡] Center of information system, University of Montenegro, Podgorica, Montenegro (lukaf@ac.me).

[§] Faculty of electrical engineering, University of Montenegro, Podgorica, Montenegro (zarkoz@ac.me).

[¶] Faculty of electrical engineering, University of Montenegro, Podgorica, Montenegro (bozok@ac.me).

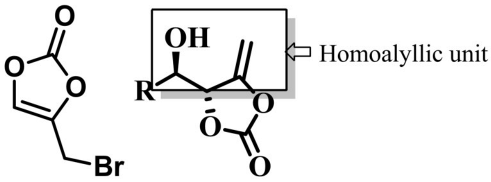


FIG. 2.1. *4-(Bromomethyl)-1,3-dioxol-2-one*

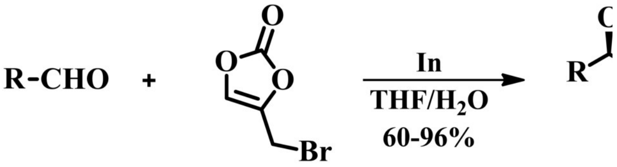


FIG. 2.2. *General view of allylation reaction of carbonyl compounds (obtaining the enol-carbonates)*

pair.

2. Enol Carbonates and their Derivatives as Simulation Models. The homoaline alcoholic unit is a very interesting and very common structural unit that characterizes a large number of physiologically active molecules. In this way, important precursors are obtained, which are further used in the synthesis of a numbered natural products with pronounced antibiotic, antimicrobial, antifungal or even cytostatic effects. Some of them are: methylenolactocin [9], myxothiazol A [10] , eupomatilone [11] , sialic acid [12] and similarly.

We have previously reported the use of 4-(bromomethyl)-1,3-dioxol-2-one [Fig 2.1] as a highly functionaliyed allylic synthon [13]. We assumed that in the reaction of this reagent with aldehydes or ketones it would provide the possibility of obtaining homoallylicacohols, obtained in the form of protected enol carbonates:

Indium mediated allylation of carbonyl compounds with this reagent in water as a solvent alows for the synthesis of α, β-hydroxy ketones, in protected or either free form. The presence of fragment with three atoms of oxygen is certainly interesting, especially in drug design industry, because it provides possibilities for further functionalization of derivatives and in organic synthesis [Fig 2.2].

After several decades of development, the preparation and application of organoindium reagents in organic synthesis have seen leaps and jumps. The feasibility of using this reagents in aqueous media permits the direct functionalization of water-soluble substrates, which is important both from the ecological and the economic point of view.

With obtained enol-carbonates, we have examined two sets of reactions: in first, we transferred them to free keto-diols using mercury salts, and in the second, after treatment with base, they were converted enol-carbonates into saturated cyclic carbonates [1] [Fig 2.3]:

Deprotection of enol-carbonates leads to free α, β-dihydroxyketone - a unit which is a common structural motive of many natural products and physiologically active compounds which recommended molecules for further analysis.The third direction of the development of our reaction was the creation of a new carbon-carbon bond. Namely, enol-carbonates undergo Heck reaction with aryl iodides in the presence of silver trifluoroacetate, to give the corresponding arylated products, obtained in moderate yields [14] [Fig 2.4].

[1] Protocol: To a 10 mL flask equipped with a glass stopper and a magnetic stirrer, 4- (bromomethyl) -1,3-dioxol-2-one 300 (50.0 mg, 0.28 mmol), THF (0.5 mL) and water (1.0 mL) , and then indium powder (32.1 mg, 0.28 mmol). The resulting suspension was intensively stirred at room temperature, whereby a white blur occurred. A carbonyl compound (0.19 mmol) was then added to the suspension. The reaction mixture was stirred for 15 minutes at room temperature, and the reaction stream was followed by thin layer chromatography (eluent: 20% acetone in petroleum ether). The reaction mixture was partitioned between dichloromethane (5 mL) and water (5 mL); the aqueous extracts were washed twice with 5 mL of dichloromethane. The combined organic extracts were dried over anhydrous magnesium sulphate, evaporated and evaporated on a rotary vacuum evaporator. The crude product was purified by column-based column chromatography, and wherever possible by subsequent recrystallization.

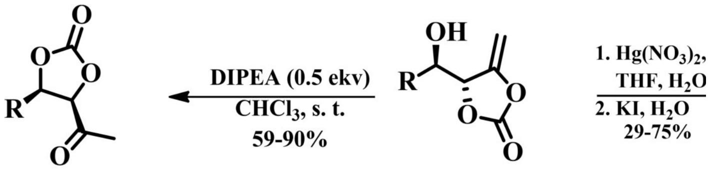


Fig. 2.3. *Functional transformations of enol-carbonates*

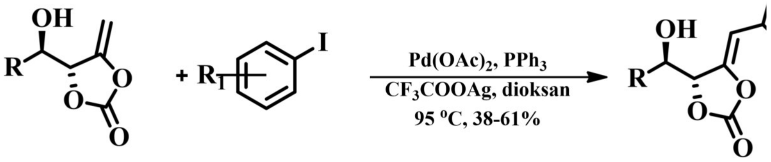


Fig. 2.4. *Heck reaction with enol carbonates- method for carbon-carbon bond formation*

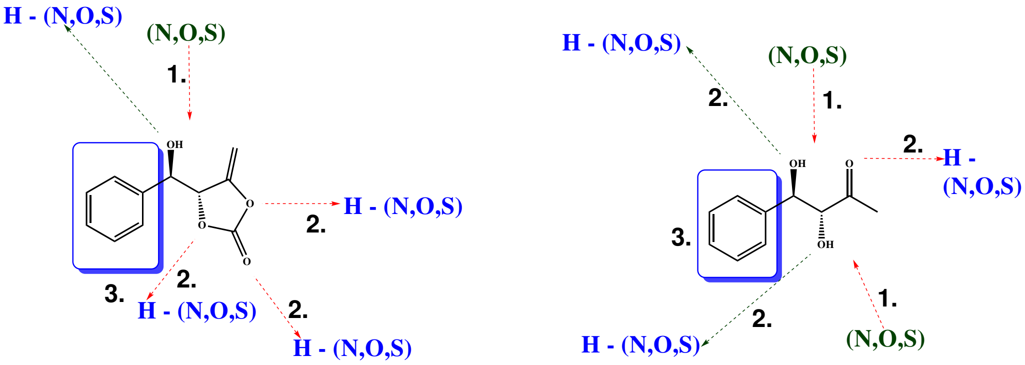


Fig. 2.5. *Potential (predicted) interactions between molecules and active sites of protein*

In view of the fact that all these molecules were newly synthesized and that no biological test was done with them, we thought that molecular-dynamic calculations and simulations could provide some important data on the interactions of functional groups that our molecule are having with active sites of proteins and receptors, which are, in fact, lateral strings of amino acids.

Based on theoretical knowledge, that according to the structural characteristics of the described molecules it should be assumed which regions could be involved in characteristic interactions with residues from active protein centers.

In the following examples of selected molecules from all of the listed reactions, we will show what possible interactions we expect, and which will be confirmed by molecular calculations [Fig 2.5].

In Fig 2.5, the meanings are the followings:

1. Forms hydrogen bonds with electronegative atoms from the active site of proteins (nitrogen, sulphur, oxygen).

2. Forms hydrogen bonds with hydrogen atoms with hydrogen atoms, which are in the active site of the protein, and are themselves bound directly to an electronegative atom (nitrogen, sulphur, oxygen).

3. π - π -non-polar interactions (part of the molecule that potentially builds these interactions is shown in a blue rectangle)

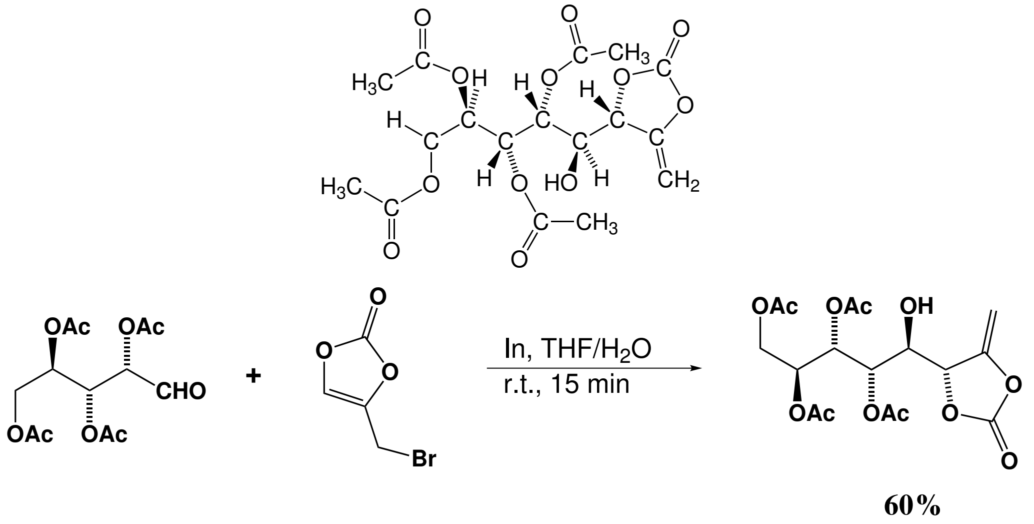


FIG. 3.1. *Structural formula of Enol carbonate ((2R,3R,4R,5S)-5-hydroxy-5-((S)-5-methylene-2-oxo-1,3-dioxolan-4-yl) pentane-1,2,3,4-tetraacetate)*

3. Selected Molecules for Molecular Simulations. The most interesting synthetized compounds for further analysis were:

1. (2R,3R,4R,5S)-5-hydroxy-5-((S)-5-methylene-2-oxo-1,3-dioxolan-4-yl) pentane-1,2,3,4-tetraacetate, see Fig 3.1
2. (R)-4-((R)-hydroxy(phenyl)methyl)-5-methylene-1,3-dioxolan-2-one, see Fig 3.2
3. (R)-4-((R)-(4-chlorophenyl)(hydroxy)methyl)-5-methylene-1,3-dioxolan-2-one, see Fig 3.3

Specified molecules were basis for further molecule modeling, visualization, simulations and comparison. Compounds were synthesized according to the standard experimental procedures and were obtained as white needle crystals. The crystals were carefully purified by the method of recrystallization from a mixture of organic solvents in order to obtain a representative monocrystal. This high quality monocrystal was used for X-ray structural analysis, which belongs to a powerful, accurate and representative technique of analysis of organic molecules. As a result, diffraction analysis provides the exact spatial model of atoms and bonds in a given molecule (Figure 3.1-3.3), which was necessary for building of the compound model for further computer simulations.

Molecules are modeled and prepared for MD simulation by VMD [15], ChemSketch [16] and OpenBabel [17]. First, the molecules are sketched by using ChemSketch a molecular modeling program used to routinely draw and modify structural formula of organic and inorganic molecules. By ChemSketch, the molecules can be saved in standard formats, such as MDL molfile, which holds all information about the atoms, bonds and their types and connectivity, as well as the coordinates of the atoms in the molecule.

After geometrical optimization in OpenBabel, molecules were converted into a standard format PDB (Protein Data Bank) which is often used in public compound databases [2]. Figures 3.4-3.6 show the resulting molecules, visualized in VMD. Computer molecule models has identical structure with molecule generated in the laboratory and it can be used as an input file for molecular dynamics simulations. We conclude that there was no breakdown of links, deformation of angles or possible replacement of individual atoms. Verification of identical models was necessary as a key step for further simulations. All effects would lead to the formation of molecule with some different valyes of angles and distances between the atoms in relation to the one with which we want to examine the simulation.

Main difference between laboratory and computer analysis is their duration and cost computer simulations are much faster when it uses many computer cores and many different combinations between molecules and proteins can be analyzed in shorter time interval. Molecular simulations have great advantages in terms of synthesis and in vivo and in-vitro biological activity testing. Namely, time is shorter, there is no need expensive

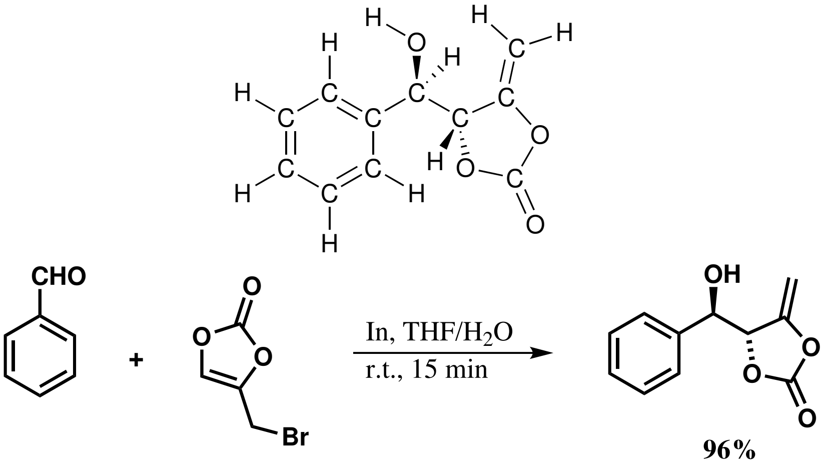


Fig. 3.2. *(R)-4-((R)-hydroxy(phenyl)methyl)-5-methylene-1,3-dioxolan-2-one*

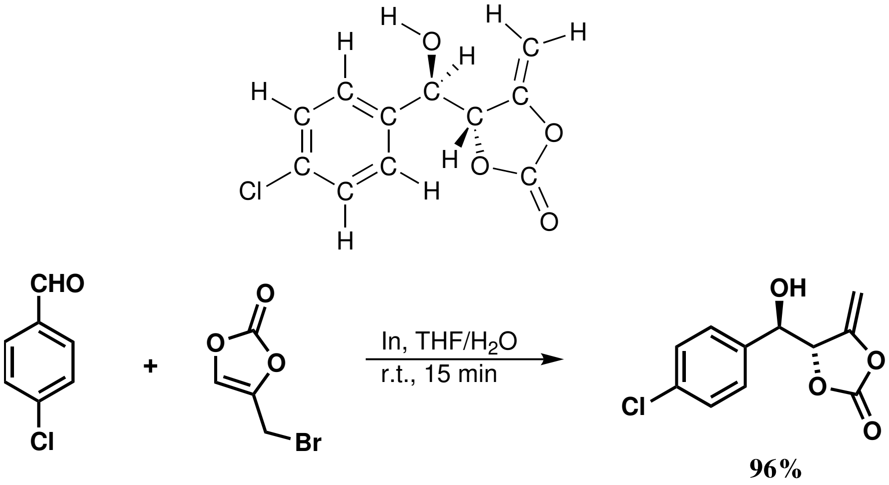


Fig. 3.3. *(R)-4-((R)-(4-chlorophenyl)(hydroxy)methyl)-5-methylene-1,3-dioxolan-2-one*

equipment and chemicals, and they have a very high accuracy.

4. Moleclar Dynamics Simulations with Protein . We have simulated a system containing a protein (T4 lysozyme L99A/M102Q) in complex with presented ligands. The simulation results related to the second molecule are presented. The similar results are obtained for the other molecules. Gromacs in conjunction with 24 CPU cores is exploited for simulations. PRODRG server [18] is used to generate a small-ligand topology for use with GROMACS force fields family. Protein-ligand interactions are widely studied based on topologies produced by this program. Simulation are performed by following the procedures described in [19], where all necessary details about system configuration can be found.

Possible steric clashes or inappropriate geometry of the system are prevented by performing the energy minimization. The convergence of the potential energy is shown in Figure 4.1, indicating the convergence in 350 iterations.

Energy minimization ensures that the starting structure is reasonable in terms of geometry and solvent orientation. Before the real dynamics starts, the solvent and ions around the protein must be equilibrated, which is performed in two phases. The first phase is conducted under isothermal-isochoric conditions (constant

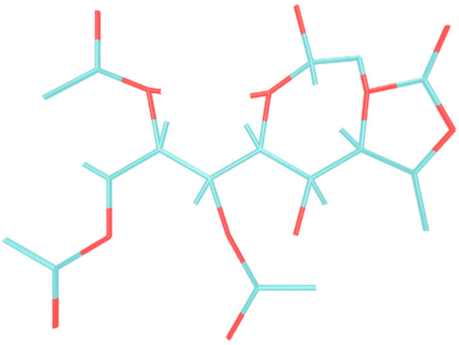


Fig. 3.4. *Structural formula of Enol carbonate ((2R,3R,4R,5S)-5-hydroxy-5-((S)-5-methylene-2-oxo-1,3-dioxolan-4-yl) pentane-1,2,3,4-tetraacetate)*

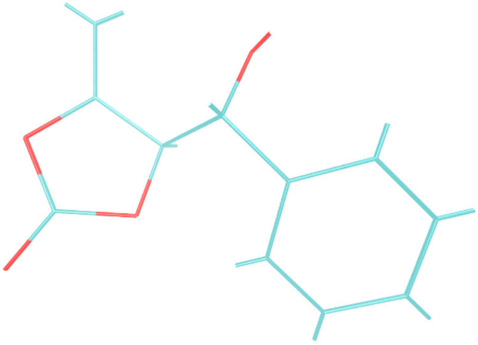


Fig. 3.5. *(R)-4-((R)-hydroxy(phenyl)methyl)-5-methylene-1,3-dioxolan-2-one*

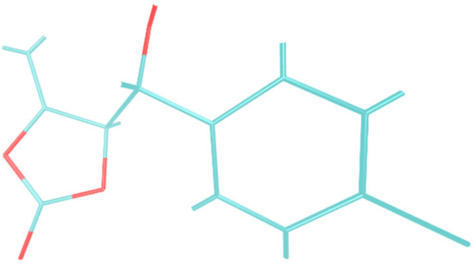


Fig. 3.6. *(R)-4-((R)-(4-chlorophenyl)(hydroxy)methyl)-5-methylene-1,3-dioxolan-2-one*

Number of particles, Volume, and Temperature, NVT). The timeframe for this simulation was set to 100ps. The system temperature after equilibration is shown in Figure 4.2. The 300K target value is rapidly reached, further remaining stable over time.

The second equilibration phase includes the system pressure stabilization. This phase is performed under isothermal-isochoric conditions (constant Number of particles, Pressure, and Temperature, NPT). The timeframe was also set to 100ps for this simulation. Figure 4.3 indicates the system pressure variations over the simulated timeframe, which was an expected behavior.

After equilibrating the system at the desired temperature and pressure, production MD phase is performed. For this computationally demanding simulation phase, high-performance computers are required. In this exam-

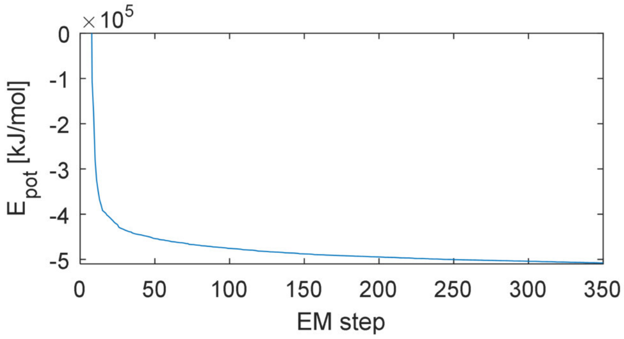


FIG. 4.1. *Energy minimization phase*

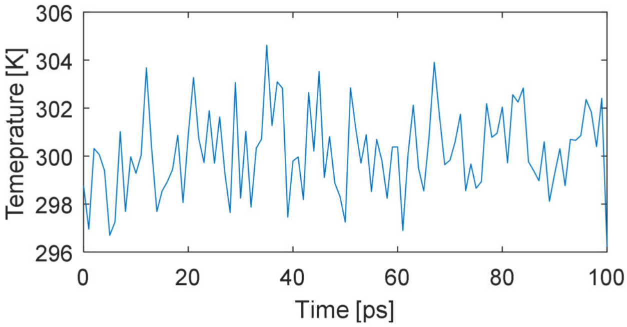


FIG. 4.2. *Temperature of the system*

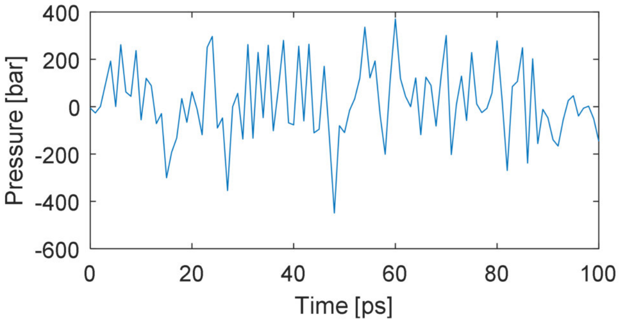


FIG. 4.3. *Pressure of the system*

ple, 1ns simulation is performed. GROMACS has some built-in tools for MD analysis. The radius of gyration (Rg) of a protein is used as the compactness measure. Based on the results presented in Figure 4.4, it can be observed that Rg has reasonably invariant values, which means that the protein remains in its folded form over the timeframe of 1ns at 300 K. Another indicator of the system compactness is Root Mean Square Deviation (RMSD). The Figure 4.5 shows that the RMSD oscillates around 0.15 nm, indicating that the structure is very stable.

Figure 4.6 shows the protein that is the result of the presented simulations. It shows parts stabilized by hydrogen bonding (secondary structure) as well as specific spatial variants of the given protein. According to GROMACS-simulation results, this is the most stable structure of the given protein.

Computer-intensive calculations were performed on FINKI HPC, as one of resources of VI-SEEM project

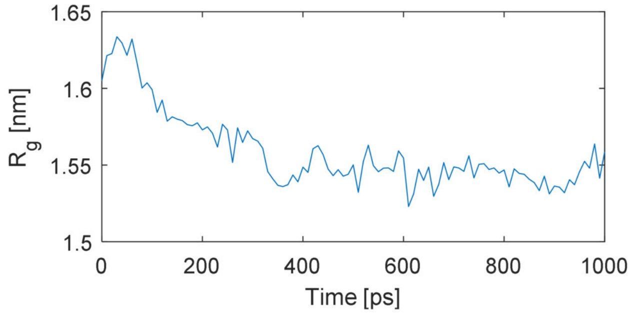


FIG. 4.4. *Radius of gyration*

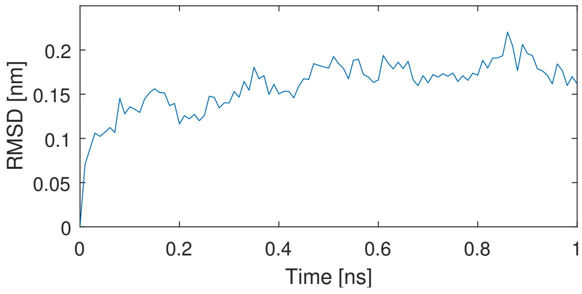


FIG. 4.5. *RMSD*

[20]. The parallel data processing noticeably accelerates the simulations and proves necessity of parallel processing in molecular dynamic simulations. The considered datasets and workflows provide the ability to test various combinations before going into testing in the laboratory.

5. Conclusion. By using computer simulations, which are characterized by high accuracy and precision, a valuable information usable in design new drugs can be provided. Moreover, it would be possible to investigate the interactions of such molecules that have not been made in the laboratory, and in terms of their structural or physiological properties, they are similar to those existing in nature (or synthesized artificially). On the other hand, the application of simulations contributes to a significant reduction in research time, and orientation to those groups of molecules that are highly likely to exhibit a physiological effect after interaction with the given receptor. In this way, the number of molecules that are planned to be tested is significantly reduced.

Finally, by exchanging some functional groups with some other, it would get useful information about the possible interactions (and activities) of such molecules, which would significantly contribute to the synthetic application of the tested reaction or methodology.

In this paper molecular dynamic simulations between series of new molecules with lysozyme are presented. The initial simulation results showed that there is no interaction between the synthesized molecules (enol-carbonate) and the lysosome. Before concluding that there is no interaction of an organic molecule with protein at all, several experiments with the same systems need to be conducted.

Acknowledgments. The research leading to these results has been co-funded by the European Commission under the H2020 Research Infrastructures contract no. 675121 (project VI-SEEM).

REFERENCES

[1] NAIR, P. C., MINERS, J. O., *Molecular dynamics simulations: From structure, function relationships to drug discovery*, in In Silico Pharmacology, 2014, 2, 4.

[2] *RCSB Protein Data Bank*, http://www.rcsb.org/

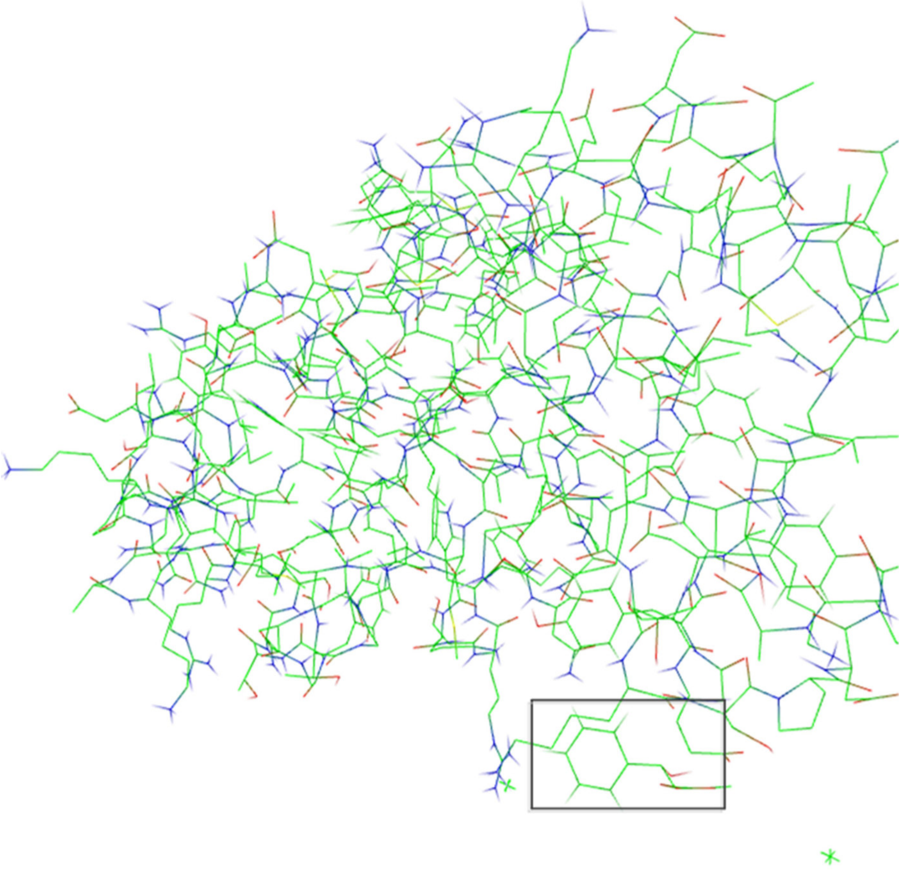


Fig. 4.6. *The most stable possible conformation of protein, presented in VegaZZ*

[3] Z. Cekovic i saradnici, *Molekuli u tajnama zivota i svetu oko nas*, Zavod za udzbenike, Beograd, 2009, pp 139-149

[4] E. Athanasiadis, Z. Cournia, G. Spyrou, *ChemBioServer: a web-based pipeline for filtering, clustering and visualization of chemical compounds used in drug discovery*, Bioinformatics Vol. 28, No. 22, November 2012, pp. 3002-3003.

[5] E. Lionta, G. Spyrou, D. K. Vassilatis, Z. Cournia, *Structure-based virtual screening for drug discovery: principles, applications and recent advances*, Curr. Top. Med. Chem., Vol.14, No. 16, Oct. 2014, pp. 1923-1938.

[6] M. Mansha, U. U. Kumari, Z. Cournia, N. Ullah, *Pyrazole-based potent inhibitors of GGT1: Synthesis, biological evaluation and molecular docking studies*, Eur. J. Med. Chem., Vol. 124, Nov. 2016, pp. 666-676.

[7] Y. Wang, P. Gkeka, J. E. Fuchs, K. R. Liedl, Z. Cournia, *DPPC-cholesterol phase diagram using coarse-grained Molecular Dynamics simulations*, Biochim. Biophys. Acta, Vol. 11, Nov. 2016, pp. 2846-2857.

[8] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A.E. Mark, H.J. Berendsen, *GROMACS: fast, flexible, and free*, J Comput Chem. Vol.16, 2005., 170118.

[9] Loh, T.P., Lye, P.L., *A concise synthesis of ()-methylenolactocin and the formal synthesis of ()-phaseolinic acid*, Tetrahedron Lett, 2001, 42, 3511.

[10] Backhaus, D, Tetharedron Lett., 2000, 41, 2087.

[11] Kabalka, G. W., Venkataiah, B., Tetrahedron Lett., 2005, 46, 7325.

[12] Chappel, M. D., Halcomb, R L., Org. Lett., 2000, 2, 2003.

[13] Bigovic, M., Maslak, V., Tokic-Vujosevic, Z., Divjakovic, V., Saicic,*A useful synthetic equivalent of a hydroxyacetone enolate*, R. Org. Lett. 2011, 13, 4720.

[14] Bigovic, M.; Skaro, S.; Maslak, V.; Saicic, R. N., *Expanding the scope of the indium-promoted allylation reaction: 4-(bromomethyl)-1.3-dioxol-2- one as a synthetic equivalent of a 3-arylhydroxyacetone enolate*, Tetrahedron Lett. 2013, 54, 6624

[15] Humphrey W., Dalke A., Schulten K., *VMD: Visual molecular dynamics*, Journal of Molecular Graphics Volume 14, Issue 1, February 1996, Pages 33-38

[16] *ACD/ChemSketch for Academic and Personal Use*, ACD/Labs.com. [online] Available at: http://www.acdlabs.com/resources/freeware/chemsketch/

[17] *Open Babel : The Open Source Chemistry Toolbox*, http://openbabel.org/

[18] A. W. Schttelkopf, D. M. F. van Aalten , *PRODRG: a tool for high-throughput crystallography of protein-ligand complexes*, Acta Crystallogr D60, 2004, 13551363.

[19] *GROMACS Tutorial - Baven Lab*, http://www.bevanlab.biochem.vt.edu/Pages/Personal/justin/gmx-tutorials/

[20] *H2020 project VI-SEEM*, https://vi-seem.eu

Edited by: Aneta Karaivanova
Received: Jan 6, 2017
Accepted: Apr 2, 2018

DOI 10.12694/scpe.v19i2.1344
Scalable Computing: Practice and Experience ISSN 1895-1767
Volume 19, Number 2, pp. 149–159. http://www.scpe.org © 2018 SCPE

SEMIEMPIRICAL ATOM-CENTERED DENSITY MATRIX PROPAGATION APPROACH TO TEMPERATURE-DEPENDENT VIBRATIONAL SPECTROSCOPY OF IRINOTECAN

BOJANA KOTESKA,[*] MAJA SIMONOSKA CRCAREVSKA,[†] MARIJA GLAVAS DODOV,[‡] JASMINA TONIC RIBARSKA,[§] AND LJUPČO PEJOV[¶]

Abstract. In the present study, a molecular dynamics study of irinotecan molecule with the atom-centered density matrix propagation scheme was carried out at AM1 semiempirical level of theory, at series of different temperatures, ranging from 5 K to 300 K. Molecular dynamics simulations were performed within the NVE ensemble, initially injecting (and redistributing among the nuclei) various amounts of nuclear kinetic energies to achieve the desired target temperatures. Subsequently to initial equilibration phase of 2 ps, productive simulations were carried out for 8 ps. The accuracy of simulations and the closeness of the generated trajectory to those at the Born-Oppenheimer surface were carefully followed and analyzed. To compute the temperature-dependent rovibrational density of states spectra, the velocity-velocity autocorrelation functions were computed and Fourier-transformed. Fourier-transformed dipole moment autocorrelation functions were, on the other hand, used to calculate the temperature-dependent infrared absorption cross section spectra. The finite-temperature spectra were compared to those computed by a static approach, i.e. by diagonalization of mass-weighted Hessian matrices at the minima located on the potential energy surfaces. Thermally-induced spectral changes were analyzed and discussed. The advantages of finite-temperature statistical physics simulations based on semiempirical Hamiltonian over the static semiempirical ones in the case of complex, physiologically active molecular systems relevant to intermolecular interactions between drugs and drug carriers were pointed out and discussed.

Key words: irinotecan, molecular dynamics, atom-centered density matrix propagation scheme, *anharmonic* vibrational frequencies, statistical physics simulations, theoretical spectroscopy

AMS subject classifications. 70F99, 82B30, 92E10

1. Introduction. Appropriate delivery of drugs and other physiologically active substances to the tissue in which they are expected to exert their activity is a fundamental issue in life sciences. We have actually witnessed a drastic paradigmatic shift in pharmaceutical sciences concerning the drug delivery issues with the advent of nanoscience [1]. The medical treatment of diseases, the very essence of drugs pharmacological activity as well as its distribution and metabolism critically depends on proper delivery. Often the delivery systems (the drug carriers) are designed such to enable a controlled release of the active ingredients as well. At the present state of the art within the field, various forms of drug delivery systems have evolved towards nanoparticles acting as encaging systems for the physiologically active components. To understand the physico-chemical basis of the encaging phenomena, often a close collaboration between theory and experiment is crucial. In-depth studies of the molecular basis of the nanoparticle-drug intermolecular interactions (often being of noncovalent type) can even lead to a much more efficient design of novel carriers. In the present study, we tackle this issue focusing on a rather interesting and important hydrophilic drug irinotecan. Irinotecan is a rather important physiologically active substance, as it is used in the treatment of colon cancer, as well as in treatment of small cell lung cancer together with cisplatin. This hydrophilic drug has been recently incorporated into nanoparticle carriers composed by the poly lactic-co-glycolic acid copolymer (PLGA) and coadsorbed PEO-PPO-PEO (polyethylene oxide polypropylene polyethylene oxide) copolymer [2]. Among the other experimental techniques, Fourier transform infrared spectroscopy has been utilized to study the structural and dynamic changes of both guest and host molecules in the course of drug-nanopartice interaction upon encapsulation. The changes in experimentally measured spectral patterns upon encapsulations were shown to be rather subtle. Therefore, in order to get in-depth insights into the spectroscopic manifestations of the

[*] Faculty of Computer Science and Engineering, "Ss. Cyril and Methodious University", Rugjer Boskovikj 16, 1000 Skopje, Republic of Macedonia (bojana.koteska@finki.ukim.mk).

[†] Institute of Pharmaceutical Technology, Center of Pharmaceutical Nanotechnology, Faculty of Pharmacy, "Ss. Cyril and Methodius University", Majka Tereza 47, 1000 Skopje, Republic of Macedonia(msimonoska@ff.ukim.edu.mk).

[‡] Institute of Pharmaceutical Technology, Center of Pharmaceutical Nanotechnology, Faculty of Pharmacy, "Ss. Cyril and Methodius University", Majka Tereza 47, 1000 Skopje, Republic of Macedonia(magl@ff.ukim.edu.mk).

[§] Institute of Applied Chemistry and Pharmaceutical Analysis, Faculty of Pharmacy, "Ss. Cyril and Methodius University", Majka Tereza 47, 1000 Skopje, Republic of Macedonia(jato@ff.ukim.edu.mk).

[¶] Institute of Chemistry, Faculty of Science, "Ss. Cyril and Methodius University", P.O. Box 162, 1001 Skopje, Republic of Macedonia(ljupcop@pmf.ukim.mk).

noncovalent interactions taking place between the incorporated drug molecule and the co-polymeric nanocarrier, substantial theoretical support is required. In the course of achieving this aim, we have recently undertaken a theoretical study of irinotecan molecule at several semiempirical levels of theory, as well as with a density functional theory (DFT) based approach [3]. We have shown in this study that theoretical treatment of this molecular system employing the semiempirical AM1 Hamiltonian is capable of reproducing most of its basic structural and spectroscopic properties computed at B3LYP/6-31G(d, p) level of theory. However, as discussed in [3], static quantum mechanical calculations inherently refer to a system at 0 K, while essentially all processes relevant to its physiological activity and incorporation into drug carrier systems occur at finite temperatures, quite above absolute zero. If one wants a reliable description of the physical phenomena in question, therefore, temperature-induced effects must be properly accounted for. To achieve this aim, in the present paper we study the temperature dependence of structure and vibrational spectroscopic properties of isolated irinotecan molecule employing molecular dynamics simulations based on the atom-centered density matrix propagation scheme [4] with a semiempirical AM1 Hamiltonian [5]. Using such methodological approach, sufficiently long simulations may be performed with a sufficiently accurate Hamiltonian for a proper description of the mentioned properties. This is a first step towards a development and implementation of rigorous theoretical approach aiming at an in-depth understanding of subtle structural and spectroscopic changes in the course of irinotecan incorporation (encapsulation) into drug nanocarrier systems.

2. Computational details.

2.1. Atom-centered density matrix propagation (ADMP) simulations. Semiempirical molecular dynamics simulations of free irinotecan molecule were carried out employing the atom-centered density matrix propagation (ADMP) scheme. This particular method belongs to the extended Lagrangian approaches to molecular dynamics, based on propagation of the density matrix, using Gaussian-type basis functions [4]. The extended Lagrangian of the studied system is written in the form:

$$L = \frac{1}{2}Tr(V^T MV) + \frac{1}{2}\mu Tr(WW) - E(R,P) - Tr[\Lambda(PP-P)] \tag{2.1}$$

In (2.1), M, R and V are the nuclear masses, positions and velocities, respectively, while P, W and μ denote the density matrix, density matrix velocity and the fictitious mass for the electronic degrees of freedom, correspondingly. Λ is a Lagrangian multiplier matrix, and is here used to impose the constraints on the total number of electrons in the system and on the condition of idempotency of the density matrix.

Applying the principle of stationary action, one subsequently arrives at the Euler-Lagrange equations for density matrix propagation, which can be written in the form:

$$\mu\frac{d^2P}{dt^2} = -\left[\left.\frac{\partial E(R,P)}{\partial P}\right|_R + \Lambda P + P\Lambda - \Lambda\right] \tag{2.2}$$

$$M\frac{d^2R}{dt^2} = -\left.\frac{\partial E(R,P)}{\partial R}\right|_P \tag{2.3}$$

For the purpose of the present study, equations (2.2) and (2.3) were integrated by the velocity Verlet algorithm. Within this algorithm, the density matrix propagation is given by:

$$P_{i+1} = P_i + W_i\Delta t - \frac{\Delta t^2}{2\mu}\left[\left.\frac{\partial E(R_i,P_i)}{\partial P}\right|_R + \Lambda_i P_i + P_i\Lambda_i - \Lambda_i\right] \tag{2.4}$$

$$W_{i+1/2} = W_i - \frac{\Delta t}{2\mu}\left[\left.\frac{\partial E(R_i,P_i)}{\partial P}\right|_R + \Lambda_i P_i + P_i\Lambda_i - \Lambda_i\right] = \frac{P_{i+1}-P_i}{\Delta t} \tag{2.5}$$

$$W_{i+1} = W_{i+1/2} - \frac{\Delta t}{2\mu}\left[\left.\frac{\partial E(R_{i+1},P_{i+1})}{\partial P}\right|_R + \Lambda_{i+1}P_{i+1} + P_{i+1}\Lambda_{i+1} - \Lambda_{i+1}\right] \tag{2.6}$$

Note that the extended Lagrangian molecular dynamics methodologies are especially well-suited for systems with very large number of the degrees of freedom, as the electronic subsystem is not treated by a full solution by e.g. a self-consistent field procedure; rather, it is propagated along with the nuclear degrees of freedom (which are, in turn, treated classically). This is achieved by an adjustment of the time scales of the mentioned motions (electronic and nuclear).

In the present study, as a starting point for the ADMP simulations, we have chosen the absolute minimum on the AM1 potential energy surface (PES) of the title molecule. This minimum has been obtained by our previous careful investigation of the potential energy landscapes of the title molecule employing a series of semiempirical Hamiltonians (AM1, PM3, PM6), as well as density functional levels of theory (e.g. B3LYP/6-31G*(d,p))* [3]. The minimum has been located employing the Schlegel's gradient optimization algorithm [6]. Subsequently to the geometry optimization phase, harmonic vibrational analysis has been performed in order to compute the harmonic vibrational frequencies (at 0 K) as well as to test the character of the located stationary point on the explored PES. Absence of imaginary frequencies (negative eigenvalues of the Hessian matrix) served as an indication that a true minimum on the PES has been reached.

Starting from the located minimum on the AM1 PES, semiempirical molecular dynamics simulations have been performed within the mentioned ADMP scheme. All ADMP simulations have been performed in the microcanonical (NVE) ensemble. Various amounts of initial nuclear kinetic energies were initially injected to the system (and distributed among the atoms) in order to reach the finally desired temperatures. No thermostats were applied to maintain a constant temperature during each of the ADMP simulations. As shown below, such approach has led to acceptable temperature fluctuations throughout the simulation. Since we want to compute the spectroscopic properties of the title system within the dynamical approach, i.e. within the time correlation function approach, the dynamics of molecular system has to be sampled properly. Introducing a thermostat to maintain constant temperature, aside from allowing for much smaller temperature fluctuations, would however, severely distort the system's dynamics. Series of ADMP simulations were carried out at target temperatures of 5 K, 100 K, 150 K and at 300 K. This temperature range was chosen to follow the temperature-evolution of the vibrational spectroscopic properties of the title molecule starting from a situation where quasi-harmonic behavior is expected, up to a situation which is often encountered in vibrational spectroscopic experiments under ambient conditions.

Upon initial velocity assignment, the system was allowed to equilibrate for 2 ps. Equilibration phase was followed by production (simulation) phase which was 8 ps long. To integrate the equations of motions, a time step of 0.2 fs was used for productive computations. The fictitious electron mass was set to 0.1 amu and the Cholesky basis for the orthonormal set was used.

Both the initial geometry optimizations with the semiempirical AM1 Hamiltonian and the subsequent ADMP simulations were performed with the Gaussian09 series of codes [7].

2.2. Time-correlation functions approach to spectroscopic properties. To compute the finite-temperature vibrational spectra of the studied irinotecan molecule from semiempirical molecular dynamics simulations, we relied on the time correlation functions approach, which is fundamentally based on the linear response formalism [8]. Within this approach, a particular autocorrelation function is computed from the collected data throughout the MD trajectory and this is sequentially Fourier-transformed to arrive at a spectrum of a given type. According to Wiener-Khintchine theorem [8], the autocorrelation function of a function of time $f(t)$ is given by:

$$\langle f(\tau)f(t+\tau)\rangle_\tau = \frac{1}{2\pi}\int\left|\int f(t)e^{-i\omega t}dt\right|e^{i\omega t}d\omega \tag{2.7}$$

It follows from (2.7) that the autocorrelation of f(t) can be obtained by first taking the Fourier transform of f(t) (which transforms it into the frequency domain), sequentially computing the square of its modulus and subsequently taking the inverse Fourier transform.

In the present study, we have used two types of autocorrelation functions to compute the vibrational spectra: autocorrelation function of the nuclear velocities (the velocity-velocity autocorrelation function) and the dipole moment autocorrelation function [9] [10].

The velocity-velocity autocorrelation function (VV-ACF) was computed from the data collected from the production (simulation) part of the ADMP trajectory as [8, 9, 10]:

$$\langle \overrightarrow{v}(t)\overrightarrow{v}(0)\rangle = \sum_i \sum_j \int_0^{T_{lag}} v_{i,j}(t') \cdot v_{i,j}(t'+t)dt \tag{2.8}$$

where i ranges from 1 to the total number of atoms, while the index j refers to the three principal Cartesian directions and ranges from 1 to 3. Subsequently, VV-ACF was normalized with respect to the initial value $\langle \overrightarrow{v}(0)\overrightarrow{v}(0)\rangle$. From the normalized VV-ACF, the rovibrational density of states spectra, which are proportional to the kinetic energy spectra were computed by [8]:

$$I_{vv}(\omega) = \lim_{T\to\infty} \int_0^T \frac{\langle \overrightarrow{v}(0)\overrightarrow{v}(0)\rangle}{\langle \overrightarrow{v}(0)\overrightarrow{v}(0)\rangle} e^{-i\omega t} dt \tag{2.9}$$

In an analogous way, also the dipole moment autocorrelation function was computed from the ADMP simulation (production) phase and the infrared absorption cross-sections were computed by a subsequent Fourier transformation, i.e.:

$$I_{\mu\mu}(\omega) \sim \lim_{T\to\infty} \int_0^T \frac{\langle \overrightarrow{\mu}(0)\overrightarrow{\mu}(0)\rangle}{\langle \overrightarrow{\mu}(0)\overrightarrow{\mu}(0)\rangle} e^{-i\omega t} dt \tag{2.10}$$

Since in the case of both types of autocorrelation functions we computed the spectrum by Fourier transforming the time series obtained from simulations of finite length ADMP simulations, we have used the Blackmans window function to account for the fact that $T < \infty$ and cause the integrand to diminish at suitable T values. Blackmans window function has the following form (in discrete notation) [11]:

$$w(n) = 0.42 - 0.5 \cdot \cos\left(\frac{2\pi n}{N-1}\right) + 0.08 \cdot \cos(\frac{4\pi n}{N-1}); \quad 0 \le n \le N-1 \tag{2.11}$$

Fourier transformations for all purposes in the present study were performed with the fast Fourier transform (FFT) algorithm.

3. Results and discussion. The starting geometry of irinotecan molecule from which the ADMP simulations were started (corresponding to the minimum on the AM1 PES) is shown in Fig. 3.1.

As already implied in the methodology section, the atom-centered density matrix propagation scheme is an extended Lagrangian molecular dynamics method in which the electronic structure is accounted for by representing the electronic subsystem with a single-particle density matrix. This, in turn, is propagated simultaneously with the nuclear degrees of freedom (which are treated classically) by introducing the fictitious inertia tensor μ which practically results in an adjustment of the nuclear and electronic time scales.

In this manner, the resulting fictitious dynamics allows controllable oscillations around the Born-Oppenheimer surface. As in the ADMP scheme the self-consistence field (SCF) convergence is not achieved, one has to analyze carefully the errors in order to be certain of the accuracy of the dynamics as well as of its physical meaningfulness. In our present study, we have thoroughly analyzed the error by following the time-evolution of the adiabaticity index, as well as of the idempotency of the density matrix [9] [10]. We have also followed the time-dependence of the total angular momentum throughout the productive part of the simulation. Fig. 3.2 depicts the time-dependence of the adiabaticity index in the production phase of the simulation (subsequent to equilibration) at two working temperatures: 10 K and 150 K.

As can be seen, the values of adiabaticity index indicate the stability of the simulations. This was also confirmed by checking out the idempotency of the density matrix, which was kept within the threshold value of 10^{-12}. The total angular momentum value was conserved to $< 10^{-13}$ $\hbar$ as well.

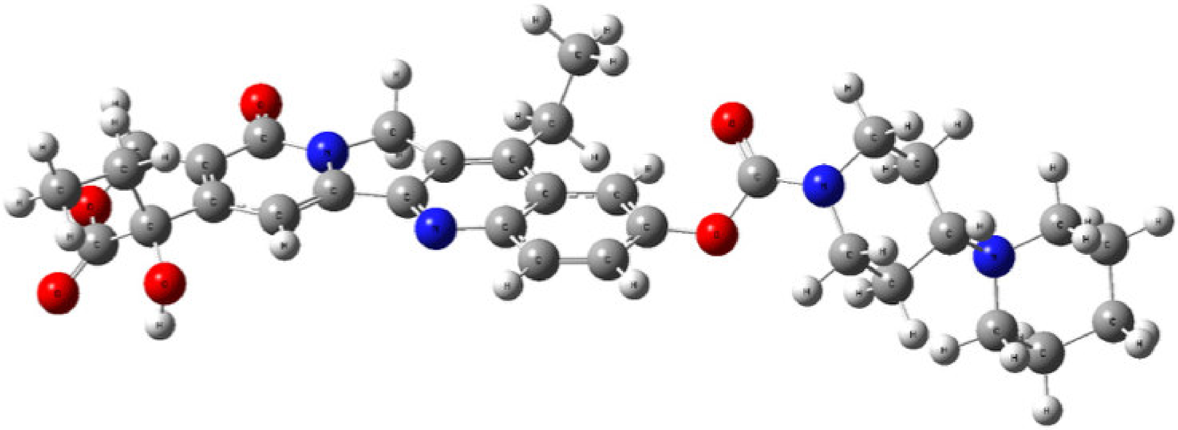


Fig. 3.1. *The minimum located on the AM1 PES of irinotecan molecule (the starting geometry for the ADMP simulations).*

The actually achieved average temperatures during the simulations compared to the target ones are given in Table 3.1. Temperature fluctuations around the target and average values presented in Table 3.1 were acceptable and in line with the statistical physics expectations for a dynamical simulation of molecular system with the current size. As mentioned in the Computational details section, however, temperature control has not been applied in the present study, since the main emphasis here is put on the computation of spectroscopic properties from dynamical simulations through the time correlation functions formalism. To do this properly, one needs to avoid the distortions of the dynamics which would be introduced by the imposed temperature control [9] [10].

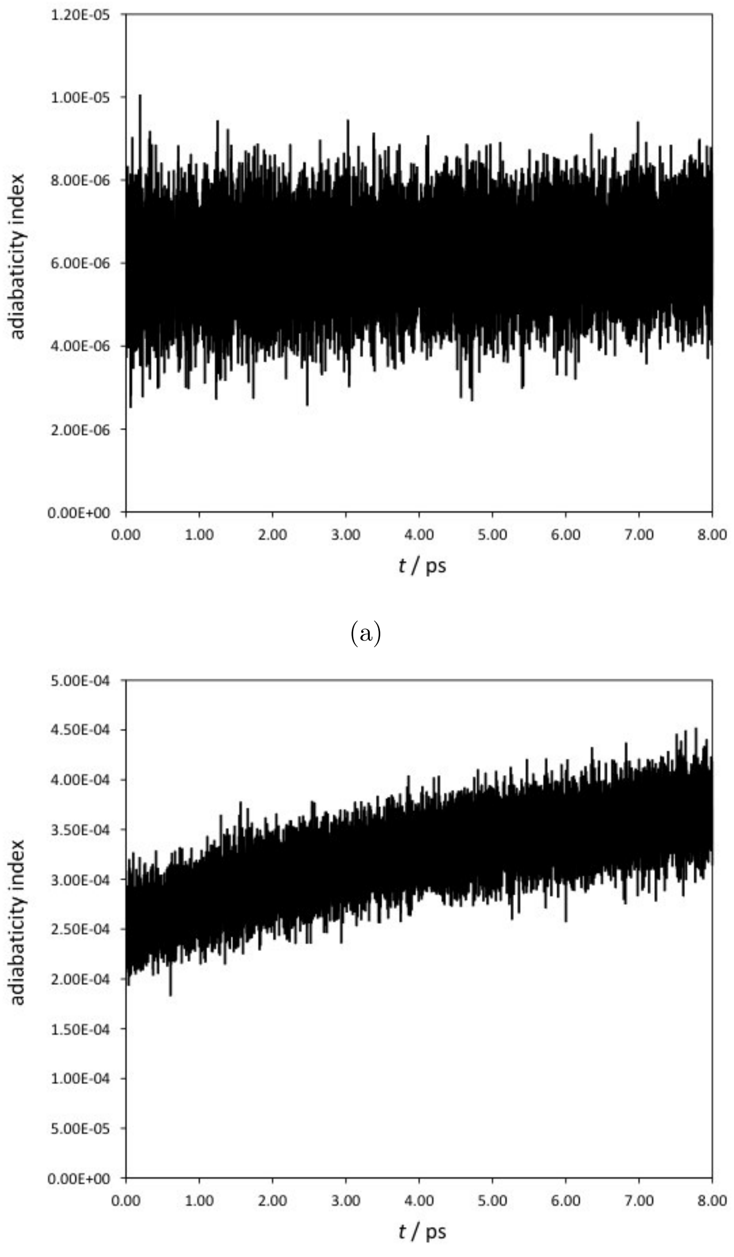
Table 3.1
Target and actual temperatures achieved during the productive part of the ADMP simulation runs.

T_{target}/K	$T_{\mathrm{sim.}}/K$
5	5.0
100	101.9
150	152.7
300	303.5

Fig. 3.3 shows the kinetic energy spectra (i.e. the kinetic energy density of states spectra) obtained by Fourier transformation of the velocity-velocity autocorrelation function for the series of simulations carried out at the four different temperatures (5, 100, 150 and 300 K). In this figure, the usually encountered frequency region in experimental studies (spanning from 500 to 4000 cm^{-1}) is shown. In Figs. 3.4 and 3.5, on the other hand, the lower- (i.e. the fingerprint region) and higher-frequency (C-H as well as O-H and N-H stretching) regions of the kinetic energy spectra are shown.

Though the intensity pattern in the kinetic energy spectra is not directly comparable to the infrared spectrum, but rather to the deep inelastic neutron scattering spectrum, it still contains valuable information concerning the spacings between vibrational energy levels of different intramolecular modes. One can therefore follow the temperature evolution of the energy level differences, i.e. albeit in rather indirect way, the temperature evolution of the molecular conformational flexibility and intramolecular vibrational energy redistribution as well. Having these data for a free molecule of physiologically active substance is of essential importance for further studies and an in-depth understanding of its interaction with nanosized drug carriers.

Throughout the present study, we will use the kinetic energy spectra in parallel with those computed from the dipole autocorrelation function due to the following reasons. These type of spectra, as already implied before, contain information about the molecular rovibrational density of states, or, perhaps even more precisely, about the existence of energy level difference at particular frequency (wavenumber) value. Existence of a frequency difference, however, does not tell us anything about the particular mode that is involved. The intramolecular



(a)

(b)

FIG. 3.2. *Time-dependence of the adiabaticity index in the production phase of the simulation (sub-sequent to equilibration) at two working temperatures: 10 K (a) and 150 K (b).*

nuclear motions corresponding to that particular mode may not involve a change in the dipole moment which would be responsible for absorption of light quanta upon interaction with the incident radiation from infrared spectral region. If one therefore restricts the analysis solely on the basis of spectra obtained from the dipole moment auto-correlation function, the thermally-induced behavior of modes which are not infrared active would be left out.

Fig. 3.6, on the other hand, shows the spectra obtained by Fourier transformation of the dipole moment

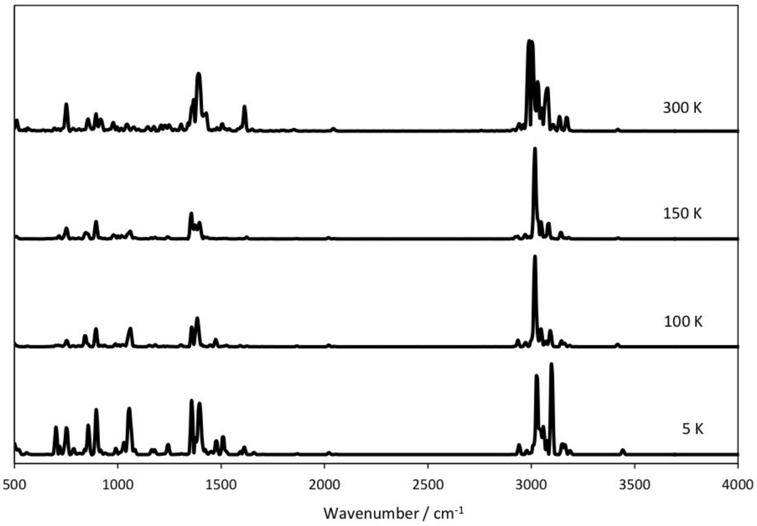


FIG. 3.3. *The kinetic energy spectra (kinetic energy density of states spectra) obtained by Fourier transformation of the velocity-velocity autocorrelation function for the series of simulations carried out at the four different temperatures in the frequency region from 500 to 4000 cm^{-1}.*

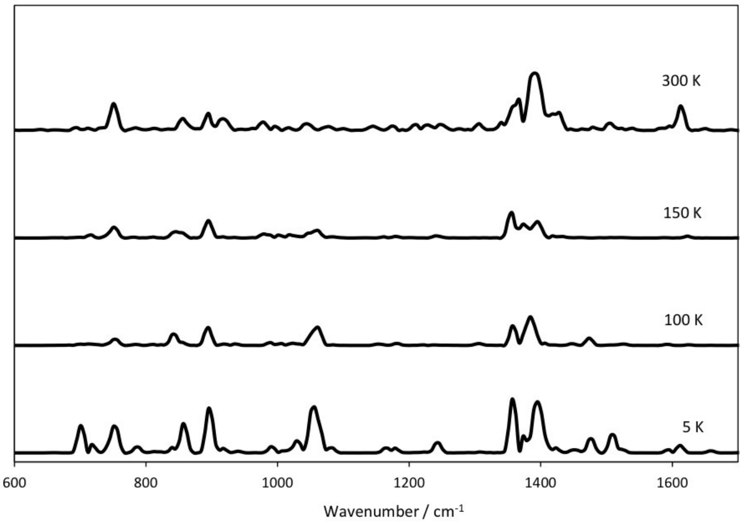


FIG. 3.4. *The kinetic energy spectra (kinetic energy density of states spectra) obtained by Fourier transformation of the velocity-velocity autocorrelation function for the series of simulations carried out at the four different temperatures in the "finger-print" frequency region from 600 to 1700 cm^{-1}.*

vector autocorrelation function for the series of simulations of free irinotecan molecule carried out at the four different temperatures (5, 100, 150 and 300 K). Analogously as in Fig. 3.3, the usually encountered frequency region in experimental studies (spanning from 500 to 4000 cm^{-1}) is shown here as well. These spectra should be directly comparable to the experimentally measured frequency dependencies of the infrared absorption cross-sections obtained by experimental infrared spectroscopic techniques. Figs. 3.7 and 3.8 depict the lower- (i.e. the fingerprint region) and higher-frequency (C-H as well as O-H and N-H stretching) regions of the dipole moment autocorrelation spectra.

The intensity patterns shown in Figs. 3.6-3.8 should, therefore, be directly comparable to the temperature-

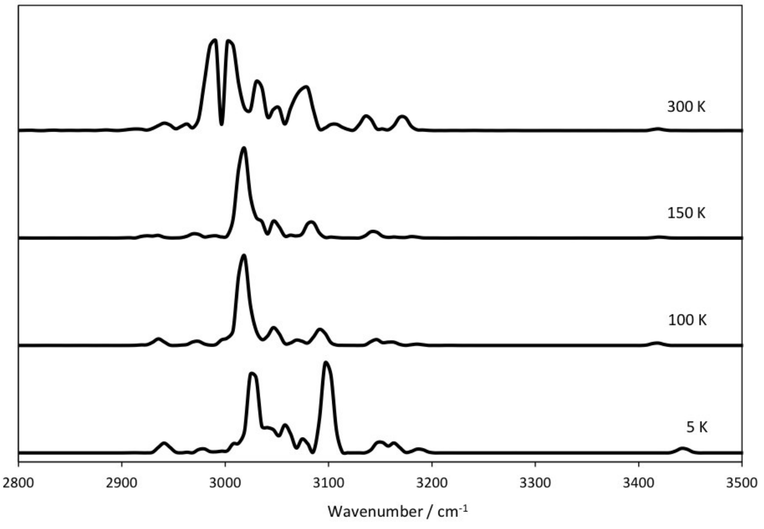


Fig. 3.5. *The kinetic energy spectra (kinetic energy density of states spectra) obtained by Fourier transformation of the velocity-velocity autocorrelation function for the series of simulations carried out at the four different temperatures in the region of C-H and O-H stretching modes.*

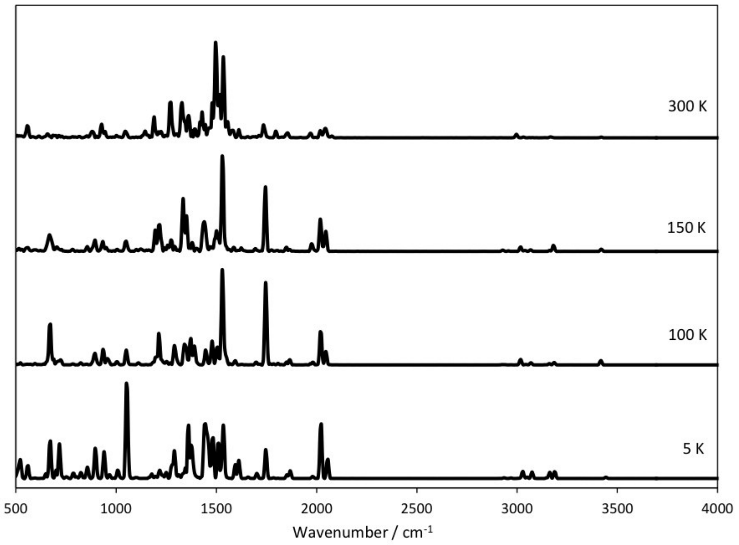


Fig. 3.6. *The dipole moment autocorrelation spectra (∼ infrared absorption spectra) obtained by Fourier transformation of the dipole moment autocorrelation function for the series of simulations carried out at the four different temperatures in the frequency region from 500 to 4000 cm^{-1}.*

dependent infrared spectra of the studied species. Aside from the information related to the vibrational energy level spacings in the case of different intramolecular modes, from these spectra one can also directly follow the temperature evolution of the infrared absorption spectra, as well as the thermally-enhanced molecular conformational flexibility and intramolecular vibrational energy redistribution.

Fig. 3.9, finally shows the harmonic vibrational spectrum of free irinotecan molecule, computed in a "static" manner, i.e. by sequential geometry optimization (location of the minimum on the considered semiempirical PES) and computation and subsequent diagonalization of the mass-weighted Hessian matrix at this particular point on the PES. Note that such analysis, aside for the computation of the IR spectrum within the harmonic

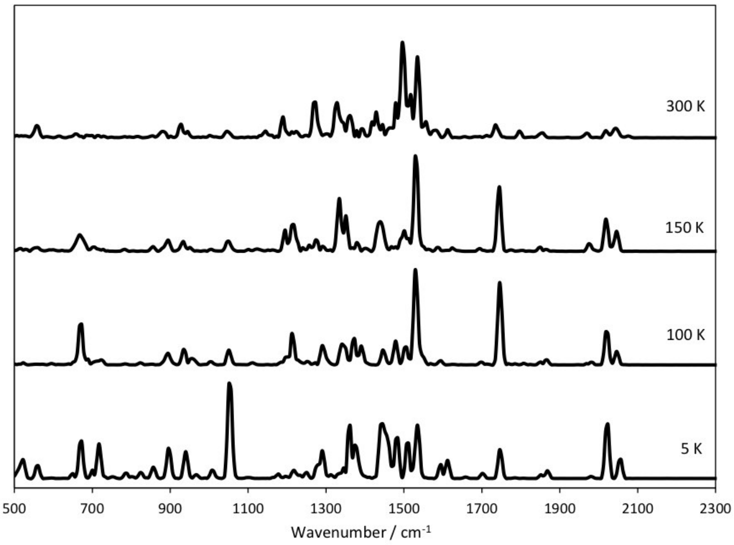


FIG. 3.7. *The dipole moment autocorrelation spectra (∼ infrared absorption spectra) obtained by Fourier transformation of the dipole moment autocorrelation function for the series of simulations carried out at the four different temperatures in the "fingerprint" frequency region from 500 to 2300 cm^{-1}.*

approximation, has also served as a test of the character of the located stationary point(s) on the studied molecular PES. Absence of negative eigenvalues of the second-derivative matrices (i.e. absence of "imaginary frequencies") indicates that a true minimum on the considered PES has been located (instead of, e.g. a saddle-point).

It is worth noting, however, that the harmonic vibrational spectrum depicted in Fig. 3.9 has been computed considering the minimum-energy structure of the studied molecular system at (implicitly assumed) temperature of 0 K. It is, therefore, most directly (although not quantitatively) comparable to the dipole moment autocorrelation spectra calculated at 5 K. The great advantage of the "dynamical spectra" is the fact that they inherently contain the influence of anharmonicity of the molecular vibrational modes, provided that the temperature is sufficiently high so that throughout the molecular dynamics simulation a sufficiently "wide" region of intramolecular motions is sampled (i.e. a sufficiently wide configurational space volume).

Both types of spectra computed from the ADMP simulations in the present study are based on the autocorrelation functions of averaged nuclear velocities or of the dipole moment vector. It is therefore worth recalling at this point that these are statistical quantities, which have been obtained as a statistical average from numerous different configurations spanned by the MD simulation.

The computed resultant spectrum should, therefore, correspond to a dynamically averaged picture of the studied molecular system. This, on the other hand, would result in an overall lowering of the intensity of the peaks (especially in the intermediate spectral region). At the same time, in the course of dynamical simulation, the immediate surroundings of each of the vibrational modes changes in a quite anisotropic manner. This, in consequence, leads to both broadening and flattening of particular spectral regions. Such observations are in line with previous results reported in the literature [9] [10].

The previously outlined theoretical results as well as the theoretical explanations behind such observations suggest that thermally-induced dynamical effects in the rovibrational density of states as well as in the infrared absorption cross-sections of individual molecular systems may lead to notables changes in comparison to the corresponding "static" properties (e.g. those computed for a particular stationary point on the studied PES or from a snapshot from a statistical physics simulation). This is especially important when one compares theoretical with experimental spectroscopic data, considering the fact that the later are often obtained at finite temperatures, much above the absolute 0 (which is the effective temperature at which "static" *ab intio* or semiempirical computations are often carried out).

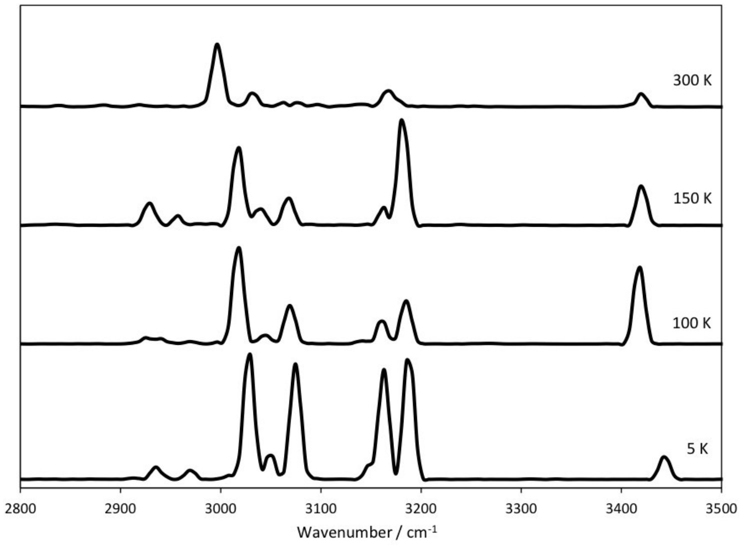


Fig. 3.8. *The dipole moment autocorrelation spectra (∼ infrared absorption spectra) obtained by Fourier transformation of the dipole moment autocorrelation function for the series of simulations carried out at the four different temperatures in the region of C-H and O-H stretching modes.*

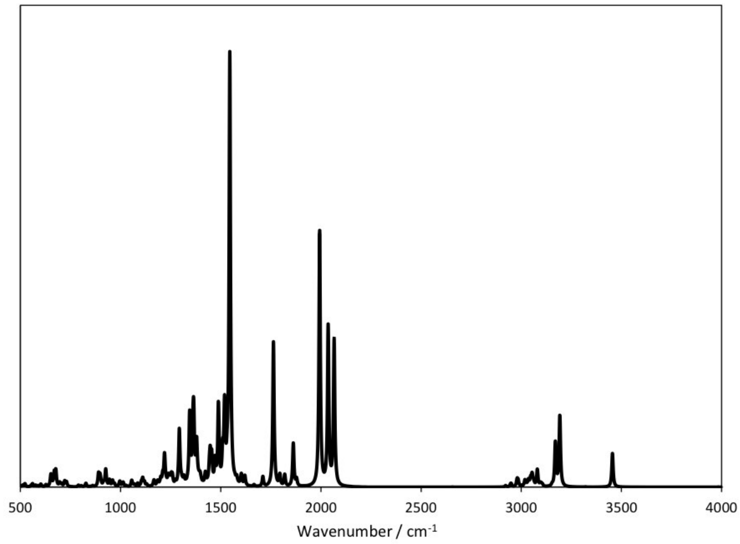


Fig. 3.9. *The "static" harmonic vibrational spectrum computed by diagonalization of the mass-weighted Hessian matrix for the minimum located on the AM1 PES of the free irinotecan molecule.*

In case when several conformers of a complex, large molecular system are close in energy, thermally-induced intramolecular motions and intramolecular vibrational energy redistributions may cause effective dynamical transitions between the corresponding wells on the molecular PES. The computed statistically averaged spectra from statistical physics simulations by the time correlation functions formalism therefore account for such dynamical effects and inherently account for the intramolecular conformational flexibility of the studied molecule. Such aspects are expected to be of essential importance especially when intermolecular interactions of noncovalent type are in questions. Such are, e.g. the interactions between drug molecules (as the presently studied irinotecan molecule) and drug carriers, cellular receptors, enzymes and other systems relevant in the biomolecular context.

4. Summary and conclusions. In the present study, semiempirical molecular dynamics simulations of hydrophilic drug irinotecan were carried out employing the atom-centered density matrix propagation scheme at series of temperatures ranging from 5 K to 300 K. From the computed molecular dynamics trajectories, various types of spectra were computed within the time correlation functions formalism. These included the rovibrational density of states spectra, which were computed from the velocity-velocity autocorrelation functions, as well as the infrared absorption cross section spectra, computed from the dipole moment autocorrelation functions. The thermally induced changes in the single-molecule spectroscopic properties were deduced and the reasons behind them were analyzed and discussed. This work is the basis for the development and implementation of an accurate and plausible statistical physics model for the drug nanocarrier intermolecular interactions and their spectroscopic manifestations.

Acknowledgments. This work was supported by the European Unions Horizon 2020 research and innovation programme, project Virtual Research Environment for Regional Interdisciplinary Collaboration in Southeast Europe and Eastern Mediterranean VI-SEEM [675121].

REFERENCES

[1] M.W. Tibbitt, J.E. Dahlman, and R. Langer. *Emerging frontiers in drug delivery.* Journal of the American Chemical Society, 138(3):704–717, 2016.

[2] M. Simonoska Crcarevska, N. Geskovski, S. Calis, S. Dimchevska, S. Kuzmanovska, G. Petruševski, M. Kajdžanoska, S. Ugarkovic, and K. Goracinova. *Definition of formulation design space, in vitro bioactivity and in vivo biodistribution for hydrophilic drug loaded plga/peo–ppo–peo nanoparticles using ofat experiments.* European Journal of Pharmaceutical Sciences, 49(1):65–80, 2013.

[3] B. Koteska, A. Mishev, L. Pejov, M.S. Crcarevska, J.T. Ribarska, and M. G. Dodov. *Computational Vibrational Spectroscopy of Hydrophilic Drug Irinotecan.* In Proceedings of the Eighth International Conference on Advances in System Simulation - SIMUL), pages 11–16, 2016.

[4] H.B. Schlegel, J.M. Millam, S.S. Iyengar, G.A. Voth, A.D. Daniels, G.E. Scuseria, and M.J. Frisch. *Ab initio molecular dynamics: Propagating the density matrix with gaussian orbitals.* The Journal of Chemical Physics, 114(22):9758–9763, 2001.

[5] M.J.S. Dewar and W. Thiel. *Ground states of molecules. 38. the mndo method. approximations and parameters.* Journal of the American Chemical Society, 99(15):4899–4907, 1977.

[6] H.B. Schlegel. *Optimization of equilibrium geometries and transition structures.* Journal of Computational Chemistry, 3(2):214–218, 1982.

[7] M.J.E.A. Frisch, G.W. Trucks, Hs.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G.A. Petersson, et al. Gaussian 09, revision d. 01, 2009.

[8] M. Thomas, M. Brehm, R. Fligg, P. Vöhringer, and B. Kirchner. *Computing vibrational spectra from ab initio molecular dynamics.* Physical Chemistry Chemical Physics, 15(18):6608–6622, 2013.

[9] S.S. Iyengar. *Dynamical effects on vibrational and electronic spectra of hydroperoxyl radical water clusters.* The Journal of chemical physics, 123(8):084310, 2005.

[10] S.S. Iyengar, M.K. Petersen, T.J.F. Day, C.J. Burnham, V.E. Teige, and G.A. Voth. *The properties of ion-water clusters. i. the protonated 21-water cluster.* The Journal of chemical physics, 123(8):084309, 2005.

[11] F.J Harris. *On the use of windows for harmonic analysis with the discrete fourier transform.* Proceedings of the IEEE, 66(1):51–83, 1978.

Edited by: Aneta Karaivanova
Received: Dec 21, 2017
Accepted: Mar 12, 2018

DOI 10.12694/scpe.v19i2.1348
Scalable Computing: Practice and Experience ISSN 1895-1767
Volume 19, Number 2, pp. 161–174. http://www.scpe.org © 2018 SCPE

ENABLING VIRTUAL COLLABORATION IN DIGITAL CULTURAL HERITAGE IN THE SEEM REGION

PANAYIOTIS CHARALAMBOUS[*] AND GEORGE ARTOPOULOS[†]

Abstract. It has been observed that many researchers in the humanities do not use digital tools to their full extent for their research. Some of the most pressing needs of researchers in Digital Cultural Heritage (DCH) are data storage and handling and large scale computing. Linking these researchers to experienced groups will significantly improve productivity and research innovation in DCH. This work presents our efforts in enabling virtual collaboration for research in the South East and Eastern Mediterranean region and more specifically the deployment of the Clowder CMS system and the development of extraction services to handle, manage and automatically process DCH data. We give technical descriptions of the system and provide some results and discussions of our efforts to enable virtual collaboration between regional level DCH researchers in the context of the Horizon 2020 funded VI-SEEM project.

Key words: Content Management System; Digital Cultural Heritage; Data Processing; Online Visualization

AMS subject classifications. 68M14

1. Introduction. Cultural heritage is a key factor of European identity. Heritage is anything that helps us collectively to better understand the present and think of our future. The Southeast Europe and Eastern Mediterranean region (SEEM) is renowned for its ancient civilizations. It is also an area of major socioeconomic and cultural developments during the medieval and early modern periods. The rich heritage in the region is in risk due to climate and human factors, such as war and conflicts, and the VI-SEEM project[1] is dedicated to its preservation, which includes activities such as heritage documentation, artefacts' analysis, conservation and preservation that spans from the conservation practices of objects to the scale of whole archaeological sites; to facilitate wider dissemination and provide access to knowledge for everyone, as well as to preserve artefacts in digital form by means of virtual reconstructions.

Learning about the history of a place is a good way of bringing communities together through a shared understanding of the unique cultural identity that heritage places give to an area. In our contemporary societies there is a pressing need to deal with issues of intercultural dialogue, social identity, and collective memory more than ever.

The most current need of researchers and scholars operating in the field of Digital Cultural Heritage (DCH) is how to produce quality from quantity, how to devise critical methodologies that produce meaning and generate knowledge out of big data, and VI-SEEM contributes with computation methods to help in achieving that. The process of interpretation is cross-disciplinary in nature and involves various faculties of human activity that rely on data processing, such as logical reasoning, associative analysis, descriptive capacity, linguistics and semiotic processes, decoding, and therefore cognition, abstraction and visualization, in order to reveal patterns and narratives, address the whole and provoke affection. After more than a decade of large-scale digitization processes spurring from most museums and libraries and archives, the next big challenge that all CH stakeholders are facing is to make sense, to add value and establish methods of interpretation and are common, comprehensive, sharable and easily applicable to the vast archives of data and complexity of digital assets in big data.

Throughout Europe, heritage sites, artefacts, texts, works of art are being electronically documented and subsequently archived. This is an on-going process due to the enormous number of artefacts and the continuous growth and development of digitizing technologies. However, the available datasets and related repositories remain fragmented, of varied quality, while access to data is still widely limited. One major effort to unite all data is the European Commission's effort for a digital library for European culture under the name of Europeana [16].

[*] Associate Research Scientist, CaSToRC, The Cyprus Institute, Cyprus. (ps.charalambous@cyi.ac.cy).
[†] Assistant Professor, STARC, The Cyprus Institute, Cyprus. (g.artopoulos@cyi.ac.cy).

[1] https://vi-seem.eu/

Challenges of using computational tools for heritage studies. The majority of researchers in the humanities do not use digital tools to their full extent for their research. Large potential is identified for research groups that have not used large scale computing before. Linking these to experienced groups will significantly improve productivity and research innovation in DCH. Data storage and handling is one of the most pressing and challenging needs of the Cultural Heritage community. The VI-SEEM project focuses on the provision of tools and resources for regional scientists to cast their data into Content Management Systems (CMS), hence offering the stepping-stone to join larger initiatives on the longer run.

Arguably the field of digital cultural heritage has still to undergo the computation paradigm shift that characterises other fields of human activity and the sciences, such as biology, climate, geography, physics and many more. Today we should be moving into a new era of computation in DCH that goes beyond digitisation of artefacts and into the interpretation of data. However the reality is different; medium to small-scale cultural operators and museums in the region do not have the knowledge, resources and capacity to digitise their huge collections. VI-SEEM offers training for those users to facilitate them on how to digitise their assets and then curate them in order for the produced big datasets of digital assets to be *accessible*, *findable (searchable)* and *interoperable* (for ingestion in larger repositories and databases), following the FAIR policies as defined in the H2020 roadmap. So VI-SEEM both enables research in the region and facilitates the integration of locally generated results into larger initiatives at the European level.

The latest developments in interacting with big data scientific visualizations rely on intensive data mining that necessitate a shift of computational tools from the traditional off-line computer cluster to High Performance Computing capable of real-time parallel processing of multiple inputs. State of the art facilities invest in bringing together humanities and science, creative industries, art and engineering, in order to study and disseminate, and ultimately contribute to the preservation of tangible and intangible heritage (cf. AlloSphere [2] and Media Lab Helsinki [24]). Additionally, science has benefited greatly of advanced visualization methods of complex systems - a process that relies heavily on HPC.

We start by giving an overview of related work (cf. Sect. 2), we give an emphasis on the needs of the VI-SEEM communities and how our choice of a CMS (Clowder) helps these communities (cf. Sect. 3), we continue with a description of our system (cf. Sect. 4), some example applications of VI-SEEM and its impact (cf. Sects 5 and 6) and end with some discussion and future directions (cf. Sect. 7).

2. Related Work. There are various definitions of these software environments and platforms for collaboration, which, according to their context, are described as Virtual Research Environments [35, 9], Science Gateways [52], or Digital Libraries [6]. In their more general form, they comprise digital infrastructure and services which enable research to take place [17]. These platforms respond to the aims of e-Infrastructures [27] and cyberinfrastructures [15]. They respond to the needs of the collaborating communities in various ways combining multiple approaches, features, services and protocols, including portals, repositories to content management systems, such as for example Clowder [13] (formerly named Medici [39]), which is used in this work. The latter case offers a wide variety of services and tools that are integrated in a comprehensive way for users to exploit the resources available and facilitate the access of data.

Literature supports that on average most researchers of the community are willing to share knowledge and information about their inquiries, as long as they are provided with a streamlined and intuitive experience [14]. It also highlights the differences between disciplines in the way scholars utilise the VRE or take advantage of its capacities, i.e., the epistemological approach of each field impacts the way and content of data / information is shared, archived, analysed and presented, etc. [7, 8]. In doing so, some researchers driven by the culture of the discipline might share code and/or data, whereas others might only use a VRE for training purposes. Others might prioritise security and access control to data (e.g., copyrighted cultural heritage assets of private collections in Museums), while for some scholars, ease of access to information and the learning curve of operating the environment is of high value. It is widely recognised though that different roles and occupation of researchers necessitates different needs from a VRE, and therefore varied features and tools. Additionally, flexibility in, and control of, the level of security and user access, as well as the duration and familiarity of interacting with the VRE all play important roles to the success and adoption of the platform by communities of users.

Therefore, the major challenges that VREs need to overcome in order to penetrate a variety of fields and become sustainable and inclusive of research communities can be attributed to *accessibility*, *wealth of information* (richness of data), *cybersecurity*, *findability* and *interoperability.* The literature also points to the complexity resulting from integrating big data resources and the difficulties that arise when combining data from various sources / archives, an issue that highlights the importance of taking measures for providing interoperability of data and the associated metadata [23, 29, 33, 34]. This measure is facilitated greatly by generating semantic structures of metadata that enable interaction with and query of the digital assets of each repository integrated in the VRE [37, 40, 42].

This paper presents how in the context of VREs the adaptation of a flexible CMS, such as Clowder, and the further customisation of its features to the needs of the regional communities can benefit research in the SEEM area, and promote collaboration for the preservation of the invaluable heritage of the region. Shared access and collaborative interaction with vast amounts of data is ever more important across a wide range of disciplines who seek for creative interdisciplinary discourse and investigating cross-disciplinary inquiries. Therefore providing customised access and control of large sets of data in a meaningful way; i.e., addressing the particularities of each discipline involved, is of paramount importance for the further development of Digital Heritage. This dynamic interface between data of knowledge and multiple users requires extensive processing power. Enabling archaeologists and historians, social and political scientists, engineers and natural scientists to access the same set of data in order to collaborate for the hands-on investigation of the links between nature (e.g., natural systems - weather and geo-physics) and culture (human artefacts - tangible and intangible) is one of the major challenges of society's computational futures.

3. Needs of the DCH community of SEEM and Clowder. VI-SEEM aims at strengthening links among key players in the field bringing users currently working autonomously together. Large potential is identified for research groups that have not used large scale computing before. There is a great potential in linking these groups with major activities in Europe, and thus offer access to the immense CH data in the region to pan-European initiatives.

In this context, Clowder responds to the needs of the DCH communities in the region and aims to provide the stepping-stone to join larger initiatives on the longer run. In particular Clowder is a content management system designed to support any data format and multiple research domains. It enables users to access and operate HPC infrastructure and provides a data management system for the following services and activities:

- data and associated metadata curation with user controlled access, file versioning, user authentication and assignment of Personal Identifiers (PIDs) to digital assets;
- online 3D visualization;
- curation and online access to geolocated data;
- cloud storage space; and,
- HPC processing services.

Some of the currently provided services and features on Clowder include:

- creation of digital repositories;
- management and safe access to data;
- searching and metadata integration;
- trained convolutional neural networks;
- optical character recognition for scanned documents (currently English);
- data of material analysis for conservation purposes;
- mapping metadata of archives and repositories for the creation of Digital Libraries; and
- tools for the creation of virtual museums using the Unity Game Engine[2].

Expected impact of using Clowder in VI-SEEM. As digitization of cultural heritage artefacts progresses by the museums of Europe and access to their digital archives is provided to an ever growing number of people from all around the globe, operating Digital Libraries and facilitating data-mining technologies for large repositories requires excessive amounts of computing power that only a HPC can offer. Furthermore, Grid-based solutions to inter-connect various library systems should be designed in order to allow end-users to search for content from a unique portal. The impact of the VI-SEEM VRE services and specifically Clowder features presented in this article is envisioned to benefit the following research inquiries:

[2] `https://unity3d.com/`

- *Digital libraries and interactive visualization of Cultural heritage.* Cultural heritage methodologies deal to a large extend with storage and analysis of artefacts and past knowledge. Applications include the management of large collections of scanned books and documents (like these of the Banatica Virtual Library application, cf. Sect. 5), as well as of dynamic file formats such as Reflectance Transformation Imaging (RTI) (e.g., the data from the Centre for the Study of Ancient Documents at Oxford University[50]). Providing to the user communities in the region access to these collections offers great opportunities for breakthrough contributions to art, historical and archaeological inquiries.
- *Image classification, feature extraction and machine learning techniques for image and video analysis.* Exploiting the computational capacity of HPC greatly benefits these methods as they require the analysis of large datasets. Additionally, artefact and built heritage structures' reconstruction (3D modelling) by means of photogrammetric techniques, such as structure-from-motion[19, 28], which rely heavily on image matching and feature extraction, benefit greatly from HPC infrastructure. VI-SEEM has been actively enabling and supporting these research activities which contribute to the presentation of sensitive or threatened heritage in the region (see online 3D Database System for Endangered architectural and archaeological Heritage in the south Eastern MEediterRAnea Area (EpHEMERA) [1]) due to the active engagement of various regional research groups[3]. Remote sensing image analysis is used for land cover and land use classification, built-up and clear land area detection, monitoring of urban growth, monitoring of natural disasters, etc. This is essential for assessing risks and policy making in the area of environmental and heritage protection. Communities in the region use feature learning for image classification in remote sensing, and geophysical analysis of earth subsurface. Electrical Resistivity Tomography (ERT) comprises one of the most important modern techniques of near surface applied geophysics; HPC infrastructure enables for accurate automated resistivity modelling and inversion schemes.
- *Immersive and interactive visualisation of archaeological sites, artefacts and virtual visits to museums*, such as the VirMuf application that is purused by the Biblioteca Alexandrina, Egypt [51]. This is a dynamically growing area of research that exploits the capacity of grid- and cloud- computing for real-time rendering of sophisticated visual representations of inaccessible, sensitive, remote or destroyed objects and structures. Interaction opportunities are offered to the visitor of the virtual space for research and educational purposes [3]. Advanced human computer interfaces are developed to enable users to better interact with information and knowledge. Also spatially distributed narratives, storytelling, and elaborate playful ways (e.g., serious games) are developed in order to sustain longer user engagement in the virtual space [4].

DCH Data in the VI-SEEM project. Data, in the case of VI-SEEM and more specifically in the field of Digital Cultural Heritage, can be of very diverse types. More specifically users can upload entire datasets or individual files of:

- Scanned books and their metadata
- 3D Models
- Image, video, text and sound files and their metadata, organised in collections.
- Advanced documentation data, such as Reflectance Transformation Imaging, and analysis of material properties of structures, works of art and artefacts.
- Code and workflows with sample files to share computational tools and methods (e.g., trained convolutional neural networks, photogrammetric techniques, interactive real time rendering environments for virtual museums and more).
- Semantic referencing of metadata.

A total of 17 applications of user communities are currently serviced in the region, and they have been offered 30,000 CPU-hours and 3,030,000 GPU-hours at 2 HPC sites, and 51 Cloud VMs at 7 Cloud sites in total, reaching approx. 18 TB of storage space - consumed by all applications.

[3] e.g., Foundation for Research and Technology Hellas, Science and Technology in Archaeology Research Centera and The Cyprus Institute

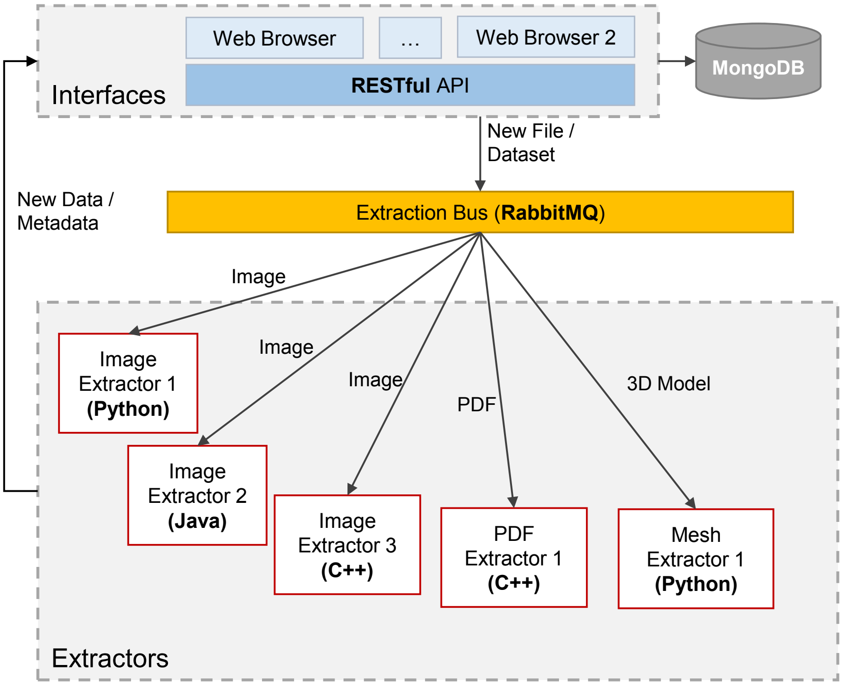


Fig. 4.1. **Clowder Architecture**. *Users of the Clowder CMS can upload data using either a web interface or the RESTful API. Depending on the type (dataset, single file) and filetype of the data, the data are forwarded to* **extraction services** *which process the data to generate both new data and metadata that are associated with the source data. These new data are then processed similarly to the input data.*

4. Methodology/Framework. We built our Digital Cultural Heritage management system on top of Clowder, which is a Web 2.0-based general multimedia content management system capable of semantic content management and service/cloud-based workflows [13]. It supports a broad range of research techniques and allows for community data management. Clowder provides scalable storage and media processing, simple straightforward user interfaces, search, social annotation capabilities, user management, preprocessing and previewing/visualization of various types data and metadata extension and manipulation. All of these features allows for the secure searchable access to large amounts of DCH data satisfying the needs of the DCH communities. At the core of Clowder are extraction services that allow for the *preprocessing*, *processing* and generation of *previews* for the data; these services are developed through collaboration of the various participating organizations.

4.1. Functionality. When users add new data to the system, whether this is through web front-end, or through the RESTful API, preprocessing is off-loaded to *extraction services* (cf. Fig. 4.1). These extraction services attempt to both extract metadata and generate new data based on the type of the data, e.g., to create image previews for videos or 3D files. These metadata are then associated with the uploaded data and presented to the user in the Clowder web interface. Newly generated data are uploaded back on the platform resulting in the call of different extraction services and so forth. We note that users can manually add and define other metadata at a later stage by using the web interface (or the RESTful API). Metadata can be defined at both the file and dataset level.

Users can upload and manage datasets in a variety of formats such as 3D, RTI, images, videos, text and audio (cf. Sect. 3); more formats can easily be integrated. Previews of large datasets in a variety of formats are also extracted and viewed to avoid the need of downloading the whole content on the user's system or finding the needed software to examine the contents of a file.

Clowder's scalability/parallelization, flexibility, and robustness, as well as its overall performance, are improved by decoupling the extraction services from the main server; i.e., multiple instances of the same extractor can run on different machines in a distributed manner. We note that extraction services can be developed in a variety of programming languages and systems as long as they use the RabbitMQ message broker for communication with the Clowder instance (cf. Sect. 4.2.3). Currently, most of the extractors are written in Python and Java.

In the following paragraphs we give a brief overview of the technologies that are at the core of Clowder (cf. Sect. 4.2), how data and access control are managed (cf. Sects 4.3 and 4.4),and finally we give a description of the currently deployed extraction services (cf. Sect. 4.5).

4.2. Supporting technologies for Clowder. Clowder relies on various technologies to get the required flexibility to handle heterogeneous DCH data and metadata (cf. Fig. 4.1). These include the web server written using the Play framework (cf. Sect. 4.2.1), the MongoDB Database Management System (DBMS) (cf. Sect. 4.2.2) and the RabbitMQ Message Broker (cf. Sect. 4.2.3).

4.2.1. Web Server. The web server is built using the Play web application framework [4] which supports both Java and Scala [25]. The Play framework provides minimal and predictabe resource consumption which is really important for highly scalable applications. The server uses the model-view-controller (MVC) architectural pattern. It relies on a number of plugins for communication with the RabbitMQ broker and the MongoDB database, and user authentication. It uses dynamic HTML (ver.5) for webpage generation (e.g., views of the data) according to the results of input processing, search, etc. The models are closely associated with collections in the database. Preprocessors and scripts (i.e., previewers) running on users' browsers communicate with the server using a REST api [18].

4.2.2. Data Storage: MongoDB DBMS. The NoSQL MongoDB DBMS system [32] is used for the storage of both data and metadata in a flexible manner. It is a schema-less database [10, 21]; i.e., it does not require a rigid schema for the duration of the lifetime of the database, it does not enforce data type limitations, it can store both structured and unstructured data and administrators do not need to add additional layers on top to abstract the relational model into a more user friendly object oriented format. The choice of a schema-less database allows for more flexibility in handling the heterogeneous nature of DCH data and easier expansion such as community-generated metadata and new data formats.

4.2.3. Communication: RabbitMQ Message Broker. The role of the RabbitMQ message broker [36] is to take preprocessing messages from the web server that are sent once a dataset or file is uploaded and distribute them to the extractors that can then handle the jobs (cf. Fig. 4.1). The role of a message broker is to mediate the communication between applications [20, 26, 38]; this is done by validating, transforming and routing messages. RabitMQ implements the Advanced Message Queuing Protocol (AMQP) [22]. In the case of Clowder, any extractor that is implemented must register one or more delivery queues on RabbitMQ the moment it is activated. Each queue is associated with a particular routing key set, which defines which routing keys a job can have in order for it to be routed to that queue. The extractor then continuously listens to the queue and acts accordingly.

4.3. Data Organization. Users can organize their data using a plethora of approaches; *datasets*, *collections* and *spaces* (cf. Fig. 4.2):

- Datasets contain related data; e.g., scanned books from a specific era or 3D models of objects scanned from an architectural side. Users can add metadata, tags or even comment on the data on a per file or dataset level.
- Collections are sets of related datasets such as sets of scanned books and audio transcripts of these books.
- Spaces can contain many datasets and collections and in addition users with different roles that can access and/or modify these data (cf. Sect. 4.4).

4.4. Access Management. Users can select the level of access to the data that they upload. This can be set using a variety of approaches. At a coarse level, data can be set as *public* or *private*; public data can be seen and downloaded by everyone that has access to the system, private data are restricted to selected users.

[4] `https://www.playframework.com/`

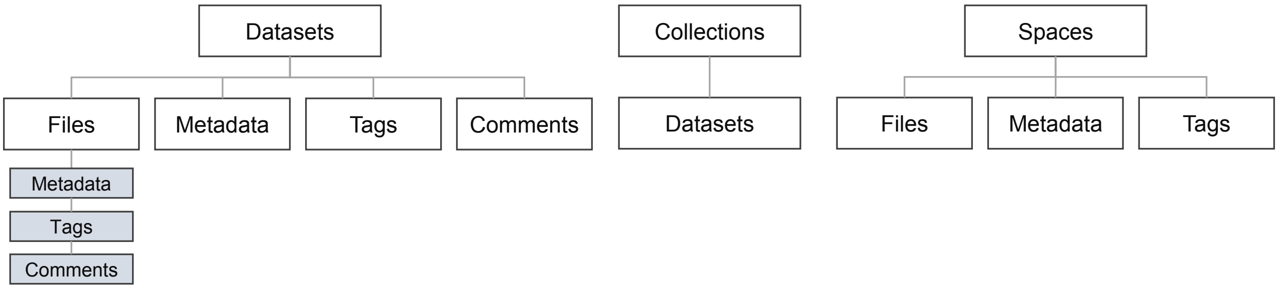


FIG. 4.2. **Data Organization in Clowder.** *Data can be organized in datasets, collections of datasets and spaces associating users and their level of access to the data.*

Additionally, in a finer level of control, data can be associated with *specific users* using spaces (cf. Fig. 4.3); these users can be assigned different roles such as administrators, editors or viewers of the data. The administrator of the system can define new roles and also controls who can register on the platform minimizing in the process misbehaving users (as much as possible).

4.5. The Vi-SEEM Instance of Clowder. We made the decision to create a separate Docker container [5] for each one of the extractors, the Clowder instance, MongoDB and the RabbitMQ broker [6]. Additionally, most of the containers run on a single Virtual Machine (VM); extractors that need more processing are deployed on separate VMs. Having separate containers ensures that software is isolated and that the platform can be migrated with minimal effort and minimal conflicts between dependencies of different software. Additionally, code for the extractors and the docker containers for the project are available to the VI-SEEM community through the project's code repository [47].

Deployed Extractors. Depending on the VI-SEEM application, we develop extractors that do specialized processing on the data. More specifically, we develop(ed) extractors for:

1. extraction and importing of metadata from the Banatica collection of books
2. three-dimensional (3-D) inversion of surface Electrical Resistivity Tomography
3. automatic image georeferencing tomography (ERT) data in order to automatically determine a 3-D resistivity subsurface model using the AutoGR-Toolkit[5]
4. compressed file handling, such as contents, extraction of content and running of other extractors based on contents,
5. Reflectance Transformation Imaging (RTI) previewing and 3D model generation
6. optical character recognition of English documents

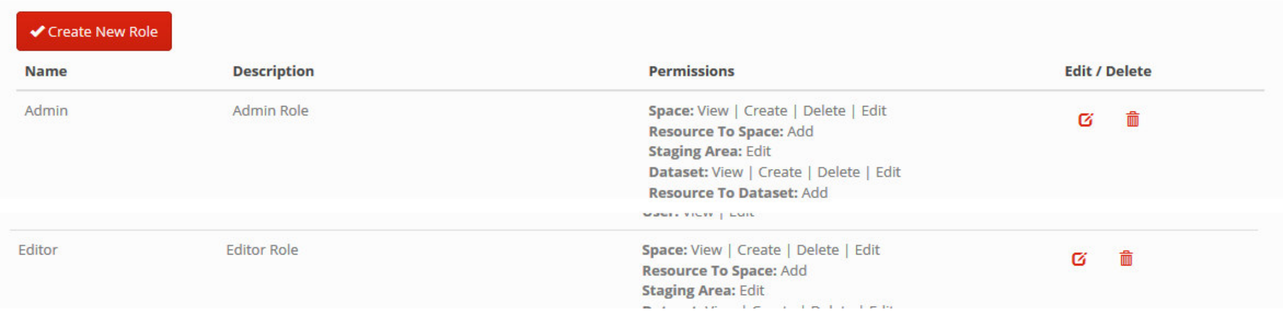
Additionally, we employ several of the readily available extractors, such as generation of previews for images, video preview generation, etc.

5. Examples of novel DCH research activities enabled by VI-SEEM. Some of the most visible research activities supported by Clowder that are also facilitated through VI-SEEM infrastructure are described below.

DataCrowds. Today more people are living in urban environments than in rural areas. It is forecasted that 70% of the global population will be living in cities by 2050. This intense urbanisation poses huge challenges in overcrowding, segregation, demographics and use of resources. The main goal of this project is to innovate in the unified area of research that is occupied with the transdisciplinary study of crowds in built environments. This project envisions a web-accessible, social platform that will allow researchers from very diverse fields, such as Crowd Simulation, Urban Modeling and Simulation, Pedestrian Dynamics, Computer Graphics, Social

[5] `https://www.docker.com/`

[6] A container is a lighweight, stand-alone, executable package that includes whatever a software package needs to run (e.g., code, executables, libraries, tools and settings).



Name	Description	Permissions	Edit / Delete
Admin	Admin Role	**Space:** View | Create | Delete | Edit **Resource To Space:** Add **Staging Area:** Edit **Dataset:** View | Create | Delete | Edit **Resource To Dataset:** Add **User:** View | Edit	
Editor	Editor Role	**Space:** View | Create | Delete | Edit **Resource To Space:** Add **Staging Area:** Edit	

FIG. 4.3. **User Roles.** *The administrator can assign user roles; each can have different access to data and the owners of data can set the role of each user in a space.*

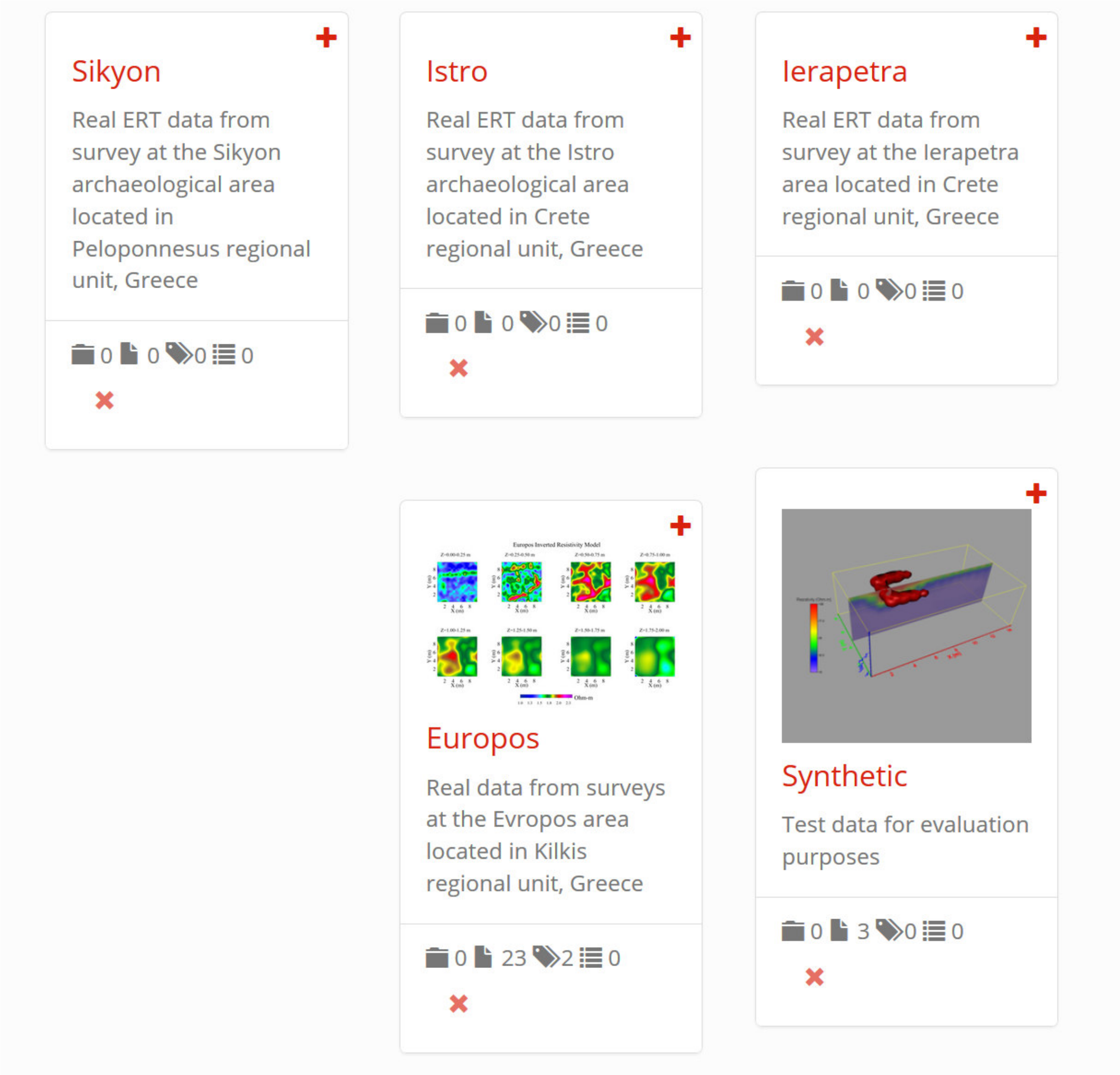
Dynamics and Architecture to collaborate, share data and take advantage of each fields breakthroughs in order to contribute more accurate crowd simulations for the future sustainability of urban environments. As a first step in the implementation of this project, tracked data of crowds from various sources (such as the ones used in [11, 12, 31]) are being uploaded to the Clowder platform of the Vi-SEEM project.

PETRA. The "PETRA: Petra Painting Conservation Project", which is pursued in collaboration with the Synchrotron-Light For Experimental Science And Applications In The Middle East (SESAME)[7] and is developed for the Department of Antiquities of Jordan by the Department of Optics and Atomic Physics at the Technical University Berlin. PETRA provides documentation, condition assessment, and characterization of Nabataean wall paintings and painted marble sculptures from Petra, with a focus on its gilded wall paintings. Characterisation methods include 2D and 3D Micro-XRF, Micro-XANES, handheld XRF, handheld FTIR in addition to various complementary lab-based characterisation techniques. An important aim of this project is to survey painted material in Petra, e.g., collecting historical and recent research material about painted walls and sculpture in Petra, including photos, descriptive documents, analysis data, etc. This activity involves not only in-situ survey of the remaining intact painted walls as well as painted fragments and painted marble sculpture in Petra, but also the study and analysis of the painted material (condition assessment) and objects. In-situ and ex-situ analysis work is taking place in Petra and Berlin. The use of Clowder to store, access, share, link and compare the data, and even visualise them at a later stage, will certainly strengthen this research, time-wise and money-wise.

HaPPen. Another application that is currently under development is "High Performance Photogrammetry (HaPPen)" pursued at the Science and Technology Research Center at the Cyprus Institute. The installation of photogrammetric tools for running structure-from-motion methods for the digital reconstruction of monuments and artefacts from large collections of high-resolution photographs, i.e., of massive datasets of images, acquired from underwater, terrestrial and aerial survey systems, can be better served by HPC infrastructures. This is a computationally intensive process of repetitive image matching and feature extraction operations, and is currently widely used by DCH communities, where budget constraints and requests for high accurate models are ever rising. The project will test and implement a set of commercial and open source software to be used for image based 3D reconstruction processes. It will also conduct an assessment on performance and usability of the available software and methods, and benchmarks will be provided/shared to the DCH communities for further exploitation.

3DInv and AutoGR. The same computational logic (i.e., feature extraction) is exploited by yet another set of significant for the region applications that involves the massive georeferencing of aerial images [43], and the 3D reconstruction of subsurface conditions and structures by large sets of imaging data including satellite images and electrical resistivity tomography[44], respectively (cf. Fig. 5.1). These applications are pursued by different groups of the Foundation for Research and Technology Hellas (FORTH), Institute for Mediterranean Studies (IMS), Laboratory of Geophysical Satellite Remote Sensing and ArchaeoEnvironment (GeoSat ReSeArch Lab). Electrical Resistivity Tomography (ERT) involves the reconstruction of a subsurface resistivity distribution

[7] http://www.sesame.org.jo/sesame/



+
Sikyon
Real ERT data from survey at the Sikyon archaeological area located in Peloponnesus regional unit, Greece
📁 0 📄 0 🏷 0 ☰ 0
✖

+
Istro
Real ERT data from survey at the Istro archaeological area located in Crete regional unit, Greece
📁 0 📄 0 🏷 0 ☰ 0
✖

+
Ierapetra
Real ERT data from survey at the Ierapetra area located in Crete regional unit, Greece
📁 0 📄 0 🏷 0 ☰ 0
✖

+
Europos
Real data from surveys at the Evropos area located in Kilkis regional unit, Greece
📁 0 📄 23 🏷 2 ☰ 0
✖

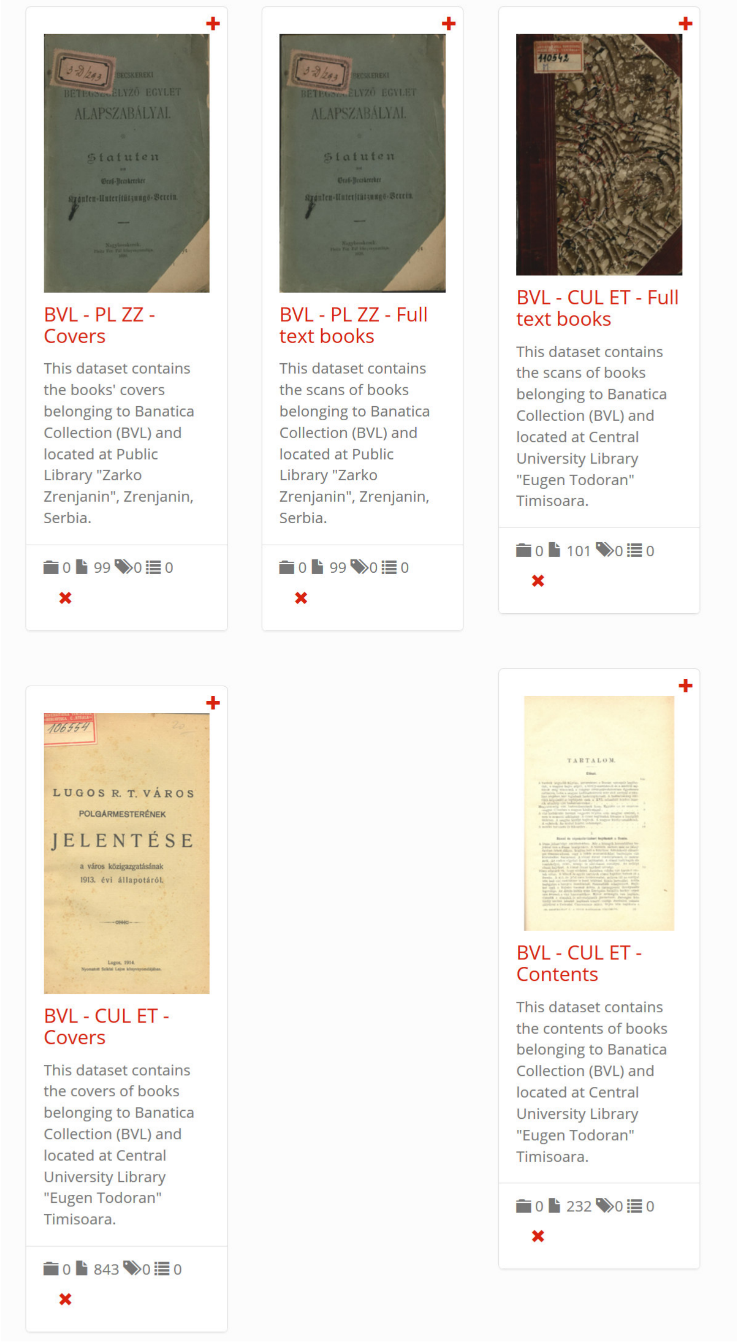
+
Synthetic
Test data for evaluation purposes
📁 0 📄 3 🏷 0 ☰ 0
✖

Fig. 5.1. **3DInv Datasets.** *Datasets of electrical resistivity tomography; a specialized extractor is run and processes these data to generate correspondances between images.*

for revealing finer archaeological details hidden in the original data through the reconstruction of truly 3-D resistivity models of the hidden archaeological relics. The knowledge gained and information acquired by the interpretation of the experimental data is expected to contribute to the update of the relevant policies and the revision of management plans of archaeological sites in Greece.

In all these applications VI-SEEM provides access to HPC infrastructure for running the software but at the same time Clowder offers to the users the opportunity to store, access, share, link, compare and visualise the data. These descriptions showcase only but a few of the wide range of applications that are currently under development and are briefly featured on the VI-SEEM DCH collaboration platform that is enabled by the use of Clowder (`http://dchrepo.vi-seem.eu`).

6. The impact of VI-SEEM to DCH inquiries. An application that already resulted a significant contribution of the VI-SEEM project in the DCH communities in the region is the Banatica Virtual Library (cf. Fig. 6.1), which is developed by the West University of Timisoara in collaboration with the IT Department of



+

BVL - PL ZZ - Covers

This dataset contains the books' covers belonging to Banatica Collection (BVL) and located at Public Library "Zarko Zrenjanin", Zrenjanin, Serbia.

🗀 0 🗎 99 🏷 0 ☰ 0

✖

+

BVL - PL ZZ - Full text books

This dataset contains the scans of books belonging to Banatica Collection (BVL) and located at Public Library "Zarko Zrenjanin", Zrenjanin, Serbia.

🗀 0 🗎 99 🏷 0 ☰ 0

✖

+

BVL - CUL ET - Full text books

This dataset contains the scans of books belonging to Banatica Collection (BVL) and located at Central University Library "Eugen Todoran" Timisoara.

🗀 0 🗎 101 🏷 0 ☰ 0

✖

+

BVL - CUL ET - Covers

This dataset contains the covers of books belonging to Banatica Collection (BVL) and located at Central University Library "Eugen Todoran" Timisoara.

🗀 0 🗎 843 🏷 0 ☰ 0

✖

+

BVL - CUL ET - Contents

This dataset contains the contents of books belonging to Banatica Collection (BVL) and located at Central University Library "Eugen Todoran" Timisoara.

🗀 0 🗎 232 🏷 0 ☰ 0

✖

Fig. 6.1. *The Virtual Banatica Library collection on Clowder.*

the Central University Library ”Eugen Todoran” Timisoara. This application result is significant as it combined the use of HPC for running computationally intensive processes and Clowder for storage, access and sharing of the resulting data [46]. In doing so the Banatica Virtual Library makes rare and old publications accessible again for a wider scientific audience in the SEE region and enables further machine processing to be applied on the entire manuscript collection.

The BANATICA collection gathers together all the printed products considered monographs (e.g., brochures, books, yearbooks, calendars in volumes, prints with an individual cover, atlases, book-like printed scores etc.), which represent documentation sources for the culture and civilization on the Banat region. This collection was jointly created by the VI-SEEM partner, Central University Library "Eugen Todoran" Timisoara (Romania), and "Zarko Zrenjanin" public library (Serbia) throughout the *Biblio-Ident* IPA funded project. The collection comprises over 1000 bibliographic descriptions and 200 full-text scanned books. On Clowder the entire collection was organized into five datasets (cf. Fig. 6.1): two containing the covers of the publications from Banatica collection, two datasets each of each containing 100 full book scans (one for books owned by BCUT and another one for books hosted by Public Library Zarko Zrenjanin), plus an additional dataset of table of contents for the books. During the upload process, metadata was added to the documents to enable searching and association of the data.

In order to make the content machine readable, an environment to run optical character recognition (OCR) on the documents of the collection was setup by the developers. In the first stage of processing, noise was removed from the documents, and the scanned photographs of the pages were sharpened. The second stage involved the OCR process with the use of an open source engine [41]. In parallel the code and scripts behind this OCR pipeline were uploaded on the VI-SEEM's code repository [49]. The developers shared their experience which pointed to the CPU intensity of the process - the initial benchmark for a dataset of 200 digitized prints took 4 days (on a dedicated Virtual Machine). Eventually in order to speed up the process, the operation was replicated on multiple VMs, each getting a subset of PDF documents in a round robin fashion, a master being responsible for distributing the work to multiple workers that run the processing pipeline.

7. Discussion / Future Directions. The overarching goal of the VI-SEEM project is to facilitate cross-fertilisation of research activities between fields and disciplines in order to promote and accelerate interdisciplinary inquiries in the region. The VRE portal of the project, as well as Clowder for the DCH communities, will hopefully contribute cross-thematic activities between the three scientific communities of the project. The services that the VI-SEEM provides to the communities in order to enable interdisciplinary inquiries include data visualization, simulation data, data analytics and processing, geographic description of datasets and curation (e.g., Levante, Balkan regions, etc.), analytical studies and access to source code and the relevant training material. First examples of initialised activities that could result in interdisciplinary research involve the following cases:

- Impact of climate change on the experience of built heritage: visualise the impact of climate anomalies on historic sites and landscapes (under development by The Cyprus Institute);
- Remote sensing and preservation of heritage (developer: FORTH);
- Computer vision and documentation of heritage (developer: FORTH and The Cyprus Institute);
- Machine learning and documentation of heritage, e.g., CNN for satellite images [48] (developer: University of Banja Luka, Faculty of Electrical Engineering);
- Impact of climate on tangible heritage: for conservation purposes (PETRA); and,
- Climate and life sciences for the study of the impact of climate change on evolution (e.g., the "Aharoni" Digitized Collections: Past, present and future of the southern Levant biodiversity by the National Natural History Collections, at the Hebrew University of Jerusalem).

Collaboration of the DCH and Climate communities. The first successful cross-disciplinary research activity between the Digital Cultural Heritage and Climate communities that capitalised on recourses offered by the VI-SEEM project was recently exhibited at a popular international venue. The Seoul International Biennale on Architecture & Urbanism was a large-scale public event organized by the Seoul Metropolitan Government and Seoul Design Foundation and received 4m visitors over the course of its duration. Titled "Imminent Commons", the Biennale provided a forum for debate to policy makers, experts and citizens at large.

Following the UN's World Urbanization Prospect Report of 2014, 54% of the world's population now live in metropolitan areas. By 2050, this percentage will increase to 86% in advanced countries, and 64% in developing nations. Already now, the MENA (Middle East and North Africa) region, renown for its wealth of cultural heritage, ancient civilizations' monuments and major sociocultural developments during medieval and early modern times, is experiencing a high degree of urbanization. Climate change will have particularly strong manifestations in the lived experience of urban settings (e.g., Lelieveld et al., 2014 [30]), and will pose great challenges to the material integrity as well as use of built heritage in these environments.

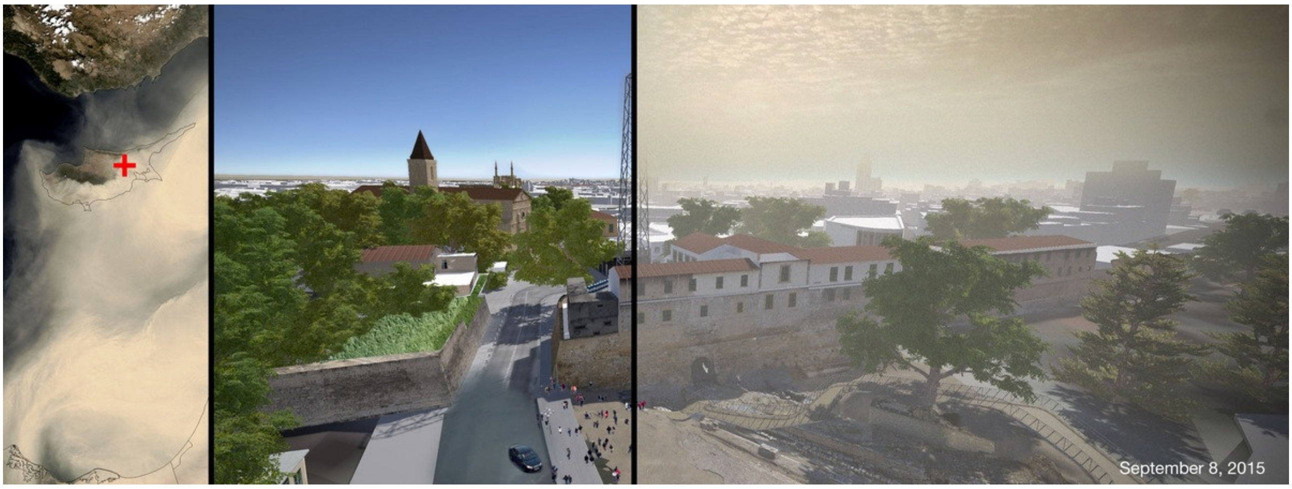


Fig. 7.1. *Visualization of the extreme dust event that took place on September 8, 2015 as seen in virtual reality using Nicosia simulation model. (Credits: Georgios Artopoulos, Theodoros Christoudias, Panayiotis Charalambous, Colter Wehmeier, Charalambos Ioannou, Charis Iacovou, Harry Varnava, Adriana Bruggeman, Panos Hadjinicolaou, Katerina Charalambous, Jonilda Kushta).*

Nicosia, the capital of the Republic of Cyprus and the only major inland city of this Eastern Mediterranean island has been continuously inhabited for over 4500 years. Estimated to become a climate change *'hotspot'* in the foreseeable future, the people of this city already face the effects of the region's changing weather patterns and climate trends. The presented collaborative research activity involved an interactive audiovisual exhibit of immersive simulations that illustrate possible futures of this city, visualising forthcoming conditions of heat, dust and floods using scientific data of climate observations and (computationally) simulated projections (cf. Fig. 7.1). The long-term objective of this activity is to contribute towards the achievement of an integrated climate change adaptation strategy for all of the evolving 'hot spot' cities of the region, and to safeguard the well-being of people living in these locations including both their social structures and the conservation of the built environment.

Finally an application that shows great potential for drawing links with other fields and disciplines within the VI-SEEM project is the “Aharoni” online digitized collection [45], which aims at creating a digital repository for presenting and preserving the greatest Levantine faunal collection from the beginning of the 20^{th} century. The application is pursued by the National Natural History Museum at the Hebrew University [8] and focuses on the promotion and study of collections and archival content of the unique fauna (avian, amphibian, reptiles and mammalian species) of the Levant region, and are the sole direct evidence of that region's biodiversity. The 3D documentation with the support and training of VI-SEEM and online access on Clowder to the generated digital models of the content of these collections of specimens, linked with all relevant metadata will serve as a high-quality database of the southern Levant fauna both for the academic community, as a key biological resource, and for the general public as a repository of knowledge about this unique region.

Most importantly though the developers of this application envision to exploit the capacity of the VI-SEEM VRE for pursuing interdisciplinary collaborative activities. They propose to complement the Clowder repository of specimens with the available analytic platforms in biodiversity science, such as the Open Tree of Life, iDigBio, Lifemapper, Arbor and other complex post-tree analyses (e.g. niche modeling, niche diversification). The aim of this interdisciplinary activity is the use of HPC support for the analysis of the big data of biological studies in order to better understand the complex patterns conniving biodiversity loss in order to promote future land management and wildlife conservation programs.

[8] `https://nnhc.huji.ac.il/`

Acknowledgments. This work was supported by the European Union's Horizon 2020 research and innovation program, project Virtual Research Environment for Regional Interdisciplinary Collaboration in Southeast Europe and Eastern Mediterranean VI-SEEM [675121].

REFERENCES

[1] Abate, D., Avgousti, A., Faka, M., Hermon, S., Bakirtzis, N., and Christofi, P., *An online 3D database system for endangered architectural and archaeological heritage in the South-Eastern Mediterranean*, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLII-2/W3, 1-8, https://doi.org/10.5194/isprs-archives-XLII-2-W3-1-2017, (2017).

[2] `http://www.allosphere.ucsb.edu/`, *AlloSphere Researach Group*, (Page retrieved 12 March 2018).

[3] Artopoulos, G. and Bakirtzis, N., *Virtual Narratives For Complex Urban Realities: historic Nicosia as Museum*, Electronic Media and Visual Arts / Elektronische Medien und Kunst, Kultur, Historie (e-book; open access) (Germany: Heidelberg University), 190–99, (2006).

[4] Artopoulos, G. and Condorcet, E., *House of Affects ' Time, immersion and play in digital design for spatially experienced interactive narrative.*, Digital Creativity Journal, vol. 17,4, 213–20, (2006).

[5] `http://ims.forth.gr/AutoGR`, *AutoGR Toolkit*, (Page retrieved 12 March 2017).

[6] Candela, L., Castelli, D., and Pagano, P. , *History, Evolution and Impact of Digital Libraries*, In I. Iglezakis, T.-E. Synodinou, & S. Kapidakis, E-Publishing and Digital Libraries: Legal and Organizational Issues (pp. 1–30), IGI Global. (2011).

[7] Candela, L., Castelli, D., Pagano, P., and Simi, M., *From Heterogeneous Information Spaces to Virtual Documents.*, Digital Libraries: Implementing Strategies and Sharing Experiences, 8th International Conference on Asian Digital Libraries, ICADL 2005, Bangkok, Thailand, December 12–15, 2005, Proceedings, Springer, (2005).

[8] Candela, L., Castelli, D., and Thanos, C., *Making Digital Library Content Interoperable* Digital Libraries - 6th Italian Research Conference, IRCDL 2010, Padua, Italy, January 28-29, 2010, 13–25, (2010).

[9] Carusi, A., and Reimer, T., *Virtual Research Environment Collaborative Landscape Study*, JISC (2010).

[10] Catell, R., *Scalable SQL and NoSQL data stores*, SIGMOD Rec. 39, 4 (May 2011), 12–27, (2011).

[11] Charalambous, P., and Chrysanthou, Y. , *The PAG Crowd: A Graph Based Approach for Efficient Data-Driven Crowd Simulation.*, In Computer Graphics Forum (Vol. 33, No. 8, pp. 95–108). (2014).

[12] Charalambous, P., Karamouzas, I., Guy, S. J., and Chrysanthou, Y., *A Data-Driven Framework for Visual Crowd Analysis.*, In Computer Graphics Forum (Vol. 33, No. 7, pp. 41–50). (2014).

[13] *CLOWDER: Open Source Data Management for Research.* , (Page retrieved 15 December 2017), `https://clowder.ncsa.illinois.edu`

[14] Connaway, L.S. and Dickey, T.J., *Common Themes Identified in an Analysis of JISC Virtual Research Environment and Digital Repository Projects.*, OCLC Research, (2009).

[15] Cyberinfrastructure Council., *Cyberinfrastructure Vision for the 21st Century Discovery*, National Science Foundation, (2007).

[16] *Europeana Collections*, (Page retrieved 12 March 2018). `https://www.europeana.eu/portal/en`

[17] Fraser, M. *Virtual research environments: overview and activity.*, Ariadne, (44), (2005).

[18] Fielding, R. T., and Taylor, R. N. *Architectural styles and the design of network-based software architectures (p. 151)* , Doctoral dissertation: University of California, Irvine. (2000).

[19] Forsyth, D., and Ponce, J. *Computer vision: a modern approach.*, Upper Saddle River, NJ; London: Prentice Hall. (2011).

[20] Gamma, E,; Helm R., Johnson, R., and Vlissides, J., *Design Patterns: Elements of Reusable Object-Oriented Software.*, Addison-Wesley, (1995).

[21] Han, J., Haihong, E., Le, G., and Du, J. , *Survey on NoSQL database*, In Pervasive computing and applications (ICPCA), 2011 6th international conference on (pp. 363–366), IEEE, (2011).

[22] O'Hara, John, *Towards a Commodity Enterprise Middleware*, Queue, Vol. 5, No. 4, pp. 48–55, (2007).

[23] Heath, T., and Bizer, C., *Linked Data: Evolving the Web into a Global Data Space*, Morgan & Claypool, (2011).

[24] *Media Lab Helsinki*, (Page retrieved 12 March 2018). `https://medialab.aalto.fi/research`

[25] Hilton, P., Bakker, E., and Canedo, F. (Early Access Edition), *Play for Scala*, Shelter Island, NY: Manning, (2012).

[26] Hohpe, G., and Woolf, B., *Enterprise integration patterns: Designing, building, and deploying messaging solutions.*, Addison-Wesley Professional. (2004).

[27] e-Infrastructure Reflection Group, *Blue Paper*, E-IRG, (2010).

[28] Koenderink, J. J., and Van Doorn, A. J., *Affine structure from motion.*, JOSA A, 8(2), 377–385. (1991).

[29] Le Boeuf, P., Do'rr, M., Ore, C.E., and Stead, S., *Definition of the CIDOC Conceptual Reference Model*, Available at `http://www.cidoc-crm.org/official\_release\_cidoc.html`, (2012).

[30] Lelieveld, J., Hadjinicolaou, P., Kostopoulou, E., Giannakopoulos, C., Pozzer, A., Tanarhte, M. and Tyrlis E., *Model projected heat extremes and air pollution in the eastern Mediterranean and Middle East in the twenty-first century*, Reg Environ Change, 14, 1937-'1949. (2014).

[31] Lerner, A., Chrysanthou, Y., and Lischinski, D., *Crowds by example.*, In Computer Graphics Forum (Vol. 26, No. 3, pp. 655-664). Blackwell Publishing Ltd. (2007).

[32] *MongoDB*, (Page retrieved 18 December 2017). `https://www.mongodb.com/`

[33] Paepcke, A., Chang, C. K., Winograd, T., and Garc'a-Molina, H. , *Interoperability for Digital Libraries Worldwide*, Communications of the ACM, 41, 33–42. (1998).

[34] Park, J., and Ram, S., *Information Systems Interoperability: What Lies Beneath?*, ACM Transactions on Information Systems, 22, 595–632. (2004).

[35] Piccoli, G., Ahmad, R., and Ives, B., *Web-based virtual learning environments: A research framework and a preliminary assessment of effectiveness in basic IT skills training.* , MIS quarterly, 401–426. (2001).

[36] *RabbitMQ*, (Page retrieved 18 December 2017). `https://www.rabbitmq.com/`

[37] Ronzino, P., Niccolucci, F., and D'Andrea A., *Built Heritage metadata schemas and the integration of architectural datasets using CIDOC-CRM*, Online proceedings of the International Conference Built Heritage 2013, Monitoring Conservation Management, 18–19 November 2013, Milan. (2013).

[38] Schmidt, D. C., Stal, M., Rohnert, H., and Buschmann, F. , *Pattern-Oriented Software Architecture, Patterns for Concurrent and Networked Objects (Vol. 2).*, John Wiley & Sons. (2013).

[39] Sophocleous, C., Marini, L., Georgiou, R., Elfarargy, M., and McHenry, K., *Medici 2: A scalable content management system for cultural heritage datasets.*, Code4Lib Journal, (2017).

[40] Stephens, R.T., *Utilizing metadata as a knowledge communication tool*, In Professional Communication Conference, PCC 2004, International Proceedings, (2004).

[41] *Tesseract OCR*, (Page retrieved 12 March 2017). `https://github.com/tesseract-ocr/tesseract`

[42] Vassallo, V. and Piccininno, M., *Aggregating Content for Europeana: A Workflow to Support Content Providers*, Lecture Notes in Computer Science, Volume 7489, Theory and Practice of Digital Libraries, Rasmussen and F. Loizides (eds.), Springer. (2012).

[43] *ViSEEM, 3DInv Application*, Georeferencing of Aerial Images, (Page retrieved 12 March 2017). `http://dchrepo.vi-seem.eu/datasets/591184d5e4b03cc97586975d`

[44] *ViSEEM, 3DInv Application*, Electrical Resistivity Tomography Reconstruction, (Page retrieved 12 March 2017). `http://dchrepo.vi-seem.eu/datasets/58e4a6e6e4b06113f7bf4c81`

[45] *ViSEEM, "Aharoni" Collection*, (Page retrieved 12 March 2017). `http://dchrepo.vi-seem.eu/spaces/59ec8ebce4b013b4ad5aa0db`

[46] *ViSEEM, Banatica Virtual Library*, (Page retrieved 12 March 2017). `http://dchrepo.vi-seem.eu/spaces/58f1bf66e4b02fd8f8f22e16`

[47] *ViSEEM Code, Clowder Code Repository*, (Page retrieved 12 March 2017). `https://code.vi-seem.eu/totis77/extractors-cyi`

[48] *ViSEEM Code, Convolutional Neural Networks for Satellite Images*, (Page retrieved 12 March 2017). `https://code.vi-seem.eu/Risojevic/cnn-features.git`

[49] *ViSEEM Code, Distributed OCR*, (Page retrieved 12 March 2017). `https://code.vi-seem.eu/bogconst/viseem-distributed-ocr`

[50] *ViSEEM Project, Digital Cultural Heritage*, (Page retrieved 12 March 2017). `https://vi-seem.eu/cultura-heritage/`

[51] *VirMuf*, (Page retrieved 12 March 2017). `http://dchrepo.vi-seem.eu/spaces/5900b3fde4b02fd8247bab63`

[52] Wilkins-Diehr, N. *Special Issue: Science Gateways - Common Community Interfaces to Grid Resources.* , Concurrency and Computation: Practice and Experience, 19 (6), 743–749, (2007).

Edited by: Aneta Karaivanova
Received: Dec 22, 2017
Accepted: Mar 12, 2018

DOI 10.12694/scpe.v19i2.1349
Scalable Computing: Practice and Experience ISSN 1895-1767
Volume 19, Number 2, pp. 175–180. http://www.scpe.org © 2018 SCPE

VIRMUF: THE VIRTUAL MUSEUM FRAMEWORK

MOHAMMED ELFARARGY AND AMR RIZQ*

Abstract. With the immergence of 3D object digitization technologies, many museums are digitizing their collections using 3D scanning, photogrammetry and other techniques. These large 3D collections are not only great for documentation and preservation, but they are also a great means for introducing these collections to a wider audience worldwide through virtual museums. However, developing Virtual Museums can be a costly process considering that it needs a team of talented software developers, 3D designers and other software/hardware tools. In this paper we present VirMuF (Virtual Museum Framework), which is a set of tools that can be used by non-developers to easily create and publish 3D virtual museums in a very short time. This way, Museum staff doing collection digitization can also publish 3D virtual museums to exhibit these collections. VirMuF is open-source; hence, teams including software developers can further extend VirMuF to fit their needs.

Key words: Digital Cultural Heritage; Data Processing; Online Visualization

AMS subject classifications. 68M14, 00A66

1. Introduction and Previous Work.

A virtual museum can be thought of in many ways. It should provide means to visit a museum that cant be visited in real world, either because it is difficult or because this virtual space does not exist in reality. The virtual museum should also complement the real one by giving users the ability to do thing they cannot do in real museums, such as freely manipulating objects or measuring them. Finally, the virtual museum should act as medium for interactive story-telling so as to provide more interesting ways of introducing history and archaeology to public. A virtual museum should have tools that concern users of a wide spectrum of expertise, ranging from casual visitors to experienced researchers. With these points in mind, a virtual museum pilot project was developed in Bibliotheca Alexandrina (BA) to digitize and exhibit over 700 pieces in the BA antiquities museum. The project aimed at recreating the real space of the museum as it exists in reality, including all showcases and artifacts. The objects were 3D scanned using Artec Eva handheld scanner. Each 3D model went down a pipeline of operations to make it ready for using inside a real time interactive environment. 3DS Max was used for modeling the museum space and Unity 4 game engine [1] was used to program the logic of the application. The project took two years to finish with a team of three software developers and one designer.

The development of a virtual museum is a demanding and costly task. This was clear through the pilot project at BA. This raised the need for a software solution that can simplify this task and make it possible for people with no software development background. This fact triggered the development of VirMuF. VirMuF was written in a modular manner to make it easier to improve or extend any module independent of the rest of the system. This also allows adding new tools and modules easily. VirMuF was written inside Unity 5, which one of the most popular modern graphics engines with a very big community. Unity is multiplatform which means that museums developed with VirMuF can simply be distributed over most operating systems and platforms, including mobile and web platforms.

There is a great diversity in the approaches used to make virtual museums. The ARCO system [1] provides museum curators with software and interface tools to develop web-based virtual museum exhibitions by integrating augmented reality (AR) and 3D computer graphics. While web-based approaches can achieve a good content exposure, they are relatively limited in terms of visual quality and immersion level, compared to what can be achieved using high-end graphics engines and Virtual Reality (VR). Mata et al. [3] introduced an experimental setup that combined navigation facilities with augmented reality. The approach is based on a semantic model of a museum environment that reflects its organization and spatial structure. Augmented reality is a great choice for complementing physical museums; however, it can't be effectively used to create a complete virtual museum used by those unable to visit the real museum. MNEME [4] project used a similar digitization approach to the one described in this paper. However, the system was not built in a way that allows people with no software development knowledge to build their own museums. Petridis et al. [5] used a mini game to attract and educate younger audience, thus expanding their potential user base. In [6], unity

[*] Bibliotheca Alexandrina, Egypt (mohammed.elfarargy@bibalex.org, amr.rizq@bibalex.org).

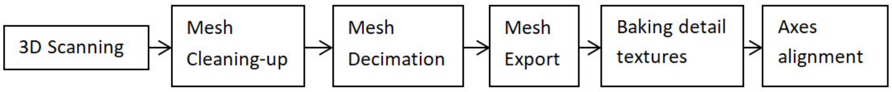


FIG. 2.1. *Artifact digitization pipeline*

game engine was used through a web player to create a virtual recreation of an archeological site. The approach however lacked some of the tools user might need to use through their virtual tour.

2. Digitization Pipeline. The output of artifact digitization operations such as scanning or photogrammetry is often an extremely high-details 3D mesh. While such a mesh is good for documentation purposes, it is not suitable for real-time applications such as virtual museums with hundreds of objects on display. VirMuF uses simplified models that keep the original look and feel of the highly-detailed models through baking multiple texture maps that encode surface properties. In some cases, when using a highly detailed model is necessary to show important surface details, VirMuF uses two versions of the model. A simplified version is used during virtual museum navigation along with the rest of displayed objects. The highly detailed version is used only when inspecting the artifact, hiding the rest of displayed items and focusing only on the inspected artifact, thus saving computational power. As shown in Figure 2.1, results of digitization operations must undergo a sequence of processing steps. Most of these steps are straight forward and can be done using many available software packages, both commercial and free.

It is not uncommon for digitized object meshed to contain some areas that need manual fixing. This is mainly due to sensitivity errors during digitization or some tight areas that scanning devices cannot reach. Artec Studio, or equivalent software, can be used to automatically fix most of these issues. Some parts might need manual. Big holes in the areas that are impossible to scan, such as statue bases glued to their bases, are left with blank colors indicating absence of scanning data.

A typical digitized artifact would initially contain millions of triangles. Models suitable for real-time applications preferably contain 5,000 $\approx$ 10,000 triangles. This means that geometry for high frequency details such as fine surface relief will be lost during the process. These details will be later substituted with normal and displacement maps. For models that still need details geometry, error based decimation is used. An error tolerance not greater than 0.5 millimeter was usually able to keep all the important details while keeping overall geometry under 300,000 triangles. Next, the mesh has to be exported for editing in other 3D software packages. OBJ file format is the best choice here because it is supported by almost all 3D software and because it separates mesh and texture in two files, making it easier to refine texture colors in any image processing software.

Normal mapping and tessellation require normal texture maps and displacement maps, respectively. To produce these textures, XNormal software is used. XNormal uses a highly detailed model and a low-res model for the same object, and by subtraction, the required texture maps are produced. Surface colors obtained through 3D Object digitization will not always be accurate because it is usually affected by factors like surrounding environment lighting and sensitivity error. In order to fix that, a photo editing software is used for gamma correction, white balance adjustments and manual fixing of color value inconsistencies in some areas.

By default, digitized object orientation is random making it difficult to manipulate the object and apply CPU and shader codes. VirMuF assumes that object is oriented so its forward vector points towards positive X axis, right vector points towards positive Y axis and up vector points towards positive Z axis. 3DS max was used to align object axes.

3. System Design.

3.1. Data Storage. VirMuF is based on Unity Game Engine scripting API. Figure 3.1 shows an overview of VirMuF's main components. Because VirMuF is directed towards users with less technical knowledge, Unity's ScribtableObjects were used to store the museum database. This nullified the extra complexity of setting up an external database outside unity development environment. Museum items can either be a stand-alone piece, a piece displayed within a showcase or that belongs to a collection. The MuseumItem class is the base class for all these classes that work as data containers. The data entry for the museum is done directly inside the unity editor, through engine extension, again, to avoid the complexities of using extra tools and backend applications.

3.2. Artifact. Actual virtual 3D game objects that user interacts with are Artifact, MuseumShowcase and MuseumCollection. These objects use MuseumItem objects for data storage and the actual 3D meshes of the displayed artifacts. User should add one Artifact object to the scene for each new item added to the museum. ArtifactBrowser and ArtifactSearcher are responsible for the museum hierarchy browsing and item searching functions, respectively. By default, one object of each will be instantiated in the scene.

3.3. MainCore. MainCore is the main controller class of the application. This singleton is responsible for general tasks such as keeping track of application running time, application termination, and interaction with artifacts. The crosshair class is responsible for managing the different states of the cross hair and using it to inspect artifacts. When user hovers the mouse cursor over an artifact, the crosshair changes indicating the ability to do an in-depth inspection. Clicking an inspect-able object will display a small popup menu with the options to inspect the object, add it to a favorites list or exiting. When Inspection is selected, the ArtifactWorkspace object will take over.

3.4. ArtifactWorkspace. ArtifactWorkspace is responsible for handling the artifact inspection mode. When an object is inspected, the object is isolated from the rest of the museum and is given the main focus. If a higher quality model exists, it's loaded in this mode as well. Regular museum walk-though is replaced with fixed camera, with mouse left and middle clicks on the object used for rotation and panning, respectively. Mouse scroll wheel is used for zooming In/out. Also the regular user interface layout is replaced by a set of buttons to activate various inspection tools, known inside VirMuF as Modules.

3.5. Modules. Modules are Unity GameObjects that perform certain operations on the object under inspection, and work independent from each other. By implementing OnGUI() and Update() methods in Unity's monoBehavior class, each module is responsible for drawing its own GUI window and performing its operations. Also, multiple modules can work simultaneously on the same object. The following subsections describe the available modules.

3.5.1. Information. This tool toggles a window where basic information about the artifact is displayed.

3.5.2. Reconstruction. This module is used to show/hide reconstructed parts for an artifact that is damaged, missing parts or generally deteriorated. Reconstructed parts have to be drawn manually and then added to VirMuF inside unity editor. The module permits adding multiple reconstructions for the same missing part. This is to cover various possibilities proposed by multiple researchers working on the same piece.

3.5.3. Related Items. RelatedItems module connects two artifacts that have some relation due to being found in the same place, belonging to the same era of following the same pattern, for example. This module will display the two related items side-by-side for thorough comparisons. All other modules can be used on either, or both, of the two pieces being compared. Sometimes, zooming in/out the two pieces will lead to user's inability to perceive the right proportions. To solve this, the module contains an option that locks the camera movement for the two viewports used to show the items in comparison in a way that each camera will exactly follow the other one being manipulated by the user. This leads to both cameras being always at the same distance from their targets and exactly at the same viewing angle.

3.5.4. Cross Section. This tool allows user to take cross sections through the artifact in any direction. The cross-section can be taken in any of the three main directions (X,Y, Z) in both positive and negative directions and at any depth. There is also the ability to use an arbitrary cross section plane in both directions.

3.5.5. Related Web Links. In case there is further online information resources related to this artifact web links can be added to each artifact. Clicking on URL button will open the webpage in users default browser (outside the virtual museum application).

3.5.6. Measure. The Measure module (Figure 3.2) allows user to make measurements on the artifact. By clicking anywhere on the artifact a new 3D point is added at the click position. The user can add any number of the 3D points and the tool will display the length between each two consecutive points as well as the overall length of all segments.

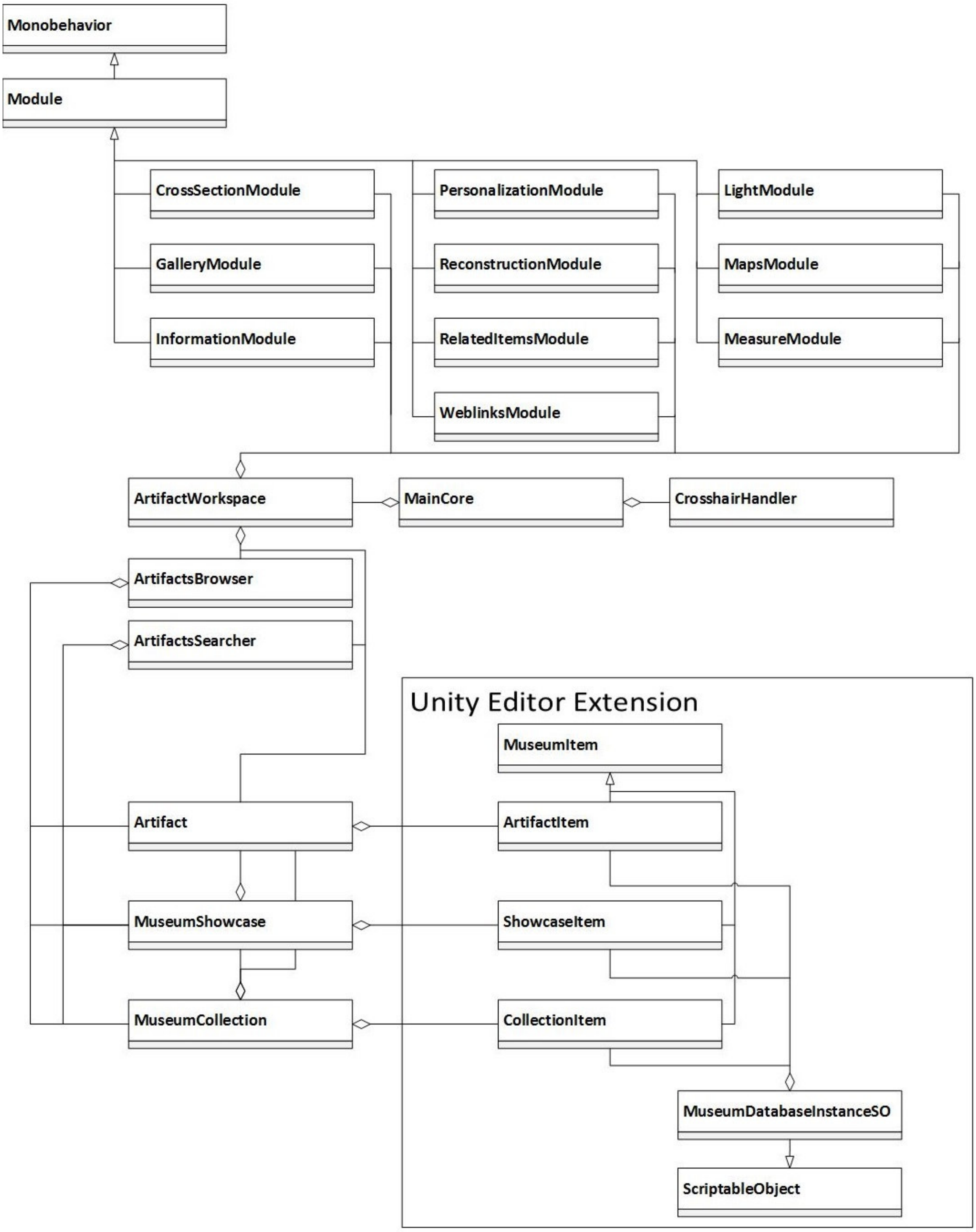
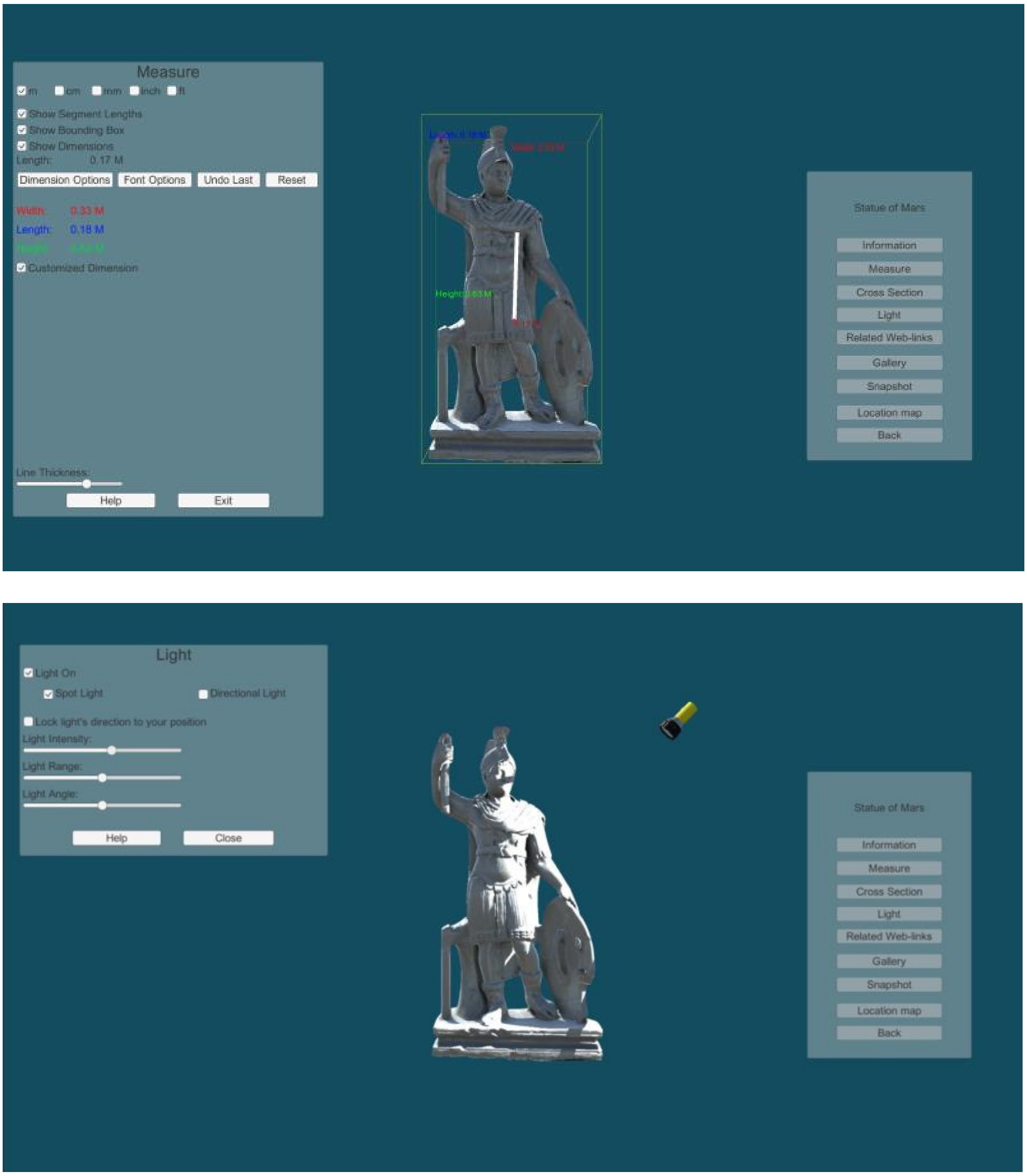


Fig. 3.1. *VirMuF class diagram*



Measure

☑ m □ cm □ mm □ inch □ ft

☑ Show Segment Lengths

☑ Show Bounding Box

☑ Show Dimensions

Length: 0.17 M

Width: 0.33 M

Length: 0.18 M

Height: [illegible]

☑ Customized Dimension

Line Thickness:

Light

☑ Light On

☑ Spot Light □ Directional Light

□ Lock light's direction to your position

Light Intensity:

Light Range:

Light Angle:

FIG. 3.2. *VirMuF modules activated on an artifact: measure module (top) and light module (bottom)*

3.5.7. Light. Light Module allows user to use a virtual torchlight. The light source has so many parameters that can be adjusted to help reveal the fine surface details. Torchlight position is change using right mouse button and it is always pointed towards the artifact. There is also the option to lock light position to viewer camera position.

3.5.8. Gallery. Through the unity editor data entry, each artifact can have a gallery of multiple images and/or videos. When the Gallery module is activated, a window of all gallery images thumbnail is opened.

When user clicks a thumbnail, it gets opened in a separate window in full size.

3.5.9. Location Map. Through Unity editor data-entry, each artifact has longitude and latitude fields that specify where the artifact was discovered and where it is currently exists. When virtual museum visitor activates the Map Location module, Google map will be loaded to show both locations.

3.5.10. Snapshot. This module saves a snapshot of the current view on the local hard drive.

3.5.11. Personalization. Personalization module has many artifact work space customization options for user convenience. This includes options to change background colors, toggle grids and change font size and color.

4. Future Work. More modules are to be added to VirMuF. One experimental module that was tested in the pilot project is the "original context module", where user is taken out of the virtual museum context to a virtual reconstruction of the original place and time where the archeological item was used. Another planned module will display text and multimedia information next to the 3D model in a slide show manner.

Online capabilities are amongst the most important features that would enrich a virtual museum experience. By allowing users to use 3D avatars and interact with each other will turn the virtual museum into a collaborative environment that is optimal for educational purposes. Virtual tour guides can be granted extra privileges and functionalities such as being able to share their view with other users when explaining a particular artifact.

Finally, VirMuF currently supports interaction through mouse and keyboard only. Support for VR input devices such as VIVEcontrollers is to be added for better and more intuitive VR virtual museum experiences using VirMuF.

Acknowledgments. This work was supported by the European Union's Horizon 2020 research and innovation program, project Virtual Research Environment for Regional Interdisciplinary Collaboration in Southeast Europe and Eastern Mediterranean VI-SEEM [675121].

REFERENCES

[1] Abate, D., Avgousti, A., Faka, M., Hermon, S., Bakirtzis, N., and Christofi, P., *An online 3D database system for endangered architectural and archaeological heritage in the South-Eastern Mediterranean*, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLII-2/W3, 1-8, https://doi.org/10.5194/isprs-archives-XLII-2-W3-1-2017, (2017).

[2] *Unity*, https://unity3d.com/

[3] Sylaiou, Stella, et al. *Exploring the relationship between presence and enjoyment in a virtual museum.* International journal of human-computer studies 68.5 (2010): 243-253.

[4] Mata, Felix, Christophe Claramunt, and Alberto Juarez. *An experimental virtual museum based on augmented reality and navigation.* Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2011.

[5] Bruno, Fabio, et al. *From 3D reconstruction to virtual reality: A complete methodology for digital archaeological exhibition.* Journal of Cultural Heritage 11.1 (2010): 42-49.

[6] Petridis, Panagiotis, et al. *The herbert virtual museum.* Journal of Electrical and Computer Engineering 2013 (2013): 16.

[7] Pecchioli, Laura, et al. *Browsing in the Virtual Museum of the Sarcophagi in the Basilica of St. Silvestro at the Catacombs of Priscilla in Rome.* Virtual Systems and Multimedia (VSMM), 2012 18th International Conference on. IEEE, 2012.

Edited by: Aneta Karaivanova
Received: Dec 22, 2017
Accepted: Mar 6, 2018

DOI 10.12694/scpe.v19i2.1350
Scalable Computing: Practice and Experience ISSN 1895-1767
Volume 19, Number 2, pp. 181–187. http://www.scpe.org © 2018 SCPE

SOLUTIONS FOR DATA DISCOVERY SERVICE IN A VIRTUAL RESEARCH ENVIRONMENT [*]

VLADIMIR DIMITROV[†] AND STILIYAN STOYANOV[‡]

Abstract. Scientific computing requires many and large volumes of complex structured data and metadata that are scattered across data centers. Researchers find it difficult to discover the specific data they need for their research. Traditional search engines, such as Google, are not effective in most of these cases. Moreover, some of the scientific data are confidential and are not publicly indexed. Therefore, it is necessary to develop a custom solution that satisfies the researchers' need for flexible search. This article introduces the Data Discovery Service designed to serve the Virtual Research Environment (VRE) during the VI-SEEM project. The solution is based on a specially configured and upgraded version of the CKAN platform. The main installation procedure is described, as well as the authors' contributions for the purpose of regularly harvesting data from different sources and updating the available content on which user queries are to be executed.

Key words: data service, data harvesting, search engine, Virtual Research Environment

AMS subject classifications. 68P20

1. Introduction. Nowadays, researching often requires the processing and analysis of many and large volumes of data and metadata. One of the frequent difficulties for researchers is to find exactly the data they need for the analysis. Usual and freely available search engines are seldom useful because scientific data is highly specialized and complexly structured. We will present the VI-SEEM Data Discovery Service (VDDS), which is a solution to search scientific data and it is developed during the VI-SEEM project [1]. The main purpose of this project is to create a useful virtual environment for the researchers located in Southeast Europe and the Eastern Mediterranean. Researchers are organized into thematically targeted communities which are: Climate, Life sciences and Digital Cultural Heritage.

The VRE (Virtual Research Environment) consists of several services which are available to the scientific users [2]. The services for VI-SEEM project are integrated in a public catalog which is accessible at this link: [3]. There are three main groups of services: Data storage, Computing, Application specific and Authentication/ Authorization. The presented Data Discovery Service belongs to Data storage group.

There are several popular platforms for hosting and searching across a variety of metadata, Open Data Portals that are mainly used in scientific areas. We looked at the most commonly used ones, such as DSpace, CKAN, Zenodo and Figshare [4] [5] [6] [7]. Our main goal is to have a flexible, easily upgradable search service for scientific metadata, with a rich and well-documented API (Application Programming Interface). Other requirements we have are that the chosen platform must be long-established, open source and has big community. Zenodo [8] and Figshare [9] dropped out of our choice because they are mainly focused on processing scientific papers and documentation. However, we need a platform that hosts and performs search in any metadata, while offering a convenient editing interface. DSpace [10] is designed to create and manage repositories for large volumes of diverse data, and there is also a search engine that is not as flexible in complex metadata. In addition, it is already used to host the project's main repository as a part of the VRE and therefore it has been dropped as a choice and finally we selected CKAN [11].

We present the main points of the installation and configuration of the CKAN platform, which is the basis of the VDDS and the additional resources such as Data Synchronization Tool that are developed and added to form the complete solution. Primary performance tests and evaluation are presented too.

2. Data Discovery Service design and implementation. As a basis for our solution, the CKAN open source platform is chosen. It is a comprehensive system with rich functionality for building large and complex data sites, including a flexible search engine. CKAN has convenient, easy to operate interface on one hand and very rich API on the other. Having user friendly interface is very important because the target users are

[*] This work is supported by EC programme H2020, Grant Agreement No 675121, VI-SEEM Project.

[†] Institute of Information and Communication Technologies - Bulgarian Academy of Sciences, Bulgaria (vgd@acad.bg)

[‡] Institute of Information and Communication Technologies - Bulgarian Academy of Sciences, Bulgaria (sstoyanov@parallel.bas.bg)

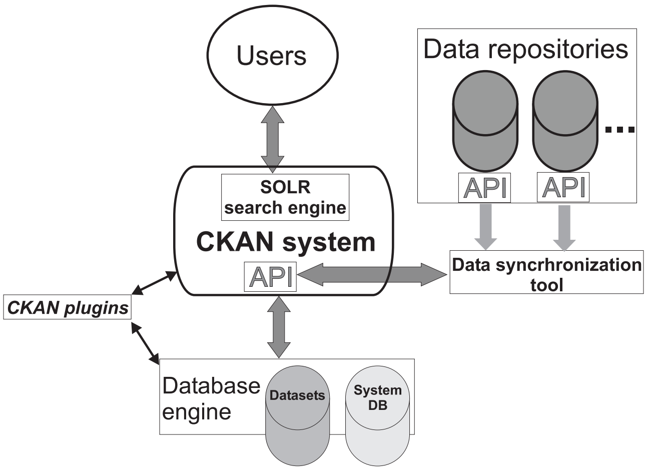


Fig. 3.1. *VDDS common architecture.*

researchers who are competent in their specialized field of study but not necessary an advanced IT proficient. Moreover, it provides well documented and supported API [12] and options for third party extensions (plugins). Many research, governmental and non-governmental organizations use CKAN for building and managing their open data sites.

On the back end the programs which form CKAN are written mainly in Python and uses Pylons web framework and SQLAlchemy [13], the database toolkit as its Object Relational Mapping (ORM). The front end is implemented in JavaScript. The database engine is the popular PostgreSQL.

In our case, special interest is SOLR [14], an open source enterprise search engine, written in Java, developed by Apache Software Foundation, which is an integral part of the system.

3. Software architecture. The common architecture of VDDS is depicted on Figure 3.1.

Customized **CKAN system** is the core of our VDDS. It is configured to support several organizations which correspond to the VI-SEEM project participants and three groups which correspond to the main research communities in the project: Climate sciences, Life sciences and Digital Cultural Heritage. Each organization and community can maintain metadata for their datasets. In addition there are common purpose communities for generic usage, software projects and testing purposes.

SOLR is the main search engine which is accessed by the end users.

API is the documented CKAN Application Programming Interface which is used internally by **Data Synchronization Tool** to harvest external metadata from **Data Repositories**. Also there are different **API**s for the external **Data Repositories**. This tool is written in Python.

Data repositories. These are external sites containing data that are of particular interest to us.

CKAN plugins. These are additional programming modules that extend the functionality of the system.

Database engine. Software components and storage space that hold and maintain the datasets.

4. Data Synchronization Tool. In order to search for data uploaded on the VI-SEEM Repository Service (VRS) or VI-SEEM Simple Storage Service (VSS) using the data discovery service (VDDS) data synchronization needs to be made of VDDS's database and the databases of the other services. More precisely this synchronization process is one-directional data flow from the source of the original data to the server running the data discovery service.

In order VDDS to be useful for its users, then the results of each search query have to be correct and complete. Therefore, this data flow process has to be executed regularly, i.e. the operation should be fully automated.

One proposed way to automatically extract the necessary metadata is via a harvester. DSpace [10], which is used to create the repository service, has such functionality implemented as OAI-ORE/OAI-PMH Harvester [15]. The main advantages of this solution are that no additional software is required to collect the data from

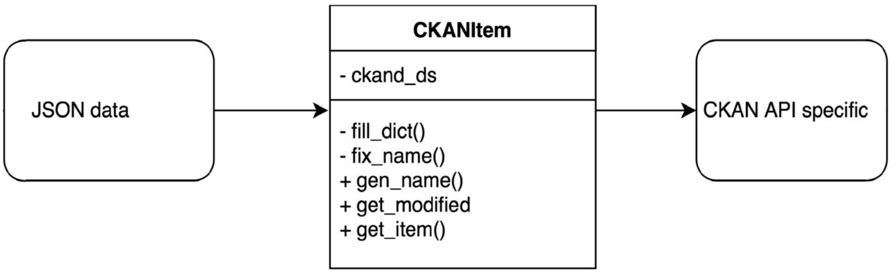


FIG. 4.1. *CKANItem core class*

VRS and also it is supposed that the procedure is more efficient. However, the output of the harvester's work usually is designed to be imported to similar interoperable repository or another DSpace software e.g. as a backup or redundant node part of more complex high availability system. It is not suitable format for our purpose and it needs significant transformations before it is ready to be uploaded on the data discovery service. Moreover, this solution is not very generic as it is very specific to the technology compiling VRS.

We decided to create our own software tool that solves the current problem and is universal to that degree so that it could easily be expanded in future to support more data sources if required. Accordingly, our solution must have modular design.

As the core of that solution one class is responsible for uploading data to VDDS and it is presented on Figure 4.1. Its role is to process some kind of generic data e.g. Java Script Object Notation (JSON) and return as output a specially formatted dictionary meeting the requirements of VDDS's CKAN based API. CKAN's Action API is RPC-style and has very rich functionality exposing all of CKAN's core features to API clients.

A function is implemented in another module that accepts the output and sends it to the data discovery service using the action call to the CKAN API package_create.

With these two modules available the task of synchronizing VDDS with any data repository is reduced to mapping the metadata tags and obtaining array of JSON formatted data.

Python version 3 is chosen as a programming language for implementing the tool. The main reason behind that decision is because of the fact that the CKAN software is written in Python and also has very good documentation and user guides with examples in that language on how to use the API. Also tools algorithm doesnt contain any compute-intensive part, neither execution time is of such importance to the performance of our system as the tests we conducted showed normal responsiveness and stable operation of the search service under greater than average system load. So using lower level programming language than Python to yield more efficient resource usage is not necessary at all. Although CKAN is written in the older version 2 of the language, we went for the newer version because it is actively supported by its core developers meaning that it is getting tha latest and better features. Third factor is that Python 3 interpreters are built in every Linux distribution available and deploying the tool doesn't require much effort.

VRS is a REST-compliant service and provides all the required operations to download the necessary metadata. We've made extensive tests of the load level and performance of the repository service when querying the REST interface because a total number of:

$$queries = 2\left(1 + \sum_i C_i\right) \tag{4.1}$$

where C_i is the number of collections of i-th community, are sent every time data is being synchronized and thus it is probably less efficient than harvester. But test results are very consistent and there are no signs of performance downgrade.

Figure 4.2 shows an example of the workflow of the tool when data is being synchronized between VRS and VDDS.

Dspacemetadownloader module executes the queries and does some additional conversion of the JSON as it is in nested format without unique keys that have to be combined before the JSON is passed to the

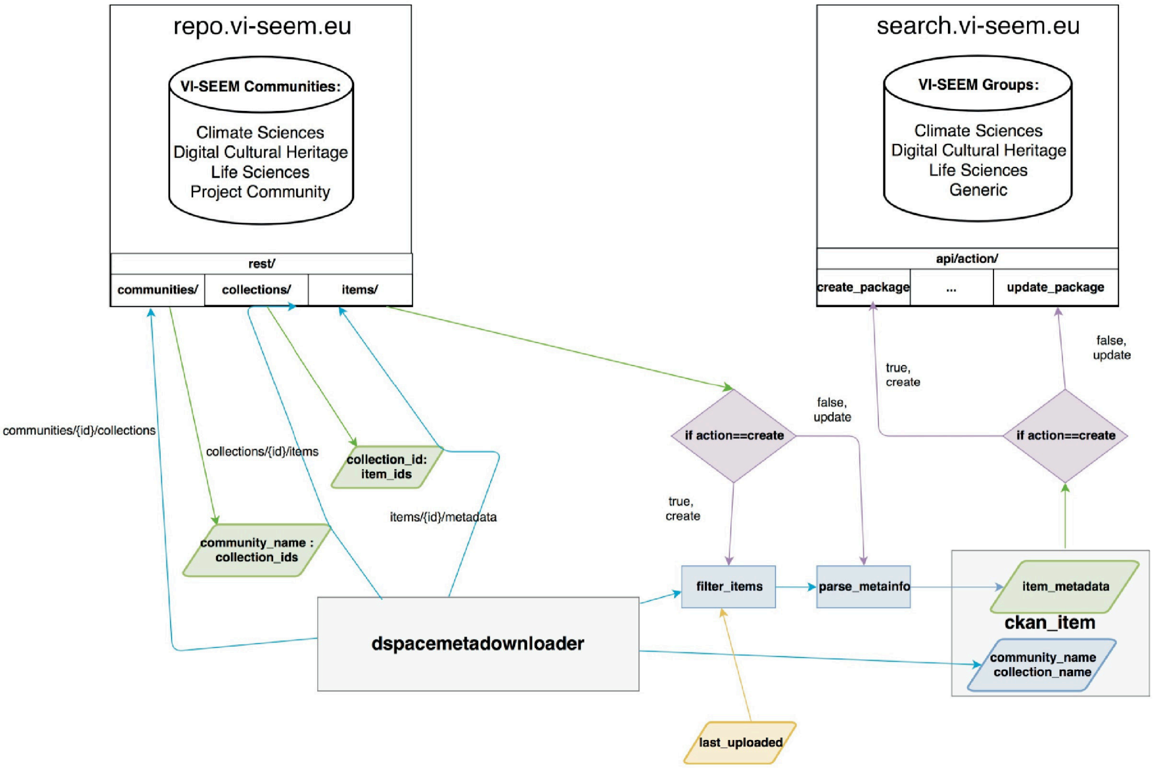


Fig. 4.2. *Workflow of the tool when synchronizing data between VRS and VDDS*

CKANItem class.

The tool runs in either create or update mode. Create is the default behavior and filters already synchronized items, uploading only new items to VDDS. If update mode is specified the metadata of all items will be checked and updated. Both modes operate absolutely independent of each other.

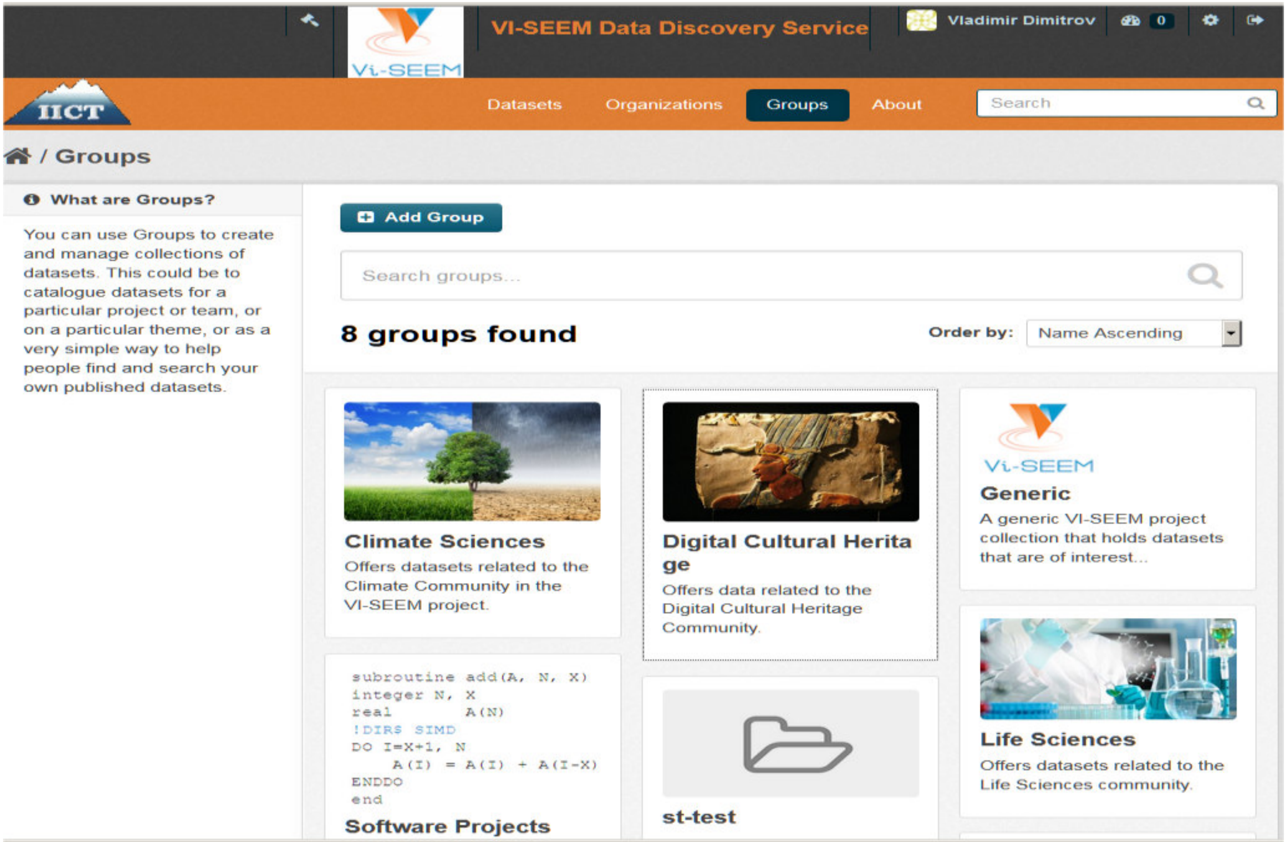
Unfortunately CKAN has very strict rules about the allowed characters uploaded using its "Action API". Hereof the main module **ckanitem** parses every metadata string and substitutes appropriately every unaccepted symbol. In case that such character is not caught by the module or for another reason the operation fails, unsuccessful upload is recorded in a log file with timestamp and error description.

5. Installation. The installation procedure of such kind of software is usually very tricky and capricious. We tried several different methods to install, including compiling the source code. Most of these methods had minor or big problems as of the middle of 2016 using the existing production versions of software components at this time. The only method that worked is recommended by the developers of CKAN and finished successfully. It requires OS Ubuntu 14, 64 bit, kernel 4.2.0-42-generic, Python 2.7.6, CKAN 2.5. For Data Synchronization Tool On the same machine Python 3 is installed too. Our experiences with other, even newer distributions and versions were painful.

The installation of Data Synchronization Tool is simple. It is a bunch of Python scripts which are copied in a directory and configured in **crontab** to be executed twice a day: around noon and around midnight.

Example screenshot of the VDDS front end is presented on Figure 5.1.

6. Solution evaluation. We conducted numerous load and performance tests after the installation of VDDS. For this purpose, we first need to import demo data in the database of the server [16]. These demo examples are available at the official source code repository of CKAN and contain many and diversified data. This is important, which is important for the tests that they are capable of generating significant stress on our server. For example, the dataset named Newcastle City Council: Payments over 500 contains over 23,000 rows



You can use Groups to create and manage collections of datasets. This could be to catalogue datasets for a particular project or team, or on a particular theme, or as a very simple way to help people find and search your own published datasets.

8 groups found

Climate Sciences
Offers datasets related to the Climate Community in the VI-SEEM project.

Digital Cultural Herita ge
Offers data related to the Digital Cultural Heritage Community.

Generic
A generic VI-SEEM project collection that holds datasets that are of interest...

```
subroutine add(A, N, X)
integer N, X
real        A(N)
!DIRS SIMD
DO I=X+1, N
     A(I) = A(I) + A(I-X)
ENDDO
end
```

Software Projects

st-test

Life Sciences
Offers datasets related to the Life Sciences community.

Fig. 5.1. *VDDS example screenshot. User groups section.*

Table 6.1
Averaged results obtained by running various requests during 60-second active connection to VDDS API.

	Latency (ms)				Requests per second				
No.	Avg	Dev	Max	Dev(%)	Avg	Dev	Max	Dev(%)	TX,KBs
1	32.21	12.65	179.85	97.24	31.94	5.79	40.00	66.33	18.61
2	111.48	14.90	247.78	96.88	9.19	1.69	10.00	89.80	429.63
3	133.82	15.74	264.15	97.11	7.95	2.60	10.00	67.41	367.15
4	67.40	11.10	217.30	97.11	14.82	5.09	20.00	46.86	354.97
5	655.05	33.43	757.82	89.01	0.96	0.20	1.00	95.73	38.66
6	655.68	31.88	772.43	91.21	0.94	0.24	1.00	93.87	40.10
7	610.92	31.70	747.05	90.82	0.97	0.18	1.00	96.63	37.74
8	70.83	17.04	313.68	96.44	14.26	5.04	20.00	52.90	366.99
9	268.53	25.50	485.26	94.26	3.20	0.63	6.00	76.44	116.36
10	108.40	65.87	504.06	66.84	54.44	7.67	60.00	92.62	492.12

in six CSV files for each month from September 2011 to February 2012. The tests use **wrk** program [17]. **wrk** is a HTTP benchmarking tool producing different levels of load on the server depending on the used command line arguments. We run it in multithreaded mode and adjust connection and duration parameters to tune the intensity of each test. Every additional connection simulates another active user working with VDDS. Produced outputs contain measurements of main performance metrics such as latency, bandwidth and number of requests.

Test results are shown in Table 6.1.

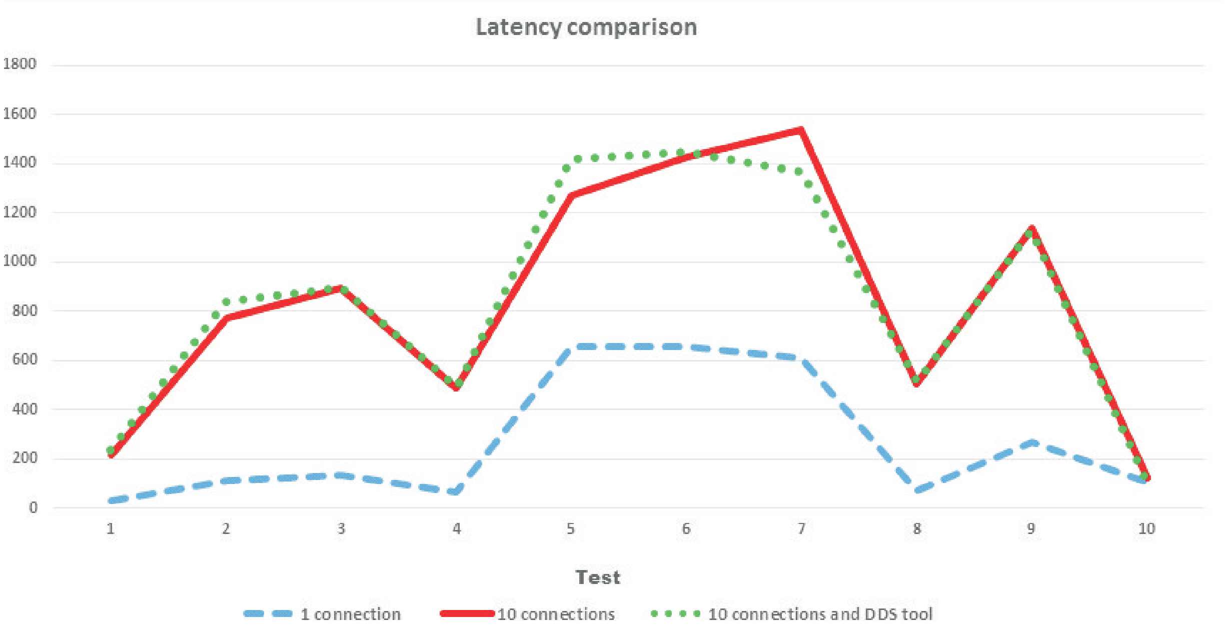


Fig. 6.1. *Latency comparison with different number of connections.*

The test suit consists of 10 HTTP requests to the following URLs:

1. /
2. /dataset
3. /dataset?page=2
4. /dataset?tags=transparency
5. /dataset/adur_district_spending
6. /dataset/us-national-foreclosure-statistics-january-2012
7. /dataset/activity/afghanistan-election-data
8. /dataset/adur_district_spending/resource/resource_id
9. /group/data-explorer
10. /package_list

Three main sections in the table are the corresponding metrics for latency, bandwidth and number of requests. For the first two sections we present statistical data in the columns for average(Avg) value, standard deviation (Dev), maximum (Max) and percentage of the requests within standard deviation. Last section is the average achieved transfer speed (TX) for every test. We increase the value of connection argument to increase load on server as more users would induce. That way we analyze how the performance scales with more stress and discover at which point the server reaches its limit. As we approach more than 20 active users at once, we cant find anything unusual and the evaluation of results from tests points the same. Non-linear dependence is observed between number of connections and performance drop across all tests. Image 6.1 shows this relation.

The blue dashed line represents average latency of all requests during 1 active connection. Continuous red solid line is for 10 simultaneous connections to the same APIs. Latency delay approaches a maximum 7.24 times higher with an average value of 4.68 times more across all tests for 10 times increase in load. Despite having 10 concurrent users acting on the system, the number of requests made by each one is only about 10 percent less. At the same time transfer speed increases up to 2.39 times during test eight. An average of 1.75 on the whole test suite indicates that the system doesnt starve for bandwidth. The dotted green line is present on the chart to visualize the impact of the used resources by data synchronization tool while the server has significant load at the same time. The previous 10 active connection tests are run together with dds tool. It is clear that performance differs by a small margin and it does not affect user experience at all.

7. Conclusion. VI-SEEM Data Discovery Service uses metadata mapped onto standardized facets and which can be collected from various research and other repositories and provides the users with the possibility for flexible search and browsing. It is possible to search for keywords, partial phrases, creator, organization, publisher, time of publishing, versions, tags, research areas and communities etc. The results are presented in a user friendly form. Searching of a particular dataset is performed using easy to use command-line Python scripts or a simple web accessible form. The search task can be either different types of free-text search or so-called faceted search, concerning tags stored in the metadata accompanying the data. The users may refine their searches inside the received results.

By virtue of its nature, VDDS offers its services relying on the data of the original sources. For this reason, completeness and correctness of its data is vital and that is why it needs to be reconciled properly and regularly. We decided to implement our own tool to take care of this task. We called it **Data Synchronization Tool**. It is developed specifically for this purpose with possibility for future upgrades in mind. After first extensive tests had been made, the tool is deployed in production mode and now is repeatedly updating VDDS's database providing its intended services to VI-SEEM users.

Acknowledgments. This work is supported by EC programme H2020, Grant Agreement No 675121, VI-SEEM Project.

REFERENCES

[1] *VI-SEEM project: Virtual Research Environment (VRE) in Southeast Europe and the Eastern Mediterranean (SEEM)*, https://vi-seem.eu

[2] Atanassov, E., Karaivanova, A. and Gurov, T., *Services And Infrastructure For Virtual Research Environments For Use By Science And Business*, International Scientific Journal Industry 4.0, Issue 2, 2016, pp. 110-113, Published by Sci Tech Union of Mechanical Engineering, ISSN:2543-8582, (open access).

[3] *VI-SEEM Services Catalog*, https://services.vi-seem.eu

[4] Jeffery, K. G. and A. Asserson, *Data Intensive Science Shades of Grey*, Procedia Computer Science (2014), pp. 223.

[5] Amorim, R. C., J. A. Castro, J. R. da Silva, and C. Ribeiro, *A Comparative Study of Platforms for Research Data Management: Interoperability, Metadata Capabilities and Integration Potential.*, Advances in Intelligent Systems and Computing (2015). Vol. 353. Springer, Cham.

[6] Neumaier, S., J. Umbrich, and A. Polleres., *Automated Quality Assessment of Metadata Across Open Data Portals*, Journal of Data and Information Quality 8 (1). doi:10.1145/2964909.

[7] Kubler, S., J. Robert, Y. L. Traon, J. Umbrich, and S. Neumaier, *Open Data Portal Quality Comparison using AHP*, (2016), doi:10.1145/2912160.2912167.

[8] *Zenodo platform*, https://zenodo.org

[9] *Figshare platform*, https://figshare.org

[10] *DSpace software for open digital repositories*, http://www.dspace.org

[11] *CKAN platform*, https://ckan.org

[12] *CKAN Application Programming Interface*, http://docs.ckan.org/en/latest/api

[13] *Python SQL Toolkit and Object Relational Mapper*, https://www.sqlalchemy.org

[14] *SOLR open source enterprise search platform*, http://lucene.apache.org/solr

[15] *Open Archives Initiative Protocol for Metadata Harvesting*, https://www.openarchives.org/pmh/tools

[16] *CKAN demo data*, https://github.com/ckan/ckan-demo-data

[17] *wrk - HTTP benchmarking tool*, https://github.com/wg/wrk

Edited by: Anastas Mishev
Received: Dec 22, 2017
Accepted: Mar 30, 2018

DOI 10.12694/scpe.v19i2.1399
Scalable Computing: Practice and Experience ISSN 1895-1767
Volume 19, Number 2, pp. 189–201. http://www.scpe.org © 2018 SCPE

DICOM DATA PROCESSING OPTIMIZATION IN MEDICAL INFORMATION SYSTEMS

ALEXANDR GOLUBEV[*], PETER BOGATENCOV, AND GRIGORE SECRIERU

Abstract. The problem of storage and visualization of medical images collected by various medical equipment is actual for latest 10 years for every medical institution. On the other hand, access to the medical investigation datasets and solving the problem of personal patient data security is important for scientific community and institutions that require this data. "DICOM Network" project was developed for solving these problems for different actors in the system based on the various customized roles. This article describes the problems and possible solutions for optimization of medical images storing, providing stable and secure access, based on the distributed warehouse for huge volumes of data with different levels of access.

Key words: DICOM; Distributed storage system; HPC; Processing algorithms; Radiology; e-Health;

AMS subject classifications. 68M14

1. Introduction. Twenty years ago, after the patient's visit, the doctor had only data on several indicators: general information about the disease, weight, pressure and symptoms. Today a large amount of digital information coming from different sources - from x-ray images to telemetry from implantable devices, such as cardiac monitors. As medical institutions abandon the paper records of the disease, these data are collected more electronically. The availability of large datasets of digital medical information allows doctors to improve diagnoses and health care in general, there is even a new concept - research "in silicon." Information technologies (IT) are now used at all stages of health care, from basic research to the provision of health services, and includes many specializations, such as bioinformatics, clinical informatics and biomedical informatics

Clinical informatics is devoted to the use of IT to improve health care and covers such processes as the preparation of medical bills, the planning of patient care, the allocation of resources for patient care. Clinical decision support systems can, for example, alert the attending physician about the potential interactions of medications, based on the patient's medical history and known allergic reactions.

Modern medical information systems integrate various types of medical equipment. This article describes the actual problems and solutions for optimizing processing and storage of medical radiography investigations. The standard for working with medical images is the DICOM format, which allows storing studies in good quality with the patient's personal data included. The main problem in storing data in DICOM format [1] is caused by the fact that one study can produce more than one gigabyte of data and consists of thousands of images.

DICOM (Digital Imaging and Communications in Medicine) - the standard for processing, storing, printing and transmitting information in medical imaging systems. It includes a description of the file format and network protocol. The network protocol uses TCP / IP in its core for communication between systems. Also, systems that support reading and writing DICOM files can exchange files in DICOM format among themselves. The owner of the standard is the American Organization National Electrical Manufacturers Association (NEMA). It is developed by the Committee of the DICOM standard, consisting of several working groups (WG).

DICOM allows the integration of scanners, servers, workstations, printers, and networking equipment from many different manufacturers into a single PACS (picture archiving and communication system). Different devices are delivered with a document called the DICOM conformance statement, which describes how and what functions the supplied device performs.

The "DICOM Network" [2] project was launched in Moldova in 2015 having a goal is to provide access to investigations for medical staff with the appropriate access rights and as well as patients themselves to the personal radiography investigations. The current realization of the project allowed connecting eleven types of medical equipment to the deployed system. Today the system collects and processes more than 1000 gigabytes of data per month.

[*] Research and Educational Networking Association of Moldova Str. Academiei 5, office 324, MD-2028, Chisinau, Republic of Moldova (galex@renam.md).

One of the problems that should be solved in DICOM Network project is the problem of storing medical investigations archive on national level that can be considered as Big Data issue. Solution for this issue should take into account the different data access levels. On the one hand, a medical investigation contains personal patient data, which means that data access should be restricted and secured. This could be reached by permission-based categories of users and individual investigations access on supervised approval. On the other hand, data should be accessible by any authorized user, like patient or doctor from any location. One of the main priority is system performance that should allow high speed access to the huge amount of data. Thus, the process of storing and transferring large volumes of medical data can be divided into several components: archiving and storage, retrieval and accessibility of data, data transfer to the end user's workstation and processing the data on the client application.

When medical investigation is completed and DICOM image set imported from the equipment raises the problem of data archiving and data storing. If X-ray photography usually does not contain more than 2-3 images, a tomographic survey can contain up to 1000 slices and take up to several gigabytes on the physical storage. It is easy to calculate that for a large hospital or a large diagnostic center with a daily flow of 500 or more patients, the data volumes will be terabytes per month, but the archive of investigations should be saved for a minimum of three years by law (and even for longer period for practical use). As a result, medical institutions could not archive such a volume of data, because in many cases there are no sufficient capacity available for their storage at the level of one institution.

As far as "DICOM Network" was selected as pilot application for integration into distributed regional VI-SEEM platform [3], the DICOM Network system architecture was adjusted to the project needs that make possible to increase the number of users and offer access to the DICOM Network for research community. Thanks to VI-SEEM project and gaining access to RENAM[7] computing infrastructure the storage capabilities and processing power is increased in many times and now DICOM Network collecting about 300 investigations per day and over 1TB of data per month. All this is possible by using of a number of the developed components: DICOM Server, DICOM Portal, DICOM Viewer, DICOM Audit and others that are enough flexible to cover the requirements of any medical institution or specific scientific community.

2. DICOM File Format Overview. The standard using in "DICOM Network" system for working with medical images is the DICOM format, which allows storing images in good quality. The main problem of data storing in DICOM format in integrated medical data processing systems is caused by generation of multi-gigabyte volume of data and necessity to work with thousands of images. The "DICOM Network" project provides access to investigations for medical staff, patients, and medical researchers with the appropriate access rights to radiography investigations. In the article we describe and analyze possible solutions for optimizing data storing and processing workflows.

2.1. DICOM data structure. DICOM files simultaneously contain both images and additional information about associated patient. Information about the patient and the study cannot be separated from the image itself. This reduces the number of possible errors. The JPEG format file is similarly organized, which can also have additional information describing the image in the file. Any DICOM object consists of a set of attributes, such as the name of the patient, its identifier, the date of the study, etc. Also, one special attribute contains image data (pixel data). Thus, there is no separate header for the DICOM file - only a lot of attributes, including image data. Attributes in the standard are called tags, each tag consisting two fields - the group number and the element number. For example, in the tag numbered 0010, 0010 (tag numbers are written in hexadecimal notation), the patient's name is always included. Each tag has a standard name. 0010, 0010 is called 'Patient's Name'. A list of all standardized tags can be found in the 6th section of PS 3.6: Data Dictionary [4].

The 7FFE, 0010 'Pixel Data' tag can contain one or more images. In the case where the Pixel Data contains more than one image, it is said that the file contains a multi-frame image. In the case of a multi-frame image, one file will contain three or four (for example, several scanned sequences in several places, but at different times) dimensional image. Digital X-ray machines, digital cassette readers give information in the form of a single-frame image. Apparatus like ultrasound, angiographs often give multi-frame images. Old X-ray and magnetic resonance tomography could give single-frame images. Modern tomography (after standardization of the expanded formats DICOM-CT Enhanced and MR Enhanced) can give both single-frame and multi-frame images.

Image data can be color and monochrome. Color can be in different color coding - RGB, YBR, Palette Color (color palette). Monochrome can be of different depths of gradation of gray (1 - 16 bits). Image data can be packed. The following packing algorithms are standardized: RLE, JPEG, JPEG Lossless, JPEG LS and JPEG 2000. For the whole file it can be applied archiving using LZW (zip) algorithm, however, the realization of such packaging in programs and equipment is rare.

DICOM uses three different schemes for encoding the tags (transfer syntax). The encoding of the file is marked with the corresponding tag in the same file. Schemes are obtained from combinations of two parameters - data representation and byte order coding.

The data representation can be Explicit and Implicit. You need to know how to interpret the data contained in the tag, because this is a simple sequence of bytes. And what kind of data is there - a string, a number, or a sequence of tags (SQ-sequence) is not known in advance. For certainty, the content of each tag has been standardized. Each tag has a standardized representation of tag data (VR, Value Representation) - OB, OW, OF, SQ, UT, UN, etc. When the data is explicitly represented in tags, the VR of the tag is explicitly written. With an implicit view, VR is not written, but is taken from the program table that works with this image.

The order of bytes can be from the oldest to the youngest (big-endian), the record begins with the older one and ends with the youngest, and from the youngest to the oldest (little-endian, "pointed"). The DICOM uses three of the four possible combinations of parameters: Implicit little endian, Explicit little endian, and Explicit big endian. Each tag consists of: group number (2 bytes), element number (2 bytes), VR (2 bytes, in explicit data representation, not implicitly used), tag length (2 or 4 bytes, depending on VR). Some standard VR tags: DA - Date, date; DS - Decimal String, a string representing a decimal fraction; FL-Floating Point Single (4 bytes), a floating-point number of ordinary accuracy; IS - Integer String, integer string; UL - Unsigned Long, unsigned double word, etc. Full information on VR and the principles of encoding tags can be found in the standard - PS 3.5: Data Structure and Encoding [4]. In addition to presenting data (to Value Representation), there is the concept of a plurality of values (VM, Value Multiplicity). VM is not marked in real files, it's just an indication of how much data it is supposed to contain a specific tag. For data represented by strings, the elements are separated by a backslash character (’́). Numeric data simply goes sequentially byte - for example, if VM = 2, the tag with VR = FL will consist of 8 bytes - these are two numbers of usual accuracy. In tables with lists of tags, the VM of each tag is specified. For example, a tag containing one coordinate of a two-dimensional point will have VM = 2. The point containing n coordinates will have VM = 2 * n, n¿ 0 (in the standard 2-2n is written).

2.2. DICOM Protocol and services OVERVIEW. DICOM consists of many different services, most of which involve the exchange of data over the network. The file exchange was added to the standard later and is only a small part of the standard.

DICOM service 'Store' is intended for transferring images or other objects (for example, structured reports - structured reports) between two DICOM devices.

Storage Commitment - This service is used to confirm that relocated objects are successfully stored in the information store. With this message, the person who receives the data - PACS, or the station informs the transmitting device - or the station that the data has been successfully saved and can be deleted.

Query / Retrieve-a service for searching and delivering whole studies or individual objects on a remote DICOM device. You can search for specific filters (for example - research date, patient's name, etc.) of the research or object of interest (most often objects in DICOM are images, but not only) and request its transfer to the local computer.

Modality Worklist allows the device (often devices in DICOM are called modalities, however the same device, for example, a lithotripter, can have several modalities - US, DX) to obtain a list of planned studies. The data on the planned studies contain information about the patients, this allows you to reduce the re-entry of the same information and the accompanying errors.

Modality Performed Procedure Step additional to the Modality Worklist service, which allows the modality to send a report on the success of the research, about the images received, the time of the beginning and the end of the study, the dose received by the patient, etc. Service allows you to get the hospital management more accurate data on the use of the resources of the device. The service, also called MPPS, allows improving the interaction of the modality and the PACS system by providing the system with a list of objects that will be sent before the parcel itself.

Printing service allows to send images to print to the DICOM printer, to get a hard copy of the images, most often on tapes. There is a method of standard calibration of printers and monitors, which helps to obtain identical images on different monitors and on a hard copy of images.

2.3. DICOM Files and services OVERVIEW. In addition to directly patient data and research, data points of the image, the files must necessarily contain the so-called meta-information (File Meta Information, tag group 0002). The meta-information indicates how to interpret correctly the contents of the file.

DICOM limits the length of file names to 8 characters, file extensions are not allowed. The names of the files must be such that no information can be obtained from them. These are historically established requirements for maintaining backward compatibility with older systems. On the medium, except files, the dicomdir file should be placed in the root directory. Dicomdir provides general indexed information about all the DICOM files on the media. Dicomdir provides more information about each file than it is possible to fit into the file name.

DICOM files that are not on the media, usually have a .Dcm extension, the media should contain files without an extension.

Some common modalities:

CT - Modality of type Computed Tomography
DX - Modality of type Digital Radiography
MR - Modality of type Magnetic Resonance,
OT - Modality of type Other
US - Modality of type Ultra Sound
XA - Modality of type X-Ray Angiography

3. "DICOM Network" Database Architecture. DICOM Network project based on a number of self-installed components that are build using a number of modules. The assemble of components and modules provides a flexible architecture of the DICOM "Network" that is consolidated by DICOM DATA Interface into one radiology investigations database. The system includes Data Storage and Data Processing components distributed between different processing units and storages, which could be customized using specific interfaces. The general architecture of the "DICOM Network" system presented in Fig. 3.1.

Project started from one node located in the Institute of Urgent Medicine (IMU) www.urgenta.md located in Chisinau, Moldova that served only one Institution and had storage element only of 2TB. Then thanks to VI-SEEM project and availability of RENAM computing infrastructure the storage capabilities and processing power increased in many times and now we are collecting about 300 investigations per day and over 1TB of data per month.

- Data from equipment are collecting by the modules integrated in "DICOM Server" that can be installed in the same location with the used medical equipment or can be distributed through different institutions, or even cities or countries.
- All investigations (DICOM Images) are archiving at DICOM Servers, but information about investigation is stored in DICOM Portal (like www.dicom.md) database. Usually various DICOM Servers connected to one DICOM Portal.
- DICOM Portal stores all data like users description, access info, system settings, DICOM Server settings and some others, but not DICOM images it salves. Each institution can deploy DICOM Portal internally using one or several own DICOM Servers.
- DICOM DATA Interface collects information about users and investigations from all DICOM Portals and provides functionality for data exchange and unification.

The functionality of the implemented system covers all necessary workflows for processing and documentation of medical investigations - from images collecting directly from equipment to archiving investigations in the patient medical record [5]. DICOM Network offers extended functionality for enhancing quality of medical management and secured access to investigations. This helps doctors, specialists and penitents to gain access to structured database of medical images, allows documenting images that are collecting from various medical apparatus. At institutional level, the system helps to reduce costs of investigation, raise the quality of provided services.

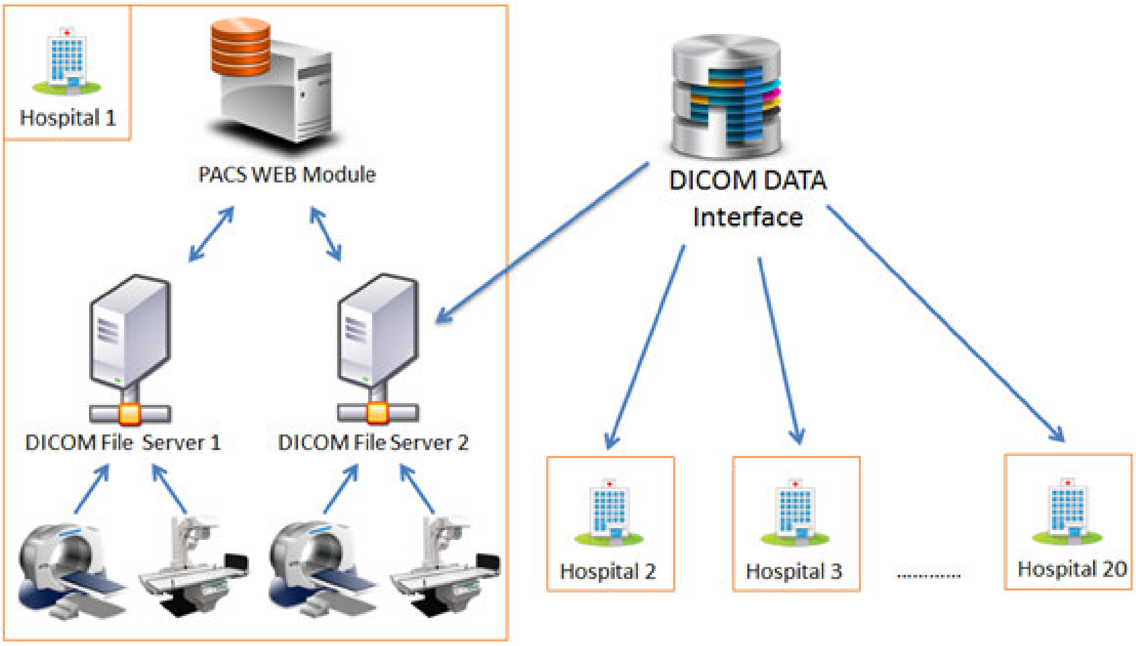


Fig. 3.1. *General scheme of DICOM Network architecture*

The considering DICOM Network system is already in production operation and can be accessed by link http://www.dicom.md/. The Graphical User Interface (GUI) is presented in the Fig. 3.2.

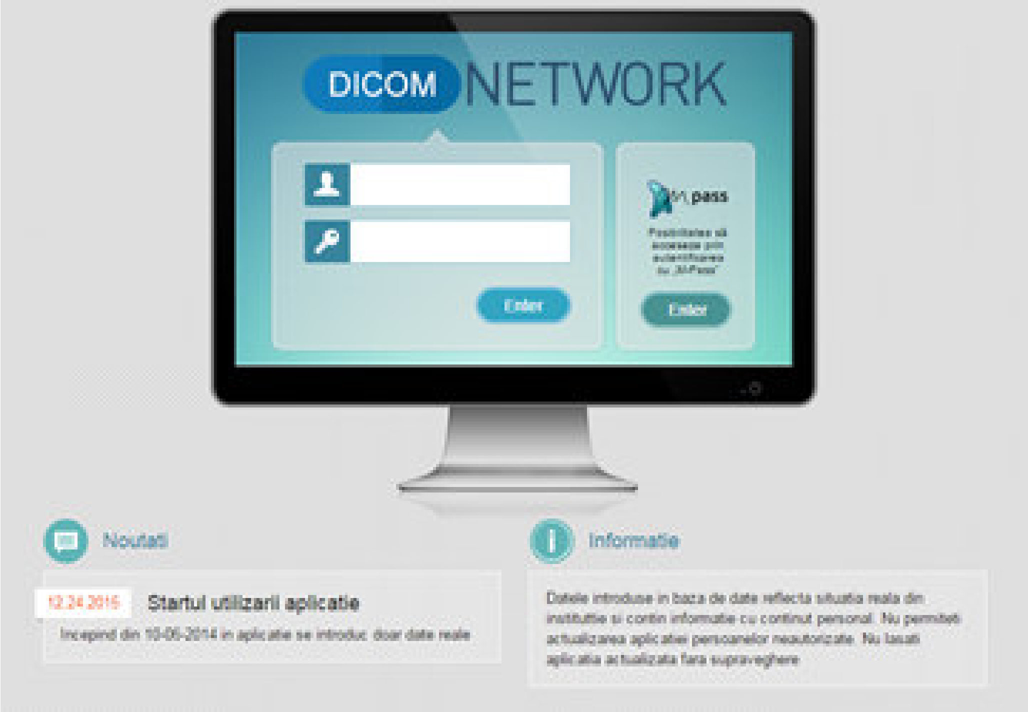
The system initially was deployed at the National Centre of Ambulance of Moldova (the Institute of Emergency Medicine) and during the first year of functioning has shown it effectiveness and attractiveness for medical personnel that is resulting in:

- Three DICOM Portals operational
- Four DICOM Servers installed
- Eleven types of medical equipment were connected to the "DICOM Network".
- About 300 investigations per day are collecting by the system.
- Over 700 doctors have access to their patients investigations from their working place.
- Over 170 000 investigations were searched and downloaded by doctors and radiology specialists during 36 months period.
- Additional budget savings ensured for hospitals due to refusing of printing investigations using expensive consumables.

"DICOM Network" system is actively developing, as far as doctors are interested in operative accessing to radiographic image sets directly from their workplace and institutions want to save money and get technological support for operative and precise decisions. A new types of equipment connected to the system that increases the number of imported investigations. These developments expand the main problem of the systems working with medical images - extreme increasing of data amount that must be preserved.

As far as database is storing personal patient data, security is one of the main key features of the "DICOM Network". We implemented Permission/Role/User system that allows a flexible access to the data. No one investigation or patient can be unauthorized. Each users activity starting with basic access to patient data is logged in the system in separate databased and could not be removed. Each image download is also logged and IP address of the data requester is saved. All of this provides possibility to track and audit each investigation. Access to audit is provided not only for system administrators, but also to responsible persons that have privileged users roles inside the institutions.

The main levels of access are based on following principles:



Posibilitatea să accesezi prin autentificarea cu „M-Pass”

12.24.2015 Startul utilizarii aplicatie

Incepind din 10-06-2014 in aplicatie se introduc doar date reale

Datele introduse in baza de date reflecta situatia reala din institutie si contin informatie cu continut personal. Nu permiteti actualizarea aplicatiei persoanelor neautorizate. Nu lasati aplicatia actualizata fara supraveghere

Fig. 3.2. *DICOM Network GUI*

- Patient can access his investigations based on national ID.
- Doctors can access all the patients investigations in case these are patients that are treated by this doctor.
- Department or Chief of section have access for all patients from his section.
- Specialists from radiology department have access to all data from their institution.
- Scientist and researches have access to the investigations shared by institution, but only in anonymized mode.
- Lawyers and courts can have temporary access to the investigations based on the request of the patient and institution.
- Patient can grant registered access to investigation to another doctor.

Today the data is stored on different DICOM servers and one Storage element. DICOM servers store an archive for the last month and on a common storage is storing a common archive for the last 2 years. This architecture allows providing the fastest access to more recent data that is used more often, while access and downloading images from Storage archive take longer time.

"DICOM Network" is only one information system that works with medical images data storing from different sources on national and international level in this region. Application is based on cloud technologies that makes possible to distribute it on the different locations and dynamically increase resources on demand. One of the benefits of this system is that it is using resources of scientific cloud infrastructure that makes possible to access the investigations both by specialists for daily-based activities and by scientists for research. The other applications that are working in this domain are restricted by their using in one specific institution and do not provide tools for secure data exchange inside the application or organizations. Most of the Information systems based on the open source software and usually these products have insufficient quality or limited functionality, or specialized PACS provided by the equipment supplier that is adopted only to the target equipment without ability of extension in large informational systems.

One of the similar to DICOM Network system is "NeoLogica" software solution. "NeoLogica" designs and develops advanced medical imaging software solutions since 2002. The benefit of "NeoLogica" software is

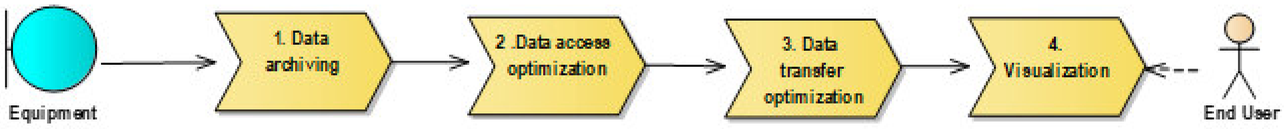


Fig. 4.1. *DICOM Network distribution using VI-SEEM platform*

also cloud-based architecture that is using cloud hosting solution in Amazon cloud. However, national laws of personal data protection in the most countries, organizations and hospitals is not accepting storing the sensitive data abroad, on Amazon servers, also it increasing costs of data transfers.

4. Main System Optimisation Objectives. First, to describe the solution for DICOM data optimization, here we overview of the algorithms and processes that are necessary for any information system working with radiography images. To obtain the source images, DICOM provides functionality for the import of data directly from the equipment. After importing the image, you must process and save it on commonly accessible storage. As a result, a medical research database available to the authorized user is created. The user accessing the database should receive data through the local or wide area network, and then using the visualization application to work with DICOM images on his workplace.

Thus, the problem of data handle optimization for this kind of information systems can be divided into three stages. First, when you import and write the source files on storage, you need to archive the data to minimize it volume saving the quality of the images. Secondly, when accessing data, it is necessary to transfer the data to user as quickly as possible while optimizing the data format to reduce the amount of transmitted images. Thirdly, the data should be optimized to speed up its loading and processing on the local processing facility.

On the diagram below Fig.4.1 are showed full process of images optimization that is representing by four sequence steps - Archiving, Access, Transfer and visualization:

In the subchapters below are described assembly of solutions for each image lifecycle step optimizations. The proposed algorithm saves 100% original image quantity, optimize the storage elements and reduce the transferred size of the thumbnails, that could be zoomed on demand. On the final step of visualization server prepares the image set in such way that it could be visualized using low performance processor such as PC or mobile device.

4.1. Data archiving. As it was mentioned above, one radiographic examination can vary from 1 to more than 1000 images. Thus, the investigation can consist (has a size) of more than one image. Of course, the most interesting and useful is archiving of large tomographic studies with a large number of slices. First, let us look at the structure of the DICOM file. In general, each file contains lines of the patient's metadata and the image itself in the raw format. Since during DICOM file processing all meta-information is written in the database, then when archiving we can discard metadata and save only the image. If necessary, it will be possible to restore this data and generate a DICOM file on client request, or use the new file format when transferring it to specialized application like DICOM Viewer. Of course metadata is only a small part of the file and takes only a couple of kilobytes, but it should be taken into consideration, that metadata for one investigation is almost identical, so for a set of data in 1000 slices it will be possible to save significant amounts of physical storage space. On the other hand, excluding the patient's personal data from the original files significantly reduces the possibility of personal data leakage. This approach also allows to create and transfer of impersonated data.

Archiving of image itself is already a much more complicated task that requires complex data compression algorithms. It should be taken into account that it is necessary to exclude loss of image quality since each pixel is important for data processing and visualization. Based on analysis produces we found that archiving an individual slice without loss of image quality does not give a tangible result, but it is worth noting that the proportion of individual slices is small enough and refers mainly to x-rays, when one investigation consists of 1-2 images. The quantity of this data type that is stored does not exceed 1%. Analysis of more complex tomographic surveys with 100 or more slices shows that standard algorithms of video image compression could

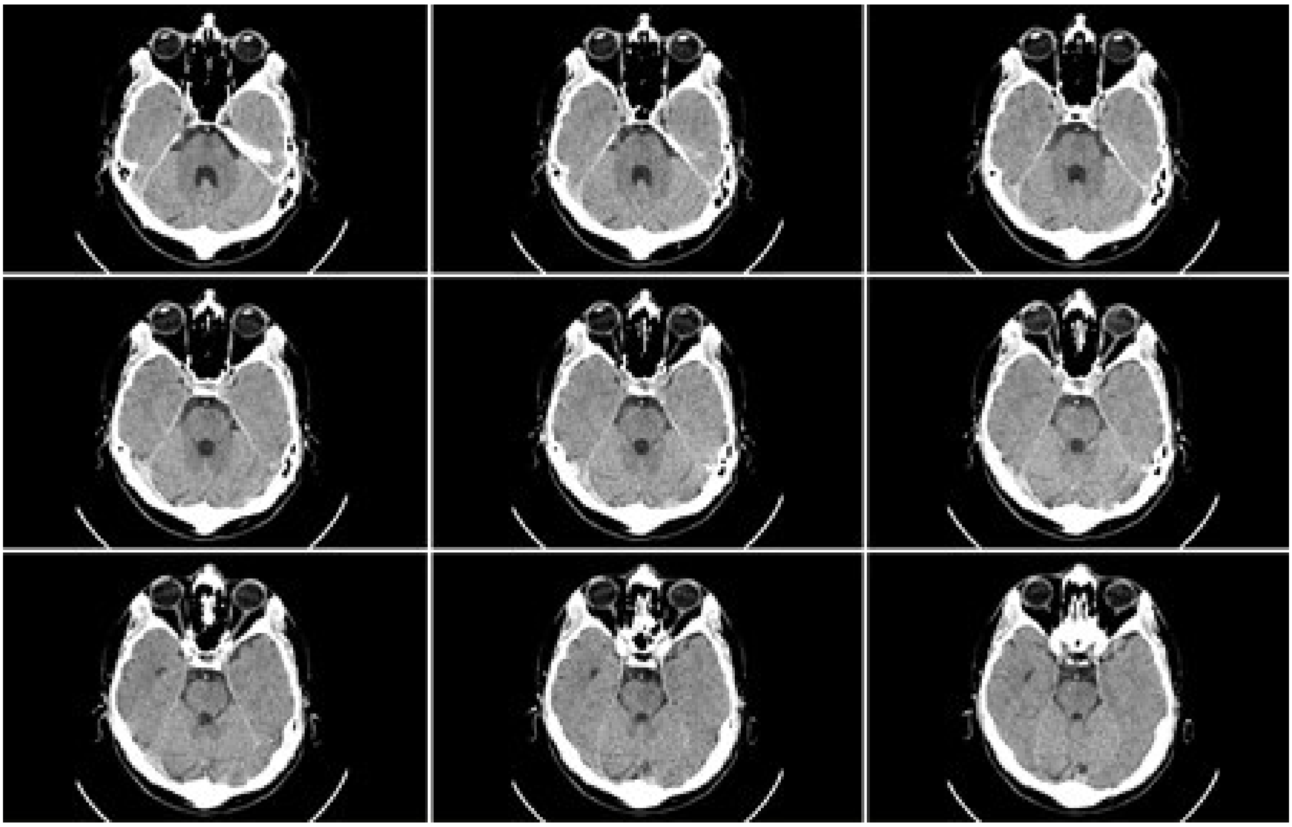


Fig. 4.2. *Similar radiology investigations slices*

be applied for their archiving, since the changes between slices (frames) are insignificant. In Fig. 4.2 presented the "neighboring" nine slices, the differences in adjacent slices are so small that you can confidently convert these slices into video stream where only the differences for pictures are retained and not the raw format for each image. This way, you can reduce the amount of images with multiple slices by 10 or more times. Of course, you need to take into account that for archiving you will need considerable computing capacity, which involves the use of high-performance computing systems such as HPC[8] or computing GRID[10].

4.2. Data access optimization. Another problematic aspect of data storage is the distribution of data among various data storage systems with different levels of security[10] and access speed. As already mentioned above, storing data in one centralized system is not effective for a number of reasons. To solve this problem, we proposed a distributed data storage system showed in Fig. 4.3.

The main components of the proposed distributed storage system architecture are:

- Local Storage Since the equipment connected to front-end DICOM server directly, the server must be located both in the local network and in the wide-area network. The medical equipment should be located only in the local network. For solution of this problem is using a local server, physically located near to the equipment (one per cabinet or laboratory). This server receives the DICOM data and performs the primary data processing. On one hand, this server provides the fastest access to data; on the other hand, a real storage capacity of this server is not exceeds 1-2 terabytes, which implies the stored data availability within 1-2 months. This allows high-speed access to the latest data. After a predefined period, the data archived and transferred to the next institutional storage level ("Laboratory Storage") or to "External Storage" level.
- Laboratory Storage Storage element for one institution, available in the local network of the organization. This storage has extended parameters, but usually they are limited too and can store data for

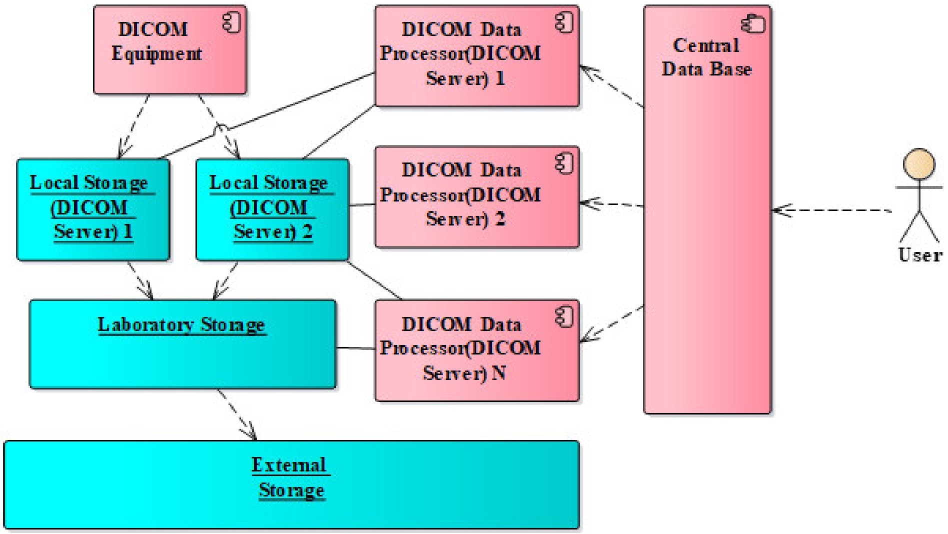

DICOM Equipment
DICOM Data Processor(DICOM Server) 1
Central Data Base
Local Storage (DICOM Server) 1
Local Storage (DICOM Server) 2
DICOM Data Processor(DICOM Server) 2
User
Laboratory Storage
DICOM Data Processor(DICOM Server) N
External Storage

FIG. 4.3. *Distributed Big Data Storage*

1-2 years.

- External Storage Distributed storage, located partly inside and mainly outside of the organization where the investigations were created and interoperable through the global network infrastructure. This type of storage has the lowest access speed, but it can store a huge distributed archive of data at the national or international levels

4.3. Data transfer optimization.. Regardless of the storage location for their visualization, images must be uploaded to the end user's computer. At the same time, regardless of the transmission channel, which can be a high-bandwidth local network or the global Internet network with unpredictable QoS, data transactions can be very large, and the speed of downloading may influence on quality of medical services and on convenience for medical doctor of the information system using.

The main solution to minimize delays is certainly implementation in the system procedures of archiving and decompressing of data described above, but it is worth taking into account the following opportunities for transmission optimization:

- Preparing a data packet taking into account the permission from client application of the receiving specific data sets. That is, if a user views a survey on a low-resolution mobile device, there is no sense in sending full-screen images at the maximum resolution.
- Ability to load a specific slice in the maximum resolution. At the request of the client application, it is necessary to provide the possibility of sending a particular image at the maximum resolution, for example, if necessary to use zoom in / zoom out.
- Loading data in background with a separate thread. In this case, the end user can begin to visualize not a complete set of data, while the full set will be loaded asynchronously and displayed as the load is finished.
- Caching data at the client and server level.

All of the above can significantly speed up the transfer of data and make the application for visualization more user friendly.

4.4. Preprocess of data for visualization.. After uploading data to the end user's computer, the application that implements the DICOM image rendering must process the data for visualization. In this case, all data must be loaded into the operative memory for fast processing. It should be noted that the DICOM Viewer (the application that displays the DICOM images) should not only display the image and change the slices, but also perform more complex operations - from drawing to building 3D models, modeling and

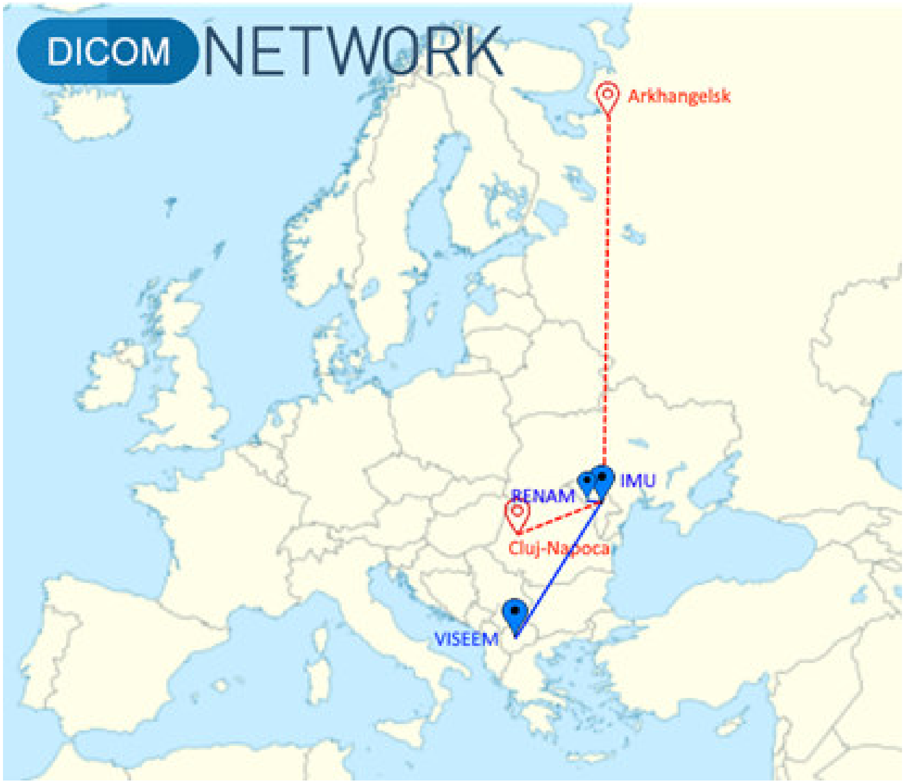


Fig. 5.1. *DICOM Network distribution using VI-SEEM platform*

representing the tissues. Not every desktop computer, much less of mobile devices have satisfactory computing capabilities. That is why the above-mentioned operations may take a long time.

The solution for this problem should be preparation of data for visualization on the server part of the system. Considering the use of high-performance systems, on the multiprocessor server can be build and transferred ready-made models that will not require complex conversions on the client facility.

5. Integration of "DICOM Netwrk" in VI-SEEM Platform. As far as "DICOM Network" was selected as pilot application for integration into distributed regional VI-SEEM platform [5], the DICOM Network system architecture was adjusted to the project needs that made possible to increase the number of interested users and provide access to the system for wider community for research purposes. The current number of collected investigations in the system is over 187 000, that are regularly requested. Many medical institutions and different medical staff are working with this system. Two medical institutions refused using any other PACS and archiving system and now are using exclusively DICOM Network for patient workflow.

Now installed three DICOM Portals and four DICOM Servers in two countries: Moldova and Macedonia. The clinic from Cluj-Napoca in Romania and Federal University named after M.V. Lomonosov in Russia expressed their interest to install DICOM Network. The current state of "DICOM Network" distribution is show in the Fig. 5.1.

At present available the following DICOM Portals:

- http://dicom.md/ in IMU
- http://renam.dicom.md/ in RENAM
- http://viseem.dicom.md/ in Macedonia (VI-SEEM project).

As was shown above, the solution for storing such large volumes of data is the distributed storage for images archiving. One solution is to participate in international projects and use storage elements from the resources available for the project realization. This cooperation is mutually beneficial for both "DICOM Network" and

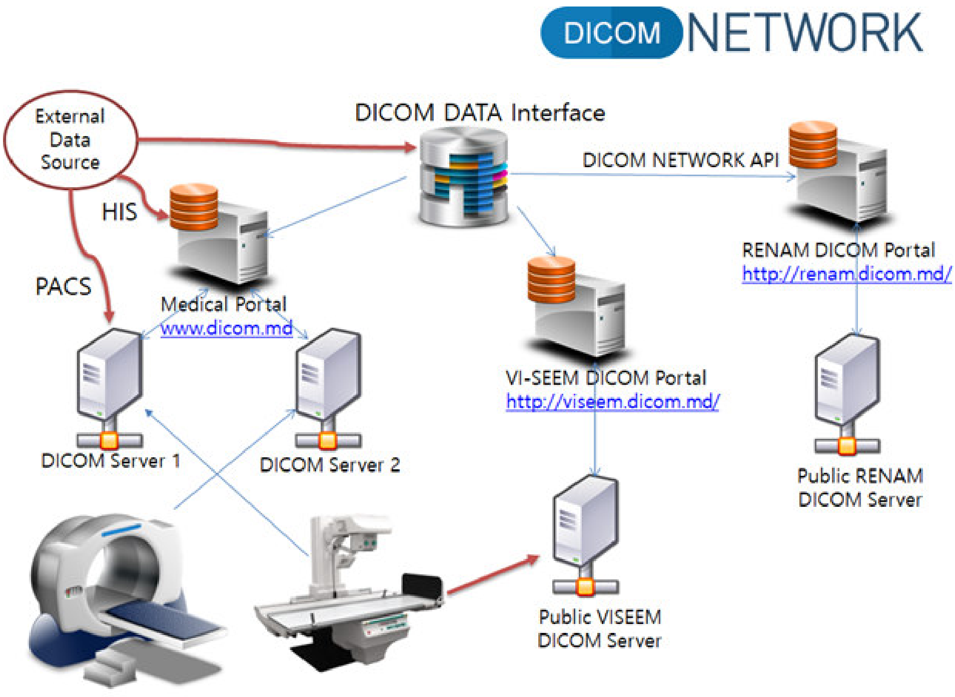


Fig. 5.2. *DICOM Network architecture for integration with VI-SEEM platform*

the VI-SEEM platform. VI-SEEM provides project resources while "DICOM Network" datasets for research community. Of course, all data collecting on the VI-SEEM servers is impersonated.

In the framework of VI-SEEM project an instances of DICOM Portal and DICOM Server were installed on the Macedonian server and are available on using following link http://viseem.dicom.md/. This server, that is available for research community have been configured for both collecting (DICOM Server installed) and distributing the medical images using DICOM Portal. Each organization can register on this server their equipment and then medical images data will be uploaded and accessible using DICOM Portal. Based on assigned role and permissions data can be anonymized, shared or distributed through other portals according to user defined rules in the system management interface of the portal that is accessible from the web browser. Now 900 gigabytes with anonymized datasets were already uploaded to the server and are available for researchers.

VI-SEEM project deploys and offering user-friendly integrated e-Infrastructure platform for Scientific Communities in Climatology, Life Sciences and Cultural Heritage for the South-Europe and Mediterranean regions by linking compute, data and visualization resources, as well as generalized services, software and tools. The regional infrastructure deployment concept is presented in Fig. 5.2

In the Fig. 5.2 presented the concept of connecting national DICOM Network application, that it is containing existing DICOM Portal http://dicom.md/, with DICOM Portal installed by using VI-SEEM platform resources[8]. DICOM DATA interface grants the interconnectivity for different users of the both portals and allows displaying DICOM investigations using the both portals interfaces. Public DICOM Server grants possibility for any authorized VI-SEEM platform user to add and retrieve the investigations from DICOM Network application and use the developed facilities based on configured and granted access rules.

VI-SEEM platform will offer possibility to install and configure publically available DOCOM Portal that can be used by any interested institutions to store, access and share medical images. Setting up public DICOM Portal instance will increase the level of access to DICOM investigations and will help to make DICOM Network services available to regional medical research and practicing community.

6. Concusions. Now "DICOM Network", although actively developing, is still far from realization the whole potential incorporated into the system. Taking into account the growing number of medical equipment and the trend towards modernization and computerization of various health facilities, the system will be able to receive and will have to process dozens of terabytes of source information. The storage and subsequent transfer of such large amounts of data is an expensive process, which effective development impossible without optimization and new approached realization. On the other hand, for the successful development of the system it is necessary to accumulate and provide access to the archive not only for 3 years, as requested by law, but for ten or more years to monitor the patient's condition changes and maintain a full medical record. It is also necessary to take into account the need for backup copies of such important information. It is easy to calculate that even for rather small country as Republic of Moldova the volumes of data are too large to store them in an unprocessed form. Thus, the issue of data optimization and using effective archiving algorithms are the key factors for development of e-Health systems.

The benefits of implementation of new effective archiving algorithms are determining the following four directions of medical information systems development:

- Reducing the costs of medical data storage and maintenance;
- Reducing internet traffic for data access and in such way reduce the costs for data transfer;
- Increasing the quality of radiology services for patients.
- Solving the main problem for DICOM images database growth insufficient space for permanently increasing amount of imagistic data.

Radiology medical investigations services are offered by majority of medical organization starting from located in small villages to huge laboratories in the specialized medical centers. In any medical institution effective archiving algorithms should be in a great demand, because on the one hand they allow reduce the costs, save organizational budget [6], and on other hand increase quality of offered services, open new opportunities for collaboration with other institutions and making joint research. Of course, the most perspective consumers of the proposed solution are large diagnostics centers in governmental and private medical organizations that have modern equipment and make huge number of investigations that require archiving and transferring to other medical institutions. However, the implemented solutions will be also interesting for small hospitals that do not have their own equipment but anyway need to have access to the radiology investigations for their patients. Using the proposed solution these small hospitals will obtain possibility to have access to the investigations that were done in external institutions like specialized diagnostics centers.

Datasets collecting in the "DICOM Network" system will provide new opportunities for researchers. Although the system is now in production stage, functionality of the "DICOM Network" is permanently enhancing. During the process of the system implementation beneficiaries have ability to specify their specific necessities for providing additional features and services, such as:

- Studying and realization of new methods for optimization of data transfer and archiving.
- Image preprocessing and detection of anomalies.
- Incorporation of expert systems to help making diagnoses for doctors.
- Development of open APIs for "Dicom Network" to collect, archive, access and jointly process medical images at international level using distributed computing infrastructure.

Algorithms and solutions described in the paper are open for joint development and can be applied to various medical information systems. They do not depend on any specific project, since the developed approach involves tight integration with the open DICOM standard and mostly complements it rather than modifies.

REFERENCES

[1] A. Golubev, P. Bogatencov, G. Secrieru. *Optimal Methods of Storage, Transfer and Processing of DICOM Data in Medical Information Systems*, International Conference on Distributed Computer and Communication Networks 2017, p 269-280

[2] A. Golubev, N. Iliuha, P. Bogatencov *DICOM Network services DICOM data exchange solution integrated in the regional VI-SEEM infrastructure*, Smart Technologies, IEEE EUROCON 2017-17th International Conference on, 558-563

[3] A. Golubev, P. Bogatencov, G. Secrieru. *Updating DICOM Network Architecture for its Integration at International Level*, Networking in Education and Research, 15th RoEduNet IEEE International Conference, Bucharest, Romania, 7-9 September 2016, pp. 161-166. ISSN 2068-1038.

[4] DICOM *Digital Imaging and Communications in Medicine*, Published by National Electrical Manufacturers Association. PS 3.6-2011.

[5] P. Bogatencov, N. Iliuha, G. Secrieru, A. Golubev. *DICOM Network for Medical Imagistic Investigations Storage, Access and Processing*, "Networking in Education and Research", Proceedings of the 11th RoEduNet IEEE International Conference, Sinaia, Romania, January 17-19, 2013, pp. 38-42. ISSN-L 2068-1038

[6] A. Golubev, P. Bogatencov, G. Secrieru, N. Iliuha. *DICOM Network - Solution for Medical Imagistic Investigations Exchange*, International Workshop on Intelligent Information Systems. Proceedings IIS. 13-14 September, Chisinau, IMI ASM, 2011, pp. 179-182. ISBN 978-9975-4237-0-0

[7] P. Bogatencov, G. Secrieru, N. Iliuha, N. Degteariov, G. Horos *New developments of Distributed Computing Technologies in Moldova*, CEUR Workshop Proceedings. Selected Papers of the 7th International Conference Distributed Computing and Grid-technologies in Science and Education. Dubna, Russia, July 4-9, 2016; Vol-1787, urn:nbn:de:0074-1787-5, pp. 20-25. ISSN 1613-0073

[8] N. Degteariov, Bogatencov P., Iliuha N., Horos G *DEPLOYMENT OF THE SCIENTIFIC CLOUD COMPUTING INFRASTRUCTURE IN MOLDOVA*, Proceeding of the 5th International Conference "Telecommunications, Electronics and Informatics", May 20 23, 2015, Chisinau, UTM, 2015, pp. 27-29, ISBN 978-9975-45-377-6

[9] A. Golubev, P. Bogatencov, *New Trends in Research and Educational Networks Security Teams Operation. An Overview of Automated Tools for CERT*, Proceedings of ITSEC-2012 International Conference on Information Technologies and Security, 15-16 October 2012, Chisinau: NCAA, 2013, 181-187. ISBN 978-9975-4172-3-5

[10] N. Iliuha, A. Altuhov, P. Bogatencov, G. Secrieru, A. Golubev, *SEE-HP Project Providing Access to the Regional High Performance Computing Infrastructure*, Proceedings IIS "International Workshop on Intelligent Information Systems", September 13-14, 2011, Chiinu, 183-186. ISBN 978-9975-4237-0-0.

Edited by: Aneta Karaivanova
Received: Jan 5, 2018
Accepted: Apr 12, 2018

DOI 10.12694/scpe.v19i2.1354
Scalable Computing: Practice and Experience ISSN 1895-1767
Volume 19, Number 2, pp. 203–214. http://www.scpe.org © 2018 SCPE

IMPROVING SERVICE MANAGEMENT FOR FEDERATED RESOURCES TO SUPPORT VIRTUAL RESEARCH ENVIRONMENTS

ANASTAS MISHEV[*], SONJA FILIPOSKA[*], OGNJEN PRNJAT[†] AND IOANNIS LIABOTIS[†]

Abstract. Virtual research environments provide an easy access to e-Infrastructures for researchers by creating an abstracted service-oriented layer on top of the available resources. Using the portal, researchers can focus on the research workflow and data analysis while being provided with a consolidated unified view of all tools necessary for their activities. The sustainable lifecycle of a virtual research environment can only be achieved if it is going to be used with high quality of experience by a large body of users. Aiming for this goal, in this paper we analyse the requirements and implementation of a cross-community virtual research environment that brings together researchers from three different domains. Promoting interdisciplinary research and cooperation, the federated virtual research environment is based on the service orientation paradigm, offering anything as a service solutions. Thus, the main pillar for a successful implementation of this solution is the careful design and management of the underlying elementary services and service compositions. The rest of the paper discusses the challenges of the service management implementation focusing on interoperability by design and service management standards.

Key words: e-Infrastructures, federation, service management, virtual research environment

AMS subject classifications. 68M14

1. Introduction. The vision of European Open Science towards pursuing excellent science sustainable in the long-term needs to be developed as a common research infrastructure serving the needs of scientists [1]. The European Open Science Cloud (EOSC) emphasizes the reusability of the existing e- and research infrastructures, and it should be a mixture of infrastructures, tools and services presented as interoperable virtual environments, "cloud of services", available to the researchers Europe-wide to store, manage, process, analyse and re-use research data across border and domains. This infrastructure should provide both common functions and localised services delegated to community level, where the EOSC will federate existing resources across data centres, e-Infrastructures and research infrastructures. In this context, e-Science is a paradigm for distributed networked, computationally- or data-intensive science, aimed at research problems that require collaboration using computational tools and infrastructures [2]. However, the future infrastructure landscape is developing into a high complexity entity: data systems of larger scale, cloud computing services, and highly heterogeneous technologies including new networking technologies (including software defined networks). Hence, performing research and experimentation has become ever more challenging and there is a growing need to lower the barrier for accessing and using the available infrastructure.

Different models have been proposed for providing shared access to the e-Infra-structure, where Grids and Clouds paving the way for solving the virtual environment sharing problem. A Grid is defined as an infrastructure that enables flexible, secure, coordinated resource sharing among dynamic collection of individuals, institutions and resources [3], where institutions that agree to share the resources are forming Virtual Organisations (VO) [4] and coordinating resource allocations based on agreed resource management rules within the VO domain. The problem with this model is that it places the resource requesters and providers in the same VO. The designed workflows in these systems are not always adapted to the methods of work preferred by the domain scientists making them hard to take advantage of. Cloud computing provides a different solution to these concerns, by separating scientists as resource requesters from cloud providers, and by allowing service owners to create user groups and manage access to deployed services. The on-demand, pay-per-use resource consumption model of cloud services, enables efficient use of equipment and flexibility for scientists [5]. The problem with cloud providers is that each provider has its own platform for accessing the services, which makes it hard for scientists that are in need of a uniform way of requesting services so that they can focus on their primary research activities, not wasting time on learning how to access and provision the services that they want to use.

Virtual Research Environments (VRE) [6] are innovative, web-based, community-oriented, comprehensive,

[*] FCSE, UKIM ({anastas.mishev,sonja.filiposka}@finki.ukim.mk)
[†] GRNET ({oprnjat,iliaboti}@grnet.gr)

flexible, and secure working environments conceived to serve the needs of modern science. Despite of their location, scientists should be free to use their browsers to seamlessly access data, software, and processing resources that are managed by various systems in different administration domains via Virtual Research Environments. The major challenges that need to be resolved to fully achieve a transparent VRE include large-scale integration and interoperability, sustainability, and adoption.

Service management in this complex VRE landscape becomes of great importance in order to provide users ease of access and use in a collaborative federated environment. With the growing demand for ever more complex e-Infrastructures, there is a strong requirement to sustain federated cross-domain experimental facilities ensuring the latest cutting-edge technologies are available to a large and experienced set of established research communities offered via centralised services.

The VI-SEEM H2020 funded project (VRE for regional interdisciplinary communities in the Southeast Europe (SEE) and the Eastern Mediterranean (EM)) focuses on bringing together the regional e-Infrastructures to build capacity and better utilise synergies, for an improved service provision within a unified Virtual Research Environment for the inter-disciplinary scientific user communities in the combined SEE and EM regions (SEEM) [7]. The main objectives of the VI-SEEM project include providing scientists with access to state of the art e-Infrastructure - computing, storage and connectivity resources - available in the region, and integrating the underlying e-Infrastructure layers with generic/standardised as well as domain-specific services for the region. The latter are leveraging on existing tools (including visualisation) with additional features being co-developed and co-operated by the Scientific Communities and the e-Infrastructure providers, thus proving integrated VRE environments.

In this paper, we present the VI-SEEM effort in the creation of the targeted communities VREs focusing on the establishment of an integrated service oriented approach in a federated interdisciplinary environment enabling end users to browse, access and use common and specific domain services in a unified manner. One of the main challenges in the project has been the development of a service management approach from defining the foundations of policies and practices to the implementation of a fully functional service catalogue and portfolio.

The paper is structures as follows: section 2 reviews the challenges and approaches in federated environments, while the section 3 describes the building blocks of service oriented VREs. Section 4 is dedicated to the service management with focus on the federated approaches in service management, along with the design and implementation of the management of the service lifecycle in the VI-SEEM VRE. Section 5 overviews the related work, while the conclusions and future work are provided in the section 6.

2. Challenges in virtual research environments. The main issues when aiming to implement a sustainable Virtual Research Environments are large scale integration and interoperability, sustainability, and adoption.

Since VREs are built as a collection of existing systems and resources, interoperability is a must in order to fully utilise the potential of all available resources. It is fundamental to rely on a rich array of systems and resources, both in terms of variety and size, that can be seamlessly accessed and combined in innovative ways to satisfy the evolving needs of the targeted community. The challenges affecting Virtual Research Environments are very broad and include every aspect of an e-Infrastructure as they represent the application layer that is built on top of one or more layers offering at raw resources, communication and authentication protocols, protocols for publication, discovery, negotiation, monitoring, and accounting of services. The infrastructure itself should address most of these challenges. Through a collection of mechanisms put in place, it should enable interoperability with existing systems to build a federated space of services. To address this issue, VI-SEEM has developed a set of open APIs for managing the service offerings in an unified manner that can then be integrated with additional systems and partners. In this way, all targeted communities are treated in a uniform federated way creating a completely transparent federated VRE.

Sustainability is another major challenge when aiming to develop a federated Virtual Research Environment. VREs require effort and money to be built and maintained according to the needs of the targeted communities. As proposed in [8], there are three key strategies for sustainability that can be put in place: (i) acquire further funding from diverse research bodies; (ii) develop business models aiming at self-sustainability; and (iii) rely on community support. Given the volatile nature of communities of practice the sustainability issue remains a challenging problem for singled out targeted communities. However, pursuing an interdisciplinary approach as fostered in VI-SEEM not only enables deeper research collaboration putting in place the initial tools needed for innovation, it also ensures a long term community support due to the increased coverage. The VI-SEEM exploitable assets including data management services, application specific services, computational and access services, and knowledge-based services, form a base line for the project sustainability.

Although several Virtual Research Environments have been developed in various application domains, the majority of these systems are not yet fully integrated into standard practices, tools, and research protocols daily used by real life communities. One of the hardest obstacles to cross is the unwillingness of the scientists to migrate from traditional and consolidated research practices and facilities to the novelty ones promoted by VREs. As recognised by [8], among the factors causing this issue are: (i) the lack of support of both technical and educational nature; (ii) the gap between the target community needs and the actual service implemented by the VRE; (iii) the reliability of the technology; (iv) legal, ethical, and cultural issues; and (v) internationality. All of these identified problems are addressed within VI-SEEM by organising educational training events that target all communities of interest, liaisoning with research communities representatives that provide input on their specific needs in terms of resources and services, and by working in a highly international environment that acknowledges different national issues and practices.

The federated approach taken in VI-SEEM in terms of multiple target communities is seen as a means to ensure the creation of an interdisciplinary virtual research environment that can be used as a blueprint for expanding to additional target communities thus expanding the user base and providing a tested environment that can further grow and follow the concept of a truly unified virtual research environment.

3. Service oriented VRE. All underlying e-Infrastructure can be offered to the VRE end users via a set of well defined services that are constructed in a way that enables the end users to achieve their requirements with minimum additional overhead in terms of service management and workflow definition.

In this context, a service offered in the VRE is a means of delivering value to end users by facilitating outcomes end users want to achieve without the ownership of specific costs and risks [9]. As long as the task or function being provided is well defined and can be relatively isolated from other associated tasks it can be distinctly classified as a service.

In the South East Europe and Eastern Mediterranean area, the established regional e- Infrastructure resources and application support enable the foundation for building service oriented VREs that will foster high-quality research and help reduce the digital divide and brain drain in Europe.

The VI-SEEM consortium brings together e-Infrastructure operators and Scientific Communities in a common endeavour: to provide an improved service provisioning for researchers within a unified Virtual Research Environment. One of the major features of the project is that it targets inter-disciplinary scientific user communities, building a single VRE platform for the Life Sciences community, the Climate community and the Cultural Heritage community that aims to support multidisciplinary solutions, advancing the community research, and bridge the regional development gap with the rest of Europe.

In other words, the project objective is to provide user-friendly integrated e-Infrastructure platform for regional cross-border Scientific Communities in Climatology, Life Sciences, and Cultural Heritage for the SEEM region. The resulting VRE portal provides access to computing infrastructure, data sets and data storage facilities, and cloud applications, as well as supporting services, models, software and tools, see Fig. 3.1.

The VRE supports the full lifecycle of collaborative research: workflows that provide support for simulations, data exploration and visualisation, possibility to access and share relevant research data, provided code and tools that can be used over the data sets to carry out new experiments and simulations on large-scale e-Infrastructures, optimized applications and libraries used for specific purposes. The newly produced knowledge and data can be stored and shared in the same VRE.

Moreover, the potential VRE users are not only researchers, but also students and SMEs that can benefit from using the services provided in the portal, see Fig. 3.2. For students, the VRE portal can be seen as a starting point to join the research communities by accessing training information and material and applying to participate on various related events. Also, the access to data sets, scientific workflows and source code repository enable students to familiarize with the state of the art research and analyze the results. For SMEs, on the other hand, the VRE platform can be seen as a place that provides possibilities for collaboration with the academic research institutions. The joint proposals are also provided with access to the infrastructure, data and

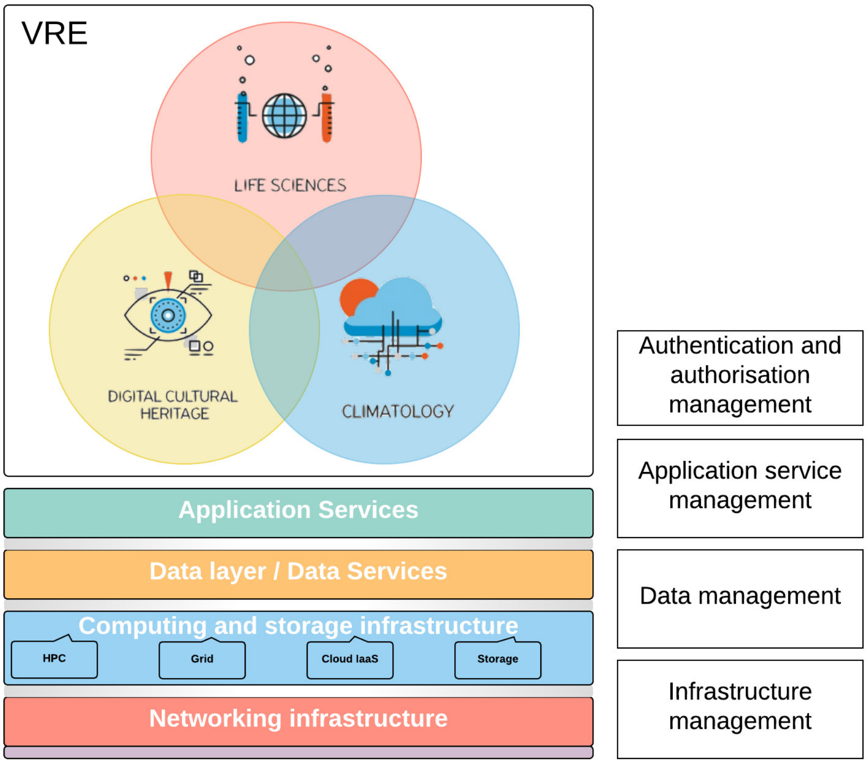

VRE
LIFE SCIENCES
DIGITAL CULTURAL HERITAGE
CLIMATOLOGY
Application Services
Data layer / Data Services
Computing and storage infrastructure
HPC
Grid
Cloud IaaS
Storage
Networking infrastructure
Authentication and authorisation management
Application service management
Data management
Infrastructure management

Fig. 3.1. *Cross-community VRE in VI-SEEM*

application services. The presented use cases can be used as successful examples that illustrate the potential and benefits from the joint academic/industrial efforts, representing one of the pillars for future sustainability of the VRE portal.

One of the major principles when developing VREs is the concept of service orientation. Service orientation is the ability and desire to anticipate, recognise and meet others' needs, sometimes even before those needs are articulated. This essentially means that the VRE needs to be developed so that is contains all of the required services by the target community in order to further empower endusers. Service orientation is based on a number of principles with the most important ones being [10]:

- Services are loosely coupled..
- Services abstract underlying logic.
- Services are autonomous.
- Services can be composed.
- Services are reusable.
- Services are stateless.
- Services are discoverable.

The service orientation principles are incorporated into the developed VI-SEEM VRE from several aspects. The federative type of organization implies loose coupling of the offered services. Similarly, all services are abstracted on a higher level so that the federated partners can deal with the local details. While the federated VRE presents a high level view of the available resources and services, the federated management tools still maintaining a more in-depth information about the way the services are implemented in order to provide high service availability via efficient service management, as it is discussed in the next section. The federated service management approach also takes into account the need for autonomy of services by maintaining a detailed level of service composition and tracking the status of all service components and interdependencies.

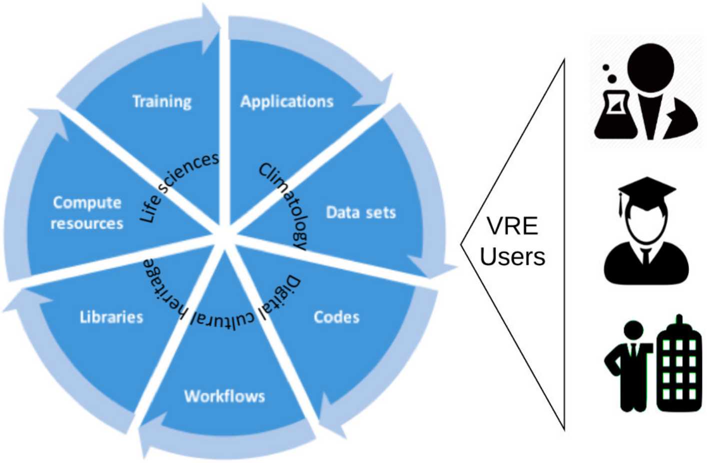


Fig. 3.2. *VRE content and users*

Furthermore, the scientific workflows section in the VRE provides information on how to compose the elementary services, reusing the same service in different ways for different workflows and scientific processes. The stateless nature of the services is a requirement when building scientific workflows where services can be reused in different settings for various users and domains.

There are multiple aspects of service discoverability that are implemented in the VI-SEEM VRE:

- Services are discoverable by the VRE users via the VRE portal - these are the end user services that can be elementary or composed complex workflows of services.
- Services are described in a service catalogue and portfolio - the catalogue component not only provides detailed service description, but also supports further discovery and integration of the services into a common European e-Infrastructure Services Gateway [19] using open APIs
- Services are internally enumerated in a service registry - enabling a coherent approach to service quality monitoring, accounting and support

4. Service management. It is evident that when implementing a service oriented VRE, one of the the key elements for a successful service orientation are the underlying components for service composition, management and monitoring encompassed in the IT service management.

IT Service Management (ITSM) [11] are the activities that an organisation performs to plan, deliver, operate and control IT services that are offered to customers. Such activities facilitate the offering of quality IT services that provide value to customers meeting their needs. IT Service Management provides the necessary guidance for an IT organisation to plan, design, develop, deploy and support business aligned IT Services. These services include the hardware, software and other IT assets necessary as well as the overall guidance for the IT organisation in the provision of these services.

In terms of federated environments, such as the ones that are fostered by the VI-SEEM project, the term IT organisation translates to a loose federation of the project partners, which makes the whole process of IT service management much more complex to define and implement. Thus, the first steps towards implementing service management in the federated VRE is choosing a suitable IT service management standard that will be used as a guidebook.

4.1. Service Management Standards and Best Practices. To successfully implement the service management process into an e-Infrastructure, there are several systems and processes that have to be put in place. The key elements are:

- Service management system, consisting of service catalogue and portfolio - the Service Portfolio is used to manage the entire Lifecycle of all services, and includes three Categories: Service Pipeline (proposed

or in development); Service Catalogue (live or available for deployment); and Retired Services.
- Service registry - The service registry is a database populated with information on how to dispatch requests to service instances.
- Monitoring system - real-time observation of and alerting about health conditions (characteristics that indicate success or failure) in an IT environment, ensuring that deployed services are operated, maintained, and supported in line with the service level agreement (SLA).
- Operational and service level definitions SLAs define what the organisation as a whole is promising to the customer, while OLAs define what the functional internal groups promise to each other.

The most important role of IT service management processes is to support the delivery of IT services. In many cases, the provisioning of one IT service requires several processes. All of these processes need to be successfully operating and interacting to deliver an IT service. This requires that each IT service management process, is defined using its standard elements, including:
- Goals and objectives
- Clearly defined inputs, triggers and outputs
- Set of interrelated activities
- Roles and responsibilities

There are many standards and best practices that cover the area of IT service management defining in details the common systems and processes that need to be put in place. ITIL [12] is one of the most frequently used, especially in the large enterprises. The IT Infrastructure Library (ITIL) provides a descriptive framework of best practices for the delivery of the components of the IT infrastructure as a set of services to the enterprise. Although ITIL is being used by a large number of IT organizations for efficient delivery of services to their users, it cannot be as successfully implemented in a case of federated environment, similar to the one built by most of the e-Infrastructure international projects including VI-SEEM. The latest efforts toward addressing these issues led to the establishment of a new, lightweight standard, with special focus on the federated environment, FitSM (federated ITSM).

The goals and activities of the FitSM standard series [13] are aimed at supporting effective, lightweight implementation of IT service management processes in an organization (or part of an organization delivering IT services to customers), and harmonizing ITSM across federated computing infrastructures. The main goals of FitSM are:
- Create a clear, pragmatic, lightweight and achievable standard that allows for effective IT service management.
- Offer a version of ITSM that can cope with federated environments, which often lack the hierarchy and level of control typical for enterprises.
- Provide a baseline level of ITSM than can act to support management interoperability in federated environments where disparate or competing organisations must cooperate to manage services.

Having in mind the federated nature of the VI-SEEM consortium, the FitSM was the clear choice for governing the IT service management of the offered services.

The IT service lifecycle is governed in FitSM using a set of 14 processes, where the standard defines the specific models for implementing them into the service management system. The process models are listed in Table 4.1.

4.2. Implementing VI-SEEM Service management. As described in the previous section, a service management system is an overall management system that controls and supports management of services within an organisation or federation. According to FitSM, the first process model that is the key element of the service management system is 1: Service portfolio management. This process model addresses the activities necessary to define and maintain a service portfolio.

The Service Portfolio (SP) is an internal list that details all services offered by a service provider including those in preparation, live and discontinued. The service portfolio includes meta-information about services such as their value proposition, target customer base, service descriptions, technical specifications, cost and price, risks to the provider, service level packages offered etc. The Service Portfolio is the basis for the Service Catalogue. The Service Catalogue (SC) is a customer facing list of services that are in production and provide value to the customers of the service provider. The SC, among others, provides also information on service

Table 4.1
FitSM requirements for a service management system

#	Process model
1	Service Portfolio management (SPM)
2	Service Level Management (SLM)
3	Service Reporting Management (SRM)
4	Service Availability & Continuity Management (SACM)
5	Capacity Management (CAPM)
6	Information Security Management (SM)
7	Customer Relationship Management (CRM)
8	Supplier Relationship Management (SUPPM)
9	Incident & Service Request Management (ISRM)
10	Problem Management (PM)
11	Configuration Management (CONFM)
12	Change Management (CHM)
13	Release & Deployment Management (RDM)
14	Continual Service Improvement Management (CSI)

options including the various SLAs available for each service. At a high level the service catalogue is a subset of the service portfolio, both in terms of the number of services that they contain and also in terms of the number of fields or attributes each one holds.

4.2.1. Requirements. The VI-SEEM service portfolio management system has been developed to support the service portfolio management process within VI-SEEM as well as being usable for other infrastructures if required. The main requirements for the creation of this tool have been collected from the service management process design that includes the infrastructure services, storage services and application level services. The service management system has been designed to be compatible with the FitSM service portfolio management. Requirements gathered in the context of EUDAT2020 [14] project have also been considered for compatibility and completeness.

The main functional requirements that were used as a foundation of the service portfolio management system development are as follows:

- The user roles that should have access to the tool functionalities are:
  - The potential customers or end users of the services listed in the service catalogue.
    These users should be able to see the list of the services that are currently in production or beta stage and are offer for use. Public details about each offered service should also be available. These users should be able to order and use the services listed in the service catalogue, as well as interact with the help desk or any other dedicated support channel for that service. The specific services features and use cases of using the service should also be available to the users.
  - The service managers within the VI-SEEM environment.
    These users should be able to see all service details about service that can be found in the service portfolio. This includes services that are not offered to the end users, as well as additional service information that is stored for management purposes only.
  - The service owners.
    Service owners are the persons responsible for each service listed in the service portfolio. Users with this role have the full responsibility for the content that is provided within the service catalogue and portfolio regarding the services under their responsibility. The service owners are assigned by the service providing partner institution.
- The service catalogue contains only public information about services that are to be provided to the potential customers and end users of the services. The service portfolio contains all services, including the ones not currently offered to end users, together with detailed information about each service (public details and management details).

- Each service can have multiple versions in the service portfolio. Each version can have a different readiness status i.e. concept, under development, beta, production, retired etc.
- The service owners and the service managers should be able to state which service versions should be available in the service catalogue.
- The service can be either customer facing or resource facing, internal to the organisation.
- The dependencies between services should be modeled in the service portfolio in order to facilitate efficient implementation of the SLA management process.
- The components that are required for deploying each service should be detailed in the service portfolio providing information needed for operations to deploy such services.
- The service portfolio should be accessible via a RESTful API to accommodate 3rd party application development and different views of the service catalogue depending on the needs of the consumer of this information (organisation, federation, external party, global e-Infrastructure catalogue).
- The service portfolio/catalogue should have a default web UI providing the main service catalogue view for the VI-SEEM project end users as well as to be used as a management UI for adding and editing information by service managers and service owners.

4.2.2. Design. FitSM defines a set of activities for the initial setup and maintenance of the service portfolio and the corresponding catalogue. For the service portfolio such activities are:

- Initial Process Setup
  - Define a way to document the service portfolio;
  - Define a way to describe / specify a specific service;
  - Set up an initial service portfolio (including service specifications) covering at least all live services provided to customers, as far as they are in the scope of the service management system;
  - Create a map of the bodies / parties (organisations, federation members) involved in delivering services.
- Ongoing process execution
  - Manage and maintain the service portfolio;
  - Manage the design and transition of new or changed services;
  - Manage the organisational structure involved in delivering services.

For the service catalogue the defined activities include:

- Initial Process Setup;
  - Define the structure and format of the service catalogue, and create an initial service catalogue based on the service portfolio;
  - Define a basic/default SLA valid for all services provided to customers, where no specific/individual SLA are in place;
  - Define templates for individual SLAs, OLAs and UAs;
  - Identify the most critical supporting service components, and agree OLAs and UAs with those contributing to delivering services to customers;
  - Agree individual SLAs with customers for the most important/critical services.
- Ongoing process execution;
- Maintain the service catalogue;
- Manage SLAs;
- Manage OLAs and UAs.

These processes for the Service Portfolio and Service Catalogue management have been implemented in VI-SEEM by developing a specialised service management component. The design model of the component takes into consideration similar approaches on service portfolio management and service catalogue that are being performed in projects such as EUDAT2020 and PRACE-4IP [15]. The information model of the component is presented in Fig. 4.1.

The central element of the information model is the service. Each service can be dependent on other VI-SEEM services, or external services. This relationship enables the construction of complex services that are designed as a composition of multiple elementary services. Each service is defined using a number of attributes that describe the services internally for the federation and externally for users. Service area and service type

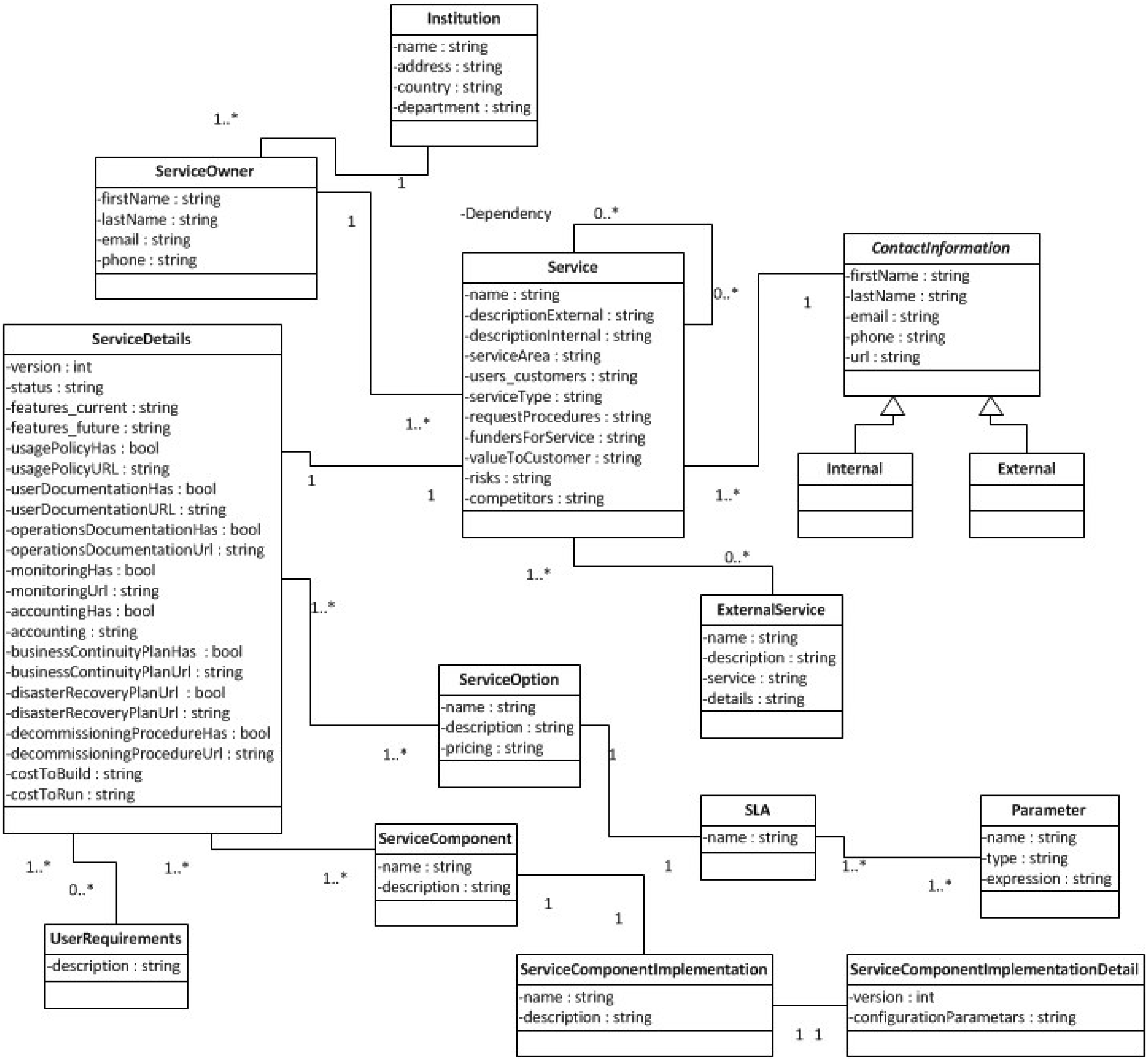


Fig. 4.1. *Information model for the VI-SEEM service catalogue and portfolio*

attributes are used for service grouping on the public UI. Each service is linked to a request procedure for the service. The service is associated with a service owner and service contact information (ex. help desk for the particular service). The service version is part of the service details attributes. The service status is used to filter the active services that are presented in the service catalogue. Additionally, attributes such as current and planned features, as well as, the policies, user documentation, and management information (monitoring, accounting, business continuity plan, disaster recovery) links are stored for each service version.

The service is built using multiple service components that represent the elementary resources needed to run the service. Each service component can have multiple service component implementations represented with different versions and parameters. In example, the VI-SEEM simple data storage service version 1.0 is designed using the File hosting component implemented using OwnCloud version 8.1. Each service can be offered with different service options, where options represent the information about value models, and service levels and

Table 4.2
Roles and responsibilities

Role	**Responsibilities**
Service Portfolio/Catalogue Process Owner	To control the SPM and SLM processes, maintain the catalogue and portfolio and report to senior management
Service Technical Coordinator / Architect	Has the overall view of services being developed or operated in the organization from the technical point of view
Customer Relationship Manager	Gathers requests for new features from feature / service requestor, Initiates a new service / service change to the service portfolio, identifies services that need decommissioning
Service Portfolio Approval Committee	Review and approves new services or changes to services
Service Owner	Has the overall responsibility for one specific service which is part of the service portfolio, Acts as the primary contact point for all (process-independent) concerns in the context of that specific service
Service Design Team	The team that is responsible for the design, implementation and maintenance of a service

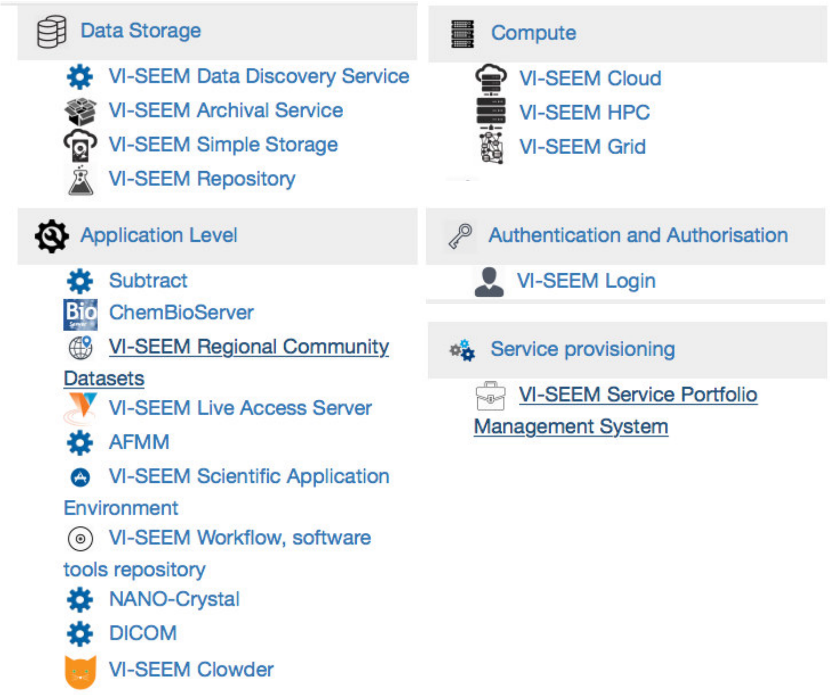
associated parameters.

The definition of roles and responsibilities within the SPM/SLM processes is presented in Table 4.2.

4.2.3. Implementation / Current status. Fig. 4.2. illustrates the current state of the the service catalog, offering production services to the scientific communities. The services are grouped in 5 service areas: data storage, application level, computer, authentication and authorization and service provisioning. The data storage services include simple storage, archival, repository and data discovery services, covering the the full lifecycle of relevant research data. The services in the application level area include domain specific services, such as the ChemBioServer [16] for the Life science, LAS [17] for the Climate and Clowder [18] for the Digital Cultural Heritage scientific community. The computing services, including grid, HPC and cloud resources are enlisted under the computer service area.

5. Related work. IT service management has been recognized as the preferred methodology for the operations and sustainability of the e-Infrastructures. In [20], the authors propose a novel methodology for creating an assessment process of the capabilities of service management systems in federated e-Infrastructures, useful for introduction or improvement of service management in the relevant domains. Various IT service management standards and/or best practices are successfully engaged to enhance the service capabilities of the exiting or future e-Infrastructures. ITIL was successfully applied in the merging process of two large e-Infrastructures, as shown in [21]. The PL-Grid infrastructure employed ITSM to maximize the efficiency of its operations [22]. By following the FitSM standard, their federated infrastructure is able to implement all required processes and serve the scientific community in highly efficient manner.

6. Conclusions. The growing complexity of the research e-Infrastructures and various resources available have driven the need to support the next generation of researchers by providing them with a consolidated access to all resource at their disposal. Implementing a virtual research environment portal on top of the underlying resources provides the researchers with an easy access to the tools, data sets and workflows of interest, enabling them to focus on the research process alone. However, such facilities are typically difficult to sustain in the long term, particularly if they are focused on a small set of users. The VI-SEEM project is a cross-domain federation of e-Infrastructures from the South Eastern Europe and Eastern Mediterranean region that aims to support



Data Storage
- VI-SEEM Data Discovery Service
- VI-SEEM Archival Service
- VI-SEEM Simple Storage
- VI-SEEM Repository

Compute
- VI-SEEM Cloud
- VI-SEEM HPC
- VI-SEEM Grid

Application Level
- Subtract
- ChemBioServer
- VI-SEEM Regional Community Datasets
- VI-SEEM Live Access Server
- AFMM
- VI-SEEM Scientific Application Environment
- VI-SEEM Workflow, software tools repository
- NANO-Crystal
- DICOM
- VI-SEEM Clowder

Authentication and Authorisation
- VI-SEEM Login

Service provisioning
- VI-SEEM Service Portfolio Management System

Fig. 4.2. *Services offered in the VI-SEEM VRE service catalogue*

researchers from three different communities: life sciences, climate, and digital cultural heritage and support their interdisciplinary activities seeking to lower the barrier to access and use complex e-Infrastructures.

This paper explored the different requirements and aspects of building a successful and sustainable federated virtual research environment. We indicated the different challenges and solution aspects of our VRE solution proposal, focusing on the service orientation paradigm and its importance for the end users. Another important aspect that was considered especially in terms of future growth and sustainability is the possibility to extend the offered services in a higher level environment by using open interoperable APIs.

The paper focused on the implementation of services offered within the VRE with the service management aspects being the most crucial link in the process of the creation of the VRE. The workings of the service portfolio and catalogue that are offered and operated by the project have been discussed. The complete service management approach is based upon the FitSM framework, where we presented the core processes and components that were put in place. The detailed description of the design of the service catalogue and portfolio can be used as a blueprint for building a service management system based on a federated approach that incorporates all service management aspects and clearly defines the roles and responsibilities of the federation members.

Acknowledgments. This work was supported by the European Unions Horizon 2020 research and innovation program, project Virtual Research Environment for Regional Interdisciplinary Collaboration in Southeast Europe and Eastern Mediterranean VI-SEEM [675121].

REFERENCES

[1] Giannoutakis KM, Tzovaras D. *The European Strategy in Research Infrastructures and Open Science Cloud.* International Conference on Data Analytics and Management in Data Intensive Domains 2016 Oct 11 (pp. 207-221). Springer, Cham.

[2] Andronico G, Ardizzone V, Barbera R, Becker B, Bruno R, Calanducci A, Carvalho D, Ciuffo L, Fargetta M, Giorgio E, La Rocca G. *E-Infrastructures for e-science: a global view.* Journal of Grid Computing. 2011 Jun 1;9(2):155-84.

[3] Foster I. *The grid: A new infrastructure for 21st century science.* Grid Computing: Making the Global Infrastructure a Reality. 2003 Mar 1;51.

[4] Alfieri, Roberto, et al. *VOMS, an authorization system for virtual organizations.* Grid computing. Springer Berlin Heidelberg, 2004.

[5] Beloglazov, Anton, Jemal Abawajy, and Rajkumar Buyya. *Energy-aware resource allocation heuristics for efficient man- agement of data centers for cloud computing.* Future generation computer systems 28.5 (2012): 755-768.

[6] Candela L, Castelli D, Pagano P. *Virtual research environments: an overview and a research agenda.* Data Science Journal. 2013;12:GRDI75-81.

[7] VI-SEEM project, `https://vi-seem.eu/`, 30 12 2017.

[8] Reimer TF, Carusi A. *Virtual research environment collaborative landscape study.*

[9] Probst J. *Anatomy of a Service.* Toronto ON Canada: Pink Elephant. 2009 Sep.

[10] Erl T.*Service-oriented architecture: concepts, technology, and design.* Pearson Education India; 2005.

[11] Galup SD, Dattero R, Quan JJ, Conger S. *An overview of IT service management.* Communications of the ACM. 2009 May 1;52(5):124-7.

[12] ITIL: IT Service Management books, `http://www.itil.org.uk/`, 30 12 2017.

[13] The FitSM Standard, `http://fitsm.itemo.org/`, 30 12 2017.

[14] EUDAT2020 Research Data Services, Expertise & Technology Solutions, `https://www.eudat.eu/`, 30 12 2017.

[15] PRACE Fourth Implementation Phase (PRACE-4IP) project `http://www.prace-ri.eu/` 30 12 2017.

[16] ChemBioServer `http://chembioserver.vi-seem.eu/` 30 12 2017.

[17] Live Access Server `http://las.vi-seem.eu/` 30 12 2017.

[18] Clowder `http://dchrepo.vi-seem.eu/` 30 12 2017.

[19] eInfra Central project `http://einfracentral.eu` 30 12 2017.

[20] Serrat J, Szepieniec T, Belloum A, Rubio-Loyola J, Appleton O, Schaaf T, Kocot J. *gSLM: The Initial Steps for the Specification of a Service Management Standard for Federated e-Infrastructures.* InProceedings of the 16th International Conference on Information Integration and Web-based Applications & Services 2014 Dec 4 (pp. 529-536). ACM.

[21] Marten H, Koenig T.*ITIL and Grid services at GridKa.* InJournal of Physics: Conference Series 2010 (Vol. 219, No. 6, p. 062018). IOP Publishing.

[22] Radecki M, Szymocha T, Szepieniec T, Roazska R.*Improving PL-Grid Operations Based on FitSM Standard.* In eScience on Distributed Computing Infrastructure 2014 (pp. 94-105). Springer International Publishing.

Edited by: Aneta Karaivanova
Received: Jan 8, 2017
Accepted: Mar 20, 2018

DOI 10.12694/scpe.v19i2.1396
Scalable Computing: Practice and Experience ISSN 1895-1767
Volume 19, Number 2, pp. 215–221. http://www.scpe.org © 2018 SCPE

VI-SEEM DREAMCLIMATE SERVICE

DUŠAN VUDRAGOVIĆ[*] LUKA ILIĆ[†] PETAR JOVANOVIĆ[‡] SLOBODAN NIČKOVIĆ[§] ALEKSANDAR BOGOJEVIĆ[¶] AND ANTUN BALAŽ[‖]

Abstract. Premature human mortality due to cardiopulmonary disease and lung cancer is found in epidemiological studies to be correlated to increased levels of atmospheric particulate matter. Such negative dust effects on the human mortality in the North Africa – Europe – Middle East region can be successfully studied by the DREAM dust model. However, to assess health effects of dust and its other impacts on the environment, a detailed modelling of the climate for a period of one year in a high-resolution mode is required. We describe here a parallel implementation of the DREAM dust model, the DREAMCLIMATE service, which is optimised for use on the high-performance regional infrastructure provided by the VI-SEEM project. In addition to development and integration of this service, we also present a use-case study of premature mortality due to desert dust in the North Africa – Europe – Middle East region for the year 2005, to demonstrate how the newly deployed service can be used.

Key words: DREAM model, dust effects, human mortality, VI-SEEM project, application service

AMS subject classifications. 68W10, 68M14, 68N30

1. Introduction. Exposure to airborne mineral dust particles can significantly influence human health. Atmospheric dust particles are primarily driven by mesoscale and synoptic processes, and may be present in high concentrations near the sources and carried over long distances while having adverse health effects. Drought and desertification, as climate-related changes and human activities such as changes in land use, affect potential dust sources of fine particulate matter in arid areas. Therefore, numerical modelling with sufficiently high resolution of the processes of the atmospheric dust cycle that drive dust emissions and transport is a useful approach to assessment of the potential health effects of exposure to dust.

The previously developed Dust REgional Atmospheric Modeling (DREAM) system [1] is a component of a comprehensive atmospheric model designed to simulate and predict the atmospheric cycle of mineral dust aerosols. The DREAM provides a climatology of dust based on long-term re-analysis of the model. It is widely used by the research and operational dust forecasting communities in more than 20 countries, including its recent use in a series of NASA-funded projects [2, 3, 4, 5] dealing with health aspects of dust suspended in the air. The Institute of Physics Belgrade group, which is a partner in the Sand and Dust Storm Warning Advisory and Assessment System (SDS-WAS) project of the World Meteorological Organization, uses DREAM to provide daily dust forecasts to the SDS-WAS model inter-comparisons and validation activities. Also, it is used for investigation on how fine particulate matter contributes to air pollution in North Africa – Europe – Middle East region.

To assess health effects of dust in the region and other dust impacts on the environment, it is usual to consider at least a one-year modelling climatology for the given region. In this case this was achieved by solving the DREAM model in a high-resolution mode with the horizontal grid resolution of 15 km. Such a high resolution model is capable to accurately describe the behaviour of small-scale dust sources in the desert areas (Sahara, Middle East), as well as the mesoscale atmospheric conditions. However, due to numerical complexity it requires a parallelised version of the DREAM code, which we created and optimised for usage on high-performance computing infrastructures available today.

[*] Scientific Computing Laboratory, Center for the Study of Complex Systems, Institute of Physics Belgrade, University of Belgrade, Serbia (dusan.vudragovic@ipb.ac.rs).

[†] Environmental Physics Laboratory, Institute of Physics Belgrade, University of Belgrade, Serbia (luka.ilic@ipb.ac.rs).

[‡] Scientific Computing Laboratory, Center for the Study of Complex Systems, Institute of Physics Belgrade, University of Belgrade, Serbia (petar.jovanovic@ipb.ac.rs).

[§] Environmental Physics Laboratory, Institute of Physics Belgrade, University of Belgrade, Serbia (slobodan.nickovic@ipb.ac.rs).

[¶] Scientific Computing Laboratory, Center for the Study of Complex Systems, Institute of Physics Belgrade, University of Belgrade, Serbia (aleksandar.bogojevic@ipb.ac.rs).

[‖] Scientific Computing Laboratory, Center for the Study of Complex Systems, Institute of Physics Belgrade, University of Belgrade, Serbia (antun.balaz@ipb.ac.rs).

In parallel to development of the DREAM model, a number of initiatives were crucial for enabling high-quality climate research in the region. This was achieved by providing e-Infrastructure resources, application support and training through the VI-SEEM project [6], funded by the EU H2020 programme. The project brings together regional e-Infrastructures in order to build capacity and better utilise synergies, as well as to provide improved service within a unified virtual research environment for several inter-disciplinary scientific user communities. The overall aim is to offer a user-friendly integrated e-Infrastructure platform for regional cross-border scientific communities in climatology, life sciences, and cultural heritage. This includes integration of computing, data, and visualisation resources, as well as services, models, software solutions and tools. The VI-SEEM virtual research environment provides the support to scientists in a full lifecycle of collaborative research.

By efforts of the DREAM code developers and the VI-SEEM support team, the DREAM model was successfully refactored and tuned for usage on high-performance computing infrastructures in a form of the DREAMCLIMATE service, presented here. Section 2 briefly describes the DREAM model, which is capable of producing results in the required high-resolution mode for a one year period. The DREAMCLIMATE service is presented in detail in Section 3, while Section 4 describes produced datasets and main results. By using an order of magnitude finer DREAM model grid than available before, we perform a detailed analysis of dust impacts to public health.

2. DREAM model. Premature human mortality due to cardiopulmonary disease and lung cancer is found in epidemiological studies to be correlated to increased levels of atmospheric particulate matter, in particular to long-term exposure to particulate matter with an aerodynamic diameter smaller than $2.5\,\mu$m. In order to estimate the premature mortality caused by the long-term exposure to airborne desert dust, we use results of the DREAM gridded model dust climatology of fine particulate matter and dust concentrations. This analysis follows the previous study [7] that indicates that there is a large number of premature deaths by cardiopulmonary disease and a significant number of deaths by lung cancer, mostly in the dust belt region neighbouring Sahara and Middle East deserts.

The DREAM model is developed as an add-on component of a comprehensive atmospheric model and is designed to simulate and/or predict the atmospheric cycle of mineral dust aerosols. It solves a coupled system of the Euler-type partial differential nonlinear equations for dust mass continuity, one equation for each particle size class, which is one of the governing prognostic equations in an atmospheric numerical prediction model [8, 9, 10]. The DREAM model takes into account all major processes of the atmospheric dust cycle. During the model simulation, calculation of the surface dust emission fluxes is made over the model cells declared as deserts. A viscous sub-layer parameterisation regulates the amount of dust mass emission for a range of near-surface turbulent regimes. Once injected into the air, dust aerosols are driven by the atmospheric dynamics and corresponding physical quantities: by turbulence in the early stage of the process, when dust is lifted from the ground to the upper levels; by winds in later phases of the process, when dust travels away from the sources; and finally, by thermodynamic processes, rainfall and land cover features that provide wet and dry deposition of dust over the Earth surface.

The model is implemented as a bundle of Fortran programs and libraries. These components are divided into three groups: the preprocessing system, the model operational system, and post-processing and visualisation tools. The preprocessing consists of two phases. The first is the setup in which the simulation domain, model configuration and interpolation of terrestrial data are defined. These parameters are mostly hard-coded and any change to parameters in this phase requires recompilation. The second stage of preprocessing is interpolation of the meteorological input data from the global meteorological model to the current simulation domain, as well as a setup of initial boundary conditions for the dust model. The model operational system is the main component, and it runs the numerical integration program. Post-processing and visualisation tools include GrADS [11] with conversion from Arakawa E-grid to geo-referenced grid and plots.

The code is predominantly written in the style of the Fortran 77 standard. Some of the more pressing constraints of the standard were the lack of support for dynamic memory allocation and command line arguments. These two constraints required for a number of parameters to be hard-coded. As a consequence, this limited the number of users who could use the application independently, and the number of parallel tests that could be ran at once. Recompilation also requires a deep technical knowledge of the implementation itself, which reduces usability and dissemination of the model.

3. DREAMCLIMATE service. Within the framework of the VI-SEEM project, the DREAM model was successfully re-factored and tuned for usage on high-performance computing infrastructures. The DREAMCLIMATE service was developed and deployed using the VI-SEEM infrastructure modules. Configuration of the considered physical system is separated from the source code of the application, and all relevant parameters are grouped into a single configuration file. Such an improved configuration approach enabled more user-friendly way to configure various model setups, without the need for each user to dive into the code and technical details of the implementation. This also enables multiple users to run their model instances independently. Important additional improvements include significant reduction of the disk-space consumption, as well as standardisation of its usage through an environment-module approach.

Configuration files follow the format of the Python configuration parser, which is a convenient, flexible, and powerful way for parsing configuration files. It uses simple INI style configuration syntax, i.e., a text file with a basic structure composed of sections, properties, and values. Parameters are divided into sections which are designated by square brackets. Within one section, each parameter is specified in a separate line and its name and value are delimited by the equals sign. In-line comments are also permissible and corresponding lines begin with a semicolon. In addition to this, a support for variable interpolation is included as well.

The DREAM processing stages remain similar to the original version of the code, and consist of the preprocessing, the model operational processing, and the post-processing phase. Majority of changes are related to reducing the complexity of configuration in the setup stage of preprocessing. In a typical use-case, a user begins the simulation project by loading the environment module for the DREAMCLIMATE service, which sets the environment paths for the commands used to initialise and prepare the DREAM model simulation. Afterwards, by invoking the `dreamclimate_init` script the default configuration file is created in the working directory and files needed for a configuration of the local simulation instance are created in the `.dreamclimate` subdirectory. After the parameters are set in the configuration file, the `dreamclimate_reconfig` script is called to execute the setup stage, which encapsulates recompilation of the components, depending on the parameters changed. The resulting binaries, which are used to run simulation, are placed in the `.dreamclimate/bin` directory. This step isolates each user's simulation instance from others and enables multiple instances to work without interference. The next step in this stage generates and interpolates vegetation and soil texture for the forecast domain, by calling the `gt30mounth`, `gt30source`, `gt30vegetadirect`, `text4eta`, and `texteta` components.

After the setup, preprocessing continues by invoking the `dreamclimate_preproc` script whose role is to prepare input data for the Eta model grid. This script invokes the following components:

- `climsst` – horizontal grid (IMT, JMT) Eta model indexing from the SST as a function of the month,
- `anecw` – horizontal grid (IMT, JMT) Eta model indexing from global initial data,
- `pusiWRF` – set of the vertical variables and vertical interpolation of the pressure to sigma surfaces,
- `const` – conversion of the initial fields in Eta model coordinates from 2D horizontal (IMT, JMT) indexing to 1D (IMJM), definition of dummy initial boundary soil moisture and temperature values, and calculation of the constants needed for the 1D version of the soil model,
- `dboco` – creation of the boundary condition files,
- `gfdlco2` – interpolation of the transmission functions grid, for which the transmission functions have been pre-calculated, to the grid structure.

This preprocessing step produces binary files interpolated to the model grid (i.e., Arakawa E-grid) in the output directory specified in the configuration file. All the routines of the model itself, which describe atmospheric processes including the dust cycle, are built into the main executable file. This is a parallel MPI program that runs the simulation and is submitted to the job scheduling system using the job description script, which is automatically generated earlier in the setup stage. The post-processing includes the conversion of the main GrADS output file from the Arakawa E-grid to the GrADS grid. These steps are handled by the `dreamclimate_post-process` script.

Many of the configuration parameters in the generated configuration file have sensible default values, to minimise the need for users to search through lengthy lists of output file locations. The domain parameters of interest for configuring the model itself, inside the ALLINC section, are:

- TLM0D – longitude of the centre point of the domain,

- TPH0D – latitude of the centre point of the domain,
- WBD – western boundary of the domain with respect to the centre point (always less than 0),
- SBD – southern boundary of the domain with respect to the centre point (always less than 0),
- DLMD – longitudinal model grid resolution,
- DPHD – latitudinal model grid resolution,
- DTB – time step of the model, which depends on DLMD and DPHD values by means of the Couranf-Friedrichs-Lewy (CFL) criteria,
- LM – the number of vertical levels.

Another set of commonly changed model parameters are dimensions of the model grid. These are grouped in the PARMETA section of the configuration file:

- IM – the number of mass grid points along the first row, essentially half of the total number of grid points in the west-east direction, due to the horizontal staggering of mass and wind points,
- JM – the number of rows in the north-south direction.

These parameters also influence the number of processes and the topology of the MPI parallel execution.

The rest of the parameters in the configuration file specify paths for input, output and intermediate files. With these paths defined during configuration, a significant reduction in disk space usage was achieved, as the data files no longer need to be copied together with the code, and no longer have to be in fixed relative locations.

The DREAMCLIMATE service is deployed during the first VI-SEEM development access call at the PARADOX high-performance computing cluster [12], hosted by the Scientific Computing Laboratory, Center for the Study of Complex Systems of the Institute of Physics Belgrade. This cluster is part of the VI-SEEM infrastructure, and consists of 106 working nodes. Working nodes (HP ProLiant SL250s Gen8) are configured with two Intel Xeon E5-2670 8-core Sandy Bridge processors, at a frequency of 2.6 GHz and 32 GB of RAM. The total number of CPU-cores available in the cluster is 1696, and each working node contains an additional GP-GPU card (NVIDIA Tesla M2090) with 6 GB of RAM. The peak computing power is 105 TFlops. The PARADOX provides a data storage system, which consists of two service nodes (HP DL380p Gen8) and 5 additional disk enclosures. One disk enclosure is configured with 12 SAS drives of 300 GB each (3.6 TB in total), while the other four disk enclosures are configured each with 12 SATA drives of 2 TB (96 TB in total), so that the cluster provides around 100 TB of storage space. Storage space is distributed via a Lustre high-performance parallel file system that uses Infiniband QDR interconnect technology, and is available on both working and service nodes.

Although the DREAMCLIMATE code is a copyright-protected software, it can be obtained for research purposes with the permission of the principal investigator (S. Ničković). Therefore, the DREAMECLIMATE service source code is only internally available at the VI-SEEM code repository [13], as well as a module at the PARADOX cluster software repository. Transfer of the software to third parties or its use for commercial purposes is not permitted, unless a written permission from the author is received.

4. Produced datasets and results. Using the DREAMCLIMATE service at PARADOX during the first VI-SEEM call for production use of resources and services, we produced a dataset with the aerosol optical thickness and surface dust concentration for the one-year period. We selected the year 2005 for this analysis, which serves as an example and demonstrates usability of DREAMCLIMATE service. The dataset covers wide region of North Africa, Southern Europe and Middle East in 30 km horizontal resolution with 28 vertical levels, and is made publicly available via the VI-SEEM data repository [14].

In addition to this initial dataset, we also produced a dataset with a higher resolution of 15 km for the same region and period of time. The global mean DREAMCLIMATE-modelled dust concentration for year 2005 is presented in Fig. 4.1.

Using the human health impact function introduced in Refs. [15, 16], we can relate the changes in pollutant concentrations to the changes in human mortality, and estimate the global annual premature mortality due to airborne desert dust. For this, we use as a baseline the mortality rate estimated by the World Health Organization (WHO) Statistical Information System on the country-level based on the International Classification of Diseases 10th Revision (ICD-10) classification, and regional data from the WHO Global burden of disease for countries with no data. Population statistics we used for the year 2005 is based on the United Nations Department of Economic and Social Affairs (UNDES 2011) database, while gridded global population numbers

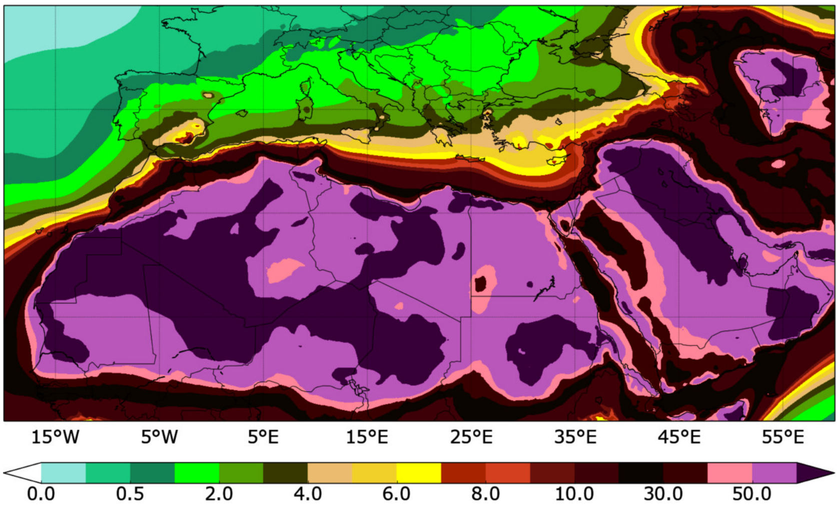


FIG. 4.1. *Calculated mean dust concentrations in* μg/m^3*, obtained from the DREAMCLIMATE model. The model integration area covers region of North Africa, Southern Europe and Middle East, with 15 km horizontal resolution in 28 vertical levels for the year 2005.*

TABLE 4.1
Total CPD and LC premature mortalities for the threshold concentrations between 0 and 10 μg/m^3.

Baseline concentration (in μg/m^3)	0	5.0	7.5	10
CPD premature mortality (in thousands)	765	615	567	524
LC premature mortality (in thousands)	14.8	10.2	9.1	8.4

are taken from the Columbia University Center for International Earth Science Information Network (CIESIN) database. We used the population cohort of 30 years and older in the health impact function.

Applying the health impact function to the considered population, the DREAM model output suggests a significant contribution of desert dust to premature human mortality. For the global background of dust concentration of 7.5 μg/m^3 i.e., threshold below which no premature mortality occurs, the estimated premature mortality (per grid cell) by cardiopulmonary disease (CPD) and lung cancer (LC) is illustrated in Fig. 4.2. In total, around 570,000 premature deaths in the model domain are predicted to occur during a one-year period, as a negative consequence of dust. According to our results, top five countries with the highest induced CPD-mortality in the year 2005 are: Egypt with 74,000; Iraq with 67,000; Iran with 50,000; Nigeria with 46,000; Sudan with 45,000. On the other hand, top five countries with the highest induced LC-mortality in the same year are: Iraq with 1,200; Iran with 900; Sudan with 800; Egypt with 800; Uzbekistan and Turkey with 500 premature deaths each.

We also investigated the sensitivity of our results on the value of the threshold concentrations, which is above assumed to be 7.5 μg/m^3. Table 4.1 gives the obtained total CPD and LC premature mortalities for the threshold concentrations between 0 and 10 μg/m^3. This analysis is presented to showcase capabilities of the model and the developed DREAMCLIMATE service, and can be efficiently used to study desired regions and time periods if the required input data are provided.

5. Conclusions. Using the VI-SEEM project infrastructure and services, we have successfully re-factored the DREAM atmospheric model. We have developed and implemented the DREAMCLIMATE service, which is tuned for usage on high-performance computing infrastructures available today. In order to demonstrate a typical use-case, we have produced a dataset with the aerosol optical thickness and surface dust concentration for the one-year period for the wide region of North Africa, Southern Europe and Middle East. We have used

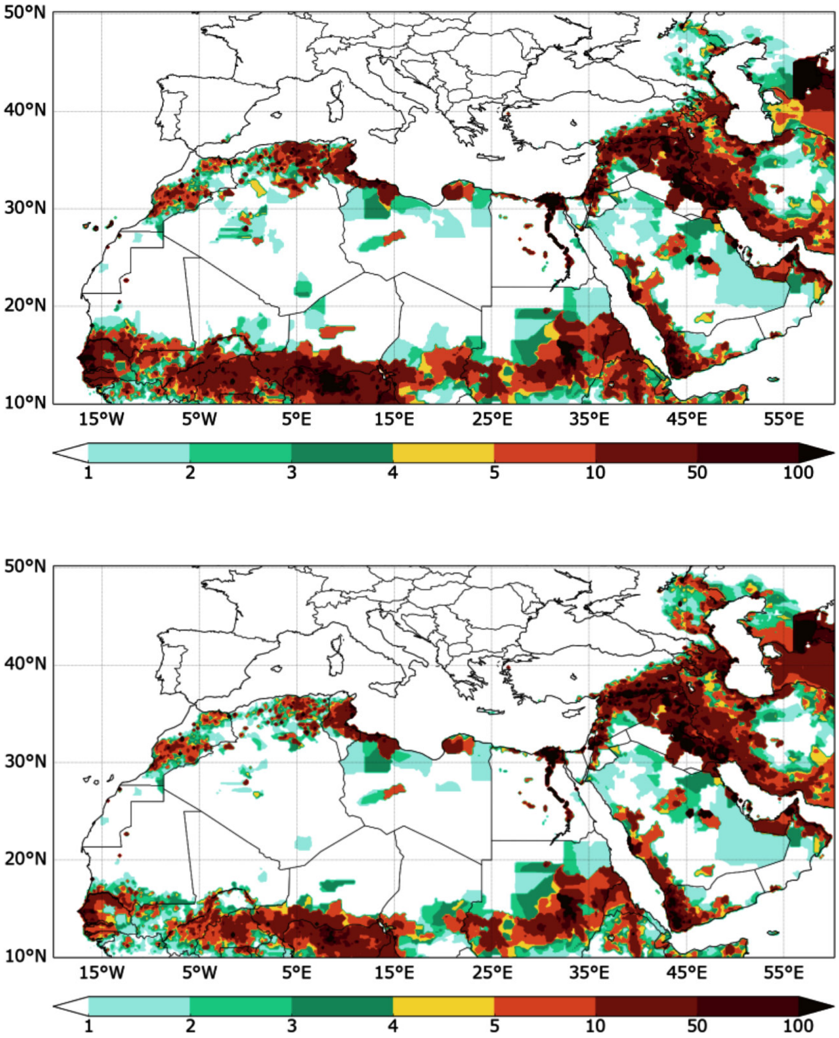


FIG. 4.2. *Estimated global premature mortality per grid cell by cardiopulmonary disease (top) and lung cancer (down) due to the long-term exposure to desert dust with an aerodynamic diameter smaller than 2.5* μm, *calculated by the VI-SEEM DREAMCLIMATE service.*

both the 30 km and the 15 km horizontal resolution, with 28 vertical levels. To showcase how results of the DREAMCLIMATE service can be applied, using the human health impact function and calculated global fine particulate matter concentrations, we have estimated the premature mortality caused by the long-term exposure to airborne desert dust with an aerodynamic diameter smaller than 2.5 μm for the year 2005 in the considered region. The results show that the large total number of premature deaths (around 570,000) in the model domain is mainly due to cardiopulmonary disease, but a significant number of deaths is also caused by lung cancer. The model also shows high sensitivity of the results on the threshold concentration, which is a significant parameter of relevance to public health.

Acknowledgments. This work was supported by the Ministry of Education, Science, and Technological Development of the Republic of Serbia under projects ON171017 and III43007, and by the European Commission under H2020 project VI-SEEM, Grant No. 675121. Numerical simulations were run on the PARADOX supercomputing facility at the Scientific Computing Laboratory of the Institute of Physics Belgrade.

REFERENCES

[1] S. Ničković, G. Kallos, A. Papadopoulos and O. Kakaliagou, *A model for prediction of desert dust cycle in the atmosphere*, J. Geophys. Res., 106 (2001), pp. 18113-18129.

[2] D. Yin and W. A. Sprigg, *Modeling Airbourne Mineral Dust: A Mexico - United States Trans-boundary Perspective*, in Southwestern Desert Resources, W.Halvorson, C. Schwalbe and C. van Riper, eds., University of Arizona Press, Tucson, AZ, (2010), pp. 303–317.

[3] D. Yin, S. Nickovic and W. A. Sprigg, *The impact of using different land cover data on wind-blown desert dust modeling results in the southwestern*, Atmos. Environ., 41 (2007), pp. 2214–2224.

[4] W. A. Sprigg, S. Ničković, J. N. Galgiani, G. Pejanović, S. Petković, M. Vujadinović, A. Vuković, M. Dacić, S. DiBiase, A. Prasad and H. El-Askary, *Regional dust storm modeling for health services: The case of valley fever*, Aeolian Res., 14 (2014), pp. 53–73.

[5] A. Vuković, M. Vujadinović, G. Pejanović, J. Andrić, M. R. Kumjian, V. Djurdjević, M. Dacić, A. K. Prasad, H. M. El-Askary, B. C. Paris, S. Petković, S. Ničković and W. A. Sprigg, *Numerical simulation of "an American haboob"*, Atmos. Chem. Phys., 14 (2014), pp. 3211–3230.

[6] D. Vudragović, P. Jovanović and A. Balaž, *VI-SEEM Virtual Research Environment*, 10th RO-LCG Conference, Sinaia, Romania, 26-28 October 2017.

[7] D. Giannadaki, A. Pozzer and J. Lelieveld, *Modeled global effects of airborne desert dust on air quality and premature mortality*, Atmos. Chem. Phys., 14 (2014), pp. 957–968.

[8] Z. I. Janjić, *The Step-Mountain Coordinate: Physical Package*, Mon. Wea. Rev., 118 (1990), pp. 1429-1443.

[9] Z. I. Janjić, *The Step-mountain Eta Coordinate Model: Further developments of the convection, viscous sublayer and turbulence closure schemes*, Mon. Wea. Rev., 122 (1994), pp. 927-945.

[10] S. Ničković and S. Dobricić, *A model for long-range transport of desert dust*, Mon. Wea. Rev., 124 (1996), pp. 2537–2544.

[11] Grid Analysis and Display System (GrADS), http://cola.gmu.edu/grads/

[12] V. Slavnić, *Overview of Grid and High Performance Computing activities in Serbia*, 7th RO-LCG Conference, Bucharest, Romania, 3–5 November 2014.

[13] DREAMCLIMATE code at the VI-SEEM Code Repository, https://code.vi-seem.eu/petarj/dreamclimate

[14] DREAMCLIMATE datasets at the VI-SEEM Data Repository, https://repo.vi-seem.eu/handle/21.15102/VISEEM-86

[15] S. C. Anenberg, L. W. Horowitz, D. Q. Tongand and J. J. West, *An estimate of the global burden of anthropogenic ozone and fine particulate matter on premature human mortality using atmospheric modeling*, Environ. Health Perspect., 118 (2010), pp. 1189-1195.

[16] J. Lelieveld, C. Barlas, D. Giannadaki and A. Pozzer, *Model calculated global, regional and megacity premature mortality due to air pollution*, Atmos. Chem. Phys., 13 (2013), pp. 7023-7037.

Edited by: Aneta Karaivanova
Received: Jan 5, 2018
Accepted: Apr 12, 2018

AIMS AND SCOPE

The area of scalable computing has matured and reached a point where new issues and trends require a professional forum. SCPE will provide this avenue by publishing original refereed papers that address the present as well as the future of parallel and distributed computing. The journal will focus on algorithm development, implementation and execution on real-world parallel architectures, and application of parallel and distributed computing to the solution of real-life problems. Of particular interest are:

Expressiveness:

- high level languages,
- object oriented techniques,
- compiler technology for parallel computing,
- implementation techniques and their efficiency.

System engineering:

- programming environments,
- debugging tools,
- software libraries.

Performance:

- performance measurement: metrics, evaluation, visualization,
- performance improvement: resource allocation and scheduling, I/O, network throughput.

Applications:

- database,
- control systems,
- embedded systems,
- fault tolerance,
- industrial and business,
- real-time,
- scientific computing,
- visualization.

Future:

- limitations of current approaches,
- engineering trends and their consequences,
- novel parallel architectures.

Taking into account the extremely rapid pace of changes in the field SCPE is committed to fast turnaround of papers and a short publication time of accepted papers.

INSTRUCTIONS FOR CONTRIBUTORS

Proposals of Special Issues should be submitted to the editor-in-chief.

The language of the journal is English. SCPE publishes three categories of papers: overview papers, research papers and short communications. Electronic submissions are preferred. Overview papers and short communications should be submitted to the editor-in-chief. Research papers should be submitted to the editor whose research interests match the subject of the paper most closely. The list of editors' research interests can be found at the journal WWW site (`http://www.scpe.org`). Each paper appropriate to the journal will be refereed by a minimum of two referees.

There is no a priori limit on the length of overview papers. Research papers should be limited to approximately 20 pages, while short communications should not exceed 5 pages. A 50–100 word abstract should be included.

Upon acceptance the authors will be asked to transfer copyright of the article to the publisher. The authors will be required to prepare the text in $\LaTeX 2_\varepsilon$ using the journal document class file (based on the SIAM's `siamltex.clo` document class, available at the journal WWW site). Figures must be prepared in encapsulated PostScript and appropriately incorporated into the text. The bibliography should be formatted using the SIAM convention. Detailed instructions for the Authors are available on the SCPE WWW site at `http://www.scpe.org`.

Contributions are accepted for review on the understanding that the same work has not been published and that it is not being considered for publication elsewhere. Technical reports can be submitted. Substantially revised versions of papers published in not easily accessible conference proceedings can also be submitted. The editor-in-chief should be notified at the time of submission and the author is responsible for obtaining the necessary copyright releases for all copyrighted material.