Hydrodynamic And Water Quality Surrogate Modeling For Reservoir Operation

Aguilar Lopez, Juan; Andel, Schalk Jan Van; Werner, M; Solomatine, Dmitri

Publication date
2014

Document Version
Final published version

Published in
Proceedings of the HIC 2014 - 11th international conference on hydroinformatics

Citation (APA)
Aguilar Lopez, J., Andel, S. J. V., Werner, M., & Solomatine, D. (2014). Hydrodynamic And Water Quality Surrogate Modeling For Reservoir Operation. In M. Piasecki (Ed.), *Proceedings of the HIC 2014 - 11th international conference on hydroinformatics* (pp. 1540-1547)

International Conference on Hydroinformatics

8-1-2014

Hydrodynamic And Water Quality Surrogate Modeling For Reservoir Operation

Juan Aguilar

Schalk-Jan Van Andel

Micha Werner

Dimitri P. Solomatine

11th International Conference on Hydroinformatics
HIC 2014, New York City, USA

HYDRODYNAMIC AND WATER QUALITY SURROGATE MODELLING FOR RESERVOIR OPERATION

J.P. AGUILAR LÓPEZ (1), S.J. VAN ANDEL (2), M. WERNER(2), D.P. SOLOMATINE(2,3)
(1): CTW-WEM, University of Twente, Drienerlolaan 5, 7522 NB Enschede,Netherlands
(2): UNESCO-IHE Institute for Water Education, Westvest 7, 2611 AX Delft, The Netherlands
(3): Water Resources Section, Delft University of Technology, The Netherlands

In order to include the water quality state in the reservoir release policy for long and short term optimization, a forecasting surrogate model was developed using M5 model trees algorithm. This data-driven model is able to recreate the state of the system while reducing the number of input variables from the original numerical model. After testing the methodology it was concluded that the model can represent the discharges in the system accurately based on a single year routing calculation. In terms of the main stream water quality it can also represent the seasonal trends. However it is also concluded that more extreme pollution events need to be included inside the training data set in order to achieve the correct representation of high pollution inflows in the system.

INTRODUCTION

Three main solutions for water quality management are commonly used for environmental remediation of polluted streams. The first approach is the construction of new infrastructural solutions for future projects, like for example pre-treatment plants for tributaries with high organic load. The second one is the development of environmental regulations and policies by the environmental authorities for each particular case, and the last one is by improving the system efficiency in terms of water quality remediation by changing the system operation. For this particular research an operational change for an existing reservoir release schedule was required in order to improve downstream water quality while satisfying the constrains for hydropower production and flood risk management. In order to achieve this goal, a simulation of the downstream water quality was needed as input for the dam operation optimization.

Studies for improving operational management of water systems are more common every day, as land use change, climate change and environmentally un-friendly agricultural practices redefine the initial water quality and quantity required values. The main challenge of this kind of studies consist in generating modelling frameworks that can be used for decision making within reasonable timeframes. For the reservoir release case, decisions are commonly taken in the range of a daily basis framed inside optimal reservoir water level guide curves estimated in a monthly scale. For water quality purposes, the release decisions are normally taken based on the actual state of the upstream system e.g. reservoir water levels, water quality, temperature

and expected inflow towards and from the reservoir. Once the decision is taken, the release is discharged downstream and the water quality inside the stream is assumed to be improved given the upstream optimized decision. However downstream water quality should be taken into account if possible in order to ensure a more holistic approach to the system operation.
Hydrodynamic and water quality latest models are capable of recreating the different scenarios required to estimate the impact downstream for any particular release decision. However, reliable results can only be achieved by highly time consuming modelling procedures. If an optimized result is needed, it might not be possible to estimate with only a 24 hour timeframe. Then surrogate modelling becomes a very low-cost solution for calculation time improvement. It can be used for dimension reduction if possible without compromising the accuracy to a large extent Razavi [7].

Different emulation of systems have been studied mostly for optimization routines, as optimization requires multiple runs which make the process very time consuming. With the use of surrogate modelling, computational time was reduced substantially and created the opportunity of including more reservoirs and water quality parameters for optimization processes in the study performed by Soltani [8]. However, the obtained reservoir operation guidelines, only considering water quality inside the reservoirs. This might not be the optimal solution for if the main drivers of pollution are located downstream. Altunkaynak [1] studied the possibility of forecasting monthly dissolved oxygen concentrations in the Golden Horn estuary in Istanbul based on observed time series in two different stations using fuzzy logic algorithms after identifying possible trends in the data. This methodology was successful for predicting monthly average dissolved oxygen concentrations, but for the present research also short-term predictions are required and therefore the capability of forecasting extreme events as well. Seasonal models can also be built by the use of Artificial intelligence methods, if the trends in the data are correctly identified before the model training process. Preis and Ostfeld [5] studied water flow and water quality forecasting based on coupled model trees for load predictions in a catchment in Northern Israel. As a useful result from this study, it was concluded that the application of model trees for water quantity and quality is significant if long time meaningful data series are available. Data driven models have a good performance in forecasting events that fall within the range of the time series used as training data. However, the determination of the main drivers that influence the system behavior is the most important and complex task in surrogate modeling. Nerini and Durbec [4] were able to predict the concentration of dissolved oxygen in the "Etang de Berre" lagoon using one day lead time, based on only two measured variables (wind speed and inflow) for training binary decision trees algorithms. The study demonstrated that this artificial intelligence methods can be capable of predicting such water quality parameters if sufficient data for training and validation is available.

For this particular study an existent highly detailed MIKE11 model of the river Cauca is available which allows to produce the required training data for building a surrogate model. It’s important to stress that surrogate models are only as good as the original models. Therefore the present study will only focus on evaluating the capacity of artificial intelligence to represent the original model behavior and not the real system behavior. The surrogate model described in the present paper was used for quality prediction based on upstream reservoir releases. The coupling of the reservoir system and the downstream surrogate model where used for short term release optimization purposes. The scope of the present paper only includes the construction and performance evaluation of the present model.

CASE STUDY

The case study was located in the River Cauca in the department of "El Valle del Cauca", Colombia. It has an average discharge of 310 m^3/s measured 220 kilometers downstream of the reservoir in the gauging station “Media Canoa”. The configuration of the system consists on the main stream with more than 40 main tributaries which pollute the water body in different proportions. Some of these tributaries are polluted due to rural activities and some others are used as drainage from urban areas. The system is regulated by an upstream reservoir which makes it a suitable case, for implementing the combined decision making (upstream-downstream) dilution based release approach.
With the use of artificial intelligence models, the system configuration was recreated by building a data driven surrogate model based on an existent MIKE11 model for water quality prediction. This surrogate model reduces the amount of input required for predicting the water quality in the system which makes it faster in order to estimate the required release for a certain event. Finally a forecasting system is implemented in order to estimate the tributary inflows. These predicted discharges are used afterwards as input for the surrogate water quality model used to predict the resultant water quality (DO values) downstream.

Existent model

Between 1975 and 1985 several models were developed for predicting the hydrodynamic and morphological behavior of the river. These models were developed in association with the Delft Hydraulics under cooperation based on the Colombian National Found for Research. This model also helped to study the feasibility and design of the Salvajina Dam in 1985. Between 1977 and 1980 a series of field measurement campaigns were developed for estimating hydrometric and water quality values and parameters in order to build the model. The second phase of the project consisted on evaluating the sediment transport and the definition and implementation of a mathematical model capable of estimating the scouring and settling process occurring in the river Corporación Autónoma Regional del Valle del Cauca – CVC. [3].
In year 2000 the requirement of a water quality model development was included by the CVC (Cauca Valley Environmental Corporation) as part of the PMC (Cauca River Modelling Project). The main goal in terms of water quality was develop a tool able to predict and reflect future change in the river for better management purposes based on the application of local regulations. The hydrodynamic model was first built using the MIKE11 one dimensional package including the solution of advection dispersion equations for water quality modelling. In order to build and calibrate such model the most influential water quality parameters of the system had to be selected. This selection was done based on the available data and the degree of impact in the system in terms of pollutant load and discharge. 31 main tributaries, 12 industrial sources and 5 municipal areas are included as input to the model. The flow routing is done in 2 minute time steps across 414 cross sections along the main stream. The 48 included pollution sources have hourly input time series such as BOD5, OD, water temperature and discharge.

Surrogate Model Input Selection

Because of the complexity of the model in terms of input data and calculation time cost, the development of surrogated model (data-driven surrogate model) was proposed. The main goals are to improve the model efficiency and without compromising the accuracy.

The first step for building the surrogate model is to identify the most important input for the model. Different methodologies exist such as the one at the time, regression coefficients or variance based methods which allows to quantify which input influences the output in a higher degree. However for this case, the input was selected based on the data availability and the amount of organic the tributary carries towards the main stream.
In principle the pollution problem can be classified by the its origin from certain catchment types. The first selected input is the dam outflow. In average the Salvajina dam discharges 130 m^3/s equivalent to 42% of the total discharge measured 220 kilometers downstream. In terms of water quality, Salvajina reservoir has an average dissolved oxygen concentration (OD) of 4 mg/l. This value fluctuates between 2 and 6 mg/l due to the currents generated by temperature gradient inside the reservoir.
From the 31 minor tributaries included in the existent model, the south channel is selected as one of the main inputs to the model. Its located 11 kilometers upstream before the gaugin station of “Juanchito”. This channel flushes part of the water coming by the pluvial drainage system of the city of Cali, the main urban area inside the Cauca river catchment. A real time measuring campaign demonstrated that high peak discharge of organic matter coming from this channel impacts the water quality in the Cauca river (Fig.1). It’s important to note that even though the average discharge of the south channel is 350 times smaller than the average discharge of the Cauca river, it can flush 140 tons/day of organic matter. In average, the Cauca river transports 180 tons/day until this point of discharge.

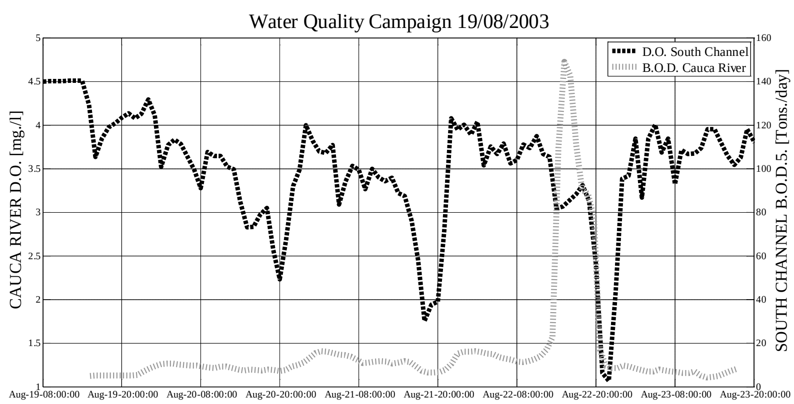

Figure 1. Organic matter transport flushed from the Canal Sur (Cali city) vs. the dissolved oxygen levels of Cauca River at the station Juanchito 8/19/2003 Velez C.A. [9]

In terms of the rural pollution contribution, the Palo River tributary is chosen as the representative input in the emulated model. According to EMCALI (Public utility company of Cali) for the year 2003, the Palo river is contributing to the 26% of the total BOD load to the Cauca river. The mean discharge of this river (35.9 m3/s), also makes it a very important input to the system, taking into account that the mean released discharge from the dam is around 130 m3/s.

The 3 previously mentioned tributaries account for more than 50% of the system discharge and more than 80% of the total organic matter that enters the system during extreme events. Because of this reason and because the availability of complete discharge daily records since

1985, the 3 inputs could have the largest impact in the system behavior in terms of discharge and water quality.

Surrogate Model Output

Once the input points for the model have been chosen, the selection of points of interest where discharge and water quality values are going to be selected as well. The original MIKE 11 model describes 444 kilometers of the Cauca river bounded upstream by the previously mentioned reservoir and downstream by a Q vs. h rating curve. Cross sections of the river are specified almost every 1 kilometer which allows to derive time series for training data in every almost interest point. In addition water levels can also be derived as the rating curve for each station is also available. However only 8 points of interest where selected due to the presence of gauging stations with long enough time series that could be used for validation in the future.

The model works as a cascade configuration, which means that the upstream emulated information is going to be used to predict the next downstream interest point. For each of the selected interest points data driven models are based on previously generated training data. The model works as a cascade configuration, which means that the upstream emulated information is going to be used to predict the next downstream interest point. It is expected that the calculation error from the upstream emulation point is going to propagate downstream but also the downstream point has more information to predict the value than the upstream value. Also water abstractions are implicit in the MIKE 11 model which has to be reflected in the lower stations. As a final result the model will predict for that certain combination of inputs (Dam release, Palo river inflow and South channel discharge) the discharge and dissolved oxygen value at the emulated interest points (Fig 2.)

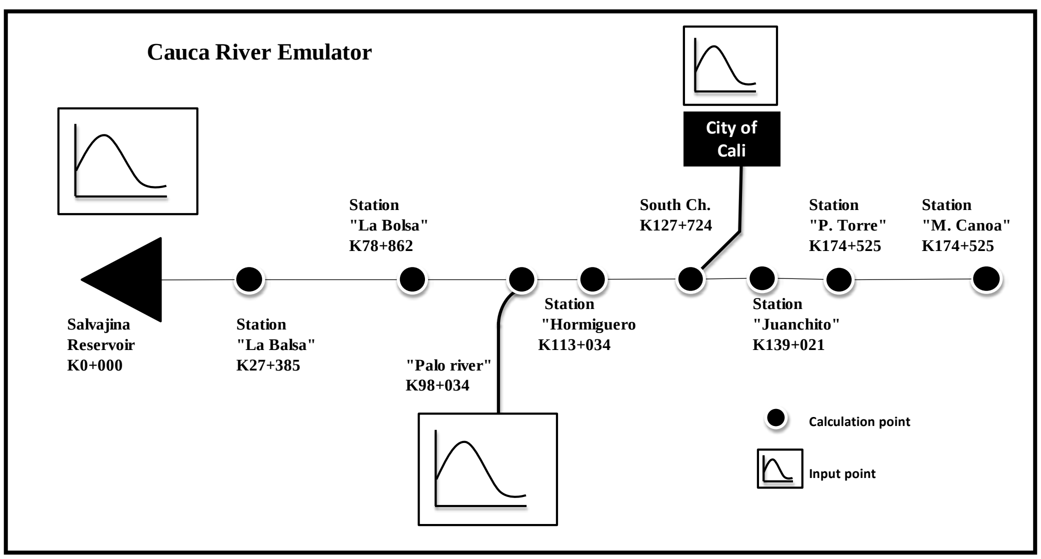

Figure 2. Emulator schematization with input points for discharge and water quality calculation in different points of interest .

This produced output in one point is then used as input by the next emulated calculation point located downstream. This model approach is also known as a "Cascade emulator". For this reason the average travel time is estimated in order to understand which is the required lagging that better represents the system between each downstream emulation.

Surrogate Model Training data

Each interest point was emulated by implementing an M5 model tree regression algorithm. For training each of this emulators, data was generated from the Mike11 original model, by changing the boundary conditions for 3 main selected inputs (Tributaries) with measured data for the year 2002. The rest of the tributaries was regionalized using discharge versus drainage area relations. The hydrodynamics of the system implicitly involve a lagging time for the input signal to become output. This is represented in the routing as the peak wave travel time. The correct lagging in the training data, will ensure higher correlation between the time series which is traduced in a better representation of the system. The wave travel times are influenced by different factors such as roughness coefficients, inertia, cross section geometry and hydrograph base time for hydrograph. All of these factors will influence the response of the system in terms of travel time and peak attenuation. For this reason Triangular hydrographs where routed with different base times and peak values.

For the hydrodynamic data generation, the Cauca River model was modified in its input in order to be able to reproduce the whole seasonal variability encountered in one year. Recorded daily time series of discharges where available for the dam release input. Additional data generated from an HBV hydrological model where also used for the discharge input in the Palo river and south channel. The time series from January 1st of 2002 until 31st of December was routed through the original model. The time series for the small tributaries where generated by simple drainage area regionalization. The simulation generated 8760 data points for each of the 389 cross sections of the model. Surrogate modeling is highly dependent on the initial design experiments in order to cover the most of the input feasible space. However the idea of the present study is to develop a model with the least possible original modelling. In order to build the emulators by M5P tree algorithms Quinlan [6], 70% of the data was used for training and 30% for validation. As predicted values depend on the previously estimating travel times, the training data preserved the chronological order. This decision also ensured to have low and high flows in both data sets.

In terms of water quality, the actual Cauca river Mike11 model is calibrated and configured to simulate only 15 days of summer based on the observed data of the campaign done on the year 2003 which only has low discharge values Coorporación Autónoma Regional del Valle del Cauca – CVC. [2]. For generating sufficient water quality data the procedure was changed from the one used for the hydrodynamic data generation. The water quality input for the Cauca River Mike 11 model is more complex as it works with parameters as BOD, dissolved oxygen and water temperature. These parameters are the result of several biochemical reactions from the catchment produced organic matter as a result of Agro-industrial and domestic processes.
The required input data was generated by estimating the amount of organic matter the enters the system (represented as BOD) as a function of the tributary discharges. The regressions where built using the measured values during the 2003 campaign. Then this regressions where used for the complete discharge time series of year 2002.

One of the main assumptions for this model is that based on the routed discharges and dissolved oxygen prediction are correlated and is mainly affected by volume dilution. The input used for the water quality data generation assumes constant temperature for the three main inputs, and a constant dissolved oxygen concentration of 7.3 mg/l for the "Balsa" boundary and 0 mg/l for the Palo River and South Channel during the whole simulation. The only parameter which is fluctuating is the BOD and is estimated from a measured constant load concentration for the Palo River. The simulation for water quality data generation was done for small period of time as the calibration period for water quality is too short. The second assumption is that the water models obtained from the short period of available data are representative for the whole year.

RESULTS

The surrogate model was cross validated by routing the observed discharge series from 2006. The results show that with such a reduction of input the results are acceptable. The correlation between the surrogated model and the original one is good for the main interest locations. However, the surrogate model has some difficulties in representing correctly the peaks of extreme events during winter months (Fig. 3). This can be attributed to the lack of occurrence of such events inside the training data. As the system is regulated by the upstream reservoir, the extreme events are generated mainly by the influence of the tributaries. This can be observed as the error in the discharge estimation whereas M5P models calculate the final values with linear relationships.

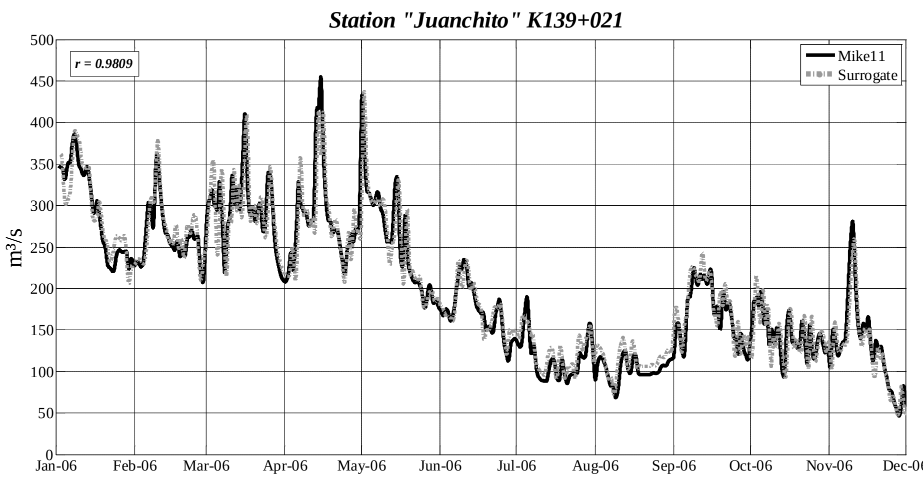

Figure 3. Comparison between original model and surrogate model predicted discharges for the year 2006 139 Kilometers downstream from the dam in the gauging point “Juanchito”.

All the locations included in the model are built based on M5P algorithms, These algorithms are not capable of dealing with discontinuities as well as Artificial Neural Networks for example. In different fields, implementation of these type algorithms (ANN) has proven to be stronger for the representation of highly non-linear systems. However the M5P models are easier to understand for decision makers as their clustering rules give additional information which the ANN don’t. For example the model tree used for representing the gauging station of “Juanchito”. In terms of water quality, the model capable of representing the seasonal trends fairly. Nonetheless, it has difficulties in representing the state of the system when high pollutant loads might enter the system. Based on the reduced amount of data for training water quality forecasting, it is also observed that only few extreme events are contained in the training set. Therefore the data driven models are not capable of representing events outside the feasible space.

CONCLUSIONS

A surrogate model for predicting discharges and water quality in points of interest was tested. The training of the model was done with one single run configured to report hourly time step output. The model was able to predict the discharges in the system while significantly reducing the number of variables. The model reduced the calculation time while representing the system fairly. Although cascade models propagate the errors and therefore is recommended to build models with the possible lowest round mean square error.

In terms of water quality prediction it is recommended to ensure several high pollution extreme events in order to train the models more accurately. It is also observed that dissolved oxygen (D.O) concentrations is highly dependent on the amount of organic matter that enters the system and so is recommended to include the Biochemical Oxygen Demand (B.O.D.) as an additional input parameter for water quality emulation. Nevertheless, the model is capable to reproduce the seasonal behavior which allows to use the actual surrogate model for optimization routines as it was initially intended. As a final remark, the authors want to stress that the present study was only intended to evaluate the capabilities of emulating a system with M5P algorithms and not to evaluate the accuracy of the original emulated model.

ACKNOWLEDGEMENTS

The Authors would like to thank the FORESEE project in the Netherlands for sponsoring this research, the “Corporación Autónoma Regional del Valle del Cauca” and the CINANRA institute in Colombia for the information and data sharing and finally the collaboration of Dr. Carlos Velez from UNESCO-IHE in the Netherlands .

REFERENCES

[1] Altunkaynak, A., Özger, M., and ÇakmakcI, M. (2005). "Fuzzy logic modeling of the dissolved oxygen fluctuations in Golden Horn." *Ecological Modelling*, 189(3-4), 436-446.

[2] Coorporación Autónoma Regional del Valle del Cauca – CVC., U. d. V. (2000). "Caracterización y Modelacion Matematica del Rio Cauca - Tramo Salvajina - La Victoria - PMC fase I." *Caracterización del río Cauca - Tramo Salvajina - La Victoria*, Universidad del Valle, Colombia, Cali.

[3] Corporación Autónoma Regional del Valle del Cauca – CVC., U. d. V. (2001). "Proyecto de Modelación del Río Cauca - PMC. Fase I.", M. M. d. R. C.-T. S.-L. Victoria, ed.

[4] Nerini, D., Durbec, J. P., and Manté, C. (2000). "Analysis of oxygen rate time series in a strongly polluted lagoon using a regression tree method." *Ecological Modelling*, 133(1-2), 95-105.

[5] Preis, A., and Ostfeld, A. (2008). "A coupled model tree-genetic algorithm scheme for flow and water quality predictions in watersheds." *Journal of Hydrology*, 349(3-4), 364-375.

[6] Quinlan, J. R. (1987). "Simplifying decision trees." *International Journal of Man-Machine Studies*, 27(3), 221-234.

[7] Razavi, S. T., Bryan A.; Burn, Donald H.; (2012). "Review of surrogate modeling in water resources." *Water Resources Research*, 48(7), W07401.

[8] Soltani, F., Kerachian, R., and Shirangi, E. (2010). "Developing operating rules for reservoirs considering the water quality issues: Application of ANFIS-based surrogate models." *Expert Systems with Applications*, 37(9), 6639-6645.

[9] Velez C.A., C. A. G., Ramirez C. A. , Baena L. M. (2006). "Cauca River Water Quality Model Hydroinformatics Application in a Developing Country." *7th International Conference on Hydroinformatics HIC 2006* Nice, FRANCE.