International Environmental Modelling and Software Society (iEMSs)
2010 International Congress on Environmental Modelling and Software
Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada
David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.)
http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings

Integration of mathematical models in a decision support system for control of priority pollutants in urban catchments

Natasa Atanasova [a], Matej Cerk [b], Primoz Banovec [b], Mateja Skerjanec[b], Webbey De Keyser [c], Lorenzo Benedetti [c]
[a] *University of Algarve, International Centre for Coastal Ecohydrology (ICCE), Av. 16 de Junho 8700-311 Olhão, Portugal (natasa.atanasova@icce.com.pt)*
[b] *University of Ljubljana, Faculty of Civil and Geodetic Engineering, Jamova 2, 1000 Ljubljana, Slovenia (matej.cerk@fgg.uni-lj.si; primoz.banovec@fgg.uni-lj.si; mateja.skerjanec@fgg.uni-lj.si)*
[c] *Ghent University, Department of Applied Mathematics, Biometrics and Process Control, Coupure Links 653, B-9000 Ghent, Belgium (webbey@biomath.ugent.be; lorenzo.benedetti@ugent.be)*

Abstract: A decision support system (DSS) for systematic control of priority pollutants (PP) sources, based on economic activities and production (release) processes in urban catchments was recently developed. One of the crucial functionalities of the DSS is evaluation of source control measures, which is based on mathematical models, used for simulating the fate of the PPs in urban catchments under different conditions. This work presents a methodology for efficiently building and integrating dynamic mathematical models into the DSS. A combination of two modelling approaches is proposed: empirical or data-driven and mechanistic or knowledge-driven. Data-driven methods, particularly those from the area of machine learning (ML), are proven to build simple and accurate models, but require a lot of measured data for their construction, which is a problem in the case of PP. Mechanistic models can overcome the data requirement problem by integrating expert domain knowledge in the model formulation. However, they tend to be too complex and computationally slow and thus, not appropriate for DSS. Within the proposed methodology a mechanistic dynamic integrated urban water system (IUWS) model for PPs is used independently of the DSS to simulate various scenarios in observed catchment. Simulated data are used by a ML algorithm for induction of rule-based regression model, which performs similarly as the mechanistic model and is integrated in the DSS. The procedure of model construction, integration, and use in the DSS is successfully illustrated based on semi-hypothetical data.

***Keywords**:* decision support system; integrated urban water system; priority pollutants; rule based regression model

1. INTRODUCTION

Management of PPs is a very complex task as it is highly interdisciplinary and includes numerous sectors (industry, economy, water, agriculture, etc.), decision makers, and stakeholders. For such complex decisions, decision support systems (DSSs) are highly appreciated. In order to be used by decision makers, a DSS should be simple and efficient but at the same time comprehensive enough, taking into account various aspects when supporting decisions.
This is particularly true when supporting the implementation of the EU Water Framework Directive (WFD), as the basic principles encouraged by the WFD lead to an integrative approach, including drivers and pressures on the water resources caused by multiple sectors. To make the implementation easier, IMPRESS [2002] issued the framework of DPSIR (drivers, pressures, state, impacts and responses), also adopted by the European

Environmental Agency (EEA). The DPSIR framework provides an overall mechanism for analyzing environmental problems.
A DSS that follows the DPSIR framework for the management of PPs emissions was recently developed within the EU project ScorePP (www.scorepp.eu). This DSS has a modular structure, where every module plays an important role in the final selection of suitable control measures, (Cerk e al. [2009]).
This paper focuses on one of the key elements of the DSS, namely the integration of mathematical models for simulating the fate of PPs. Mathematical water quality models are used to simulate and evaluate the system's responses to decision makers' actions. However, their integration in the DSS is a challenging task, due to the complexity (1) of the problem itself, (2) of the DSS, and (3) of the models, as they include many sub-models at urban catchment scale, dynamic processes, and (sometimes unknown) parameters. In principal, various types of simulation models can be integrated in the DSS. Two different modelling approaches were evaluated on their feasibility for integration into the DSS: mechanistic models and empirical or data driven models.
Data driven models constructed by machine learning algorithms would be appropriate for such a system, as they are proven to be simple and efficient (e.g. Atanasova and Kompare, [2002], Bhattacharya and Solomatine, [2005]). However, sufficient amount of measured data is needed for such models to be induced, i.e., time series of PP measurements at various points in the catchment, so that the algorithm can be trained to relate the PP fate to the other parameters in the catchment (e.g. PP releases, rain events). Unfortunately this sort of data is extremely scarce, and such an approach cannot be used.
Mechanistic models can overcome the problem of data scarcity by incorporating expert knowledge in the model formulation. However, they appear to be very complex and very demanding in computational resources, which is not efficient for a DSS.
To tackle the problem, a combination of mechanistic and data-driven model construction is proposed, where the mechanistic model is used to simulate various scenarios and to provide sufficient amount of data, which is then used by a machine learning algorithm to synthesize those data and to induce a simple model that performs similarly as the mechanistic model, i.e. a model that requires less input information, is computationally fast, contains simple rules or equations, and produces similar output results as the mechanistic model.

2. THE STRUCTURE OF THE DSS FOR CONTROL OF PPs

A modular approach was used to construct a DSS for management and control of PPs in urban catchments. The DSS is composed of 9 modules, communicating between each other through data exchange in a central database (CDB) that integrates all crucial information for controlling PPs. The modules are:

- *Priority Pollutants module*, which contains the list of all PPs and their inherent properties,
- *Emission Strings module*, which contains classified knowledge about sources of pollution obtained from literature. The module is describing the *driving forces* and in combination with the *GIS module* also the *pressures* (i.e. emissions) in the DPSIR framework.
- *Emission Barrier (EB) module*, which contains classified treatment options and corresponding removal efficiencies. The module also contains information about costs for some treatment options and options for substitution of PPs. It represents *responses*.
- *GIS module* containing geographical information about the city or catchment with additional specific information (called the Adaptation Matrix (AM)) needed for linkage to other modules of the CDB. The module can be used to visualize the *pressures* to the environment at catchment level according to the DPSIR framework.
- *Mathematical module* containing mathematical functions and the data needed to simulate the behaviour of pollutants in the receiving environment. The simulated data represents the *state* of the environment.
- *Source Control Measures (SCM) module*, which is a framework that serves as a container for potential SCM that need assessment. It holds information about existing European legislation concerning PPs. The SCM module represents possible *responses* to improve the environmental *state*.

- *Economical module* currently containing both, 20 and 60 sector classification to be used in economical analysis. The module is used in addition with the Emission Barrier SCM module to help deciding which SCMs are economically feasible and which are not. Therefore it is affecting the *responses* within the DPSIR framework.
- *Monitoring module* containing monitoring information that can be used to provide measured data as charts and maps. The data within the module represents the *state* of the environment.
- *Classifications warehouse,* which contains all the classifications that are used in the project. Among others the classifications are: Chemical Abstracts Service (CAS), Classification of Economic Activities in the European Community (Nomenclature des ACtivites Economiques - NACE), Nomenclature of Units for Territorial Statistics (NUTS), etc.

This paper focuses on the Mathematical module which is indispensable since all the crucial functionalities connected with decision making are linked to it. The main role of the Mathematical module is to calculate the PP loads in the catchment, based on the emission data provided by the other modules in the CDB and to calculate the concentrations of the PPs in the receiving water. It can be also used for the evaluation of different emission control strategies.

3. INTEGRATION OF MATHEMATICAL MODELS INTO THE DSS FOR PP CONTROL

In principle it is possible to integrate various types of models, including the most popular mechanistic urban catchment models, into the CDB's Mathematical module. However, integration of mechanistic models can make the DSS inefficient and difficult to use, due to (1) complex data requirements and preparation to run the model, (2) long computational time and (3) requirement of external tools and software to use the DSS. Therefore, it is proposed to use the mechanistic model independently of the DSS for simulating a set of possible scenarios in the catchment and to provide a sufficient amount of simulated data, from which a simple data-driven model can be induced. This model is later integrated in the DSS and used for simulation and evaluation of various control measures.

4. MECHANISTIC MODELLING OF PPs IN URBAN CATCHMENTS VS. RULE-BASED REGRESSION MODELS

A mechanistic integrated urban water system (IUWS) model for PPs typically consists of different unit process models for each part of the urban water cycle: the simplified KOSIM model (Solvi [2007]) as hydrological catchment runoff and sewer transport model, the ASM2d (Henze et al. [2000]) for activated sludge processes, the Takacs et al. [1991] model for secondary settling, the stormwater treatment unit model for PPs (STUMP) (Vezzaro et al. [2010]) for stormwater infiltration ponds, and the RWQM1 described in Reichert et al. [2001] for river water quality. These state-of-the-art water quality models were recently extended with the fate of PPs (Benedetti et al. [2009]). Additionally, a multimedia fate and transport model was added to the configuration to allow for integrated environmental assessment (cfr. De Keyser et al. (accepted for publication)).

In contrast to the mechanistic modelling approach, where basic theoretical knowledge about the domain is used to formulate a model, regression based models belong to the data-driven modelling domain. The goal of these methods is to learn the dependencies between the inputs and the outputs of the observed system from measured data only.

While the multivariate regression method calculates one equation (one weight vector) for the entire data set, piecewise regression divides the data set to several subsets on which *uniform class value* or *linear equation* can be applied. The division to subsets is based on tests of the values of the input attributes. Two typical representations of the piecewise regression are *regression trees* and *rule-based regression* models.

Rule based regression models have the form of a set of IF THEN rules, where each rule is associated with a multivariate linear model. A rule indicates that, whenever an example satisfies all the conditions, the linear model is appropriate for predicting the value of the

target attribute. The algorithms for rule induction mostly represent different variations of the M5 algorithm introduced by Quinlan [1992]. For this research, the algorithm implemented in the software package Cubist by RuleQuest [2010] was used, where the basic M5 algorithm was enhanced by combining the model-based and instance-based learning (Quinlan [1993]).

5. A SEMI HYPOTHETICAL CASE STUDY

As a semi-hypothetical case study to demonstrate the use of the DSS, the fate of the PP bis(2-ethylhexyl) phthalate (DEHP) was simulated in a simple urban environment consisting of a rural catchment, three urban sewer catchments (A, B and C) connected to an intercepting combined sewer system, a simple activated sludge plant including primary settling, two aerated tanks and secondary settling. The WWTP and the combined sewer overflow (CSO) structures at the three urban catchments discharge to a river, modelled as a series of five completely mixed tanks, each of them in contact with river sediment. The CSO structures were implemented as reactive basins with a given buffer volume. Additionally to the urban water system models, also a multimedia model consisting of seven compartments (air, soil, groundwater and upstream and downstream water compartments each in contact with sediment) was included in the configuration. A scheme of the conceptual setup is shown in Figure 1. Key properties of this integrated environmental model can be found in De Keyser et al. (accepted for publication) and Cerk et al. [2009].

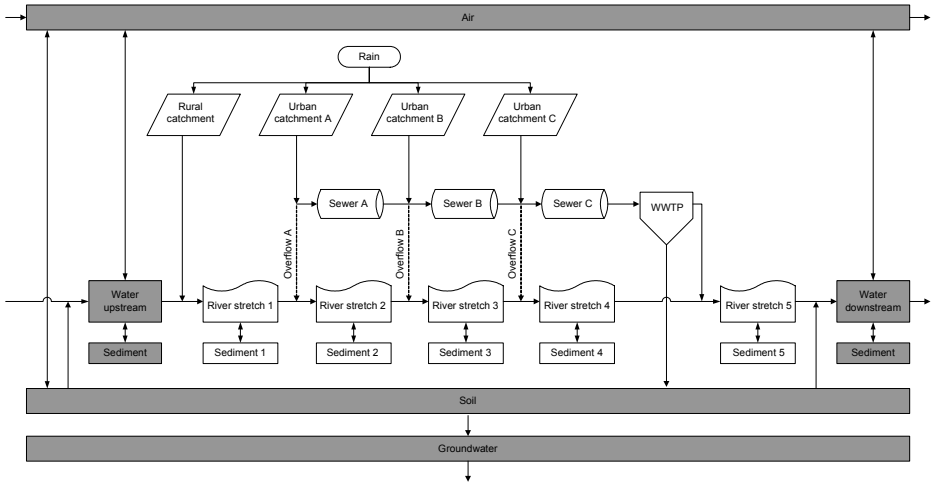

Figure 1. Representation of the integrated environmental model configuration (IUWS model (white blocks) + multimedia model (grey blocks)).

5.1 Using the mathematical module of the DSS to calculate the DEHP loads in the catchment

Querying the CDB for the potential DEHP emissions returns 90 ES, of which 21 were actually identified in the catchment. Table 1 shows some of the identified ES in a generic way. In order to be quantified for the specific case, an adaptation matrix (AM) is needed, which contains (1) the coordinates of each ES (one ES can appear in several points in the catchment), (2) the release factor of each emission, and (3) the compartment to which the ES contributes. The following emission receiving compartments were used: water indirect (WI; emission flows through sewer system and WWTP to the receiving surface water), urban impervious (UI; emission flow paths through impervious surfaces), air (AIR; emission to air), urban permeable (UP; emission flow paths through fields and lawns) and water direct (WD; emission flows directly to the receiving water).

Table 1. Some ESs identified in the catchment

No	Source	ES_ID	ES_type
1	Municipal waste incineration (plastics like PVC).	1416	dumping grounds
2	Production of DEHP at production site.	1462_6	facilities
3	Manufacture of plastic products. Use of DEHP as plasticizer in polymers 97-98% (mainly PVC). Release during industrial use as polymer - formulation and processing	1450_1	facilities

Regarding the localization of the ES, no specific coordinates for emissions are applied. Instead, they are aggregated at sub-catchment level, i.e., each ES can be localized at one of the three locations (urban catchments A, B and C) and later be summarized by sub-catchment and compartment. Aggregated load variables used in the models are presented in Table 2.

Table 2: Load data, aggregated by compartments and sub-catchments

Name	Description	Unit
PREC	Precipitation; summed within the selected time interval (4 hours).	**Mm**
DEHP_WI_ A, DEHP_WI_ B, DEHP_WI_ C	Sum of all DEHP emissions connected to the sewer system in urban catchments A, B and C respectively. This load contributes to the dry weather flow.	**g/day**
DEHP_UI_A, DEHP_UI_ B, DEHP_UI_ C	Sum of all emissions in urban catchments A, B and C respectively that are accumulated on the surface. This load is washed from the surface during rain events.	**kg/(ha·day)**
DEHP_AIR	Sum of all emissions going to the air.	**g/day**
DEHP_UP	Sum of all emissions going to the soil.	**g/day**

Given the AM, the DEHP loads can graphically be shown by using the ES module and the GIS module in the DSS. Figure 2 shows the DEHP loads in the sub-catchment A. Note that the loads can be represented by ES (Figure 2, left), where all emissions in the sub-catchment are shown, or aggregated by compartments (Figure 2, right), where only the amounts of emissions going to a specific compartment are shown.

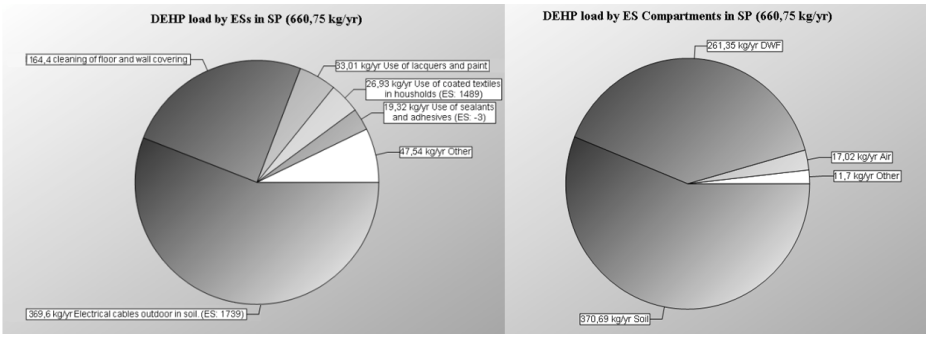

Figure 2. Visualized loads in the sub-catchment A. The loads can be aggregated by ES (left) or by compartments (right).

Calculated loads are further used as inputs to the mechanistic mathematical model to simulate the DEHP concentrations in the urban environment. Like this, a sufficient amount of data is obtained, which is later used by the ML algorithm, to construct a simple rule-based regression model.

5.2 Simulated DEHP concentration in the river with the IUWS model

Given the input data described in the previous section, the IUWS model simulates DEHP loads and concentrations at various locations in the catchment. The model takes the emission data as aggregated loads at sub-catchment level, partitioned for each compartment. Six different scenarios were simulated by changing the input loads and the rainfall time series (Table 3).

Table 3: Simulated scenarios

Scenario	PP load	Rainfall
1	Base scenario with the load of 21 ES converted into dynamic loads and aggregated according to the compartment they contribute to (WI, UI, AIR, UP and WD)	Base case (801 mm/year)
2	the entire load is decreased by 50 %	Base case (801 mm/year)
3	the entire load is decreased by 80 %	Base case (801 mm/year)
4	WI increased by 50%	Base case (801 mm/year)
5	Same as base case	Dry year (539 mm/year)
6	UI reduced by 20%	Dry year (539 mm/year)

Figure 3 presents the simulated DEHP concentrations in river stretch 5 (downstream of WWTP discharge) for three selected scenarios (no. 1, 3 and 4). As expected, the highest peaks appear with the highest input loads (scenario 4 – WI increased by 50%) and the lowest with the loads reduced by 80% (scenario 3).

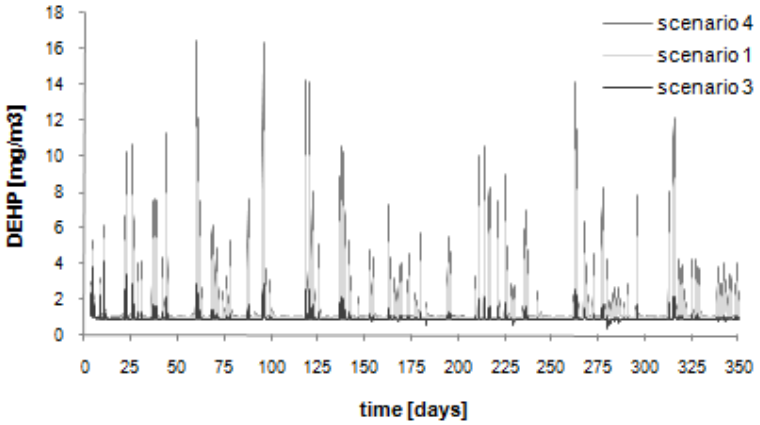

Figure 3. Simulated DEHP concentrations in the river stretch 5 for three scenarios.

5.3 Induction of the rule based model

Using the loads, the precipitation data and the simulated DEHP concentration data, the Cubist algorithm was employed to induce (learn) a rule based regression model to simulate the DEHP concentrations in the river. The attributes, i.e. the load variables were converted from yearly constant loads to yearly dynamic loads using a dynamic load generator developed by De Keyser et al. [2010]. Thus, the data set used by the Cubist algorithm is composed of examples containing the load and the precipitation data (attributes) and the corresponding DEHP concentration in the river (target variable) at each time step. Additionally, attributes were introduced that include six hours of history in the data, i.e. the present DEHP load in the river section depends on the present load from the catchment as well as on the load that was released one to six hours ago. Historical attributes were given additional notation X, where X represents the value of the attribute X hours ago. For example, DEHP_WI_A_3 represents the load from the sub-catchment A that contributes to the dry weather flow, generated 3 hours before. After introducing these attributes, the data set comprised 70 attributes and the class DEHP_river. The data from scenarios 1 to 5 was used as a learning data set. The remainder of the data (scenario 6) was used for testing the model performance.

The model induced on the training data set is composed of seven IF THEN rules each associated with a multivariate regression model. The model achieves high accuracy when simulated on the training data (the correlation coefficient is 0.78) as well as on the test data, where the correlation coefficient is even higher than the one on the training data (0.91). Higher correlation coefficient is observed because the test data set comprises 'only' 8651

points, i.e. simulated data every 5 minutes over a period of one year, whereas the train data set contains 5 minutes data over a period of five years (43260 points). The lower correlation coefficient on the training data is the result of averaging the error (correlation coefficient) over a long period of data. In Figure 4 the performance of the rule based model is presented on the segment of the test data by comparing the data simulated by the IUWS model and by the rule based model.

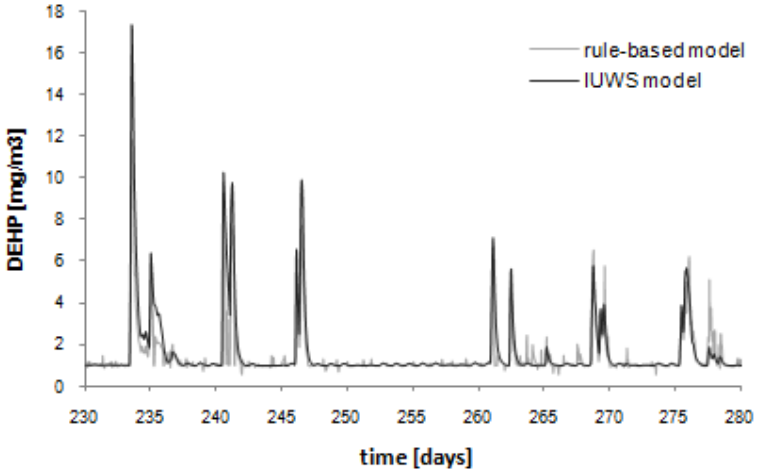

Figure 4. Comparison of DEHP concentrations in the river stretch 5 for scenario 6 simulated with IUWS (blue line) and rule-based model (red line).

5.4 Integration and use of the rule-based regression model in the DSS

A web-based interface was developed to connect all functions of the Mathematical module including calculation of DEHP loads in the catchments, simulating the DEHP concentration in the last river stretch with the rule based model, and evaluating emission barriers by manually changing specific loads and simulating the concentrations again.

The Mathematical module Web Interface is used as a tool in the DSS for interpreting the effects of emission barriers (by reducing specific load in the catchment). The tool is PP and catchment specific. Therefore each different PP or catchment needs the following preparation steps to be performed (1) evaluation of the emission sources in the catchment, i.e. providing quality AM, (2) formulating and simulating a conceptual model for the catchment, (3) construction of data-driven model that mimics the behaviour of the conceptual model, and (4) implementation of the data-driven models into the Mathematical module.

The models that run behind the module are specific to PP, catchment, range of emissions and the end-of-pipe treatment solution. Therefore changing any of these parameters leads to the modification or re-creation of the Mathematical module. Nevertheless, once the models are prepared for a specific catchment and integrated into the catchment-specific DSS, they can serve as an efficient tool for the decision making process in the catchment.

6. CONCLUSIONS

A methodology was presented for efficient integration of dynamic integrated urban water system (IUWS) models for simulating PPs into a modular and multi-objective DSS for control of PPs in an urban environment. The methodology is based on combining the results from this mechanistic model with a data-driven modelling approach. The Cubist machine learning method was used to synthesise the provided simulation data and to construct a simple rule-based regression model, which is easily integrated into the DSS. The methodology was successfully applied to a semi-hypothetical case study to construct a model for DEHP concentrations in the receiving river. The data from previously simulated scenarios with the IUWS model were used to induce a rule based regression DEHP model and integrated in the DSS. The model matches with good accuracy the simulated data by

the mechanistic model. As it is typical for data-driven models, it is case and substance specific, i.e. controlling another substance in the same catchment would require repeating the procedure and induction of another model. However, once a model is induced and integrated into the DSS, it can serve as an efficient tool for the decision making process concerning PPs in the catchment.

ACKNOWLEDGEMENTS

This research has been conducted within the ScorePP - "Source Control Options for Reducing Emissions of Priority Pollutants" project, contract no. 037036, funded by the European Community's Sixth Framework Programme. Mateja Skerjanec is part-financed by the European Union, European Social Fund. Lorenzo Benedetti is a post-doctoral researcher of the Special Research Fund (BOF) of Ghent University.

REFERENCES

Atanasova N. and B. Kompare, Modelling of wastewater treatment plant with decision and regression trees. In: Proceedings of the Workshop in Binding Environmental Sciences and Artificial Intelligence, ECAI, Lyon, 6-1 – 6-9, 2002.

Benedetti, L., W. De Keyser, L. Vezzaro, N. Atanasova, N. Gevaert, F. Verdonck, P.A. Vanrolleghem and P.S. Mikkelsen, Integrated dynamic urban scale sources-and-flux models for PPs, ScorePP project deliverable D7.4, available at: http://www.scorepp.eu, 2009.

Bhattacharya, B. and D.P. Solomatine, Neural networks and M5 model trees in modelling water level-discharge relationship, *Neurocomputing*, 63, 381-396, 2005.

Cerk, M., N. Atanasova and W. De Keyser, Implementation of distributed mathematical models, ScorePP project deliverable D6.4, available at: http://www.scorepp.eu, 2009.

De Keyser, W., V. Gevaert, F. Verdonck, B. De Baets and L. Benedetti, An emission time series generator for pollutant release modelling in urban areas. *Environmental Modelling & Software*, 25(4), 554-561, 2010.

De Keyser, W., V. Gevaert, F. Verdonck, I. Nopens, B. De Baets, P.A. Vanrolleghem, P.S. Mikkelsen and L. Benedetti, Combining multimedia models with integrated urban water system models for micropollutants, *Water Science and Technology* (accepted for publication).

Henze, M., W. Gujer, T. Mino and M. van Loosdrecht, Activated Sludge Models ASM1, ASM2, ASM2d and ASM3. STR No. 9. IWA Publishing, London, UK, 130p., 2000.

IMPRESS, Guidelines for the analysis of pressures and impacts in accordance with the Water Framework Directive, CIS Working Group 2.1, Office for Official Publications of the European Communities, 156 p., 2002.

Quinlan, J.R., Learning with continuous classes. In: Adams & Sterling (eds.), Proceedings of the Australian Conference on AI (AI'92), World Scientific Singapore, 343-348, 1992.

Quinlan, J.R., Combining instance-based and model-based learning, In: Proceedings of 10th Machine Learning Conference, Morgan Kaufmann, San Mateo, California, 236-243, 1993.

Reichert, P., D. Borchardt, M. Henze, W. Rauch, P. Shanahan, L. Somlyódy and P.A. Vanrolleghem, River Water Quality Model No. 1. STR No. 12. IWA Publishing, London, UK, 144 p., 2001.

RuleQuest, Data mining with Cubist, available at: http://www.rulequest.com/cubist-info.html, 2010.

Solvi, A-M, Modelling the sewer-treatment-urban river system in view of the EU Water Framework Directive. PhD Thesis, Ghent University, Belgium, 218 p., available at http://biomath.ugent.be, 2007.

Takacs, I., G.G. Patry and D. Nolasco, A dynamic model of the thickening/ clarification process., *Water Research*, 25(10), 1263-1271, 1991.

Vezzaro L., E. Eriksson, A. Ledin, and P.S. Mikkelsen, Dynamic stormwater treatment unit model for micropollutants (STUMP) based on substance inherent properties, *Water Science and Technology* (In press), 2010.