Early Stopping Technique Using a Ga for Calibration in a Storm Runoff Model

Journal:	*The International Journal of River Basin Management*
Manuscript ID	JRBM-2020-0033.R1
Manuscript Type:	Research Paper
Date Submitted by the Author:	16-Sep-2020
Complete List of Authors:	Cu, Phuong; University of Technology, Sydney, School of Civil and Environmental Engineering Ball, James; UTS, Civil & Env Eng Dao, Ngoc Hung; National Education University, Faculty of Physical Geography
Keywords:	flood, SWMM, Calibration, catchment model

EARLY STOPPING TECHNIQUE USING A GENETIC ALGORITHM FOR CALIBRATION OF AN URBAN RUNOFF MODEL

PHUONG CU THI[(a)], JAMES E BALL[(b*)], NGOC HUNG DAO[(c)]

[(a)] *Faculty of Hydrology and Water Resources, Thuy Loi University, Tay Son Street, Hanoi, Vietnam. E-mail: phuongcuthi@yahoo.com*

[(b)] *School of Civil and Environmental Engineering, University of Technology Sydney, Broadway, NSW, 2007, Australia. E-mail: james.ball@uts.edu.au*

[(c)] *Faculty of Physical Geography, National Education University, Hanoi, Vietnam. E-mail: daongochung69@gmail.com*

ABSTRACT

Identifying suitable parameter sets for use in catchment modelling remains a critical issue in hydrology. This paper describes an early stopping technique (EST) for use during calibration of a multi-parameter urban catchment modelling system. The proposed method takes advantage of MODE and lower confidence limit (LCL) functions in statistical analysis of spanning set of objective function values. The paper also introduces a monitoring process and regularization techniques to avoid under/overfitting during the calibration and to enhance generalisation performance. The methodology is assessed using SWMM and linked with a Genetic Algorithm for calibration of a Powells Creek catchment model in Sydney, Australia. Results demonstrate that the statistical spanning set analysis approach overcomes issues of poor interpretation and deterioration in the model's generalisation properties. By stopping early, the calibration process avoided overfitting; this was indicated by too closely fitting to the calibration dataset and a failure to fit to the monitoring dataset.

Keywords: catchment model, early stopping, SWMM, flood estimation, calibration

INTRODUCTION

Management of storm runoff in urban areas remains a serious problem for many water facility managers. The flow data needed for this management generally are obtained either from model studies or field monitoring. Model-based studies have become very popular for catchments where there are no or insufficient flow data for management purposes. However, the reliability of model predictions depends on how well a model replicates the catchment response; in other words, the reliability of model predictions depends on the selection of suitable model parameter values.

The process of selecting parameter values is referred to as model calibration and involves the minimizing of an objective function over a search space; the search space represents the ranges of possible values for the parameters. A number of studies (Kuczera et al., 2006; Jin et al., 2010; Refsgaard and Storm, 1996; Cu and Ball, 2016) have discussed the potential for errors in the input and field data to influence the selection of parameter values resulting in over or under-fitting of the model.

Underfitting occurs when a model cannot adequately capture the structure of the data and therefore, suitable parameter sets are unable to be found. Overfitting, on the other hand, occurs when parameter set(s) are fitted too closely, or exactly, to a particular dataset and, therefore, fail to fit to data from other events, or fails to reliably predict future observations (Choi and Ball, 2002a).

There are alternative approaches to catchment simulation; these alternatives can be classified as regression based (pattern recognition) and process based. The issue of over and under-fitting of simulation models has received greatest attention in those models based on Artificial Neural Networks (ANN) which are a form of regression model (e.g. Coulibaly et al. 2000 and Piotrowski and Napiorkowski, 2013). While a number of different approaches have been attempted, Piotrowski and Napiorkowski (2013) reported that an Early Stopping Technique (EST) can be considered a good strategy to avoid over-fitting a catchment model.

The idea of an early stopping technique was first introduced by Nelson and Illingworth (1991), who wanted to avoid the problem of overfitting in large feed-forward ANNs. Applications of this technique appear in systems affected by noise or fitting signals (Zhang and Yu, 2005; Mahsereci et al., 2017). Furthermore, the technique has become one of the more common approaches applied to deal with overfitting issues in ANN models due to the good model performance and the quicker convergence (Piotrowski and Nappiorkowski, 2013). Finally, Coulibaly et al. (2000) states that, in an ANN application to a complex catchment system, EST makes it possible to avoid overfitting since the calibration stops as soon as the criteria are met.

While ANN based approaches have been found to provide a useful approach to modelling of catchment response to storm events, the more commonly used approaches would be classified as process based. Similar to the ANN based approaches, over-fitting of process models during calibration is possible. However, the focus during calibration of these catchment models has been parameter uncertainty (Beven and Binley, 1992; Refsgaard and Storm, 1996; Kuczera et al., 2006; Fang and Ball, 2007; Cu and Ball, 2016) rather than assessing whether the model has been over-fitted. These studies have demonstrated that, rather than a single parameter value, or set of parameter values, being suitable for replication of catchment response, there are multiple alternative parameter values, or sets of parameter values capable of replicating recorded catchment response to one or more storm events. As the number of parameters in a distributed catchment model usually is greater than the available data for their estimation, the potential for over-fitting of the parameter value pdfs has not been addressed in these studies. This poses the question of generalization of the pdfs of the parameter values.

Choi and Ball (2002b), however, considered potential over-fitting of a process-based catchment model through application of an EST in the search for a single optimal parameter set of values. In their study, optimal was defined as the set of parameter values that had the best generalization; in other words, the set of parameter values that resulted in the best replication over a range of events. An EST was employed to define when calibration of parameter values should cease; validation of parameter values was attempted then using independent events. Of the validation events considered, it was found that use of the EST resulted in the minimum difference between the predicted and field data occurred 81% of the time at an *early stop point* rather than after the specified number of calibration iterations had been completed. In other words, the EST approach avoided over-fitting the parameter values in 81% of the models tested. The study of Choi and Ball (2002b), however, did not consider parameter uncertainty. Hence, there is a need to investigate if an EST approach should be used when estimating parameter uncertainty; in other words, can the parameter value pdfs be enhanced using an EST during calibration. The investigation reported herein aims at addressing this question. An EST approach is applied during calibration, using a Genetic Algorithm, of an urban flood model applied to the Powells Creek catchment in Sydney, Australia.

CASE STUDY CATCHMENT DETAILS

General

Description of the case study catchment can be divided into three components, namely the catchment, the modelling software, and the available data. Each of these components will be presented in the following sections.

Catchment

The Powells Creek catchment (see Figure 1), sometimes referred to as the Strathfield catchment, is an 8.41km^2 catchment situated 10km west of Sydney's central business district. The drainage network comprises a closed piped system that opens out to a lined channel and then into the Parramatta River. The main open channel was established in 1892 and the closed pipe system was established in the 1920's. From a topographic perspective, the catchment is classified as having gentle slopes between 4% and 6% with a maximum elevation of 40m AHD; the minimum elevation is governed by the tidal regime of the Parramatta River.

Between 1958 and 2005, a gauging station on the main Powells Creek Stormwater Channel was operated by UNSW. The catchment area draining to this gauging station was 2.3km^2 of the total 8.41km^2. Details of the data collected are described later.

Insert Figure 1 Here

Model Software

There are numerous alternative software systems suitable for process-based modelling of existing and potential urban catchments. After considering these alternatives, SWMM (Rossman, 2005) was used for simulation of the Powells Creek catchment. This model has received extensive application; examples of applications are presented by Sun et al. (2014) and Shahed Behrouz et al. (2020).

For purposes of modelling, the catchment upstream of the gauging station was divided into 38 subcatchments based on topographic and drainage system characteristics. As discussed by Choi and Ball (2002a), these parameters can be categorised arbitrarily as:

- Measured parameters. These are parameters that are physically measured such as pipe diameters, catchment areas, rainfall depth, etc.; and
- Inferred parameters. These are parameters that are not measured and are determined from the application of a model. Examples of inferred parameters are Manning's roughness for catchment surfaces or channels, depression storage, catchment or subcatchment imperviousness, and the infiltration loss parameters.

In discussing inferred parameters (i.e. those parameters more likely to be adjusted during calibration) for a typical application of SWMM, Fang and Ball (2007) noted that there are 10 parameters. These parameters are the subcatchment width, the impervious percentage of the subcatchment, the percentage of the impervious area with zero depression storage, depression storage of the impervious and previous areas, the Manning's roughness for impervious and previous areas, and three infiltration parameters for Horton's infiltration equation.

Available Data

As previously noted, a gauging station was in operation during the period 1958 to 2005. In addition to the flow data, continuous rainfall data was collected at two locations within the gauged portion of the catchment; these locations were at the centroid of the gauged catchment and at the flow gauging station. While this rainfall data was collected for the same period as the flow data, only rainfall data for the period 1981 to 1998 from the flow gauging station was available for this study.

In addition to the flow data, continuous rainfall data was collected at two locations within the gauged portion of the catchment. While this rainfall data was collected for the same period as the flow data, only rainfall data for the period 1981 to 1998 was available for this study.

Details of the events used for the calibration, monitoring, and validation are shown in Table 1 and Figure 2.

Insert Figure 2 here

Insert Table 1 here

METHODOLOGY

Objective Function

Calibration of a catchment model involves selection of parameter values and the testing of these values using an objective function to assess the model performance. In this study, a Modified Nash Sutcliffe Efficiency (MNSE) was used as the objective function. Following Podger (2004), the MNSE can be expressed as

$$MNSE = 1 - \left[\frac{\sum_{n}^{i} \left(\left(Y_n^{obs}\right)^{\lambda} - \left(Y_n^{sim}\right)^{\lambda} \right)^2}{\sum_{n}^{i} \left(\left(Y_n^{obs}\right)^{\lambda} - \left(Y^{mean}\right)^{\lambda} \right)^2} \right] \qquad (1)$$

where: Y_n^{obs} and Y_n^{sim} are the observed and simulated flows at the nth increment of the hydrograph, and Y^{mean} is the mean observed flow over the n increments. As prediction of peak flow is the major concern when modelling flood events, a value of λ =2 was used to give more weight to high flows as recommended by Podger (2004).

Uncertainty in Modelling Parameters

Numerous previous studies (Beven and Binley, 1992; Refsgaard and Storm, 1996; Kuczera et al., 2006; Fang and Ball, 2007; Cu and Ball, 2016) have investigated parameter uncertainty. These studies have shown that there is a range of possible values for a given parameter and that, for each potential value for a parameter, there is an associated probability.

A common problem encountered when modelling individual events in a catchment is the need to ensure generality of selected parameter values. For this study, this concept is expanded to generality of parameter values and their associated probabilities.

Search algorithm

A Genetic Algorithm (GA) served as the search algorithm to estimate the range and likelihood of parameter values necessary for operation of SWMM. GAs were first developed by Holland (1975) and are a class of metaheuristic search algorithms that make use of evolutionary ideas to generate solutions for optimization problems. The principle of "survival" is accomplished by evaluating each candidate's fitness through an appropriate objective function and a biased random selection procedure of individuals for "reproduction", wherein higher rated candidates are more likely to be selected. This reproduction is undertaken using stochastic transformations inspired by natural evolution, such as inheritance, mutation, selection, and crossover (Goldberg 1989).

Linking a GA with a catchment modelling system for calibration means that each individual set of parameters is represented by an individual in the generated population of chromosomes. In this case, the chromosomes consist of 342 genes representing the 342 model parameters. To use the GA, a population of chromosomes will be generated from a random combination of values within the search space; for the generation of these parent chromosomes, the values were assumed to have uniform likelihoods. Based on the likelihood that use of a chromosome will result in replication of the catchment response, individual chromosomes were selected from the current population to be parents and reproduction; a child chromosome is accepted for the next generation only if the objective function (MNSE) is better than the parent.

The reproduction of children is conducted by changing chromosomes based on the following genetic features:

(i) real-value coding - the chromosomes consist of parameter values;

(ii) tournament selection with replacement for chromosomes (Miller et al., 1995); and

(iii) cross-over (with elitism) operations with a random level at uniform probabilities (Fang and Ball, 2007).

GAs can be used to produce an infinite number of generations. However, the under/overfitting problem exists, and hence there is a need to define when production of generations should cease. An EST approach is used for this purpose.

For testing the EST approach in calibration of a model for the Powells Creek catchment, a linkage between a GA and SWMM was developed. This linkage is illustrated in Figure 3.

Insert Figure 3 here

Using this linkage, the steps in the calibration process are:

1. An initial population of 600 parameter sets (referred to as the parent data sets) are developed. These parameter sets are uniformly distributed across the search space.

2. These 600 parameter sets are used with SWMM to generate 600 MNSE for the calibration and monitoring datasets.
3. The 300 datasets that provided the best MNSE were selected to produce 300 new datasets, known as children, for the next generation.
4. The 300 children datasets were used with SWMM to generate the MNSE for the new datasets. This provided a total of 600 parameter sets and MNSE at the new generation.
5. Steps 3 and 4 are repeated until the calibration process ceases.
6. For the purposes of this study, the maximum number of generations considered was 50.

Early Stopping Technique

The basic concept of the EST is to split the available data for calibration and validation into three parts: (1) a calibration (or training) set, used to determine the catchment modelling parameters; (2) a monitoring (or testing) set, serving to estimate the generalisation of the catchment modelling system performance and to decide when calibration must cease; and (3) a validation set, used to verify the effectiveness of the stopping criteria and to estimate expected performance in the future. Typical error curves of calibration and monitoring data are shown in Figure 4.

Insert Figure 4 here

During the initial phase of the calibration process, the objective function values for both the calibration and monitoring data will decrease. When the objective function values calculated for calibration data decrease, while those for the monitoring data increase, the calibration process has reached a critical point and should terminate. Further calibration is likely to result in the parameters being fitted to errors present in the data, rather than to the reduction of prediction errors arising from inappropriate parameter values, in other words – overfitting occurs.

Early Stopping Technique Application

A novel aspect of the research reported herein is analysis of the objective function for each generation as part of an EST. To achieve this aim, regularization techniques were applied to the vector of MNSE values to find convergence points of the generalization process and, consequently, identify stopping points for the calibration process. Use of two statistical characteristics, i.e. the mode and the confidence limits of the MNSE distribution, as a regularization technique is proposed. The use of these statistical characteristics as part of an EST process follows.

Use of the statistical characteristics in an EST process requires the following definitions:

- The possible values of MNSE are the space (R) which a set of real numbers;
- A sample of the population in each generation of the GA is defined by I (I = {1,2,3,… }), a set of integer numbers;
- The number of a dataset (flood event) is defined by J (J = {1,2,3,… }), a set of integer numbers; and
- The generation in the GA is defined by K (K = {1,2,3,… }), a set of integers.

Using these definitions, the MNSE is a vector of values comprising j components at generation k. The statistical characteristics of the vector MNSE are used in the EST process.

The calibration procedure can be evaluated using a span function of MNSE generated from j datasets. The Span function is a combination of all members from all subspaces which meet objective function constraints and can be expressed by Eq. (2) (modified from Zhang 2005):

$$\mathrm{Span}(S) = \{MNSE^{i,j,k}: MNSE^{i,j,k} \in R \text{ , } i \in I, j \in J, k \in K\} \qquad (2)$$

where i is the population number in a GA generation, j is the number of datasets, and k is the number of iterations (generations). Convergence of the calibration process occurs if a concave/convex function (A) of Span S for k generations exists. The early stopping point can be identified by the Suprema/Infimum function A (sup/inf A) of span S and can be expressed in the following form:

$$\text{Stopping point} = \sup/\inf A(f); f \in \mathrm{Span}(S) \qquad (3)$$

The aim of this study is to find a concave function (A) which indicates convergence of the calibration metric (MNSE). To find this convergence point, two statistical functions were tested, namely MODE and lower confidence limit (LCL) of the vector MNSE.

RESULTS AND DISCUSSION

Objective function analysis

At each generation, the value of the objective function (MNSE) was calculated for each chromosome in the population (300 chromosomes) for 2 events. The two statistical attributes MODE and LCL were determined for analysis of the MNSE distributions to identify suitable stopping points. A moving average method was employed also.

Method 1: Trend analysis of MODE values

At each generation k and for j events, MNSEs were fitted by a distribution and MODEs were calculated. The Span X, therefore, consists of MODEs at k generations for the calibration and monitoring events. Presented in Figure 5 are MODEs over 50 generations for the 2 selected events. As can be seen from this figure, MNSE strongly fluctuates during the calibration process. Overall MNSE increased during the first 2 steps from 0.85 to more than 0.90, followed by a downward trend for both validation and monitoring events. However, individual values are largely scattered around these trends. This indicates the absence of convergence in the calibration process and hence stopping points were undefined.

INSERT FIGURE 5 HERE

To identify a convergent point a moving average method was applied with step sizes of 3, 4 and 5. Figure 5 shows evolution of the calibration metric over the generations during the calibration process. During calibration, the maximum function value is obtained virtually from the starting point (after the first step), and meanwhile, the monitoring process produces 3 stages according to different reduction rates. Nevertheless, this process resulted in poor interpretation and hence the early stopping point was undefined.

Method 2: Analysis of confidence limits

Analysis of the confidence limits is aimed at improving estimation of the stopping point compared to the first method. At each generation, the 90% LCL of MNSE (i.e. 95% of the MSNE were greater than this value) in the spanning set S for each event was calculated (see Figure 6).

INSERT FIGURE 6 HERE

Figure 6 illustrates the changes in LCL for the 3 datasets (calibration, validation and monitoring). The maximum value of the LCL of MNSE was 0.906 occurring at the 15th step. This confirmed that the best stopping point was the 15th generation which resulted in 95% of the population having a MNSE equal to or exceeding 0.906. The MODE values at step 15 were 0.90 and 0.98 in calibration and monitoring, respectively. The distribution of MNSE at the 15th generation is illustrated in Figure 7.

INSERT FIGURE 7 HERE

Parameter analysis

This section critically examines how the calibration process performed by evaluating the statistical properties of the SWMM model parameter values at different stages. Based on an analysis of the objective function shown in Figure 6, the fitting process can be divided into 3 stages, namely:

- Stage 1: From the beginning to generation 15.
- Stage 2: From generation 15 (stopping point) to generation 31.
- Stage 3: From generation 31, where there is a dramatic drop in the monitoring process performance, to the end of the fitting process (generation 50).

Subcatchment weighted average values over the catchment were calculated for each parameter category. These values were fitted by a normal distribution using the closed-form Maximum Likelihood Estimation of Pandey and Nguyen (1999). Statistical parameters (mean and standard deviation) of each parameter at 3 points (first; generation 15; and the last generation) are shown in Table 2 and are illustrated in Figure 8.

INSERT FIGURE 8 HERE

INSERT TABLE 2 HERE

Statistical differences in parameter sets at these four critical points were tested by 2 criteria; the first test was a Welch Two Sample t-test (Welch 1951), and the second test was an F-Test (Lomax 2007). The Welch Two Sample t-test is a test of difference in mean between two normal distributed data sets based on assumptions of unequal variance and application of the Welch distribution modification with confidence limits of 90%. The F-test seeks to compare the two variances. Both these tests use p-values to accept or to reject the null hypothesis. If the p-value is less than the chosen alpha level, then the null hypothesis is rejected and there is evidence that the data tested do not originate from the same population. Commonly, an alpha value of 0.05 is accepted (Welch 1951). An alpha value of 0.05 indicates there is less than 5 chances out of a hundred that a sample came from a population where that was not true. Hence, a p-value of more than 0.05 means that two data sets can be assumed to be similar. Results of the t-test and F-test for each stage are shown in Tables 3 and 4.

INSERT TABLE 3 HERE

INSERT TABLE 4 HERE

From inspection of Figure 8 and Tables 3 and 4, it can be seen that many parameters attained stability after several generations. The mean and standard deviations during the 3 stages were similar indicated by the p-values in the t-test and the F-test were more than 0.05 (Tables 3 and 4). There was a significant shift in the distributions of two parameter classifications, namely the weighted average catchment width and percentage of impervious area. These two parameter classifications influence the catchment lag and rainfall loss (impervious areas in SWMM are assumed to have no continuing losses) respectively. These two parameter categories warrant further analysis. This analysis is presented in terms of the stages previously noted.

Stage 1: From the starting point to the 15th generation

During this stage of the calibration process, the mean values were moved and the standard deviations were narrowed (less scattered values) (Figure 8b and 8c). This is illustrated by differences in mean and standard deviations between the distributions resulted from the starting point and the 15th generation. The p-values in the t-test and F-test were less than 0.05 (Tables 3 and 4) and ratios of variances were 1.99 and 8.13 for catchment width and catchment impervious area, respectively. This proves that the change in parameter values during the calibration process resulted in improvement of the model's performance. This was indicated by robustness in objective function values (MNSE). The MODE of the MNSE increased from 0.839 to more than 0.95 (Figure 6a). The LCL rose from 0.737 to more than 0.906 for the calibration, validation, and monitoring events, respectively (see Figure 7).

Stage 2: From the 15th generation to the 31st generation

As the calibration process continued there was a shift in the parameter means. Use of an F-test showed significant differences in variance; this is illustrated by a narrowing in the parameter distributions (Figure 8c, 8c). The p-values in the F-test were less than 0.05 and the ratios of variances increased to 4.01 for the impervious area parameter, 2.21 and 2.36 for width and pervious area depression storage, respectively. This indicated that changes had occurred in these sensitive parameter distributions.

However, there was a slight improvement in the model's calibration performance. The MODE fluctuated at around 0.9 (Figure 6a) and a decline in the LCL was observed (Figure 7). This signal indicated that the calibration process had started to fit the noise/error of the outputs rather than the model structure. As a result, despite the calibration process slightly improving there was, however, a decline in validation and monitoring performance.

Stage 3: From the 31st generation to the end

In this stage, the parameter distributions were similar as illustrated by p-values in the t-test and F-test more than 0.05 (Tables 5 and 6). Furthermore, the distributions did have a similar shape (Figure 8b, 8c). The variances were slightly different with the ratio being around 1.0. This similarity suggested that the fitting process in fact presented poor performance in terms of refining the model parameter values (no significant change in parameters). However, it continued to fit the flow errors/noise. As a result, the process only slightly improved the calibration and validation. It, however, failed to fit the monitoring dataset. In this case, the MODE in the monitoring event dramatically dropped to 0.71, and the LCL during monitoring fell to 0.723. This can be deemed an example of poor performance.

INSERT TABLE 5 HERE

INSERT TABLE 6 HERE

From the above analysis, it can be concluded that the first stage was the best calibration stage while the best point at which to stop calibration was the 15th generation.

CONCLUSION

The focus of this study was the testing of an EST applied during the calibration process of a catchment modelling system. Of particular concern was the feasibility of using an EST during the estimation of the parameter uncertainty. The EST approach was applied successfully to modelling an urban catchment for flood estimation. Furthermore, the EST approach overcame the problem of overfitting in parameter estimation.

The EST approach applied in this study was based on statistical analysis of the objective function (MNSE) at each generation of the GA. The two statistical characteristics considered were the Mode and LCL. It was found that the LCL allowed easier identification of a stopping point and the point where values of the parameters started losing generalisation through over-fitting.

While the approach was found to be viable, the analysis undertaken required a number of generations of the GA to be completed. Hence, the stopping points could be identified by analysis of the MNSE only after the simulation process had been completed; the analyses undertaken relied on post-processing data from the GA and not processing the data during the GA. Further development of the approach is needed to mitigate this issue.

ACKNOWLEDGMENTS

The authors would like to thank Faculty of Hydrology and Water Resources, Thuy Loi University, School of Civil and Environmental Engineering, University of Technology Sydney, Faculty of Physical Geography, National Education University for supporting this study. The authors sincerely thank the editor and the anonymous reviewers for reviewing the manuscript and providing valuable comments to improve the manuscript.

REFERENCES

Beven, K. and Binley, A., (1992), The future of distributed models: Model calibration and uncertainty prediction, Hydrological Processes, 6:279-298. doi:10.1002/hyp.3360060305

Choi, K.-S. and Ball, J.E., (2002a), Parameter estimation for urban runoff modelling, *Urban Water*, 4(1):31-41.

Choi, K.-S. and Ball, J.E., (2002b), A Generic Calibration Approach: Monitoring the Calibration, *In Proc. Hydrology and Water Resources Symposium 2002*, Melbourne, Australia, Pub Institution of Engineers Australia, Barton, A.C.T 650-657.

Coulibaly, P, Anctil, F, and Bobee, B, (2000), Daily reservoir inflow forecasting using artificial neural networks with stopped training approach, *Journal of Hydrology*, 230:244–257.

Cu, P. and Ball, J.E., (2016), The influence of the calibration metric on design flood estimation using continuous simulation, *International Journal of River Basin Management* 15(1):9-20.

Fang, T. and Ball, J.E., (2007), Evaluation of spatially variable control parameters in a complex catchment modelling system: A genetic algorithm application, *Journal of Hydroinformatics*, 9:163-173.

Goldberg, D.E., (1989), *Genetic algorithms in Search, Optimisation and Machine Learning*, Addison-Wesley Publishing Co. Inc., Reading, Mass., USA

Holland, J. H., (1975), Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, MI, USA

Jin, X., Xu, C.-Y., Zhang, Q. and Singh, V.P., (2010), Parameter and modeling uncertainty simulated by GLUE and a formal Bayesian method for a conceptual hydrological model, *Journal of Hydrology*, 383:147-155.

Kuczera, G., Kavetski, D., Franks, S. and Thyer, M., (2006), Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters, *Journal of Hydrology*, 331:161-177.

Mahsereci M., Balles L., Lassner C., and Hennig P., (2017), Early Stopping without a Validation Set, CoRR abs/1703.09580(1703.09580).

Nelson, M.C. and Illingworth, W.T., (1991), *A Practical Guide to Neural Nets*, Addison-Wesley, Reading, MA, USA

Piotrowski, A. P. and Napiorkowski, J.J., (2013), A comparison of methods to avoid overfitting in neural networks training in the case of catchment runoff modelling, *Journal of Hydrology*, 476:97-111.

Podger, G., (2004), Rainfall Runoff Library User Guide, Unpublished Report, CRC for Catchment Hydrology, Australia.

Refsgaard, J.C. and Storm, B., (1996), Construction, calibration and validation of hydrological models, In Abbott, M.B., Refsgaard, J.C. (Eds.), *Distributed Hydrological Modelling*, Kluwer Academic, 41-54.

Rossman, L., (2015), Storm Water Management Model User's Manual Version 5.1 – Manual, US EPA Office of Research and Development, Washington, DC, *EPA/600/R-14/413* (NTIS EPA/600/R-14/413b).

Shahed Behrouz, M., Zhu, Z., Matott, L.S. and Rabideau, A.J., (2020), A new tool for automatic calibration of the Storm Water Management Model (SWMM), *Journal of Hydrology*, 581, doi.org/10.1016/j.jhydrol.2019.124436

Sun, N., Hall, M., Hong, B. and Zhang, L.J., (2014), Impact of SWMM Catchment Discretization: Case Study in Syracuse, New York, *Journal of Hydrologic Engineering*, 19(1):223-234.

Welch, B. L., (1951), On the Comparison of Several Mean Values: An Alternative Approach, *Biometrika*, 38:330–336. doi:10.2307/2332579.

Zhang, T. and B. Yu., (2005), Boosting with Early Stopping: Convergence and Consistency, *Annuals of Statistics*, 33(4):1538-1579.

No.	Comment	Response
Reviewer 1		
1	Line 1, Title, it is better to use "Genetic Algorithm" instead of "GA"	Agreed
2	Line 12 "this paper descrided" better "describes". Same line 15 "introduces"	Agreed
3	Figure 1. Replace y label "No of Function evaluations" by "Number (or No.) of function evaluations	Figure caption corrected to provide correct citation. Figure is copied from source.
4	Figure 2. Improve this figure, it looks like it was cut from the reference not properly	No longer relevant
5	Please check your Introduction, you have a mix of past and present tense in the way you cite references.	The abstract has been revised and corrected.
6	Line 112, Please introduce MNIST dataset for those who are not familiar with it	No longer relevant
7	Line 117, which three techniques?	Clarified in new version of manuscript
8	Line 122, you already said this in the previous line 119.	No longer relevant
9	Line 131, what is the most relevant part of Nguyen's work (2017)?	Nguyen reference deleted
10	Lines 133-138, A table should serve better to introduce the model's parameters. The table has to shows parameter, a brief description, and range of values that parameters can take.	Rewritten to cite references where the suggested approach has been used
11	Lines 145-147 Please introduce these events and mention the most relevant of them.	Agreed. Section rewritten.
12	The Methodology is not clear. The description of Methodology section needs considerable improvement. The steps followed should be mentioned; perhaps a scheme could help you to do this. Why do you only use Modifier Nash Sutcliffe coefficient (MNSE) as the objective function?	Methodology clarified in a rewrite of this section of the manuscript.
13	Please improve the Application section, this is not a technical report. Any other researcher interested in replicating your work must have all the elements to do it.	Methodology clarified in a rewrite of this section of the manuscript.
14	In the results and discussion, I think the following elements should be discussed: -Among the methods to avoid overfitting, early stopping is the most used in practice, according to Piotrowski and Napiorkowski (2013). Although the manuscript does not mention or address other techniques, what are the main advantages of the proposed technique over others? -What are the main drawbacks of applying the technique presented in the paper? -What are the main difference and advantages of the technique proposed in the paper with the called optimized approximation algorithm introduced in Liu et al. (2008)?	Results and discussion sections have been modified to clarify the important aspect.
15	In your references, the most recent is from 2018 (only one), being the topic that you	Literature review has been revised, 2020 references now included.

	present in the paper of high relevance, there are no more recent references in this regard?	
16	Line 386, What was the purpose of your research?	Clarified in the rewritten introduction
17	Line 387, To whom or what does "Its" refers?	Clarified in the rewritten introduction.
18	Lines 389, What methods? Did you make a comparison between different methods? Why do you mention this?	Clarified in the new manuscript
Reviewer 2		
1	Language has to be thoroughly revised. At the moment it does not real well. The literature review also misses some key papers on hydrological model calibration, which are essential to your discussion. Also, the figures require significant rework for clarity.	Language has been checked and major changes made. Figures have been reworked for clarity.
2	Model calibration is not an issue about performance, but rather to estimate the un-observable parameters of a system. This is conceptual difference between regression models and conceptual/physically-based models. The issue of overfitting refers to models whose complexity (namely number of parameters) is larger than what the available data supports. This is the reason why there are hardly reported "overfitting" issues in conceptual model calibration, as you accurately report on your literature review.	While the Reviewer makes a valid point, the practice of catchment modelling commonly requires the use of complex models where the relationship between the model process and real process lies only in the terminology. For example. A 2D model of flow on a road surface will require different values of Manning's 'n' depending on the alignment of the grid with the road – the model and the real world energy losses must be similar if the flow propagation is to be reliably simulated. However, the flow lengths will differ between the real world and the model depending on the grid orientation leading to the need for differing values of the same parameter name. Consequently, there is a need to investigate this practical problem.
3	I think the literature review should be more profound, by having more references to model calibration issues, as data-driven models (as is a big chunk of your literature) does have over/under fitting issues as parameters are added in function of the available dataset. In conceptually and physically-based hydrological models, the number of parameters depend on the process and its approximation and not on the data availability, thus the same model does not increase in complexity depending on the available data. The issue with conceptual/physically-based type of models is usually in the parameter identifiability (due to equifinality), as its number is usually fixed. In other words, the number of model parameters in a conceptual model does not change once more data is available.	I disagree with the reviewer and particularly with respect to the definition of a model. The model in the context of this manuscript comprises the modeller's concept of hydrologic and hydraulic flow paths, the available data, and the processes embedded in the software. In this context, a change in data does force a change in the model – for example, a new DEM may change the subcatchments due to greater resolution. As the parameters are spatially distributed and functions of land-use, etc, a change in DEM may change parameter values. However, the Reviewer is correct in that equifinality is an issue in identification of parameter values. In this paper, this issue is addressed through consideration ensembles of parameter values. As the desire is to use the ensemble for prediction, there is a need to ensure generalisation of the adopted parameter values is not comprised due to systematic and random errors in the model system.
4	In addition, you are using data from only 1 storm for calibrating the model. In my experience, this is quite little data to draw conclusions, considering the size of the	The Reviewer is correct that more event data would enhance the paper. However, the problem being dealt with is one of flood prediction. Hence, a greater number of events

	problem you are dealing (over 300 parameters). I would recommend using a case study with more data availability.	would require a longer period of monitoring. As described in the paper, only a 20 year period of data was available for the study. Hence, in that period the expectation of a 1%AEP event is less than 0.2 and for a 5%AEP is approx. 0.6. As a result, meaningful events are rare. A similar result would exist for any other urban catchment – considering a longer period has the potential for non-stationarity in catchment development to influence model parameters.
5	Also, the methodology can be clearer, as it is really hard to follow. For example, in l212 I is defined as the population number in each generation of GA, while in l221 i the number of GA generations and, later in l299 each generations is defined by k. Similarly, you present the index J, which indicates the number of datasets, while I suppose you should stick to calibration set to generate the new parameterisation, validation to tests its performance, and testing (or monitoring if I understand correctly) to evaluate the performance on unseen data (which should not be used to retro-fit the model as it mimics "operational conditions").	Agreed. Rewritten to clarify this point.
6	l14 There is not indication of what is MODE or LCL	MODE and LCL are defined in the revised version – note that these two characteristics are standard statistical parameters.
7	l31-32 I don't think the only reason for models is lack of data. Actually you do require data to build models	Agreed. Rewritten to clarify.
8	l151 why in particular using MNSE instead of simply NSE? Is it the intention to get the peak value right, or is it important to understand the dynamics of the system as well? If so, perhaps a MINMAX performance function would be more informative	MNSE was used to bias the metric to higher flows rather than the more frequent low flows that were not of interest in this study. This has been clarified in the text.
9	l175-196 Please avoid using metaphors (such as 'survival', 'reproduction', 'chromosomes', 'genes') to describe GA	These terms are standard in the GA literature. However, the use of these terms has been minimised.
10	l188 Are there 342 model parameters using only 3 storms for the model calibration?	Yes, a common problem in practice.
11	l197 it is named a metaheuristic	corrected
12	l227-231 Is quite hard to read but seems to be essential in describing the methodology. I have two interpretations of this 1) you will find a set of parameters that yield a value of 1, thus indicating the "convergence" of the calibration process. 2) Find a concave function "A" that contains said value (MNSE = 1). In both cases, having a value of 1 in the calibration set is not overfitting already?	This section has been clarified in the revised text.
13	l260 Why an initial population of 600?	Previous knowledge – see Fang and Ball (2007).
14	l267 How did the parent generate the children?	Details are beyond the scope of the current paper but some clarification has been added.
15	l269 How did the children and parents were combined?	Details are beyond the scope of the current paper but some clarification has been added.

16	l273 Why to stop at 50 generations?	The point was identification of an EST, 50 generations were considered sufficient for that purpose.
17	Table 2 Why to calibrate percentage of impervious area? This is an observable parameter	I disagree. The model assumes effective impervious area – 100% of rainfall becomes runoff from the effective impervious area, not the total impervious area. This has been discussed extensively in the urban drainage literature. Furthermore, impervious areas need not be effective impervious areas as some runoff from impervious areas may flow onto pervious areas and not contribute to the subcatchment runoff.
18	Figures require considerable rework	Figures have been reworked.
19	Figure 5 Why to use NSE instead of MNSE?	MNSE was used as it is not a traditional NSE approach.
20	The notation of the equations can be greatly simplified	This has been considered with additional text to clarify the equations.

List of Tables

Table 1: Details of selected events for testing the model

Events	**Start date**	**End date**	**Peak flow (m^3/s)**
Calibration	13:05 23/10/1985	17:35 23/10/1985	11.89
Monitoring	08:02 07/10/1997	15:00 17/10/1997	5.93
Validation	08:02 29/06/1997	13:57 29/06/1997	6.59

Table 2: Statistical parameters of the model parameters

Parameter	**1st generation**		**15th generation**		**50th generation**	
	Mean	STD	Mean	STD	Mean	STD
Width (m)	739.7	65.00	566.4	46.03	439.9	28.16
Percentage of impervious area (%)	43.7	0.72	44.4	0.38	44.6	0.39
Average Catchment Roughness	0.039	0.0036	0.039	0.0032	0.040	0.0032
Imp. area depression storage (mm)	3.04	0.28	3.15	0.29	3.40	0.22
Pervious area depression storage (mm)	28.9	2.74	29.3	2.87	28.4	2.16
Max. Infiltration rate (mm/hr.)	362.8	26.18	367.1	28.10	345.3	22.12
Min. Infiltration rate (mm/hr)	21.3	1.55	21.2	1.56	20.3	1.43
Infiltration decay rate	7.03	0.41	6.29	0.38	6.44	0.40

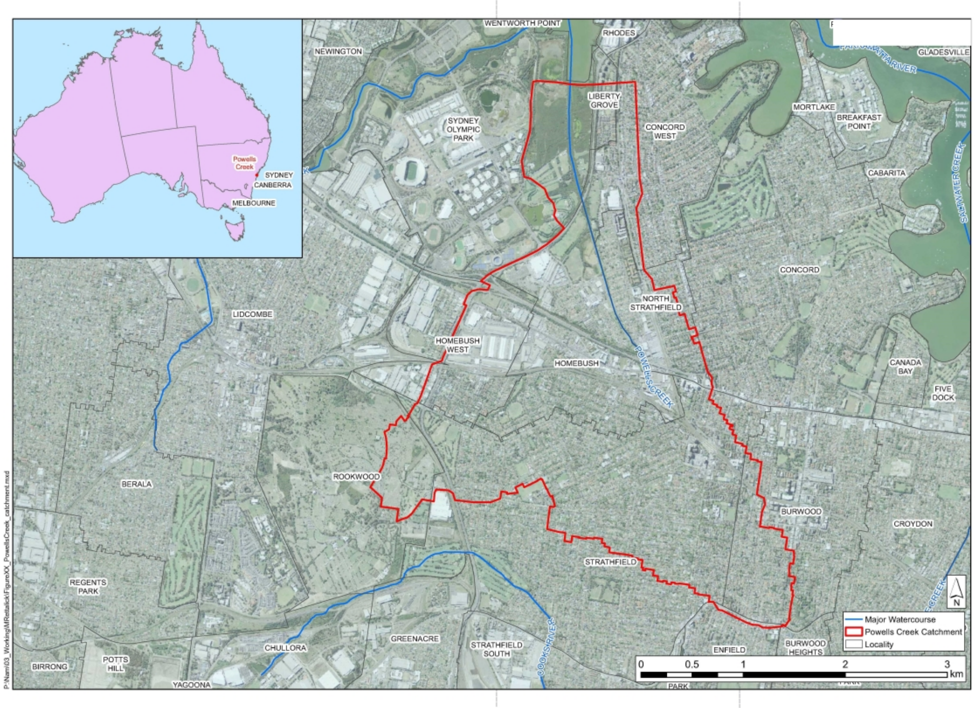
Table 3: t-test for the differences in model parameter distribution at selected generations

Parameters	p-value		
	1st -15th generation	15th -31st generation	31st-50th generation
Width (m)	2.2E-16	2.20E-16	2.20E-16
Percentage of impervious area (%)	4.54E-06	2.20E-16	0.256
Average Catchment Manning	0.185	1.21E-05	0.882
Imp. depression storage (mm)	2.46E-06	7.03E-14	7.75E-05
Pervious area depression storage (mm)	0.073	2.20E-16	2.20E-16
Max. Infiltration rate (mm/hr.)	0.055	0.000125	8.39E-10
Min. Infiltration rate (mm/hr)	0.172	1.38E-05	0.004
Infiltration decay rate	3.384E-13	8.66E-12	0.150

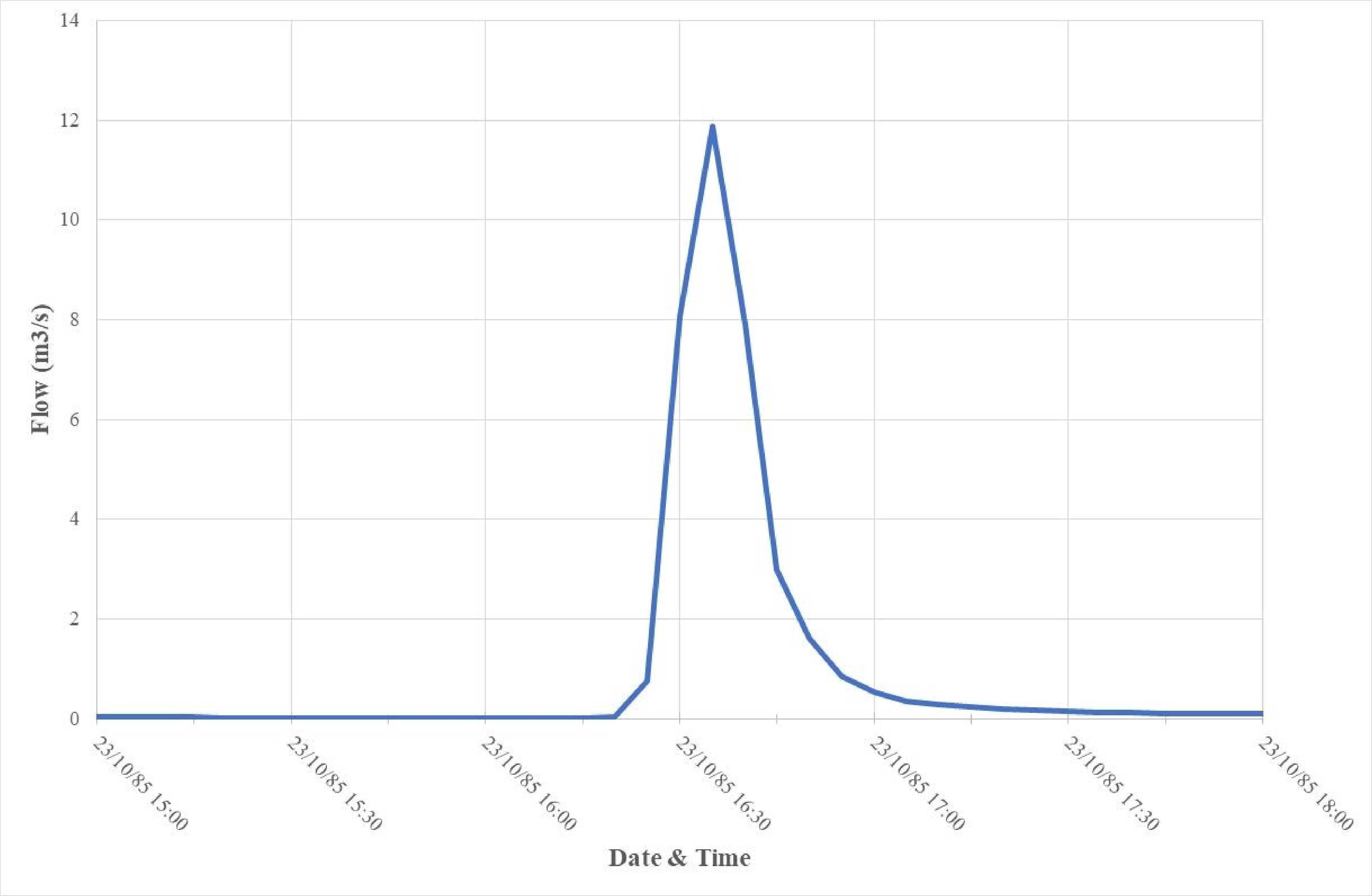
Table 4: F-test for the differences in model parameter distribution at selected generations

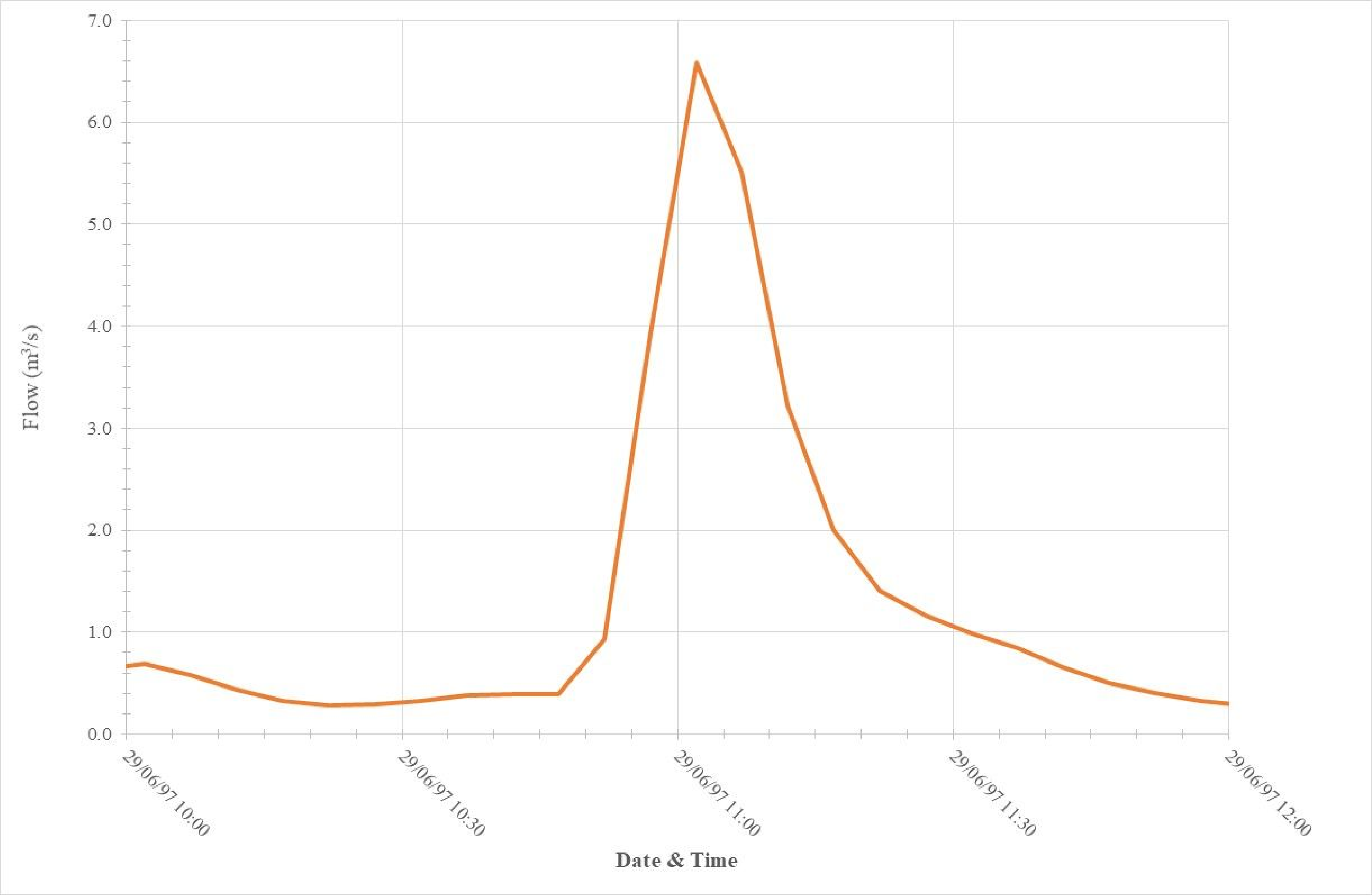
Parameters	**Between 1st -15th generation**		**Between 15th -31st generation**		**Between 31st -50th generation**	
	p -value	Ratio	p-value	Ratio	p-value	Ratio
Width (m)	3.56E-09	1.99	1.25E-11	2.21	0.104	1.21
Percentage of impervious area (%)	2.20E-16	8.13	2.20E-16	4.01	0.188	0.86
Average Catchment Manning	0.066	1.24	0.048	0.80	0.045	1.26
Impervious area depression storage (mm)	0.758	0.96	0.136	1.19	0.002	1.43
Pervious area depression storage (mm)	0.429	0.91	2.80E-13	2.36	0.012	0.75
Maximum Infiltration rate (mm/hr.)	0.222	0.87	0.981	1.00	3.56E-05	1.62
Minimum infiltration rate (mm/hr.)	0.924	0.99	0.843	0.98	0.087	1.22
Decay rate of infiltration in Horton’s equation	0.265	1.14	0.877	0.98	0.610	0.94

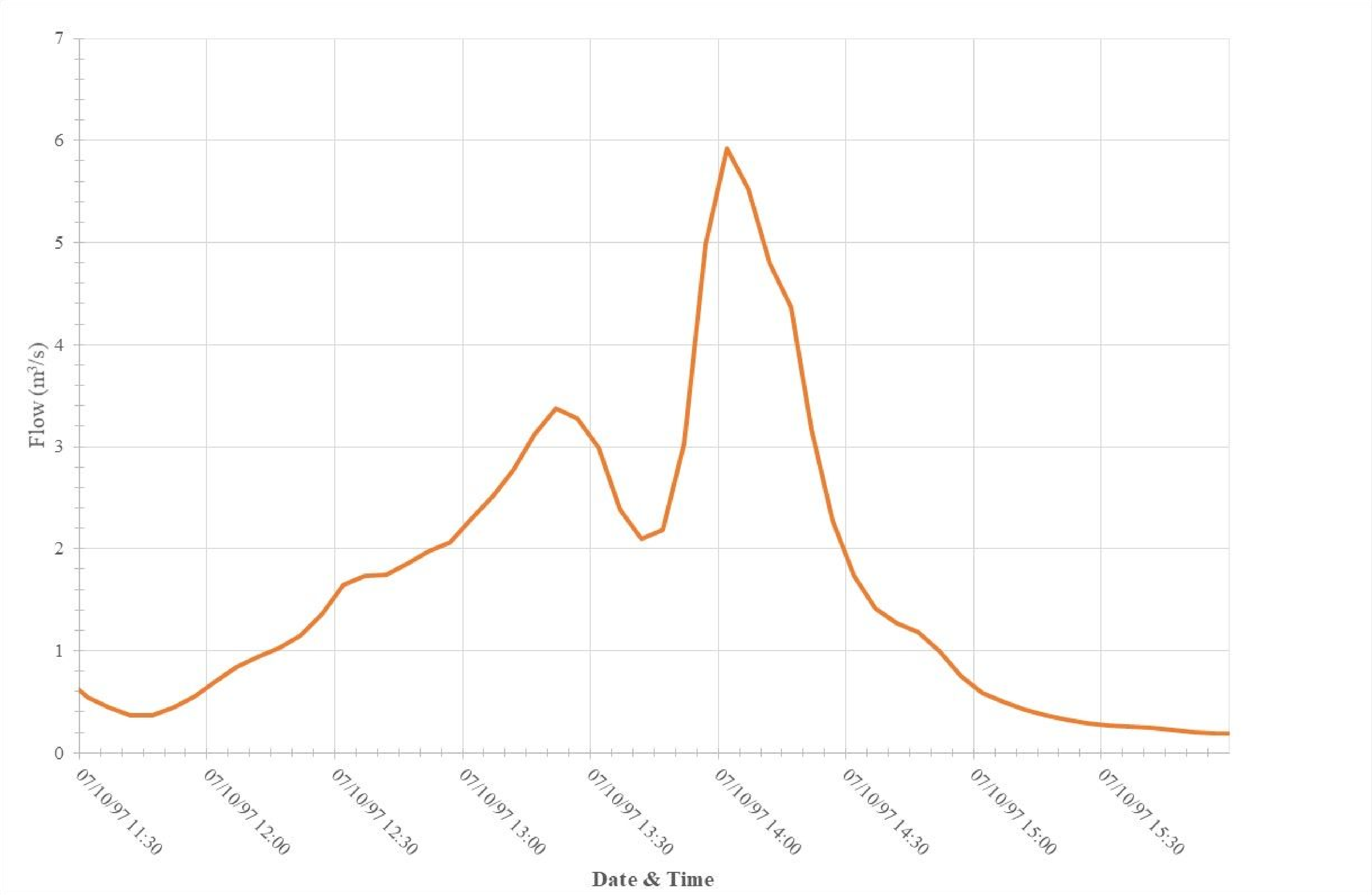
LIST OF FIGURES

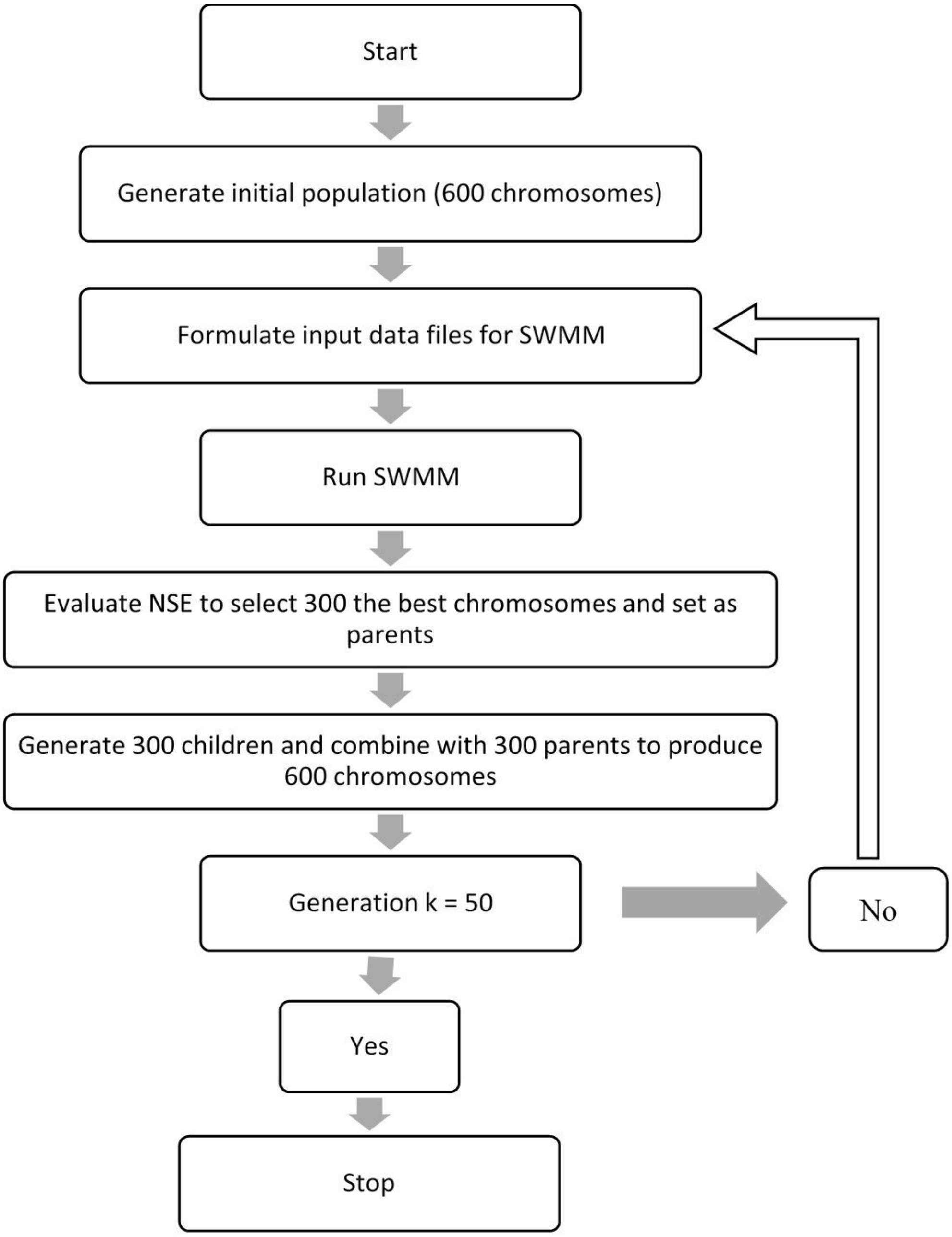


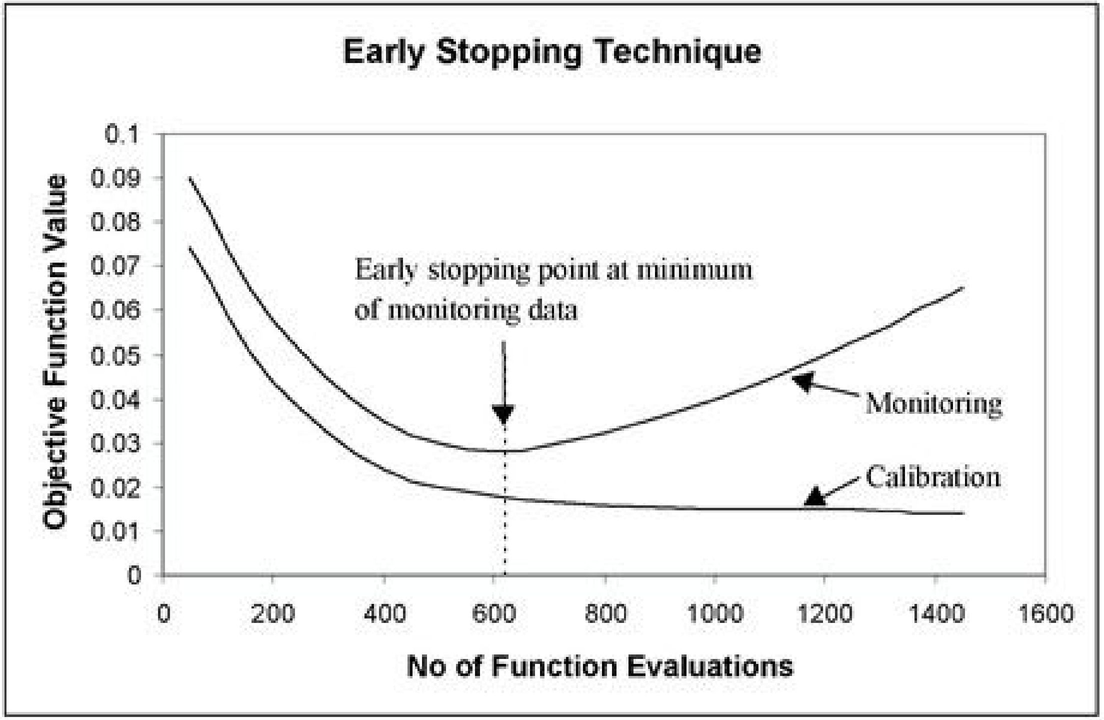
1307x925mm (96 x 96 DPI)

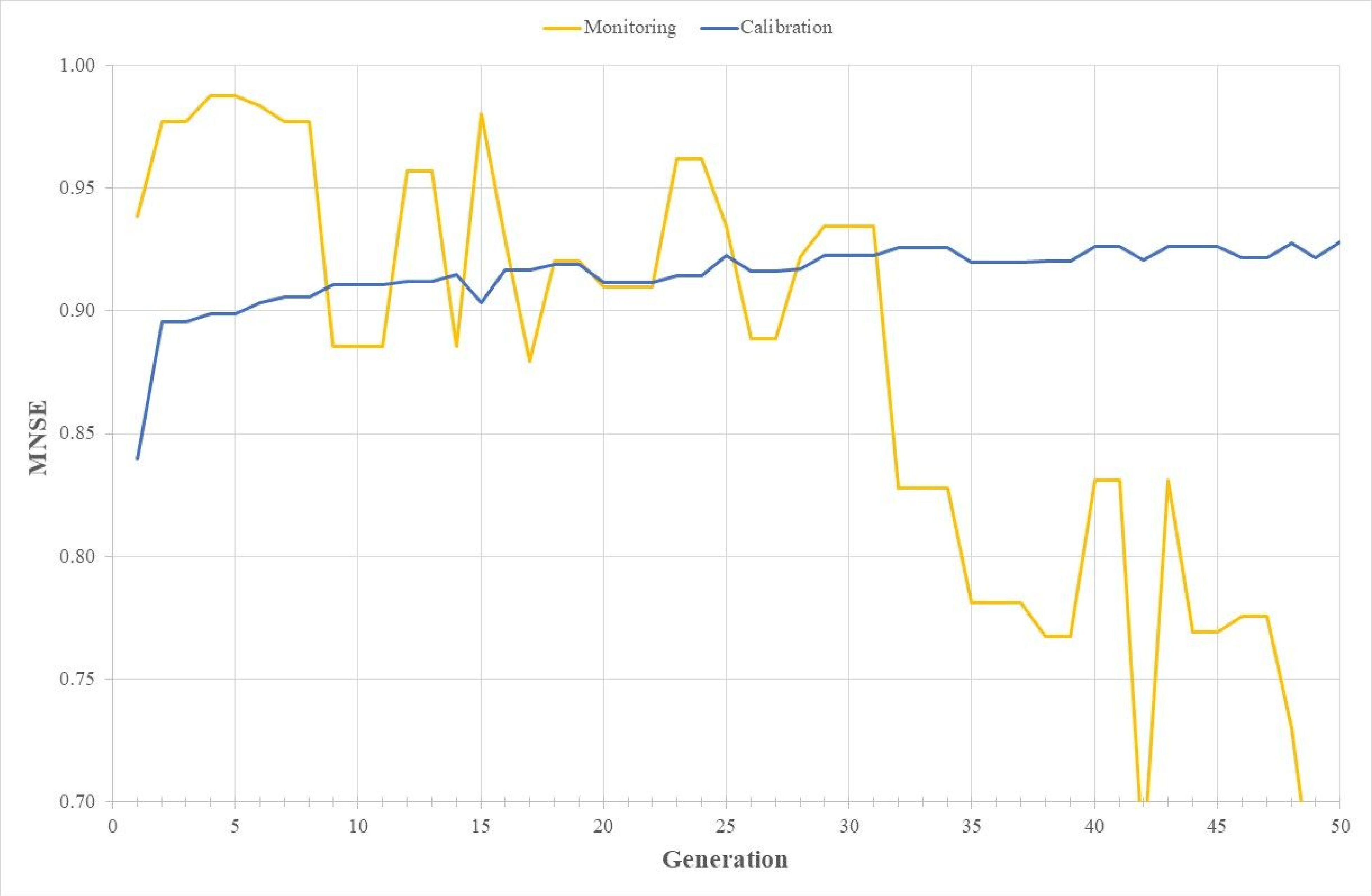

14
12
10
8
6
4
2
0
Flow (m3/s)
23/10/85 15:00
23/10/85 15:30
23/10/85 16:00
23/10/85 16:30
23/10/85 17:00
23/10/85 17:30
23/10/85 18:00
Date & Time

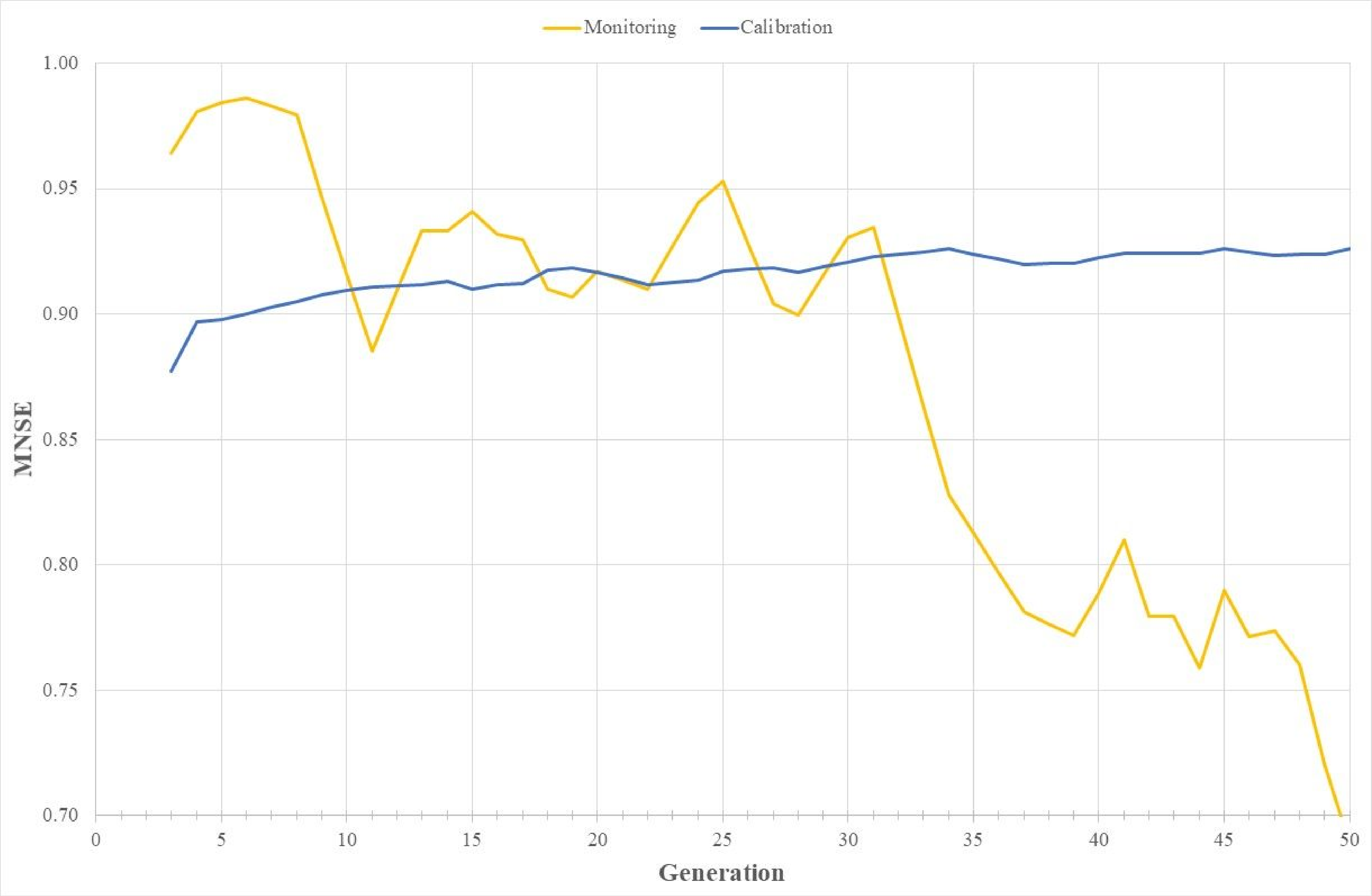

7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
Flow (m³/s)
29/06/97 10:00
29/06/97 10:30
29/06/97 11:00
29/06/97 11:30
29/06/97 12:00
Date & Time

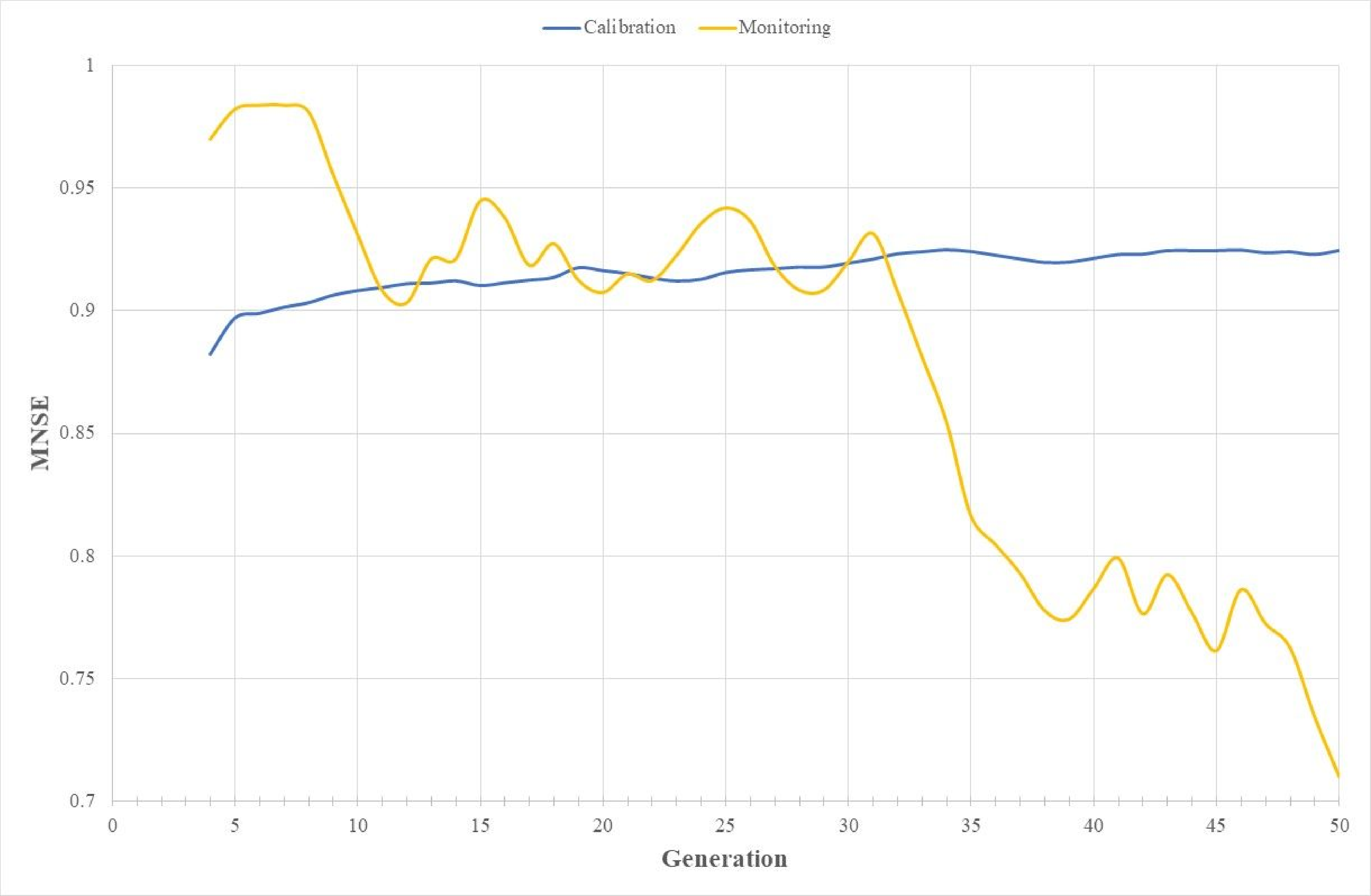

7
6
5
4
3
2
1
0
Flow (m³/s)
07/10/97 11:30
07/10/97 12:00
07/10/97 12:30
07/10/97 13:00
07/10/97 13:30
07/10/97 14:00
07/10/97 14:30
07/10/97 15:00
07/10/97 15:30
Date & Time

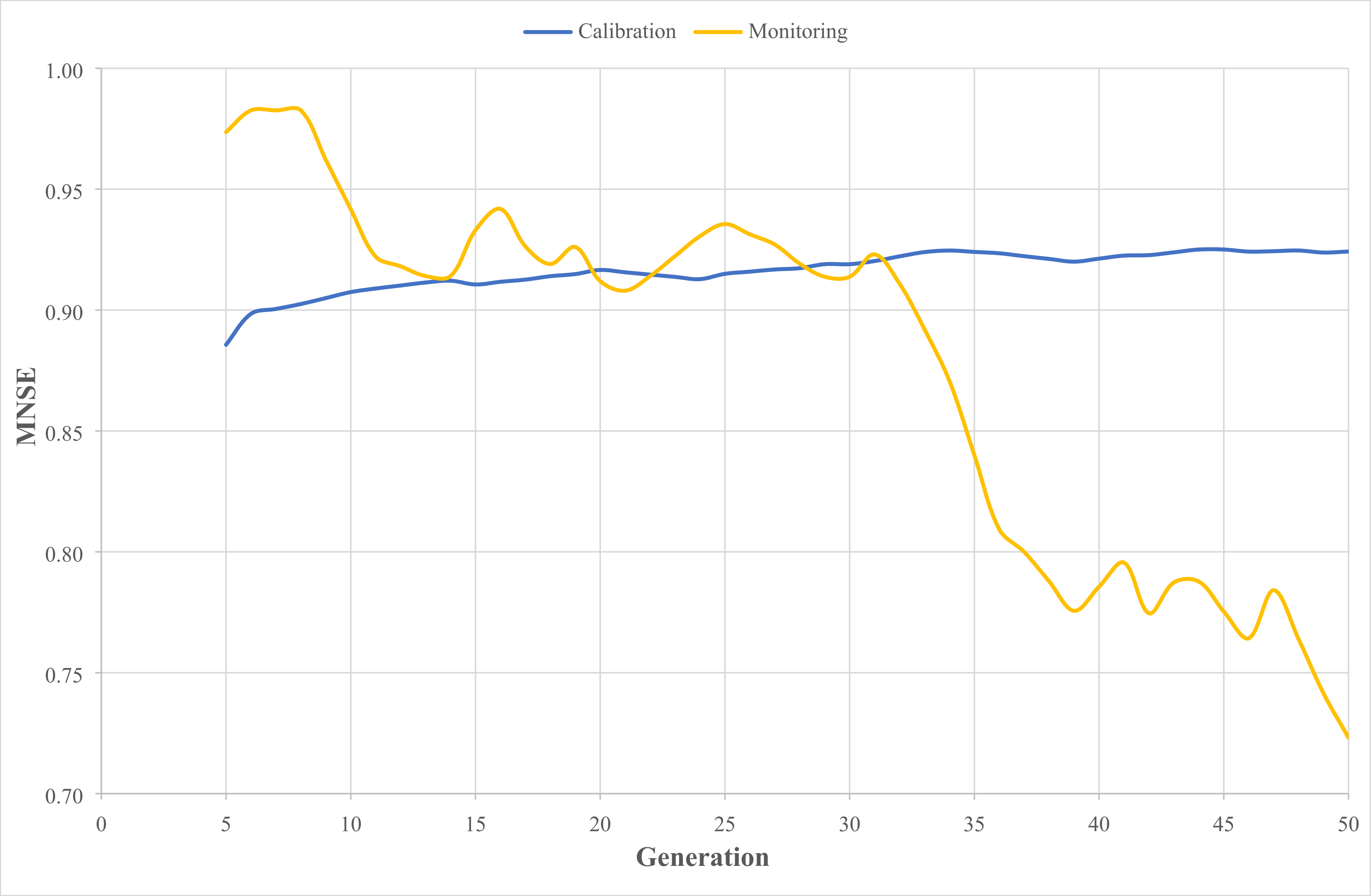

Start
Generate initial population (600 chromosomes)
Formulate input data files for SWMM
Run SWMM
Evaluate NSE to select 300 the best chromosomes and set as parents
Generate 300 children and combine with 300 parents to produce 600 chromosomes
Generation k = 50
No
Yes
Stop

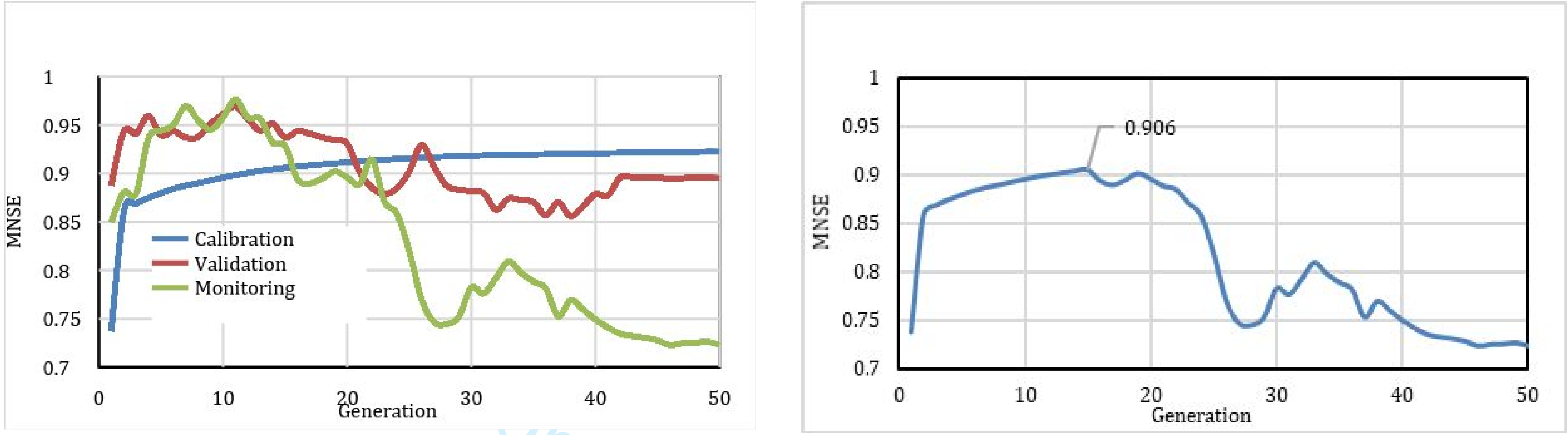

Early Stopping Technique
Objective Function Value
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
200
400
600
800
1000
1200
1400
1600
No of Function Evaluations
Early stopping point at minimum of monitoring data
Monitoring
Calibration


Monitoring
Calibration
MNSE
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0
5
10
15
20
25
30
35
40
45
50
Generation


Monitoring
Calibration
MNSE
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0
5
10
15
20
25
30
35
40
45
50
Generation

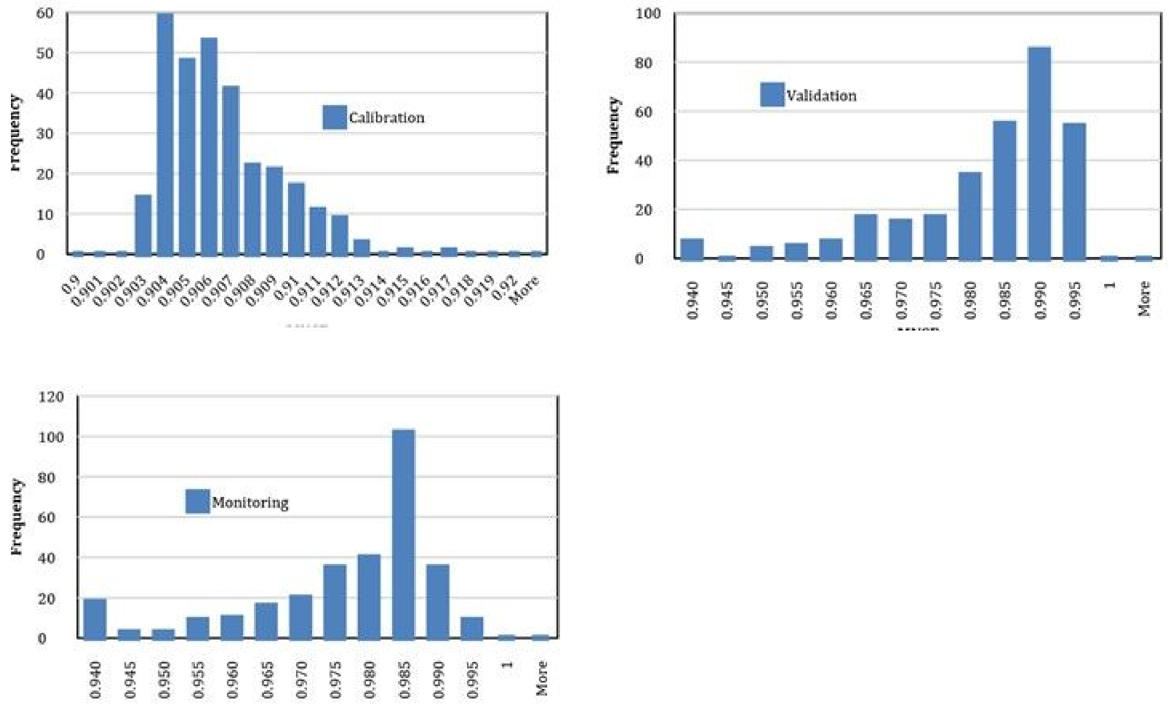

Calibration
Monitoring
MNSE
1
0.95
0.9
0.85
0.8
0.75
0.7
0
5
10
15
20
25
30
35
40
45
50
Generation

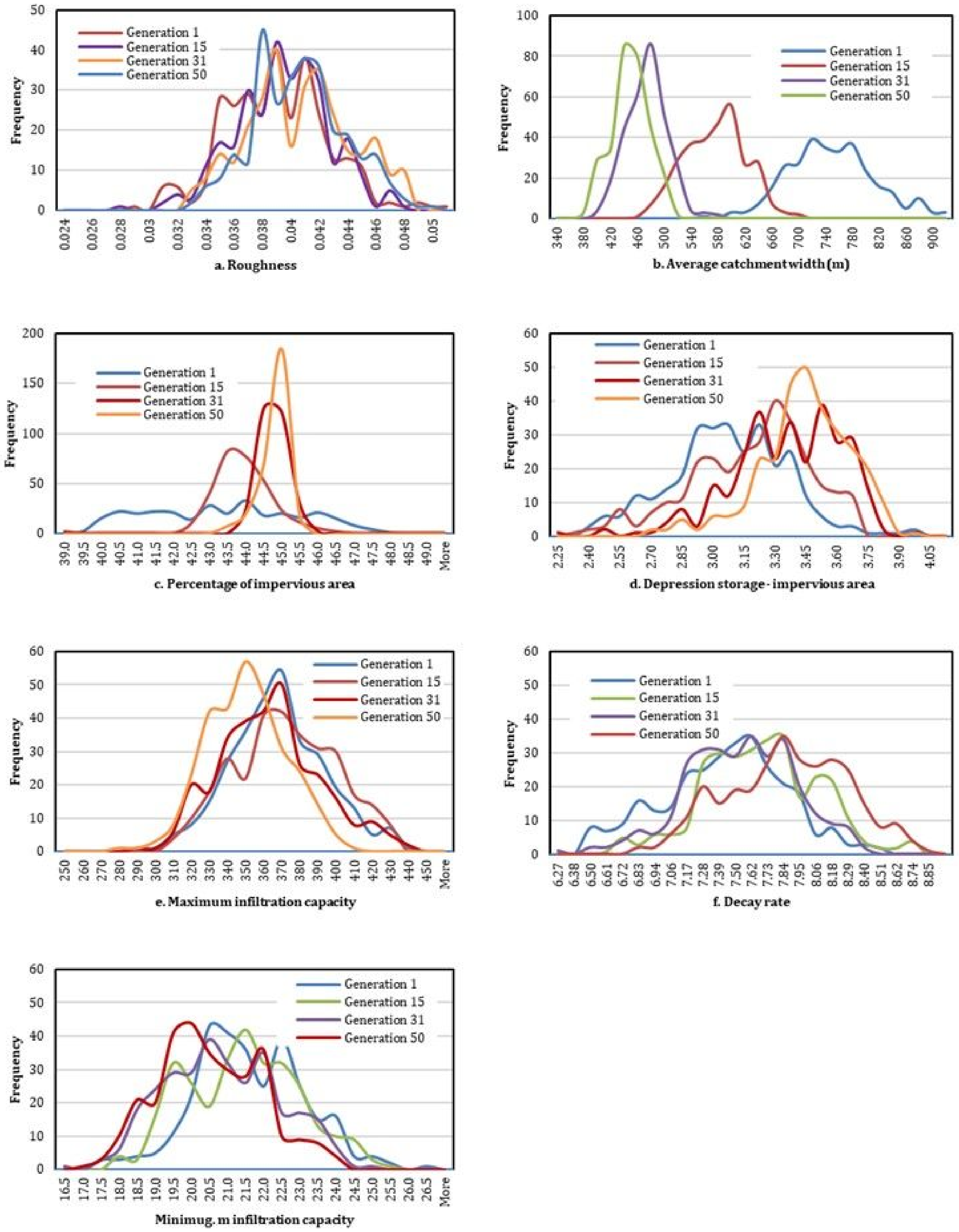

Calibration
Monitoring
1.00
0.95
0.90
0.85
0.80
0.75
0.70
MNSE
0
5
10
15
20
25
30
35
40
45
50
Generation


1
0.95
0.9
0.85
0.8
0.75
0.7
MNSE
Calibration
Validation
Monitoring
0
10
20
30
40
50
Generation
0.906

(a)

(b)


60
50
40
30
20
10
0
Frequency
Calibration
0.9
0.901
0.902
0.903
0.904
0.905
0.906
0.907
0.908
0.909
0.91
0.911
0.912
0.913
0.914
0.915
0.916
0.917
0.918
0.919
0.92
More
100
80
60
40
20
0
Frequency
Validation
0.940
0.945
0.950
0.955
0.960
0.965
0.970
0.975
0.980
0.985
0.990
0.995
1
More
120
100
80
60
40
20
0
Frequency
Monitoring
0.940
0.945
0.950
0.955
0.960
0.965
0.970
0.975
0.980
0.985
0.990
0.995
1
More



a. Roughness

b. Average catchment width (m)

c. Percentage of impervious area

d. Depression storage - impervious area

e. Maximum infiltration capacity

f. Decay rate

Minimug. m infiltration capacity