International Environmental Modelling and Software Society (iEMSs)
2012 International Congress on Environmental Modelling and Software
Managing Resources of a Limited Planet, Sixth Biennial Meeting, Leipzig, Germany
R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.)
http://www.iemss.org/society/index.php/iemss-2012-proceedings

A Nearest-Neighbour Surrogate Model for the Simulation of Rainwater Tanks

J. Dugge [a], **P. Pedruco**[b] **and M.J. Hardy**[a]

[a] *eWater CRC, Innovation Centre, Building 22, University of Canberra, ACT 2617, Australia (jdugge@ethz.ch)*

[b] *BMT WBM, Level 5, 99 King Street, Melbourne, Victoria 3000, Australia (Philip.Pedruco@bmtwbm.com.au, Matthew.Hardy@bmtwbm.com.au)*

Abstract: Mathematical modelling of decentralised water supply options requires a finer temporal resolution than models of centralised regional systems. To avoid having to run a coarse scale model with the timestep needed for the fine scale model, the latter can be upscaled to a larger timestep. One way of doing this is to use a surrogate model.

This paper presents a nearest-neighbour surrogate model for upscaling a detailed cluster scale urban water model from a sub-daily to a monthly time scale. The approach consists of running a detailed cluster model with a fine temporal resolution using a historical rainfall timeseries, adjusting the daily rainfall values by clipping peaks and assigning rainfall occurring near the end of a month to the following month, and then aggregating the output and corrected input timeseries to monthly timesteps. This set of monthly values is then used as a surrogate model in subsequent simulations, using nearest-neighbour sampling to select an appropriate output value for a given monthly rainfall.

It was found that the surrogate model performs well in emulating the detailed model at a monthly timestep, producing a good model fit and succeeding in reproducing autocorrelation while running faster than the detailed model by several orders of magnitude.

Keywords: Rainwater Tanks; Upscaling; Surrogate Modelling

1 INTRODUCTION

The last three decades have seen many important changes to the manner in which the Australian water sector approaches the planning, provision and management of urban water cycle services, with new approaches taken within the stormwater sector [Lloyd et al., 2002; Mouritz, 1996; Wong and Eadie, 2000] . In more recent years, drought and dwindling storage levels have also become major drivers of water resource efficiency and the development of new and innovative ways to meet our urban centres needs. A key element of this change has been the emergence (and in some cases the re-emergence) of decentralised management strategies such as rainwater tanks, stormwater harvesting ponds and greywater systems for the provision of water cycle services such as water supply, stormwater management and waste water disposal. While the localised benefits of these systems are well understood, this is less true for the impacts of their widespread adoption at the regional scale.

Regional scale models typically operate on a monthly or annual scale, with long simulation periods or replicate simulation runs required for regional water resource planning.

Models of on-site residential rainwater tanks, however, require temporal scales of no more than daily timesteps to capture the system dynamics of the fill, spill and draw down cycles that occur [Mitchell et al., 2008]. Even if only average monthly output values are of interest, as would be the case when investigating the impact of distributed rainwater tanks on a regional system, rainwater tank models can not simply be run with average monthly input values. Instead, the output of a model run with a fine timestep has to be aggregated to monthly values, leading to long model run times. This temporal scale modelling requirement presents an important challenge for the incorporation of the water supply savings from rainwater tanks into regional scale headworks models.

To avoid this problem, the rainwater tank model itself, rather than its output variables, can be upscaled. One way of doing this is to generate a *meta-model* or *surrogate model* [Bierkens et al., 2000, p. 105], that captures the emergent properties of the detailed model but takes less time to run.

One surrogate model for upscaling a rainwater tank model from a daily to a monthly timestep is presented in Kuczera [2008]: The approach consists of running a model of a house with a rainwater tank at a daily timestep using a long simulation period, then aggregating the model input and output to monthly values. These aggregated sets of inputs and outputs are then used as a lookup table in monthly simulations: Instead of re-running the detailed model, the monthly input value is used to select the corresponding monthly output value. This constitutes a nearest-neighbour surrogate model [Altman, 1992; Fonseca et al., 2010]. In order to preserve some of the variance in the output values, the surrogate model described in Kuczera [2008] randomly selects one of the k nearest neighbours in the lookup table.

In this paper, we present extensions of the nearest-neighbour surrogate model described in Kuczera [2008] that improve its performance.

2 METHODS

For this study, a detailed model of a house with a rainwater tank and a surrogate model emulating the detailed model, both described in Kuczera [2008], were recreated. The performance of the model was analysed, and two additional input variables – a daily rainfall clipping value and a carryover period – were investigated. An extended surrogate model using the additional input variables was proposed and its performance in emulating the detailed single-house model was compared to that of the basic surrogate model. The extended surrogate model was then tested for its ability to emulate detailed models of five residential clusters in Canberra, Australia with monthly varying demand.

2.1 Detailed Model

Initially a model of a single house with a rainwater tank and constant daily water demand was created using the modelling platform Urban Developer [eWater Cooperative Research Centre, 2011]. This model was designed to replicate one of the configurations used in Kuczera [2008], that is, a harvestable roof area of 75 m^2, a tank size of 2500 L, a total daily water demand of 150 L cap^{-1} d^{-1} with 86 % of the water demand potentially being satisfied by tank water and an occupancy of 2.7 cap. The demand was varied throughout the day according to a pattern with a morning peak at 6 am and an evening peak at 7 pm. The simulation was run using the 6 minute pluviograph data from the Sydney Observatory Hill weather station for the period from 1959 to 2010 [Bureau of Meteorology, Commonwealth of Australia, 2011].

2.2 Basic Surrogate Model

The tank yield calculated using the detailed model was then averaged over each month of the simulation period to produce an average daily value. These results were stored in a resource file together with the corresponding total monthly precipitation. A scatterplot of the resource file is shown in Figure 1.a.

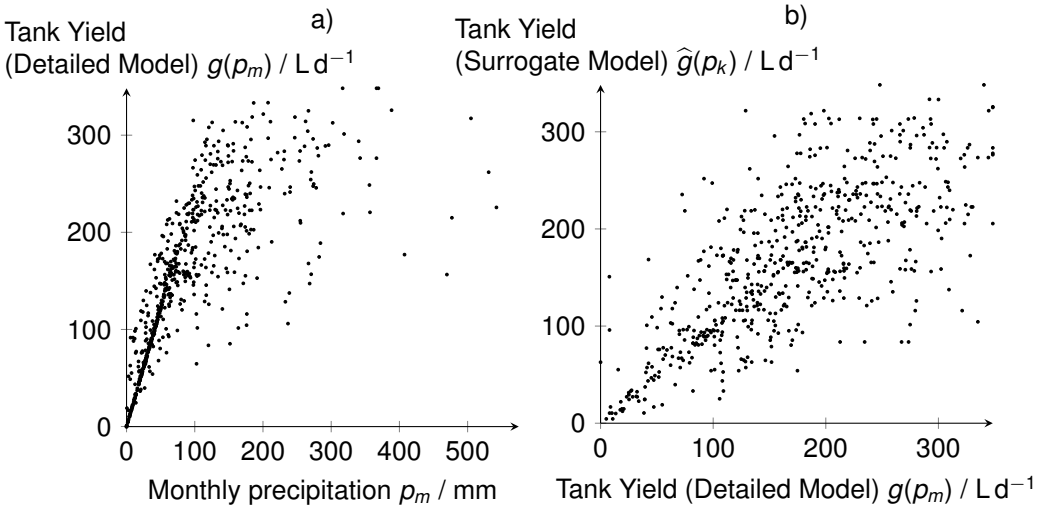


Figure 1. a) Resource file for the surrogate model, generated using the detailed model with historical rainfall data. b) Scatterplot comparing the output of the detailed model to the output of surrogate model using the resource file shown in a).

As can be seen in the scatterplot, the tank yield is positively correlated with monthly precipitation. There is a well defined linear relationship between tank yield and precipitation up to 75 mm/month: Points on this line correspond to months where all rainfall was stored in the rainwater tank and subsequently used. With increasing precipitation, the relationship becomes less well defined. For instance a monthly rainfall of 150 mm can coincide with tank yields between 75 L d^{-1} and 300 L d^{-1}.

The resource file is used in the surrogate model, which can be expressed as

$$\widehat{g}(p_m) = g(p_k)$$

where $\widehat{g}(p_m)$ is the average daily tank yield returned by the surrogate model given a monthly precipitation p_m; p_k is a monthly precipitation value drawn randomly from the k precipitation values in the resource file that are most similar to p_m, and $g(p_k)$ denotes the average daily tank yield in the resource file that corresponds to p_k. Generally, the square root of the number of data points is a sensible choice for the value of k [Lall and Sharma, 1996].

This surrogate model can be used to generate a tank yield timeseries from a rainfall timeseries at a monthly timestep. To analyse the performance of the surrogate model in emulating the detailed model, both were run with the same synthetic rainfall timeseries. The rainfall timeseries was generated using the DRIP model [Heneker et al., 2001] with a 6-minute timestep based on the historical Sydney Observatory data.

For use in the surrogate model, the rainfall timeseries was first aggregated to a monthly

timestep. For each month in the aggregated rainfall timeseries, the surrogate model then randomly returned the tank yield for one of the $k = 25$ nearest data points in the resource file, generating a monthly tank yield timeseries. The detailed model used the rainfall timeseries directly, generating a daily tank yield timeseries, which was then also aggregated to a monthly timestep. Figure 1.b) shows a scatterplot comparing the two monthly tank yield timeseries: the surrogate model output resembles the detailed model output well for lower tank yields, but the variance increases significantly for larger yields. The average Nash-Sutcliffe efficiency [Nash and Sutcliffe, 1970] of 100 runs of the surrogate model is 0.47 with a standard deviation of 0.03. Given the simplicity of the model, the performance is quite good. There is, however, still room for improvement.

2.3 Extended Surrogate Model

To improve the model performance, factors causing scatter in the resource file were investigated. It was found that departures from the linear relationship between monthly precipitation and tank yield were mainly caused by two effects: Spillage from the tank due to high intensity rainfall events, and precipitation occurring near the end of a month, which fills the tank but does not contribute to tank yield in the same month. To correct for these effects, it is proposed to use a corrected monthly precipitation in the surrogate model instead of the total monthly precipitation.

Large rainfall events: To correct for the spillage caused by high intensity rainfall events, the measured daily rainfall can be clipped, that is, all rain above a certain daily threshold is assumed to not contribute to the tank yield.

Rainfall occurring near the end of the month: Rain falling in the last days of a month would fill the tank and be used in the following month. The closer the rain occurs near the end of the month, the less it contributes to the tank yield during that month. To correct for this, a linearly decreasing weight can be applied to the daily rainfall values towards the end of a given month when calculating the monthly sum. Conversely, rain falling near the end of the preceding month can be included in the current months sum using a linearly increasing weight.

Using these corrections, the surrogate model becomes

$$\widehat{g}(q_m) = g(q_k)$$

where q_m is the corrected rainfall for month m, and q_k is a corrected monthly rainfall drawn randomly from the k corrected precipitation values in the resource file that are most similar to q_m. The corrected monthly rainfall is calculated as

$$\begin{aligned} q_m = \sum_{d=1}^{n_m} \cdot \min(c_1, p_{m,d}) \quad - \sum_{d=n_m-c_2}^{n_m} \frac{d-(n_m-c_2)}{c_2+1} \cdot \min(c_1, p_{m,d}) \\ + \sum_{d=n_{m-1}-c_2}^{n_{m-1}} \frac{d-(n_m-c_2)}{c_2+1} \cdot \min(c_1, p_{m,d}) \end{aligned}$$

where n_m is the number of days in month m, $p_{m,d}$ is the precipitation on day d of month m, c_1 is the threshold above which daily rainfall is clipped, and c_2 is the number of days at the end of the current and the preceding month that contribute to the monthly total with a lower weight. The parameters can be determined by using a steepest ascent hill climbing algorithm which searches for the values that maximise the correlation coefficient between q and g in the resource file. For the system configuration used here, with a tank-replaceable water use of 348 L d^{-1} and a 2500 L tank, the optimal parameters were found

to be c_1 = 12 mm and c_2 = 7 d. The value of c_2 is roughly equal to the ratio of tank size to tank-replaceable water use, and therefore the number of days it would take to use up all the water starting with a full tank.

Figure 2.a shows the resource file for the surrogate model using corrected precipitation with the optimal parameter set. The relationship between the yield and the corrected monthly precipitation exhibits less variance than the relationship between yield and total monthly precipitation (Figure 1.a).

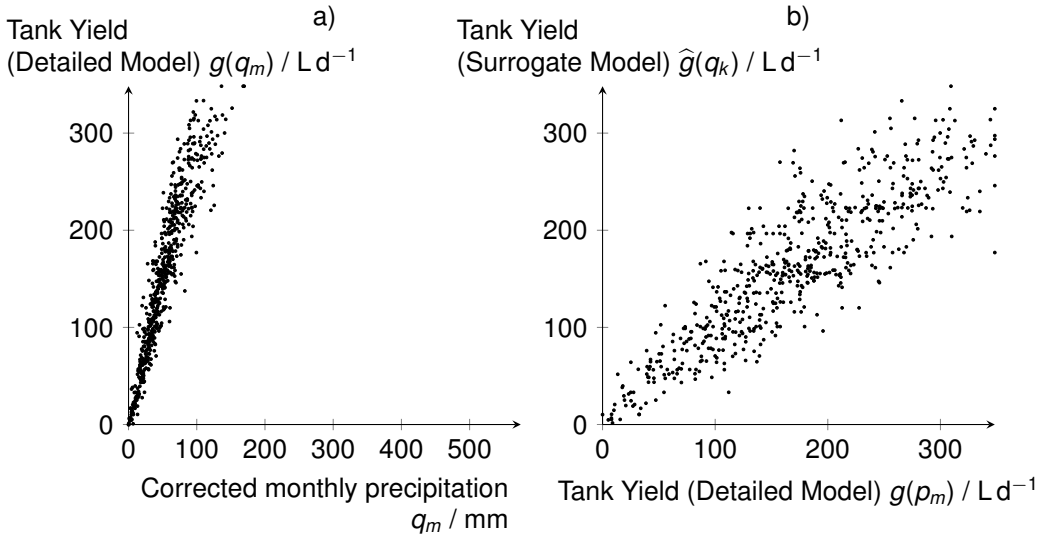


Figure 2. a) Resource file for the surrogate model, generated using the detailed model with historical rainfall data. b) Scatterplot comparing the output of the detailed model to the output of surrogate model using the resource file shown in a).

When using the corrected monthly precipitation instead of the total monthly precipitation to calculate tank yields from a synthetic rainfall timeseries, the model output is much closer to the output of the detailed model. This can be seen when comparing the scatterplot of the extended surrogate model and the detailed model (Figure 2.b) to the scatterplot of the basic surrogate model and the detailed model (Figure 1.b). There are far fewer outliers, and the variance is significantly less for larger yields. The average Nash-Sutcliffe efficiency of 100 runs of the extended surrogate model is much higher at 0.75 with a standard deviation of 0.01.

In a more complicated model setup, the water demand may vary by the month. To take this into account, instead of choosing a value from the k nearest neighbours in terms of just the corrected monthly precipitation, the month itself can be used as an additional predictor. In order to give the two dimensions of the search space (month and precipitation) equal weight, they can be normalised by dividing by their respective standard deviations [G. Kuczera, pers. comm. 2011].

2.4 Case Study

In order to see how the extended surrogate model performs in emulating more complicated model setups, five models of residential clusters were set up based on different

areas in Canberra, Australia in terms of occupancy and roof area. Model areas were chosen to reflect the actual housing mix in Canberra: Three of the five areas are predominantly made up of single houses, one of terraced flats, and one of apartment buildings. Roof areas were estimated from aerial imagery, and household sizes were taken from 2006 census data [Australian Bureau of Statistics, 2006], which for each cluster reports the number of households with a given number of occupants. Household size and roof area were randomly matched for this study. The average values used in the models are shown in Table 1. The harvestable roof area was assumed to be 50% of the total roof area, and the demand characteristics were reflect that used in Ravalico et al. [2011]: Indoor water use was assumed to be 188.4 L/cap d, 76 % of which are replaceable with tank water [Troy and Randolph, 2006], while outdoor water use was assumed to vary by month, between $17\,\mathrm{L\,d^{-1}}$ in July and $1285\,\mathrm{L\,d^{-1}}$ in January for the complete household [McMahon and Weeks, 1973], all of which was assumed to be potentially satisfied using tank water. None of the parameters used in the case study were tuned to optimise the model performance, and no study areas were omitted from the results.

Table 1. Parameters of the case study cluster models

Name	Number of Houses	Average Roof Area / m^2	Average Occupancy
Collier St	39	164.4	2.9
Mackenzie St	53	204.5	2.5
Cockle St	41	178.3	2.6
Dodgshun St	36	75.3	2.0
Kogarah Ln	8	328.8	30.2

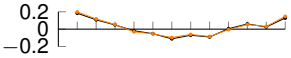
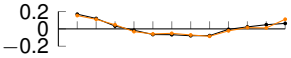
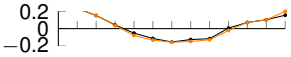
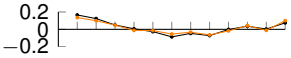
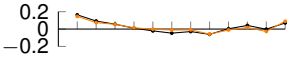
Each model was run using the historical rainfall data from the Canberra Airport weather station for the period from 1960 to 2010, and with a 50-year synthetic rainfall timeseries generated using the DRIP model. The model results from the run using historical data were used to generate the resource files for the surrogate model. For each model cluster, the parameters c_1 and c_2 for the surrogate model were determined using a steepest ascent hill climbing algorithm that optimised the correlation coefficient between the corrected monthly precipitation and the monthly tank yield. 100 realisations of yield timeseries for the synthetic rainfall timeseries were generated using the surrogate model. Those from the synthetic timeseries served as a test case for comparing the surrogate model output to the detailed model output. The detailed model was run on a 2.3 GHz CPU, taking about 90 minutes for a 50-year period; generating the 100 realisations using an R script takes only about 20 seconds on a 3 GHz CPU.

3 Results and Discussion

The goodness of fit measures of mass balance error and Nash-Sutcliffe efficiency in relation to the output of the detailed model, and the autocorrelation function up to lag 12 were calculated for each of the generated series; the results are presented in Table 2.

The mass balance error is smaller than 5 % for all clusters, which would be acceptable for many applications. For all clusters, a Nash-Sutcliffe efficiency of approximately 0.8 is obtained with low variance. For clusters with small tanks, an important factor for the good performance is the correction of monthly rainfall: If the cutoff c_1 is not used, the Nash-Sutcliffe efficiency drops to about 0.65 for the clusters with small tanks. For larger tanks with fewer spill events, reflected by larger calibrated values for c_1, the impact of the correction becomes less important.

Table 2. Parameters for the surrogate models and goodness of fit measures with standard deviations for 100 runs of the surrogate model on synthetic rainfall data. The average autocorrelation function up to a lag of 12 months of the surrogate model outputs and of the detailed model is shown in black and orange, respectively.

Cluster / Tank Size	c_1 [mm]	c_2 [–]	MBE (std dev) [%]	NSE (std dev) [–]	Autocorrelation Function [–]
Collier / 5 kL	49	4	1.89 (0.80)	0.86 (0.01)	
Mackenzie / 1 kL	10	1	3.29 (0.77)	0.81 (0.01)	
Cockle / 1 kL	21	2	1.85 (0.63)	0.80 (0.01)	
Dodgshun / 1 kL	30	1	−3.61 (0.91)	0.82 (0.01)	
Kogarah / 50 kL	45	6	−3.80 (1.09)	0.80 (0.02)	

The autocorrelation functions of the surrogate model output match that of the detailed model output well. This ability of the surrogate model to reproduce the autocorrelation function is due in part to the use of the rainfall recorded on the last c_2 days of the previous month, and in part to the use of the month as a predictor. In the Kogarah model, the total demand does not vary much between the months (see Table 1), since the variable outdoor demand is only a small fraction of the total demand in multilevel apartment buildings. In this case, if $c_2 = 0$ in the Kogarah model, the autocorrelation function of the average monthly tank yield would resemble that of the rainfall timeseries: In the synthetic timeseries used in these examples, it would drop very quickly. However, if the monthly demand varies strongly between the months, the use of the month as a predictor would lead to a reasonably good reproduction of the autocorrelation function even if $c_2 = 0$. In all cases, however, even a small value of $c_2 = 1$ d will improve the reproduction of the autocorrelation function.

4 Conclusions

The presented surrogate model is found to perform well in emulating the detailed node-link model at a monthly level, especially considering its simplicity and the much shorter run time. The consistently high NSE and the good reporduction of the autocorrelation function indicate that the surrogate model could serve as an adequate replacement for the detailed model in applications where short model run times are essential.

The surrogate model has been implemented as a plug-in for the modelling environment eWater Source IMS, which provides for a user-friendly and efficient way to upscale models without the need for processing the output of the detailed model using external tools [Fareed Mirza, pers. comm. 2011-10-28].

The general surrogate modelling approach could be adapted to be applied to other model setups, such as wastewater reuse or semi-centralised stormwater harvesting. This makes it possible to investigate the influence of innovative decentralised supply options on regional water supply systems with a high degree of confidence, allowing for the exploration of new approaches in water resources management.

REFERENCES

Altman, N. S. An introduction to kernel and Nearest-Neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

Australian Bureau of Statistics. 2006 census of population and housing, 2006. Cat. No. 2068.0 - 2006 Census Tables for Census Collection Districts: 8020310, 8012307, 8012910, 8013304, 8013804.

Bierkens, M. F. P., P. A. Finke, and P. de Willigen. *Upscaling and Downscaling Methods for Environmental Research*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.

Bureau of Meteorology, Commonwealth of Australia. Pluviograph data, 2011.

eWater Cooperative Research Centre. Urban Developer. Canberra, Australia, 2011.

Fonseca, L. G., H. J. C. Barbosa, and A. C. C. Lemonge. On Similarity-Based surrogate models for expensive single- and multi-objective evolutionary optimization. In Tenne, Y. and Goh, C., editors, *Computational Intelligence in Expensive Optimization Problems*, volume 2, pages 219–248. Springer, Berlin, Heidelberg, Germany, 2010.

Heneker, T. M., M. F. Lambert, and G. Kuczera. A point rainfall model for Risk-Based design. *Journal of Hydrology*, 247:54–71, June 2001.

Kuczera, G. Urban water supply drought security: a comparative analysis of complimentary centralised and decentralised storage systems. In *Proceedings of Water Down Under 2008*, pages 1532–1543, Adelaide, 2008. Engineers Australia/Causal Productions.

Lall, U. and A. Sharma. A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research*, 32(3):679–693, 1996.

Lloyd, S., T. Wong, and B. Porter. The planning and construction of an urban stormwater management scheme. *Water Science & Technology*, 45(7):1–10, 2002.

McMahon, T. A. and C. R. Weeks. Climate and water use in australian cities. *Australian Geographical Studies*, 11(1):99–108, April 1973.

Mitchell, V., N. Siriwardene, H. Duncan, and M. Rahilly. Investigating the impact of temporal and spatial lumping on rainwater tank system modelling. In *Proceedings of Water Down Under 2008*, pages 54–65, Adelaide, 2008. Engineers Australia/Causal Productions.

Mouritz, M. *Sustainable urban water systems: policy and professional praxis*. PhD thesis, Murdoch University, Perth, Australia, 1996.

Nash, J. and J. Sutcliffe. River flow forecasting through conceptual models part i a discussion of principles. *Journal of Hydrology*, 10(3):282–290, April 1970.

Ravalico, J., G. Dandy, and H. Maier. *Urban Water Optimisation of a cluster scale model in the Woden Valley, ACT*. eWater CRC, Canberra, Australia, 2011.

Troy, P. and B. Randolph. *Water consumption and the built environment: a social and behavioural analysis*. City Futures Research Centre, Kensington, Australia, 2006.

Wong, T. and M. Eadie. Water sensitive urban design – a paradigm shift in urban design. In *Proceedings of the Xth World Water Congress*, Melbourne, Australia, 2000.