International Journal of Greenhouse Gas Control 109 (2021) 103382

Contents lists available at ScienceDirect

International Journal of Greenhouse Gas Control

journal homepage: www.elsevier.com/locate/ijggc

Machine learning to discover mineral trapping signatures due to CO_2 injection

Bulbul Ahmmed [*,a], Satish Karra [a], Velimir V. Vesselinov [a], Maruti K. Mudunuru [b]

[a] Computational Earth Science Group, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[b] Watershed & Ecosystem Science, Pacific Northwest National Laboratory, Richland, WA 99352, USA

ARTICLE INFO

Keywords:
Unsupervised machine learning
Reactive-transport simulation
CO_2 sequestration
Hidden features
Matrix factorization

ABSTRACT

Mineral trapping is pursued as a geological CO_2 sequestration (GCS) mechanism because it permanently stores CO_2 in solid phases or minerals. However, CO_2 mineral-trapping mechanisms are poorly understood due to (1) lack of sufficient field and laboratory data characterizing these complex processes, and (2) challenges to develop site-specific reactive-transport models coupling fluid flow and geochemical reactions occurring at various temporal (from milliseconds to years) and spatial (from pore (millimeters) to field (kilometers)) scales. Reactive transport with additional complexities such as heterogeneity can make the simulation outputs even more difficult to interpret because of complex nonlinearity and multi-scale interdependencies. Furthermore, the values of model outputs such as concentrations can vary by several orders of magnitude, making it harder to correlate and characterize the impact of the variables via traditional data interpretation techniques such as exploratory data analyses. Recently, machine learning (ML) has shown promise in feature discovery and in highlighting hidden mechanisms that cannot be obtained by existing data-analytics and statistical methods. In this study, we applied an unsupervised ML approach, non-negative matrix factorization with custom *k*-means clustering (NMF*k*) to the data generated by reactive-transport simulations of GCS. The reactive-transport data consisted of 19 attributes, including four physio-chemical variables (pH, porosity, aqueous CO_2, and sequestered CO_2), six chemical species (K^+, Na^+, HCO_3^-, Ca^{2+}, Mg^{2+}, Fe^{2+}), and four carbonate minerals (calcite, dolomite, siderite, and ankerite), a feldspar mineral (albite), and four clay minerals (illite, clinochlore, kaolinite, and smectite) over a period of 200 years of simulation time. The simulation data used was for Morrow B sandstone at the Farnsworth hydrocarbon unit in Texas. Data are sampled at two locations within the model domain: (1) at the injection well and (2) 200 m west of the injection well. The injection was performed for a period of 10 years. Using NMF*k*, we estimated the temporal interdependencies among the 19 attributes over a span of 200 years. We found that NMF*k* was able to identify four reaction stages and their dominant attributes; these cannot be directly discerned through traditional visualization (e.g., line plots, Pareto analysis, Glyph-based visualization methods) or exploratory data analysis tools of the simulation data. The four stages were: reactions in the injection phase followed by short-, mid-, and long-term reactions. The NMF*k* analysis also revealed that 10 among the 19 attributes are dominant. These dominant attributes for mineral trapping include calcite, dolomite at injection well, siderite at 200 m away from the injection well, clinochlore, kaolinite, Na^+, K^+, Ca^{2+}, Mg^{2+}, pH, and aqeuous CO_2. Finally, at late times (65–200 years), our results showed that calcite plays a major role in mineral trapping with insignificant contribution from siderite, ankerite, and clay minerals. These findings make the proposed unsupervised ML-model attractive for reactive-transport sensing towards real-time GCS monitoring.

1. Introduction

Greenhouse emissions due to anthropogenic activities have caused approximately 1 °C rise in global average temperature over pre-industrial levels (Masson-Delmotte et al., 2018). At the current emission rate, it is estimated that this average is likely to reach about 1.5 °C over pre-industrial levels (Masson-Delmotte et al., 2018). If we cannot reach and sustain net-zero global anthropogenic CO_2 emissions, we have to face dire climate consequences such as sea level rise, extreme weather events (e.g., floods, hurricanes), extreme cold and hot weathers,

* Corresponding author.
E-mail address: ahmmedb@lanl.gov (B. Ahmmed).

https://doi.org/10.1016/j.ijggc.2021.103382
Received 2 January 2021; Received in revised form 23 April 2021; Accepted 8 June 2021
Available online 22 June 2021

etc. To reach net-zero CO_2 emission by 2055, it is estimated that we have to reduce ≈500 Gt CO_2/year either through reducing emission or sequestration or both (Masson-Delmotte et al., 2018). Geological carbon sequestration (GCS) that involves the injection of CO_2 into the subsurface is being explored as a solution to reduce CO_2 emission to the environment.

Injected CO_2 can be simultaneously sequestered in geologic media via several mechanisms (Bachu, 2015): (1) storing in the supercritical form; (2) dissolution; or (3) trapping by reacting with natural minerals. Supercritical CO_2 tends to move upward because of its buoyancy leading to the risk of leaking back into the environment through faults and fractures near the stored reservoir. Dissolved CO_2 stays in the aquifer system but can exsolve and come to the surface if there is an escape route (e.g., fractures). The mineral form of CO_2 is generally the longest-term contributor to CO_2 sequestration that traps and stores CO_2 in the subsurface as minerals. For effective mineral CO_2 trapping, it is critical to understand the reaction mechanisms of injected CO_2 with subsurface geochemical species to form the product minerals. Reactive-transport simulations are sought after to understand such behavior since field experiments for decades are not feasible (Ahmmed et al., 2016; Audigane et al., 2007; Chen et al., 2018; Knauss et al., 2005; Liu et al., 2011; Menad et al., 2019; Siqueira et al., 2017; Zhang et al., 2020; 2009). Reactive-transport of injected CO_2 studies the evolution of hydraulic and geological properties by coupling flow/transport with geochemical simulations (Gunter et al., 1997; Xu et al., 2006; 2011).

Upon injection, supercritical CO_2 (for simplicity, we will call it "CO_2") dissolves in brine and decreases the system's pH. Low pH induces mineral dissolution and complexation with dissolved ions such as Na^+, Ca^{2+}, Mg^{2+}, Fe^{2+}, and HCO_3^- to form $CaHCO_3^+$, $MgHCO_3^+$, and $FeHCO_3^+$ (Rosenbauer et al., 2005; Wang et al., 2013; Zhang et al., 2009). As time progresses, solubility trapping increases that in turn decreases the CO_2 solubility. (Ennis-King and Paterson, 2007; Lindeberg and Wessel-Berg, 1997). The density enhancement increases the mineral precipitation. Also, variations in the reaction rate and abundance of illite, kaolinite, smectite, clichlore, albite, oligoclase, etc. minerals significantly affect the mineral alteration and CO_2 storage by the trapping mechanism (Xu et al., 2006; 2011). This coupling of reactive transport and reservoir characteristics such as heterogeneity, and multi-scale interdependencies make it harder to interpret the simulated data. Also, coupling process solves and combines multiple non-linear equations (e.g., conservation laws, constitutive models) that increases the complexity in the outputs. Furthermore, the values of variables such as concentrations in the system vary by several orders of magnitude (Xu et al., 2011). This variation by orders of magnitude among attributes makes it harder to correlate and characterize the impact of physio-chemical variables, species, and minerals in the system with traditional data interpretation techniques such as line plots, Pareto analysis, Glyph-based visualization methods. However, the recent unsupervised machine learning (ML) methods can handle multivariate, multi-scale, and complex nonlinear datasets. Additionally, they can provide a better picture of the system's evolution by highlighting dominant features and discovering any hidden signatures (Cichocki et al., 2009; Vesselinov et al., 2019; 2018).

In this paper, we use a type of unsupervised ML called non-negative matrix factorization with customized *k*-means clustering (NMF*k*) (Iliev et al., 2018; Vesselinov et al., 2018) to correlate and characterize the evolution of critical reactive-transport attributes due to CO_2 injection and mineral trapping. NMF*k* identifies the optimal number of features (or signals) present in the data when non-negative matrix factorization (NMF) analyses are performed (Cichocki et al., 2009; Gillis, 2020; Lee and Seung, 1999). Typically, NMF involves splitting a given data matrix of size ($m \times n$) into two non-negative factor matrices of dimensions $m \times k$, $k \times n$, such that their product approximates the data matrix. However, the parameter k that determines the number of features in the data, is an unknown. Classical NMF approaches do not allow for automatic estimation of k. The NMF*k* approach overcomes this drawback and estimates k through custom *k*-means clustering coupled with regularization constraints. Thus, NMF*k* finds the optimal number of dominant signals or latent features and similarities of attributes in a dataset (Iliev et al., 2018; Vesselinov et al., 2018). Recently, NMF*k* has been successful in identifying key features and signals in various geoscience applications, such as geothermal, groundwater hydrology, and contaminant transport. For instance, Alexandrov and Vesselinov (2014) applied NMF*k* for blind source separation of transient pressure distribution for a hydrogeological system, while Vesselinov et al. (2018) detected contaminant source using NMF*k* in a complex hydrogeologic setting. More recently, Ahmmed et al. (2021) analyzed 15 chemical elements of 14351 groundwater samples of the Great Basin to find patterns between geothermal resource types and chemical characteristics of groundwater. They showed that NMF*k* successfully identified patterns between chemical elements and geothermal resource types (low, medium, and high temperature). Vesselinov et al. (2019) used a variant of NMF*k* called non-negative tensor factorization with customized *k*-means clustering (NTF*k*) to analyze reactive-mixing data. They were able to differentiate the effects of anisotropy, diffusion, and spatial components in reactive mixing. Due to the success of NMF*k* for identifying key features and signals in various geoscience applications, including reactive-transport, we chose to use this approach to analyze CO_2 mineral trapping data. Moreover, to the best of our knowledge, there are no previous works, which use NMF*k* or similar unsupervised ML tools, such as Principal Component Analysis (PCA), Independent Component Analysis (ICA), etc., to analyze either reactive-transport simulations or geochemical data from a site, related to the CO_2 sequestration application.

We chose a GCS pilot site at the Farnsworth hydrocarbon unit in Texas for our study. Openly available reactive-transport simulation data with 19 attributes from Khan (2017) was used. *NMFk was applied to this data to learn three critical aspects of reactive-transport simulations that are: (1) hidden signals that characterize the reactive-transport simulations; (2) the progression of these signals to capture the reaction stages; and (3) the dominant attributes corresponding to each hidden signal.* We note that no one performed such a detailed analysis of geochemical data to better understand the mineral-trapping mechanism due to CO_2 injection before. The above aspects make this study innovative and attractive for extracting actionable information (e.g., important attributes) from geochemical data, which is very important for cost-effective data sampling (for instance, collecting only informative variables) during a field campaign.

The paper is organized as follows. Section 2 describes the NMF*k* methodology followed by details of the reactive-transport simulation data in Section 3 and Section 4. The results and discussion are presented in Section 5. Finally, conclusions are drawn in Section 6.

2. Unsupervised machine learning methodology based on NMF*k*

Non-negative matrix factorization with *k*-means clustering (NMF*k*) is an unsupervised machine learning method that combines two unsupervised machine learning methods, non-negative matrix factorization (NMF) and customized *k*-means clustering. NMF learns from parts of an object (Lee and Seung, 1999) while *k*-means clustering partitions similar data points together and discovers underlying patterns (Wagstaff et al., 2001). Customized *k*-means clustering estimates similarity of an object to a cluster compared to other clusters (Iliev et al., 2018; Vesselinov et al., 2018). In this study, a data matrix, X of size (n, m) was first built with non-negative elements, where n is the number of time snapshots and m is the number of features/attributes.

NMF of NMF*k* algorithm decomposes $X_{n\times m}$ into two matrices, $W_{n\times k}$ (basis/mixing matrix) and $H_{k\times m}$ (coefficient/attribute matrix) as:

$$X = W \times H + \epsilon(k), \tag{1}$$

where $\epsilon(k)$ is an error between X and $W \times H$ for the specified k while k is

the unknown/hidden number of signals present in the data and is always smaller than n and m. The W matrix represents how hidden signals are related to the data. The H matrix depicts the relationship between features and hidden signals. In this methodology, the number of hidden signals (k) is an unknown and is identified by performing a series of NMF for $k = 2, 3, \ldots, d$, where $d \leq m$. The NMF process minimizes the following objective function, $\mathscr{L}$, based on Frobenius norm for a specified k, such that the entries of the resulting W, H are non-negative:

$$\mathscr{L} = \| X - W \times H \|_F \quad \text{such that} \quad W, H \geq 0 \quad \forall \quad n, m, k. \tag{2}$$

In the minimization, the following rules are used to update the W and H matrices in each iteration (Gillis, 2020; Lee and Seung, 1999):

$$W_{ij} \leftarrow W_{ij} \frac{(XH^T)_{ij}}{(WHH^T)_{ij}}, \tag{3}$$

and

$$H_{ij} \leftarrow H_{ij} \frac{(W^T X)_{ij}}{(W^T WH)_{ij}}. \tag{4}$$

where i and j are indices from $1, 2, \ldots, n$. For each k, NMF is solved for 1,000 random initial guesses for W and H matrices. The least value of $\mathscr{L}$ for a given k is assumed as the best value for the reconstruction error. After completing the NMF process, the 1,000 estimated Hs are clustered into k clusters using a customized k-means clustering. However, k is also unknown in the k-means clustering. The algorithm consecutively examines a specified k by obtaining 1,000 H matrices for each feature/variable. During clustering, the similarity between two clusters is assessed using the Silhouette width/value (Rousseeuw, 1987; Vesselinov et al., 2018), which is essentially the cosine norm:

$$\rho(p, q) = 1 - \frac{\sum_{i=1}^{n} p_i q_i}{\sum_{i=1}^{n} p_i^2 \sum_{i=1}^{n} q_i^2}, \tag{5}$$

where i is an index from $1, 2, \ldots, n$. p_i and q_i are components of vectors $\mathbf{p}$ and $\mathbf{q}$. The Silhouette value quantifies how similar an object is to its own cluster compared to other clusters and varies from -1 to $+1$; high values indicate that the object is well matched to its own cluster and poorly matched to neighboring clusters. The combination of reconstruction error ($\mathscr{L}(k)$) and the Silhouette value are used to determine the optimal number of hidden signals. If k is low, the Silhouette value will be high, but so maybe $\mathscr{L}(k)$ because of under-fitting. For high k, the Silhouette value will be low and the solution may be over-fit. So, the best estimate for k is a number that optimizes both $\mathscr{L}(k)$ and the Silhouette value. However, signal with the Silhouette value >0.5 is considered as a dominant signal.

3. Model description

The data used for NMFk-based ML analysis are from the reactive-transport simulations of CO_2 injection in the Morrow B sandstone at the Farnsworth hydrocarbon unit in Texas from Khan (2017). Figure 1 illustrates the study area. A 3D heterogeneous model with realistic topography was built to cover this area, 73 km^2 with ≈ 10m depth. The model domain is meshed using a structured grid containing 61,600 cells (i.e., 110 × 80 × 7). The Farnsworth hydrocarbon unit is a sandstone rich reservoir with a mineral composition shown in Table 1. Among these 61,600 cells, 15,357 were active for reactive-transport simulation to mimic the topography. TOUGHREACT software with the ECO2N equation of state for NaCl brine and CO_2 was used to perform the simulations (Spycher et al., 2003; Xu et al., 2006; 2011). The EQ3/6 thermodynamic database (Arnorsson and Stefansson, 1999; Kulik and Aja, 1997; Rimstidt, 1997; Wolery, 1992) with kinetic-reaction rate constants from Palandri and Kharaka (2004) were used for geochemical

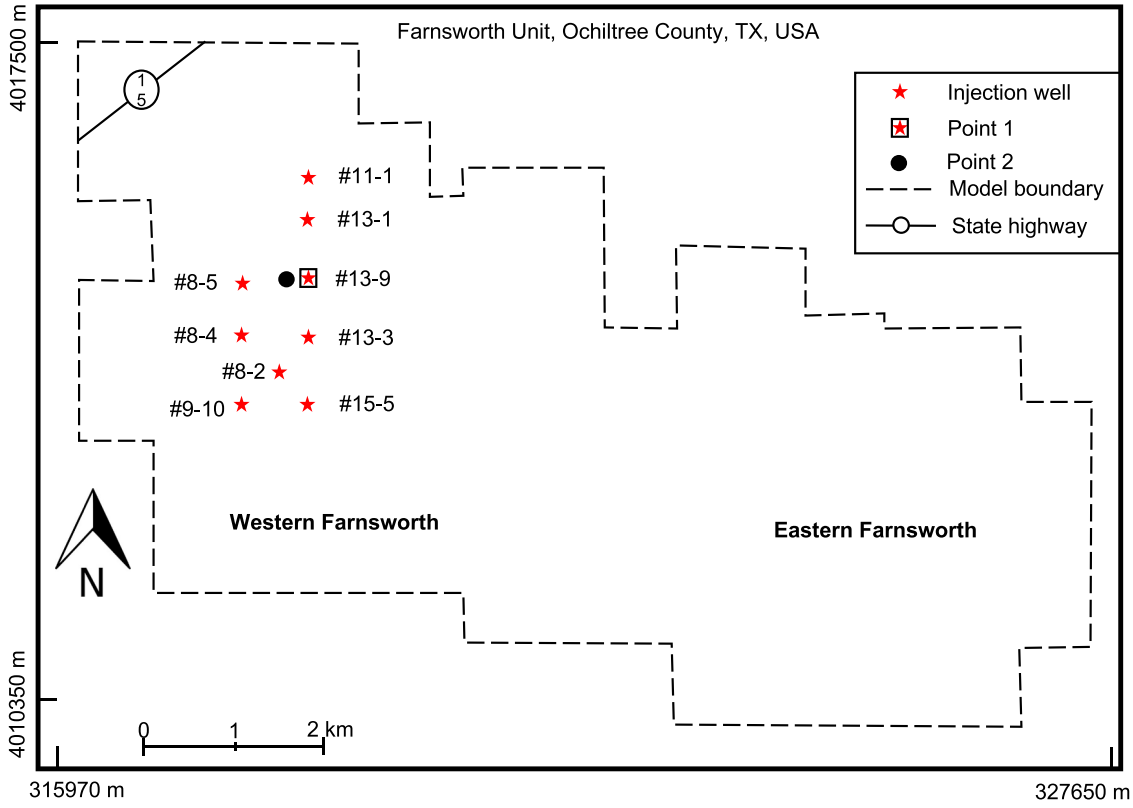

Fig. 1. Schematic of the study area. The study area includes nine wells (red filled stars). In this study, Point 1 (well 13-9) and Point 2, which is 200 m west of well 13-9 (black filled circle) were explored. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Minerals used during simulation with their chemical compositions and initial volume fractions. Zero volume fraction means that the mineral does not exist in the reservoir but expected to be precipitated, source: (Ahmmed et al., 2016).

Mineral	Chemical composition	Volume fraction (%)
Quartz	SiO_2	84.26
Albite	$NaAlSi_3O_8$	9.0
Calcite	$CaCO_3$	0.75
Ankerite	$(Ca(Fe, Mg, Mn)(CO_3)_2)$	0.25
Siderite	$FeCO_3$	0.25
Dolomite	$CaMg(CO_3)_2$	0.0
Smectite	$(Na, Ca)_{0.3}(Al, Mg)_2[Si_4O_{10}](OH)_2.nH_2O$	0.1
Kaolinite	$Al_2Si_2O_5OH_4$	2.72
Illite	$(K, H_3O)(Al, Mg, Fe)_2(Si, Al)_4O_{10}[(OH)_2, (H_2O)]$	0.88
Chlinochlore	$Mg_5Al(AlSi_3O_{10})(OH)_8$	1.79

reactions. The simulation included 10 minerals, 18 basis species, and pH. Table 1 shows the initial volume fractions of these nine minerals in the model domain. These volume fractions were uniformly distributed to the model domain. Table 2 includes the initial values of the main aq. species used in the simulation. A total of 2.67 million tonnes of supercritical CO_2 was injected for 10 years using water alternating gas (WAG) method at nine wells (Fig. 1). The simulation considered only storage scenario; therefore, no oil/gas production well was used. The total simulation time was 200 years.

The simulation tracks the mass of mineral being trapped due to CO_2 injection via the quantity $SMCO_2$ [kg]. The initial value of $SMCO_2$ is 0. After 200 years of simulation $SMCO_2$ was 0.6 kg per m^3 of the system. The primary mode of sequestration is aq. phase while a significant amount of CO_2 leaked to the environment through faults. About 10^{-3}% porosity changed after 200 years of simulation.

4. Data description

For our analysis, we used data for 19 attributes (Ahmmed, 2015; Ahmmed et al., 2016) simulated at well 13-9 and a location 200 m west of the well 13-9 (Fig. 1). For the rest of the article, we will refer to well 13-9 as Point 1 and the location at 200 m west of well 13-9 as Point 2. For both locations, the data was collected from a single grid cell at the injection depth. At both points, data matrices (X) of size 200 $\times$19 were formed using 200 yearly snapshots of the 19 attributes. The 19 attributes consisted of changes in volume fraction [–] for nine minerals, the concentration of six species, and four physio-chemical attributes collected over the 200 years simulation time. The nine minerals included carbonate and noncarbonate minerals, with carbonate minerals being calcite, dolomite, siderite, and ankerite while noncarbonate minerals were kaolinite, clinochlore, illite, and albite. Note that quartz is the dominant reaction for this system, but the change in its volume fraction was found to be insignificant for the simulation period. For this reason, we have omitted quartz for the ML analysis in this study. The six aq. species were Na^+ [mol/L], K^+ [mol/L], Ca^{2+} [mol/L], Mg^{2+} [mol/L], HCO_3^- [mol/L], and Fe^{2+} [mol/L], while the four physio-chemical attributes included pH [–], porosity [%], aqueous CO_2 [mol/L] (denoted by aq. CO_2), and $SMCO_2$. Figures 2(a) and 3(a) show the normalized volume fractions of the nine minerals at Points 1 and 2, respectively. Each attribute is individually scaled between 0 and 1 (dividing each value attribute by the maximum value of the attribute vector) in the pre-processing step. Figures 2(b) and 3(b) show the normalized concentrations of the six species, while Figs. 2(c) and 3(c) show the normalized magnitude of the four physio-chemical variables at these two locations.

Table 2
Initial values for the physio-chemical variables in the simulation, source: (Ahmmed et al., 2016).

Variables	Chemical composition
Na^+	1.245×10^3 mg/kg in fluid
K^+	7.49 mg/kg in fluid
Ca^{2+}	35.7 mg/kg in fluid
Mg^{2+}	9.285 mg/kg in fluid
Fe^{2+}	3.65×10^{-5} mg/kg in fluid
HCO_3^-	29347.5 mg/kg in fluid
Aq. CO_2	0.0 mg/kg in fluid
$SMCO_2$	0.0 mg/kg in fluid
pH	7 [–]
Porosity	2.72%

5. Results and discussion

In this section, we present the results of our unsupervised learning based on NMF*k*. Optimal number of hidden features are provided and associated signatures for mineral trapping are discussed.

5.1. Results at point 1

NMF*k* analysis was performed on the normalized geochemical data to discover the dominant attributes at Point 1. The normalized data matrix was factorized up to $k = 12$ signals and we found that $k = 4$ was optimal because of the low reconstruction error and high Silhouette width (see, Fig. 4(a)). Factorized data provided two reduced-order matrices W (mixing matrix) and H (attribute matrix) of the data matrix, X. The mixing matrix shows how signals vary over time (Fig. 4(b)) while the attribute matrix (Fig. 4(c)) shows the significance of attributes on discovered signals. We denote the four identified signals as S_{inj}, S_{str}, S_{mtr}, and S_{ltr}. The signal S_{inj} shows the highest magnitude during the injection period of 1–10 years and is close to zero after the cessation of injection (Fig. 4(b)), and hence captures the injection-term reactions. The dominant attributes of S_{inj} are dolomite, ankerite, illite, smectite, Na^+, K^+, pH, porosity, aq. CO_2, and $SMCO_2$ (see, Table 3). S_{str} captures the short-term reaction effects of CO_2 in the system with a high magnitude between 11–25 years, and we define this as a short-term reaction stage. The dominant attributes of S_{str} are dolomite, kaolinite, K^+, HCO_3^+, pH, and aq. CO_2. The magnitude of S_{mtr} increases between 25–48 years and then gradually decays to zero at the end of the simulation (Fig. 4(b)). We define this stage as the mid-term reactions in the system. The dominant attributes are calcite, dolomite, clinochlore, kaolinite, K^+, Ca^{2+}, pH, aq. CO_2. The signal S_{ltr} increases in magnitude after $\approx$ 60 years till the end of the simulation, and hence captures the long-term geochemical reactions. The dominant attributes of S_{ltr} are calcite, dolomite, clinochlore, kaolinite, Na^+, Ca^{2+}, Mg^{2+}, pH, and aq. CO_2.

5.2. Results at point 2

Similar to Point 1, NMF*k* analysis was performed to discover dominant attributes to the data at Point 2, and $k = 4$ was also found to be the optimal number of signals (Fig. 5(a)). The resulting mixing matrix is shown in Fig. 5(b), and the attribute matrix is shown in Fig. 5(c). Similar to Point 1, the four optimal signals also represent the four reaction stages at Point 2, namely, S_{inj}, S_{str}, S_{mtr}, and S_{ltr}. Overall, the four signals at Point 2 follow similar trends as the ones at Point 1 but with a lag of $\approx$5 years. This lag is due to the flow and transport of injected CO_2 from Point 1 to Point 2. The trends can be observed by comparing the peaks of S_{inj} for both points (Figs. 4(b) and 5(b)). The dominant attributes of S_{inj} are dolomite, ankerite, albite, kaolinite, smectite, Na^+, Ca^{2+}, pH, porosity, aq. CO_2, and $SMCO_2$ (see, Table 4). The dominant attributes of S_{str} are siderite, clinochlore, kaolinite, Ca^{2+}, Mg^{2+}, HCO_3^+, aq. CO_2. The dominant attributes of S_{mtr} are calcite, siderite, clinochlore, kaolinite,

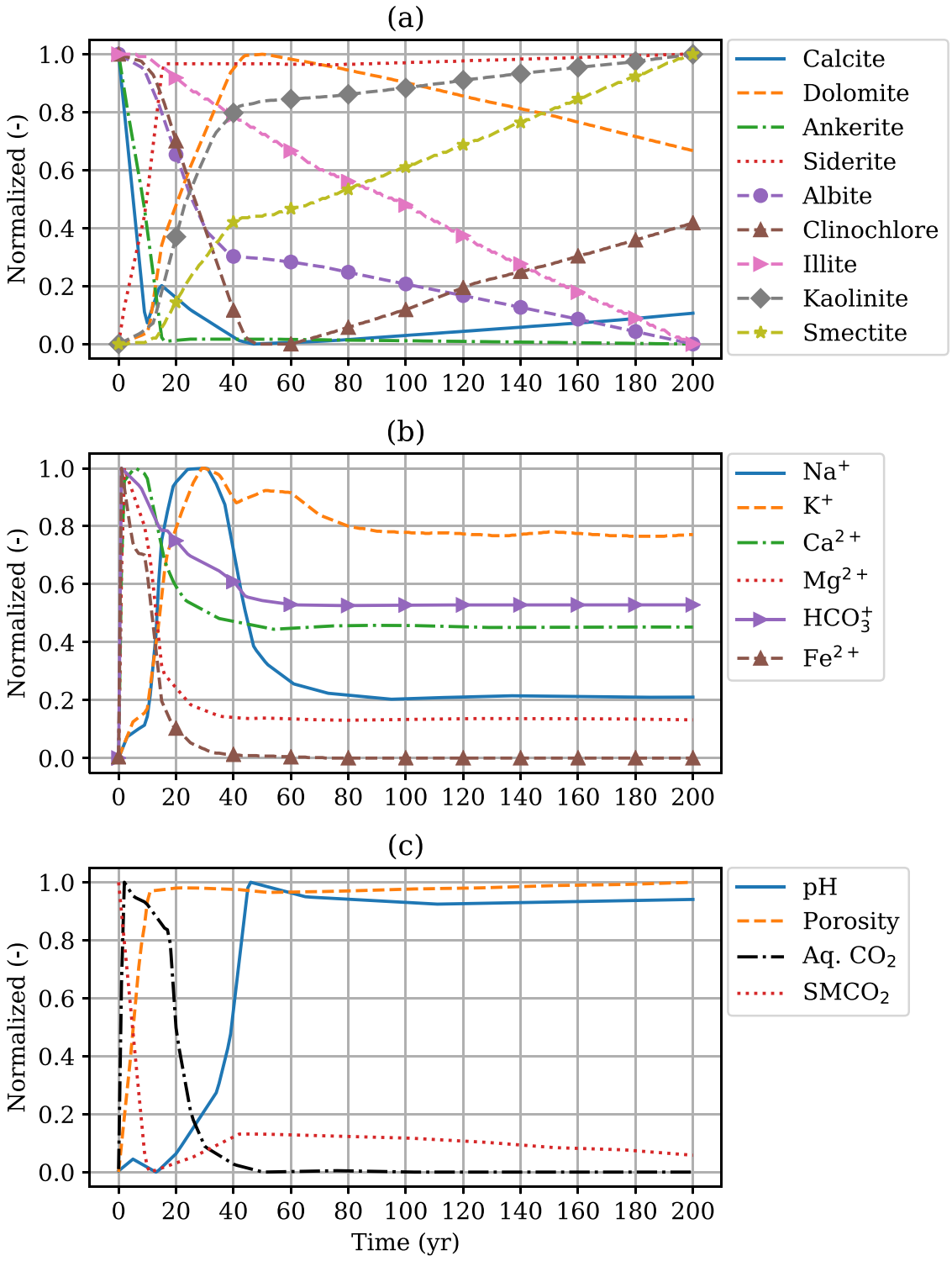

Fig. 2. Data at Point 1: Normalized ($\frac{X_i}{\max[X_i]}$) magnitude of minerals (a), species (b), and physio-chemical variables (c) over time.

K^+, Ca^{2+}, Mg^{2+}, pH, and aq. CO_2. The dominant attributes of S_{ltr} are calcite, siderite, clinochlore, kaolinite, K^+, Mg^{2+}. We will detail the significance of these four reaction stages and their dominant attributes in the next section.

5.3. Discussion

Reactive-transport modeling is a useful tool for the quantitative assessment of the geochemical-reaction processes. One can understand the integrity of caprock and wellbore and identify the key leakage pathways upon CO_2 injection. Long-term progression of reactions with host rock and fluids (e.g., water, CO_2) at a field site can also be identified (Dai et al., 2020). However, the complex coupling of various physical and chemical processes can make the reactive-transport simulation outputs and their relation to various model inputs hard to understand. The model data typically includes a lot of inputs/outputs with complex nonlinear interdependencies. We need a tool that can efficiently correlate all inputs and outputs regardless of these issues. The unsupervised

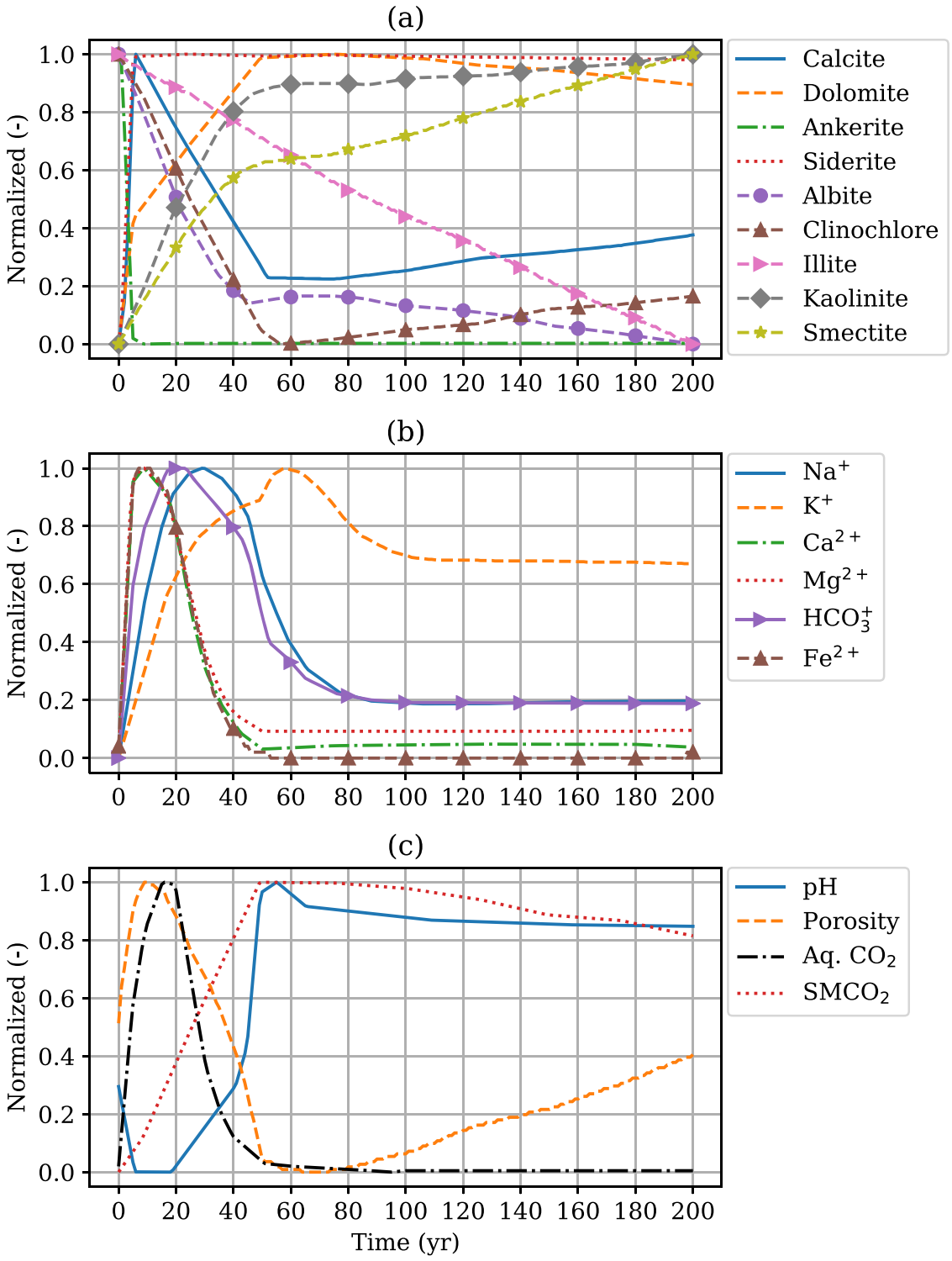

Fig. 3. Data at Point 2: Normalized ($\frac{X_i}{\max|X_i|}$) magnitude of minerals (a), species (b), and physio-chemical variables (c) over time.

ML tool, NMF*k*, fills this gap. We will highlight the main inferences from our results below.

First, this study reduced the data dimension and identified four key signals at both points (Point 1 and Point 2). These signals characterized how reactions progress over time and the four signals corresponded to reactions during the injection (S_{inj}) followed by short-term (S_{str}), mid-term (S_{mtr}), and long-term (S_{ltr}) reaction mechanisms. It is critical to mention that none of these reaction mechanisms were labeled in the dataset but NMF*k* was able to identify them. NMF*k* identified the important minerals, species, and physio-chemical variables for each of these discovered signals. Also, it successfully captured the time delay of the various reaction stages between Point 1 and Point 2. This categorical and detailed understanding is a massive gain towards obtaining insight into the mineral trapping process that is not viable through direct data visualization.

Physio-chemical variables control mineral precipitation/dissolution, which in turn controls aq. species concentration. Accurate estimation of interdependencies among these three types of attributes is critical to

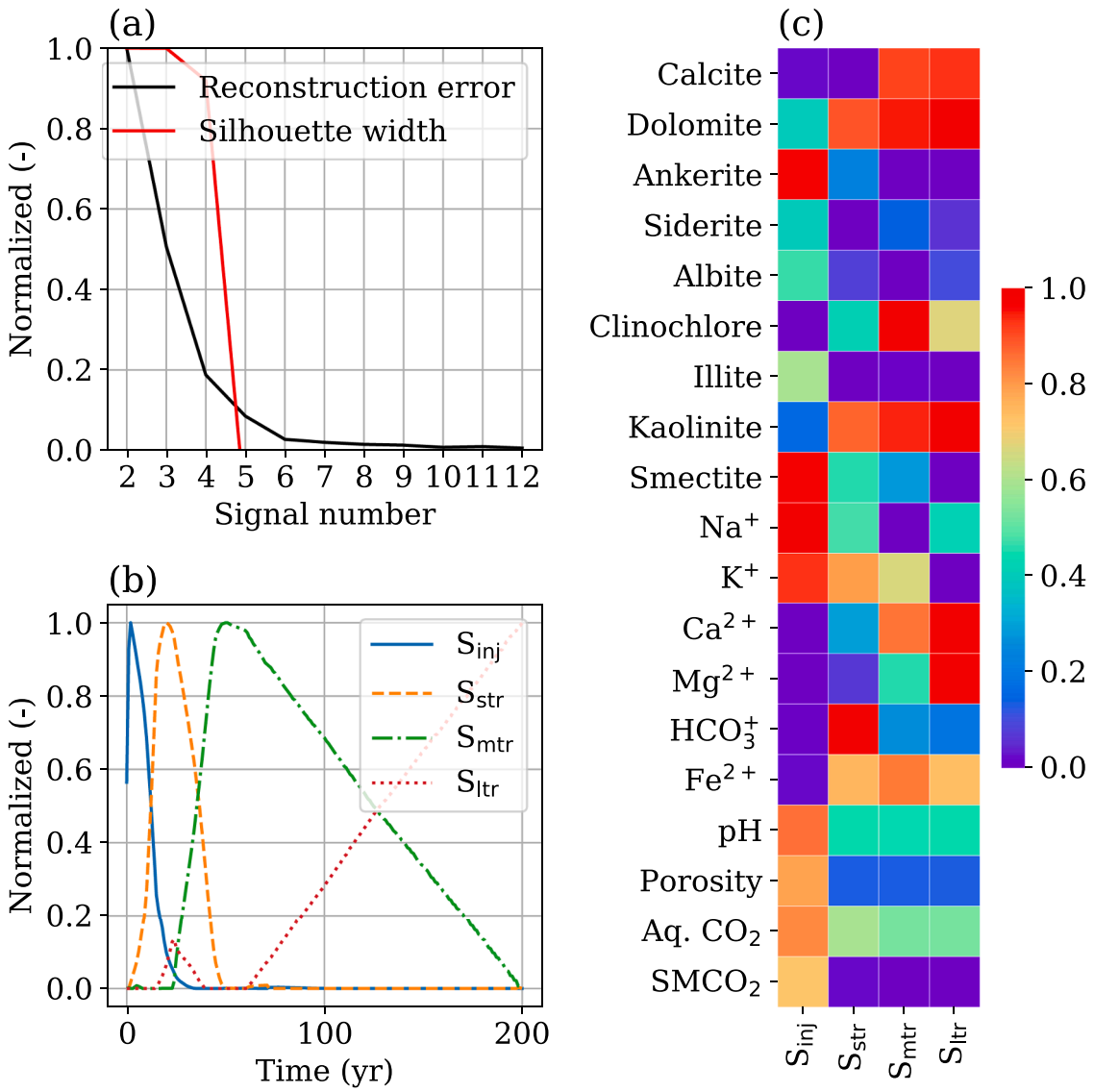

Fig. 4. NMF*k*results at Point 1: (a) Reconstruction error and Silhouette width vs. number of signals, (b) progression of four signals over time ($\frac{W_i}{\max[W_i]}$), and (c) significance of attributes of each signal ($\frac{H_i^T}{\max[H_i^T]}$). S_{inj}, S_{str}, S_{mtr}, and S_{ltr} stand for the injection activity, the short-term reaction, the mid-term reaction, and the long-term reaction, respectively.

Table 3
Dominant attributes and physical significance of each hidden signal at Point 1.

Signal	Dominant attributes	Physical significance
S_{inj}	Dolomite, ankerite, illite, smectite, Na^+, K^+, pH, porosity, aq. CO_2, and $SMCO_2$	Injection-term reactions
S_{str}	Dolomite, kaolinite, K^+, HCO_3^-, Fe^{2+}, pH, and aq. CO_2	Short-term reactions
S_{mtr}	Calcite, dolomite, clinochlore, kaolinite, K^+, Ca^{2+}, Fe^{2+}, pH, and aq. CO_2	Mid-term reactions
S_{ltr}	Calcite, dolomite, clinochlore, kaolinite, Na^+, Ca^{2+}, Mg^{2+}, Fe^{2+}, pH, and aq. CO_2	Long-term reactions

understanding the system dynamics. We know they are correlated but not the extent of their interdependencies over time. The NMF*k* analyses were able to provide these missing details of the system dynamics (see, Figs. 2–5). At Point 1, during injection, CO_2 reacted with the reservoir fluid that increased aq. CO_2 concentration, porosity, and $SMCO_2$ but quickly decreased pH, leading to precipitation of ankerite and dissolution of dolomite, albite, kaolinite, and smectite. These mineral precipitation and dissolution processes significantly changed the concentrations of Na^+ and Ca^{2+}. As a result, the four physio-chemical variables along with dolomite, ankerite, albite, kaolinite, smectite, Na^+ and Ca^{2+} were dominant in S_{inj}.

During the short-term reaction phase, system dynamics became stable, the pH increased, and the aq. CO_2 remained similar to that in the injection phase. Therefore, dolomite and kaolinite precipitated that altered concentrations of K^+ and HCO_3^+, pH, and aq. CO_2. Hence, dolomite, kaolinite, K^+, HCO_3^+, Fe^{2+}, pH, and aq. CO_2 are the dominant attributes of this signal.

During the mid-term reaction stage, the pH and aq. CO_2 remained high. Calcite, kaolinite, and clinochlore precipitated, but dolomite dissolved that changed the concentrations of K^+, and Ca^{2+}. Consequently, calcite, dolomite, clinochlore, kaolinite, K^+, Ca^{2+}, pH, and aq. CO_2 were the dominant attributes in this signal.

During the long-term reaction stage, the pH and aq. CO_2 remained high while calcite, clinochlore, and kaolinite precipitated with a higher rate while dolomite kept dissolving. These preceding effects significantly altered the concentrations of Na^+, Ca^{2+}, and Mg^{2+}. Hence,

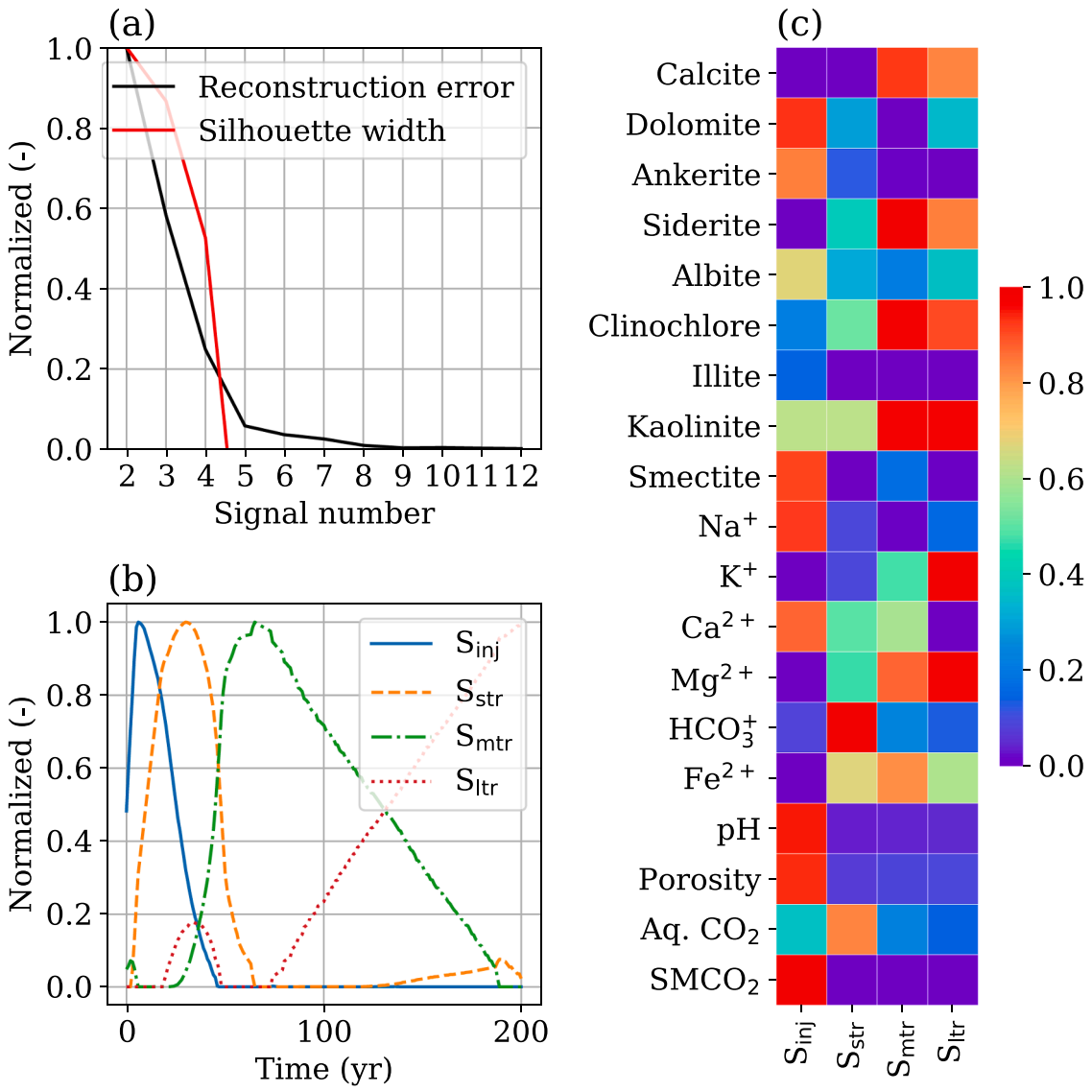

Fig. 5. NMF*k*results at Point 2: (a) Reconstruction error and Silhouette width vs. number of signals, (b) progression of four signals over time ($\frac{W_i}{\max[W_i]}$), and (c) significance of attributes of each signal ($\frac{H_i^T}{\max[H_i^T]}$).

Table 4
Dominant attributes and physical significance of each hidden signal at Point 2.

Signal	Dominant attributes	Physical significance
S_{inj}	Dolomite, ankerite, albite, kaolinite, smectite, Na^+, Ca^{2+}, pH, porosity, aq. CO_2, and $SMCO_2$	Injection-term reactions
S_{str}	Siderite, clinochlore, kaolinite, Ca^{2+}, Mg^{2+}, HCO_3^-, Fe^{2+}, and aq. CO_2	Short-term reaction
S_{mtr}	Calcite, siderite, clinochlore, kaolinite, K^+, Ca^{2+}, Mg^{2+}, Fe^{2+}, pH, and aq. CO_2	Mid-term reaction
S_{ltr}	Calcite, siderite, clinochlore, kaolinite, K^+, Fe^{2+}, and Mg^{2+}	Long-term reaction

calcite, dolomite, clinochlore, kaolinite, Na^+, Ca^{2+}, Mg^{2+}, pH, and aq. CO_2 were the dominant attributes in S_{ltr}.

Note that Fe^{2+} was also a dominant attribute in S_{inj}, S_{mtr}, and S_{ltr} because illite (a mineral with Fe^{2+}) started to dissolve after the cessation of CO_2 injection. However, illite did not show up in any of the signals because of the insignificant changes in its volume fraction compared to other minerals. Since Point 2 went through a similar set of reaction mechanisms, a similar explanation is applicable for signals at Point 2 but with a ≈ 5 year time lag. The time lag between the 2 points provides insights into the effective velocity of propagation of their geochemical processes.

Next, we will highlight the dominant reactions in each of the four reaction stages. During the injection period, dominant reactions at Point 1 and Point 2 are dolomite, ankerite, and smectite. During the short-term reaction, dolomite (at Point 1)/siderite (at Point 2), and kaolinite are dominant reactions at both points. During the mid-term reaction, calcite, dolomite/siderite, clinochlore, and kaolinite are dominant reactions at both points. In the long-term reaction, calcite, dolomite/siderite, clinochlore, and kaolinite are critical at both points. These four stages demonstrated what minerals would play a major role in the reservoir during and post injection. Especially, the long-term reaction is critical for mineral trapping because it defines the long-term fate of injected CO_2. In the long run, calcite may precipitate back after initial dissolution while siderite and dolomite might dissolve away after initial precipitation. Also, clay minerals may show insignificant changes in the long run.

All these findings are available in the data and can be surmised by subject-matter-experts. However, due to the small magnitude of these

changes, they are not easily detectable through direct visualization or application of comonly used data-analytics and statistical methods. In contrast, NMF*k* elegantly captured and synthesized them. Any site operator can use this tool without being an expert in ML using SmartTensors (https://tensors.lanl.gov/). The site operator will gain a better understanding of the complex system that will essentially assist in site selection, risk assessment, operations, and monitoring perspectives. For example, we know the critical reactions and aq. species for this site including their temporary distribution. Now, we can concentrate only on the critical variables and quantify their uncertainties for the risk assessment regarding CO_2 injection and mineral sequestration components.

6. Conclusions

This study applied NMF*k* to reactive-transport data towards improving our understanding of the mineral trapping mechanism. Datasets from two locations of the Farnsworth hydrocarbon unit in west Texas were used. Each dataset contains nine minerals, six species, and four physio-chemical variables totaling 19 attributes representing model simulation of reactive transport over 200 years. For each dataset, NMF*k* found four (S_{inj}, S_{str}, S_{mtr}, and S_{ltr}) optimal hidden signals characterizing geochemical processes have been discovered. Each signal captured unique physical significance about the reaction mechanisms for both of these datasets. For Point 1, S_{inj} spanned between 0–10 years, S_{str} spanned between 11–25 years, S_{mtr} spanned between 26–64 years, and S_{ltr} spanned between 65–200 years. Point 2 experienced similar reaction mechanisms but with a time lag of ~5 years. For Point 1, the dominant attributes for S_{inj} were dolomite, ankerite, illite, smectite, Na^+, K^+, pH, porosity, aq. CO_2, and $SMCO_2$; the dominant attributes for S_{str} were dolomite, kaolinite, K^+, HCO_3^+, pH, and aq. CO_2; the dominant attributes of S_{mtr} were calcite, dolomite, clinochlore, kaolinite, K^+, Ca^{2+}, pH, and aq. CO_2; and finally, the dominant attributes for S_{ltr} were calcite, dolomite, clinochlore, kaolinite, Na^+, Ca^{2+}, Mg^{2+}, pH, and aq. CO_2. For Point 2, the dominant attributes for S_{inj} were dolomite, ankerite, albite, kaolinite, smectite, Na^+, Ca^{2+}, pH, porosity, aq. CO_2, and $SMCO_2$; the dominant attributes for S_{str} were siderite, clinochlore, kaolinite, Ca^{2+}, Mg^{2+}, HCO_3^+, and aq. CO_2; the dominant attributes of S_{mtr} were calcite, siderite, clinochlore, kaolinite, K^+, Ca^{2+}, Mg^{2+}, pH, and aq. CO_2; finally, the dominant attributes for S_{ltr} were calcite, siderite, clinochlore, kaolinite, K^+, and Mg^{2+}. Results also demonstrated that calcite would play a major role in mineral trapping in the long run with insignificant contribution from siderite, ankerite, and clay minerals.

The reaction stages over time and the dominant attributes provided a better understanding of the mineral-trapping mechanism due to CO_2 injection. Such a study can be implemented to any GCS site for a comprehensive understanding of the response of a reservoir during and post CO_2 injection. This study analyzed data at two points that demonstrated part of the story of the specified system but not the whole system. To capture the whole system, we need to capture data at every mesh cell in the reactive-transport model simulation. This will result in the formation of a tensor, which is a multi-dimensional array of geochemical data. One can then use the extension of NMF*k* to tensors, non-negative tensor factorization with customized *k*-means clustering (NTF*k*) (Vesselinov et al., 2019), to extract hidden spatio-temporal patterns in the data. NTF*k* analyses will demonstrate the system evolution more comprehensively. Our approach can be easily extended to look at the fate of mineral trapping for an extensive period of time (e.g., million years).

CRediT authorship contribution statement

Bulbul Ahmmed: Data curation, Formal analysis, Writing - original draft. **Satish Karra:** Supervision, Writing - review & editing. **Velimir V. Vesselinov:** Software, Investigation. **Maruti K. Mudunuru:** Software, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they do not have conflict of interest.

Acknowledgments

The authors thank Riaz H. Khan of CEGIS, Bangladesh and Martin Appold of University of Missouri, USA for making the reactive-transport simulation thesis and data for Farnsworth hydrocarbon unit available online at https://mospace.umsystem.edu/xmlui/handle/10355/66744. The authors also thank Christopher Holle for assisting in generating data. BA and SK thank Center for Space and Earth Science and Information Science & Technology Institute at Los Alamos National Laboratory. MKM is partially supported by ExaSheds project during the paper writing process, which was supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Earth and Environmental Systems Sciences Division, Data Management Program, under Award Number DE-AC02-05CH11231. The Los Alamos publication number is LA-UR-21-20069.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ijggc.2021.103382.

References

Ahmmed, B., 2015. Numerical modeling of CO_2-water-rock interactions in the Farnsworth, Texas Hydrocarbon Unit, USA. University of Missouri. Master's thesis.
Ahmmed, B., Appold, M., Fan, T., McPherson, B., Grigg, R., White, M., 2016. Chemical effects of carbon dioxide sequestration in the Upper Morrow Sandstone in the Farnsworth, Texas, hydrocarbon unit. Environ. Geosci. 23 (2), 81–93.
Ahmmed, B., Vesselinov, V., Mudunuru, M., Middleton, R., Karra, S., 2021. Geochemical characteristics of low-, medium-, and hot-temperature geothermal resources of the Great Basin, USA. World Geothermal Congress, Reykjavik, Iceland.
Alexandrov, B., Vesselinov, V.V., 2014. Blind source separation for groundwater pressure analysis based on nonnegative matrix factorization. Water Resour. Res. 50 (9), 7332–7347.
Arnorsson, S., Stefansson, A., 1999. Assessment of feldspar solubility constants in water in the range of 0° to 350°C at vapor saturation pressures. Am. J. Sci. 299 (3), 173–209.
Audigane, P., Gaus, I., Czernichowski-Lauriol, I., Pruess, K., Xu, T., 2007. Two-dimensional reactive transport modeling of CO_2 injection in a saline aquifer at the Sleipner site, North Sea. Am. J. Sci. 307 (7), 974–1008.
Bachu, S., 2015. Review of CO_2 storage efficiency in deep saline aquifers. Int. J. Greenhouse Gas Control 40, 188–202.
Chen, L., Wang, M., Kang, Q., Tao, W., 2018. Pore-scale study of multiphase multicomponent reactive transport during CO_2 dissolution trapping. Adv Water Resour 116, 208–218.
Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.I., 2009. Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation. John Wiley & Sons.
Dai, Z., Xu, L., Xiao, T., McPherson, B., Zhang, X., Zheng, L., Dong, S., Yang, Z., Soltanian, M., Yang, C., Ampomah, W., Jia, W., Yin, S., Xu, T., Bacon, D., Viswanathan, H., 2020. Reactive chemical transport simulations of geologic carbon sequestration: methods and applications. Earth Sci. Rev. 208, 103265.
Ennis-King, J., Paterson, L., 2007. Coupling of geochemical reactions and convective mixing in the long-term geological storage of carbon dioxide. Int. J. Greenhouse Gas Control 1 (1), 86–93.
Gillis, N., 2020. Nonnegative Matrix Factorization. SIAM.
Gunter, W.D., Wiwehar, B., Perkins, E.H., 1997. Aquifer disposal of CO_2-rich greenhouse gases: extension of the time scale of experiment for CO_2-sequestering reactions by geochemical modelling. Mineral. Petrol. 59 (1–2), 121–140.
Iliev, F.L., Stanev, V.G., Vesselinov, V.V., Alexandrov, B.S., 2018. Nonnegative matrix factorization for identification of unknown number of sources emitting delayed signals. PLoS ONE 13, e0193974.
Khan, R., 2017. Evaluation of the geologic CO_2 sequestration potential of the Morrow B Sandstone in the Farnsworth, Texas hydrocarbon field using reactive transport modeling. University of Missouri–Columbia. Master's thesis.
Knauss, K., Johnson, J., Steefel, C., 2005. Evaluation of the impact of CO_2, co-contaminant gas, aqueous fluid and reservoir rock interactions on the geologic sequestration of CO_2. Chem. Geol. 217 (3–4), 339–350.
Kulik, D., Aja, S., 1997. Hydrothermal stability of illite: implications of empirical correlations and Gibbs energy minimization. Proceedings on the Fifth International Symposium on Hydrothermal Reactions, Gatlinburg, Tennessee, pp. 228–292.

Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791.
Lindeberg, E., Wessel-Berg, D., 1997. Vertical convection in an aquifer column under a gas cap of CO_2. Energy Convers. Manage. 38, S229–S234.
Liu, F., Lu, P., Zhu, C., Xiao, Y., 2011. Coupled reactive flow and transport modeling of CO_2 sequestration in the Mt. Simon sandstone formation, Midwest USA. Int. J. Greenhouse Gas Control 5 (2), 294–307.
Masson-Delmotte, V., Zhai, P., Pörtner, H.O., Roberts, D., Skea, J., Shukla, P.R., Pirani, A., Moufouma-Okia, W., Péan, C., Pidcock, R., Connors, S., Matthews, J.B.R., Chen, Y., Zhou, X., Gomis, M.I., Lonnoy, E., Maycock, T., Tignor, M., Waterfield, T., 2018. Global Warming of 1.5°C. An IPCC Special Report on the Impacts of Global Warming of 1.5°C above Pre-industrial Levels and Related Global Greenhouse Gas Emission Pathways, in the Context of Strengthening the Global Response to the Threat of Climate Change, Sustainable Development, and Efforts to Eradicate Poverty. Technical Report. IPCC.
Menad, N., Hemmati-Sarapardeh, A., Varamesh, A., Shamshirband, S., 2019. Predicting solubility of CO_2 in brine by advanced machine learning systems: application to carbon capture and sequestration. J. CO_2 Util. 33, 83–95.
Palandri, J., Kharaka, Y., 2004. A Compilation of Rate Parameters of Water-mineral Interaction Kinetics for Application to Geochemical Modeling. Technical Report. Geological Survey Menlo Park CA.
Rimstidt, J., 1997. Quartz solubility at low temperatures. Geochim. Cosmochim. Acta 61 (13), 2553–2558.
Rosenbauer, R., Koksalan, T., Palandri, J., 2005. Experimental investigation of CO_2–brine–rock interactions at elevated temperature and pressure: implications for CO_2 sequestration in deep–saline aquifers. Fuel Process. Technol. 86 (14–15), 1581–1597.
Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.
Siqueira, T., Iglesias, R., Ketzer, J., 2017. Carbon dioxide injection in carbonate reservoirs–a review of CO_2-water-rock interaction studies. Greenhouse Gases Sci. Technol. 7 (5), 802–816.
Spycher, N., Pruess, K., Ennis-King, J., 2003. CO_2–H_2O mixtures in the geological sequestration of CO_2. I. Assessment and calculation of mutual solubilities from 12 to 100 °c and up to 600 bar. Geochim. Cosmochim. Acta 67 (16), 3015–3031.
Vesselinov, V., Mudunuru, M., Karra, S., O'Malley, D., Alexandrov, B., 2019. Unsupervised machine learning based on non-negative tensor factorization for analyzing reactive mixing. J. Comput. Phys. 395, 85–104.
Vesselinov, V.V., Alexandrov, B.S., O'Malley, D., 2018. Contaminant source identification using semi-supervised machine learning. J. Contam. Hydrol. 212, 134–142.
Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S., 2001. Constrained *k*-means clustering with background knowledge. Proceedings of the Eighteenth International Conference on Machine Learning, Williams College, Williamstown, MA, USA, June 28–July 1, pp. 577–584.
Wang, T., Wang, H., Zhang, F., Xu, T., 2013. Simulation of CO_2–water–rock interactions on geologic CO_2 sequestration under geological conditions of China. Mar. Pollut. Bull. 76 (1–2), 307–314.
Wolery, T., 1992. EQ3/6, a software package for geochemical modeling of aqueous systems: package overview and installation guide (Version 7.0).
Xu, T., Sonnenthal, E., Spycher, N., Pruess, K., 2006. TOUGHREACT–a simulation program for non-isothermal multiphase reactive geochemical transport in variably saturated geologic media: applications to geothermal injectivity and CO_2 geological sequestration. Comput. Geosci. 32 (2), 145–165.
Xu, T., Zheng, L., Tian, H., 2011. Reactive transport modeling for CO_2 geological sequestration. J. Pet. Sci. Eng. 78 (3–4), 765–777.
Zhang, J., Feng, Q., Zhang, X., Shu, C., Wang, S., Wu, K., 2020. A machine learning approach for accurate modeling of CO_2-brine interfacial tension with application in identifying the optimum sequestration depth in saline aquifers. Energy Fuels.
Zhang, W., Li, Y., Xu, T., Cheng, H., Zheng, Y., Xiong, P., 2009. Long-term variations of CO_2 trapped in different mechanisms in deep saline formations: a case study of the Songliao Basin, China. Int. J. Greenhouse Gas Control 3 (2), 161–180.