PROCEEDINGS, 43rd Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 12-14, 2018
SGP-TR-213

Reservoir Characterization and Prediction Modeling Using Statistical Techniques

Halldora Gudmundsdottir and Roland N. Horne

Department of Energy Resources Engineering, Stanford University, Stanford, CA 94305

halldora@stanford.edu

Keywords: Reservoir characterization, interwell connectivity, clustering, tracer, regression, direct forecasting

ABSTRACT

Reservoir characterization and prediction modeling have long been among the more challenging tasks in geothermal reservoir engineering. The main reason is the presence of fractures and faults, which control the mass and heat transport in the subsurface. In this work, the applicability of using statistical methods for reservoir characterization as well as prediction modeling was explored. Three methods were analyzed and applied on a synthetic library of fracture networks. First, the Alternate Conditional Expectation (ACE) algorithm was used to estimate well-to-well connectivity between injection and production wells using tracer return and temperature data. The results obtained with tracer data were in good agreement with tracer transit times, for 80.5% of the fracture networks the ACE connectivity was within ±0.05 of the connectivity implied by transit time, while temperature data showed much less correlation to connectivity with the ratio reduced to 58.3%. Second, k-means clustering was applied where fractures of similar character were grouped together and interwell connectivity and thermal behavior estimated. The method displayed potential but the main limitations were deciding on the number of clusters and the growing complexity with added producers. Third, preliminary results using direct forecasting with Canonical Functional Component Analysis (CFCA) were presented. A significant reduction in the predicted range of thermal responses for production wells was obtained but introducing more complex data is likely to cause data-prediction relationships to become less linear.

1. INTRODUCTION

An essential part of a sustainable exploitation from geothermal reservoirs is the ability to understand and predict the reservoirs' response, such as premature cooling and pressure drawdown. This commonly involves characterizing the reservoirs and building geological models. Characterizing of fractures and flow paths is a challenging task, fractures control the mass and heat transport in the subsurface and the details of their location and orientation are often uncertain. Nowadays, with a substantial increase in data due to advances in computer power and measuring equipment, the oil and gas as well as the geothermal industries are presented with some of today's most complex data science problems. Therefore, statistical methods are becoming increasingly popular as tools for predictive analysis in the exploration, production and delivery phases.

Characterization of fractured reservoirs has been the subject of numerous papers. Production and injection rates, which are among the most readily available data, have been used to infer well-to-well connectivity to optimize injection schedules. Heffer et al. (1997) calculated the Spearman's rank correlation coefficient between oil production rates, reflecting communication through the reservoir. Refunjol and Lake (1999) applied the Spearman's analysis in combination with geological features and tracer data to determine preferential flow paths in oil reservoirs. Tian and Horne (2016) proposed a new method, the modified Pearson's correlation coefficient, to calculate correlation between injection and production wells and compared to the Spearman's analysis, highlighting the method's advantages in identifying faults within the reservoir. Other approaches include the capacitance-resistance model, described for example by Yousef et al. (2006). The model was based on governing material-balance equations at reservoir conditions and used to estimate two coefficients, one for the degree of connectivity and one for the amount of fluid stored between the wells. Recently, due to their nonlinear capabilities, attention has been given to Artificial Neural Networks (ANN). Demiryurek et al. (2008) generated an ANN with injection rates and production data and applied sensitivity analysis to quantify interwell connectivity. Sarkheil et al. (2009) presented a method of predicting natural fracture distribution between two wells using ANN, based on image logs and core measurements. Furthermore, Alizadeh et al. (2015) used ANNs to predict the fractures dip inclination degree of a well based on image log and other geological log data from two other wells nearby.

Traditionally, the geothermal industry has sought inspiration to applications used within the oil and gas industry due to their similar nature as well as longer and more extensive history of research. Sullera (1998) isolated short-term variations in produced chloride concentration and injection rate using Wavelet transformations. These short-term variations were then used in multiple regressions to quantify connections between injectors and producers. Horne and Szucs (2007) further developed this idea by using nonparametric regression on the same data set without making assumptions about the underlying form of the relationships. Juliusson (2012) developed a novel method for characterizing fractured reservoirs using flow rate and tracer data. The flow rate data were used to calculate an interwell connectivity matrix, describing how injected fluid is divided between producers. A method referred to as M-ARX was used to compute the interwell connectivity, but the main advantage of the method was that producer-producer interactions could be estimated as well as injector-producer interactions.

In reservoir engineering, statistical methods for directly predicting the reservoir response without inferring about its character are gaining popularity. Liu and Horne (2013) proposed to use a convolutional kernel based algorithm to forecast pressure response in oil wells based on flow rate histories. They applied the method successfully on multiple synthetic and real field cases, recovering the underlying reservoir

models. Tian and Horne (2015) formulated this approach as a feature-based linear regression, leading to a closed solution and thus reducing the computational cost dramatically. Additionally, Tian and Horne (2017) applied a Recurrent Neural Network (RNN) on synthetic and real flow rate and pressure data sets, showing the RNN's ability to forecast the reservoir performance without making assumptions about the physical model.

In this work, several statistical methods for reservoir characterization and predictions were analyzed and applied on a synthetic dataset of two-dimensional fracture networks representing geothermal reservoirs. The alternate conditional expectation (ACE) algorithm was used to infer well-to-well connectivity with production data, namely flow rates, tracer returns and temperature. The possibility of using clustering for reservoir characterization was studied. Fracture networks of similar character were grouped together and from each group interwell connectivity and temperature response ranges estimated. Lastly, the concept of direct forecasting in geothermal settings was introduced and the benefits and limitations of statistical methods investigated.

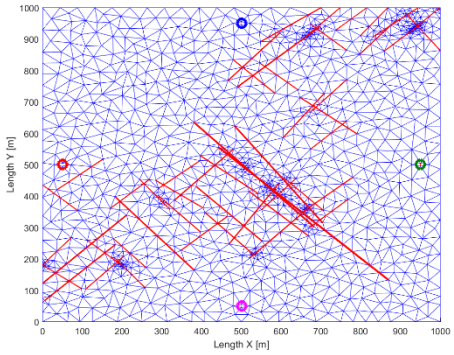
2. METHODOLOGY

2.1 Library of Fracture Networks

A library of two-dimensional fracture networks was used for this work. The library was generated by Magnusdottir (2013) and included different realizations of a geothermal reservoir. Of the 1200 networks available, 800 were used for this study. The discrete fracture networks were generated from a distribution of several parameters such as fractal dimension, orientation and maximum length. The initial temperature of the reservoir was defined at 200°C and injection temperature at 100°C. Fluid was injected at 10 kg/s over 1000 days or approximately 3 years. A detailed description of the design and generation of the fracture networks can be found in Magnusdottir (2013). A summary of main reservoir parameters and an example a fracture network from the library are shown in Table 1.

Table 1: Main reservoir parameters and an example of a fracture network. The red lines indicate fractures, blue triangle simulation mesh and colored circles wells (blue: Injector, red: Producer 1, green: Producer 2 and pink: Producer 3)

	SI-Units
Dimension (xyz)	1000x1000x1 m^3
Spatial fractal dimension (D)	1.0-1.6
Fracture orientation (θ_f)	45°-135°
Maximum fracture length (L)	600 m
Fracture aperture (w)	0.002x$L^{0.4}$ m
Fracture porosity (φ_f)	0.9
Fracture permeability (k_f)	$w^2/12$ m^2
Matrix porosity (φ_m)	0.1
Matrix permeability (k_m)	10^{-10} m^2



2.2 The ACE Algorithm

Regression analysis is a statistical process for estimating the relationship between variables. More specifically, its objective is to explain or reveal the effect of one or more independent variables (predictors) on a dependent variable (response). Traditionally, multiple regression techniques require making assumptions about the functional form of the relationship between response and predictor variables. In practice, this relationship is commonly unknown, which can lead to inaccurate assumptions causing erroneous and misleading results. The Alternating Conditional Expectations (ACE) algorithm is a nonparametric regression method that seeks to build a model, fitting the data without assuming an underlying form of the relationship between variables.

The ACE algorithm was developed by Breiman and Friedman (1985) and its objective is to find transformations $\varphi_1, \dots, \varphi_p$ of the predictor variables $X_1, \dots, X_p$ and a transformation θ of the response variable Y that maximizes the correlation between $\theta(Y)$ and $\varphi_1, \dots, \sum_{i=1}^{p} \varphi_i(X_i)$. The general form of the ACE regression model is written as:

$$\theta(Y) = \sum_{i=1}^{p} \varphi_i(X_i) + \epsilon \tag{1}$$

where ϵ is the error that is not explained by the regression. The error is related to the maximum correlation coefficient by $\epsilon^2 = 1 - \rho^2$. Therefore, maximizing the correlation is equivalent to minimizing the error variance (under the constraint that $E[\theta^2(Y)] = 1$):

$$\epsilon^2(\theta, \varphi_1, \dots, \varphi_p) = E[(\theta(Y) - \sum_{i=1}^{p} \varphi_i(X_i))^2] \tag{2}$$

In order to minimize the error, optimal transformations θ^* and φ^* are defined. The ACE algorithm gives estimates of these transformations by the following iterative process.

1. Set $\theta(Y) = Y/\|Y\|$
2. Compute new φ with $\varphi(X) = E[\theta(Y)|X]$. By defining $\varphi(X)$ as the conditional expectation of $\theta(Y)$ given X, $\varphi(X)$ becomes the function of X most correlated with $\theta(Y)$.
3. Compute new θ with $\theta(Y) = E[\varphi(X)|Y]/\|E[\varphi(X)|y]\|$. To fulfill the constraint of $E[\theta(Y)] = 1$, the new $\theta(Y)$ needs to be divided with the standard deviation.
4. Repeat Step 2 and 3 until ϵ^2 fails to decrease.
5. Now, θ and φ are the solutions to θ^* and φ^*.

As can be seen, the procedure involves *alternating conditional expectations,* hence the name of the algorithm (ACE). For the multiple predictor case in Eq. (2), the transformations of $\varphi(X)$ and $\theta(Y)$ can be expressed as:

$$\varphi_i(X_i) = E[\theta(Y) - \sum_{j\neq i}^{p} \varphi_j(X_j)|X_i] \tag{3}$$

$$\theta(Y) = E[\sum_{i=1}^{p} \varphi_i(X_i)|Y]/||E[\sum_{i=1}^{p} \varphi_i(X_i)|Y]|| \tag{4}$$

After the iterations, the final transformation functions $\varphi_1(X_1), \dots, \varphi_p(X_p)$ and $\theta(Y)$ are assumed to be the optimal transformations $\varphi_1^*(X_1), \dots, \varphi_p^*(X_p)$ and $\theta^*(Y)$. Finally, the predictor and response variables are related in the transformation space through:

$$\theta^*(Y) = \sum_{i=1}^{p} \varphi_i^*(X_i) + \epsilon^* \tag{5}$$

where ϵ^* is the minimization error not captured by the ACE transformations and is assumed to have a normal distribution with zero mean (Nguyen-Cong and Rode, 1995; Wang and Murphy, 2004).

2.3 K-means Clustering

Clustering involves finding subgroups in a dataset where observations within a subgroup are quite similar to each other, while observations in different groups are quite dissimilar from each other. Numerous clustering methods exist but one of the most popular is k-means clustering, which is a simple algorithm where observations are partitioned into prespecified number of clusters. Let $C_1, \dots, C_K$ denote sets containing the indices of the observations in each cluster. These sets must fulfill:

$$C_1 \cup C_2 \cup \dots \cup C_k = \{1, \dots, n\} \quad \text{and} \quad C_k \cap C_{k'} = \emptyset \text{ for all } k \neq k' \tag{6}$$

or in other words, each observation belongs to one and only one cluster. The goal of clustering is to minimize variation within a cluster as much as possible. That can be done by solving the following optimization problem:

$$\underset{C_1,\dots,C_K}{\text{maximize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\} \tag{7}$$

where $W(C_k)$ is a measure of distance between observations. There are many ways to define a distance measure but the most commonly used is the *Euclidian distance*, expressed as:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i'\in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \tag{8}$$

An approximation to the optimization problem in Eq. (7) is found with the k-means algorithm by an iterative process. First each observation is assigned an initial cluster number from $1, \dots, K$. For each cluster K, the cluster centroid is computed. Now each observation is assigned to a new cluster, whose centroid is closest, in terms of the Euclidian distance, to the observation. This procedure is repeated until cluster centroids stop shifting and cluster assignments stop changing. Because the k-means algorithm finds a local minimum it is important to run the algorithm multiple times with different initial conditions (James et al., 2013).

2.4 Canonical Functional Component Analysis (CFCA)

Canonical Functional Component Analysis (CFCA) involves building a statistical relationship between historical and forecasting data which can then be used along with observed data to make predictions. Here, a brief overview is given but a more detailed description of the procedure can be found in Satija et al. (2015, 2017) and Li (2017). Let d represent data variables, sometimes referred to as the historical data, and m the model parameters, such as porosity or permeability, of the subsurface model g. Similarly, let h represent prediction variables, or the forecasting data, and r another unknown forward operator of the subsurface model. Data variables, in the form of temperature and tracer concentration, at three different production well locations, illustrated in Table 1, are calculated after x days of injection as:

$$d_i(t) = \left[d_{P1,i}(t);\ d_{P2,i}(t);\ d_{P3,i}(t)\right] = g(m_i) \tag{9}$$

where i $= 1, \ldots, 800$ represents the number of models, or more specifically fracture networks in the library. In similar fashion, prediction variables are calculated after y days of injection, where y $>$ x, at the same locations as:

$$h_i(t) = \left[h_{P1,i}(t);\ h_{P2,i}(t);\ h_{P3,i}(t)\right] = r(m_i) \tag{10}$$

Applying Functional Component Analysis (FCA), which presents time-series as linear combinations of basis functions, on the data and prediction variables results in:

$$d(t) \cong \sum_{i=1}^{K} d_i^f \vartheta_{d,i}(t) \qquad h(t) \cong \sum_{i=1}^{K} h_i^f \vartheta_{h,i}(t) \tag{11}$$

where d^f and h^f are the functional components, representing the time series in a K-dimensional functional space. Significant reduction in dimensions can be obtained with FCA but the relationship between data and prediction variables may still be complex and nonlinear. Therefore, Canonical Component Analysis (CCA) is used to maximize the correlation between the nonlinearly related variables by applying the following transforms:

$$d^c = A^T d^f \qquad h^c = B^T h^f \tag{12}$$

where A and B are the canonical variates of d^f and h^f. As well as maximizing the correlation, CCA constrains all intercomponent correlations to be zero. To apply CFCA on the library of fracture networks, networks are chosen from the library to serve as observed data sets $d_{obs} = \left[d_{obs,P1};\ d_{obs,P2};\ d_{obs,P3}\right]$ and predictions made on those. If CCA results in a linear relationship, linear Gaussian regression can be used to forecast h directly from d. First, to make sure h^f follows a Gaussian distribution, simple normal score transformation is applied. Then the posterior distribution is formed, which is constrained to the canonical form of the observations $f(h^c|d_{obs}^c)$. This distribution is sampled and the samples are then back transformed into time-series.

3. RESULTS AND ANALYSIS

3.1 Inferring Well Connectivity Using ACE

Geothermal reservoirs are normally highly fractured with short circuit paths controlling the fluid flow, potentially leading reinjected fluid to producers without optimal thermal sweep of the reservoir. Using chloride concentration of produced fluid as a response variable and injection histories as predictor variables has proven successful in estimating well-to-well connectivity in actual fields as shown by Horne and Szucs (2007). In this study, other types of response data were investigated and their ability, in conjunction with nonparametric regression, to estimate meaningful well connectivity indices. These responses are namely tracer and temperature histories in producers due to injection.

Water at 100°C with NaCl concentration equal to 22 wt% was injected into the reservoir which was at 200°C and production wells modeled to deliver against fixed bottom hole pressure. Due to the injection, the temperature at the production wells decreased over time with a slope depending on the connection between the injector and producers. In this case, the subject of investigation was the effect of a constant injection at one location on three production wells located within the reservoir boundaries. Therefore, for the ACE regression technique, the constant injection rate is defined as the response variable Y and the resulting signals at the three producers as predictor variables. The ACE algorithm is applied on the library, one fracture network a time, and the results then normalized where values close to 1 are believed to exemplify strong connections and values close to 0 weaker ones.

First, the temperature response at the three producers due to injection over 1000 days was analyzed. The well-to-well connectivity indices obtained with the ACE algorithm for all fracture networks can be seen in Figure 1. Three temperature curves are presented for each fracture network, Producer 1, Producer 2 and Producer 3, which are colored according to the strength of the connection. A trend can be seen for Producer 1 and Producer 2, temperature curves with high connectivity (in red) drop faster than temperature curves with low connectivity (in blue). This is consistent with the physical processes taking place because for highly connected paths, injected fluid travels faster towards producers resulting in more rapid cooling. For Producer 3, a pattern is present in the data, but the situation is reversed, slow cooling curves are characterized with high connectivity. Furthermore, for all producers there is some overlap in the data which could indicate that there is not a clear correlation between temperature and connectivity.

Investigating tracer concentration at the three producers due to injection over 1000 days yields different results. The well-to-well connectivity indices from the nonparametric regression, using the tracer data, can be seen in the left column of Figure 2. These connectivity indices are also used to label corresponding temperature curves of each producer, illustrated in the right column of Figure 2. Note that for clearer presentation, tracer curves are only plotted up to 200 days even though data up to 1000 days is used for the regression. In this case, a clear trend in tracer curves is present. Rapid decrease in tracer concentrations correspond to high connectivity flow paths. Furthermore, little to no overlap of colored curves can be seen, indicating that using tracer data could potentially yield meaningful well connectivity indices. The corresponding temperature curves, labeled with connectivity indices obtained using tracer data, are also in good accordance with physical processes.

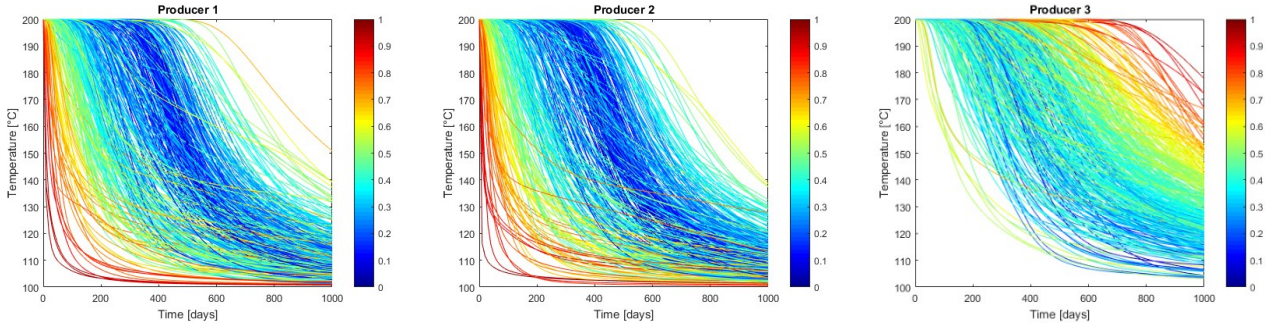


Figure 1: Temperature response curves at each producer for the library of fracture networks. Each curve is colored according the well-to-well connectivity indices calculated with ACE and temperature data.

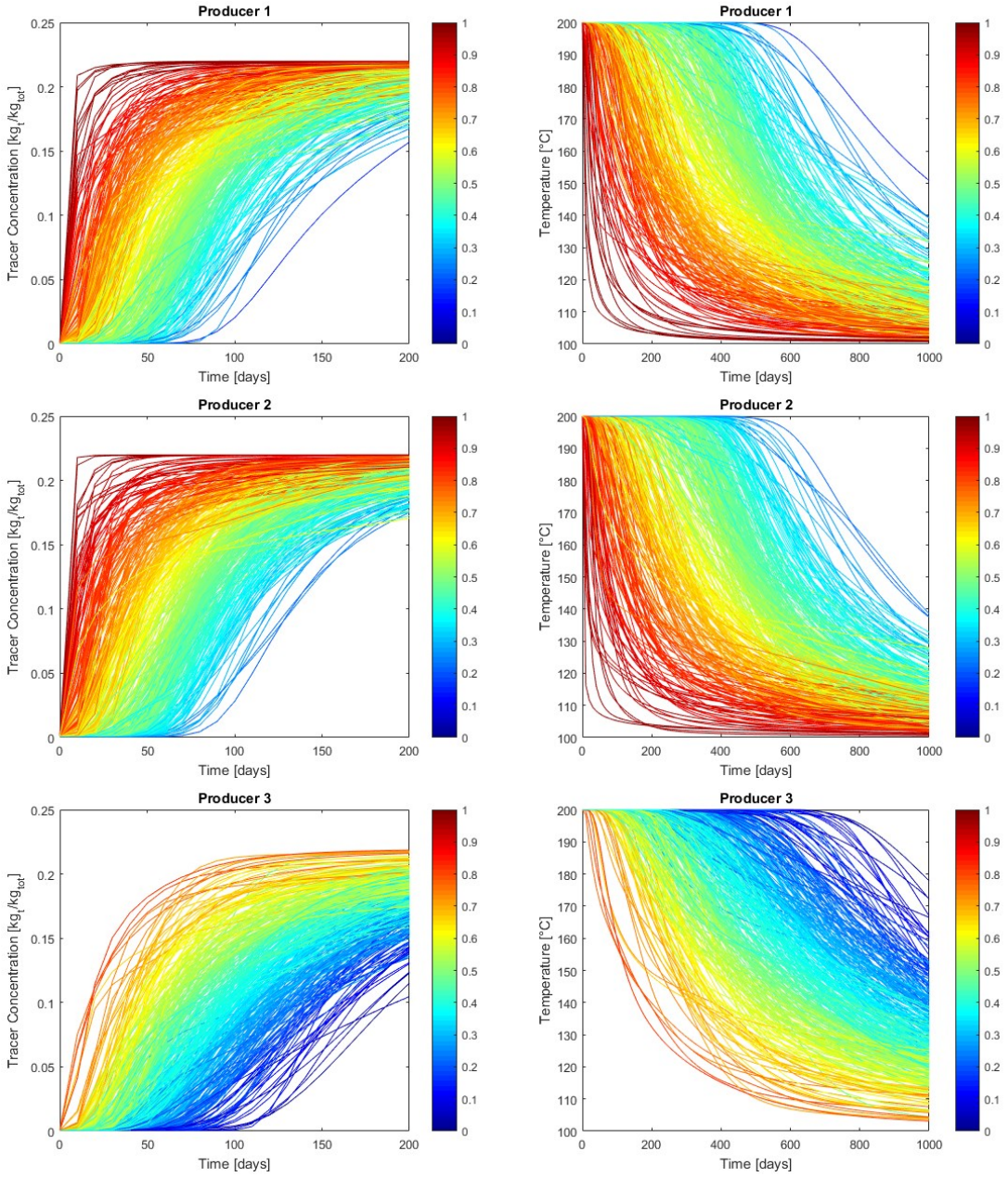


Figure 2: Tracer curves, colored with connectivity indices obtained using tracer data (left column). Corresponding temperature curves, colored with the same connectivity indices (right column). Note that tracer data is only plotted up to 200 days.

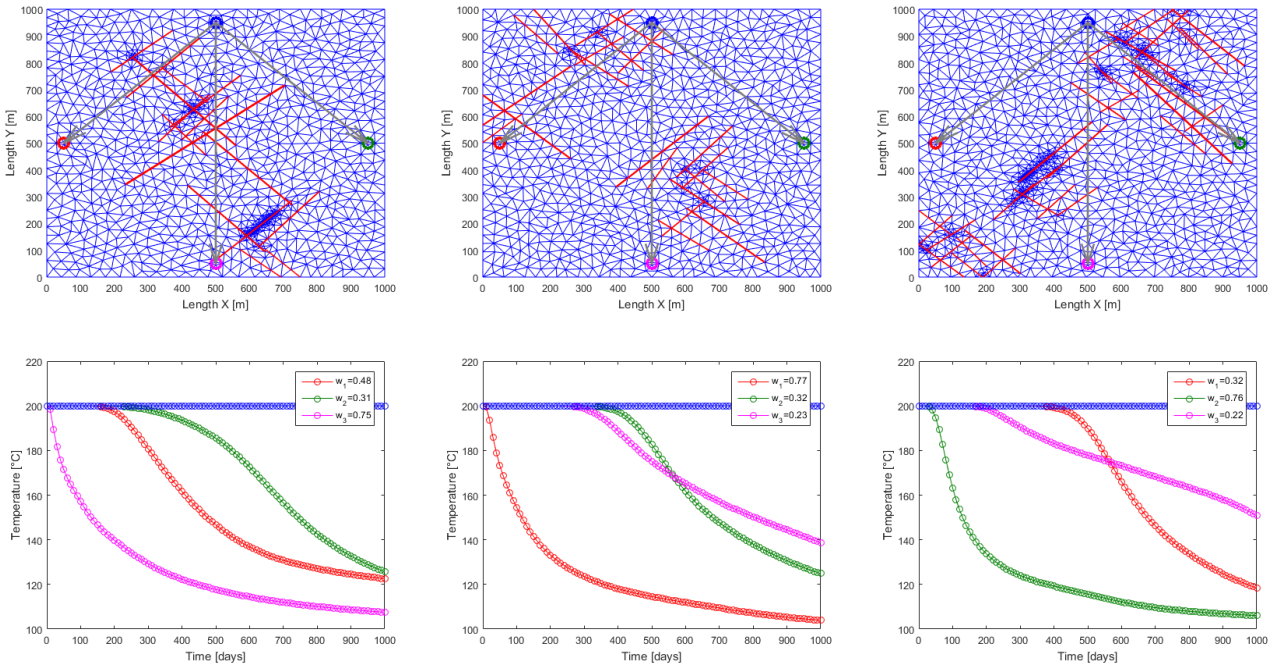


Figure 3: Examples of fracture networks (Network 1, 2 and 3) with grey arrows indicated direction of flow from the injector to the producers (top row). Corresponding temperature response curves for Network 1, 2 and 3 colored with well type along with well-to-well connectivity value between injector and each producer (bottom row).

To quantify if the interwell connectivity calculated with ACE can be related to the physical fluid motions in the subsurface, a comparison to tracer tests was made. For the library of fracture networks, both the thermal behavior as well as the chemical front due to injection were simulated, and the first arrival or transit time of the injected fluid noted. Quick arrival represents strong connectivity and thus the inverse of the transit time is used as an indicator of the connectivity. To compare the ACE connectivity indices to the ones obtained from tracer transit times, the indices were normalized so that they sum to one for each individual fracture network. For 80.5% of the networks in the library, the ACE connectivity was within ±0.05 of the tracer connectivity when the nonparametric regression was performed with tracer data. Furthermore, for 54.5% of the networks the strength of connection was correctly ordered (e.g. P2: highest, P1: lower, P3: lowest). Using temperature data instead of tracer data, these ratios were reduced to 58.3% and 18.8%, demonstrating the greater predictive abilities of tracers.

3.2 Reservoir Characterization Using Clustering

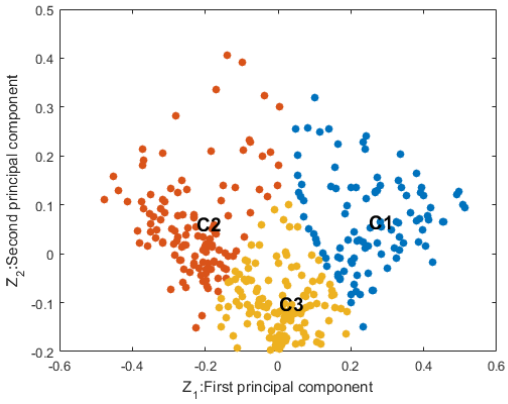
Many types of fracture networks are found in geothermal reservoirs. With exploration methods, geophysicists are able to map major faults and fractures, but the detailed character of the networks is not easily interpreted. With the true nature of reservoirs usually unknown, the exploratory capabilities of unsupervised learning could potentially be helpful in reservoir characterization. Here we discuss the possibility of using clustering to group fracture networks and thus infer about the character of geothermal reservoirs as well as their thermal behavior.

To simplify the examples, only response values from Producer 1 (P1) and Producer 2 (P2) were used and values from Producer 3 (P3) excluded. This is equal to dividing the fracture networks shown in Table 1 horizontally in half, considering exclusively the upper half. The response of the producers due to injection are in the form of time series. This can add complexity because typical machine learning methods assume that the data are independent and identically distributed, which is not the case for time series data. Therefore, Principal Component Analysis (PCA) was performed on the time series to produce independent principal components and reduce the dimensionality of the problem. For this library of fracture networks, the first two principal components were found to explain 85% of the total variance in the data.

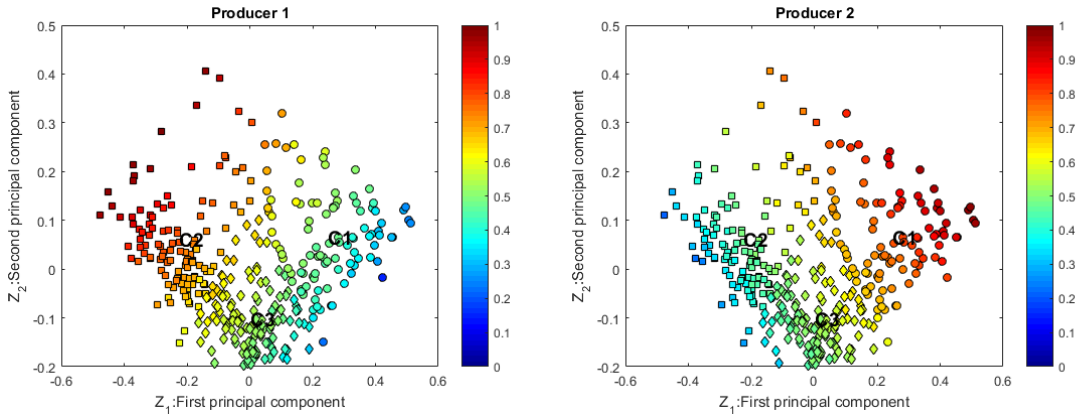
The tracer return curves from P1 and P2 were used as inputs to the k-means clustering algorithm and for robustness, ten-fold cross validation was performed. This procedure involves dividing the data set randomly into ten groups of roughly equal size. The first group is treated as a validation set, while the clustering algorithm is used on the other nine parts. Then this is repeated for all groups. Robustness comes from the fact that slightly different datasets are being used for the clustering and therefore in the end ten slightly different clustering schemes will be obtained. For the final product, the average is taken to represent the P10-P50-P90 quantiles of the temperature curves for each producer.

Results from k-means clustering on tracer data from P1 and P2 using three clusters are shown in Figure 4. In Figure 4(a), the two principal components of the data are plotted along with a cluster label where each point represents a fracture network. Figure 4(b) has the same configuration as Figure 4(a) but with each cluster having a different marker: Cluster 1 is a circle, Cluster 2 a square and Cluster 3 a diamond. The markers are further colored according to the well connectivity indices of P1 and P2, obtained with the ACE algorithm in

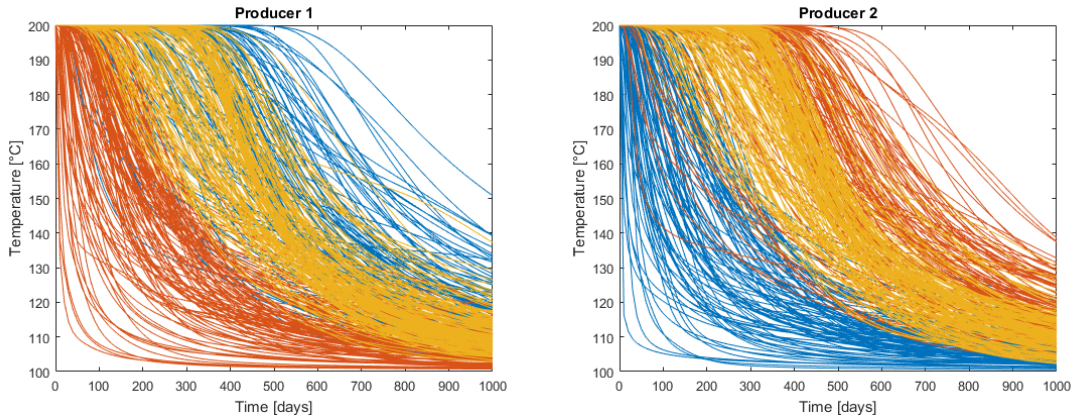
the previous section. Lastly, in Figure 4(c), the temperature curves for each producer are plotted and curves colored to identify to which cluster they belong. Taking a closer look at Figure 4(b), for networks within Cluster 1, P1 has rather low well connectivity while P2 is highly connected to the producer. For Cluster 2, this is reversed, P1 has high connectivity and P2 low connectivity, and for Cluster 3 both producers have intermediate connectivity. This is further shown in Figure 4(c), where for Cluster 1 temperature drops slowly in P1 but rapidly in P2, for Cluster 2 this is reversed and for Cluster 3 the decrease in temperature is similar for both producers.



(a) Cluster 1 (blue), Cluster 2 (red) and Cluster 3 (yellow).



(b) Cluster 1 (circle), Cluster 2 (square) and Cluster 3 (diamond), colored with well connectivity indices.



(c) Temperature curves for Producer 1 and 2; Cluster 1 (blue), Cluster 2 (red) and Cluster 3 (yellow).

Figure 4: Results from k-means clustering with k = 3, using tracer data from P1 and P2.

For comparison, analyses of Networks 1, 2 and 3 from the previous section are shown. The networks were assigned to a cluster according to the smallest distance between the networks and the cluster centroids. The results are illustrated in Figure 5: Cluster 1 (blue), Cluster 2

(red) and Cluster 3 (yellow) using tracer data along with temperature curves from Network 1, 2, and 3 (in pink)., Network 1 falls within Cluster 3, Network 2 within Cluster 2 and Network 3 within Cluster 1. Table 2 lists the actual temperature of each network after 1000 days of simulation, along with temperature boundaries found by clustering. Now, instead of having only one best fit to the 'true' reservoir, a range of networks has been found. The widths of these temperature boundaries range from 10°C to 25°C. Temperature curves of Network 2 and 3 fall within the boundaries of their clusters, while Network 1 is harder to match. The poor match to Network 1 can be explained partly by the limited number of fracture networks in the library, as there are not enough networks with similar behavior. Adding more producers increases the complexity and more clusters are needed to describe the data. Another example is considered where curves from all producers, P1, P2 and P3, are clustered into five groups. Results are reported in Table 3.

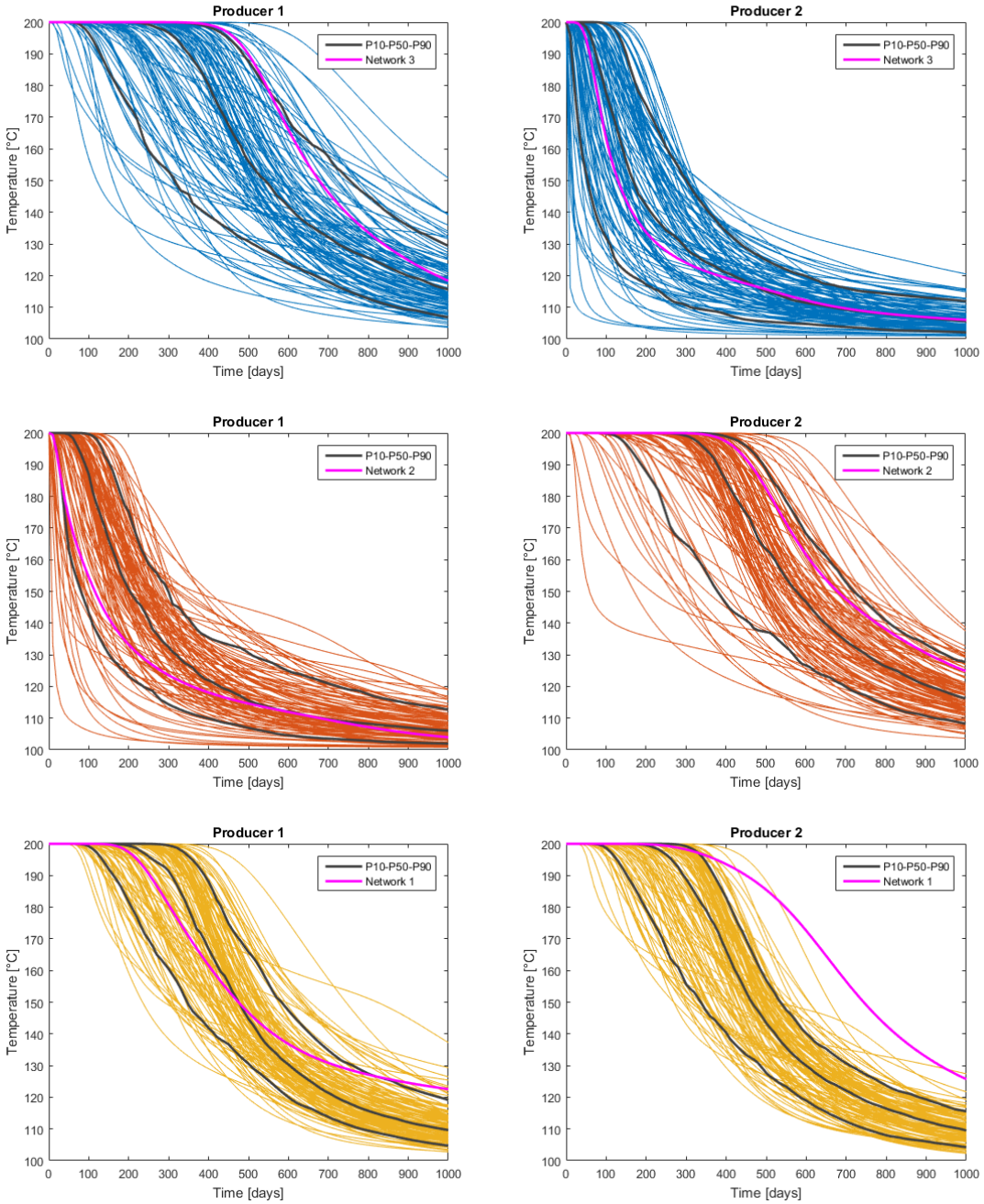


Figure 5: Cluster 1 (blue), Cluster 2 (red) and Cluster 3 (yellow) using tracer data along with temperature curves from Network 1, 2, and 3 (in pink).

Table 2: The actual temperature of Network 1, 2 and 3 after 1000 days of injection along with temperature ranges found with k-means clustering (using P1 and P2).

Actual temperature	Network 1	Network 2	Network 3
Producer 1	123°C	104°C	118°C
Producer 2	126°C	125°C	106°C
k-means temperature, k = 3	**Network 1**	**Network 2**	**Network 3**
Producer 1	105-118°C	102-112°C	107-131°C
Producer 2	104-116°C	108-128°C	102-112°C

Table 3: The actual temperature of Network 1, 2 and 3 after 1000 days of injection along with temperature ranges found with k-means clustering (using P1, P2 and P3).

Actual temperature	Network 1	Network 2	Network 3
Producer 1	123°C	104°C	118°C
Producer 2	126°C	125°C	106°C
Producer 3	108°C	139°C	151°C
k-means temperature, k = 5	**Network 1**	**Network 2**	**Network 3**
Producer 1	107-124°C	102-111°C	109-132°C
Producer 2	106-123°C	109-129°C	102-111°C
Producer 3	106-121°C	118-144°C	118-144°C

The greatest challenge in cluster analysis is to determine the optimal number of clusters. The choice of k often seems highly subjective to the desired clustering resolution of the user. Increasing k will always reduce the error in the resulting clustering to the point that each data point becomes its own cluster. However, increasing k without thought, can cause overfitting problems where the generalization of the clustering model reduces. One way to validate the number of clusters is the elbow method. In short, the method involves applying k-means clustering on a wide range of values of k, and for each k calculate the sum of squared errors. Plotting the sum of squared error as a function of k, the curve should look like an arm, where the "best" value of k is the elbow. This method does not always work well, especially if the data are not particularly clustered. Another method is the so-called gap statistics, where the user finds a standardized comparison of $\log W_k$ with a null reference distribution of the data, that is a distribution with no obvious clustering. The elbow method and the gap statistics were applied on the dataset. When using only P1 and P2 the methods gave an optimum around 3-4 clusters but for the extended case, P1, P2 and P3, no distinct optimum was found. This can be due to several reasons. First, the data curves are of similar shapes, resulting in unclear clustering structure where clusters are very close to each other. Second, the limited number of fracture networks used can be problematic.

3.3 Direct forecasting using CFCA

The goal of developing a model of a geothermal field is not necessarily to obtain the model parameters themselves but to attain forecasts made by the model along with uncertainty quantification. A Prediction Focused Approach (PFA) has been suggested by Scheidt et al. (2014) and Satija et al. (2015) that builds a statistical relationship between historical and forecasting data that could help avoid full model inversion. Satija et al. (2015) proposed a Canonical Functional Component Analysis (CFCA) that in short consists of: (1) reducing the dimension of time-series data; (2) building a linear relationship between historical and forecasting data; and (3) using regression to quantify forecast uncertainty.

Here, the possibility to make predictions about thermal behavior of three production wells due to cold fluid injection was investigated. These predictions were based on historical data either in the form of tracer concentration or temperature at the producers. To make reliable predictions, good correlation needs to be established between data variables (historical data) and prediction variables (forecasting data). If poor correlation exists between the data, predictions often have no physical meaning and turn out to be wrong. For this study, tracer return curves up to 500 days at the three producers were used as historical data and temperature curves from 500 to 1000 days as forecasting data. Firstly, tracer concentration at P1 was used to predict the producer's temperature at later time. The data used for the CFCA and corresponding results are illustrated in Figure 6. The data-prediction relationship is highly linear which allows for the linear Gaussian regression to be used to sample prediction curves (posterior) from a Gaussian distribution. Comparing P10-P90 quantiles of the forecasting data to the P10-P90 quantiles of the predictions, a great uncertainty reduction may be obtained. Similar prediction results were obtained for P2 and P3, shown in Figure 7.

Additionally, an exploration whether correlation exists between data of different wells was undertaken, for example to understand if data from P1 can be used to predict the response of P2 and so on. Figure 8 shows one these scenarios, where data from P1 and P3 is used to predict the response of P2. The data-prediction relationship becomes less linear but is still high enough for the prediction to be meaningful. Furthermore, the P10-P90 quantiles of the forecasting data increased compared to using data from P2 to predict on itself. The same analysis was performed using temperature data as both data and prediction variables which resulted in a lower data-prediction correlation in all

cases. The result is not surprising since the chemical front travels orders of magnitude faster than the thermal front, granting tracers their predictive capabilities.

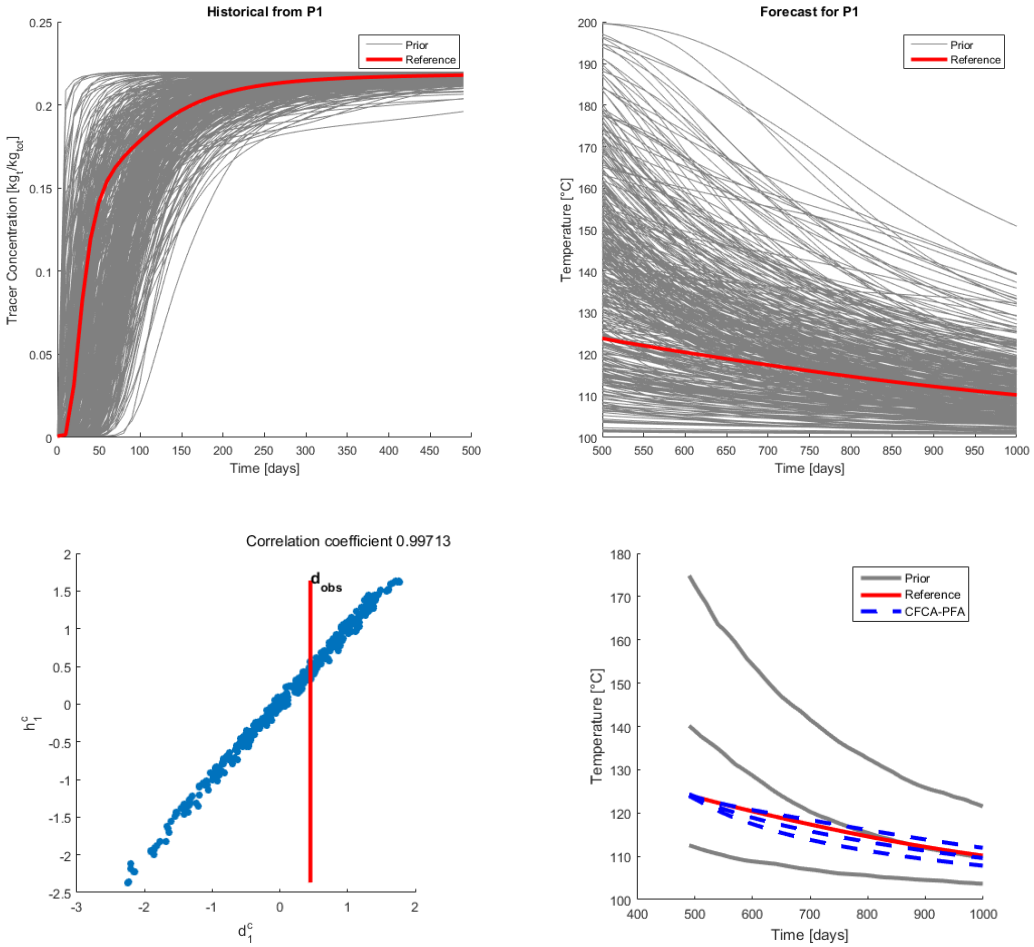


Figure 6: Data (tracer return curves) and prediction (temperature curves) variables (top). Data-prediction correlation coefficient (bottom left) and forecast quantiles for Producer 1 (bottom right).

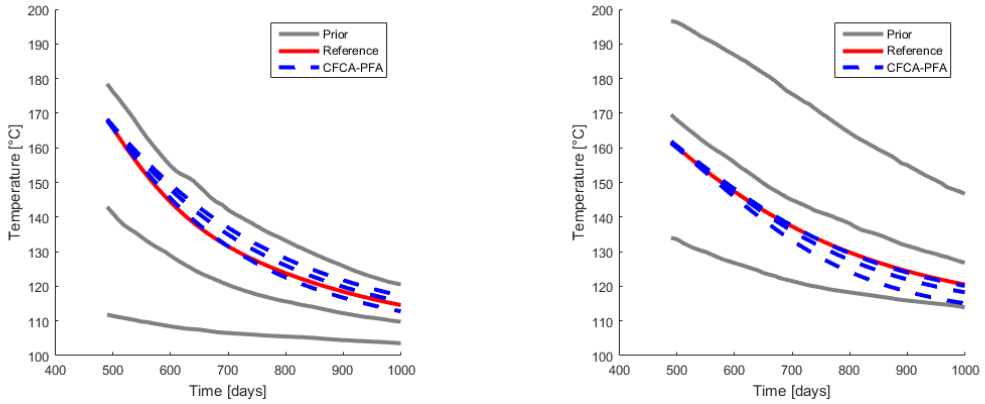


Figure 7: Forecast quantiles for Producer 2 (left) and Producer 3 (right).

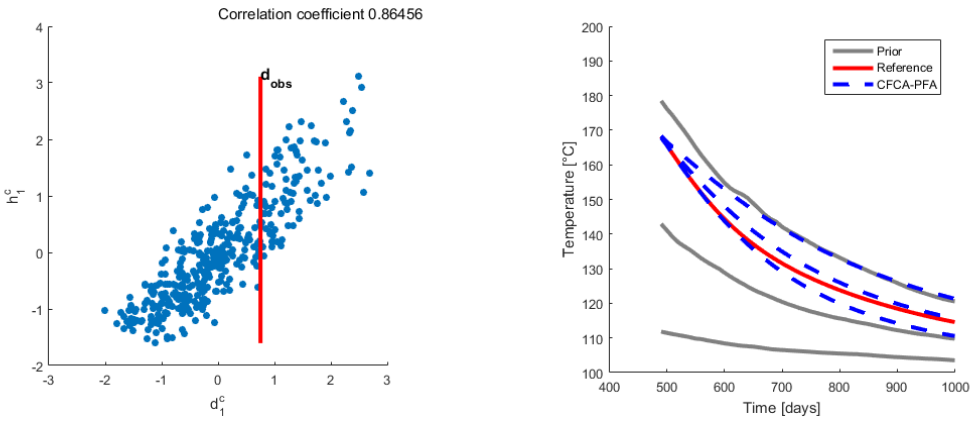


Figure 8: Correlation coefficient (left) and forecast quantiles (right) for Producer 2 when using data from Producer 1 and 3 for the analysis.

4. CONCLUSIONS

The cornerstone of a successful geothermal power generation is to sustain fluid or steam flow with high energy content. The supply is maintained by reinjecting waste water into the reservoir, which in turn brings the possibility of a thermal breakthrough in producers. The disadvantages of conventional methods for predictions, which involve building reservoir simulation models based on exploration data, inversions and iterations, are the computational efforts needed and the prior knowledge of the underlying physical models required. The field of data science is steadily growing, and many data-driven models and machine learning techniques have shown potential in analyzing large volumes of data from multiple wells, overcoming some of the limitations of preceding methods.

In this study, three statistical methods for reservoir characterization and prediction modeling were introduced. A nonparametric regression with the ACE algorithm was applied on a library of fracture networks and the well-to-well connectivity between injector and producers estimated. Results obtained using tracer data for the regression were in good agreement with tracer transit times whereas temperature data displayed much less correlation to connectivity. K-means clustering showed promise in being able to group fracture networks in agreement with connectivity between wells and CFCA was able to make direct forecasts for the producers based on tracer or temperature data. However, some complications were encountered. For the clustering technique the complexity grows with added producers, and therefore the number of groups increases in order to adequately describe the character of the fracture networks. Furthermore, the challenge of deciding how many clusters should be used is often difficult to solve. The CFCA method assumes data is smooth and its success depends on finding a linear relationship between data and prediction variables in the canonical space. Introducing real data, this relationship will most likely become less linear due to added noise. Moreover, the time-series data will not be smooth and perhaps even discontinuous.

To further support the use of these methods in geothermal settings, a broader analysis needs to be conducted. More realistic realizations of geothermal reservoirs will be generated, and the complexity of injection schedules increased. Furthermore, field scale data are considered and other statistical learning methods that can overcome the aforementioned shortcomings investigated.

ACKNOWLEDGEMENTS

The authors would like to thank the Department of Energy Resources Engineering at Stanford University and Landsvirkjun, National Power Company of Iceland, for the financial support during this work. Also, we would like to acknowledge Lewis Li (Stanford University) for all his help and allowing us to use his code formulation of the direct forecasting with CFCA method.

REFERENCES

Alizadeh, M., Junin, D., Mohsin, R., Movahed, Z., Alizadeh, M., & Alizadeh, M. (2015). Application of Artificial Neural Networks in Fracture Characterization and Modeling Techniques. In *Proceedings of the 17th International Conference on Mathematical and Computational Methods in Science and Engineering, Kuala Lumpur, Malaysia* (pp. 162–169).

Breiman, L., & Friedman, J. H. (1985). Estimating Optimal Transformations for Multiple Regression and Correlation. *Journal of the American Statistical Association*, *80*, 580–598.

Demiryurek, U., Banaei-Kashani, F., Shahabi, C., & Wilkinson, F. (2008). Neural-Network Based Sensitivity Analysis for Injector-Producer Relationship Identification. In *SPE Inteeligent Energy Conference and Exhibition, Amsterdam, The Netherlands*.

Heffer, K. J., Fox, R. J., McGill, C. A., & Koutsabeloulis, N. C. (1997). Novel Techniques Show Links between Reservoir Flow Directionality, Earth Stress, Fault Structure and Geomechanical Changes in Mature Waterfloods. In *SPE Journal* (Vol. 2, pp. 91–98).

Horne, R. N., & Szucs, P. (2007). Inferring Well-to-Well Connectivity using Nonparametric Regression on Well Histories. In *Proceedings, 32nd Workshop on Geothermal Reservoir Engineering, Stanford University*. Stanford, California.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer-Verlag New York.

Juliusson, E. (2012). *Characterization of Fractured Geothermal Reservoirs Based on Production Data*. *Stanford University*. PhD thesis, Stanford University.

Li, L. (2017). *A Bayesian Approach to Causal Evidential Analysis for Uncertainty Quantification Throughout The Reservoir Forecasting Process.* PhD thesis, Stanford University.

Liu, Y., & Horne, R. N. (2013). Interpreting Pressure and Flow Rate Data from Permanent Downhole Gauges Using Convolution-Kernel-Based Data Mining Approaches. In *SPE Western Regional & AAPG Pacific Section Meeting, California, USA*.

Magnusdottir, L. (2013). *Fracture Characterization in Geothermal Reservoirs Using Time-lapse Electric Potential Data*. *Stanford University*. PhD thesis, Stanford University.

Nguyen-Cong, V., & Rode, B. M. (1995). Application of Alternating Conditional Expectations Method to Quantitative Electronic Structure-Activity Relationships (QESAR). *Quantitative Structure-Activity Relationships*, *14*, 512–517.

Refunjol, B. T., & Lake, L. W. (1999). Reservoir Characterization Based on Tracer Response and Rank Analysis of Production and Injection Rates. *Reservoir Characterization - Recent Advances, AAPG Memoir 71*, 209–218.

Sarkheil, H., Hassani, H., & Alinia, F. (2009). The Fracture Network Modeling in Naturally Fractured Reservoirs Using Artificial Neural Network Based on Image Logs and Core Measurements. *Australian Journal of Basic and Applied Sciences*, *3*.

Satija, A., & Caers, J. (2015). Direct forecasting of subsurface flow response from non-linear dynamic data by linear least-squares in canonical functional principal component space. *Advances in Water Resources*, *77*, 69–81.

Satija, A., Scheidt, C., Li. L. & Caer, J. (2017). Direct forecasting of reservoir performance using production data without history matching. *Computational Geoscience,* 21, 315–333.

Scheidt, C., Renard, P., & Caers, J. (2014). Prediction-Focused Subsurface Modeling: Investigating the Need for Accuracy in Flow-Based Inverse Modeling. *Mathematical Geosciences*, *47*(2), 173–191.

Sullera, M. (1998). *Inferring Injection Returns from Chloride Monitoring Data*. Master thesis, Stanford Univeristy.

Tian, C., & Horne, R. N. (2015). Applying Machine Learning Techniques to Interpret Flow Rate, Pressure and Temperature Data From Permanent Downhole Gauges. In *SPE Western Regional Meeting, California, USA*.

Tian, C., & Horne, R. N. (2016). Inferring Interwell Connectivity Using Production Data. In *SPE Annual Technical Conference and Exhibition, Dubai, UEA*.

Tian, C., & Horne, R. N. (2017). Recurrent Neural Networks for Permanent Downhole Gauge Data Analysis. In *SPE Annual Technical Conference and Exhibition, San Antonio, Texas.*

Wang, D., & Murphy, M. (2004). Estimating Optimal Transformations for Multiple Regression Using the ACE Algorithm. *Journal of Data Science*, *2*, 329–346.

Yousef, A. A., Gentil, P., Jensen, J. L., & Lake, L. W. (2006). A Capacitance Model To Infer Interwell Connectivity From Production and Injection Rate Fluctuations. *SPE Reservoir Evaluation & Engineering*, *9*(6).