The DRIHM Project: a Flexible Approach to Integrate HPC, Grid and Cloud Resources for Hydro-Meteorological Research


*CNR - Institute of Applied Mathematics and Information Technologies
Email: {dagostino, clematis, galizia, quarati, danovaro, roverelli, zereik}@ge.imati.cnr.it
†Ludwig-Maximilians-Universität München
Email: {kranzlm, schiffer, felde, straube}@nm.ifi.lmu.de
‡CNRM-GAME (CNRS, Météo-France), Email: olivier.caumont@meteo.fr
§CNRS-University of Toulouse III, Email: evelyne.richard@aero.obs-mip.fr
¶Universidad Politécnica de Madrid, Email: garrote@caminos.upm.es
∥HR Wallingford Ltd., Email: q.harpham@hrwallingford.com
**Deltares, Email: bert.jagers@deltas.nl
††Republic Hydrometeorological Service of Serbia
Email: {vladimir.dimitrijevic, ljiljana.dekic}@hidmet.gov.rs
‡‡CIMA Research Foundation
Email: {elisabetta.fiori, fabio.delogu, antonio.parodi}@cimafoundation.org

Abstract—The distributed research infrastructure for hydro-meteorology (DRIHM) project focuses on the development of an e-Science infrastructure to provide end-to-end hydro-meteorological research (HMR) services (models, data, and post-processing tools) by exploiting HPC, Grid and Cloud facilities. In particular, the DRIHM infrastructure supports the execution and analysis of high-resolution simulations through the definition of workflows composed by heterogeneous HMR models in a scalable and interoperable way, while hiding all the low level complexities. This contribution gives insights into best practices adopted to satisfy the requirements of an emerging multidisciplinary scientific community composed of earth and atmospheric scientists. To this end, DRIHM supplies innovative services leveraging high performance and distributed computing resources. Hydro-meteorological requirements shape this IT infrastructure through an iterative “learning-by-doing” approach that permits tight interactions between the application community and computer scientists, leading to the development of a flexible, extensible, and interoperable framework.

I. INTRODUCTION

Computational earth and atmospheric sciences such as hydro-meteorological research (HMR), play a key role in guiding the design and implementation of prediction tools devoted to the safety and prosperity of humans and ecosystems from highly urbanized areas to coastal zones and agricultural landscapes. Severe hydro-meteorological (HM) events are increasing in frequency and magnitude, often with severe societal and economic implications [1]. The analysis carried out within the FLASH European Project estimated the material damages produced by floods in the Mediterranean region during the 1990-2006 period at over 29 billion euros [2]. The total number of casualties has been estimated to be over 4,500, concentrated in the Mediterranean African countries. The peculiar nature of severe HM processes occurring in complex orography areas such as these makes this difficult to predict [3], [4]. Indeed, Mediterranean regions are characterized by steep mountains and small to medium-size catchments. Moreover, storms do not respect country boundaries [5] and therefore an international approach is necessary.

For these reasons, forecasting severe storms and floods could be considered as one of the main challenges of the 21st century. Meeting this challenge requires improvements in the way predictions are obtained [6]–[8]. At the heart of this challenge lies easy access to HM data repositories [9], [10], models and computing resources and facilitating collaboration between meteorologists, hydrologists, and earth scientists.

Such forecasts, from atmospheric phenomena leading to cloud formations to the assessment of impact created by flood events, require the combined use of meteorological, hydrological, hydraulic and impact modeling tools in well-orchestrated workflows, such as those depicted in Figure 1. Assembling and executing such workflows is, indeed, a non-trivial task. HM scientists have two main concerns about models and their composition into workflows: validation of simulation results and access to sufficiently powerful computing resources. Validation aims at improving the prediction
quality of each single step in a workflow, as well as of the overall chain. This validation process has to compare forecasting results with observed data and crosscheck them with those produced by other models. Accurate predictions require the execution of highly detailed, large scale and computationally demanding simulations. An e-Infrastructure capable of effectively supporting the set of heterogeneous requirements posed by this model workflow is needed.

Presently, the scientific and technical challenges are considerable, but within range. Models that can be executed as consecutive steps of a specific workflow usually depend on different execution environments and organizational constraints, and also have to deal with incompatible data formats and semantics. The same issues occur with alternative model choices at each step. Therefore, in general, most of the forecasting workflows run by civil protection agencies or research centers are hard-wired, ad-hoc solutions developed for specific and well-defined execution environments.

In this paper we present how the expertise of HM and information and communications technology (ICT) scientists has been combined within the DRIHM European Project [11] to enable a step beyond the state-of- the-art in modeling and testing HM probabilistic forecasting chains. The objectives include easier improvement of the HM models, and the processes supporting provision of reliable and timely information to decision makers about the location and timing of extreme HM events.

Several socio-economic benefits can be derived from the project. Firstly, a deeper understanding of severe HM events in complex topography regions will be gained. This will give an important contribution to the development of new modelling, post-processing, and analysis tools, resulting in medium-term improvements in operational facilities for the prediction, prevention and risk mitigation of extreme events. Secondly, a more comprehensive cooperation will arise at the European (and possibly global) level in the study of such events and other climate change effects. Thirdly, the project will encourage a further exploitation of the ICT infrastructures, in part designed to assure a secure society.

To succeed in meeting those general objectives, DRIHM has to tackle a set of technical problems. The adopted solutions may not always be innovative from an ICT point of view, but they require the definition of best practices [12], [13] and, in several cases, the problems are still open. In particular, this paper focuses on the design and development of a flexible, extensible and interoperable e-science infrastructure that:

- is based on the seamless integration of HPC, Grid and Cloud resources, all of which could be required to execute just a single HMR workflow;
- provides ICT technologies and HM models through a unified access point with a user-friendly interface based on the science gateway paradigm;
- adopts solutions adhering, as far as possible, to existing and emerging standards;
- results in the creation of a comprehensive ecosystem of HM services able to effectively improve the way simulations are performed.

The main challenges faced in bringing together multidisciplinary teams have been (1) gaining a common understanding of how the different modelling tools work, (2) translating the HMR requirements into effective ICT services, and (3) resolving differing terminologies used by the teams. This was supported by our iterative learning-by-doing approach, where the HM scientists carry out HMR activities while simultaneously working with ICT researchers to develop or adapt the necessary tools and practices. Learning-by-doing has been a well-chosen and positive methodology as each group begins to appreciate the issues faced by the others. We would strongly recommend it to future teams who are conducting similar tasks and who also may wish to extend the concept.

The paper is structured as follows: Section II overviews related scientific projects; Section III introduces the HMR experiment suites describing workflows and their requirements; Section IV details the DRIHM infrastructure; Section V supplies some insight about the science gateway developed in the project; a proof of concept is given in Section VI; the last Section concludes and presents future work.

II. RELATED WORK

The literature describing the parallelization of specific HM models for current multicore and heterogeneous architectures is rich with examples, not least papers presented at the “Programming weather, climate and earth-system models on heterogeneous multi-core platforms” workshop, which has been running since 2011 at the National Center for Atmospheric Research in Boulder, Colorado [14].

However, the focus of this paper is slightly different. Many attempts have been made in different countries to build up sound HM probabilistic chains, notably in Southern Europe and the US [15], [16]. Moreover, there is an increasingly pressing need, driven by societal and economical expectations, to predict possible impending floods with a quantitative measure of uncertainty [17]. In most cases they comprise of one, or a few, fixed chains of models as stated in the Introduction.

By contrast, one of the aims of the DRIHM project is the provision of advanced services to integrate heterogeneous HMR models, tools, and data [18]. For this reason we consider an analysis of the most important related projects and infrastructures in Europe and in the US as far as they relate to HMR.

The community earth system model (CESM) [19] is a coupled climate model for simulating the Earth’s climate system.

![Fig. 1. A complete HM chain. The integration bridges are represented with arrows. They implement data interchange standards to allow coupling of I/O data between models.](image-url)
Composed of separate models (atmosphere, ocean, land, land-ice, sea-ice) plus one central coupler component, CESM allows researchers to conduct fundamental research into the Earth’s past, present and future climate states. CESM is funded by the US National Science Foundation (NSF) and the U.S. Department of Energy (DOE). Because CESM aims at focusing on the Earth System Model in order to better understand the Earth’s dynamics over long time periods, there is an ecosystem of supporting services to make these simulation runs happen. Although CESM is not devoted to HMR, HMR simulation may benefit from the data and models made available. In particular, CESM’s standards-based approach may simplify the integration of these data and models into HMR applications and workflows.

Similar to CESM, the earth system modeling framework (ESMF) [20] collaboration aims at building a high performance, flexible software infrastructure to improve climate research, numerical weather prediction, data assimilation and corresponding Earth science applications. ESMF defines an architecture for composing complex, coupled modeling systems and includes data structures and utilities for developing individual models. The basic idea behind ESMF is that complicated applications should be broken up into smaller pieces, or components. A component is a unit of software that has a coherent function and that exposes a standard interface and behavior. Components can be assembled to create multiple applications and different implementations of a component may be available. ESMF also includes toolkits for building components and applications, such as re-gridding software, calendar management, logging, error handling, and parallel communications. As CESM, ESMF provides a modeling framework with focus on geophysical phenomena rather than those belonging to HM.

The CUAHSI Hydrologic Information System (HIS) [21] is an Internet-based system for sharing hydrologic data. It is comprised of databases and servers, connected through Web Services to client applications, allowing for the publication, discovery and access of data. HIS is operated through graphical user interfaces (HydroDesktop, HydroExcel) and Web Services in the context of pure hydrologic science. Accordingly, HIS provides standardized ways to convert scientific data between models (WaterML), to couple models (OpenMI), to retrieve and integrate data using standard templates (Observations Data Model - ODM). The centre of the CUAHSI-HIS architecture is HIS Central, a Web application that provides interfaces for adding and managing registered water data services and the HIS Central Metadata Catalog. The catalogue is designed to maintain observation series information including site information, variable information, the period of record, as well as project metadata for all registered data sources of hydrologic observations.

The multiscale applications on european e-infrastructures (MAPPER) [22] is a project that aims to deploy a computational science environment for distributed multiscale computing on and across European e-infrastructures. By taking advantage of existing software and services, MAPPER develops tools, software and services that permit loosely and tightly coupled multiscale computing in five representative scientific domains (fusion, hydrology, clinical decision making, systems biology, nano science). Besides the hydrological use case, the resulting tools and methods for distributed multiscale computing can beneficially be applied to HMR specific applications requiring a multiscale modelling approach.

In 2008 the Australian Bureau of Meteorology and Australia’s national science agency, the Commonwealth Scientific and Industrial Research Organization (CSIRO), joined their research forces to establish the Water Information Research and Development Alliance (WIRADA) [23]. Since then, WIRADA has delivered significant achievements in improving Australia’s water information systems, water accounting and water forecasting, with national and global impacts. WIRADA developed a reliable seasonal streamflow forecasting service, a high resolution digital elevation model for landscape and water resources, a national water resources modeling system and a common format within an international water data standard. The primary access to all WIRADA services is the Hydrologists Workbench, facilitating an automated management of the flows of exponentially increasing data volumes through models, while ensuring auditability and compliance with standards (WaterML, Water Data Transfer). Built on commercial off-the-shelf scientific workflow software which provides the workflow, audit and governance utility, the Workbench draws together public domain and proprietary hydrological, statistical and geo-information system toolkits with tailored workflows to provide an extensible portal for the provision and management of (one-off or routine) modeling exercises.

Among other initiatives, the WRF for GRID (WRF4G) [24] is a system developed by the Santander Meteorology Group for the execution and monitoring of WRF experiments on Grid resources by using the science gateway customization methodology developed within the scientific gateway based user support (SCI-BUS) European project [25].

Past initiatives include the linked environments for atmospheric discovery projects (LEAD I and II) [26], [27], funded by NSF, which aimed to create an integrated framework for distributed computing with a main focus on the identification, access and assimilation of a broad array of meteorological data and model output in an independent way with respect to format and physical location.

With respect to these projects DRIHM: a) aims to exploit existing software solutions and infrastructures without the development of a new framework/middleware, but only specific application-oriented services mainly for data interoperability; b) is a joint activity between ICT and HMR scientists therefore the solutions are designed together and not designed by ICT partners and afterwards customized for HMR applications; c) aims to integrate/couple several models to run scientific HMR workflows where the meteorology is only one component.

Regarding a), several general purpose technological solutions allow the development of advanced environments. In particular the project adopted the gUSE/WS-PGRADE science gateway toolkit [28], [29], one of a number of solutions in-
cluding the Apache Airavata [30], the Vine Toolkit [31] frameworks, the RADICAL Cybertools [32], and the commercial EnginFrame software [33]. gUSE has the advantage to support a large number of distributed computing infrastructures, as the most important Grid middlewares plus Web Services and Cloud infrastructures. Moreover, it has an active cooperating community also in the context of the above mentioned SCI-BUS project, where a central repository of application-specific portlets and workflows have been implemented.

III. THE HMR EXPERIMENT SUITES

One of the main goals of DRIHM is to enable a step beyond the state-of-the-art in the modelling of a forecasting chain composed by a Rainfall layer, a Discharge layer and a Water level, Flow and Impact Layer (see Figure 2). This addresses the interdisciplinary challenges of HMR in forecasting severe HM events over complex orography areas, and assesses their impact with the formulation of flood risk scenarios relevant to early-warning and Civil Protection.

For each layer, the HM models and tools adopted were chosen as part of a survey process undertaken in a preliminary phase of the DRIHM project. This choice was subsequently refined on the basis of the expertise available within the consortium and also through a set of thematic sessions at the European Geosciences Union General Assembly (a major HMR event). As a result, the selected models and tools shown in Figure 2 are representative of the state-of-the-art.

A. Composing HMR Workflows

The DRIHM project based the design, implementation and deployment of its workflow framework on a set of specific use cases, divided into three main experiment suites [34]:

Experiment suite 1 - Rainfall: a combination of different Numerical Weather Prediction (NWP) models to form a high resolution multi-model ensemble together with stochastic downscaling algorithms to enable the production of more effective quantitative rainfall predictions for severe meteorological events. These models have considerable HPC requirements.

Experiment suite 2 - Discharge: a fusion of rainfall predictions (potentially from experiment suite 1) with corresponding observations, which are input into multiple hydrological models, to enable of the production of more accurate river discharge predictions. These models are often based on ensemble forecasting: a numerical prediction method that is used to generate a representative sample of the possible future states. This means that multiple numerical predictions are created by perturbing certain initial conditions and/or physical parameters. This creates a need for high throughput computing (HTC) facilities, such as the Grid.

Experiment suite 3 - Water Level, Flow and Impact: execution of hydraulic model compositions in different modes to assess the water levels, flow and impact created by the flood events. Indeed, this process can be driven from data produced by experiment suite 2. The most important aspect for this suite is that most of the models are commercial and Windows-based, therefore the Cloud represents one possible solution to ensure scalability in case of a large number of simulations.

B. HMR workflows requirements

The probabilistic forecasting chain is implemented as a workflow of HM models, as depicted in Figure 2. The analysis of data exchange between models and the study of the composition techniques identified the need for improvement in their integration at three levels: model interfaces, user interfaces, and computing infrastructures.

Model Interfaces: HM model chain components are often linked in highly tailored connections applicable only to the specific models required in a use case or product: only model $i$ (of Rainfall layer) and model $j$ (at Discharge layer) and model $k$ (at the Water level layer). Adding another data format, replacing model $j$ by model $j_2$, exploring sensitivities etc., can involve considerable re-engineering and analysis and thus hampers progress. The goal is that each model has to be configured to interact and exchange data with the preceding and subsequent models in the forecasting chain. This has to be the case for all the alternative models (i.e. WRF-NMM or Meso-NH for the meteorological forecast). Such an
HM model chain components are usually configured in different ways such as graphical user interfaces giving some help and support to a list of parameters created with a text editor requiring an extensive knowledge of each model. Moreover HM scientists recognized that general-purpose workflow management systems provide a minimal support in the definition of the experiment [38]. HMR user interfaces have to be enhanced to reduce the risk of mistakes in workflow configuration such as non-overlapping time intervals between consecutive models and unrelated spatial domains. This is especially relevant for newcomers but is also applicable for HM scientists, who cannot necessarily be expected to be familiar with all available models. For these reasons, a vertical integration is also necessary among the three levels in order to provide end users with a seamless environment able to support their experiments on the available resources. From the technological point of view, these represent the most important aspects driving the development of the three levels of experiment suites.

The present work focuses on Computing Infrastructures detailed in Section IV, while User Interfaces and the vertical integration are briefly presented in Section V.

IV. THE E-INFRASTRUCTURE FOR HMR

The heterogeneity of the HM modelling chain components is a key issue, in particular for research projects that are not based on a proprietary or pre-defined computing infrastructure as in the case of DRIHM. On the other hand, this represents an interesting case study about the enhancement and integration of software tools and problem solving environments for computational earth and atmospheric sciences applications that require not only high performance resources but also Cloud resources. The high level architecture of the DRIHM distributed computing infrastructure (DDCI) is presented in Figure 3. It is based on a portal, briefly described in Section V, that provides seamless access to two sets of computing resources.

The first set groups the Grid resources. The Grid is the natural choice to run rainfall downscaling models, such as RainFARM and models belonging to the Hydrologic and partially Hydraulic levels depicted in Figure 2. This is due to the fact that most of them do not involve elaborate physics or solve equations over large grids and they run sequentially. Moreover, ensemble forecasting requires running multiple instances of a single model using slightly different initial conditions and/or physical parameters. Most of the Grid resources are granted to DRIHM by the European grid infrastructure (EGI) [39], a federation of National grid infrastructure (NGIs) set up to deliver sustainable, integrated and secure computing services to European researchers and their international partners. Besides the Grid, performance and functionality requirements demand the inclusion of additional services which are not part of the core Grid resources. In particular, these services include supercomputers, data repositories, Linux/Windows dedicated nodes, specialized hardware, Cloud and Web services. The main component of this second set is represented by supercomputers, that in Europe are provided within the framework of the partnership for advanced computing in Europe (PRACE) project [40], that has the mission to enable high-impact scientific discovery and engineering research by offering powerful computing and data management resources through a peer review process. The use of such resources is motivated by the fact that execution of massive, fine-grained meteorological simulations requires much more computational power than that provided by the HPC systems available in a Grid infrastructure. The main drawback in the use of PRACE resources concerns the access policy: the use of an HPC system in PRACE is granted only to a small group of scientists whereas access by community (such as the extreme science and engineering discovery environment (XSEDE) community or the EGI robot certificate policies) is really what is required. DRIHM was designed to manage resource access by specific groups of users.

The other resources belonging to this set provide the following functionality:

- **Dedicated nodes**: computational resources used to run models that have more restrictive licenses and/or cannot be installed on the Grid due to certain functional restrictions e.g. compilers, non-portable software libraries, specific operative systems or large databases. An example is represented by the Windows resources, specific instances of dedicated nodes that can run the Windows-based models or services.
- **Data repositories**: storage and/or access to large data sets,
such as the global circulation models, and datasets related to the critical cases defined in the project.

- Specialized hardware providing capabilities for specific services, such as remote or 3D visualization.
- Cloud services, that provide cloud resources at the different IaaS/PaaS/SaaS levels. Currently the project is considering the adoption of Cloud resources for the IaaS level.
- Web services allowing access to models and facilities from outside of the Cloud and/or Grid. One example is the cumulonimbus tracking and monitoring (Cb-TRAM) algorithm for nowcasting [41], an HMR service that can be used to validate forecasts.

A. The DDCI in details

The resources of the DDCI are available through grants, such as EGI or PRACE, or they can be provided by DRIHM partners. The most important components of the present implementation of the DDCI, shown in Figure 4, are briefly described here.

Non-Grid resources - CIMA: The first step of the chain depicted in Figure 2 is the Rainfall layer. The meteorological models considered in DRIHM are nested in Global Circulation Models, which are usually run by the main weather forecasting centers like the European Centre for Medium-Range Weather Forecasts (ECMWF) or the national oceanic and atmospheric administration (NOAA). This means that the initial and boundary conditions are selected from their results. While the NOAA databases are accessible for free, this is not true in general for those from ECMWF. For the purpose of the project activities ECMWF allowed the initial and boundary condition files of a limited number of cases to be freely shared on the DDCI. The project used a dedicated machine hosted at CIMA foundation to store them and to run fast, sequential preprocessing tools (i.e. WPS for WRF-NMM and WRF-ARW) that extract the initial and boundary conditions, in order to avoid unnecessary large data movement.

PRACE resources: The weather forecast models are the most computationally demanding components of the workflow. For example, a very detailed three-nested domain on the Ligurian sea can require 500 seconds per simulated hour using 3072 cores of a BlueGene/Q system. Three PRACE resources were sometimes used to run such large scale simulations. Generic users will use the Grid resource to run them instead.

Grid resources: The production Grid resources are granted to DRIHM by EGI. They are located in Croatia, Germany, Greece, Italy, Macedonia, the Netherlands, Serbia and Spain [42]. The most powerful resource among them is the cluster provided by the Leibniz Supercomputing Centre (LRZ) in Germany, equipped with the Globus Toolkit 5 and able to provide up to around 4,000 cores. This resource is only used to run the large scale meteorological simulation for the users that cannot exploit PRACE resources. The other resources are equipped with gLite and the access is transparently managed by a WMS metascheduler which cooperates with CREAM job managers at the resource level. gLite resources are used for the other Linux-based HMR models. For monitoring purposes the EGI BDII information service is used and about 100 TB of data storage is available. The deployment of preliminary versions of software tools is performed using a testbed provided by the European globus community forum (EGCF, formerly initiative for globus in Europe - IGE) [43].

Cloud Resources: Some hydraulic models and the impact model are Windows-based. There are three main options to provide them in DRIHM. The first is applicable to a few .NET based applications, that is to re-compile them on the Mono software platform. The second is to install them on a fixed set of resources and make them available via Web services. The third is to exploit the Cloud. While the first solution is not general, the second has the disadvantage of a limited scalability. The Cloud may require a modification of the model user interface, but it can be adapted according to user requirements, thus granting the availability and scalability for hydraulic and impact models. The project is investigating a solution based on the EGI Federated IaaS Cloud platform [44]. All the models and related tools were installed on a Virtual Machine template and an interface able to launch the simulation via the command line is being built.
B. The “DRIHMification” process

Each HMR model of a forecasting chain can (theoretically) be executed on a different DCCI resource. Each resource (potentially) is owned, managed and maintained by a set of autonomous resource providers and uses different operating systems, software, and libraries. Therefore, it is of paramount importance to adhere to the transparency principle underlying distributed systems in general and Grids in particular. Consequently, HMR models, tools and data sets need to be deployed in such a way that both location and migration transparency is achievable for data sets, models and application level workflows.

In particular the models may exhibit different levels of maturity: may be written in a multitude of programming languages, may depend on specific libraries (sometimes even specific library versions), and on specific execution environments. Adaptations are therefore inevitable. While the general task is called “gridification”, the project adopted the term “DRIHMification” for the process of preparing a particular HMR model for usage on the DCCI. A DRIHMrified model can be used together with others and executed on DCCI resources, in the sense that it can be executed in an environment other than the one in which it was created without requiring (major) rework.

A central DRIHM binary repository (see Figure 4) is used to hold the DRIHMrified and versioned Linux-based model binaries and their library prerequisites. For Windows-based models a Virtual Machine template containing all the required software was prepared and deployed on Cloud repositories. The DRIHM binary repository also provides mechanisms for a secure and reliable transfer of files to a particular Grid resource. A repository holding static data for particular models, e.g. soil moisture and geographic information system (GIS) data for the models belonging to experiment suites 2 and 3, complements the binary repository (see again Figure 4).

V. THE DRIHM SCIENCE GATEWAY

DDCI, as a general heterogeneous DCI, is able to provide the required resources to perform the experiments designed by the users, but it also presents some issues. In particular, it is complex and can be hard to access and use, due to the geographical distribution and heterogeneity of the resources. Therefore great attention has to be paid in designing solutions that provide easy, secure and consistent access to the underlying infrastructure, which have to be tailored to the expected user communities. This is even more compelling in the case of non-IT experts, most of the HM researchers, who have to focus on the model results and not on technicalities such as job scheduling and data movements.

Grid portals [45] (and Science gateways [46] that can be considered their technological evolution), represent a feasible solution to hide all these complexities. A science gateway toolkit (or framework) is a solution that, in addition to a general purpose portlet container, is able to mediate between the user at the front-end and DCI services at the back-end. Moreover, scientists orchestrate their simulation components by coupling appropriate models on various scales using ad-hoc interfaces and execute the designed experiments on resources typically unknown in advance.

The DRIHM Portal (http://portal.drihm.eu) is the scientific gateway developed by the project to shape this vision. It is based on the gUSE/WS-PGRADE science gateway toolkit and provides access to all the resources of Figure 4. The portal architecture is shown in Figure 5. In particular four groups of dedicated portlets have been developed. The first is devoted to the definition of the experiment. The interface shown in Figure 6 lets the user compose their desired model chain (at present only experiment suites 1 and 2 are supported) and select or upload their input data. The second group is composed of a set of model-specific portlets that guide the user through parameter specification of the selected models. Presently some basic consistency checks are performed in order to avoid wasting computer time for a misconfigured time interval or spatial domain. The third group performs some administrative tasks for specific models, i.e. the insertion of the data set in the static data repository for the hydrologic and hydraulic models. The last group is composed of the portlets for experiment execution. These portlets generate all configuration files and schedule the execution of the models on DCCI following the allocation policies with respect to the resources described in Section IV-A.

The actual execution of workflow steps and data movement is managed using the available gUSE/WS-PGRADE services. In particular the DCI-Bridge is responsible for translating each job submission into the format suitable for the selected resource type, and the Data Avenue portlet provides all the necessary data transfer protocols.

The DRIHM portal provides two facilities for analysing and inspecting large simulation datasets (mainly netCDF files). The ncWMS tool [47] is fully integrated in the web application and allows remote inspection of 2D variables. In the case of more complex data analysis and visualization (e.g. stereo rendering) DRIHM can offer access through the data access protocol (also known as OPeNDAP). In this scenario, the user can employ their preferred tool (e.g. Vapor, IDV) and simply connect to the remote dataset. Immersive dataset inspection has been
VI. PROOF OF CONCEPT: THE GENOA CASE

On November 4th, 2011 in Genoa, Italy a flash flood event occurred [18], [48]. It was chosen as one of the test cases to assess how HM experiment suites can take advantage of the DRIHM services. About 500 millimeters of rain, a third of the average annual rainfall, fell in six hours. The torrential rainfall inflicted the worst disaster Genoa has experienced in over four decades and six people were killed.

At the beginning of the project, only one of the partner institutions was easily able to carry out full HM simulations of the Genoa tragedy and these simulations involved only three models coupled with traditional hand-crafted links. This example represents what has been termed the “baseline HM experiment suites” i.e. starting capabilities (upper panel of Figure 7). In particular, only the WRF-ARW model and RainFARM were available to produce rainfall fields used as input for the DRiFt hydrological model which is used to predict the discharges at the outlet of the Bisagno river, as shown in the upper panel of Figure 7. In each panel of Figure 7, the upper row represents rainfall sources, including models and observations, and the lower row represents hydrological components, also including models and observations. In the initial available setup, WRF-ARW was executed once, then the predicted hourly two-dimensional fields of rainfall at ground level were either used to drive DRiFt directly, or disaggregated through RainFARM, to produce an ensemble of higher-resolution rainfall fields. In the latter workflow, RainFARM rainfall ensemble forecasts were then used to drive one DRiFt instance per RainFARM ensemble member. It was also possible to drive DRiFt with rain gauge observations, as indicated by the purple arrow. In addition to this limited choice of models, each was configured with tailored, esoteric tools requiring specific knowledge and so no consistency check was done across the whole workflow.

One of the first DRIHM services was the interface that successfully showcased at the Virtual Reality and Visualization Center (V2C) of LRZ, through isosurface extraction from remote results in netCDF format.
converts the outputs of atmospheric models into a common format acceptable by all hydrological models. This allowed enlargement of the list of meteorological models to include Meso-NH, AROME and WRF-NMM. The RIBS and HBV models were added to the available hydrological models. Notably, Meso-NH, AROME and RIBS were run as ensemble prediction systems (EPSs). Through the use of the DRIHM portal, it was easy for scientists to configure the different HM workflows to execute 58 atmospheric runs and more than 1800 hydrological runs (Table 1).

The result is that DRIHM users are not tied to their own in-house models and tools anymore, but are now able to carry out the same set of advanced experiments as schematized in the lower panel of Figure 7. Here the resources depicted in the upper panel are reproduced and complemented by arrows that represent newly available workflows. For example, unlike in the baseline experiment suites, DRiFT is not restricted to being driven by WRF-ARW, RainFARM and observations, but also by Meso-NH and AROME ensembles. Indeed, all these rainfall sources can now be used to drive the RIBS hydrological model. In theory, all meteorological and hydrological models can be combined in every feasible way. Furthermore, through the use of standard, common interfaces (netCDF-CF and WaterML 2, represented as boxes in the lower panel of Figure 7), the interpretation of simulation results from heterogeneous resources can be compared in a consistent framework and therefore is considerably eased, as illustrated in Figures 8 and 9. Figure 8 shows hourly rainfall accumulation in the upper panel and the wind direction and near-ground speed at the same time for three different meteorological models (from left to right AROME, WRF-ARW and Meso-NH). Although these models have very different characteristics in terms of physics, resolution and projection of computational grids and output file formats, the use of a common format (netCDF-CF) and associated bridges allows features, such as the effect of the horizontal resolution on rainfall patterns, to be distinguished at a glance. AROME predicts much smoother fields than Meso-NH. Differences in the location of the low-level convergence line over the Ligurian Sea can also be seen where the north-south white line over the sea is more easterly in Meso-NH than in the other models. All this provides unprecedented HM research capabilities such as component sensitivity studies where modular replacement of one of the components is enabled by the DDCI, as illustrated in Figure 9. This kind of study enables new research paradigms and gives new insight into, for example, the performance of different models or, the propagation of errors in HM chains.

The detailed and complete analysis of the Genoa critical case will be the subject of a dedicated HM paper, presently in preparation. However the early results can be briefly summarized as follows: the diversity of models helps understand the key physical processes responsible for the event; the ensembles based on Meso-NH perform best for this case; the simulated discharges are sensitive to meteorological modelling errors. Ultimately, the overall situation depicted by the collection of model ensembles would have suggested that the meteorological situation was potentially very dangerous.

### VII. Conclusions and Future Work

A key challenge for current hydro-meteorological research is to develop and validate new tools and methods for investigating severe events. The setup of complex workflows, in fact, is an interdisciplinary endeavor, requiring collaboration among measurement scientists and agencies collecting data, meteorological scientists forecasting the weather, hydrological and hydraulic scientists predicting runoff and flooding and public services such as national environmental and civil protection agencies to protect people.

There is a growing landscape of services and tools, which individually are valuable for HMR scientists but are, in themselves, not interoperable. In particular, there are three main challenges in combining different models: 1) the definition of common interchange data formats (like NetCDF-CF and WaterML) and protocols for coupling atmospheric and hydrologic models; 2) legacy model reuse with minimal adaptation, to setup interdisciplinary workflows and exploit distributed resources; 3) the development of effective portals and other user interfaces which simplify access to the different models and the availability and reliability of distributed computing resources for running them.

This paper presents the e-Science infrastructure developed for HMR within the DRIHM European project, an initiative that tackles these issues enabling the proper management of different kinds of software and hardware resources, from models and data to newly deployed services and infrastructures. DRIHM is currently two-thirds complete and therefore

### TABLE I

<table>
<thead>
<tr>
<th>Rain source</th>
<th>Description</th>
<th>Ensemble members</th>
<th>Resolution DRiFT (km)</th>
<th>RIBS exec.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>Rain gauge measurements</td>
<td>1</td>
<td>n/a</td>
<td>1 30</td>
</tr>
<tr>
<td>WRF-ARW</td>
<td>IC+BC: IFS</td>
<td>1</td>
<td>1.0</td>
<td>1 30</td>
</tr>
<tr>
<td>WRF-NMM</td>
<td>IC+BC: IFS</td>
<td>1</td>
<td>1.3</td>
<td>1 30</td>
</tr>
<tr>
<td>AROME</td>
<td>IC: Pert. obs.; BC: PEARP</td>
<td>8</td>
<td>2.5</td>
<td>8 240</td>
</tr>
<tr>
<td>MNH-ARP</td>
<td>IC+BC: ARPEGE</td>
<td>10</td>
<td>0.5</td>
<td>10 300</td>
</tr>
<tr>
<td>MNH-IFS</td>
<td>IC+BC: IFS</td>
<td>10</td>
<td>0.5</td>
<td>10 300</td>
</tr>
<tr>
<td>RainFARM</td>
<td>Disaggregation of rainfall from WRF-ARW simulation</td>
<td>7+20</td>
<td>0.7</td>
<td>7+20 210+600</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><em><strong>58</strong></em></td>
<td><em><strong>n/a</strong></em></td>
<td><em><strong>58</strong></em></td>
<td><em><strong>1740</strong></em></td>
</tr>
</tbody>
</table>
the work described in this paper is ongoing. In particular not all the models have been “DRIHMified” and the related portlets developed. As such, the full advantages and benefits of integrating these models, tools and data are only briefly discussed by way of showing their potential importance in helping researchers to design more efficient early-warning flash flood prediction systems.

As another further contribution, it is important to note the companion European project distributed research infrastructure for hydro-meteorology to United States of America (DRIHM2US), an international cooperation between Europe and the USA. This project builds on the NSF-funded standards-based cyberinfrastructure for hydro-meteorology (SCIHM) project [49], with the aim to promote the development of an interoperable HMR e-Infrastructure across the Atlantic. In particular, one forthcoming development is the inclusion of XSEDE resources in the DRIHM distributed computing infrastructure. This will be achieved in a straightforward way due an improvement of the DCI-Bridge component, developed within the SCI-BUS project, that will provide a plugin for such an infrastructure.

ACKNOWLEDGMENT

The authors would like to acknowledge the support of the DRIHM (Distributed Research Infrastructure for Hydro-Meteorology, 2011-2015) project, funded within the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement No: RI-28356.

REFERENCES


