Computational method for prediction enhancement of a river flood simulation

Adriana Gaudiani[1,3], Emilio Luque[2], Pablo García[3], Mariano Re[3], Marcelo Naiouf[4], Armando De Giusti[4]

Abstract

The occurrence of flood events has become more frequent in many parts of the world over the past 30 years and, consequently, more people are exposed to flood damages. Modeling and computational simulation provide powerful tools which enable us to forecast, in order to reduce flood damages. This work is oriented to address input parameter uncertainty toward providing a methodology to tune the flood simulator and so achieve lower errors between simulated and observed results. Even a much reduced parameter set is considered to run the flood simulator, the search space is large. The results obtained by using a parametric simulation heuristic and a clustering technique are promising and a reduction of the search space was achieved, consequently we could reduce the computational cost of the search for the best scenario and the optimization scheme implemented enables us to get a better understanding of the problem.

1. Introduction

Flooding is one of the most common natural hazards faced by the human society. Future climate change and its impact on flood frequencies and damages, make this problem a serious environmental problem. Flood damage refers to all varieties of harm caused by flooding. The computational simulations are used extensively as models of real systems to evaluate output responses. In particular, computational models are used to reach a better understanding on inundation events and to estimate flood depth and inundation extent. For these reasons, simulation becomes a powerful tool for predicting flood events and minimizing their environmental effects.

Predictions of flood simulation extent have been made possible by advances in numerical modelling techniques and increases in computer power. Nevertheless, a series of limitations cause a lack of accuracy in forecasting, such as the case of uncertainty in the values of the input parameters to the flood model. Hydrodynamic modelling of a fluvial channel involves defining certain parameters as input variables which, for various reasons, may incorporate uncertainties in the results. Firstly, these parameters are measured or estimated in certain particular points but the value of such parameters must then be interpolated to the whole domain. For example, levees height can be measured in some sections but then it is necessary to estimate the heights for the other sections. Secondly, the parameters measurement is not direct, as it involves an estimation error associated with the estimation methodology [1]. The parameters uncertainty has an important impact on the simulation output, which is far from approaching the actual observed data [2].

To overcome this problem, in our previous work, we implemented a parametric simulation in order to find the best set of parameters, or adjusted set, which will be used as the input set for the underlying flood simulator emulating an "ideal" flood simulator as much as possible. The main objective of this work is to add an optimization process to the classical prediction approach to tune

[1] Instituto de Ciencias, Universidad Nacional de General Sarmiento, Buenos Aires, Argentina, agaudi@ungs.edu.ar
[2]Departamento de Arquitectura de Computadores y Sistemas Operativos, Universidad Autónoma de Barcelona, Barcelona, España
[3] Programa de Hidráulica Computacional, Laboratorio de Hidráulica, Instituto Nacional del Agua, Argentina
[4] Instituto de Investigación en Informática LIDI (III-LIDI), Universidad Nacional de La Plata, Buenos Aires, Argentina

input parameters, in order to minimize the difference between real and simulated result. The optimization method results in a large number of scenarios carrying out the search for the optimal, or suboptimal, set of input parameters. This process requires a huge amount of computation and it is on possible with resources in parallel programming and high performance computing.

Our work takes advantage of the results performed by the research group of High Performance Computing for Efficient Applications & Simulation at the University Autonoma of Barcelona, with close collaboration of the hydraulic engineering team at the National Institute of Water of Argentina. To conduct the research we selected the computational model EZEIZA V (Ezeiza), currently used as one of the tools of the Hydrologic Alert and Information System of the National Institute of Water (INA) at Buenos Aires, Argentina, in order to alert as early as possible on the occurrence of extreme water level events at the Paraná River basin, in South America [3]. The main limitations of this computational model are related to the reduced scale of the problem resolution (1D) and the inaccuracies in the river geometry representation. The model challenges are related to the uncertainty reduction in determining the flood peak arrival. This work goes in this direction by providing a better model calibration [4].

2. The flood wave simulator

Our work starts using a computer model of a real system such as flood events. The computational model is the conceptual model implemented on a computer, and the conceptual model is the mathematical representation of the physical problem to be modelled [5]. The selected software, Ezeiza, is a computational implementation of a one-dimensional hydrodynamic model for a flow net, based on the Saint Venant equations [6].

Ezeiza software family started to be developed in the '70s and its ongoing updating is performed by the INA staff. This computational model was chosen because of its simplicity when exporting results to output files, which can be processed by statistical and/or mathematical software, and for its convenience when running parametric simulations by changing the parameters values in the input files [4]. These features are very useful to take forward the tuning methodology.

An exhaustive study of the Paraná River model performance was carried out later by Ing. Latessa at INA, who stated the need to improve Ezeiza simulated results [7]. The utility of several efficiency criteria to evaluate hydrological performance model is addressed in [8]

3. Paraná River Model

La Plata basin is one of the most important rivers systems in the world. The Paraná River is one of the main rivers that form the basin. This is the second longest of South American rivers and it has a length of 4000 km alongside its major tributary, the Paraguay River (2550 km). The stretch of the Paraná River simulated by Ezeiza extends between the Yacyretá dam (Corrientes) to Villa Constitución (Santa Fe), both in Argentina. The Paraguay River runs from Puerto Pilcomayo (Formosa) to its confluence with the Paraná. Both river basins were divided into a number of sections, to measure rivers flow or height in each of them.

Large areas of land along the Middle and Lower Paraná margins are frequently subject to extended floods, which cause considerable damage. During the highest floods, monthly discharges at Middle Paraná exceed twice, and even three times, the mean discharge. A complete description of the highest floods at the Paraná basin and the possible climate forcing of such events are shown in [9].

The simulator Ezeiza prediction method is for height prediction in Parana River. In other words, Ezeiza is used to forecast daily water level variations at the Paraná River basin. The data required to define the modelled river system, as shown in Figure 1, is as follows:

- Initial conditions: levels and flow at every point of the river's domain.
- Boundary conditions: time series of rivers levels and flow at upstream and downstream points.
- Geometry data: data on the topography of the system.
- Input Parameters: Manning values and levees height, at every river sections.
- Observed data: water heights of Paraná River measured at each monitoring station.

This information was provided by INA, including the observed (actual) data of 1994 - 2011 period of time whose values are daily heights measured at 15 monitoring stations placed along the Paraná River basin.

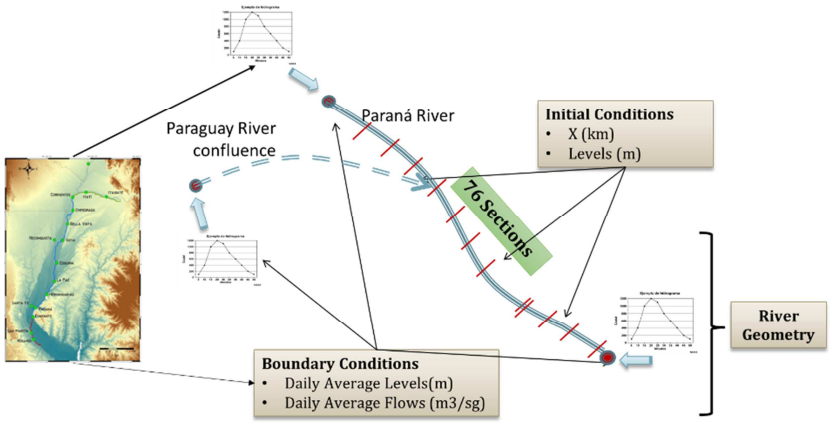

Figure 1 Flow net topology

Despite following 15 control stations, the levels are modeled throughout the length of the river. To define the flooding areas, the information of the modeled height should be crossed with ground levels. In general, every city has predefined levels of warning and evacuation.

When a simulation is run, Ezeiza returns a time series of heights values, which are calculated at each one of the 15 monitoring stations. These simulated data and the observed data, for the simulated period of time, are compared with each other to determine a *similarity index* (SI) that will be used to measure the simulation accuracy, to which we will return later.

The more sensitive input parameters of the flood routing models are the rugosity coefficient, or Manning values, and the levees height. Manning values for flood plains can be quite different from values for channels; therefore, manning values for flood plains are determined independently from Manning values for channel [10]. Finding an adjusted set of parameters is a key issue for our work, because it is the major step in order to develop a tuning methodology.

4. The Tuning Methodology

4.1. Parametric Simulation

The parametric simulation consists of changing the values of the internal input parameters and launching as many simulations as different combinations of parameters values are possible. In this kind of experiments it is possible to make deliberate changes in the parameters values. A scenario is defined by a particular setting of the set of parameters.

The number of possible scenarios is determined by the cardinality, C_i, for each of the N parameters considered. For each parameter i we define an associated interval and an increment value, which are used to move throughout the interval. For example, given the parameter i we define the associated domain and step values with the tuple: $< [Limit^i_{inf}, Limit^i_{sup}], Step_i >$

$$\#\ Scenarios = \prod_{i=1}^{N} C_i \quad (1)$$

$$C_i = ((Limit_{sup} - Limit_{inf}) + Step_i)/Step_i \quad (2)$$

We show in equation (2) the cardinality expression for parameter i, where #Scenarios, in equation (1), is the calculation of the total number of scenarios that we obtain after performing all the possible combinations of parameters values. As we perform an exhaustive parametric simulation in this phase, we define each new scenario by changing a single parameter, leaving the other fixed.

Paraná River basin, which represents the model domain, was divided into 76 sections in order to measure river flow and height in each of them. Each section is characterized by a Manning value for floodplain, another value for riverbed and a levee height. Here we itemize the parameters domain:

- Manning values for floodplain are within the [0.1, 0.2] range, with an ideal step of 0.01
- Manning values for riverbed are within the [0.015, 0.035] range, with an ideal step of 0.005.
- Levees height is within the 5m to 50m range with a step of 5m. (The step value is set according to the local geography)

4.2. The Optimization Problem

In order to tune the simulation, a specific objective of this work is to implement an optimization process to find the best set of parameters. We mean, our objective is finding the combination of input parameters that minimizes the deviation of the simulator prediction from the real scenario.

Optimization is generally defined as the process of finding the best or optimal solution for a given problem under some conditions. Formal optimization is associated with the specification of mathematical objective function (called f) and a collection of parameters that should be adjusted to optimize the objective function. Mathematically an optimization problem can be stated as:

$$\begin{aligned} &max/\min f(x) \\ &subject\ to\ x \in S \end{aligned} \quad (3)$$

Where x is the variable; f is a function $(f : S \rightarrow \mathbb{R})$; S is the constraint set, and $\exists\, x_0 \in S$ such that $f(x_0) \leq f(x)\ \forall x \in S$, for minimisation, and $f(x_0) \geq f(x)\ \ \forall x \in S$, for maximisation.

In this work, the optimization process, expressed by equation (3), can be defined as follows: We find the parameters vector $\overrightarrow{x^*} = [x_1^*, x_2^*, \ldots, x_N^*]$, N-dimensional, which optimize equation (3), where $\overrightarrow{x^*} \in S$ and the domain $S \subseteq \mathbb{R}^N$ represents the constrain set defining the allowable values for the $\overrightarrow{x^*}$ parameters. The search space of the problem is the S-dimension. In our problem, the search space consists of as many vectors as different combinations of parameter values are possible; so, we can say that S-dimension states the number of scenarios. Furthermore, we have to define a process to find a setting for the parameter vector $\vec{x}$, which provides de best value for the objective function $f(\vec{x})$. When there is no explicit form of the objective function and the parameter settings or design variables are discrete values, thus the optimization problems became discrete optimisation via simulation problems. We use the results obtained by [11] [12] [5].

[5] The Research Group in Individual Oriented Modelling (IoM) in the University Autonoma of Barcelona.

As simulation is computationally expensive, in particular when we need a greater search space, i.e., when we include more sections in the parameters vector, a new alternative was explored. The Monte Carlo (MC) based approach is one of the most popular, even though if the yield solution could be not the global optima, but rather an approximate good solution. MC is a statistical sampling method used to approximate solutions to quantitative problems.

4.3. Problem Delineation

An exhaustive search always guarantees finding a solution, if there exist a solution. In order to determine that solution, it may be necessary to test each possibility and verify if it satisfies the statement of the problem. The computational cost is high because it is proportional to the search space dimension. The search space for this model's parameters is determined tacking into account: a) the parameters corresponding to 76 sections along the river basin, b) considering that each section is divided into 3 to 5 subsections and c) the parameters domain, as we described in a previous section.

In an initial approach, the search of the optimum was done through an exhaustive search technique, even though it implies a lot of search time, it is guaranteed that the optimum is found. We combined only the Manning values, leaving aside levees heights. On this basis, if we had implemented an exhaustive search to find the adjusted parameters we would have launched 112^{76} simulations; this means that Ezeiza should have be executed 10^{154} times. With the aim to reduce the search space ($\mathbb{R}^N$), we had implemented a parametric simulation algorithm combining the possible Manning values in sections 70 – 72 – 74 and 76. The domain experimentation cardinality, which is shown in Table 1, was calculated using equation (2). We run the simulator 4096 times: $(4 \times 2)^4$. Even with this reduced setting of the parameter vector dimension, the search space is large. The observed data and the simulated data are time series of daily river heights at each monitoring station. The period simulated was 365 days (1999 year).

Manning	Interval	Cardinality	C value
Floodplain	<[0.1, 0.2], 0.1>	((0.2-0.1)+0.1/0.1	2
Riverbed	<[0.010, 0.04], 0.01>	((0.04-0.01)+0.01)/0.01	4

Table 1: Domain cardinality for Manning values

We implemented a solution to the "search problem", so we say to the "optimization problem", by using a parametric simulation technic applied to the reduced search space. We used the root mean square error (RMSE) as a metric to calculate the SI index, in order to evaluate the simulator response for each simulation scenario launched with Ezeiza and to find the minimum SI. We described in detail the steps involved in this methodology in [13]. The improvement percentage experiences, regarding simulated results come from INA's scenario currently used, ranges from 33% to 60% in the best three predicted stations. Running a full simulation under the conditions stablished lasted 2 minutes. When we run the 4096 scenarios, the execution time lasted 8192 minutes (137 hours). We used a master-worker approach to parallelize the method and reduce the computing time, this solution, however, is not sufficient when the dimension of the parameter vector grows. The computational cost grows exponentially in function of the number of section considered. In the future, it will be necessary to address this issue with a lower computational cost. Therefore, a better approach optimization technique, rather than an exhaustive search, must be used. Now, this approach, in a first phase, is using a computational process based on an iterative method of MC scheme, which is combined with a K-Means clustering method, in order to identify the regions where the optimum is. A second phase consists in a reduced exhaustive search [14].

The selected scenarios that have a better mean SI value than the previous ones are accumulated, in order to reuse past information. The MC program stops when two consecutive iterations cannot be able to improve the SI value, i.e., it becomes stationary or asymptotic, and the MC stores the last time when the average mean value was improved. We mean, it stops when the prediction error cannot be improved by the method. At each MC iteration, the parameters values of the input scenario were randomly selected to be fed into the simulator Ezeiza. To evaluate the improvement achieved by the method, we measured the SI for each scenario, which provides an adjustment rate between simulated results and observed data at each monitory station, taking into consideration the complete time series.

The final SI value is the mean of the RMSE calculated in the 15 output stations. The best SI value is compared with the SI reached by running Ezeiza with the INA scenario. The INA simulated results are the reference point to get the improvement rate achieved. This rate resulting of MC method can be expressed as follows:

$$Improvement_{Station} = \frac{abs(SI_{INA,} - SI_{\widehat{Scenario}})}{SI_{INA}} \quad (4)$$

where $\widehat{Scenario}$ represent the best scenario, the $SI_{\widehat{Scenario}}$ is the minimum prediction error and SI_{INA} is the INA prediction error.

5. Experimentation

The simulator is used as a black box, even though the more realistic the simulator is. We used the reduced search space of 4096 scenarios, which is the same domain configuration as the one used in our previous work. The same period of time was used to carry on the simulation and 4 section were selected, which are located at the lower Paraná. We remained the same conditions to evaluate the utility and reliability of this optimization method and compared both final results.

The objective of the index SI is to provide a metric to select the best scenarios. Each scenario configuration is represented by the objective function, and this function depends on the vectors $\overrightarrow{\mathrm{Sim}}$ and $\overrightarrow{Obs}$, whose components are the simulated and the observed data respectively, for each output station and for each simulation day. The restrictions are the possible ranges of values that the parameters can take. This optimization problem is expressed mathematically in equation (5)

$$\begin{aligned} &Minimize\ prediction\ error\ (SI) && f(\overrightarrow{\mathrm{Sim}},\ \overrightarrow{Obs}) \\ &subject\ to && dias\ \in [01-01-1999\,..\,31-12-1999] \\ & && 0.1\ \leq MannPlain \leq 0.2 \\ & && 0.01 \leq MannBed \leq 0.035 \\ & && Section\ \in \{\,selected\ sections\} \subseteq \{76\ sections\} \end{aligned} \quad (5)$$

Table 2 shows the scenarios resulting of minimum average of the index SI. These are the scenarios that allow us to reach better simulated results than the INA scenario results, where M-F is the Manning value for floodplain and M_R for riverbed. We mean that the improvement rate, as we show in equation (4), are the best achieved.

We are measuring the index mean for the 15 stations, so we cannot reach rates upper 15% yet for all the stations at the same time. In the other hand, some individual stations were improved in 30-40% and sometimes two stations resulted enhanced (Rosario 28% and San Martín 25%, but these enhancement were achieved in different scenarios). This situation needs to be improved in the future.

	Section 70		Section 72		Section 74		Section 76		Improv.
	M-F	M-R	M-F	M-R	M-F	M-R	M-F	M-R	rate
Sce-1	0.02	0.1	0.02	0.1	0.03	0.2	0.03	0.1	14%
Sce-2	0.02	0.2	0.02	0.1	0.03	0.2	0.03	0.1	13%
Sce-3	0.03	0.1	0.02	0.1	0.03	0.1	0.03	0.1	12%
Sce-4	0.02	0.2	0.02	0.2	0.02	0.1	0.03	0.1	11%
Sce-5	0.02	0.1	0.02	0.2	0.03	0.1	0.03	0.1	14%

Table 2: The best scenarios selected by the first phase of the optimization scheme

Section	70	72	74	76
Mann.RiverBed	0.02, 0.035, 0.03	0.02	0.02, 0.025, 0.03	0.03
Mann.FloodPlain	0.1, 0.2	0.1, 0.2	0.1, 0.2	0.1

Table 3: Restrictions to the parameters values for the second phase

This phase was successful. The search returned 4 scenarios with an improvement rate between 30% and 40% in 3 stations. We selected the best:

Section 70: (0.03, 0.2), Section 72: (0.02, 0.1), Section 74: (0.035, 0.1), Section 76: (0.030, 0.1)

with an improvement rate of : Station Rosario: 31% , Station San Martín: 35% and Station Diamante 38%. We point out that these ratios were achieved with the same scenario.

6. Results and Conclusions

In our previous work, the parametric simulation allowed us to test the goodness of the method and find 5 scenarios whose prediction error were less than the RMSE reached with INA's scenario and the improvement was greater than 30% for each station. We run all the possible scenarios and we concluded that we could not get the same good results for more stations, at the same time and with the same scenario. Now, in this work, we use a two phase scheme. Firstly, we get the best (adjusted) set of parameters using a MC asymptotic scheme. MC + K-means arrived to the end in 4 steps, is said in 800 simulator running. If the new SI is not lower than the last one stored then there is no need to run MC again. We stopped the process when the new SI is not better than previous one. Secondly, we run a reduced exhaustive search for the reduced search space resulting from the previous phase, is said 76 simulator running were added. In this step we repeat the search used in the previous approach. We got an improvement of 30% to 40% and it is worth pointing out that we reduced the time of all the process. First, we needed to launch 4096 simulations and now we needed to launch 876 simulations. We have to enhance the prediction and adjust the heuristic technique but the results are promising and a better understanding of the problem was achieved. As future work, the MC method + K-Means clustering technique must be tested for all the sections and a huge amount of parameters values should be computed. Just this situation requires high performance computing. This will be a key resource to tune the simulator Ezeiza for a more accurate forecasting.

References

[1] A. Bárdossy and S. Singh, "Robust estimation of hydrological model parameters," *Hydrology and Earth System Sciences Discussions,* vol. 5, no. 3, pp. 1641-1675, 2008.

[2] S. Balica, I. Popescu, L. Beevers and N. Wright, "Parametric and physically based modelling techniques for flood risk and vulnerability assessment: A comparison," *Environmental Modelling & Software,* vol. 41, pp. 84-92, 2013.

[3] A. Menéndez, "Three decades of development and application of numerical simulation tools at INA Hydraulics LAB," *Mecánica Computacional,* vol. 21, pp. 2247-2266, Octubre 2002.

[4] A. Menéndez, "Ezeiza V: un programa computacional para redes de canales," *Mecánica Computacional,* vol. 16, pp. 63-71, Septiembre 1996.

[5] R. Sargent, "Verification and validation of simulation models," in *Proceedings of the 37th conference on Winter simulation*, Orlando, Florida, 2005.

[6] N. Hunter, P. Bates, M. Horritt and M. Wilson, "Simple spatially-distributed models for predicting flood inundation: A review," *Geomorphology,* vol. 90, no. 4, pp. 208-225, 2007.

[7] G. Latessa, "Modelo hidrodinámico del río Paraná para pronóstico hidrológico: Evaluación del performance y una propuesta de redefinición geométrica.," INA - UBA, Buenos Aires , 2011.

[8] P. Kraus, D. Boyle and F. Bäse, "Comparison of different efficiency criteria for hydrological model assessment," *Advances in Geosciences,* vol. 5, pp. 89-97, 2005.

[9] I. A. Camilloni and V. R. Barros, "Extreme discharge events in the Paraná River and their climate forcing.," *Journal of Hydrology,* vol. 278, pp. 94-106, 2003.

[10] F. Saleh, A. Ducharne, N. Flipo, L. Oudin and E. Ledoux, "Impact of river bed morphology on discharge and water levels simulated by a 1D Saint-Venant hydraulic model at regional scale," *Journal of Hydrology,* vol. 476, pp. 169-177, 2013.

[11] M. Taboada, E. Cabrera, M. L. Iglesias, F. Epelde and E. Luque, "An agent-based decision support system for hospitals emergency," *Procedia Computer Science,* vol. 4, pp. 1870-1879, 2011.

[12] E. Cabrera, M. Taboada, M. L. Iglesias and E. Luque, "Simulation optimization for healthcare emergency departments," *Procedia Computer Science,* vol. 9, pp. 1464-1473, 2012.

[13] A. Gaudiani, E. Luque, P. García, M. Re, M. Naiouf and A. De Giusti, "Computing, a powerful tool for improving the parameters simulation quality in flood prediction," *Procedia Computer Science,* vol. 29, pp. 299-309, 2014.

[14] E. Cabrera, M. Taboada, F. Epelde, M. L. Iglesias and E. Luque, "Optimization of emergency departments by agent-based modeling and simulation," in *Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference*, Las Vegas, 2012.