Simultaneous estimation of groundwater recharge rates, associated zone structures, and hydraulic conductivity values using fuzzy c-means clustering and harmony search optimisation algorithm: A case study of the Tahtalı watershed

M. TAMER AYVAZ1 & ALPER ELÇİ2
1 Department of Civil Engineering, Pamukkale University, 20070 Denizli, TURKEY
2 Department of Environmental Engineering, Dokuz Eylül University, 35160 Buca-Izmir, TURKEY
alper.elci@deu.edu.tr

Abstract The aim of this study is to present a linked simulation-optimisation model to estimate the groundwater recharge rates, their associated zone structures, and hydraulic conductivity values for regional, steady-state groundwater flow models. For the zone structure estimation problem the fuzzy c-means clustering (FCM) method was used. According to our current knowledge, this is the first time that this method was implemented for groundwater recharge zonation problems. The association of zone structures with the spatial distribution of groundwater recharge rates was then accomplished using an optimisation approach where the heuristic harmony search (HS) algorithm was used due to its efficiency in finding global or near global optimum solutions. Since the solution was obtained by a heuristic algorithm, the optimisation process did not require an initial solution, which is an advantage of the proposed approach. The HS based optimisation model determines the shape of zone structures, their corresponding recharge rates and hydraulic conductivity values by minimizing the root mean square error (RMSE) between simulated and observed head values at observation wells and springs, respectively. To determine the best recharge zone structure, the identification procedure starts with computation of one zone and systematically increased the zone number until the optimum zone structure is identified. Subsequently, the performance of the proposed simulation-optimisation model was evaluated on the Tahtalı watershed, an urban watershed for which a seasonal steady-state groundwater flow model was developed for a previous study. The results of our study demonstrated that the proposed simulation-optimisation model is an effective way to calibrate the groundwater flow models for the cases where tangible information about the groundwater recharge distribution does not exist.

Key words groundwater recharge; zone structure estimation; optimisation; harmony search

INTRODUCTION
The spatial distributions of groundwater recharge rates and hydraulic conductivities are key properties of groundwater flow models. For occasions when field data or measurements for these parameters are absent and cannot be obtained during the timeframe given for the modelling job, numerical estimation methods can be implemented. It is the objective of this study to propose a procedure to estimate groundwater recharge rates with the associated zone structure and hydraulic conductivities for steady-state groundwater flow models. The proposed procedure involves the adaptation of algorithms that were in the past applied in other fields of science. Here, the harmony search algorithm is used in combination with the fuzzy c-means clustering algorithm to determine zone structures and values for hydraulic conductivity and recharge. The applicability of the entire procedure was demonstrated on the semi-urban Tahtalı watershed in Izmir-Turkey, which is a key component of Izmir’s water supply system. The Tahtalı dam reservoir (38°08’ N; 27°06’ E) is located 40 km south of Izmir and meets about 36% of the city’s total water demand. The watershed of the reservoir has an area of 550 km² and is a sub-watershed of the larger K. Menderes river watershed (Fig.1). Elçi et al. (2010) previously presented results of a seasonal, steady-state groundwater flow model, for which model parameters were obtained with the parameter estimation code, PEST (Doherty, 2004). Therefore, another objective of this study is to compare parameters obtained with the proposed procedure to previously obtained ones.
MODEL DEVELOPMENT

Optimisation Model: Harmony Search Algorithm (HS)

HS, first proposed by Geem et al. (2001), is a heuristic optimisation algorithm which gets its basis from the musical processes. It is well known that the purpose of the musical processes is to seek a musically pleased harmony through making several improvisations (Yang, 2009). Although HS is a newly proposed optimisation algorithm, it has been applied to many different problems including water-related applications, structural design, information technology, transport related problems, thermal and energy problems, and many other applications. The state-of-the-art in the structure of HS algorithm and overview of its applications and developments can be found at Ingram and Zhang (2009) and Geem (2010). The mathematical statement of HS is as follows:

Let HMS be the harmony memory size, $N$ be the number of decision variables, $\{x_i\}_{i=1}^{HMS}$ be the solution vectors, and $\{x'_j\}_{j=1}^{N}$ be the newly generated solution vector.

Using these parameters, solution of an optimisation problem is performed based on the following scheme:

1. Initialization of HM: Generate initial solution vectors as many as HMS, $x^1, \ldots, x^{HMS}$.
2. Generate a new solution vector $x'$ for each $x'_j$:
   - with probability HMCR select $x'_j$ from memory, $x'_j = x^{\text{best}[1:HMS]}_j$
3. Pitch adjustment: For each $x'_j$:
   - with probability PAR change $x'_j$ as, $x'_j = x'_j \pm bw \times \text{Rnd}(0,1)$.
   - with probability $(1-\text{PAR})$ do nothing.
   - with probability $(1-\text{HMCR})$ select a new random value from the possible range.
4. If $x'$ is better than the worst $x^i$ in harmony memory, replace $x^i$ with $x'$.
5. Repeat step 2 to 5 until the given termination criterion is satisfied.
As can be seen from the computational scheme given above, HS requires some solution parameters which are Harmony Memory Size (HMS), Harmony Memory Considering Rate (HMCR), Pitch Adjusting Rate (PAR), and distance bandwidth (bw). Note that the HM is a matrix where decision variables and corresponding objective function values are stored. The HMCR is the probability of selecting any harmony from HM. If HMCR is selected too low, only few elite harmonies are selected and the algorithm can converge too slowly. On the other hand, if HMCR is selected too high, the pitches in HM are mostly used and other possibilities are not explored well (Yang, 2009).

If the generated decision variable is selected from the HM, an evaluation for the requirement of pitch adjustment is necessary. This evaluation is performed using the PAR parameter which is the probability of making pitch adjusting. Pitch adjusting is a process that is analogous of taking the slightly neighbour value based on the predefined bandwidth (bw). The pitch adjusting process is similar to the mutation operator in genetic algorithm, which maintains the diversity of population (Geem et al., 2001). Based on the experience of the authors, HMS=10, HMCR=0.95, and PAR=0.50 are appropriate values to solve many optimisation problems dealing with groundwater modelling (Ayvaz, 2009; 2010).

Estimation of the Groundwater Recharge Zone Structure

The recharge zone structure of the model domain is determined using fuzzy c-means (FCM) clustering algorithm (Bezdek, 1981). In FCM algorithm, fuzzy membership values are assigned to each data point which is related to the relative distance of that point to the cluster centers. FCM provides a procedure to group the data points that populate some multidimensional space into a specific number of different clusters (Ayvaz, 2007). Although the FCM algorithm is extensively used in many pattern classification and image processing studies, to our knowledge there is no published application example for the groundwater recharge zone structure estimation problem. The mathematical statement of FCM which is modified for the groundwater recharge zone structure estimation problem can be summarised as follows:

Let \( n_x \) and \( n_y \) be the number of finite difference grid points of the MODFLOW model in \( x \) and \( y \) directions, respectively, \( \mathbf{X} = \{ X_i \}_{i=1}^{n_x} \) and \( \mathbf{Y} = \{ Y_j \}_{j=1}^{n_y} \) be the vectors that contain the locations of grid points in \( x \) and \( y \) directions, respectively, and \( c \) be the number of clusters in which recharge rates are assumed to be homogeneous (hereafter the term of “zone” is used instead of “cluster”). Zonation of the groundwater recharge distribution is performed by using the 3-dimensional fuzzy partition matrix \( \mathbf{u} = \left[ u_{ijk} \right]_{i=1}^{n_x} \times _{j=1}^{n_y} \times _{k=1}^{c} \) such that:

\[
0 \leq u_{ijk} \leq 1 ; \quad \sum_{k=1}^{c} u_{ijk} = 1 ; \quad \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} u_{ijk} > 0
\]  

(1)

where \( u_{ijk} \) represents the fuzzy membership value between the \( (i,j)^{th} \) grid point and \( k^{th} \) zone structure. Let \( \mathbf{X} = \{ \hat{X}_k \}_{k=1}^{c} \) and \( \mathbf{Y} = \{ \hat{Y}_k \}_{k=1}^{c} \) be the zone centers to be determined by the optimisation model in \( x \) and \( y \) directions, respectively. The elements of the fuzzy partition matrix are updated using the determined zone centers as follows:
\[
U_{ij} = \left( \sum_{k=1}^{c} \left( \frac{\|X_i - \tilde{X}_k\| + \|Y_j - \tilde{Y}_k\|}{\|X_i - \tilde{X}_k\|^2 + \|Y_j - \tilde{Y}_k\|^2} \right)^{\frac{1}{m-1}} \right)^{-1} 
\]  

(2)

where \(\|\cdot\|\) is the Euclidean norm, and \(\hat{m}\) is the degree of fuzzification (\(\hat{m} = 2\)). It should be noted that if the calculated membership value of a grid point using Equation (2) has a maximum value, then this grid point is assigned to this zone (Wang and Xue, 2002). By applying this procedure to all the finite difference grid points, the flow domain can be partitioned into \(c\) zones. After this partitioning process, homogeneous groundwater recharge rates are assigned to each zone by the optimisation model, and the aquifer’s response is determined by performing a MODFLOW run.

**Problem Formulation and Search Procedure**

The purpose of applying the proposed simulation-optimisation procedure to the Tahtalı watershed model is to simultaneously estimate the groundwater recharge zone structure, associated recharge rates, and uniform hydraulic conductivity values within the six geological formations shown in Fig. 2. This problem can be formulated as an optimisation problem in which HS randomly generates the zone centers; FCM builds up the zone structures; and finally, randomly generated recharge rates and hydraulic conductivity values are assigned to the corresponding zone structures. Based on the errors for calculated hydraulic head values, zone centers, associated recharge rates, and hydraulic conductivity values are modified by the HS-based optimisation model. The objective of the optimisation model is to minimise the root mean squared error \(\mathfrak{R}\) between the simulated and observed hydraulic head values at the monitoring wells shown in Figure 2. This problem can be mathematically stated as follows:

\[
\min \mathfrak{R}(\Omega_c) = \mathfrak{R}(\Omega_c) + \sum_{k=1}^{c} P\left(\tilde{X}_k, \tilde{Y}_k\right) 
\]  

(3)

\[
\mathfrak{R}(\Omega_c) = \frac{1}{n_w}\sum_{i=1}^{n_w} \left( h_i(\Omega_c) - \tilde{h}_i \right)^2 
\]  

(4)

\[
\frac{\partial}{\partial x} \left( K h \frac{\partial h}{\partial x} \right) + \frac{\partial}{\partial y} \left( K h \frac{\partial h}{\partial y} \right) = W - R \quad \text{for } \Omega_1 \rightarrow \Omega_2 \rightarrow \cdots \rightarrow \Omega_c 
\]  

(5)

\[
P\left(\tilde{X}_k, \tilde{Y}_k\right) = \begin{cases} 
0 & \text{if } (\tilde{X}_k, \tilde{Y}_k) \text{ is in active cell} \\
\lambda \kappa & \text{if } (\tilde{X}_k, \tilde{Y}_k) \text{ is in inactive cell} 
\end{cases} 
\]  

(6)

where \(\Omega_c\) is the solution of the problem with \(c\) recharge zones, \(\mathfrak{R}(\Omega_c)\) is the root mean square error for the solution of \(\Omega_c\), \(\mathfrak{R}(\Omega_c)\) penalised objective function for the solution of \(\Omega_c\), \(h_i(\Omega_c)\) is the simulated hydraulic head value at observation well \(i\) for the solution of \(\Omega_c\), \(\tilde{h}_i\) is the observed hydraulic head value at observation well \(i\), \(n_w\) is the number of observation wells (\(n_w = 51\)), \(h\) is the hydraulic head over the flow domain, \(W\) is the sinks/source term due to pumping, \(R\) is the set of groundwater recharge rates to be estimated such that \(R \in \{R_1, R_2, \cdots, R_c\}\), \(\widetilde{K}\) is the set of hydraulic...
Simultaneous identification of groundwater recharge rates and zone structures using fuzzy c-means clustering and harmony search optimization algorithm: A case study of the Tahtalı watershed

conductivities to be estimated such that $\bar{K} \in \{K_1, K_2, \ldots, K_6\}$, $P(\hat{X}_k, \hat{Y}_k)$ is the penalty function depending on the locations of zone centers, $\lambda$ is the penalty parameter, and $\kappa$ is the nearest distance to the model boundary in terms of the row and column numbers of the finite difference grid (Figure 3). As can be seen from Equations (3) to (7), the groundwater flow process enters the problem in Equation (5) for unknown $\bar{R}$ and $\bar{K}$ distributions. These distributions are determined by the optimisation model and passed on to MODFLOW to obtain the solution for groundwater flow in the study area.

It should be noted that the reason for using the penalty function given in Equation (6) is the irregular shape of the modelling domain. All the grid cells outside the model boundary are specified as inactive cells, which are shown as the dark shaded area in Fig. 3. Although these cells appear in the finite difference grid structure of the MODFLOW model, they are excluded from the numerical solution. Therefore, these inactive cells must be also excluded from the search space of the zone centers. Equation (6) states that if a zone centre is located in an inactive cell, the calculated objective function value is penalized with $\lambda \kappa$. The values of $\kappa$ increase as the zone centers move away from the model boundaries. The value of $\lambda$ is mostly arbitrary and problem dependent. Our trials show that $\lambda = 100$ can be used for the implementation of the penalty function given in Equation (6). The decision variables of the optimisation model are the locations of the zone centers, associated recharge values for each zone, and uniform hydraulic conductivity values for each of the six pre-defined geological zones.

![Fig. 3](image-url) The model domain and $\kappa$ values used for the penalty function given in Equation (6)

Although the proposed simulation-optimisation model may solve the problem based on the solution scheme given in Equation (3) to (7), this mathematical formulation is valid only for cases where the number of zones ($c$) is known a priori. However, recharge zone structures, their numbers, and the associated recharge rates are unknown for the most cases. Therefore, it is necessary to determine the number of zones such that the eventually identified zone structure optimally represents the field data. With this purpose, the zone structure estimation problem starts with one zone, and then, systematically increases the zone number until the best solution is obtained.

Furthermore, each successive solution for different zone numbers requires three additional decision variables (one is for recharge rate and two for zone centre coordinates). However, when the number of decision variables increases, there is a greater chance of producing local optimum
solutions due to the increased dimension of solution space (Huang and Mayer, 1997). This situation also leads to less reliable solutions. For such cases, the final value of the objective function may increase although \( R \to 0 \), while \( c \to \infty \) (Ayvaz, 2007). Therefore, final identified parameter values, zone structures, and objective function values are evaluated altogether to decide which successive zone structure best represents field conditions.

Our trial runs indicate that the value of the objective function does not improve significantly after about 15,000 iterations of HS. Therefore, the maximum number of iterations is set to 20,000. Completing 20,000 iterations of HS takes about 12 hours on a workstation with Intel Xeon 3.07 GHz processor and 6 GB RAM.

**IDENTIFICATION RESULTS**

Fig. 4 shows the identified groundwater recharge zone structures for the solutions of \( \Omega_2 \) to \( \Omega_6 \) and the recharge zone structure originally used by Elçi et al. (2010). As can be seen from Figure 4, centres of the identified zones remain inside the flow domain by virtue of the penalty function implementation. This result also implies that the final objective function values do not include any penalty term (i.e. \( R^* = 0 \)).

Fig. 4 Comparison of the identified zone structures; (a) to (e): for \( \Omega_2 \) to \( \Omega_6 \) (small circles correspond to zone centers); (f): The zone structure used by Elçi et al. (2010) (shaded area represents the Tahtali watershed).

Summary of the identified hydraulic conductivity values, recharge rates, and final \( R^* \) values for solutions \( \Omega_1 \) to \( \Omega_6 \), and the results by Elçi et al. (2010) are given in Table 1. The simulation-optimisation model calculates hydraulic conductivity values that are comparable between all solutions. On the other hand, the identified recharge rates are all different because the zone structures for recharge evolve during the optimisation, while the zone structure for hydraulic conductivity is fixed. Regarding final \( R^* \) values after 20,000 iterations, it can be observed that the largest \( R^* \) value (16.18) is obtained for \( \Omega_1 \) where it is assumed that the flow domain takes a
Simultaneous identification of groundwater recharge rates and zone structures using fuzzy c-means clustering and harmony search optimization algorithm: A case study of the Tahtalı watershed

uniform recharge with a rate of $2.65 \times 10^{-4}$ m/d. For this solution, the number of decision variables is seven, six for conductivities and one for uniform recharge rate. After this solution, the value of $R_1$ decreases as the solution approaches $\Omega_4$, and increases again for $\Omega_5$ and $\Omega_6$. As mentioned earlier, theoretically the increase in the zone numbers should result in the decrease in the corresponding $R_1$ values. For such cases, however, the reliability of the identified parameters generally decreases since more unknown parameters need to be determined (Ayvaz, 2007). Therefore, by considering the identified parameter values, zone structures, and the final $R_1$ values, it can be concluded that the reliability of the identified solutions after $\Omega_4$ tends to decrease. Thus, the four-zone structure ($\Omega_4$) is selected as the best zone structure for the estimation problem discussed here (Figure 4(c)).

**Table 1** Summary of the identified hydraulic conductivity values, recharge rates and final $R_1$ values

<table>
<thead>
<tr>
<th>Identified Parameters</th>
<th>$\Omega_1$</th>
<th>$\Omega_2$</th>
<th>$\Omega_3$</th>
<th>$\Omega_4$</th>
<th>$\Omega_5$</th>
<th>$\Omega_6$</th>
<th>Elçi et al. (2010)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hydraulic Conductivities (m/d)</td>
<td>$K_1$</td>
<td>20.81</td>
<td>22.12</td>
<td>20.16</td>
<td>25.87</td>
<td>26.35</td>
<td>24.38</td>
</tr>
<tr>
<td></td>
<td>$K_2$</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>$K_3$</td>
<td>1.00</td>
<td>1.00</td>
<td>0.54</td>
<td>0.95</td>
<td>0.91</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>$K_4$</td>
<td>5.74</td>
<td>6.07</td>
<td>7.49</td>
<td>7.57</td>
<td>6.99</td>
<td>6.46</td>
</tr>
<tr>
<td></td>
<td>$K_5$</td>
<td>5.00</td>
<td>5.00</td>
<td>5.00</td>
<td>4.98</td>
<td>4.38</td>
<td>4.47</td>
</tr>
<tr>
<td></td>
<td>$K_6$</td>
<td>3.58</td>
<td>3.62</td>
<td>2.61</td>
<td>2.56</td>
<td>1.64</td>
<td>3.06</td>
</tr>
<tr>
<td>Groundwater Recharge Rates (m/d)</td>
<td>$R_1$</td>
<td>2.65E-04</td>
<td>1.05E-03</td>
<td>1.00E-10</td>
<td>9.53E-04</td>
<td>4.02E-05</td>
<td>3.23E-06</td>
</tr>
<tr>
<td></td>
<td>$R_2$</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>$R_3$</td>
<td>1.00</td>
<td>1.00</td>
<td>0.54</td>
<td>0.95</td>
<td>0.91</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>$R_4$</td>
<td>5.74</td>
<td>6.07</td>
<td>7.49</td>
<td>7.57</td>
<td>6.99</td>
<td>6.46</td>
</tr>
<tr>
<td></td>
<td>$R_5$</td>
<td>5.00</td>
<td>5.00</td>
<td>5.00</td>
<td>4.98</td>
<td>4.38</td>
<td>4.47</td>
</tr>
<tr>
<td></td>
<td>$R_6$</td>
<td>3.58</td>
<td>3.62</td>
<td>2.61</td>
<td>2.56</td>
<td>1.64</td>
<td>3.06</td>
</tr>
<tr>
<td>Final $R_1$ value</td>
<td>16.18</td>
<td>15.97</td>
<td>12.96</td>
<td>11.90</td>
<td>12.55</td>
<td>13.30</td>
<td>16.40</td>
</tr>
</tbody>
</table>

*a* This value equals to $R_1$ and does not include the penalty function in Equation (7)

*b* These values were calculated using the PEST model for fixed recharge and conductivity zone structures

*c* This value was calculated based on a external transient precipitation-runoff model

In the previous modelling study by Elçi et al. (2010), the calibration of the same groundwater flow model was performed by adjusting the recharge rates and hydraulic conductivity values using the PEST parameter estimation code, while keeping the recharge rate for the zone representing the Tahtali watershed constant at a value that was obtained by an independent precipitation-runoff model. For that study the hydraulic conductivity zone structure was based on the geology of the study area and the four-zone recharge zone structure was manually created based on land use/land cover and lithology information. Comparison of results obtained by Elçi et al. (2010) with the results for $\Omega_4$ shows that optimised hydraulic conductivity values in this study are in the same order of magnitude, except for zone 2 ($K_2$). However, this is not the case for recharge rates, as they are different for both studies. This difference can be explained by the different outcome of zone structures in both studies. It had to be assumed by Elçi et al., that the recharge rate for the entire watershed (zone 4 in Figure 4(f)) is uniform since the precipitation-runoff model was a lumped model. In the current model, however, this part of the model domain was split into other zones, each allowed to have different recharge rates. Elçi et al. (2010) obtained a final $R_1$ value as 16.40 for the four-zone structure given in Figure 4(f), which indicates a less optimised solution compared to the $R_1$ value (11.90) of the $\Omega_4$ solution given in Figure 4(c). As can be seen from these results, the final $R_1$ value decreases by 27% through the use of the simulation-optimisation procedure when compared to Elçi et al. (2010). Based on the error evaluation, it can be concluded that the groundwater flow model is improved with the proposed procedure.

**CONCLUSIONS**

A coupled simulation-optimisation model is developed for the simultaneous estimation of groundwater recharge zone structure, their associated recharge rates and hydraulic conductivity
values. In this model, MODFLOW is used to perform the steady-state groundwater flow calculations. The association of recharge zone structures with the recharge rates is accomplished by linking the FCM clustering algorithm with the MODFLOW in the simulation model. This model is then integrated to an optimisation model where heuristic HS algorithm is used. The main objective of the HS based optimisation model is to minimize the root mean square error which is calculated between the simulated and observed hydraulic head values at available observation wells by adjusting the zone centres, associated recharge rates within each generated zone, and uniform hydraulic conductivity values. The applicability of the developed model is evaluated in a case study for the Tahtalı watershed (Izmir-Turkey) and the estimation results are compared to previous modelling results for the same model domain that were obtained with a different optimisation approach. Comparison of the results indicate that the developed model may be an effective way in calibrating steady-state groundwater flow models, where tangible information about the groundwater recharge distribution does not exist.

Acknowledgements This study is based on the work supported by The Turkish Academy of Sciences (TÜBA) - The Young Scientists Award Programme (GEBIP). The first author would like to thank TÜBA for their support of this study.

REFERENCES