International Environmental Modelling and Software Society (iEMSs)
2010 International Congress on Environmental Modelling and Software
Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada
David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.)
http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings

A Novel Model Calibration Technique Through Application of Machine Learning Association Rules

Simon Hood[1] and David Swayne[2]
[1]simon.hood@gmail.com, [2]dswayne@cis.uoguelph.ca

Abstract: Recent work involving attempts to calibrate the Soil and Water Assessment Tool (SWAT) non-point source model has led to application of machine learning techniques (specifically the Apriori algorithm) to test runs of the model for particular watersheds. A new Parsimonious Explicit Apriori Reduction (PEAR) method for model calibration is evinced, and details of experiments demonstrating the improved efficiency and accuracy, as opposed to both the manual approach and a well-known Genetic Algorithm (GA), are outlined. The PEAR method overcomes difficulties intrinsic to three current classes of multi-objective optimization.

***Keywords**: PEAR, multi-objective optimization, model calibration, SWAT, SCE*

1. INTRODUCTION

Computer based models must be calibrated to match initial and ongoing conditions in order to produce successful predictions for the systems they simulate. The application of machine learning techniques to any model calibration greatly reduces the cost and time of expensive and intensive expert human effort. The proposed PEAR method of calibration through machine learning of association rules potentially overcomes several key stumbling blocks and increases the precision, timeliness, and utility over techniques already in use.

1.1 Motivation

Contemporary models require numerous input parameters be estimated or calibrated for the model to produce output that corresponds with measured data. Madsen [2003] states the input parameters are often related to summative process descriptions, and are not typically reproducible through observation of the natural system alone. Model equations are often generated through statistical correlation of data sets, with input values for correlation polynomials representing modifying factors existing solely as abstracts in relation to their derived formulae. These intangible input factors greatly influence the model output, even though they are fundamentally unknowable from the environment alone.

1.2 Current Calibration Methods

Multi-objective automatic calibration or optimization is a method of managing the size of the solution space while attempting to calibrate multiple input values simultaneously according to Deb et al. [2002], and three approach classes have arisen which reduce the human time and effort, each with varying weaknesses. The three main methods available today are still lacking, however. GAs typically produce only single solutions despite extensive calculation, and must overcome the problem of local optima. Probability-based methods do not necessarily produce well-defined results, and there is a danger of

oversampling the system producing spurious phenomena. Multi-dimensional clustering techniques are developing but remain in their infancy, they also suffer from problems of oversampling, and they do not necessarily guarantee a workable solution with infinite amounts of computation. The PEAR method was designed in an attempt to overcome these limitations.

2. PEAR METHOD COMPONENTS

The PEAR Method is comprised of a number of existing technologies. Essentially, PEAR applies Probably Approximately Correct (PAC) learning with the Apriori Algorithm. In order to successfully calibrate a model, however, two additional steps must be taken: the input and output data must be roughened, and the results must be grouped. When properly executed, the PEAR method is capable of accurately and efficiently calibrating models where parameter sets form multi-objective functions that must be optimized.

2.1 PAC Learning

PAC machine learning reduces the enormous potential size of the solution sample space. Concepts from decision theory and statistical pattern recognition were combined, with simple computational complexity being the overarching goal. Haussler [1990] built upon a polynomial method by Valiant [1984] by defining a formula asserting the minimum number of training examples required to achieve a specific probability a hypothesis is correct, within a margin of error, given an intractably large sample space, hence, probably approximately correct.

Minimum sample size required to learn statistically valid rules from large systems is shown in (1). The minimal sample size m required for proof of a hypothesis with δ accuracy, given a margin of error ε, and a total solution space of size H_n is provided. In applying the formula to large calibration solution spaces, it can be shown that the number of training examples necessary to prove a generated rule or hypothesis is approximately true is miniscule when compared to the total number of possibilities.

$$m \geq \frac{1}{\varepsilon}\left(\ln(H_n) + \ln\left(\frac{1}{\delta}\right)\right)$$

Equation 1: Minimum sample size by Russel and Norvig [1995]

There are two main criticisms of PAC machine learning: the worst case emphasis in the results, and the requirement of noise-free training data. According to Pazzani and Sarrett [1992b], PAC learning overestimates the hypothesis error as a function of the actual data distribution, given the training set size. They argue that if (1) is solved for ε, a curve can be produced for specific H_n and δ values that will be overly pessimistic. In the PEAR method, an overestimation of the statistical error involved lends further credibility to the results. Pazzani and Sarret also note that error bounds over the minimum training set size for PAC learning can be improved. For the purposes of the PEAR method, we prefer to err on the side of caution and include a small number of potentially redundant examples in the training set. Second, it can be argued that noise-free training sets and well-defined target concepts are unrealistic. Bergadano et al. [1988b] provide a PAC framework compensating for noisy environments by using both hypothesis examples and counterexamples, which the PEAR method mirrors through optimization of both the highest and lowest scoring results. Angluin and Laird [1988a] also argue that reliable rules can be generated through PAC learning from noisy data by selecting the most consistent rules, or the rules with the highest levels of backing evidence. The Apriori algorithm defines this concept as the level of support, and the PEAR method sets minimum levels of support that must be achieved for rules to be considered valid. The first major criticism of PAC learning actually works in

favour for the PEAR method, and the results from the exploration of the second criticism are thus incorporated.

PAC learning reduces the size of the training set required to achieve calibration solutions with high precision by many orders of magnitude. A simple calibration might use nine variables, each of which can be ten possibilities. Such an example may be impractical in reality, but the use of the rangifying and grouping process in the PEAR method discussed later makes it pragmatic. The calibration therefore has a total of 10^9, or 1,000,000,000, possible solutions, and all must be evaluated to achieve 100% accuracy 100% of the time. For the PEAR method, we generally accept that hypotheses need be 99.9% accurate nine times out of ten. Table 1 shows the dramatic reduction in sample space size possible by slightly reducing the accuracy or error level in the logarithmic function.

REDUCED SAMPLE SIZE FOR 1,000,000,000 POSSIBLE SOLUTIONS			
	5% Error	**1% Error**	**0.1% Error**
95% Accuracy	475 examples	2372 examples	23719 examples
99% Accuracy	507 examples	2533 examples	25329 examples
99.9% Accuracy	553 examples	2763 examples	27631 examples

Table 1: Minimum examples required for 1,000,000,000 possibilities

Given that a quick model by contemporary standards might take 5 minutes to execute per run, the amount of execution time is reduced from 3,802 years to just 23 hours on a single processor, without significant compromise on the correctness of the results, and parallel processing can reduce the time required from 23 hours to mere minutes.

2.2 The Apriori Algorithm

The Apriori algorithm was coined in a highly cited paper by Agrawal et al. [1993] in an effort to present an efficient method for generating association rules between items in a database. It is designed to operate on databases containing transactions, e.g. databases containing purchased items or website visitation; in other words, sets of records where a clear result is present. Apriori attempts to find subsets common among a given set of database records or *training set* using a breadth-first search tree structure. Frequent subsets are extended using a bottom-up approach, and are tested against the complete training set repeatedly to verify whether they are still valid rule candidates. Essentially, the algorithm tries to expand each found subset recursively with all remaining criteria in turn. If a new subset is still true, given the minimum support and confidence levels, it is added as a new association rule, and the search continues until there are no other options.

$$CoE = 1.0 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2}$$

Equation 2: Nash-Sutcliffe Coefficient of Efficiency

The Apriori algorithm requires transactional databases for operation, or that the parameters involved produce a single result, therefore a fitness score is included for each model run using the PEAR method. Essentially, any measure of the goodness-of-fit of the calibration output data will suffice according, but we have found empirically that the PEAR method performs best using the Nash-Sutcliffe Coefficient of Efficiency (CoE) pairwise comparison, shown in (2), of observed output data (O) against calculated or perceived output data (P).

Before each model run, criteria being evaluated are randomly generated within reasonable ranges and stored as individual items or columns in a single record of a database. After the

model run, a score measuring the effectiveness of the parameters in producing output that corresponds with observation is added as a final column to the same database record. Repeating the process builds a relational table where different random sets of input criteria result in higher or lower scores, depending on the actual data used. With the item sets in transactional form, the Apriori algorithm can be used to mine the database and look for patterns.

The minimum support and confidence levels set are critical to the Apriori algorithm's operation in the PEAR method. Support refers to the ratio of subsets where a rule is true compared to the total number of item sets, and confidence refers to the ratio of subsets containing the rule that are true compared to the total item sets containing the subset. The support value should be adjusted so that it reflects an appropriate number of true cases among the data for a rule to be considered reasonable. Likewise, the confidence level should be positioned so that rules are at least reasonably accurate given the database.

NUMBER OF RULES CREATED IN THE RAISIN RIVER WATERSHED FOR VARIOUS SUPPORT AND CONFIDENCE LEVELS			
	60% confidence	**80% confidence**	**99% confidence**
2 examples	8675 rules	5234 rules	5194 rules
3 examples	819 rules	169 rules	129 rules
4 examples	112 rules	50 rules	10 rules

Table 2: Rules created with various support and confidence

In Table 2, the balance between high and low confidence and support levels is illustrated by the number of rules generated. If either the support or confidence is set to too high a value, obvious rules may be overlooked, but setting the values too low results in noise, or biased sampling, among the rule set. The PEAR method experiments in this paper use a support level requiring approximately two examples are present among the data for a rule to be considered valid and a confidence level above the probability of a simple coin flip at 60%. The optimum levels may be highly relative however, and may require some tweaking by practitioners of the PEAR calibration method. Work by Scheffer [2005] suggests methods for selecting only the most probable rules. We recommend setting the levels as low as is reasonable to generate the largest number of workable rules possible for the next step in the process, grouping.

The most important relationships found by the Apriori algorithm in the PEAR method are ones in which a rule results in either the highest possible score, or the lowest possible score. If the values for a set of criteria can be shown to produce a high measure of fitness in the output, they are strong candidates for the best calibration values. Conversely, if values for criteria are shown to consistently produce low fitness scores, they should be avoided. In this way, the PEAR method utilizes the Apriori algorithm to mine association rules of both superlative and slipshod calibration values.

2.3 Roughening and Grouping

The Apriori algorithm considers similar potential calibration values as separate subsets, and assumes each is capable of generating independent rules, rendering it incapable of successfully mining the model execution database. As far as Apriori is concerned, remembering the Latin definition of a priori as knowledge requiring no experience, the values 1.0 and 1.00 are separate entities. In order to successfully mine calibration values with the PEAR method, the parameter and score data must be roughened and grouped into concise blocks so that meaningful associations can be made. Ranges are dynamically calculated by dividing the data set into approximately equally sized divisions. With the data properly rangified, the Apriori algorithm can be run.

After the Apriori algorithm is run and all the association rules are mined, the rules leading to the highest transaction score range and the lowest transaction score range are grouped. A

typical iteration of the PEAR method, continuing with the example, might produce 8675 rules in total, of which 161 lead to the highest fitness score range, and 153 lead to the lowest transaction score range. In effect, the parameter range subsets that produce the best and worst calibration candidates are stripped out of the total rule set. Part of a typical rule set might read as follows in Table 3.

The first rule states that a combination of the SWAT parameters SMFMX between 6.91 to 7.81, SFTMP between -0.66 to -0.06, and TIMP between 0.42 to 0.53 results in a fitness score between 0.656 and 1, or the highest goal range. All three parameters (SMFMX, SFTMP, and TIMP) must be present within the ranges specified for the goal (in this case, a CoE between 0.656 and 1) to be produced. Individual rules can contain two, three, four, or more parameters at a time, though the rules shown below all contain three. The rules are then grouped; the SMFMX range 6.91 to 7.81 is present in two of the three rules, so is a promising candidate for an SMFMX range that will lead to a successful calibration.

	Criteria	Goal	Support	Confidence
1	SMFMX : 6.91 - 7.81, SFTMP : -0.66 - -0.06, TIMP : 0.42 – 0.53	CoE : 0.656 - 1	0.006	1
2	SURLAG : 2.98 - 4.08, SMFMX : 4.01 - 5.11, SMFMN : 0.96 - 1.51	CoE : 0.656 - 1	0.004	1
3	SURLAG : 8.59 - 9.69, SMFMX : 6.91 - 7.81, TIMP : 0.42 – 0.53	CoE : 0.656 - 1	0.004	0.667

Table 3: Typical rule set

2.4 The PEAR Method Step-by-Step

1. Calculate the required training set size based on the accuracy and number of parameters desired following (1).
2. Execute the model using random parameters within reasonable limits at least the required number of times (see step 1) and calculate a fitness function for each.
3. Dynamically roughen the data.
4. Execute the Apriori algorithm on the roughened data.
5. Group the results of the Apriori algorithm.
6. Determine the new optimal ranges for each parameter in the calibration.
7. Repeat as necessary using the new ranges until no further reduction is practical.

3. EXPERIMENT AND OBSERVATIONS

The experiment detailed below illustrates the hypothesis the PEAR method is both more effective and more efficient at computer model calibration than either the standard manual and common automatic Shuffled Complex Evolution (SCE) approaches. Calibrations were performed on two separate watersheds, Raisin River and Fairchild Creek, using the SWAT model with all three methods, to ensure a fair comparison is made. The PEAR method was then contrasted against the other two calibrations, and the amount of time, number of model executions, and overall accuracy of the results was evaluated.

3.1 Experimental Procedure

Two sets of data for a computer model were calibrated using the manual method, the SCE method, and the PEAR method independently, and the results were analyzed to reinforce or disprove the hypothesis presented. It should be known that each calibration was performed independently, and any domain knowledge gained from previous calibrations was discarded in the next attempt. As such, the manual calibrations were performed first to avoid potential bias. It was felt a comparison of the time required for a manual calibration

provided the most effective measure of efficiency, while the total number of runs was more appropriate for SCE, because the two approaches differ greatly. A manual calibration might only require 50 model executions but take months, while an SCE calibration might require mere hours but take thousands of model runs.

Watershed data for the Raisin River and Fairchild Creek watersheds in eastern and central Ontario, Canada respectively were compared against the SWAT model. The immense size and complexity of the SWAT model has made it the de-facto standard used in calibration experiments; many papers today compare the accuracy of new calibration techniques against the SWAT model using a format developed by Eckhardt and Arnold [2001]. The performance of a technique is usually calculated against the SCE algorithm based on Yapo et al. [1996] because it was proven to be the most effective method in common use at the time by Duan et al. [1992a]. The manual calibrations were performed independently by two separate experts, while the SCE and PEAR calibrations were produced without deep knowledge of the SWAT model or the watersheds.

3.2 Manual Calibration vs. PEAR

In both the Raisin River and Fairchild Creek watersheds, the amount of time required for a manual calibration greatly exceeds that required using the PEAR method. The manual calibration performed by McCrimmon [2010b] of Environment Canada on the Raisin River watershed required roughly 2,880 minutes assuming eight hour work days. The PEAR method calibrated the same parameters on the Raisin River watershed in about an hour. The calibration of the Fairchild Creek watershed required approximately one work day using a combination of manual and assisted methods by Liu [2010a], but also approximately one hour using the PEAR method. Liu and McCrimmon are considered world-class experts on SWAT, so their calibration times are relatively quick, but Liu modestly stated an untrained calibrator may require months to do the same work. Liu also mentioned he made use of SWAT's internal auto-calibration scheme, but did not use any GA for his work. A comparison of the amount of time required to run both the manual and PEAR methods can be found in Figure 1.

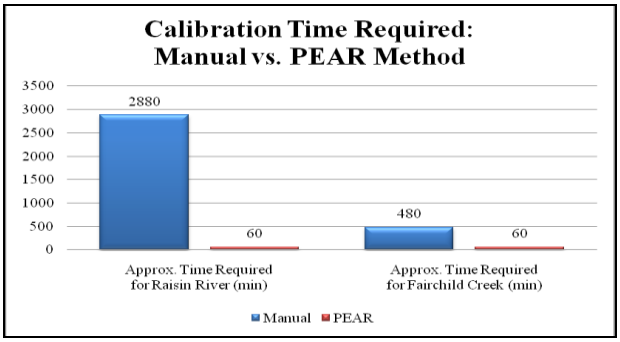

Figure 1: Calibration time required

3.2 SCE Calibration vs. PEAR

The SCE method calibrations of both the Raisin River and Fairchild Creek watersheds terminated themselves after the specified maximum of 2000 iterations. In contrast, the PEAR method used only 1,273 model executions on the Raisin River watershed and 1,497 on the Fairchild Creek watershed. The nature of the SCE algorithm implies that a better answer might be found with more model runs, but SWAT's upper limit and SCE's only occasionally random evolution process does not make this a certainty. Figure 2 shows the PEAR method required fewer model runs for a successful SWAT calibration, and was, in fact, still more accurate than SCE.

The PEAR method required about an hour from start to finish including runtime, while the SCE method required several hours of computation. Both methods were essentially run in serial, although a Serial Farm technique was employed in the PEAR method. The number

of model runs necessary for the PEAR method on the watersheds could also be made considerably smaller if maximal parameter accuracy levels were set. It is often unnecessary to calculate parameters to four or five decimal places of accuracy given many models, SWAT in particular, cannot distinguish between parameters beyond a certain precision.

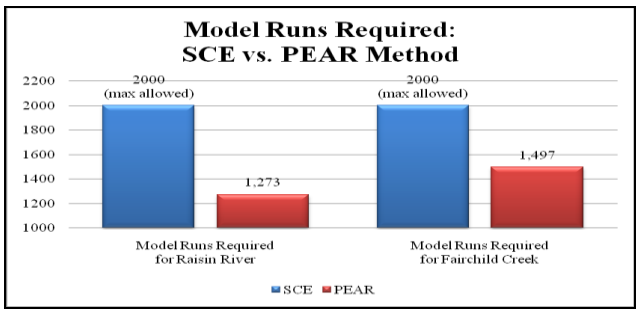

Figure 2: Model runs required

3.4 Accuracy Comparison

The PEAR method was more accurate than both the manual and SCE calibrations in both the Raisin River and Fairchild Creek watersheds. The higher the CoE score, the closer the pairwise comparison of average monthly calculated flow values to actual monthly average flow values. Moreover, although the measured data may be accurate enough according to the model, calculating flow levels that are more precise potentially results in a worse CoE score on the same data; small errors from increased precision factor into the pairwise comparison and reduce the overall score. Therefore, it is probable that future calibrations with the PEAR method will require fewer model runs and produce even higher CoE scores than those of the experimental results shown in Figure 3.

3.5 Observations

Calibration through the PEAR method overcomes many of the difficulties associated with the three modern methods of calibration: GA, probability-based, and multi-dimensional clustering. The PEAR method generates ranges with precisely the desired level of statistical accuracy, so is better able to cope with unknown automatic pattern recognition. Potential bias is eliminated because the PEAR method provides a mechanism in the form of the logarithmic reduction formula for calculating exactly how many examples should be included in the training set for any desired certainty with any level of error, assuming stationarity holds; the sample size will always be smaller than the whole. Finally, the PEAR method will always produce a solution in the form of a parameter range, even if that range is not far removed from the initial conditions.

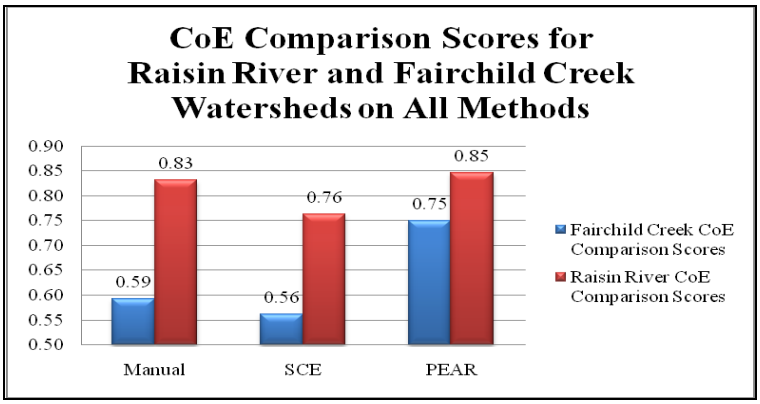

Figure 3: CoE comparison scores for all methods on both watersheds

The PEAR method does not generally improve upon the best score that can be generated through simple repeated random parameter selection, despite the fact that the range for each parameter is narrowed through successive iteration. Instead of the upper end, or highest CoE scores, increasing, it is the bottom end of the CoE score range that is affected. When used properly, a high CoE score based on the ranges selected is assured, and the possibility a random set of parameters will produce an ineffectual calibration is eliminated. For example, an additional 500 runs on the Fairchild Creek watershed using the final parameter ranges generated an average CoE of 0.73, a minimum score of 0.69, and a standard deviation of 0.0075. Similar results were seen in the Raisin River watershed. Therefore, it is a safe assumption the best comparison that can be produced by chance is only slightly improved by the PEAR method, but the method guarantees that the vast majority of output will produce scores which are equal, or at least similar, to the highest score.

4. CONCLUSIONS

The combination of technologies associated with the PEAR method represent new and original thinking, and have come to fruition in the form of a calibration technique that is quick and exact. The Apriori algorithm is well-suited for dealing with the lack of general knowledge machine learning routines must face. The PAC logarithmic reduction formula reduces the size of the sample space by vast margins and provides precise statistical accuracy in the process. Last, the rangifying and grouping procedures help the PEAR method produce workable solutions with relatively little cost. These technologies come together in the PEAR method, resulting in an improved calibration technique which is more effective and efficient than manual calibration and the standard automatic approach.

REFERENCES

Agrawal, R., T. Imieliński and A. Swami. Mining association rules between sets of items in large databases, *ACM SIGMOD Record, 22*(2), 207-216, 1993.

Angluin, D. and P. Laird. Learning from noisy examples, *Mach.Learning, 2*(4), 343-370, 1988a.

Bergadano, F., A. Giordana and L. Saitta. Automated concept acquisition in noisy environments, *IEEE Trans.Pattern Anal.Mach.Intell., 10*(4), 555-578, 1988b.

Deb, K., A. Pratap, S. Agarwal and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II, *Evolutionary Computation, IEEE Transactions on, 6*(2), 182-197, 2002.

Duan, Q., S. Sorooshian and V. Gupta. Effective and Efficient Global Optimization for Conceptual Rainfall-Runoff Models, *Water Resour.Res., 28*(4), 1015-1031, 1992a.

Eckhardt, K. and J.G. Arnold. Automatic calibration of a distributed catchment model, *Journal of Hydrology, 251*(1-2), 103-109, 2001.

Haussler, D. Probably approximately correct learning, 1101-1108, 1990.

Liu, Y. Re: Fairchild Calibration Stats, *Email,* 2010a.

Madsen, H. Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives, *Adv.Water Resour., 26*(2), 205-216, 2003.

McCrimmon, C. Re: Raisin River SWAT Calibration Statistics, *Email,* 2010b.

Pazzani, M.J. and W. Sarrett. A framework for average case analysis of conjunctive learning algorithms, *Mach.Learning, 9*(4), 349-372, 1992b.

Russell, S. and P. Norvig. Artificial Intelligence: A Modern Approach, *New Jersey,* 1995.

Scheffer, T. Finding association rules that trade support optimally against confidence, *Intelligent Data Analysis, 9*(4), 381-395, 2005.

Valiant, L.G. A theory of the learnable, *Commun ACM, 27*(11), 1142, 1984.

Yapo, P.O., H.V. Gupta and S. Sorooshian. Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data, *Journal of Hydrology, 181*(1-4), 23-48, 1996.