Chapter 7: DECISION VARIABLE REDUCTION IN LARGE SCALE MULTI-OBJECTIVE GROUNDWATER OPTIMIZATION PROBLEMS

7.1 Introduction

As discussed in previous chapters, groundwater management is a complex decision-making challenge that demands comprehensive social, economic, and environmental evaluations of quality and quantity-related aspects. Integrating meta-heuristic evolutionary algorithms with simulation optimization has greatly expanded the capabilities of S-O in tackling complex groundwater management problems. Despite its effectiveness in addressing complex groundwater management problems, simulation optimization suffers from two significant drawbacks: convergence issues in large decision variable problems, which are often referred to as the "Curse of dimensionality" (D. Chen et al., 2016; Crevillén-García, 2018; Hou and Behdinan, 2022) and determining optimal decision variable values at an administrative level to facilitate definitive policy implementation. High-dimensional GSOP is characterized by an exponential increase in the search space domain when the dimensions of the problem are increased linearly. As a result, algorithms and techniques that are effective in low-dimensional problems become infeasible when exploring the vast search space of high-dimensional problems (Du et al., 2019). Furthermore, tuning the algorithm's hyper-parameters to perform better in high-dimensional search space is computationally

expensive. The decision variables in case of groundwater management problems are often discharge rates and pumping well locations (quantitative aspects), the cost associated with installing a single well (cost minimization), the injection flow rate in a well (aquifer recharge), initial concentration of contaminants at each well (quality considerations), and net benefit of the crop (agricultural concerns).

This chapter focuses on two primary objectives. The first objective is to introduce a novel approach that utilizes aquifer parameters for clustering, effectively reducing the number of DVs. Furthermore, this approach finds the impact of the clustering techniques on the evaluation metrics of the Pareto fronts. The pareto front obtained by solving multi-objective GSOPs is compared using the convergence, diversity, and uniformity indices. The optimal results' behavior when aquifer parameters are incorporated into the clustering variables is also investigated. The S-O model is developed at the sub-basin scale; however, GW management policies are typically established at the administrative level. To assess the spatial variability of decision variables at the administrative level and determine their significance, the optimal discharge rates for each district (commune) were evaluated. Thus, the second objective aims to calculate the optimal discharge rates for specific communes at the administrative level and achieve a holistic approach to groundwater management. It has been demonstrated that certain aquifer properties provide huge variability to the pareto front and thus may not be omitted without sacrificing generality.

7.2 Methods

The research commenced by developing a GSOP model. The simulation model was constructed using MODFLOW, while the MOPSO algorithm was chosen as the optimization approach. The integration of the S-O model was carried out in Python. The study initially considered a total of 319 wells as DV. Given the absence of prior knowledge regarding the optimal number of clusters for initial reduction, the decision variables were

reduced into 20, 40, and 80 distinct well management zones or clusters. Three clustering algorithms (Ward, K-means, and Affinity Propagation) were employed to ensure the consistency of the obtained results. First, the Euclidean distance between the wells was selected as the pivotal parameter for the clustering process, facilitating identifying spatial relationships and patterns among the decision variables. Further, an optimal number of clusters for the GSOP was finalized using several intrinsic clustering evaluation measures, including Calinski-Harabasz, Davies-Bouldin, and Silhouette values. After the optimal number of DV was finalized, multi-parameter clustering was introduced based on fundamental aquifer properties that govern groundwater flow. They incorporate four key hydrogeological parameters: top layer elevation and topography of the aquifer (top elevation), hydraulic conductivity of the aquifer material (HK), deep aquifer recharge from precipitation (RCH), and initial piezometric head data of wells in the study area (Initial Head). After obtaining the converged solution, potential changes in the accuracy of Pareto fronts were assessed considering convergence, diversity, and uniformity criteria. Three points along the pareto front were selected, with one representing the central region and the other representing the periphery. The total discharge for each commune was calculated by multiplying the number of wells within that zone by the corresponding decision variable (DV) value derived from the S-O model. Specifically, two discharge values were computed for each district: one before the implementation of aquifer parameter clustering and the other after incorporating aquifer parameter-based clustering, facilitating groundwater management at the administrative level. The study also investigated the change in discharge values before and after implementing aquifer parameter-based clustering. For a visual representation of the methodology employed, refer to Figure 7.1.

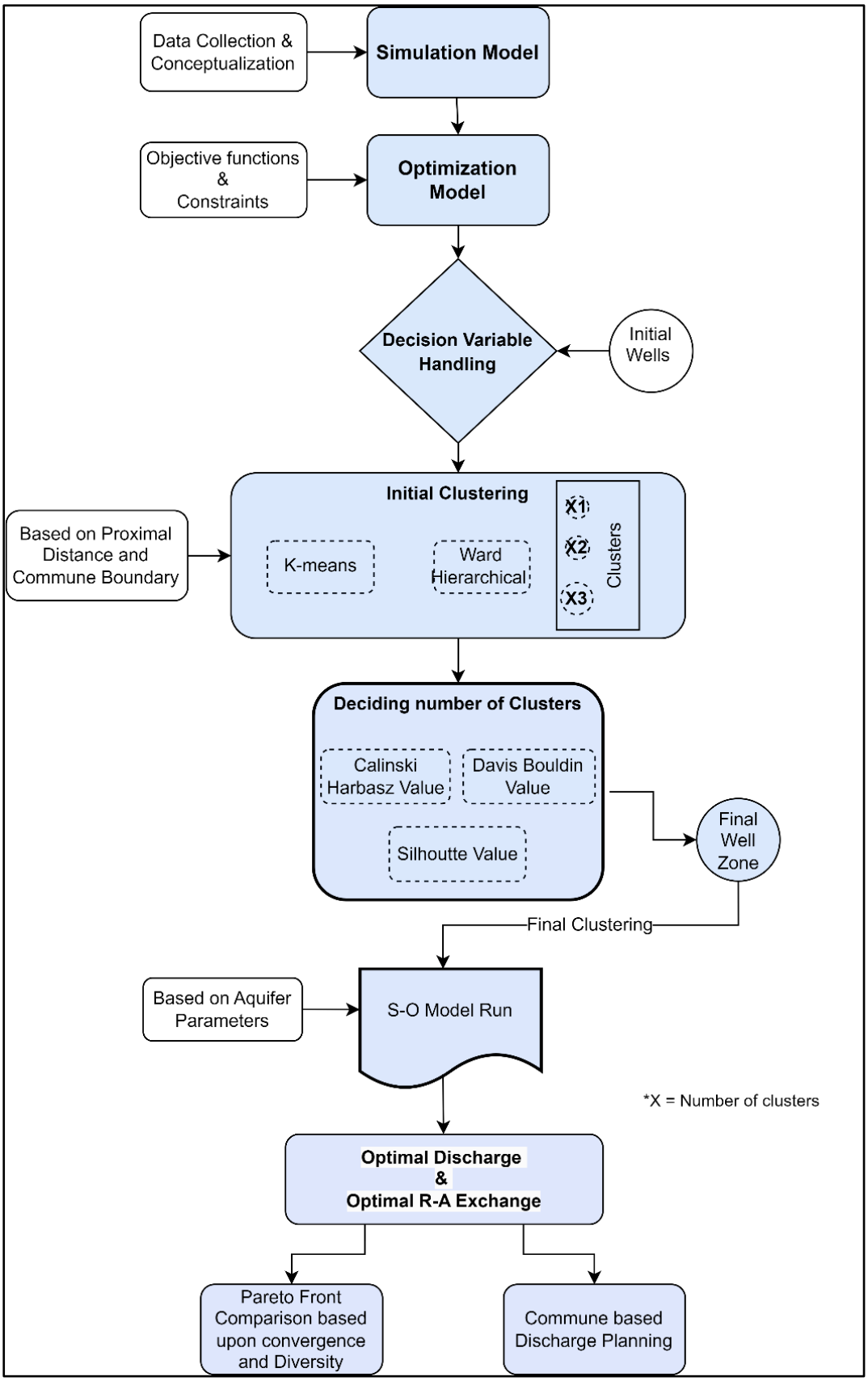


Figure 7.1 Schematic diagram of the workflow.

7.2.1 Clustering Algorithms

7.2.1.1 K-means Clustering

K-Means clustering, introduced by (MacQueen, 19967) is an unsupervised learning algorithm widely applied in fields such as groundwater management and beyond. The algorithm operates iteratively to partition a dataset into K predefined clusters, aiming to minimize the sum of squared distances between data points (e.g., well locations) and the centroids of their respective clusters. The number of clusters, K, is a critical hyperparameter and must be determined before applying the algorithm. Selecting the optimal value of K is crucial for accurate clustering and typically involves calculating various indices and elbow methods. In the elbow method, the total within-cluster sum of squares (WSS) is plotted against increasing values of K. The point where the reduction in WSS begins to taper off—forming an elbow—is chosen as the optimal K. Calculating the indices, on the other hand, evaluates how well each point lies within its cluster by computing the different coefficient, with higher values indicating better-defined clusters.

Once K is chosen, the algorithm starts by randomly initializing K centroids. It then assigns each data point to the nearest centroid based on Euclidean distance and recalculates the centroids as the mean of the points assigned to each cluster. This process—assignment and updating—is repeated iteratively until convergence, which occurs when the centroids stabilize or the assignments of data points no longer change. The algorithm's objective is to minimize the sum of squared errors (SSE) or the inertia of the clusters. Supposing that the target object is x, x_i indicates the average of the cluster, C_i criterion function is defined as follows:

$$E = \sum_{i=1}^{k} \sum_{x \epsilon C_i} |x - x_i|^2 \qquad (7.1)$$

where E is the total SSE across all clusters, K is the number of clusters, and Ci represents the points in cluster i. The Euclidean distance, which measures the similarity between points and centroids, is calculated as:

$$d(x_i, y_i) = \left[\sum_{i=1}^{n} (x_i y_i)^2\right]^{1/2} \quad (7.2)$$

where n is the number of dimensions.

Another important consideration is the initialization of the centroids, as random initializations may lead to different final clusters. The K-Means algorithm assumes spherical clusters and equal variance within clusters, so its performance may degrade when dealing with clusters of various shapes and densities. Despite these caveats, K-Means remains highly efficient for large datasets due to its linear time complexity and ease of implementation, making it a popular tool for clustering tasks in both research and applied settings.

7.2.1.2 Ward method

Ward's method, an agglomerative hierarchical algorithm employed in cluster analysis, treats the clustering task as an inspection of the variance problem, distinguishing it from conventional clustering approaches that predominantly rely on distance metrics or measures of association (Liu et al., 2020). This approach adopts an analysis of variance (ANOVA) based framework, wherein one-way univariate ANOVAs are performed for each well, with groups delineated by the clusters at that specific stage of the clustering process. A key advantage of Ward's method lies in its suitability for quantitative variables, making it a well-suited choice for extracting meaningful insights and patterns from complex water resource management datasets. The objective function of the algorithm is defined as:

$$ESS_j = \sum_{i=1}^{k} \left\| X_{ij} - \bar{X}_j \right\|^2 \quad (7.3)$$

Where X_{ij} is the i[th] well in the j[th] cluster and k is total wells and $\bar{X}_j$ is the well at average mean of the dataset.

7.2.1.3 Affinity Propagation

Traditional clustering methods often rely on the preselection of "K" centroid points, making them sensitive to initial parameter choices and necessitating multiple runs to obtain optimal clusters. However, a novel approach known as affinity propagation (Frey and Dueck, 2007) considered all points as potential cluster centers or exemplars. The similarity measure reflects the suitability of a given well to serve as a cluster center, while the preference parameter influences the number of clusters formed. A damping factor also controls message availability, mitigating numerical instability during the message updating phase. This approach offers a data-driven solution for clustering, eliminating the need for manual selection of initial centroids and providing improved accuracy and reliability in cluster formation. In the spatial clustering, 6 S-O models were computed, while 12 S-O models were generated based on the type of aquifer parameter-based clustering and choice of clustering algorithms. A total of 18 S-O models were generated for the analysis. A list summarizing the combinations is given in Table 7.1.

Table 7.1 The total combination of S-O models run.

	Type of clustering	Cluster numbers	Algorithm(s)
1.	Spatial	20,40,80	Ward, K-means
2.	Aquifer Property		
	a) Top elevation + distance from the river	40	Ward, K-means and Affinity Propagation
	b) Hydraulic conductivity + distance from the river	40	
	c) Recharge + distance from the river	40	
	d) Initial head + distance from the river	40	

7.2.2 Clustering Performance Metrics

7.2.2.1 Calinski Harabasz Value

The Calinski-Harabasz criterion is particularly well-suited for clustering solutions derived from k-means clustering, and when squared, Euclidean distances are used to calculate distances between points. The Calinski-Harabasz criterion also called the Variance Ratio Criterion (VRC), is a widely used metric to evaluate the quality of clustering solutions. The Calinski-Harabasz index is defined as:

$$VRC_k = \frac{Cb}{Sw} \times \frac{(N-k)}{(k-1)} \tag{7.4}$$

where:

Cb represents the overall between-cluster variance,

Sw denotes the overall within-cluster variance,

k is the number of clusters, and

N is the total number of observations in the dataset.

The between-cluster variance Cb is calculated as:

$$Cb = \sum_{i=1}^{k} n_i \|m_i - m\|^2 \tag{7.5}$$

where:

k is the number of clusters,

n_i is the number of observations in cluster iii,

m_i is the centroid of cluster iii, and

m is the overall mean of the entire dataset.

Here, $\|m_i - m\|$ represents the L2 norm (Euclidean distance) between the centroid of cluster i and the overall mean.

The within cluster variance (Sw) is calculated as:

$$Sw = \sum_{i=1}^{k} \sum_{x \in c_i} \|x - m_i\|^2 \quad (7.6)$$

where:

x is a data point,

c_i is the i[th] cluster, and

m_i is the centroid of cluster i.

A clustering solution is considered well-defined when it exhibits high between-cluster variance and low within-cluster variance. The higher the VRC ratio, the better the cluster separation and, consequently, the more effective the clustering solution. To identify the optimal number of clusters (k), one should aim to maximize the value of VRC_k. The optimal solution corresponds to the number of clusters that yield the highest Calinski-Harabasz index.

7.2.2.2 Davis Bouldin Index

The Davies-Bouldin criterion is a widely used metric for assessing the quality of clustering solutions. It measures the ratio of within-cluster dispersion to between-cluster separation, indicating how well-separated and compact the clusters are. The Davies-Bouldin index is defined as:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max\{D_{i,j}\} \quad (7.7)$$

where:

k is the number of clusters, and

$D_{i,j}$ represents the within-to-between cluster distance ratio for clusters i and j.

The within-to-between cluster distance ratio $D_{i,j}$ is given by:

$$D_{i,j} = \frac{d_i + d_j}{d_{i,j}} \tag{7.8}$$

where:

d_i is the average distance between each point in the i-th cluster and its centroid,

d_j is the average distance between each point in the j-th cluster and its centroid, and

$d_{i,j}$ is the Euclidean distance between the centroids of clusters i and j.

In this formulation, the numerator $(d_i + d_j)$ represents the within-cluster scatter or dispersion for clusters i and j, while the denominator $d_{i,j}$ captures the between-cluster separation, i.e., the distance between their centroids. For each cluster i, the maximum value of $D_{i,j}$ across all clusters j (where j ≠ i) represents the worst-case within-to-between cluster ratios. This value reflects the worst-case scenario of how well cluster i is separated from the others. The Davies-Bouldin index is calculated as the average of these worst-case ratios across all clusters. A lower Davies-Bouldin index indicates better clustering performance, implying smaller within-cluster dispersion relative to the between-cluster distances. Thus, the optimal clustering solution is the one that minimizes the Davies-Bouldin index, yielding well-separated and compact clusters. This criterion is often employed to evaluate clustering algorithms, including k-means, particularly when the Euclidean distance is used to measure distances between cluster centroids and data points.

7.2.2.3 Silhouette index

The silhouette value for each point provides a quantitative measure of how well that point fits within its assigned cluster compared to other clusters. It evaluates the quality of clustering by assessing the similarity of each point to points relative to points in different clusters. The silhouette value for the i-th point, denoted as s_i, is defined mathematically as:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{7.9}$$

where:

a_i is the average distance from the i^{th} point to all other points within the same cluster (representing how closely related it is to points within its cluster), b_i is the minimum average distance from the i^{th} point to points in a different cluster, calculated by minimizing this value across all clusters (representing how distinct the point is from points in other clusters). If the i^{th} point is the only point in its cluster, the silhouette value s_i is set to 1 by default, as it cannot be compared to other points within the cluster.

The silhouette values fall within the range of -1 to 1: a value close to 1 indicates that the point is well-clustered, meaning it is significantly closer to points in its cluster than in other clusters. A value near 0 suggests that the point lies on the boundary between clusters. A negative value indicates that the point is likely assigned to the wrong cluster, as it is closer to points in a different cluster than to points within its own. To evaluate the overall quality of a clustering solution, the silhouette values of all points are averaged. A high average silhouette value implies a well-defined clustering solution, where most points are well-matched to their respective clusters. Conversely, if many points have low or negative silhouette values, this may indicate an inappropriate clustering solution, potentially due to an excessive or insufficient number of clusters.

The silhouette analysis can be used as an evaluation criterion for clustering with any distance metric, such as Euclidean or Manhattan distance. Each cluster contributes equally to the overall silhouette score, regardless of size. The optimal number of clusters corresponds to the solution with the highest average silhouette value, indicating the most cohesive and well-separated clusters.

The combined analysis of the three abovementioned clustering performance matrices was done in the Python platform. The optimal clustering number was derived from this analysis.

7.3 Case study

The S-O models were developed for the lower Ain River basin, a significant right-bank tributary of the Rhône River, as shown in Figure 7.2. Detailed descriptions of the study area can be found in Chapter 3.

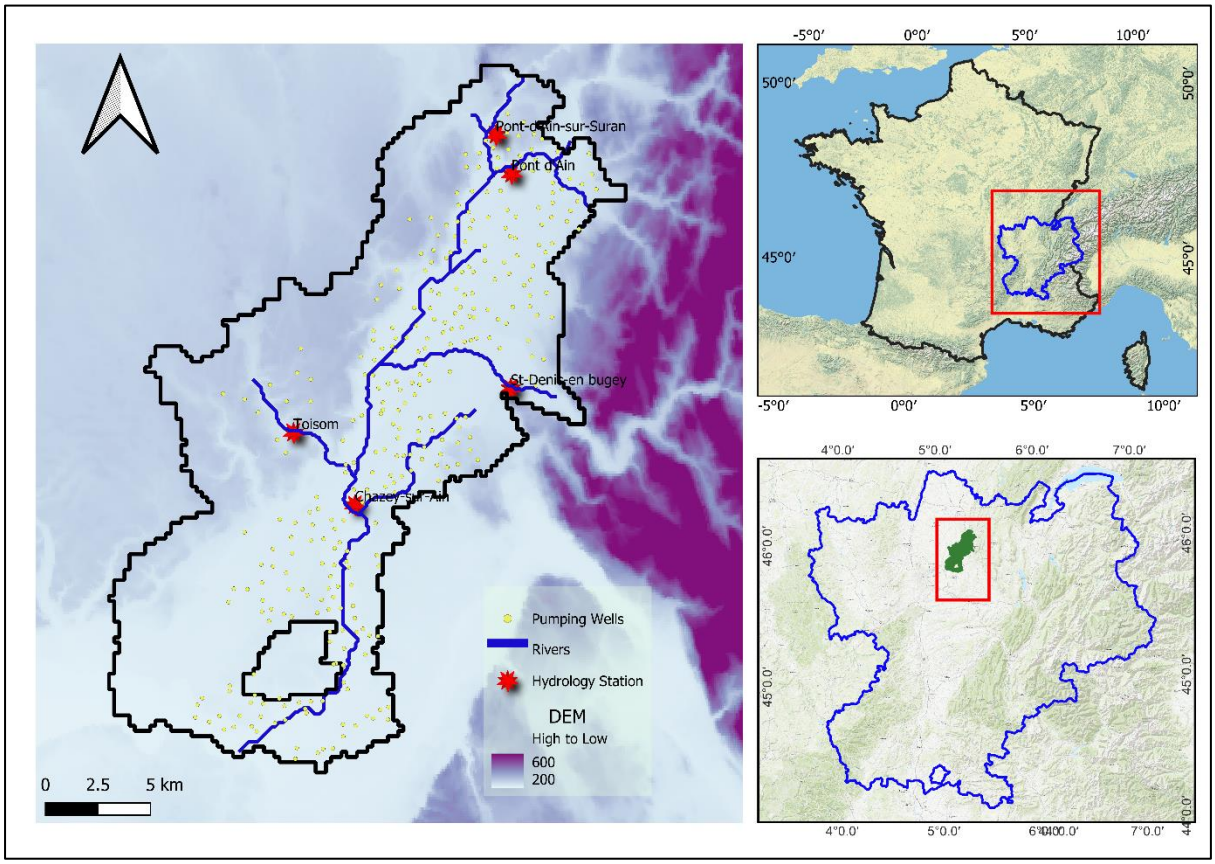


Figure 7.2 Case study of the Lower Ain River basin

The optimization algorithm used in the chapter is MOPSO (Lalwani et al., 2013). The efficacy and robustness of MOPSO have already been discussed in Chapter 4.

7.4 Results

The comparative analysis of spatial clustering using both the Ward hierarchical and K-means algorithms yielded the optimal number of decision variables to be 40. In the case of the Ward algorithm, employing 40 DVs consistently yielded favorable outcomes characterized by low Davies Bouldin values. Conversely, for the K-means algorithm, the Calinski Harabasz values associated with 40 clusters (804.036) were higher than other

clusters (751.986, 770.831). The Silhouette values further supported this observation, notably higher for 40 clusters. However, concerning Davies Bouldin values, 80 number of DVs (0.7203) were better relative to 40 (0.7723) and 20(0.7664). The decision variables were thus reduced to 40 based on the abovementioned indices. Hence, selecting the smallest possible number of DVs is not always viable for GSOPs. Figure 7.3 depicts the three primary criteria for cluster analysis, and its representation in both algorithms.

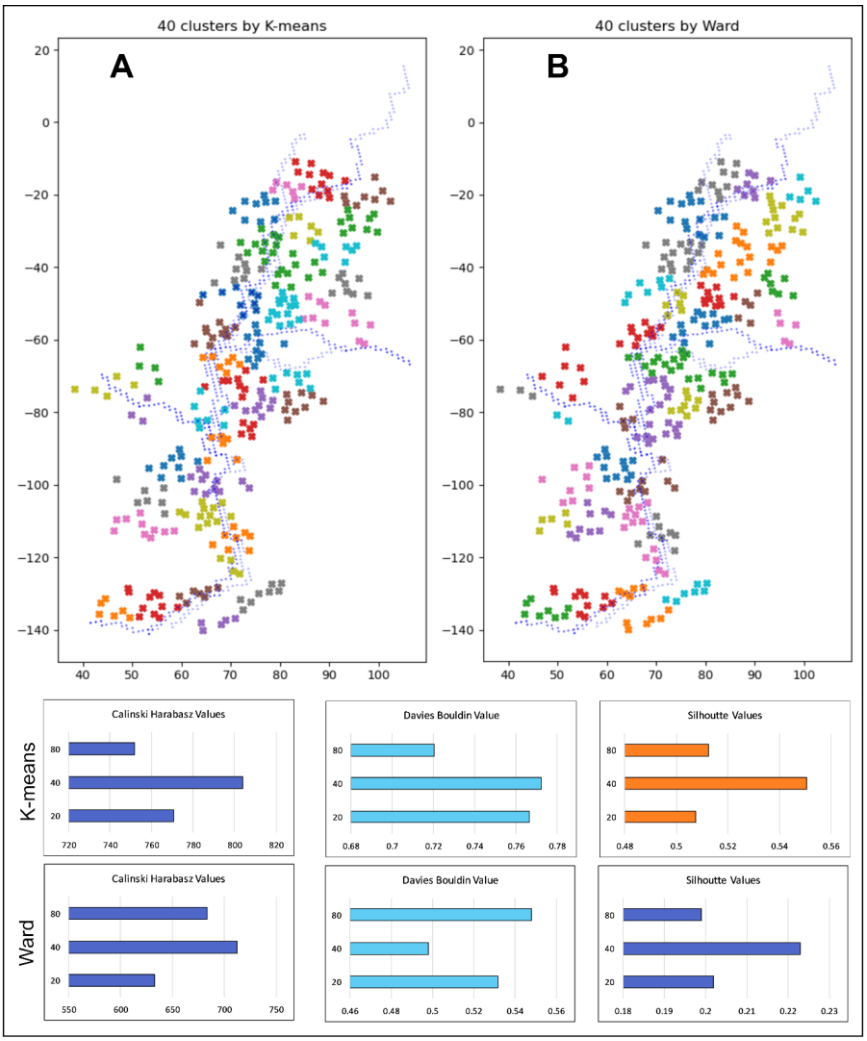
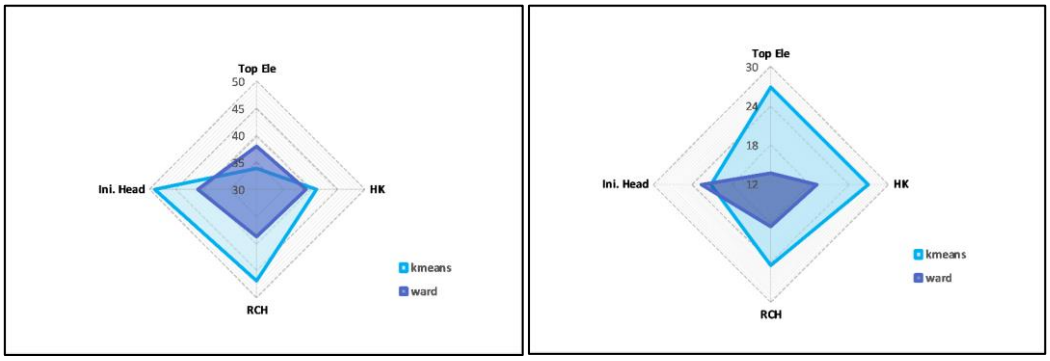


Figure 7.3 An optimal number of clusters using three clustering metrics. Figures A and B represent the 40 well clusters by K-means and Ward algorithm.

The subsequent analysis of the results revealed that the pareto front obtained through aquifer parameter-based clustering, using a fixed number of decision variables, exhibited superior convergence and diversity compared to the pareto front obtained solely based on

the spatial distribution as the clustering criterion. The introduction of the parameter-based clustering approach resulted in a significant increase in hypervolume and a decrease in inverted generational distance, irrespective of the clustering algorithm or hydrogeologic parameters used. Increase in the hypervolume was maximum for initial head based-clustering by K-means approach (46%) while decrease in IGD was maximum for recharge-based clustering in K-means (28%). In the ward hierarchy, the initial head showed a maximum increase in hypervolume (35%) and a maximum decrease in IGD values (22%). However, in the case of K-means clustering, the change in performance metrics was more pronounced compared to ward hierarchical clustering, as depicted in Figure 7.4.



a b

Figure 7.4 Percentage (a) increase in hypervolume (b) decrease in IGD, after introduction of aquifer parameter-based clustering.

In the analysis of various pareto performance metrics (Figure 7.5), Hypervolume starts at approximately 0.6, showing a steep decline as Epsilon increases, while Spread shows an upward trend from 1.1 to 1.3. Meanwhile, IGD remains relatively low and steady, varying between 0.002 and 0.014, indicating better convergence. In terms of individual aquifer properties, the blue line representing *Top Ele* and the yellow line representing *HK* show diverging behaviors with increasing Epsilon, with *Top Ele* decreasing as Epsilon rises. Hypervolume declines slightly from 0.54 to 0.48 in the K-Means algorithm, reflecting a smaller Pareto front as the optimization progresses. Epsilon exhibits an increase up to 200,000, while the Spread metric shows considerable fluctuation, peaking at 1.15. For Ward

Clustering, the Spread remains relatively flat, hovering near 1.0 throughout. It can be observed from the analysis that the aquifer properties, particularly *Top Ele* and *RCH* (red and yellow lines), demonstrate significant variability in their impact on the clustering results, with *HK* (orange line) showing the least fluctuation.

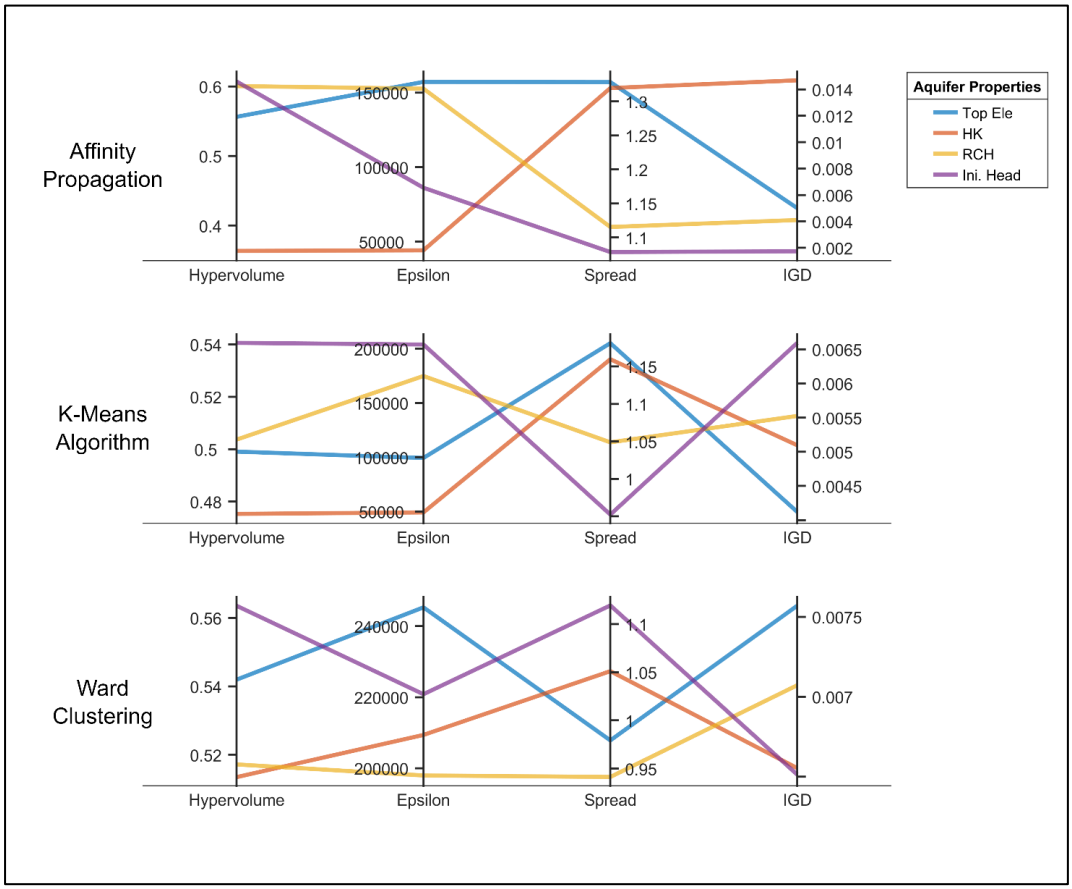


Figure 7.5 Variation of Pareto performance metrics for different clustering algorithms after decision variable reduction.

The comparison among the solutions derived from different aquifer properties indicated that initial head-based clustering yielded a better pareto front in terms of convergence and diversity. This finding underscores the significance of using accurate initial water table data in modeling groundwater flow. The various pareto fronts obtained can be visualized in Figure 7.6. In Affinity Propagation, the exchanges between the river and aquifer (objective 1) increased from 2.2 $\times 10^5$ to 2.85 $\times 10^5$ between *HK* and the initial head, respectively. Furthermore, the total discharge (objective 2) also exhibited a significant increase

compared to the other parameters, with an approximate difference of 10.6% when transitioning from *HK* to *RCH* as the clustering parameter. On the contrary, K-means clustering yielded results within a close range for both objectives, varying from as low as 5% to as high as 15% between them for all the aquifer parameters. In the case of ward hierarchical clustering, the pareto front obtained from recharge performed better in terms of epsilon and spread criteria, whereas the initial head-based pareto front exhibited superior performance in terms of hypervolume and epsilon. Although the pareto front is diverse in HK and RCH-based solutions, the solutions fail in uniformity and spread.

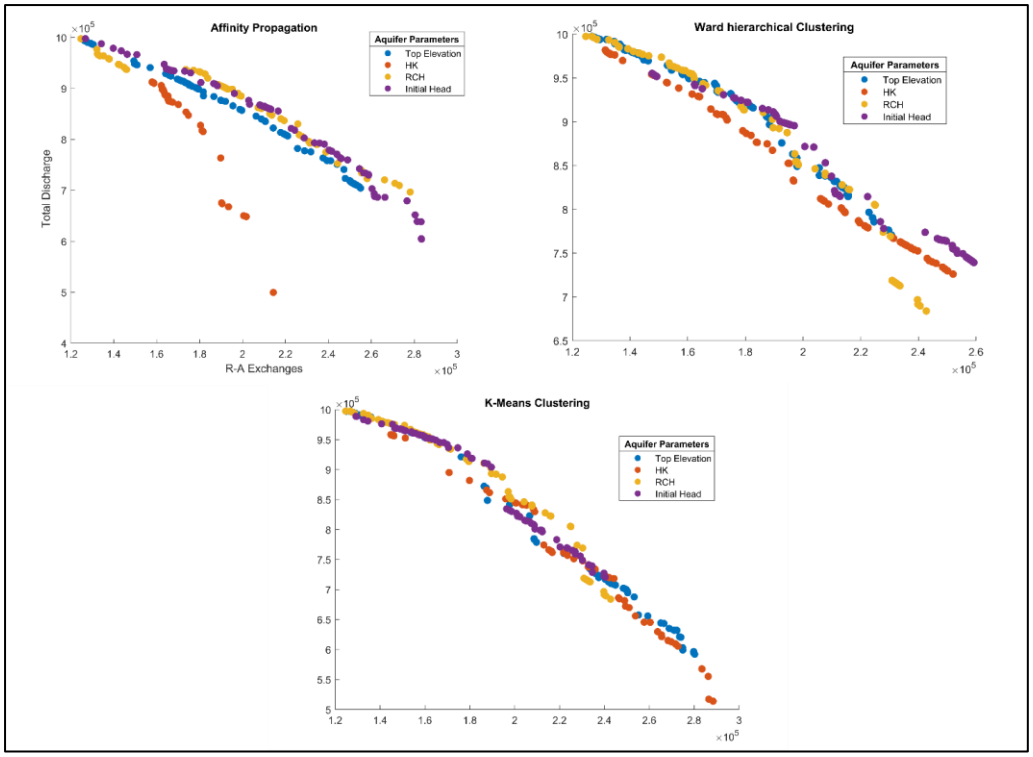


Figure 7.6 Pareto fronts obtained after convergence of S-O model based upon aquifer parameter clustering. The X-axis and Y-axis represent the objective functions 1 and 2, respectively.

The determination of total discharge, while ensuring adequate R-A exchanges at the commune level, revealed significant differences between the old pareto front generated using distance-based clustering and the new pareto front obtained through aquifer parameter-based clustering. In particular, Chazey-sur-Ain, located near the Ain River,

exhibited a substantial increase in discharge (12659.13 m^3/d). Conversely, the communes near the study area's boundary, namely Douvres, and Jujurieux, experienced only marginal increases in discharge (500 m^3/d and 130.33 m^3/d, respectively). However, as we move toward the periphery of the pareto front, the differences in discharge become more pronounced. Chazey-sur-Ain, for instance, demonstrated significantly higher discharge values (34000 m^3/d and 14031.68 m^3/d), while the boundary communes of Douvres (2000 m3/d) and Jujurieux (4000 m3/d) also exhibited notable variations. The pareto points and the discharge difference maps are depicted in Figure 7.7.

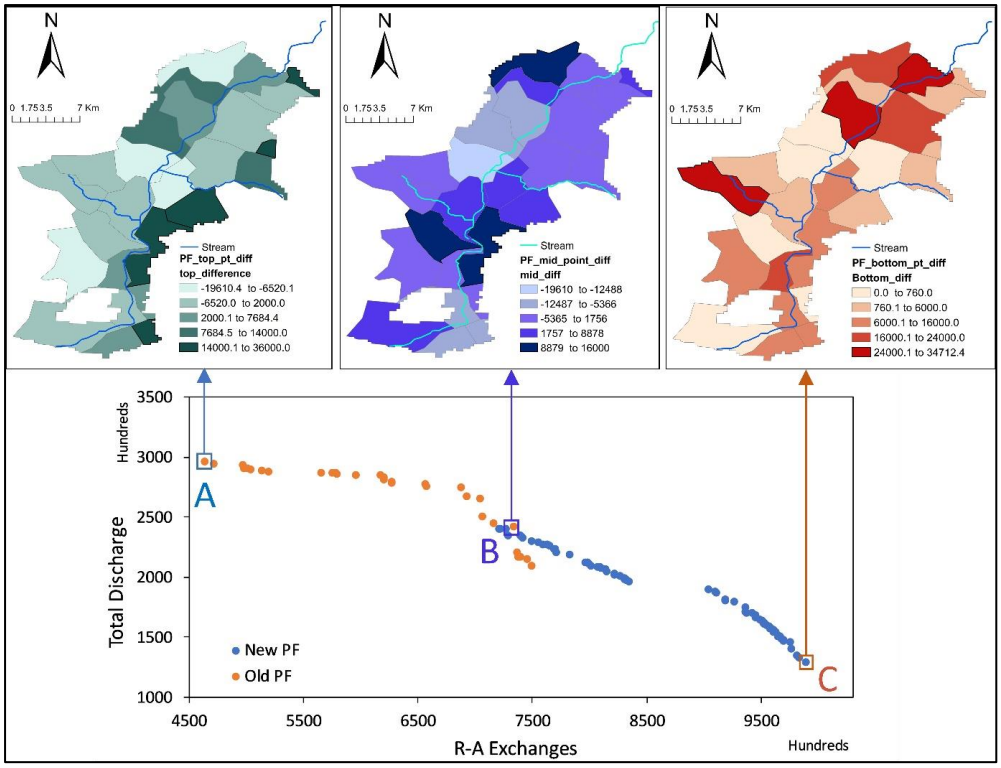


Figure 7.7 Differences in total discharge of all communes with and without aquifer parameter-based clustering.

Notably, the discharge results at the commune level further highlight the value of aquifer parameter-based clustering. The results are shown in Table 7.2, and for communes located near the Ain River, such as Chazey Sur Ain, the clustering approach yielded a substantial increase in groundwater discharge (12659.13 m^3/d). In contrast, communes near the study

area's boundary, such as Douvres and Jujurieux, observed only marginal increases in discharge (500 m³/d and 130.33 m³/d, respectively).

Table 7.2 The commune-wise discharge is calculated before and after parameter clustering.

	Top		Mid-point		Bottom	
Commune	Total discharge OLD	Total Q new	Total discharge OLD	Total Q new	Total discharge OLD	Total Q new
Ambérieu-en-Bugey	4500	18500	10500	10500	10500	12500
Ambronay	7731.146	9500	23177.66	23000	31000	55000
Blyes	1000	5000	5598.169	2015.623	10000	10000
Charnoz-sur-Ain	9311.457	13000	15009.22	15000	15000	35000
Château-Gaillard	10500	10500	51000	51000	54908.3	55000
Châtillon-la-Palud	32812.12	23000	35000	15390.02	35000	35000
Chazey-sur-Ain	15000	51000	32340.87	45000	30968.32	45000
Crans	500	500	9886.66	10500	9783.88	12500
Douvres	22001.84	52500	500	500	500	2500
Druillat	8535.711	2015.6	4500	18500	4500	22500
Jujurieux	11000	45000	1000	5000	1000	5000
Leyment	21939.79	52500	47221.64	52500	1000	5000
Loyettes	6569.58	14254	20000	14000	36500	52500
Meximieux	4569.58	4254	14000	30000	20000	20000
Pérouges	35000	15389	24645.91	25500	14000	30000
Pont-d'Ain	4862.619	10500	8376.08	9500	9500	44212
Priay	3500	9500	37874.8	31250	8946.66	37500
Rignieux-le-Franc	15779.11	15000	14515.22	11500	10493.66	42500
Saint-Jean-de-Niost	30000	30000	13397.25	4023.4	49000	62413
Saint-Jean-le-Vieux	25655.82	25500	17214.69	13000	11500	17500
Saint-Maurice-de-Gourdans	11509.65	11500	53825.78	62500	20000	20000
Saint-Maurice-de-Rémens	15516.62	4192.769	49788.53	52500	23374.96	35000
Saint-Vulbas	32500	62500	9849.669	4000	52500	62500
Varambon	7500	17500	18500	22500	50500	52500
Villette-sur-Ain	3500	17500	17500	9500	9239.95	10000
Villieu-Loyes-Mollon	8500	10500	12776.73	17500	18500	22500

Future research should explore the integration of sensitivity analysis into the clustering process, potentially offering a more refined approach to parameter selection. Additionally, while this study focused on groundwater extraction and river-aquifer exchanges, future work could expand the scope to include other critical groundwater management objectives, such as contaminant transport and water quality preservation. Incorporating these additional considerations into the clustering framework could provide a more comprehensive tool for sustainable groundwater management.

Aquifer parameter-based clustering proves to be a suitable approach for effectively achieving overall accuracy. Therefore, conducting a sensitivity analysis on hydrogeological properties holds both fundamental and practical significance in solving multi-objective GSOPs for groundwater management strategies.

7.5 Summary

The chapter's findings highlight the efficacy of parameter-based clustering techniques in reducing large DV in a GSOP while achieving management objectives at the administrative level. First, distance-based clustering was adopted to determine an optimal number of clusters in this GSOP. Three clustering algorithms (Ward, K-means, and Affinity Propagation) were applied to ensure the consistency of the result. In the second phase, aquifer parameters such as top layer elevation, recharge, hydraulic conductivity, and initial starting head were utilized for clustering, and their accuracy was tested. The GSOP was solved using the MOPSO algorithm. The resulting pareto fronts were evaluated based on hypervolume, epsilon, spread, and inverted generational distance to assess their diversity, convergence, and spread. Furthermore, a comparison between distance-based clustering and aquifer property-based clustering was presented. The chapter concludes that clustering algorithms offer a simple and robust method for reducing decision variables in large-scale groundwater optimization problems. At last, the total optimal discharge rate for each

commune was calculated and it was found to be much higher than the distance-based clustering pareto front. Addressing these high-dimensional GSOPs is vital for effectively managing groundwater resources and bridging the gap between sub-basin or hydrologic response unit (HRU) scale conceptualization and administrative-level planning and decision-making processes. Finally, the findings from this study are not limited to the Ain River basin, and it can be extended to other GSOPs that are highly vulnerable to converge because of large decision variables.