Jun 18th, 10:40 AM - 12:00 PM

Using surrogate modelling for fast estimation of water budget component in a regional watershed

Aurelien HAZART
Geosphere Environmental Technology, auhazart@getc.co.jp

Koji MORI
Geosphere Environmental Technology, mori@getc.co.jp

Kazuhiro TADA
Geosphere Environmental Technology, tada@getc.co.jp

Hiroyuki TOSAKA
The University of Tokyo, tosaka@sys.t.u-tokyo.ac.jp

International Environmental Modelling and Software Society (iEMSs)
7th Intl. Congress on Env. Modelling and Software, San Diego, CA, USA,
Daniel P. Ames, Nigel W.T. Quinn and Andrea E. Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2014-proceedings

Using surrogate modelling for fast estimation of water budget component in a regional watershed

Aurelien HAZART[1], Koji MORI[1], Kazuhiro TADA[1] and Hiroyuki TOSAKA[2]

[1]Geosphere Environmental Technology, 2-1 Kanda-Awajicho, Chiyoda-ku, Tokyo 101-0063, JAPAN
auhazart@getc.co.jp, mori@getc.co.jp, tada@getc.co.jp

[2]Department of Systems Innovation, Graduate School of Eng., The University of Tokyo. 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 JAPAN
tosaka@sys.t.u-tokyo.ac.jp

Abstract: We developed a method based on surrogate modelling to estimate the water flow in both surface and subsurface sections of a basin watershed. The real case of a regional watershed near Tokyo with a catchment area of 100km^2 was targeted with observations on a period of about 16 years. Our purpose was to propose a tool of water resources management that can be operated daily by the person in charge of a watershed without expert knowledge of the numerical modelling process. Replacement of complex physical models by surrogate models like artificial neural networks (ANN) has been proved useful in several applications in hydrology. We think that it is a primary choice when the computation time and the complexity of the physical model become prohibitive for a usage by non-specialists on a standard computer. Here, five surrogate models were tested with a focus on ANN. A feed forward neural network and a radial basis neural network were trained with the Levenberg-Marquardt algorithm to emulate a general purpose fluid flow simulator with fully-coupled surface/subsurface capability used to simulate water flow volume that cannot be observed directly. The comparison with others surrogate modelling techniques has shown the relative superiority of ANN in this application. One configuration of input datasets was made by selecting meteorological (e.g. temperature, rainfall) and hydrological (e.g. streamflow, groundwater level) time series. A data selection study showed the effectiveness of considering river flow rate and rainfall over the groundwater level as input data. Regarding the output variables, six water budget components were selected to fully characterize the water balance of the watershed. The performances were evaluated by mean of the coefficient of determination (R2). Estimation of the unobserved water balance component provides indicators of the watershed condition and guides the sustainable management of the watershed.

Keywords*:* surrogate model; water budget; estimation; regional watershed; artificial neural network

1 INTRODUCTION

In parallel to the development of complex physical models for simulating hydrological process, “black box” models that see only the physical quantities as dimensionless input and output variables have been continuously used. Each modelling approach (physically based and “black box”) aims at estimating or predicting unobserved quantities but have different usages and limitations. “Black box” model have been proved useful when there is no need to discover the underlying physical process, when collecting and calibrating the numerical model parameters becomes prohibitive, or when the model must be operated quickly by a non-expert person on a standard computer. However, this comes generally with a loss of accuracy when new input data are given to the model. A careful parameter setting is needed to avoid overfitting of existing data and to ensure an efficient generalization capability. The two main usages of a “black box” model are the prediction of quantities ahead of time based on past measurements (i.e. time series prediction) and the emulation of physical models. In the latter case, the model is often called surrogate model or meta-model as it aims to reproduce the results of the physical modelling (a model of a model). Concept and application of surrogate model in water resources has been extensively reviewed in Razavi et al. (2012). The scope of applications covers automatic model calibration, optimization, uncertainty and sensitivity analysis. The methods include polynomial regression, kriging, radial basis function, artificial neural network

(ANN). Recently, ANN has been extensively used in hydrological modelling, as replacement of costly traditional model or for prediction and forecasting. When not coupled with a physical model, the most frequent applications are forecasting of rainfall, streamflow and groundwater level.
Although meta-models in general and ANN in particular are widely used in hydrology, their application in water resource management tool for decision-support of stakeholders is rather limited. In such application, the challenge is to select and estimate global indicators of the watershed condition from local monitoring data. In Holzkämper et al. (2012), the authors developed a Bayesian network (BN) to integrate data from various sources at the basin scale. In Rogers et al. (1994), an ANN is training to mimic a solute transport code, and then used to predict the optimal pumping scenario for groundwater remediation. Here, the goal is to obtain an estimation of the water balance without having to prepare, calibrate and execute a costly hydrological model. Trained to reproduce six water balance component from 14 input time series, the capability of the ANN surrogate model to generalize to new data is used to obtain an estimation of the water balance of a given watershed in a fraction of the time required by the numerical model.

2 STUDY AREA

The basin of Hadano city near Tokyo is well-known for the quality and the relative abundance of its water stored in several gravel aquifers in the subsurface. The resources in water are used for the main part by the local government to supply the local population in drinking water while the second main usage is to supply the industrial demand. Figure 1 shows the location of the study area.
The land is mostly composed of urban area in the valley and forest in the mountainous part. Other land use includes farming field, paddy and golf course. Thanks to about 100 boreholes locations, the geology of the site is well-known. The site is composed to several layers of loams and five principal aquifers. Regarding the topography, digital elevation data are available up to 5 meters precision, with a 10 meters precision in the mountainous area. Several monitoring points provide measurement of groundwater level, river streamflow and underground temperature profiles. A few rainfall and temperature stations managed by the Japan meteorological agency give access to the meteorological data.
In average, it was observed that spring and groundwater level slightly increased in the past 30 years while rainfall remained constant so the regional watershed does not suffer from water scarcity. Still, an accurate and easy-to-use water resource management is important to control the heavy use of the resource in the region and the possible higher water demand owing to change in the pumping regulation. An increase of the temperature of about 1.5^{o}C in the last 40 years has been observed and, because of the climate change, a similar tendency is expected in the future. The proximity to Tokyo-Yokohama (about 50km) also causes constant changes of the landscape due to civil engineering construction (tunnel excavation, highway construction, etc.) that could affect the equilibrium of the water cycle.

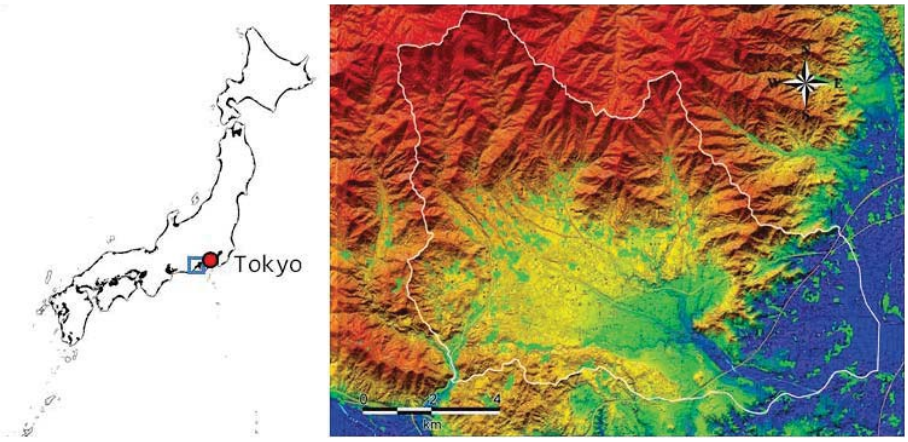

Figure 1: Topographical map of Hadano city, Japan. Location of the study area is marked with the blue square.

3 METHOD AND MODELLING

3.1 Method

The method developed in this study aim to be a new support-decision method for regional watershed stakeholders without advanced numerical modelling knowledge. In the first step, a numerical model is built to simulate the physical process occurring in the watershed. As we want to obtain global indicator of the water balance that cannot be directly measured (global recharge and discharge of the whole watershed), this computationally expensive step cannon be avoid.
In practice, the hydrological model GETFLOWS (Tosaka et al. 2000, 2010) is applied to simulate the water cycle of the watershed. After GETFLOWS has been calibrated to match some monitoring data, the following steps consists in selecting the computed time-series that will be used as input for the surrogate model and extracting the relevant water budget component from the numerical results in the whole watershed that will be used as outputs. In this study, the input data are made of 14 time series data, that is monthly recording of rainfall (one), groundwater level (two) and river streamflow (eleven). Among these time series, one consists of meteorological observation (rainfall) and the other consist in results computed by the hydrological model. A representation of the input data is shown in Figure 2. Data from 1994 to 2005 are used for training the surrogate models while the period [2006,2011] is used for testing the surrogate models on unseen data.
The six components shown in Figure 3 are selected as quantities related to the global water balance of the watershed: Surface recharge and discharge (except river), surface recharge and discharge from the river, subsurface boundary inflow and outflow. Then, a surrogate model is trained to reproduce each water balance component, leading to six models (with same structure but different weights).

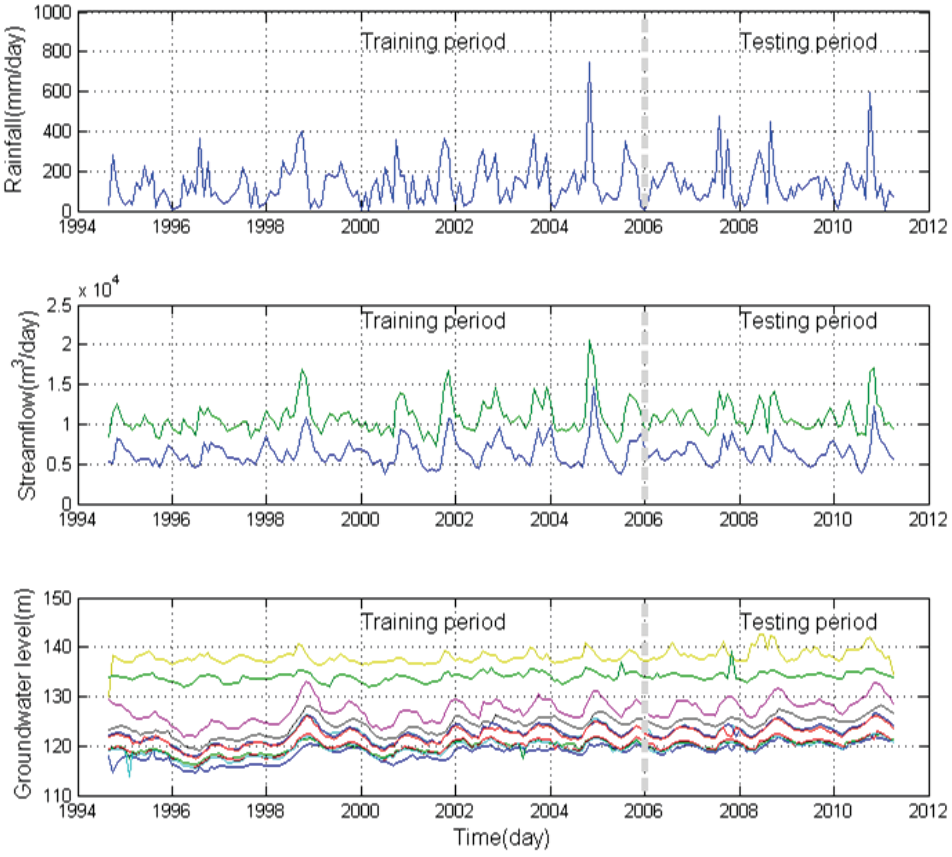

Figure 2: Monthly time series data used as input of the surrogate models. It consists of meteorological observation (rainfall) and results computed with the high fidelity hydrological model GETFLOWS (streamflow and groundwater level). Data in the period [1994,2005] are used for the training of the surrogate models and the testing is done with data from the period [2006,2011].

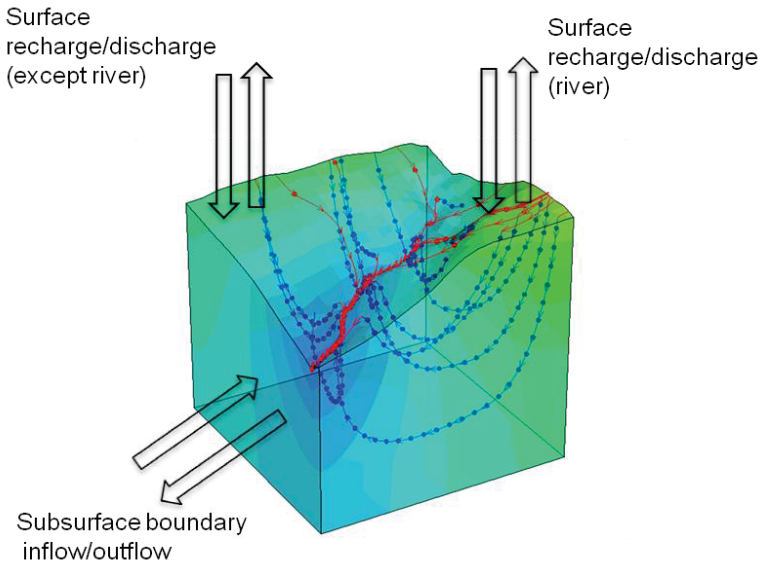

Figure 3: The six components of the water balance used as indicator of the watershed condition.

3.2 Numerical modelling

The purpose of the numerical modelling step is to compute the water balance component that will be used to train the surrogate model. Because we want to obtain global water balance indicator, it is necessary to use an integrated hydrological model capable of simulating the main physical process occurring in the surface but also in the underground. We use the integrated two-phase flow model GETFLOWS with non-isothermal condition proposed in Tosaka et al. (2000, 2010) (free version for non-commercial use available). It supports multiphase multi-component flow, gas dissolution to water, solute transport, heat transport and fully coupled surface–subsurface fluid flow. The governing equations are based on the generalized Darcy law, the equation of continuity and the shallow water equation for the surface flow. The resulting system of equations is solved simultaneously with a fully implicit integrated finite difference scheme. With a good knowledge of the watershed geology, topography and land use as well as precise meteorological data, such model can produce an accurate estimation of the water saturation, the air pressure and the temperature in the whole watershed, including the surface and the subsurface.
In the modelling of the Hadano basin, a catchment area of about 100km^2 was defined and discretized on a non-uniform grids with elements of about 50 meters length. Calibration of the numerical model where conducted by matching observations and computed results of river streamflow and groundwater level at three areas where the accuracy was considered the most important (about 10 stations in each area were used). After calibration, the numerical model is used to compute the global recharge and discharge components of the catchment area. Comparison of GETFLOWS results after calibration and observations at two stations is shown in Figure 4 (streamflow and groundwater level). Only the results of the hydrological model are used as input/output for the surrogate model, the original streamflow and groundwater level observations are discarded.

3.3 Surrogate modelling

Following some recent works of surrogate modelling applications in hydrology, artificial neural network models were selected as they usually give the best results. For the sake of comparison, three other surrogate modelling methods were tested, (kriging, polynomial surface response and radial basis function). In the following, a brief description of the methods is presented (see Razavi et al. (2012) and following references for more detailed descriptions).

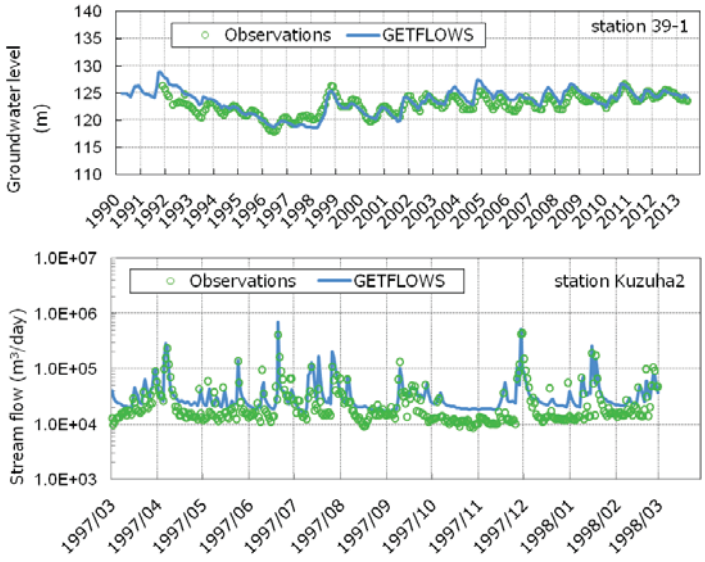

Figure 4: Comparison between computed results with the hydrological model (GETFLOWS) and the observations at two stations.

An artificial neural network (ANN) consists in the association of many simple operators called neurons and linked into layers. The name of the technique evokes the functioning of the human brain but the parallel with the biology is somehow limited. Each neuron models a non-linear bounded parametric function that depends on input variables and weights. Once connected, the neurons realize a universal functional approximation, that is, they can approximate any bounded function with an arbitrary precision and a finite number or neurons in a single hidden layer. ANN are highly flexible, suitable for many input and capable of modelling non-linear interaction. The determination of the optimal structure for a particular problem can be prone to over-fitting if not correctly designed. Among the possible network structures, radial basis neural network (RBNN) and feed forward neural network (FFNN) were selected. FFNN is one of the simplest types of artificial neural network, as it contains no loop and the data only propagate forward in the network. RBNN is a particular case of neural network with radial basis neurons. It can enforce more constraints than FFNN and can sometimes allow a better generalization.

The design of an ANN requires to set up the general structure of the network, the number of hidden layers, the number of neuron in each layer, the non-linear activation function, the size of the input and output layers and to select a suitable training algorithm to estimate the optimal weights. Following the recent literature on ANN, a feed-forward ANN (FFNN) and a radial basis neural network (RBNN) with one hidden layers made of 10 neurons was built and Levenberg-Marquardt algorithm was selected as it is usually the one that gives the best generalization results. After testing several activation functions (linear, log sigmoid), the hyperbolic tangent function was selected as in Diakopoulos et al. (2005).

A polynomial surface response (PRS) formalizes the relation between inputs and output by a polynomial function. The parameters to adjust are then the coefficients in a polynomial regression, usually estimated with the standard least square regression method. Simple to build and useful for screening the important factors, the method is rigid and usually not suitable for complex model with many interaction between variables or for many input variables. The kriging method is an interpolation method developed in the geostatistic field. It consists in a combination of a known functions (e.g., a linear model such as a polynomial trend). As the method is based on a probabilistic framework, it produces naturally uncertainty interval of the output data. Finding the appropriate correlation parameters is sometime difficult and the correlation matrix can be singular making the optimization stage non-trivial. Finally, radial basis function (RBF) models consist of a weighted sum of radial basis functions (also called correlation functions) and a polynomial function. Owing to the high constraints, is usually efficient even with few input variables. However, each variable is assumed equally important.

Within the tested surrogate models, some like kriging, or ANN can emulate multiple input, multiple output system (MIMO) and some can handle only one output (multiple input single output, MISO). Here, to help the results interpretation between surrogate models and to assess the difference among

the six water budget component, one surrogate model is trained for each output, leading to six surrogate models.

4 RESULTS

To evaluate the matching performance of the surrogate model during the training and the testing period, we compute the coefficient of determination R^2, given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} {y_i}^2 - \frac{1}{N}\sum_{i=1}^{N} {\hat{y}_i}^2}$$

where y_i is the data computed by the hydrological model, $\hat{y}_i$ is the data computed by the surrogate model and N is the number of points in the time series. To ease the presentation of the results, the output data are designed with the following notations:

(1) WB1: River recharge,
(2) WB2: Surface recharge except river,
(3) WB3: Underground recharge,
(4) WB4: River discharge: WB4,
(5) WB5: Surface discharge except river,
(6) WB6: Underground discharge.

The results of the five surrogate models are summarized in Table 1 where the coefficient of determination for the training period and the testing period are shown. The methods based on ANN are usually the one that perform the best in this application. Obviously the kriging and RBF methods suffer from overfitting, with a perfect reproduction of the training data and a relatively poor reproduction of the testing data, particularly for the components WB1 and WB6. Globally, three components (WB1, WB4, WB6) show less accurate reproduction than the three others (WB2, WB3, WB5). Although these results could be the sign of relevant input data for predicting these components, it can point out a difficulty of the surrogate models to emulate the underlying process. Comparison of expected output results from the hydrological model (GETFLOWS) and the output of the radial basis neural network surrogate model (RBNN) is shown in Figure 5. A good matching is observed even in the testing period, which shows that the trained RBNN generalizes well on unseen data.
To analyse the optimality of the results, we conducted an input selection analysis and applied a pre-processing technique. In the former, one type of input data is removed one by one from the input data set (rainfall, streamflow and groundwater model). The results with RBNN model, summarized in Table 2, shows that the performances improves in only one case (rainfall + streamflow, component WB5) and that the two streamflow time series have a significant impact and should be considered. In the latter, the principal component analysis (PCA), a projection method that allows reducing the number of inputs without losing much information, was applied with or without rescaling.
The performances after reducing from the original 14 inputs to 2 inputs, shown in Table 3, are slightly decreasing or constant. Moreover, the previously inaccurate estimation of WB6 is improved. Such result indicates that pre-processing techniques can help to extract relevant information from various observed data.

Table 1: Coefficient of determination for the five surrogate models (training period / testing period).

	Kriging	RBF	PRS	FFNN	RBNN
WB1	1.0 / 0.35	1.0 / 0.49	0.83 / 0.34	0.91 / 0.77	0.90 / 0.77
WB2	1.0 / 0.84	1.0 / 0.85	0.91 / 0.88	0.95 / 0.93	0.95 / 0.95
WB3	1.0 / 0.89	1.0 / 0.89	0.97 / 0.84	0.96 / 0.91	0.97 / 0.92
WB4	1.0 / 0.70	1.0 / 0.68	0.94 / 0.73	0.93 / 0.85	0.93 / 0.82
WB5	1.0 / 0.75	1.0 / 0.83	0.96 / 0.81	0.95 / 0.91	0.94 / 0.90
WB6	1.0 / 0.45	1.0 / 0.56	0.88 / 0.34	0.95 / 0.77	0.90 / 0.80

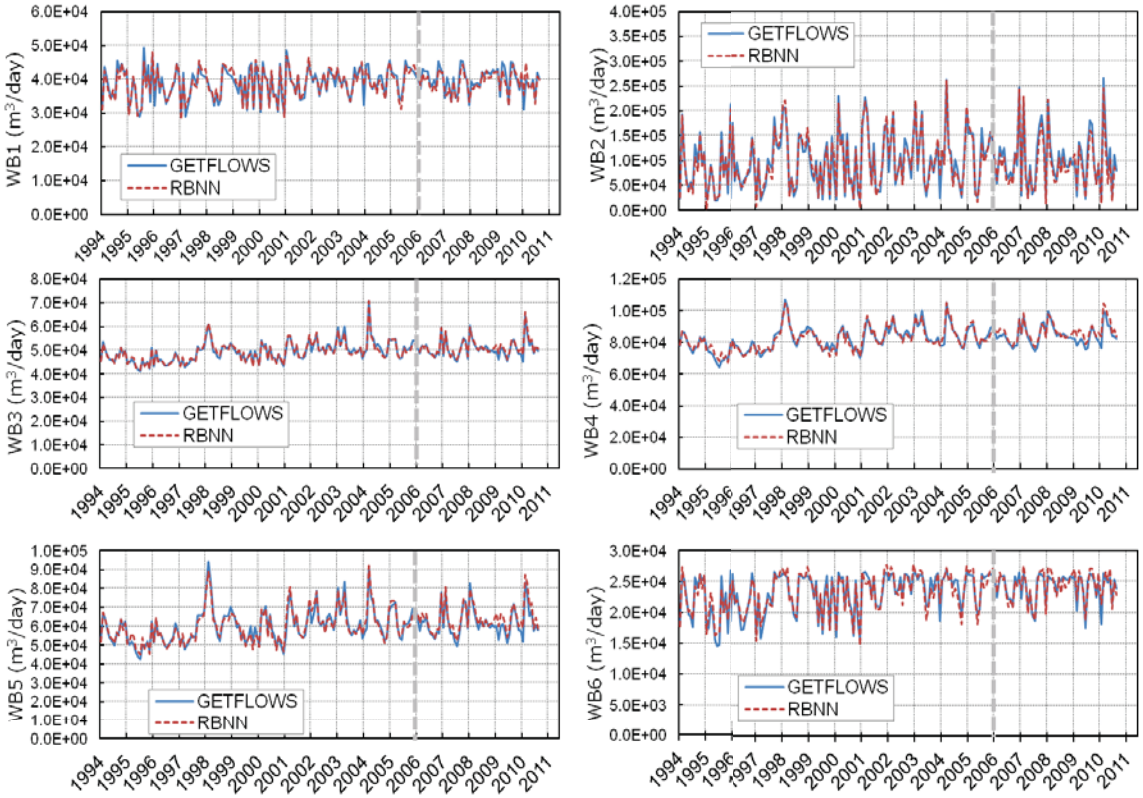

Figure 5: Comparison between expected output (GETFLOWS) and obtained output (RBNN) for the six water budget components. The training period is [1994,2005] and the testing period is [2006,2011].

Table 2: Input sensitivity analysis. The arrows indicate an increase, stagnation or decrease of the performance of RBNN model compared with the results in Table 1 (testing period).

	Streamflow + groundwater level	Rainfall + groundwater level	Rainfall + streamflow
WB1	↘	↘	→
WB2	→	↘	→
WB3	→	↘	↘
WB4	→	↘↘	→
WB5	↘	↘↘	↗
WB6	↘	↘↘	↘

Table 3: Results after pre-processing of the inputs by principal component analysis and scaling. The arrows indicate an increase, stagnation or decrease of the performance of RBNN model compared with the results in Table 1 (testing period).

	PCA pre-processing (2 inputs)	PCA+scaling pre-processing (2 inputs)
WB1	→	↘
WB2	→	→
WB3	↘	↘
WB4	↘	↗
WB5	↘	↘
WB6	↗	↗

5 CONCLUSION

We proposed a method for fast water balance component estimation using a surrogate model trained to emulate a complex numerical model with surface-subsurface full coupling. The method requires calibrating and executing the numerical model in order to obtain an accurate knowledge of hydrologic cycle in the watershed and to compute the water budget component. Then, a surrogate modelling method is applied to connect input data made of time series of meteorological and hydrological observations with the output time-series data made of the water balance component. The method is applied on a regional watershed near Tokyo, where, due to climate change, civil engineering construction (e.g. tunnel excavation, highway construction) and regulation changes that allow groundwater pumping from authorized companies, the management of the basin requires a fast tool to evaluate the usage of the groundwater and the water balance of the entire watershed. Among the five surrogate modelling techniques tested, the feed forward neural network and the radial basis neural network trained with the Levenberg-Marquardt algorithm achieved the best performance. Significant differences of performances were obtained depending on the component. An input selection analysis showed the relative optimality of the input data set considered and estimation of the underground discharge could be improved by applying a pre-processing technique. Limitation of using surrogate models in estimating the water balance component is the difficulty to find pattern of success and failure that would help the reproduction of the results in other watershed application. A method to obtain uncertainty on the estimated component would certainly improve this fact.

REFERENCES

Daliakopoulos, I. N., Coulibaly, P., & Tsanis, I. K., 2005, Groundwater level forecasting using artificial neural networks. Journal of Hydrology, 309(1), 229-240.

Holzkämper, A., Kumar, V., Surridge, B.W.J., Paetzold, A., Lerner, D.N., 2012, Bringing diverse knowledge sources together – A meta-model for supporting integrated catchment management, Journal of Environmental Management, Volume 96, Issue 1, 15 April 2012, 116-127

Hung, N.Q., Babel, M.S., Weesakul, S., & Tripathi, N. K. (2009). An artificial neural network model for rainfall forecasting in Bangkok, Thailand. Hydrology & Earth System Sciences, 13(8).

Maier, H.R., & Dandy, G.C. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environmental modelling & software, 15(1), 101-124.

Razavi, S., Tolson, Bryan A., Burn, D.H., 2012, Review of surrogate modelling in water resources., Water Resour. Res. 48, no. 7.

Rogers, L.L., Dowla, F.U., 1994, Optimization of groundwater remediation using artificial neural networks with parallel solute transport modelling, Water Resources Research, Water Resour. Res., 30, 2, 1944-7973, 457-481

Tosaka, H., Itho,K., Furuno, T., 2000., Fully coupled formation of surface flow with 2-phase subsurface flow for hydrological simulation. Hydrological Process, 14, 449– 464.

Tosaka, H., Mori, K., Tada, K., Tawara,Y. Yamashita, K., 2010., A general-purpose terrestrial fluids/heat flow simulator for watershed system management. In: IAHR International Groundwater Symposium.