Computing and Informatics, Vol. 34, 2015, 1065–1088

INTEGRATED SYSTEM FOR HYDRAULIC SIMULATIONS

Giang Nguyen, Viera Šipková, Peter Krammer
Ladislav Hluchý, Miroslav Dobrucký
Viet Tran, Ondrej Habala

Institute of Informatics
Slovak Academy of Sciences
Dúbravská cesta 9, 845 07 Bratislava, Slovakia
e-mail: {giang, sipkova, krammer, hluchy, dobrucky}.ui@savba.sk

Abstract. The work described in this paper is aimed at applying and co-operating of modern information technologies and mathematical modeling to make a risk analysis of the water-supply in big cities. It is instrumental in the investigation of the hydraulics of water-supply systems using the simulation model EPANET executed on the underlying high-performance computing infrastructure. The simulation process is integrated with the GIS environment in order to correct input data and visualize the simulation output. Input data for the model can be modified directly within the designed scientific gateway which enables hydraulic domain experts to interact comfortably with the HPC capacity. Furthermore, the system includes some data mining capabilities forming bridges between the hydraulic data storage and available hydrological measurements focused on water consumption modeling and predictions. In simulating the main emphasis is given to optimize the measure of a similarity between the mathematical model and the real system in order to obtain reliable results.

Keywords: Hydraulic simulations, water-supply system, high performance computing, scientific gateway, data mining, hydrology scenarios

Mathematics Subject Classification 2010: 68U35

1 INTRODUCTION

Mathematical modeling and computer simulations play a significant role not only in scientific and technical disciplines, but also in the economics, social, public and private spheres. In order to protect the environment and population safety there is an increasing pressure from various organizations on enforcing computer simulations of natural disasters, as well as various risky technical, economic and urban events that are associated with a damage of property or health or life endangering. Computer simulations represent some kind of the approximation of real systems and provide one of the most effective tool for examining the nature of a problem and its rational explanation. They are beneficial in case of investigating different scenarios that are impractical, financially expensive, unfeasible or too dangerous to run in real life.

In the last years, geographic information systems (GIS) have become an integral part of our living. The rapid development of information technologies along with the vast amounts of data observed shifted the GIS employment to the higher level. GIS systems enable to join graphic representations with entries stored in a database. Based on such data structures, many kinds of actions can be performed since the GIS technology combines common database operations, e.g. task assignment and statistical calculations, with the unique capabilities of displaying and spatial analysis provided by the map. The data constitutes naturally a good base for mathematical modeling and computer simulations.

The term "simulation" refers to the process of using the mathematical representation of a real system, called *mathematical model*. Constructing a mathematical model is a challenging, systematic work that needs many skills and employs the higher cognitive methods of interpretation, analysis and synthesis. It includes the observation and investigation of phenomena, followed by designing the concept of the problem, its parameterization, and the creation of the computer simulation model. The role of computer simulations is to find a solution to problems that makes the given scenario able to predict the behavior of the system from a set of parameters and initial conditions. The correctness of the mathematical model and the simulation itself is assessed on the basis of experimental results, verified theory or a long-term observation of the reality.

Simulation models can take many forms, they may range from simple regression expressions up to complex numerical solutions. Generally, the complexity of a simulation model is a compromise between the model simplicity and its exactness and reliability. Along with the improvement of the model accuracy also requirements on the precision of underlying data are gradually growing, as well as the computing power and the time needed for simulating. An imperative part of the simulation process is the verification and validation phase, in which it is found out whether the mathematical model describes the real system with the sufficient accuracy, that is, whether results obtained from the simulation are identical with results observed from the real system in a declared degree of similarity. For this purpose it is necessary to investigate and solve calibration and parameter setting of the model which is based primarily on the analysis and comparison of the simulation output and

results obtained from actual measurements and experiments. To achieve the objective, a large number of simulation runs must be performed, operating with many input scenarios, varying input parameters and boundary conditions. This represents a challenging task requiring high performance computing resources, a considerable amount of memory and storage space, experience of domain experts and also the employment of supporting software tools, whose main function is to manage the whole simulation process in a simple and effective way.

The body of this paper is organized in five sections. After the introduction, Section 2 outlines the mathematical approach to describe the water-supply system. Section 3 presents the technological integrated concept of the system including interconnected frameworks. Section 4 aims to the work of each component and illustrates the potential of hydraulic simulations running on high performance computing infrastructures (HPC/Grid/Cloud). The conclusion of the work is given in Section 5.

2 MATHEMATICAL MODEL OF WATER-SUPPLY SYSTEMS

A water-supply system (particularly, the water-supply system in Bratislava analyzed in our work) can be defined formally as an undirected graph

$$G = (V, E)$$

containing a set of edges (links) E and a set of vertices (nodes) V. Each vertex $v \in V$ represents a junction where links join together, a water source point or a water delivery endpoint. An edge $e = (u, v) \in E$ connects a pair of vertices u and v, where $u \neq v$. The graph may contain a lot of open endpoints, e.g. connections to each house. A water-supply system can also be described as a directed graph, where starting and ending vertex are marked for each edge.

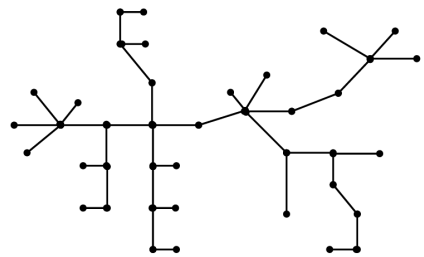

Figure 1. Graph of the water-supply system with a number of endpoints

The graph G is a vertex-evaluated graph. To each vertex $v \in V$ the geographical position is assigned as a triplet:

$$g = (x, y, z)$$

where:

- x, y are the real coordinates for the co-ordination system globe (particularly, in Slovakia the coordinate system is *"S-JTSK Krovak EastNorth"*)
- z is a coordinate which expresses the height above the sea level in meters

Additional features representing the set of node properties may be assigned to each vertex $v \in V$ as an $n-$tuple:

$$ov = (ov_1, ov_2, \ldots, ov_n)$$

where:

- n is the number of features assigned to v
- ov_1, ov_2, ..., ov_n are various features, e.g. the elevation, demand, head, pressure, quality, etc.

The graph G is also the edge-evaluated graph. Additional features representing link properties may be assigned to each edge $e \in E$ as an m-tuple:

$$oe = (oe_1, oe_2, \ldots, oe_m)$$

where:

- m is the number of features assigned to e
- oe_1, oe_2, ..., oe_m are various features, e.g. the length, diameter, material, pipe roughness, flow, velocity, headloss (pressure), status (open, active, closed), chemical reaction rate, etc.

The geometry of the water-supply system is transformed into the geometry between points and lines (see Figure 2). In routine practice, it often happens that during the digitalization process various impurities are carried into the GIS data. For example:

- lines do not intersect at the crossing point,
- small gaps/spaces are between two lines or between a line and an endpoint,
- points that do not lie on lines although they should, e.g. pumps or valves on pipes,
- duplicate lines.

Such defects induce errors and have to be eliminated. They give false information and can spread increasingly distorted or incorrect information to the further processing as well as to simulations to predict situations. To eliminate these impurities geometrical vector computations [6] (the dot product, cross product, line to point distance vector computation, etc.) are used in various implementation approaches, e.g. using scripting languagues and/or existing geographical tools offered by GIS environments.

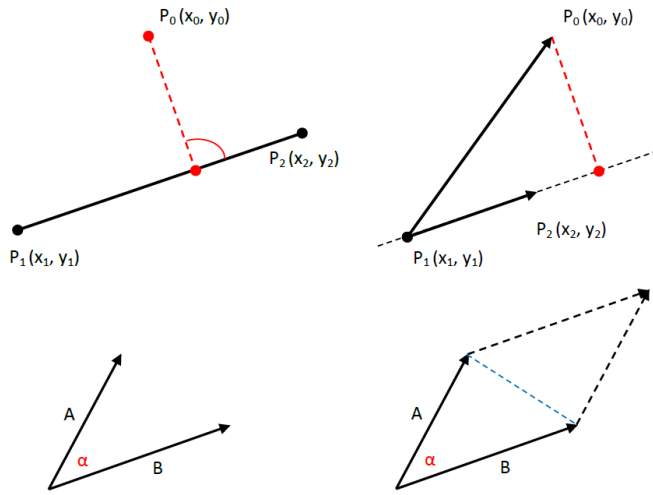

Figure 2. Geometry between points and lines

3 SYSTEM ARCHITECTURE

The purpose of the research of the water flow hydraulics in pipes of a water network is to gain insight on the behavior of all components of the water-supply system, starting from the water acquisition, water exploitation, its treatment for drinking, water transport, business issues, and solutions for water losses. A computational simulation model together with the measured data of the real water distribution system of a large city presents an effective tool for doing hydraulic research. Hydraulic models that are capable to accurately and realistically simulate reactions of the water-supply network under various conditions and for many different scenarios are becoming more and more required from the point of view of planning, design and analysis. The mathematical modeling itself consists in replacing the actual water distribution network by an idealized image in the form of mathematical-data model, and in the verification of the model substantiality based on data sources. The essential 'must' in modeling is that the idealized system fits as close as possible to the actual system, and the parameters of the system conform sufficiently with the objective behavior of the real network.

The trend in the development of software tools for simulating pressure flow is directed towards the integration of mathematical modeling with GIS technology which couples the database information with the geographic information. By computational simulations of pressurized networks the model EPANET (US Environmental Protection Agency) [7] is considered to be the world standard. It became the mathematical core of multiple commercial products, for example, MIKE URBAN [9], WaterGEMS [10], KYPIPE [11], etc. Another well-known software systems, com-

mercial and public available, designed to simulate hydraulic phenomena are: InfoWater [12], H2ONET [13], AFT Fathom [14], SynerGEE [15], ERACLITO [16], CROSS [17]. Many of them are fully integrated with GIS software tools, such as ArcGIS [19] or QGIS [20]. In our work, following the requirements and working experience of hydraulic domain experts [21, 22] the tool ArcGIS was chosen for the operating mode.

Model EPANET is a computer program that performs extended period simulation of hydraulic and water quality behavior within pressurized pipe networks. The input for EPANET is the mathematical representation of a water distribution network consisting of pipes, nodes (pipe junctions), pumps, valves and storage tanks or reservoirs. EPANET tracks the flow of water in each pipe, the pressure at each node, the height of water in each tank, and the concentration of chemical species throughout the network during a simulation period comprised of multiple time steps. In addition to chemical species, the water age and source tracing can also be simulated. EPANET is designed to be a research tool for improving our understanding of the movement and the fate of drinking water constituents within distribution systems.

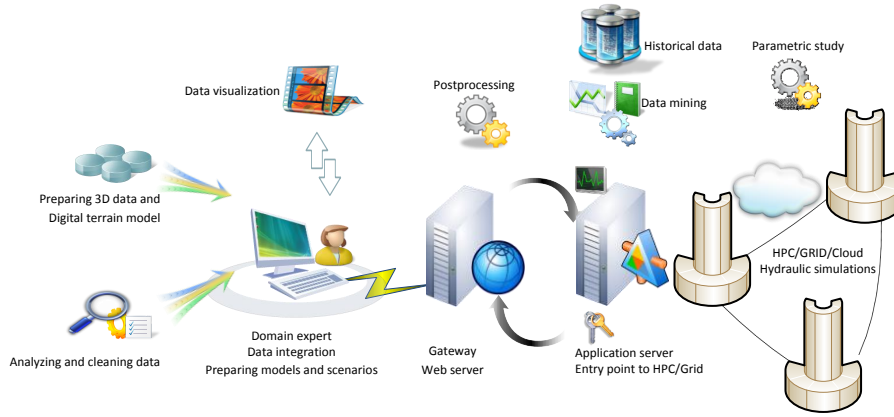

Figure 3. System components

The architecture of our system integrating several instruments to practise the hydraulic research is presented in Figure 3. The system is user-oriented and is designed primarily for hydraulic domain experts who may enter their own knowledge into models and input scenarios. The results of simulations are transferred back to the end-user for further processing, such as to interpret, analyze, visualize, or to utilize them in real operational services. The main components of the system are described in the following.

- Tools to analyze and clean GIS data. They prepare the input GIS data for simulating. They remove data impurities as described above in Section 2 applying the available geographical ArcGIS routines and designed python scripts.

- The HPC environment including the deployed simulation model EPANET.
- The tool to realize the parametric study with hydraulic simulations using the EPANET.
- The data mining tool. It employes several machine learning models operating on historical data (about pumps, reservoir, etc.) and results of hydrological measurements related to the water consumption modeling and predictions.
- The scientific gateway and the underlying web server. It provides a comfortable bridge to the HPC environment to realize hydraulic simulations. It offers functionalities to manage several computing environments, and additionally, it enables to create new scenarios through direct modifications of some input quantities.
- The monitoring daemon included in the application server. It provides a remote access to the selected HPC platform, which in common is not visible to end-users.
- Tools to post-process the results of hydraulic simulations. They transform the simulation output written in the plain text format into the CSV (Comma Separated Values) format [23], and then to MDB (Microsoft Access Database) format [24] for easy data importing back into the GIS environment.
- The system also provides the hydraulic domain experts the support in their work by tasks "Preparing 3D data and digital terrain model" and "Data visualization".

4 HYDRAULIC SIMULATIONS ON HPC ENVIRONMENTS

The simulation model EPANET 2 [8] was deployed on the compute cluster SIVVP [25] designed for high performance computing. The cluster can be accessed locally, or remotely using the grid technology [26, 27, 28], as it is integrated in the "European Grid Infrastructure" EGI [29]. The execution of simulations on the local cluster can be realized by means of the system PBS (Portable Batch System) [30] and the execution of simulations on EGI infrastructure through the Grid middleware gLite/EMI [31, 32]. EPANET 2 is also pre-installed in the virtual disk image on the Cloud infrastructure IISAS-Fedcloud which is part of the "EGI Federated Cloud Infrastructure". It was designed in cooperation with several European projects: EGI-InSPIRE [33], Helix Nebula [34] and Cloud Plugfests [35]. On IISAS-Fedcloud the middleware OpenStack [36] is installed, along with tools for the integration into the federated Cloud infrastructure.

The available middleware services offered by HPC infrastructures deliver most of functionalities necessary for the development and running of applications, however, for non-informatics scientists, they are mostly too complex and require non-trivial knowledge to be used correctly. To facilitate the process of a submission and efficient execution of simulations some software tools, working on the base of the underlying middleware, have been designed. They are capable to isolate the end-user from the middleware infrastructure taking on the entire deployment burden.

4.1 Scientific Gateway and Web Server

The scientific gateway for hydraulic simulations (look at Figure 5) is implemented using AJAX technologies. AJAX [37], an acronym for Asynchronous JavaScript and XML, represents a set of techniques useful for the development of interactive web applications enabling to change the site content without the need of a complete page reloading from the server. In comparison with traditional web applications, AJAX applications provide the user with a more comfortable experience only requiring modern (i.e. later) web browsers. AJAX is not an independent programming language or a simple technology, it is a combination of the following elements, but not necessarily of all of them.

- HTML and CSS languages for describing the structure and layout of information in the display.
- DOM (Document Object Model) interface associated with the JavaScript to dynamically display and interact with the presented information.
- Methods for data exchange between browsers and the server without the need to restore the current page, the most commonly used is XMLHttpRequest object.
- Formats for data transmission by the browser including the plain text, XML, HTML and JSON (JavaScript Object Notation). Such data can be dynamically generated by scripts on the server-side.

Hydraulic domain experts utilize the scientific gateway as one of the forms to access and exploit the HPC computational power. Using the gateway, they do not have to concern about the target compute platform, whether their job is executed directly on the HPC cluster, or it is submitted for execution to the Grid or Cloud infrastructure. Moreover, the gateway has capabilities which enable to generate new input scenarios by modifying some quantities within the input data for hydraulic simulations. It is possible to change time-step parameters for simulation, output report parameters, requirements on computational power, and water consumption demands for various consumer groups (industrial, hospitals, schools, households, gardens, etc.). One instance is shown in Figure 4. The list of gateway functionalities is extensible considering the needs of domain experts. The gateway approach is known and it has already been applied for handling grid applications [38, 39], however, in most cases it is too general and requires deep knowledge to be customized for a given scientific domain. Our approach is easy, user-oriented and applicable not only to grid, but to HPC cluster and cloud as well.

Usually, the high performance computers and our cluster inclusive, are characterized by a high hardware and operating cost. Grid, to be said very briefly, interconnects high-performance compute centres. Such centres must provide simultaneously and continuously various services to many users, therefore, they have high requirements on the security features. Generally, in the security sphere a web server is considered to be a weak element. Due to its potential vulnerability our web server was installed separately. The role of the web server is to communicate with the user

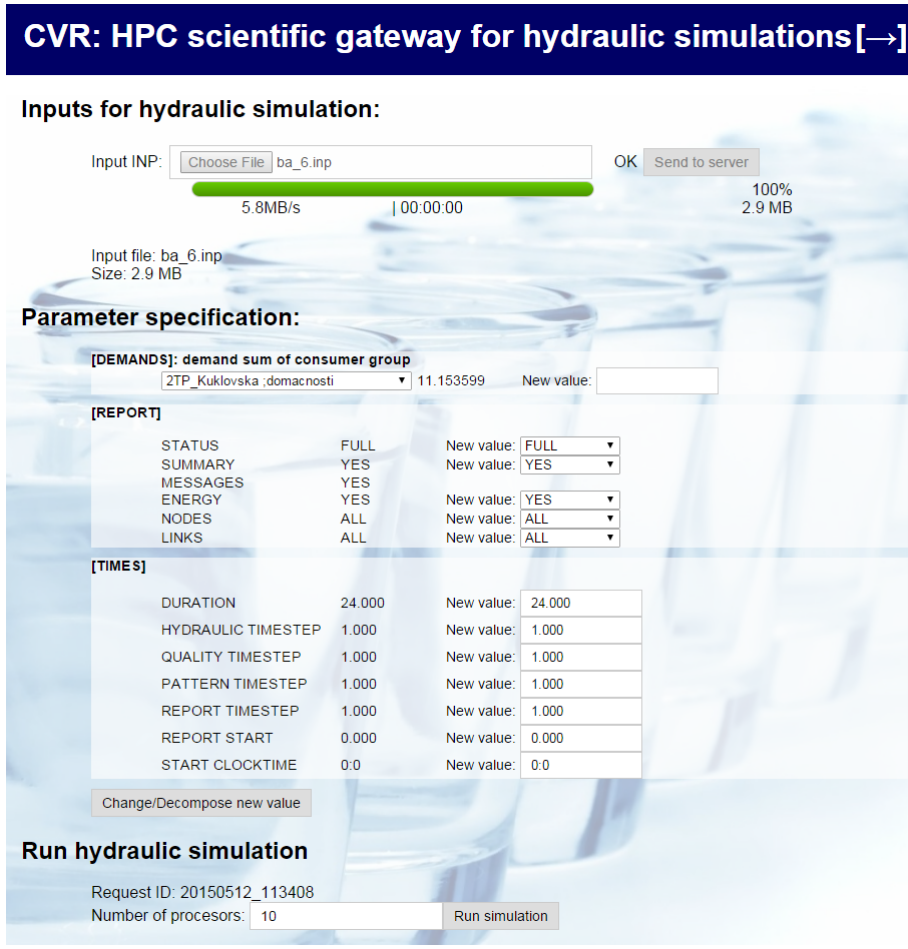

Figure 4. Scientific gateway for hydraulic simulations – modifying input parameters and demands on water consumption for households

and to collect all requirements necessary for submitting his job and performing simulations. The gateway in the web server is responsible also for the initial checking and validation of asked input values.

4.2 Application Server and Monitoring Daemon

While the function of the web server is more or less passive, the monitoring daemon running in the application server (see Figure 5) plays a more active role. It monitors the user's requests coming through the web server and provides for the job execution following the given demands. The daemon is responsible for accepting job inputs, it

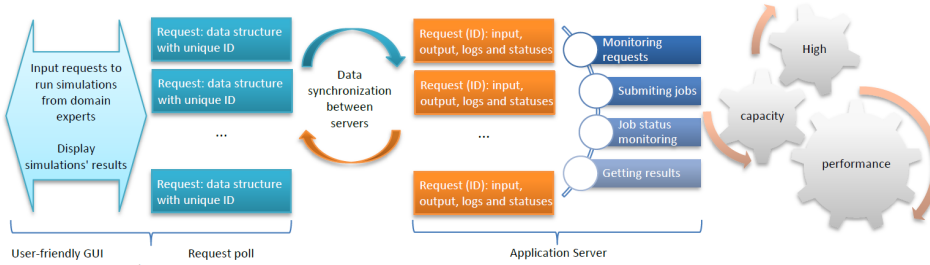

Figure 5. Communication between web server and application server

must understand how to process them, i.e. it must be able to compose and submit the application along with input data. It masters also the technique of communication with the web server. The daemon works either with the access authorization delegated by an operator or it can use the operator's grid certificate with a limited range of activities [39]. A valid authorization to access the grid resources is based on the policies of virtual organization (VO). There are also other alternatives for the management of user identities [38], but they have been not implemented yet.

The job execution process through the gateway begins by sending a user request which is placed in the request pool of the web server together with the requirements specification and input data collected by the interface. The user has the possibility to keep track of the job execution process. The states displayed through the gateway interface are the following:

Waiting: the state includes two cases – waiting in the web server's pool or pending for the submission in the application server pool. These states were joined to one since users are not very interested in implementation details.

Submitted: the state is achieved when the job is in a queue waiting for a suitable resource to run in the cluster or grid.

Running: the state indicates that the job is running.

Finished: the state indicates that the job is completed and its output results are transferred from the application server to the web server. The user can retrieve the output through the web interface of the gateway.

Cancelled: the state indicates that the job was cancelled. The user can cancel his simulation request at any time, if necessary.

Failed: the state indicates that the job execution failed, the user gets the error report as an output.

The gateway job states are a little different from the job states which appear in the cluster environment (PBS: *Held, Queued, Waiting, Running, Completed, Exiting, Suspend, Moved*) or in the grid environment (EMI: *Submitted, Waiting, Ready, Scheduled, Running, Aborted, Done, Cleared*). They represent some combinations of both in order to provide a unified and simplified cases.

Our gateway approach is relatively generic and commonly usable. It can be adopted for carrying out not only hydraulic simulations but also other applications. We have used it for nanoscale simulations [40] and also for Quantum Monte Carlo calculations.

4.3 EPANET and Parametric Study

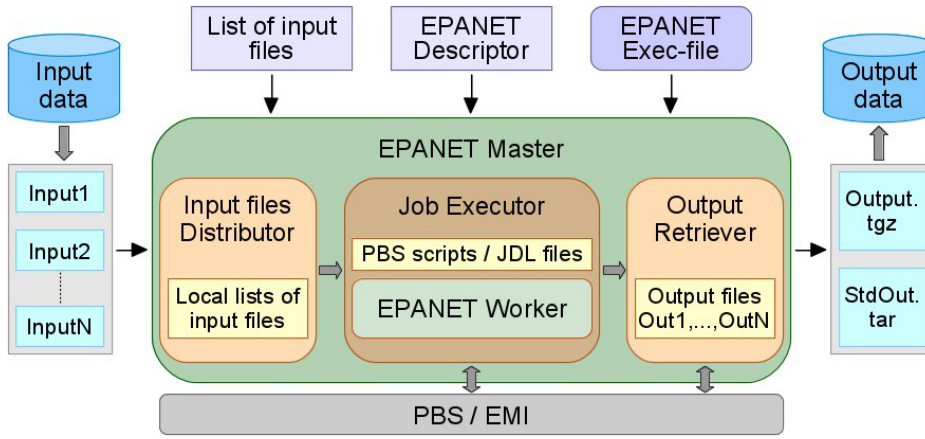

Figure 6. Parametric study for hydraulic simulations

One of the supporting tool was designed to realize the parametric study performing hydraulic simulations with the model EPANET. The tool is built on top of PBS [30] and/or EMI [32] middleware. It serves primarily as a support mechanism in the process of calibration and verification of the model. In simulations input scenarios describing the water-supply network of Bratislava and its surroundings (the owner is "Bratislava Water Company" [22]) are handled. On the cloud infrastructure the distribution of calculations in the parametric study is similar as on the local cluster, with the difference that the role of PBS is superseded by the middleware OpenStack.

A Parametric study (PS) represents an important class of applications which is the case of many scientific and engineering contexts. Typically, the PS is defined as a set of experiments, where the same program is carried out, but with a different set of input data. Due to the inherent parallelism, PS can be performed in a distributed fashion what can significantly reduce the execution time. In our case, the set of input files can be divided into a number of independent parts of arbitrary sizes which may be processed separately in any order on different hardware resources. The tool is constructed following the Master-Worker parallel processing scheme, where the role of the Master is to manage the entire process of performing the PS with the given set of input files, while the execution of the application itself is accomplished through the Worker. The tool is implemented as a command-line based

on the pair of executable scripts *EPANET-Master* and *EPANET-Worker* using one configuration file *EPANET-Descriptor* and services offered by the middleware of the target computing platform. The scheme of the tool is outlined in Figure 6.

EPANET-Descriptor – is a configuration file which defines all information about the PS. It includes names of all files and folders accessed, options for simulation logging, and the number of CPU cores used simultaneously to perform the PS – the value of this entry determines also the number of jobs in the PS. Each job will be run separately on the allocated CPU core and will handle one or more input files.

EPANET-Master – represents an executable script which is started up on a user-interface machine. It consists of the following components:

- At first, it accepts and checks all input parameters specified through the configuration file. Parameters values are passed to the subsequent operations.
- *Input files Distributor* – divides the set of input files into subsets assigned afterwards to separate PS jobs for simulating. The number of subsets is defined by the number of specified CPU cores. The robustness of subsets is computed by a strategy to ensure that each processor (except the last) one gets the same workload.
- *Job Executor* – controls the execution of simulations. For each job of the PS it prepares the *job description file* based on the input parameters and commands of the middleware, PBS for the local cluster, or JDL (Job Description Language) [41] in case of grid. Job description files serve as an input for submission commands.
- *Output Retriever* – is responsible for gathering results of all PS jobs. It monitors the execution status of individual jobs, and in case of their completion it transfers results from the computing resource to the user interface machine. At the end, all output files are integrated into one archive file representing the final output of the PS.

EPANET-Worker – is an executable script which runs on the target computing platform. Its function is to perform the simulation consecutively with each file included in the assigned local list.

4.3.1 Data Mining Scenario and Data Preprocessing

Hydrological domain represents one of the appropriate field for a data mining application. Particular hydraulic systems represent deterministic processes which contain quite enough significant patterns and data relations. In this domain, there are usually numerical attributes that are related to specific exact physical features, such as: volume, pressure, flow and water-level. It allows defining the task and using the statistical numerical prediction method. However, more error features has an influence to this domain except of these deterministic effects. These error features have

a dominant stochastic character. It is very difficult to model these errors physically but if we take them into the consideration, we are able to increase the model quality and error predictions. The comparison of physical and statistical model can provide an interesting and inspirational view of the presented problem. The use of the methods of backward reasoning from statistic models could bring the improvement of physical models in the future. In the last years, the hydrological data from rivers, water reservoirs and drinking water distribution systems are collected and stored in warehouses and databases automatically. This provides the main factor for obtaining large data set and then allows a successful statistical model application. By the collecting of hydrological data (in this case data from water distribution system), we are able to use data in data mining process for many different tasks:

- time estimation of water consumption of reservoir during the pumps are turned off,
- time estimation for filling up of reservoir to defined range,
- defining of optimal strategy for water chlorination to get the chlorine concentration in specified range,
- defining of effective strategy of water pumping into the reservoir by using of night electrical current.

Figure 7. Bratislava: Kuklovská location (yellow dot) and Koliba location (red dot).

We have used historical water consumption data from the surrounding of Bratislava, especially Kuklovska locality. These data are stored in CSV (Comma Sep-

arated Values [23]) files, each of them contains information about the consumption during the period of weeks and an appropriate time information. In the preprocessing phase, data files were integrated and chronologically sorted. In the next step, data records were resampled by using a biquadratic interpolation to arrange a constant time delay between two measurements.

It is appropriate to model water consumption for the weekdays and weekends separately, because consumption during the weekdays has a significantly different distribution in comparison with the consumption during weekends. In our experiment, we have focused on weekdays, because weekend data set contains not enough records.

A daytime attribute (represented in minutes) was chosen as a primary input attribute for the water consumption modelling. Also other extra input attribute was defined based on a part of the year. This added input attribute contains a real number values between 0 and 1; 0 represents date 1st January, and 1 represents date 31st December.

4.3.2 Model Training and Testing

The water consumption depends on time and its graph is depicted in Figure 8. Each point in the graph represents one performed measure. An evident nonlinearity of relation as well as stochastic character in the graph can be seen. It is visible that the water consumption also depends on the air temperature. Considering the fact that the air temperature data in a specific locality had not been available, therefore we had to use data from a nearby locality. WEKA tools [1] were used for models training with 20-fold cross validation for evaluations.

Following machine learning approaches were used and compared:

- Neural Network of MultiLayer Perceptron (MLP): regressor contains 32 neurons in hidden layer with 0.01 ridge penalty factor and tolerance 1.0E−6.
- Radial Basis Function (RBF): regressor has also 0.01 ridge penalty factor, and use for regression 32 basis functions and tolerance 1.0E−6
- Regression Trees M5P [4, 5] use traditional C4.5 algorithm for building the tree. M5P model were trained traditionally and with meta-learning methods
  - Additive Regression [3] and
  - Bagging [2] methods

  The maximum information gain criterion is replaced with the minimum dispersion criterion, and each leaf contains the linear regression model. Minimal records per leaf were set to 6. All three models use the pruning method and give similar performance ratings (Table 1).
- Advanced learning methods – Bagging and Additive Regression [3] method, were also used with REP-Tree (Reduce Error Pruning Tree) and Decision Stump

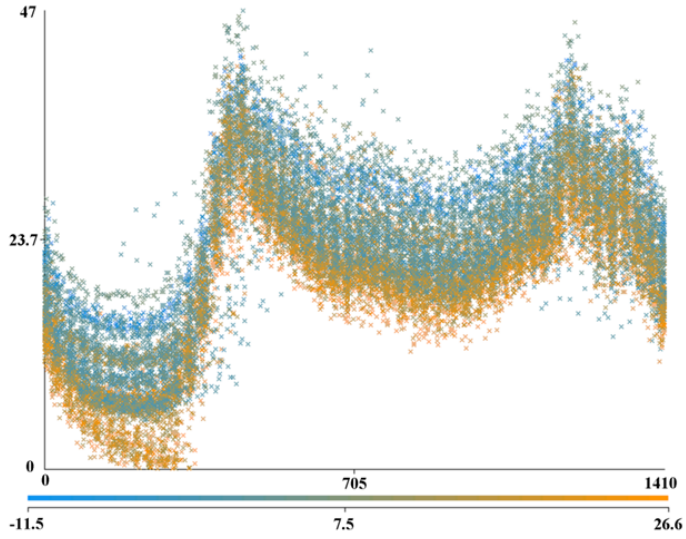

Figure 8. Water consumption distribution in l/s (y-axis) based on workday time (minutes x-axis). Color of marker indicates a temperature.

models. Training models reached the solid performance results in the evaluation phase. A relative absolute error occurence, which is slightly above 36 %, represents the positive result for modeling a significantly stochastic process.

4.3.3 Model Quality Measurements

For the performance comparison we used: Correlation Coefficient (3), Root Mean Squared Error (1) and Relative Absolute Error (2) criterions. Also the relative squared error is more often used than the relative absolute error, for the representation of relative error. But the relative absolute error is the solid criterion, if the target attribute contains some zero values.

Let variables

- r represent the predicted values,
- y represent the actual values and
- N be the number of data records

then:

$$\textit{empirical mean } \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$\textit{empirical mean } \bar{r} = \frac{1}{N}\sum_{i=1}^{N} r_i$$

$$root\ mean\ squared\ error\ RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i - y_i)^2} \tag{1}$$

$$relative\ absolute\ error\ RAE = \frac{\sum_{i=1}^{N}|r_i - y_i|}{\sum_{i=1}^{N}|y_i - \bar{y}_i|} \tag{2}$$

$$Pearson's\ correlation\ coefficient\ PCC = \frac{S_{RY}}{\sqrt{S_R S_Y}} \tag{3}$$

where:

$$covariance\ S_{RY} = \frac{\sum_{i=1}^{N}(r_i - \bar{r})(y_i - \bar{y})}{N-1}$$

$$covariance\ S_Y = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}$$

$$covariance\ S_R = \frac{\sum_{i=1}^{N}(r_i - \bar{r})^2}{N-1}$$

Other criterions such as Mean Squared Error (MSE), Relative Squared Error (RSE), Root Relative Squared Error (RRSE) were also considered for testing purposes, where:

$$mean\ squared\ error\ MSE = \frac{1}{N}\sum_{i=1}^{N}(r_i - y_i)^2$$

$$relative\ squared\ error\ RSE = \frac{\sum_{i=1}^{N}(r_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}$$

$$root\ relative\ squared\ error\ RRSE = \sqrt{RSE}$$

4.3.4 Model Performances

It is evident, that water consumption variable depends on several physical variables, it includes information about weather. Temperature of atmosphere is a solid relevant candidate for adding input attribute, which represents information about weather. But the atmosphere temperature for the analyzed range of dates from Kuklovska location was not available. So the atmosphere temperature from Kuklovska location was replaced by the atmosphere temperature from Koliba location. These two temperature sets are definitely not equivalent, or collinear, but they have a strong correlation. The quality of models is obviously increasing after adding the extra temperature attribute (from the nearby Koliba locality) into the training dataset. The temperatures for the two locations have marked a dependency.

List of the best models after adding a temperature attribute is shown in the Table 2. The precision of models is evidently on a higher level. Parameters of models are the same, as for above Table 1 . Some of models in Table 2 reached significantly

Model	PCC	RMSE	RAE
Bagging M5P Regression Tree	0.9337	2.934	0.36243
Regression Tree M5P	0.9324	2.961	0.36551
Bagging with REP-Tree	0.9319	2.972	0.37078
Additive Regression M5P Regression Tree	0.9301	3.010	0.37302
MLP-Regressor	0.9239	3.134	0.37927
RBF-Regressor	0.9210	3.190	0.38949
Additive Regression Decision-Stump	0.9117	3.382	0.38128

Table 1. Comparison of models with the best performance for water consumption prediction

Model	PCC	RMSE	RAE
Additive Regression M5P Regression Tree	0.9717	1.935	0.20932
Bagging with REP-Tree	0.9671	2.086	0.22520
Bagging with M5P Regression Tree	0.9548	2.438	0.26781
M5P Regression Tree	0.9532	2.479	0.27753
MLP-Regressor	0.9409	2.773	0.31517
RBF-Regressor	0.9008	3.556	0.49300
Additive Regression Decision-Stump	0.8706	4.043	0.49367

Table 2. Comparison of models with the best performance for water consumption prediction after adding atmosphere temperature attribute

higher quality than in Table 1. There are M5P with Additive Regression, and REP-Tree with Bagging. The improvement of relative absolute error is around 15 %; Pearson's correlation coefficient (PCC) of the best model is 0.9717, which is very positive result allowing to use the model in many related problems and tasks, such as time estimation of water consumption or time estimation for filling up the reservoir. However, some of models reached in Table 1 worse performance (RBF-Regressor and Decision Stump with Additive Regression models) probably due to an over-fitting effect. In the future, we plan to apply the model as submodel for the effective strategy of water pumping by using the night electrical current and optimal strategy for water chlorination. Water consumption model will be connected as an input feature to the chlorination model with a purpose to increase the chlorination model quality.

4.4 Post-Processing of Hydraulic Simulations

Hydraulic simulations described in this paper produce output results in the structured text format which is quite specific and not very suitable for further processing or porting back to the GIS environment. Therefore, the transformation into the format CSV (Comma Separated Values) [23] is required. CSV is a common, relatively simple file format which is widely supported by various applications and environments. Its most common use includes the tabular data exchange between programs

that use the native incompatible formats. Most of these programs support CSV, at least, as an alternative format to import and/or export data. In practice, it refers to the CSV files that:

- contain a clear text encoded e.g. ASCII, Unicode, etc. and a clear format
- consist of records which are divided into the same number of fields separated by a reserved character as a comma, semicolon or tab.

Figure 9. Water velocity (m/s) on pipe links by color range indicators

Even within these general constraints the variations of this format are used and many implementations allow users to specify a delimiter character, use quotes, etc.

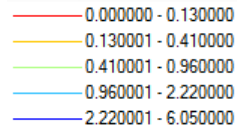

Figure 10. Color range indicators – velocity m/s

on different platforms for different purposes. CSV format is more strictly defined as generic text format and it is frequently selected as a format for data storage. The process of importing data into different IDE work environments is easier and more precise with CSV data, as the freedom in TXT format often tolerates various bugs.

In our framework, the transformation of simulation outputs into the CSV format was implemented in Python language [42] due to portability between platforms and cooperation with geographic integrated environment GIS. An example of the output transformation from the text format (.txt) through CSV format (.csv) using Python supporting scripts, and consequently to geographical database format (.mdb) using ArcGIS geographical tools. Illustrated example of color range indications for water velocity (one of simulation output attributes) in Bratislava water-supply system at 11 hours simulation time of the sixth scenario without group-consumer water demand modifications is shown in Figure 9.

5 CONCLUSIONS

The solutions for each society or a global community must withstand the requisition that professional users i.e. domain experts will consider as adopting new methods if there is an incremental path which avoids excessive learning barriers. That path must also allow them to balance a potential gain against the risks of lost time and opportunities while discovering how to apply the methods to their problems. The transferability of methods and technologies is important, the turbulence of the digital revolution means that different groups are experiencing different rates of changes in their data environment. The aim of our work is to establish the support for hydraulic domain expert category, so that they get a chance to contribute to solutions and to innovate new data-intensive strategies. The conceptual framework is intended to allow independent thinking in each context, but also to allow a collaboration and stimulation across the categorical boundaries. The technological architecture is shaped to facilitate that autonomy with a communication if beneficial. This paper briefly described a core of our research work on the enviromental domain. More details on data preparation, data visualization, collaborations among partners, sensor systems, daily terrain work, difficulties and problem solving are available in [43].

Acknowledgments

This work is supported by projects KC-INTELINSYS ERDF ITMS 26240220072, CLAN APVV-0809-11, EGI-Engage EU H2020-654142, VEGA 2/0054/12 and CVR ITMS 26240220082. Simulations and technical realization were achieved on the hardware equipment obtained within SIVVP ERDF ITMS 26230120002. We would like to thank to Dr. Gibala and Dr. Tóthová from DHI-Slovakia [21] and BVS [22] colleagues for collaboration, scenarios and consultations on hydraulic domain.

REFERENCES

[1] Hall, M.—Frank, E.—Holmes, G.—Pfahringer B.—Reutemann, P.—Witten, I. H.: The WEKA Data Mining Software: An Update. ACM SIGKDD Explorations Newsletter, Vol. 11, 2009, No. 1, pp. 10–18. Available on: http://www.cs.waikato.ac.nz/ml/weka/.

[2] Breiman, L.: Bagging Predictors. Machine Learning, Vol. 24, 1996, No. 2, pp. 123–140.

[3] Friedman, J. H.: Additive Regression – Stochastic Gradient Boosting. 1999.

[4] Wang, Y.—Witten, I. H.: Induction of Model Trees for Predicting Continuous Classes. The 9th European Conference on Machine Learning, 1997.

[5] Quinlan, R. J.: Learning with Continuous Classes. The 5th Australian Joint Conference on Artificial Intelligence, 1992, pp. 343–348.

[6] Project: Research Center for Risks of Water-Supply of a big City (ITMS code: 26240220082) Funded by ERDF Resources Supported by OP Research and Development. Available on: https://cvr.ui.sav.sk/.

[7] EPANET Software (US Environmental Protection Agency). Available on: http://www.epa.gov/nrmr/wswrd/dw/epanet.html.

[8] EPANET 2 Users Manual, EPA/600-R-00-057, September 2000. Available on: http://nepis.epa.gov/Adobe/PDF/P1007WWU.pdf.

[9] MIKE URBAN – Urban Water Modeling. Available on: http://mikebydhi.com/Products/Cities/MIKEURBAN.aspx.

[10] WaterGEMS – Water Distribution Modeling and Management. Available on: http://www.bentley.com/en-US/Products/WaterGEMS/.

[11] KYPIPE – Pipe Network Analysis. Available on: http://kypipe.com/kypipe/.

[12] InfoWater – GIS Integrated Water Distribution Modeling and Management. Available on: http://www.innovyze.com/products/infowater/.

[13] H2ONET – The Most Powerful and Complete Water Distribution Modeling, Analysis and Design Software. Available on: http://www.innovyze.com/products/h2onet/.

[14] AFT – Fathom Pipe Flow Analysis and System Modeling. Available on: http://www.aft.com/products/fathom/.

[15] SynerGEE – Advanced Water Distribution Analysis. Available on: http://www.gl-group.com/en/water/SynerGEEWater.php.

[16] ERACLITO – Modular System for the Management of Fluid Underpressure Networks and Open Channels Systems. Available on: http://www.proteo.it/prodotti/eraclito.asp.

[17] CROSS – Hydraulic Calculation of Water Supply Networks. Available on: http://www.rehm.de/produkte/waterpac/default.aspx.

[18] SCADA – Supervisory Control and Data Acquisition. Available on: http://scada.com/.

[19] ArcGIS – Software to Create, Manage, and Share Geographic Data, Maps, and Analytical Models. Available on: http://www.arcgis.com/.

[20] QGIS – A Free and Open Source Geographic Information System. Available on: http://www.qgis.org/en/site/.

[21] DHI Slovakia, s.r.o., Available on: http://worldwide.dhigroup.com/sk.

[22] Bratislava Water Company (BVS – Bratislavská vodárenská spoločnosť, a.s.) Available on: http://www.bvsas.sk/sk/.

[23] CSV – Comma-Separated Values. Available on: http://en.wikipedia.org/wiki/Comma-separated_values, 2014.

[24] MDB – Microsoft Access Database. Available on: http://en.wikipedia.org/wiki/Microsoft_Access, 2013.

[25] SIVVP – Slovak Infrastructure for High Performance Computing. Available on: http://hpc.ui.savba.sk/.

[26] Foster, I.—Kesselman, C.: The Grid 2: Blueprint for a New Computing Infrastructure. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 2003, ISBN 1558609334.

[27] Foster, I.—Kesselman, C.—Tuecke, S.: The Anatomy of the Grid: Enabling Scalable Virtual Organizations. International Journal on Supercomputer Applications, Vol. 15, 2001, No. 3, pp. 200–222

[28] Foster, I.—Kesselman, C.—Nick, J. M.—Tuecke, S.: The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. Open Grid Service Infrastructure WG, Global Grid Forum, June 22, 2002.

[29] EGI – European Grid Infrastructure. Available on: http://www.egi.eu/.

[30] PBS – Portable Batch System. Available on: http://www.mcs.anl.gov/research/projects/openpbs/ and http://www.adaptivecomputing.com/.

[31] EMI – European Middleware Initiative. Available on: http://www.eu-emi.eu/.

[32] EMI User's Guide, April 24, 2013. Available on: https://edms.cern.ch/document/674643/.

[33] EGI-InSPIRE – Integrated Sustainable Pan-European Infrastructure for Researchers in Europe, RI-261323. Available on: http://www.egi.eu/projects/egi-inspire/.

[34] Helix Nebula – the Science Cloud. Available on: http://helix-nebula.eu/.

[35] Cloud Plugfests – the Cloud Interoperabilty Plugfest. Available on: http://www.cloudplugfest.org/.

[36] OpenStack – Open Source Software for Building Private and Public Clouds. Available on: http://www.openstack.org/.

[37] Ajax programming. Available on: `http://en.wikipedia.org/wiki/Ajax_(programming)`, 2014.

[38] Kacsuk, P.—Farkas, Z.—Kozlovszky, M.—Hermann, G.—Balasko, A.—Karoczkai, K.—Marton, I.: WS-PGRADE/gUSE Generic DCI Gateway Framework for a Large Variety of User Communities. Journal of Grid Computing, Vol. 10, 2012, No. 4, pp. 601–630.

[39] Tugores, M. A.—Colet, P.: Web Interface for Generic Grid Jobs, Web4Grid. Computing and Informatics, Vol. 31, 2012, No. 1, pp. 173–187.

[40] Nguyen, G.—Hluchý, L.—Tóbik, J.—Šipková, V.—Dobrucký, M.—Astaloš, J.—Tran, V.—Andok, R.: Unified Nanoscale Gateway to HPC and Grid Environments. Proceedings of the Symposium on Information and Communication Technology, ACM, 2014, pp. 85–91, ISBN 978-1-4503-2930-9.

[41] Pacini, F.: Job Description Language Attributes Specification. Available on: `https://edms.cern.ch/document/590869/`, 2011.

[42] Welcome to Python.org Programming Language for Effective System Integrations. Available on: `https://www.python.org/`.

[43] Proceedings of the Conference: Inovatívne informačno-komunikačné technológie vo vodnom hospodárstve. Bratislava, November 2014, pp. 41–46, ISBN 978-80-89535-16-3.

[44] Paralič, J.—Hubal, M.: Flexible Support for Knowledge Discovery. Journal Automation Computers and Applied Mathematics, Vol. 12, pp. 41–52, 2003. Available on: `http://people.tuke.sk/jan.paralic/papers/ACAM-2003.pdf`.

[45] Hluchý, L.—Krammer, P.: Optimized Computing of Parameters for Functional Regression in Data Mining. Proceedings of the Conference on Fuzzy Systems and Knowledge Discovery, IEEE Computer Society, 2012, pp. 1617–1623, ISBN 978-1-4673-0023-0.

[46] Hluchý, L.—Krammer, P.—Habala, O.—Šeleng, M.—Tran, V.: Advanced Data Integration and Data Mining for Environmental Scenarios. Proceedings of the International Symposium on Symbolic and Numeric Algorithms for Science Computing, IEEE Computer Society, 2011, pp. 400–406, ISBN 978-0-7695-4324-6.

[47] Nguyen, B. M.—Tran, V.—Hluchý, L.: A Generic Development and Deployment Framework for Cloud Computing and Distributed Applications. Computing and Informatics, Vol. 32, 2013, No. 3, pp. 461–485.

Giang Nguyen is Scientific Researcher at Institute of Informatics, Slovak Academy of Sciences, her main research topics include distributed computing and knowledge discovery. She received M.Sc. and Ph.D. degree in applied informatics from the Slovak Technical University in Bratislava. She is (co-)author of numerous scientific papers and has participated in EU, international and national research projects. She also is a member of program committees and reviewer for international scientific conferences.

Viera Šipková is Researcher at the Institute of Informatics of the Slovak Academy of Sciences. She received her M.Sc. degree in mathematics and the RNDr. degree in computer science from the Comenius University in Bratislava. Her research interests are parallel and distributed computing technologies focused on developments, and porting complex scientific applications. She has more than 10 years working experience at the Vienna University and Vienna University of Technology, and she has participated in many national and international research projects.

Peter Krammer graduated from the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology in Bratislava, and is currently Researcher and the Ph.D. candidate at the Institute of Informatics of the Slovak Academy of Sciences. His research interests include data mining and machine learning. He is (co-)author of several scientific papers and has participated in international and national research projects.

Ladislav Hluchý is the head of the Parallel and Distributed Computing Department (IISAS), and he was the director of II SAS (Institute of Informatics, Slovak Academy of Sciences) for more than 10 years, R & D Project Manager, coordinator and WP leader in more than 20 EU FP projects and outstanding national R & D projects, a member of IEEE, e-IRG, EGI Council, the editor-in-chief of the CC journal, (co-)author of scientific books and numerous scientific papers.

Miroslav Dobrucký is Researcher at the Institute of Informatics of Slovak Academy of Sciences. He graduated from the Slovak Technical University in Bratislava in 1986. His main research topics include high performance and distributed computing (MPI, grid, cloud). He is (co-)author of scientific papers and has participated in EU, international and national research projects.

Viet Tran is Scientific Researcher at the Institute of Informatics, Slovak Academy of Sciences with research focused on distributed computing and cloud computing. He received his M.Sc. and Ph.D. degrees from the Slovak Technical University in Bratislava. He has participated in a number of EU, international and national research projects as a team leader, or work-package leader. He is a scientific coordinator for several national projects, (co-)author of scientific books and more than 100 scientific papers, member of program committees, reviewer for international scientific conferences.

Ondrej Habala graduated from the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology in Bratislava in 2001 and since then he is Researcher at the Institute of Informatics of the Slovak Academy of Sciences. His research interests include complex distributed systems, cloud computing and data analytics including process mining. He is (co-)author of numerous scientific papers and has participated in international and national research projects.