CHAPTER 4

Simulated annealing algorithms for water systems optimization

M.C. Cunha & A. Antunes
Department of Civil Engineering, Coimbra University, Portugal.

Abstract

Simulated annealing is a heuristic method based on the analogy with the physical annealing process. In the past 20 years, this method has been used to solve many complex optimization problems, and it has become known for its ability to produce good results in reasonable time. This chapter starts with an explanation of why heuristic methods like simulated annealing are needed to address water systems development problems. The main issues involved in the implementation of this method are examined next, followed by their discussion with respect to application to water systems development problems.

Keywords: optimization, simulated annealing, water systems development.

1 Introduction

Water system development problems usually involve the placement and operation of infrastructure. To find the optimal solution of such problems, when all the costs are considered, is a very difficult task. In fact, non-linear models that include discrete variables have to be solved. This is why such problems are often simplified either by ignoring fixed costs or by linearizing the cost functions.

A variety of optimization models have been developed for the planning and management of water systems, ranging from those relating to the optimal operation of existing infrastructure (where there are no location and construction costs to be considered) to those aiming at planning the exploitation of unused systems (where there are location, construction, and operating costs to be considered). From the point of view of their mathematical characteristics, it can be said that these problems are very different in nature, which have consequences for the degree of complexity of their resolution.
The first type usually involves linear and nonlinear functions containing continuous variables for the flows to be operated, physical management, etc. The literature shows that they can be solved through the use of the classical programming methods, even for large-scale real-world case studies.

The second type usually involves an objective function expressing cost minimization of infrastructure placement and operation. The solution of such problems implies determining the values of decision variables with respect to operating rates and the values of discrete decision variables relating to the number and the location of hydraulic infrastructure, while meeting some constraints. Therefore, the problems are mixed-integer non-linear optimization problems (MINLP), that are among the most difficult optimization problems to solve. There are many engineering fields dealing with this type of problem, which is why the literature reports a variety of attempts to solve them. The corresponding models were often simplified by disregarding the fixed costs incurred by installing the infrastructure and, as such, the discontinuous variables are not considered, resulting in the models becoming standard non-linear programs (NLP). Another way of converting these problems into an NLP is to preselect the type and the number of items of equipment and their location. This can be done in an iterative fashion, updating the type, number, and the location of the equipment items in each iteration, but with this kind of procedure, there will be many doubts about the quality of the solution (even near-optimal solutions may be difficult to obtain). In other cases, cost functions are linearized and they become standard mixed-integer linear problems (MILP).

In the last 25 years, developments in the field of heuristic optimization algorithms have made it possible to solve real-world problems involving planning and management issues without considering the aforementioned simplifications.

Simulated annealing is one of the best known heuristic optimization algorithms that, due to its conception, produce high-quality results. Simulated annealing is a random search based algorithm that allows asymptotic convergence to optimal solutions under mild conditions [1–3]. In the next sections, some theoretical bases and practical aspects concerning its application to water systems development are presented.

2 Theoretical aspects

Simulated annealing is a local search method that theoretically allows the identification of a global optimum of a function. This property is related to the fact that during the search procedure, mainly in the first phase, the algorithm accepts worse solutions. The first example of an annealing algorithm was presented by Metropolis for a thermodynamic problem regarding the evolution of the energetic state of a piece of glass subjected to a high temperature that is progressively decreased. Since then and after the pioneering work by Kirkpatrick et al. [4], many applications of annealing algorithms have been developed.

The theoretical bases of annealing algorithms are summed up below.

Consider a system $S$ represented by the set $X = \{1, \ldots, n\}$ of admissible configurations and $P$ a problem representing the minimization of a function $e = e(i) : X \rightarrow R$, that is:
According to the thermodynamic terminology, the function $e = e(i)$ is called ‘system energy’.

Consider $\pi$ to be a probability distribution defined through $X$ in terms of a function $f: R^2 \rightarrow R$ and a parameter $\tau \in R^+$ with general form

$$
\pi(\tau) = [\pi(1, \tau), \ldots, \pi(n, \tau)] = \left[\begin{array}{c}
f[e(1), \tau] \\
\sum_i [e(i, \tau)] \\
\vdots \\
\sum_i [e(i, \tau)]
\end{array}\right].
$$

According to the thermodynamic terminology, the parameter $\tau$ is called ‘system temperature’.

The expected value of the system energy at temperature $\tau$ will be $\langle e \rangle_\tau = \sum e(i) \pi(i, \tau)$.

It is possible to demonstrate the existence of distributions $\pi(\tau)$ for which $\lim_{\tau \to \infty} \langle e \rangle_\tau = e^*$ [5]. This is the case of distributions where $f[e(i), \tau] = \exp[-g(\tau) \cdot e(i)]$ if $g(\tau)$ meets the following assumptions: $\lim_{\tau \to 0} g(\tau) = +\infty$ and $g'(\tau) < 0$.

Among these distributions, there is the distribution $g(\tau) = a \cdot \tau^{-b}$ and, when $a = \frac{1}{k}$ and $b = 1$ (where $k$ is the Boltzmann constant), the Boltzmann–Gibbs distribution occurs.

To obtain a distribution with these characteristics, it is necessary, in principle, to use a Monte Carlo method of the type proposed by Metropolis because it is usually difficult to establish such distribution by analytical means. This method consists of associating the configuration space and a Markov chain having this distribution as equilibrium distribution in the following manner.

Let $V = [v_{ij}]$ be the matrix containing the neighboring configurations of the system, making it probable that a direct move (move with only one step) from configuration $i$ to configuration $j$ could occur, no matter what the configurations are; $A$ be a Markov chain defined over the system’s configurations and $A = [a_{ij}]$ be a transition matrix of this chain, giving the probability of the direct transition from configuration $i$ to configuration $j$, defined by

$$
a_{ij} = v_{ij} \min\left[1, \frac{\pi(j, \tau)}{\pi(i, \tau)}\right] \iff i \neq 0
$$

$$
a_{ii} = 1 - \sum_{i \neq j} a_{ij}.
$$

Considering that configurations are organized by the inverse order of the respective energy levels (i.e. $i(j \iff e(i)e(j))$, it is possible to prove that if the matrix $V$ is symmetric and the chain $A$ is irreducible, then chain $A$ will have unique equilibrium distribution [6]:

$$
\pi(\tau) = \pi(1, \tau) \left[\begin{array}{c}
\frac{\pi(2, \tau)}{\pi(1, \tau)} \\
\vdots \\
\frac{\pi(n, \tau)}{\pi(1, \tau)}
\end{array}\right].
$$
In particular, when \( f[e(i), \tau] = \exp[-e(i)/\tau] \), one will have

\[
\pi(\tau) = \pi(1, \tau) \left\{ 1, \exp\left[\frac{-\Delta e(2)}{\tau}\right], \ldots, \exp\left[\frac{-\Delta e(n)}{\tau}\right] \right\},
\]

where

\[
\pi(1, \tau) = \frac{1}{\sum_i \exp\left[\frac{-\Delta e(i)}{\tau}\right]},
\]

\[
\Delta e(i) = e(i) - e(1).
\]

This distribution is designated hereafter as reference distribution.

It follows that

\[
\pi(\tau) = \left(\frac{1}{n^*}, \ldots, \frac{1}{n^*}, 0, \ldots, 0\right) \iff \tau \approx 0,
\]

where \( n^* \) is the number of global minima, and

\[
\pi(\tau) = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right) \iff \tau = +\infty.
\]

For the reference distribution at equilibrium, one will have

\[
\text{prob} \left[ e(i) - e^* \leq \partial \right] \geq 1 - (n - 1) \exp\left(-\frac{\partial}{\tau}\right), \partial > 0.
\]

If a random draw sequence of transitions between configurations is performed and these transitions are accepted according to the respective probabilities, it is quite likely that, after a large number of random draws, the system will attain the minimum energy configuration (if this procedure is carried out at very low temperature, the minimum energy configuration will be attained almost certainly).

In formal terms, if the sequence of random draws is represented by \( \{S_j\}_{j=1,\infty} \), it is possible to write

\[
\lim_{j \to \infty} \text{prob} \left\{ S_j = i \mid \tau \right\} = \pi(i, \tau), \forall i \in X,
\]

\[
\lim_{\tau \to 0} \left( \lim_{j \to \infty} \text{prob} \left\{ S_j = i \mid \tau \right\} \right) = \pi(i, 0), \forall i \in X.
\]

Regarding the speed of convergence to reach equilibrium, it is known that the convergence rate is geometric and that the convergence factor is equal to the second eigenvalue of the transition matrix \([6]\). It is also known that for low temperatures and for problems with many local minima, the eigenvalue is close to one (this means a slow rate of convergence). This implies that it is useful to carry out the random draw procedure while decreasing the temperature of the system.
Following the thermodynamic analogy, the sequence \( \{ \tau_j \}_{j=1}^{\infty} \) with \( \tau_j > \tau_{j+1}, \forall j \) and \( \lim_{j \to \infty} \tau_j = 0 \), which represents the temperature levels during the random draw, is called cooling schedule. This schedule should be chosen such that the convergence of the procedure toward the equilibrium is not compromised.

It is possible to prove that if \( V \) is symmetric and \( A \) is irreducible and if \( \pi(\tau) \) is the reference distribution [7], then

\[
\lim_{j \to \infty} \text{prob}\{S_j = i | \tau\} = 0, \forall i \notin X^m,
\]

where

\( X^m \): set of local minima of \( e = e(i) \);

\[
\lim_{j \to \infty} \text{prob}\{S_j = i | \tau\} = 0, \forall i \in X' \Leftrightarrow \sum_{j=1}^{\infty} \exp\left(-\frac{d}{\tau_j}\right) = +\infty,
\]

where

\( X' : X^m \setminus X^* \);

\( X^* \): set of global minima of \( e = e(i) \);

\( d \): difference between the local minimum and the global maximum of the energy;

\[
\lim_{j \to \infty} \text{prob}\{S_j = i | \tau\} = 0, \forall i \in X^* \Leftrightarrow \sum_{j=1}^{\infty} \exp\left(-\frac{d^*}{\tau_j}\right) = +\infty,
\]

where \( d^* = \max_{i \in X'} d_i \).

This means that, to achieve the global minimum, it is necessary (and sufficient) that \( \tau_j \) implies the series

\[
\sum_{j=1}^{\infty} \exp\left(-\frac{a}{\tau_j}\right), \forall a \geq d^* \geq \left[ \max_{i \in X} e(i) - \min_{i \in X} e(i) \right]
\]

to be divergent.

A series such as \( \sum_{j=1}^{\infty} 1 / (1 + j) \), which may be obtained using the cooling schedule \( \tau_j = a / \{ \ln(1 + j) \} \), \( j \geq 1 \), fulfills this condition. Series obtained through the use of exponential or linear cooling schedule would not fulfill these conditions.

### 3 Algorithm implementation: general aspects

To implement a simulated annealing algorithm, the solution under evaluation in each iteration (current solution) is improved by generating small random changes in an iterative way. For each new solution, the value of the objective function is evaluated and the solution is accepted or not according to a given probability (Metropolis criterion). The overall procedure for generating solutions, with
each new one being built from the previous one, gives rise to a Markov chain. The broad evaluation of the solution space enables simulated annealing to evolve without being trapped in local optima.

The simulated annealing algorithm can be built using the following steps (summarizing the principles presented in Section 2):

1. choose $s_1$ \{ $s_1$ is the initial solution and current solution for the first iteration \}
2. choose $\tau_1$ \{ $\tau_1$ is the initial temperature \}
3. choose $\tau_f$ \{ $\tau_f$ is the stopping temperature \}
4. $j \leftarrow 0$
5. repeat
6. $j \leftarrow j + 1$
7. choose at random $s'_j \in \Omega(s_j)$ \{ $\Omega(s_j)$ is the candidate set of $s_j$ \}
8. choose at random $p \in [0, 1]$ \{ $p$ is a probability \}
9. if $p \leq \min \left\{ 1, \exp \left( \frac{C(s_j) - C(s'_j)}{\tau_j} \right) \right\}$ \{ Metropolis Criterion \}
   then $s_{j+1} \leftarrow s'_j$ \{ $s_{j+1}$ future current solution; $s'_j$ candidate solution \}
   else $s_{j+1} \leftarrow s_j$
10. choose $\tau_{j+1} \leq \tau_j$
   until $\tau_{j+1} \leq \tau_f$
11. end

The general algorithm presented above can be used to solve a multitude of problems, but some choices have to be made. The implementation of the annealing algorithm to solve water system problems can follow a general procedure, as depicted in Figure 1.

Once the cost function is established, two more elements are needed to implement the annealing algorithm: the procedure for generating candidate solutions (known as rearrangements) and the schedule for decreasing temperatures (known as cooling schedule). To implement a simulated annealing algorithm, it is necessary to provide an initial solution. This solution will be randomly changed while the algorithm develops. The development of the algorithm is controlled by the parameter called temperature. The initial value of this parameter is defined, and it will decrease along the iterative procedure until reaching a predefined value. In each iteration, the solution to be analyzed, called the candidate solution, depends on the neighborhood structure employed and the construction of this neighborhood is highly dependent on the specific problem to solve.

To establish the cooling schedule, four parameters have to be considered:

- $ea$: the probability of accepting a transition from the initial solution to a candidate solution whose cost is higher than that of the starting configuration by a given percentage (this parameter is used to define the initial temperature of the annealing process and is called ‘elasticity of acceptance’). In these conditions,
one might expect that the final solution would be independent of the initial solution. For practical purposes, the initial temperature $T_1$ can be determined as a function of the parameter $ea$ through the equation $T_1 = 0.1c_0/\ln ea$ (where $c_0$ is the cost of the initial solution). This expression allows us to identify a temperature for which $ea\%$ of solutions whose cost is 10% higher than the cost of the starting configuration are accepted.

- $n_1$: the minimum number of algorithm iterations that will be performed, even without an improvement of the optimum or of the average cost of the solutions analyzed, before decreasing the temperature.
• $\gamma$: the cooling factor, rate at which temperature is decreased, whenever a temperature decrease should occur.
• $n_2$: the number of temperature decreases that will be performed without an improvement of the optimum or the average cost of the solutions analyzed, before stopping the algorithm.

The Metropolis criterion, which is used to accept or reject a candidate solution, is influenced by the temperature parameter (parameter whose name is inspired by the physical process on which simulated annealing is based). A high initial temperature signifies a high rate of solution acceptance and thus a broad covering of the solution space. As the procedure continues, the temperature is lowered and the rate of acceptance decreases.

The definition of the initial and the final temperatures, of a cooling factor for changing the temperature from iteration to iteration and the number of iterations to be performed at each temperature, needs calibration to ensure that the algorithm is robust (i.e. the results of its application are not sensitive to random number generation). A detailed description of the calibration of these parameters can be found in [8] and [9]. Figure 2 shows how the different annealing parameters work together.

![Figure 2: The cooling schedule.](image-url)
Figure 3 shows how the application of simulated annealing algorithms works to solve water system development problems. It involves a main program including two solvers: the optimization solver and the water system simulator. The optimization solver, including the different steps of the simulated algorithm, calls upon the water system simulator to check the compliance of constraints for each candidate solution.

4 Algorithm implementation: specific aspects

The procedure for generating the initial solution and the candidate solutions (known as rearrangements) and the schedule for decreasing temperatures (known as cooling schedule) is problem dependent. The next sections detail some specific aspects related to the implementation of simulating annealing algorithms for three different water systems.

4.1 Aquifer systems

Let us suppose that an aquifer will be operated to satisfy the demand of a given number of consumption centers at minimum cost. Resolving this decision problem by means of a simulated annealing algorithm implies the following tasks:

- specifying all possible location for well development (previously defined through stakeholder involvement);
- specifying all cost functions;
Hydrology, Hydraulics and Water Resources Management

- generating an initial solution of the problem;
- estimating the cost of the solution;
- checking compliance of the constraints (solving groundwater simulator for drawdown evaluation);
- generating a new solution;
- comparing the new solution with the previous one;
- applying the simulated annealing requirements;
- verifying the convergence criteria.

The initial solution can be set using all the sites previously specified to satisfy the demands. The algorithm can randomly choose the number of sites to satisfy each center and then the demand is equally divided between these sites.

The generation of new solutions (candidate solutions) as the algorithm proceeds should be suggested. As examples of generation, the following options, randomly chosen by the algorithm, can be proposed:

- An open site is chosen at random and the flow pumped there is decreased by a small amount of flow $\Delta Q$. This $\Delta Q$ added to another open site is also randomly chosen.
- A randomly chosen open site is closed and the flow pumped there will be pumped at another randomly chosen site (already open or not).
- The flow pumped at a randomly chosen open site will be pumped at a randomly chosen number of open sites.

These ways of generating candidate solutions can be combined in varying proportions during the search procedure.

After the generation of each candidate solution, the effects in terms of costs (objective function) and in terms of drawdowns must be determined. A main program for managing these actions is therefore needed. The optimizer should call upon a groundwater simulator to perform the calculations mentioned.

It is important to have an efficient and robust algorithm, this means that the calibration of the parameters included in the cooling schedule has to be carried out carefully. During this calibration, sets of cooling schedule parameters are evaluated against the quality of the solution obtained and the computer running time. For example, taking the problem described in [10] (for an aquifer management problem, as described below), Figures 4 and 5 show the influence of the parameters corresponding to the minimum number of simulations or algorithm iterations (also called minimum chain length) at each temperature ($n_1$) and the stopping criteria ($n_2$) on the quality of the solution and on computer running times. Analyzing these two figures, the values to be chosen can be $n_1 = 150$ and $n_2 = 20$.

The number of simulations or algorithm iterations run at each temperature is a crucial parameter for obtaining an optimal solution. It will have a decisive influence on the computation time and on the rate of convergence.
4.2 Water distribution systems

Let us suppose that a looped gravity-fed water distribution system, whose layout is known, is to be designed at minimum cost. Solving this decision problem by means of a simulated annealing algorithm implies the following tasks:

- Specifying all possible commercial diameters that can be used to size the pipes;
- Specifying all cost functions;
- Generating an initial solution of the problem;
- Estimating the cost of the solution;
- Checking compliance of the constraints (resolving hydraulic equilibrium equations simulator for determining flow distribution, node pressures, and velocities);
- Generating a new solution;
- Comparing the new solution with the previous one;
- Applying the simulated annealing requirements;
- Verifying the convergence criteria.
The initial solution can be set by having all the pipes of the maximum diameter. The generation of new solutions (candidate solutions) as the algorithm proceeds should be suggested. As examples of generation the following options, randomly chosen by the algorithm, may be proposed:

- A pipe is chosen at random and is assigned a new diameter. The new diameter can be larger or smaller than the present diameter. The number of times that the diameter is increased or decreased is also to be calibrated.
- Two pipes are chosen at random and are assigned with new diameters. The new diameters can be larger or smaller than the present diameters. The number of times that the diameters are increased or decreased is also to be calibrated.
- Two pipes are chosen at random and are assigned new diameters. One is assigned a diameter larger than the present one, the other is assigned a diameter smaller than the present one.

After the generation of each candidate solution, the effects in terms of cost (objective function) and hydraulic behavior must be determined. A main program for managing these actions is therefore needed. The optimizer should call upon a hydraulic equilibrium equation simulator to perform the relevant calculations.

For the calibration procedure, Figures 6 and 7 taken from the calibration of the Alperovits and Shamir looped network [11] show the influence of the parameters corresponding to the cooling factor and to the number of simulations (minimum chain length) at each temperature \( n_1 \) on the computational effort. The cooling factor transforms the temperature throughout the cooling procedure in a geometric fashion. This means that higher cooling factors produce slower temperature decrease.

![Figure 6: Effect of the cooling factor and the minimum chain length on the computational effort.](image-url)
Figure 7 depicts the rate of solution acceptance as the temperature decreases. The shape obtained fits the general pattern described before. In fact, the temperature is used to control the probability of accepting worsening moves. For high temperatures, most uphill moves are accepted. As the temperature level approaches zero, most uphill moves will be rejected.

4.3 Wastewater systems

Let us suppose that the siting and sizing of the different components of a regional wastewater system is to be decided at minimum cost. The essential components of the regional wastewater system are sewer networks and treatment plants. There is a river that will receive the treated effluent.

Solving this decision model by means of a simulated annealing algorithm implies the following tasks:

- specifying all wastewater sources and their location;
- listing all possible locations for siting wastewater treatment plants (previously defined through stakeholder involvement);
- designing a basic network containing all possible links between wastewater sources and wastewater treatment plants’ possible locations;
- generating an initial solution of the problem;
- estimating the costs of the solution;
- checking compliance of the constraints;
- generating a new solution;
- comparing the new solution with the previous one;
- applying the simulated annealing requirements;
- verifying the convergence criteria.

The initial solution can be set by randomly assigning the sewers that will be linked to each possible wastewater treatment plant.
The generation of new solutions (candidate solutions) as the algorithm proceeds should be suggested. An example of generation can be proposed:

- A randomly chosen sewer included in the current configuration is replaced by a randomly chosen sewer not included in the current configuration but connected to the same wastewater source or intermediate node (Figure 8).

After the generation of each candidate solution, the effects in terms of cost (objective function) and hydraulic behavior of the sewer system and water quality in the river must be determined. A main program for managing these actions is therefore needed. The optimizer should call upon a hydraulic sewer flow simulator and an advection diffusion simulator to perform the relevant calculations.

Figure 9 depicts information about a case study taken from [13]. It includes 13 wastewater sources and 6 possible locations for wastewater treatment plants.

Figure 8: Generation of candidate solutions (adapted from [12]).

Figure 9: Wastewater system case study.
Figure 10 represents the basic network of the problem. The nodes of this network are the wastewater sources, the locations for possible treatment plants, and the possible intermediate nodes. The arcs are the possible sewers linking the different types of nodes. The final solution will be a part of this basic network. Figure 11 presents the optimal configuration for the sewer network and for the location of the wastewater treatment plants (in this case just one) when the only constraint related to the quality parameter to be observed in the river is the minimum value of the dissolved oxygen (DO). Figure 12 depicts the optimal configuration when a more demanding environmental solution is needed. This means that various additional quality parameters have to be verified: total nitrogen (N), total phosphorus (P), and Kjeldahl nitrogen (Nkj) values.
5 Conclusions

The mathematical nature of water systems planning and design problems is such that heuristic optimization methods like simulated annealing can be very helpful. Simulated annealing is one of the best known heuristic optimization algorithms, which, due to its underlying principles, produces high-quality results. It is a random search based method that allows asymptotic convergence to optimal solutions under mild conditions. Some features of the application of this method to water systems development were discussed and some particular aspects of its implementation were described with respect to different types of water systems. A careful calibration of simulated annealing parameters and a sound generation of candidate solutions are crucial aspects for the success of this method. The use of this type of heuristic method contributes to overcome the difficulties for resolving real-world problems because all their features can properly be taken into account, therefore avoiding excessive problem simplifications.

References


Figure 12: Optimal configuration (DO>7.0 mg/l; N<7.5 mg/l; P<1.0 mg/l, and Nkj<3.5 mg/l).


