Applications of the FACE-IT portal and workflow engine for operational food quality prediction and assessment: Mussel farm monitoring in the Bay of Napoli, Italy

Raffaele Montella
Uni. of Napoli Parthenope
University of Chicago
montella@uniparthenope.it

Alison Brizius
University of Chicago
abrizius@uchicago.edu

Diana Di Luccio
Uni. of Napoli Parthenope
diluccio@uniparthenope.it

Cheryl Porter
University of Florida
cporter@ufl.edu

Joshua Elliot
University of Chicago
elliot@uchicago.edu

Ravi Madduri
University of Chicago
madduri@uchicago.edu

David Kelly
Uni. of Napoli Parthenope
davidkelly@uchicago.edu

Angelo Riccio
Uni. of Napoli Parthenope
riccio@uniparthenope.it

Ian Foster
Argonne National Laboratory
University of Chicago
foster@anl.gov

ABSTRACT

Mussel farm product quality remains a challenging problem for operational marine science. In an operational scenario, the model chain, orchestrated in a workflow fashion, produces a huge amount of predicted spatially-referenced (big) data. These workflow components have been integrated into the Framework to Advance Climate, Economic, and Impact Investigations with Information Technology (FACE-IT), a workflow engine and data science portal based on Galaxy and Globus technologies. We describe how FACE-IT workflows can be used to couple many simulation/prediction models, leveraging high-performance and cloud computing resources to enable fast full system modeling in order to produce operational predictions about the impact of pollutants spilled out from both natural and anthropic sources in mussels farming high density areas.

CCS Concepts

•**Computing methodologies** → **Distributed algorithms;**
•**Applied computing** → *Environmental sciences;*

Keywords

workflows; large scale data science; data portal; mussel farms; food security; cloud computing; environmental modelling

1. INTRODUCTION

Projections of future food quality require data from climate models to forecast future conditions, coupling weather models, wind-driven sea waves models and ocean circulation/river advection models leveraging, transport and diffusion of pollutants on projections about future infrastructures, as new fishery and mussel farm installments, land-use and land cover [10]. Decisions are mainly made by coastal management planners in designing or re-designing human facilities, sea/lake fronts, ports, harbors and farms (fishery, mussels) placement using scenario analysis tools.

A system supporting this kind of decisions would require:

- the management of diffuse and point pollution sources;
- the ability to scale in terms of domain size;
- the computational performance and effectiveness needed to provide results for decision support.

We describe here a real world operational and on-demand application for mussel farm food quality prediction and assessment [5]. Users (both field scientists and food quality/human health managers and experts) interact with the FACE-IT Galaxy [11] data portal in order to evaluate the ongoing situation, generate alerts and depict future scenarios for strategic management (http://www.faceit-portal.org). This work could be considered an updated extension of a previous GPU accelerated high performance cloud computing infrastructure for a virtual environmental laboratory [4] carried out by a multidisciplinary and interdisciplinary research team.

The rest of this paper is as follows. §2 introduces the general FACE-IT infrastructure and how it has been developed in the context of agricultural modeling and food quality and extended in order to support the described application; §3 details the application workflow and how different models have been implemented to fit the proposed workflow infrastructure; §4 discusses computational and environmental issues as carried out from data analysis; and finally §5 presents conclusions and proposes future work.

2. THE FACE-IT INFRASTRUCTURE

We have previously developed the Framework to Advance Climate, Economic, and Impact Investigations with Information Technology (FACE-IT) for crop and climate impact assessments.

This integrated data processing and simulation framework enables data ingest from geospatial archives; data regridding, aggregation, and other processing prior to simulation; large-scale climate impact simulations with agricultural and other models, leveraging high-performance and cloud computing; and post-processing to produce aggregated yields

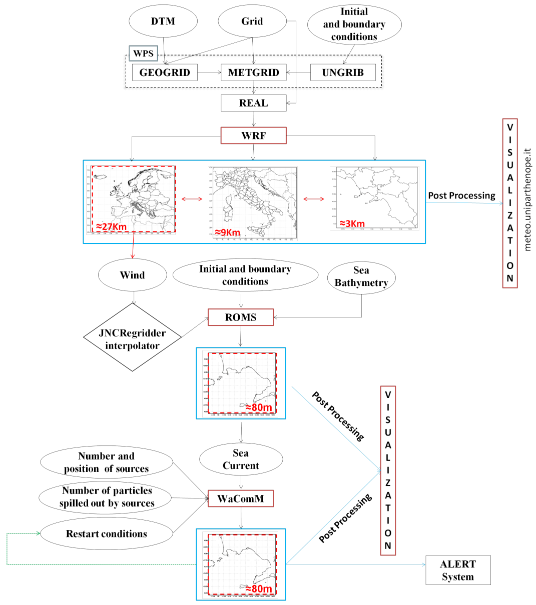

Figure 1: The computation process.

and ensemble variables needed for statistics, for model intercomparison, and to connect biophysical models to global and regional economic models. FACE-IT leverages the capabilities of the Globus Galaxies platform [7] to enable the capture of workflows and outputs in well-defined, reusable, and comparable forms.

The FACE-IT infrastructure is extended in order to support applications related to weather, sea wave conditions, ocean circulation and pollutant transport and diffusion.

FACE-IT Galaxy workflows provide a robust and effective integration of earth science features in Globus Galaxies in order to prepare a stable environment for virtually endless data workflow-based applications. Mainly leveraging on the datatype enforcement using the NetCDF Schema, the latest FACE-IT Galaxy improvements are focused on static and dynamical maps for data visualization, dedicated tool parameters and new data sources [10].

3. TOOLS AND WORKFLOWS

Improving the evaluation of pollution effects in aquatic ecosystems is important for economic profit and to improve environmental sustainability. To achieve this target we designed and the implemented WaComM, a three dimensional decision support model enabling the simulation and prediction of pollutant spills, transport and dispersion in both inshore and offshore environments.

WaComM is driven by a complex chain of outputs from observational data, weather and oceanic models. The computational process is shown in Figure 1. First, data for the region is used to initialize the Weather Research and Forecast (WRF) model. WRF is used to compute wind conditions that are one of the forcing of sea surface current forecasted by Regional Oceanic Modeling System (ROMS). The final result of WRF-ROMS coupled models is a hourly modeling simulation of the 3D hydrodynamic flow that we used as input data for WaComM to follow the pollutants Lagrangian transport.

The model chain has been integrated into FACE-IT Galaxy to be a qualitative and (semi) quantitative tool for human health risk assessment that could be helpful to achieve a better management of offshore human activities and a making decision support tool to select the best marine areas where these activities could be deployed (Figure 3).

3.1 Weather Research and Forecast

The Weather Research and Forecasting (WRF) [8] model is a next-generation mesoscale numerical weather prediction system designed for both atmospheric research and operational forecasting needs. The impressive diffusion of this tool as the base driver of most current weather/climate scenario analysis motivated our effort in implementing a FACE-IT Galaxy version of WRF. In order to fully support WRF within FACE-IT, we developed the following tools:

1. **Make WRF Experiment** is used to define a domain entering its center by latitude and longitude. The user can choose the number of domains for data preparation from 1 to 4. The space and temporal rate for each domain is 1:3. The number of domains used for preparation could differ by the two-way nested domains computed by the WRF module. This option enable the user to mix two-way nesting and nest-down based nesting in the same workflow.

2. **Get NCEP from WRF Experiment** downloads the initial and boundary conditions from the services made available by the National Center for Environmental Protection. In particular the data produced by the Global Forecast System (GFS) [13] at the resolution of 0.5 degrees are used for WRF initialization.

3. **GeoGrid** wraps the WRF software component with the same name. Its main duty is preparing the geographic domains extracting static data (i.e. elevation, albedo, land use) from a database accordingly with the WRF experiment.

4. **UnGrib** wraps the WRF software component with the same name. UnGrib decode the GFS data in an intermediate file used by other tools.

5. **MetGrid** shares the name with its regular WRF counterpart which is the wrap. It interpolates data produced by the UnGrib tool on the domains produced by GeoGrid.

6. **Real** prepares the data produced by MetGrid for the real case simulation. It is responsible of the creation of the boundary and initial conditions for the WRF tool.

7. **WRF** wraps the core model. In the current implementation it relays on a only OpenMP enabled WRF compiled binary or on a hierarchical parallelism enabled on MPI distributed memory, OpenMP shared memory and GPGPUs for some model components.

8. **WRF Aggregator** aggregates WRF outputs of a specified domain in a single NetCDF file remapped on a equal spaced regular latitude longitude grid.

9. **WRF Plot:** it is used for simply, fast and reliable WRF output plotting.

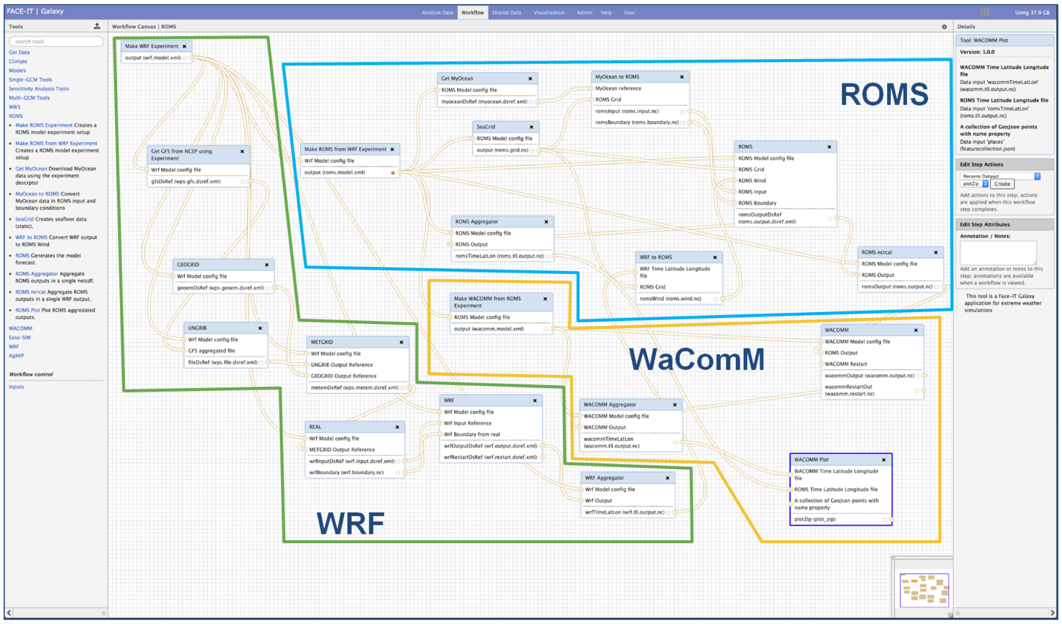

Figure 2: The workflow in FACE-IT.

Integration of WRF into the FACE-IT Galaxy framework completely abstracts the model setup complexity giving allowing field scientists to concentrate on experiment creation. We have developed a tutorial to disseminate the use of WRF/FACE-IT.

3.2 Regional Ocean Model System

Nevertheless previous experiences with the Princeton Ocean Model (POM) parallelization [3], aiming the implementation of the FACE-IT workflow application on mussel farms, we chose ROMS, a free-surface, terrain-following, primitive equations ocean model widely used by the scientific community for a diverse range of applications [15].

We developed the following tools:

1. **Make ROMS from WRF Experiment** reads a WRF experiment description dataset and produces a ROMS experiment description dataset. The ROMS experiment is created to be compliant with the WRF domains, initial and boundary conditions and the simulation starting time and duration.

2. **Get MyOCEAN:** While we are able to download initial and boundary conditions for the WRF model from NCEP/GFS, there is no analogous worldwide service for ocean modeling. Because our mussel farms application is focused on the central Tyrrhenian sea east sector, we download data from the services provided by the Copernicus European Project for marine environment monitoring.

3. **Sea Grid** is a toolbox backend that we developed to emulate the behavior of WRF's GeoGrid module. SeaGrid produces the digital bathymetry model of the computation domain defined by the ROMS experiment descriptor dataset.

4. **MyOCEAN to ROMS** acts in the same shape of WRF's MetGRID module performing a regridding and eventually an interpolation of MyOCEAN data [1] on the actual ROMS domains.

5. **WRF to ROMS:** While the initial and boundary conditions provided are provided by the Copernicus project with the "OCEAN to ROMS" tool, the wind friction data is gathered directly from WRF outputs. The wind field and other variables are regridded and eventually interpolated on the actual ROMS domains.

6. **ROMS** is a wrapper for the Regional Ocean Model System executable. For our application on the Bay of Napoli's mussel farms we compiled ROMS using the shared memory OMP-based setup.

7. **ROMS ncrcat** wraps the NCO NCRCAT utility to concatenate the NetCDF files produced by ROMS.

8. **ROMS Aggregator** aggregate ROMS outputs in a single NetCDF file remapped on a equal spaced regular latitude longitude grid.

3.3 Water quality Community Model

WaComM is an evolution of the Lagrangian Assessment for Marine Pollution 3D (LAMP3D) model [2] [6]. We strongly optimized the algorithms thanks to the development of i) deep-algorithms optimization to increase efficiency and effectiveness on High Performance Computing (HPC) environment (X86_64 and IBM P8 architectures) and ii) the implementation of novel restart feature to calculates the amount of pollutants in the water taking into account the residual particles (since the last run) and the particles released from the sources (current run). This is needed in order to have a realistic simulation.

We developed the following tools:

1. **Make WaComM experiment from ROMS** reads a ROMS experiment description dataset and produces

a WaComM experiment description dataset. The WaComM experiment is created to be compliant with the ROMS domains, initial and boundary conditions and the simulation starting time and duration.

2. **WaComM** wraps the WaComM executable. At the present it support restarts and shared memory parallel implementation based on OpenMP. This tool produces a WaComM output as a NetCDF-based datatype and a comma-separated-values file with the particles status. This file could be used as input in order to implement restarting.

3. **WaComM Aggregator** aggregates WaComM outputs of a specified domain in a single NetCDF file remapped on a equal spaced regular latitude longitude grid.

4. **WaComM Plot** is used for simply, fast, and reliable WaComM output plotting.

4. COMPUTATIONAL ENVIRONMENT

The FACE-IT application on mussel farms modeling for food quality assessment and human health protection have practical implications from both computational and environmental point of view.

4.1 Deployment scenarios

The FACE-IT application described in this paper is computationally intensive, involving loosely coupled parallelization at the workflow level and tight coupled parallelization at the tool level. The need for an operational system drove us to design and implement two different deployment scenarios.

The first deployment scenario, **FACE-IT Amazon Web Services**, is that normally used by FACE-IT. It involves EC2 machines onto which tasks are scheduled by HTCondor integrated by an ad-hoc daemon monitoring the Condor queue. This daemon analyzes the ClassAd requirements for each submitted job and manage the instantiation of the needed computation resources.

The second deployment scenario is a **Dedicated HPC Cluster**. FACE-IT Galaxy can be deployed on dedicated HPC clusters using the DRMAA interface to Torque if the system local scheduler match the interface requirements. If this is not the case of the current deployment, we developed a custom job runner interacting with the local scheduler with a custom implementation of the queue management. We used this approach successfully providing HPC resources to a distributed / shared memory and GPU accelerated version of WRF. This approach also supports shared memory parallel versions of ROMS and WaComM.

4.2 Application results

Weather conditions influence the transport of pollutants near mussel farms. Our case study focuses on mussel farms in the Punta Terone (between Capo Miseno and Punta del Poggio) and Centocamerelle (Gulf of Pozzuoli, Campania Region) areas. These mussel farms are classified as “type A” in accordance with current Italian legislation. The concentration of *Escherichia coli* must be less than 230 MPN (most probable number) and the concentration of *Salmonella* must be zero.

We ran simulations during the historical period from December 7th–21st, 2015, chosen to correspond with available field measurements. As shown in Figure 4.2, the forecasted average areal distribution of tracers falls within a region bounded by Lon_{min} 14.08, Lon_{max} 14.1, Lat_{min} 40.76 and Lat_{max} 40.81. Chemical analysis on the mussels (*Mytilus galloprovincialis*) showed that on December 9th the concentration of *E. coli* was greater than the legal limits, while on December 21st it was lower than the legal limits).

Comparison between numerical forecasts and chemical analysis show a remarkable similarity in trends, although more observations will be needed before the method can be fully assessed. These early results do confirm that the system holds promise as a decision support tool for many applications that depend on sea quality, and require forecasts in support of local observations and measurements (Figure 4).

5. CONCLUSIONS

We described our use of the Globus Galaxy-based FACE-IT technology in a project that extends FACE-IT’s initial focus on climate, agricultural and socio-economic interactions to a tactical pollutant prediction application.

We will continue to maintain and improve the FACE-IT core infrastructure [12]. In the short term, we will improve interactive visualization tools for the application shown in this paper and others in the FACE-IT data portal. In the longer term, we will implement the FACE-IT infrastructure as a Docker [14] application so that users can run large-scale science workflows on their own resources (cloud, cluster or local), while leveraging real or virtualized accelerators [9].

6. ACKNOWLEDGMENTS

We thank the Globus Galaxies, Globus, and Galaxy teams for their outstanding work on those systems and for their assistance with this project.

This work is supported in part by NSF cyberSEES program award ACI-1331782; NSF Decision Making Under Uncertainty program award 0951576; DOE contract DE-AC02-06CH11357; project IZSM ME04/12 RC/C78C 120017001, “Mapping Escherichia Coli and Salmonella pollution in mussel farm areas and model prediction comparisons”; and European Union Grant Agreement number 644312-RAPID – H2020-ICT-2014/H2020-ICT-2014-1 “Heterogeneous Secure Multi-level Remote Acceleration Service for Low-Power Integrated Systems and Devices,” using GVirtuS open source software components. EC2 resources were generously provided by Amazon.

7. REFERENCES

[1] F. Adragna, M. J. Bell, F. Jacq, J. A. Johannessen, P.-Y. L. Traon, N. Pinardi, and J. She. Ocean monitoring and forecasting core services: The European MyOcean example. In *OceanObs 9*, pages 21–25, 2009.

[2] A. Doglioli, M. Magaldi, L. Vezzulli, and S. Tucci. Development of a numerical model to study the dispersion of wastes coming from a marine fish farm in the Ligurian Sea (Western Mediterranean). *Aquaculture*, 231(1–4):215–235, 2004.

[3] G. Giunta, P. Mariani, R. Montella, and A. Riccio. pPOM: A nested, scalable, parallel and Fortran 90 implementation of the Princeton Ocean Model.

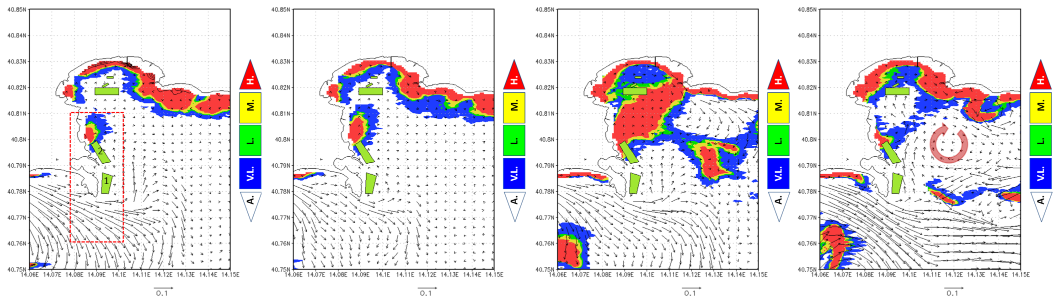

Figure 3: Sea surface currents (vectors) and pollutant concentrations in in Gulf of Pozzuoli (Campania Region, Italy) on days 08/12/2015 Z12 (figure A), 09/12/2015 Z12 (figure B), 15/12/2015 Z12 (figure C) and 21/12/2015 Z23 (figure D). The red dotted line is the area of Study with mussel farms in Punta Terone – Capo Miseno (number 1 in figure A) and Punta del Poggio – Centocamerelle (number 2 in figure A) areas. The color bar represents the concentration of tracers (A=absent, VL=very low, L=low, M=medium, H=height) in the area of study.

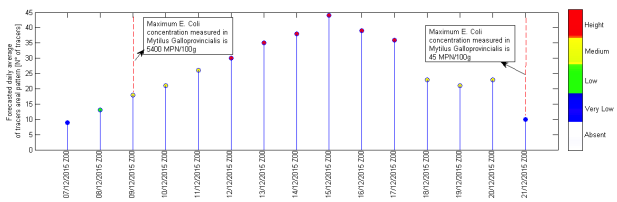

Figure 4: Simulation vs. measured results for the pollution event started at December 6th 2015.

Environmental Modelling & Software, 22(1):117 – 122, 2007.

[4] G. Giunta, R. Montella, G. Laccetti, F. Isaila, and F. J. G. Blas. A GPU accelerated high performance cloud computing infrastructure for grid computing based virtual environmental laboratory. *Advances in Grid Computing*, 2011.

[5] G. Giunta, R. Montella, P. Mariani, and A. Riccio. Modeling and computational issues for air/water quality problems: a grid computing approach. *Nuovo Cimento C Geophysics Space Physics C*, 28:215, Mar. 2005.

[6] R. D. Lauro, F. Giannone, L. Ambrosio, and R. Montella. Virtualizing general purpose gpus for high performance cloud computing: An application to a fluid simulator. In *2012 IEEE 10th International Symposium on Parallel and Distributed Processing with Applications*, pages 863–864, July 2012.

[7] R. Madduri, K. Chard, R. Chard, L. Lacinski, A. Rodriguez, D. Sulakhe, D. Kelly, U. Dave, and I. Foster. The Globus Galaxies platform: Delivering science gateways as a service. *Concurrency and Computation: Practice and Experience*, 27(16):4344–4360, 2015. CPE-15-0040.

[8] J. Michalakes, J. Dudhia, D. Gill, T. Henderson, J. Klemp, W. Skamarock, and W. Wang. The Weather Research and Forecast model: Software architecture and performance. In *11th ECMWF Workshop on the Use of High Performance Computing in Meteorology*, pages 156–168, 2005.

[9] R. Montella, G. Coviello, G. Giunta, G. Laccetti, F. Isaila, and J. G. Blas. A general-purpose virtualization service for hpc on cloud computing: An application to gpus. In *Proceedings of the 9th International Conference on Parallel Processing and Applied Mathematics - Volume Part I*, PPAM'11, pages 740–749, Berlin, Heidelberg, 2012. Springer-Verlag.

[10] R. Montella, G. Giunta, and A. Riccio. Using grid computing based components in on demand environmental data delivery. In *Proceedings of the Second Workshop on Use of P2P, GRID and Agents for the Development of Content Networks*, UPGRADE '07, pages 81–86, New York, NY, USA, 2007. ACM.

[11] R. Montella, D. Kelly, W. Xiong, A. Brizius, J. Elliott, R. Madduri, K. Maheshwari, C. Porter, P. Vilter, M. Wilde, M. Zhang, and I. Foster. FACE-IT: A science gateway for food security research. *Concurrency and Computation: Practice and Experience*, 27(16):4423–4436, 2015. cpe.3540.

[12] Q. Pham, T. Malik, I. Foster, R. Di Lauro, and R. Montella. *SOLE: Linking Research Papers with Science Objects*, pages 203–208. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[13] S. Saha, S. Nadiga, C. Thiaw, J. Wang, W. Wang, Q. Zhang, H. M. V. den Dool, H. L. Pan, S. Moorthi, D. Behringer, D. Stokes, M. P. na, S. Lord, G. White, W. Ebisuzaki, P. Peng, and P. Xie. The NCEP Climate Forecast System. *Journal of Climate*, 110, 2009.

[14] G. M. Tihfon, S. Park, J. Kim, and Y.-M. Kim. An efficient multi-task PaaS cloud infrastructure based on docker and AWS ECS for application deployment. *Cluster Computing*, 19(3):1585–1597, 2016.

[15] J. L. Wilkin, H. G. Arango, D. B. Haidvogel, C. S. Lichtenwalner, S. M. Glenn, and K. S. Hedström. A regional ocean modeling system for the Long-term Ecosystem Observatory. *Journal of Geophysical Research: Oceans*, 110(C6):2156–2202, 2005. C06S91.