Journal of Contaminant Hydrology xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Journal of Contaminant Hydrology

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

Nonnegative tensor factorization for contaminant source identification

Velimir V. Vesselinov[a,*], Boian S. Alexandrov[b], Daniel O'Malley[a]

[a] Computational Earth Science Group, Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, NM, USA
[b] Physics and Chemistry of Materials Group, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM, USA

ARTICLE INFO

Keywords:
Nonnegative tensor factorization
Tucker decomposition
Feature Extraction
Exploratory analysis
Blind Source Separation
Robustness analysis
Unsupervised machine learning
Groundwater contamination
Source identification
Advection-diffusion transport
Geochemical signatures

ABSTRACT

Unsupervised Machine Learning (ML) is becoming increasingly popular for solving various types of data analytics problems including feature extraction, blind source separation, exploratory analyses, model diagnostics, etc. Here, we have developed a new unsupervised ML method based on Nonnegative Tensor Factorization (NTF) for identification of the original groundwater types (including contaminant sources) present in geochemical mixtures observed in an aquifer. Frequently, groundwater types with different geochemical signatures are related to different background and/or contamination sources. The characterization of groundwater mixing processes is a challenging but very important task critical for any environmental management project aiming to characterize the fate and transport of contaminants in the subsurface and perform contaminant remediation. This task typically requires solving complex inverse models representing groundwater flow and geochemical transport in the aquifer, where the inverse analysis accounts for available site data. Usually, the model is calibrated against the available data characterizing the spatial and temporal distribution of the observed geochemical types. Numerous different geochemical constituents and processes may need to be simulated in these models which further complicates the analyses. Additionally, the application of inverse methods may introduce biases in the analyses through the assumptions made in the model development process. Here, we substitute the model inversion with unsupervised ML analysis. The ML analysis does not make any assumptions about underlying physical and geochemical processes occurring in the aquifer. Our ML methodology, called NTF*k*, is capable of identifying (1) the unknown number of groundwater types (contaminant sources) present in the aquifer, (2) the original geochemical concentrations (signatures) of these groundwater types and (3) spatial and temporal dynamics in the mixing of these groundwater types. These results are obtained only from the measured geochemical data without any additional site information. In general, the NTF*k* methodology allows for interpretation of large high-dimensional datasets representing diverse spatial and temporal components such as state variables and velocities. NTF*k* has been tested on synthetic and real-world site three-dimensional datasets. The NTF*k* algorithm is designed to work with geochemical data represented in the form of concentrations, ratios (of two constituents; for example, isotope ratios), and delta notations (standard normalized stable isotope ratios).

1. Introduction

Characterizing contaminated groundwater sites presents a number of major challenges, and these challenges have remained key areas of research in subsurface hydrology for several decades (Gelhar, 1993; Fetter and Fetter, 1999; Vengosh et al., 2014). One of the major challenges is the manifold of uncertainties that are present in these subsurface environments. Characterizing the source(s) of the contamination is often of paramount importance and this alone comes with uncertainties in the location, geochemical signature, and even the number of contaminant sources. Similarities in the geochemical signatures of different groundwater types, geochemical interference between groundwater types, and complex physical and geochemical processes (advection, diffusion, dispersion, sorption, retardation, precipitation, phase partitioning, biodegradation, biogeochemical reactions, etc) during transport from the source location to the observation location often make source identification extremely challenging. Furthermore, the geochemical measurement data are affected by random and systematic errors which additionally complicates the analyses.

The water at a given time and location in the subsurface is a mixture of water with different origins and geochemical signatures (which we call groundwater types) (Deutsch and Siegel, 1997). These groundwater types might be associated with (potentially contaminated) sources of groundwater recharge or different upstream regions in the subsurface

* Corresponding author.
E-mail address: vvv@lanl.gov (V.V. Vesselinov).

https://doi.org/10.1016/j.jconhyd.2018.11.010
Received 13 March 2018; Received in revised form 20 November 2018; Accepted 26 November 2018

(which can be called background sources). In addition, groundwater flows through different regions of the subsurface with different rock types and geochemical properties that can modify its geochemical signatures via physical and chemical processes (e.g., reactions and ion exchanges). Groundwater samples collected at multiple wells over time can be used to glean information about these groundwater types. The identification of these groundwater types is an important task in characterizing a contaminated aquifer site (Wagner, 1992; Böhlke and Denver, 1995; Lapworth et al., 2012). The typical approach to identifying these groundwater types utilizes numerical models that simulate flow and transport in the aquifer and model calibration techniques to enable the model to accurately reproduce the observed site data (Wagner, 1992; Neupauer et al., 2000; Atmadja and Bagtzoglou, 2001; Michalak and Kitanidis, n.d.; Guan et al., 2006; Mamonov and Tsai, 2013; Hamdi and Mahfoudhi, 2013; Murray-Bruce and Dragotti, 2014; Borukhov and Zayats, 2015). These models are often very complex requiring the simulation of numerous geochemical constituents (cf. (Hammond et al., 2014; Hansen et al., 2017)) which can make these analyses computationally expensive, often requiring compromises between fidelity to the physics/chemistry and computational efficiency. This is closely related to contaminant source zone identification for which numerous sophisticated methods have been developed. One common approach uses partitioning tracers to estimate heterogeneous permeability and nonaqueous phase liquid saturation (Jin et al., 1995; James et al., 2000; Zhang and Graham, 2001; Yeh and Zhu, n.d.; Illman et al., 2010). Additionally, these complex model-based approaches often require the set-up of site-specific grids in real-world cases (Gzyl et al., 2014). The overarching theme of these approaches is to combine a numerical model with an optimization scheme (Ayvaz, 2010), though sometimes model uncertainty is also considered (Sun et al., 2006).

Recently, methods for analyzing sources of groundwater contamination have been developed that utilize machine learning (ML) and statistical techniques (Chan and Huang, 2003; Rasekh and Brumbelow, 2012). Methods such as factor (Harman, 1976) and principal component (Jolliffe, 2002) analysis have been used to describe variations and evolution in the chemical composition of water types (Knudson et al., 1977; Helena et al., 2000). In addition, unsupervised ML techniques such as discriminant (Scholkopft and Mullert, 1999) and clustering (Diday and Simon, 1980) analysis can group objects into two or more classes (Shrestha and Kazama, 2007; Tariq et al., 2008). Unsupervised ML based on nonnegative matrix factorization (NMF) methods (Throckmorton et al., 2016; Vesselinov et al., n.d.-a) have been used to identify groundwater types and their mixing ratios.

Another approach is to use supervised ML techniques (such as neural networks (Yegnanarayana, 2009), support vector machines (Drucker et al., 1999), locally weighted projection regression (Vijayakumar and Schaal, 2000), and relevance vector machines (Tipping, 2001)) to replace or supplement the complex numerical models previously mentioned. These ML-developed models can be used to make predictions related at groundwater contamination sites (Khalil et al., n.d.). Quasi-optimal learning (Cervone et al., 2010) has been used to explore a symbolic supervised ML classification method to understand the relationship between different chemical species in ground and surface water. The drawback of the supervised ML methods compared to the unsupervised ML method is that they require extensive training based on subject-matter expertise, existing site data or physics-model outputs. The process is computationally intensive and can introduce bias in the analyses.

The unsupervised nonnegative tensor factorization (NTF) approach proposed here is similar to nonnegative matrix factorization (NMF) methods developed recently (Throckmorton et al., 2016; Vesselinov et al., n.d.-a). To understand the advance from the NMF method to the NTF method, we must first consider the structure of the data. The data that are assimilated by these methods comes from the observation of (1) chemical species at different (2) locations and (3) times. The NMF methods can only consider variability in two of these three components at once, for example, observations of different species at different locations but at a fixed time (Vesselinov et al., n.d.-a). This comes from the fact that a matrix has two indices—one can be associated with the locations and another with the species, but none remains to be associated with the different times. The NTF method allows for an arbitrary number of indices enabling it to consider variability in all three (species, location, time) providing an advantage compared to NMF. Both the NMF and NTF methods provide a means of analysis that does not rely on complex inverse models, making it less computationally expensive and with fewer assumptions built-in.

The main goal of the paper is to present and demonstrate the applicability of a novel unsupervised ML algorithm called NTF*k*. NTF*k* performs a Blind Source Separation (BSS) analysis (Belouchrani et al., 1997), based on Nonnegative Tensor Factorization (NTF) (Cichocki et al., 2009), combined with a custom clustering algorithm (Vesselinov et al., n.d.-a; Alexandrov and Vesselinov, 2014). Here, NTF*k* is applied to unmix the geochemical signatures in the observations and identify the contaminant sources. As a result, NTF*k* is capable of identifying (1) the unknown number of groundwater types (contaminant sources) present in the aquifer, (2) the original geochemical concentrations (signatures) of these groundwater types and (3) spatial and temporal dynamics in the mixing of these groundwater types. Since the problem involves mixing, NTF*k* here is implemented applying additional optimization constraints. NTF*k* is a high-dimensional extension of our existing matrix-based NMF*k* methodology developed in (Throckmorton et al., 2016; Vesselinov et al., n.d.-a; Alexandrov and Vesselinov, 2014).

Using synthetic and real-world site data, we demonstrate that NTF*k* is capable of accurately determining the unknown number of contaminant sources from observation samples of their mixtures, without any additional information. In addition, our methodology can also estimate the source locations based on the estimated mixing coefficients (at the monitoring wells) and monitoring well location coordinates. Our methodology also allows for generation of spatial and temporal maps of estimated contaminant mass distribution in the subsurface. The NTF*k* methodology is coded in Julia (Bezanson et al., n.d.) and an open-source code implementing our algorithm will be released soon. The NTF*k* algorithm works with geochemical data represented in the form of concentrations, ratios (of two constituents, for example, isotope ratios), and delta notations (standard normalized stable isotope ratios). Despite the methodological complexities discussed below, the algorithm is fast and relatively easy to implement.

2. Methodology

2.1. Blind source separation (BSS)

In the analyses discussed here, we assume that the geochemical observations are taken at several detectors (sampling points; typically monitoring wells) distributed in space. When there are multiple contamination sources in the aquifer each detector registers a mixture of contamination fields (plumes) over time originating from different sources (release locations). Our objective is to identify the unknown number of original contamination sources, which necessitates decomposing the recorded transient mixtures to their original components. Through the ML analyses, we also identify geochemical concentrations (signatures) of the original sources and characterize spatial and temporal dynamics in the mixing of contaminant sources in the aquifer. These results are obtained only based on the observed concentration data without any other site information.

Our novel unsupervised ML methodology, NTF*k*, is a method for feature extraction and exploratory analysis capable of revealing features hidden in data. NTF*k* is based on NTF, which is an emerging research area in the field of data analytics and data compression (Cichocki et al., 2009; Kolda and Bader, 2009). In addition to extracting hidden features that are buried in large high-dimensional datasets, NTF-based methods are also used in blind source separation. BSS techniques

are typically based on matrix factorization methods such as Principal Component Analysis (PCA) (Jolliffe, 1986), Independent Component Analysis (ICA) (Amari et al., 1996), and NMF (Paatero and Tapper, 1994). These techniques form a class of unsupervised machine learning (ML) methods that are instrumental in model-free feature extraction and dimensionality reduction. When a BSS technique is applied in signal processing, the extracted features are the unique original signals that form the mixtures recorded by a set of spatially distributed sensors (e.g., the voices of several speakers recorded by multiple microphones placed at different locations in a ballroom (Haykin and Chen, 2005)). However, the matrix-based methods are inherently deficient for examining high-dimensional datasets (i.e., tensor datasets), which are natural extensions of the matrix datasets. Many real-world datasets are high-dimensional and often represent one or more state variables at a discrete set of locations in space and time, and, as a result, are ideal for tensor-based analyses.

There are multiple tensor factorization methods (Hitchcock, 1927; Harshman and Lundy, 1994; De Lathauwer et al., 2000) and, among them, we utilize the Tucker decomposition (Tucker, 1966; Andersson and Bro, 2000). Examples of Tucker models for three-dimensional datasets are presented in Fig. 1; note that multiple possible Tucker models can be used (there are 7 possible Tucker models in the three-dimensional case: 1 with 3 factor matrices, 3 with 2 factor matrices, and 3 with 1 factor matrix). To apply Tucker decomposition to a given dataset, we need to find not only which of the possible models to use, but we also need to identify the size of core tensor (G in Fig. 1). Typically, we do not have *prior* knowledge about the specific Tucker model and the core size. To find the optimal decomposition model and core size, NTF*k* applies analyses of the NTF solution robustness and parsimony as discussed in Section 2.2 below.

Herein, using NTF*k*, we analyze three-dimensional data that represent the evolution in time and space of concentrations of a series of geochemical species observed at a series of monitoring wells in time by Tucker decomposition.

The analyzed data-tensor C has three dimensions: (s,w,t), where s indicates a geochemical species, w a monitoring well and t an observation time. The Tucker-3 decomposition (Fig. 1) of the three-dimensional tensor $C(s,w,t)$:

$$C(s,w,t) = G \otimes W(s) \otimes H(w) \otimes V(t) + \varepsilon(w,s,t) \tag{1}$$

where $\otimes$ denotes a tensor product. The decomposition of the tensor $C(s,w,t)$ ($C \in \mathbb{R}_{\geq 0}^{K \times M \times N}$) can be expressed by components:

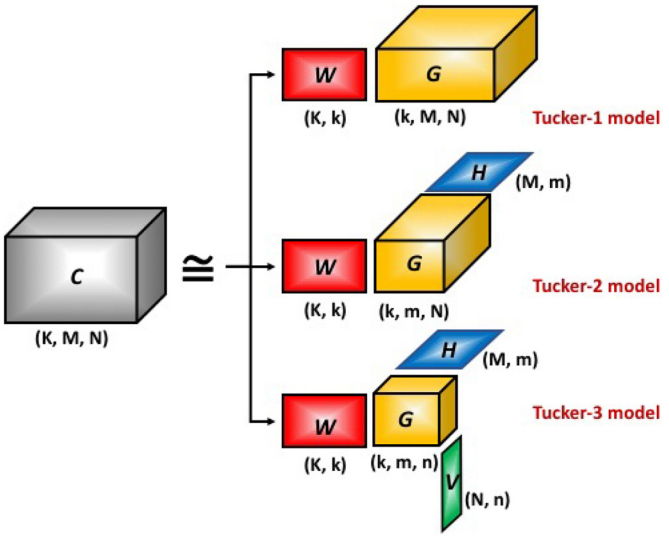


Fig. 1. Schematic representation of various Tucker-based factorization models for three-dimensional tensors. Herein, we employ Tucker-1 model to decompose the tensor $C(w,t,s)$ into a core tensor G and a factor matrix W.

$$C_{ijl} = \sum_{p=1}^{k} \sum_{q=1}^{m} \sum_{r=1}^{n} G_{pqr} W_{ip} H_{jq} V_{lr} + \varepsilon_{ijl} \quad \forall\, i,j,l \tag{2}$$

where all the elements of C, G, W, H, and V are nonnegative,

$$C_{ijl},\ G_{pqr},\ W_{ip},\ H_{jq},\ V_{lr} \geq 0 \quad \forall\, i,j,l,p,q,r. \tag{3}$$

Here, i ranges from 1 to K where K is the number of geochemical species, j ranges from 1 to M where M is the number of monitoring wells, and l ranges from 1 to N where N is the number of time frames (snapshots). The NTF*k* methodology allows the tensor C to be sparse (i.e., some of the observations can be missing).

In this case, the Tucker decomposition includes (i) a core tensor G ($G \in \mathbb{R}_{\geq 0}^{k \times m \times n}$) that represents the interactions between the s, w, and t components of $W(s)$, $H(w)$ and $V(t)$; (ii) a factor matrix W ($W \in \mathbb{R}_{\geq 0}^{K \times k}$) representing geochemical signatures of each groundwater type; (iii) a factor matrix H ($H \in \mathbb{R}_{\geq 0}^{M \times m}$) accounting for dependence on the monitoring points, and (iv) a factor matrix V ($V \in \mathbb{R}_{\geq 0}^{N \times n}$) that captures the time dependence. $\mathbb{R}_{\geq 0}$ denotes the set of nonnegative real numbers $\mathbb{R}_{\geq 0} = \{x \in \mathbb{R} | x \geq 0\}$. Additionally, ε ($\varepsilon \in \mathbb{R}^{K \times M \times N}$) in Eq. (2) denotes the unknown discrepancy between the original data C and the Tucker estimate $\widetilde{C}$ ($\widetilde{C} = G \otimes W \otimes H \otimes V$); The discrepancy ε can be caused by the presence of random measurement errors in the data tensor C. The discrepancy can also be caused by the inadequacy of the Tucker decomposition $\widetilde{C}$ to represent the data. The Tucker decomposition $\widetilde{C}$ can be viewed as linear combinations of geochemical, spatial (well location), and temporal features where each of these features can have any complex nonlinear shape. The features are represented in the factor matrices (W, H, and V) and the linear combinations among them are given by the core tensor G. If there are nonlinear interactions among the features, NTF*k* cannot resolve them in its current form. However, for the geochemical analyses presented here, nonlinear interactions were not needed to characterize data.

Mathematically, the solution of the nonnegative Tucker tensor decomposition is a solution of a multi-dimensional optimization problem with nonnegative constraints given by:

$$\min_{G,W,H,V \geq 0} \|C - G \otimes W \otimes H \otimes V\|_F^2 \tag{4}$$

To extract the unknown core tensor G, and factor matrices W, H, and V, different optimization algorithms can be applied.

To solve the geochemical problems discussed here, we reduce the nonnegative Tucker-3 decomposition presented in Eq. (2) to Tucker-1 where (Fig. 1):

$$C_{ijl} = \sum_{p=1}^{k} G_{pjl} W_{ip} + \varepsilon_{ijl} \quad \forall\, i,j,l \tag{5}$$

Now, the Tucker decomposition includes only (i) an unknown core tensor G ($G \in \mathbb{R}_{\geq 0}^{k \times M \times N}$), and (ii) an unknown factor matrix W ($W \in \mathbb{R}_{\geq 0}^{K \times k}$) representing the changes in C associated with geochemical species (the species-component). Here, the W matrix can be viewed as a "source" matrix representing concentrations of K geochemical species in k contaminant sources (groundwater types). The core, G, represents the mixing ratios of these k contaminant sources (groundwater types) at each well over time. For example, $G_{1,2,3}$ will define the mixing ratio of the source (groundwater type) 1 in well 2 at time frame 3. Therefore, here we assume that the observational data, C, is formed by a linear mixing of k original signals represented by the "source" matrix W and blended by a mixing core tensor G at each observation point and time.

In addition, we impose constraints on the core tensor elements:

$$\sum_{j=1}^{M} G_{pjl} = 1 \quad \forall\, p,l \tag{6}$$

where we require that all the mixing ratios at each time for each monitoring point (well) add up to 1. These constraints represent

conservation of mass.

To analyze the tensor C, we utilize a constrained version of the sparse nonnegative Tucker-1 decomposition model (Mørup et al., 2008). Our constraints are imposed so that the tensor decomposition accounts for the underlying mixing processes; the constraints are similar to the approach applied by (Vesselinov et al., n.d.-a) for the matrix-factorization problem. Our choice for nonnegative constraints is motivated by (i) the fact that concentrations are inherently nonnegative and (ii) our goal to relate the extracted features to easily interpretable quantities without introducing any *prior* assumptions. Indeed, a meaningful interpretation of the obtained results requires the extracted features to be parts of the original data (Lee and Seung, 1999) and the nonnegative constraints lead to extraction of strictly additive components, which are parts of the original data (Ross and Zemel, 2006). Thus, NTF*k* has the ability to identify readily understandable structure-preserving features that enable the discovery of new causal structures and unknown mechanisms hidden in the data (Cichocki et al., 2009).

2.2. NTFk algorithm

The NTF*k* algorithm starts with a random guess for W and G elements, and proceeds by minimizing the cost (objective) function, O, which in our case is the Frobenius norm,

$$O = \frac{1}{2}\|C - G \otimes W\|_F^2 \tag{7}$$

during each iteration. Minimizing the Frobenius norm (Eq. (7)) with nonnegativity constraints (Eq. (6)) is equivalent to representing the discrepancies between the observations, C, and the reconstruction, $G \otimes W$, as white noise.

Due to the constraints in Eq. (6), the classical multiplicative NTF optimization algorithms (Cichocki et al., 2009) are not applicable. Instead, a nonconvex nonlinear optimization algorithm is needed, and for this purpose, we utilized the nonlinear minimization procedure provided by Julia packages JuMP.jl and Ipopt.jl. JuMP.jl is a modeling language for mathematical optimization embedded in Julia (Dunning et al., n.d.). It supports a number of open-source and commercial solvers for a variety of optimization problems. JuMP.jl is coupled with Ipopt (Interior Point OPTimizer): an open-software package for large-scale nonlinear optimization (Wächter, 2002; Wächter and Biegler, 2005; Wächter and Biegler, 2006). Here, Ipopt is applied to perform nonconvex constrained second-order minimization.

If we knew the number of sources k, solving Eq. (7) is all we need to perform: the best solution of Eq. (7), we would estimate matrix/tensor elements and solve the inverse problem. However, the true number of sources is typically unknown, and thus the number of the sources is an unknown parameter which we have to identify from the observations.

A naive approach would be to (1) explore all of the possible solutions of Eq. (7) for a range of a possible number of sources k and (2) select the solution with the smallest norm to identify the number of sources, k_s. However, this is a flawed approach—more free parameters (higher k) will generally lead to a better fit, irrespective of how close the estimated number of sources is to the actual number of sources. This would cause the naive approach to over-estimate the number of sources:

To resolve this issue, NTF*k* considers all possible numbers of sources k ranging from 1 to d ($k = 1, 2, \ldots, d$). For each value of k, Z different factorizations are performed with different random initial guesses. NTF*k* then estimates the accuracy and robustness of the large set of solutions Z with a different number of sources. In NTF*k*, the maximum number of explored sources d should not exceed the expected number of observed geochemical components K (although, theoretically, the minimization algorithm used here can be applied for any $k > 1$).

Thus, NTF*k* performs Z sets of simulations, called NTF runs, where each run is using a different number of sources, $k = 1, 2, \ldots, d$, with random initial guesses for all the unknown matrix/tensor elements. At the end of each NTF run, we get a set of Z solutions, U_k, where each solution contains two arrays: the matrix W_k^j and the tensor G_k^j, (for k original sources, and $j = 1, 2, \ldots, Z$),

Table 1
True and estimated concentrations of four geochemical constituents (A, B, C & D) representing three synthetic sources (S1, S2 & S3).

Source	True				Estimated			
	A	*B*	*C*	*D*	*A*	*B*	*C*	*D*
S1	0.326	0.071	1.000	1.000	0.316	0.031	1.043	1.146
S2	1.000	1.000	0.368	0.026	1.106	1.121	0.301	0.004
S3	0.209	0.134	0.820	0.013	0.115	0.033	0.871	0.002

Table 2
True and estimated concentrations of the four geochemical constituents (A, B, C & D) observed at five observation points for five time frames; note that no observation errors are introduced when the true concentrations were computed.

Well	Time	True				Estimated			
		A	*B*	*C*	*D*	*A*	*B*	*C*	*D*
W1	1	0.209	0.134	0.820	0.013	0.209	0.134	0.820	0.013
W2	1	1.000	1.000	0.368	0.026	1.000	1.000	0.368	0.026
W3	1	0.995	0.994	0.371	0.026	0.995	0.994	0.371	0.026
W4	1	0.692	0.663	0.544	0.021	0.692	0.663	0.544	0.021
W5	1	0.999	0.999	0.368	0.027	0.999	0.999	0.368	0.027
W1	2	0.209	0.134	0.820	0.013	0.209	0.134	0.820	0.013
W2	2	1.000	1.000	0.368	0.027	1.000	1.000	0.368	0.027
W3	2	0.998	0.998	0.369	0.026	0.998	0.998	0.369	0.026
W4	2	0.486	0.291	0.850	0.769	0.486	0.291	0.850	0.769
W5	2	0.966	0.953	0.400	0.076	0.966	0.953	0.400	0.076
W1	3	0.966	0.955	0.398	0.069	0.966	0.955	0.398	0.069
W2	3	0.210	0.135	0.820	0.013	0.210	0.135	0.820	0.013
W3	3	0.210	0.135	0.820	0.013	0.210	0.135	0.820	0.013
W4	3	0.326	0.071	1.000	1.000	0.326	0.071	1.000	1.000
W5	3	1.000	1.000	0.368	0.026	1.000	1.000	0.368	0.026
W1	4	0.427	0.258	0.843	0.602	0.427	0.258	0.843	0.602
W2	4	0.327	0.071	1.000	1.000	0.327	0.071	1.000	1.000
W3	4	0.381	0.147	0.948	0.920	0.381	0.147	0.948	0.920
W4	4	0.210	0.135	0.819	0.013	0.210	0.135	0.819	0.013
W5	4	0.326	0.071	1.000	1.000	0.326	0.071	1.000	1.000
W1	5	0.319	0.086	0.974	0.880	0.319	0.086	0.974	0.880
W2	5	0.326	0.071	1.000	1.000	0.326	0.071	1.000	1.000
W3	5	0.338	0.116	0.952	0.834	0.338	0.116	0.952	0.834
W4	5	0.691	0.662	0.544	0.021	0.691	0.662	0.544	0.021
W5	5	0.270	0.101	0.914	0.525	0.270	0.101	0.914	0.525

Table 3
NTF*k* results for the Example problem #1; the reconstruction quality O, silhouette width S, and AIC are estimated for number of sources $k = 2,3,4$.

k	*O*	*S*	*AIC*
2	$2.300 \cdot 10^{+6}$	1.000	1100.324
3	$1.146 \cdot 10^{-7}$	0.997	−1904.693
4	$7.144 \cdot 10^{-8}$	−0.665	−1893.951

Table 4
NTF*k* results for Example problem #2; the reconstruction quality O, silhouette width S, and AIC are estimated for number of sources $k = 2, \ldots, 7$.

k	*O*	*S*	*AIC*
2	$6.072 \cdot 10^{+07}$	1.000	13,085.080
3	$2.752 \cdot 10^{+07}$	1.000	12,477.540
4	$1.176 \cdot 10^{+07}$	1.000	11,801.880
5	$6.284 \cdot 10^{-07}$	0.981	−23,099.440
6	$5.691 \cdot 10^{-07}$	0.049	−22,909.530
7	$5.689 \cdot 10^{-07}$	0.351	−22,605.940

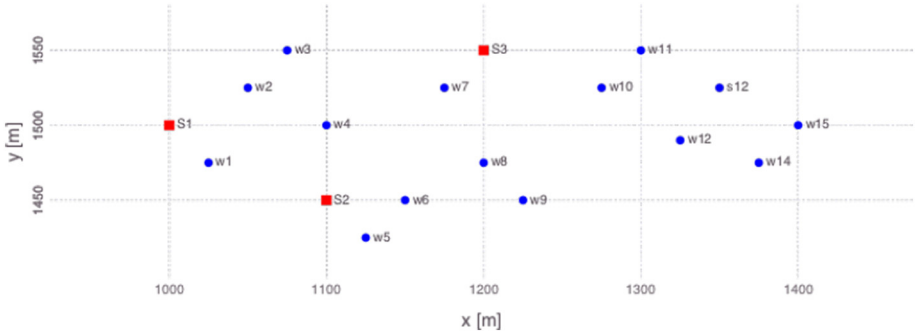


Fig. 2. Synthetic site map showing locations of unknown point sources (red rectangles) and wells (blue circles). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
True and estimated concentrations (*ppm*) of four geochemical constituents (A, B, C & D) representing four synthetic sources (three contaminant sources S1, S2 & S3 and background concentrations).

Source	True				Estimated			
	A	*B*	*C*	*D*	*A*	*B*	*C*	*D*
S1	0.000	0.000	1.000	0.500	0.000	0.000	0.830	0.415
S2	0.000	1.000	0.000	1.0000	0.000	0.865	0.018	0.875
S3	1.000	0.000	0.000	0.000	0.963	0.002	0.0240	0.014
Background	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000

$$U_k = ([W_k^1; G_k^1], [W_k^2; G_k^2], ..., [W_k^Z; G_k^Z]) \quad (8)$$

After that, NTF*k* leverages a custom clustering algorithm to assign each of these Z solutions in a given set, U_d, to one of k specific clusters. This clustering method is based on k-means clustering that keeps the number of solutions in each cluster equal to the number of NTF runs (cf. (Vesselinov et al., n.d.-a; Alexandrov and Vesselinov, 2014)). For example, for the case with $k = 2$, after the execution of $Z = 1{,}000$ NTF runs (performed with random initial guesses for the W and G elements), each of the two clusters will contain 1,000 solutions. In the cases when the NTF problem is under-parametrized (i.e., low number of sources), the final solution is generally not very sensitive to the random initial guesses. This suggests there is a single global minimum which can be identified regardless of the initial guesses for matrix/tensor elements. In the cases when the NTF problem is over-parametrized (i.e., high number of sources), the final solution is generally very sensitive to the initial guesses. This suggests that potentially there are multiple local/global minima which are identified using random initial guesses.

Note that we have to enforce the condition that the clusters have an equal number of solutions, since each NTF simulation contributes an equal number of solutions for each source. During the clustering, the similarity between sources W_{i1} and W_{i2} is measured using the cosine distance (Vesselinov et al., n.d.-a; Alexandrov and Vesselinov, 2014; Pang-Ning et al., 2006). The cosine distance measures the angle between the two sources and effectively ignores their magnitude.

The main idea for estimating the unknown number of sources in NTF*k* is to use the separation between the clusters as a measure of how good a particular choice of k is as an accurate estimate of the number of unknown sources. We estimate the degree of clustering for a different number of sources, and plot it as a function of k, and we expect a sharp drop after we cross k_s (the optimal number of sources) (Vesselinov et al., n.d.-a; Alexandrov and Vesselinov, 2014).

To quantify this behavior, after the clustering, we compute a measure, $S(k)$ (called the average Silhouette width (Rousseeuw, 1987)), of how well the solutions are clustered for a given number of original sources, k. This measure of how well-clustered the NTF*k* solutions are for different values of k can be applied to evaluate the optimal number of contaminant sources, k_s. In general, $S(k)$ declines as k increases. Theoretically, $S(k)$ varies between 1 and -1. When $S(k)$ is close to 1, that indicates that the data is well-clustered (i.e., the average distance between points within a cluster is small compared to the average distance between points in different clusters). As $S(k)$ decreases, the quality of the clusters decreases. Typically, $S(k)$ declines sharply after the optimal number of contaminant sources, k_s, is reached.

In NTF*k*, in addition to the robustness, the average reconstruction error (Eq. (7)) is used to evaluate the accuracy with which the derived average (cluster) solutions $[W_k^a; G_k^a]$ reproduce the observations C. In general, the solution accuracy increases (while the solution robustness decreases) when k goes up. Hence, the average silhouette width and Frobenius norm for each of the k cluster solutions can be used to define the optimal number of contaminant sources, k_s. Specifically, k_s can be set equal to the minimum number of sources that accurately reconstruct the observations (i.e., the Frobenius norm is less than a given value or hit a plateau) and the clusters of solutions are sufficiently robust (e.g., the average silhouette width S is bigger than 0.8).

When some of the source geochemical compositions are very close to each other or do not demonstrate clear features, it is also useful to formulate another criteriion for the NTF*k* solution robustness, which is based on the Akaike Information criterion (*AIC*) (Akaike, 2011). Specifically, to compare the NTF models with a different number of sources, we calculate for each of them the *AIC* value. To calculate *AIC*, we take from each of the sets of solutions with a different number of sources, U_k, the best NTF solution, and use the corresponding Frobenius norm, $O^{(k)}$, in the *AIC* formula:

$$AIC = 2Q - 2\ln(L) = 2(k(MN + K) - MN) + KMN\ln\left(\frac{O^{(k)}}{KMN}\right) \quad (9)$$

Here, the number of adjustable NTF*k* parameters, Q, is equal to the number of components in the matrix W and the tensor G minus the number of constraints for each observation well / time (cf. 6), which reduces the number of adjustable parameters. Thus, we have, $Q = (k-1)MN + Kk = k(MN + K) - MN$, where k is the number of sources, M is the number of wells and N is the number of the observation time frames. L is the likelihood functions of the NTF solution with given k, and we define it using the reconstruction error $O^{(k)}$ of the NTF solutions: $\ln(L) = -(KMN/2)\ln(O^{(k)}/KMN)$ (KMN is the total number of observational data points in the tensor C; if there is missing data, the empty tensor elements are not counted). The *AIC* is a measure of the relative quality of statistical models, which takes into account both the likelihood function (in our case determined by the reconstruction error) and the independent degrees of freedom needed to achieve this level of likelihood (the elements of the matrices W and H). Choosing the model that minimizes *AIC* helps avoid overfitting. In general, *AIC* decreases as the number of sources, k, increases. Typically, *AIC* substantially drops when $k = k_s$. For $k > k_s$, the *AIC* values commonly plateau and do not exhibit substantial changes. Comparisons between different solutions using *AIC* capture the parsimony principal; models with a smaller number of parameters are favored when the

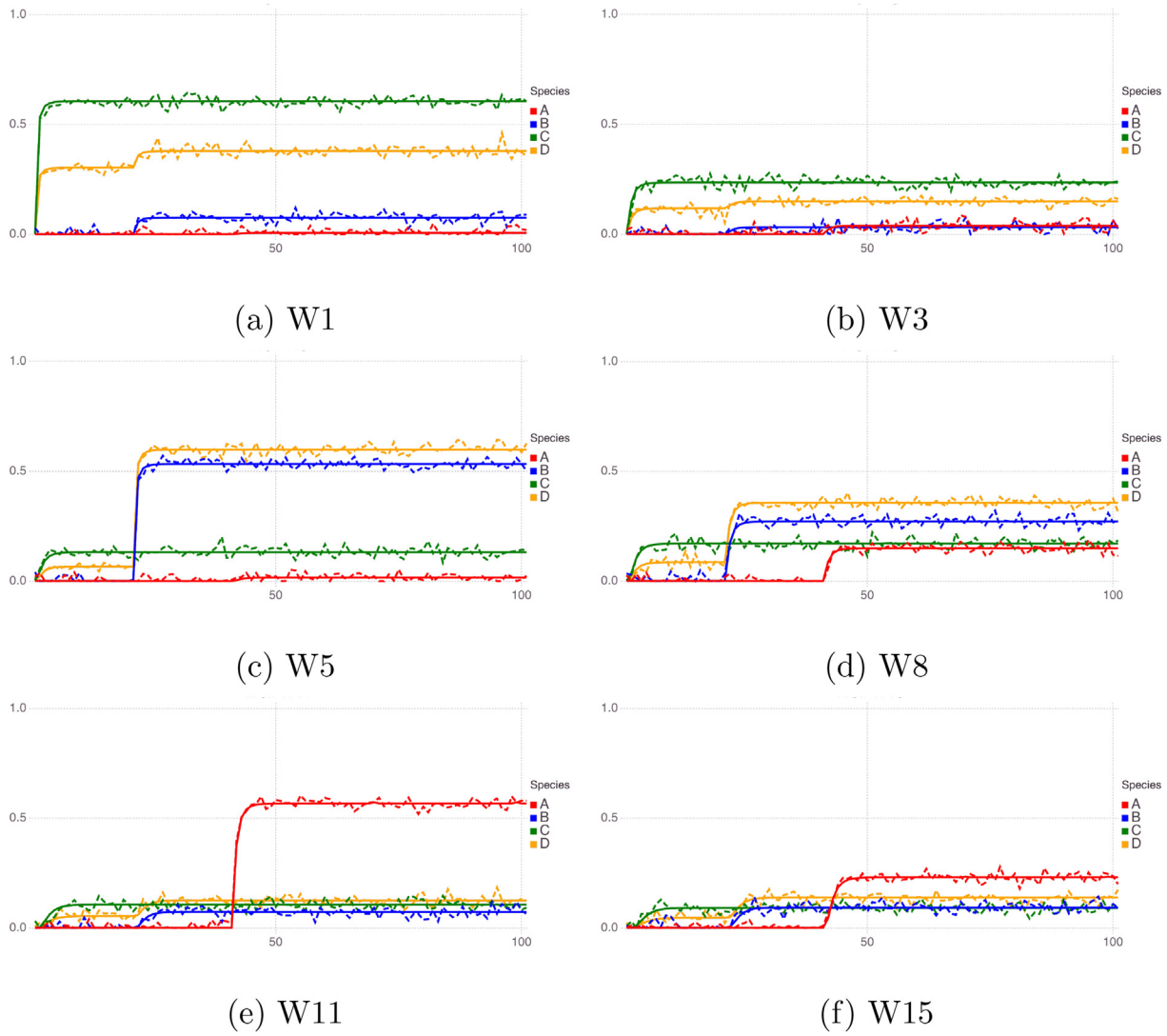


Fig. 3. Transients in the "true" (dashed lines) and estimated (solid lines) concentrations of the four contaminant sources (groundwater types) at six of the monitoring wells; the dashed lines are not seen when the curves overlap. The true concentrations include 10% measurement error. The vertical axis is concentrations in *ppm* and the horizontal axis is time in years.

Table 6
NTF*k* results for Example problem #3; the reconstruction quality *O*, silhouette width *S*, and *AIC* are estimated for number of sources $k = 2, \ldots, 5$.

k	*O*	*S*	*AIC*
2	$9.103 \cdot 10^{7}$	1.000	$6.412 \cdot 10^{4}$
3	$3.497 \cdot 10^{7}$	1.000	$6.136 \cdot 10^{4}$
4	$7.628 \cdot 10^{-7}$	1.000	$-1.262 \cdot 10^{5}$
5	$9.111 \cdot 10^{-7}$	0.710	$-1.221 \cdot 10^{5}$

reconstruction qualities of the models are similar.

In general, both the average silhouette width *S* and *AIC* should estimate the same number of sources k_s. If there is a discrepancy, *S*-based estimate is typically smaller than the *AIC*-based estimate (this type of situation is discussed in the results section below). In general, the *S*-based estimate of k_s should be preferred because the solutions for $k > k_s$ are potentially over fitting the data.

The NTF*k* algorithm is coded in Julia and will be available as open-source code soon. It is fast and easy to use with the only user's input being the processed data tensor.

3. Results

3.1. NTF*k* analysis of synthetic data

First, we apply the NTF*k* unsupervised ML algorithm described above to identify the source concentrations from two synthetic randomly-generated data sets representing scenarios generally consistent with real-world conditions in terms of number of wells, number of geochemical constituents, and the number of temporal observations. These two problems are presented in Sections 3.1.1 and 3.1.2. They are applied to test the NTF*k* algorithm and demonstrate its general applicability. Here, the concentrations are generated randomly and they do not represent an actual groundwater contaminant transport problem. Through the first two synthetic problems, we also demonstrate that NTF*k* can be applied even in situations when the concentrations (and respective mixing ratios) vary erratically. The third synthetic problem presented in the Section 3.1.3 is designed to represent a groundwater contamination problem obtained through model simulation of an advective-dispersive transport in an aquifer.

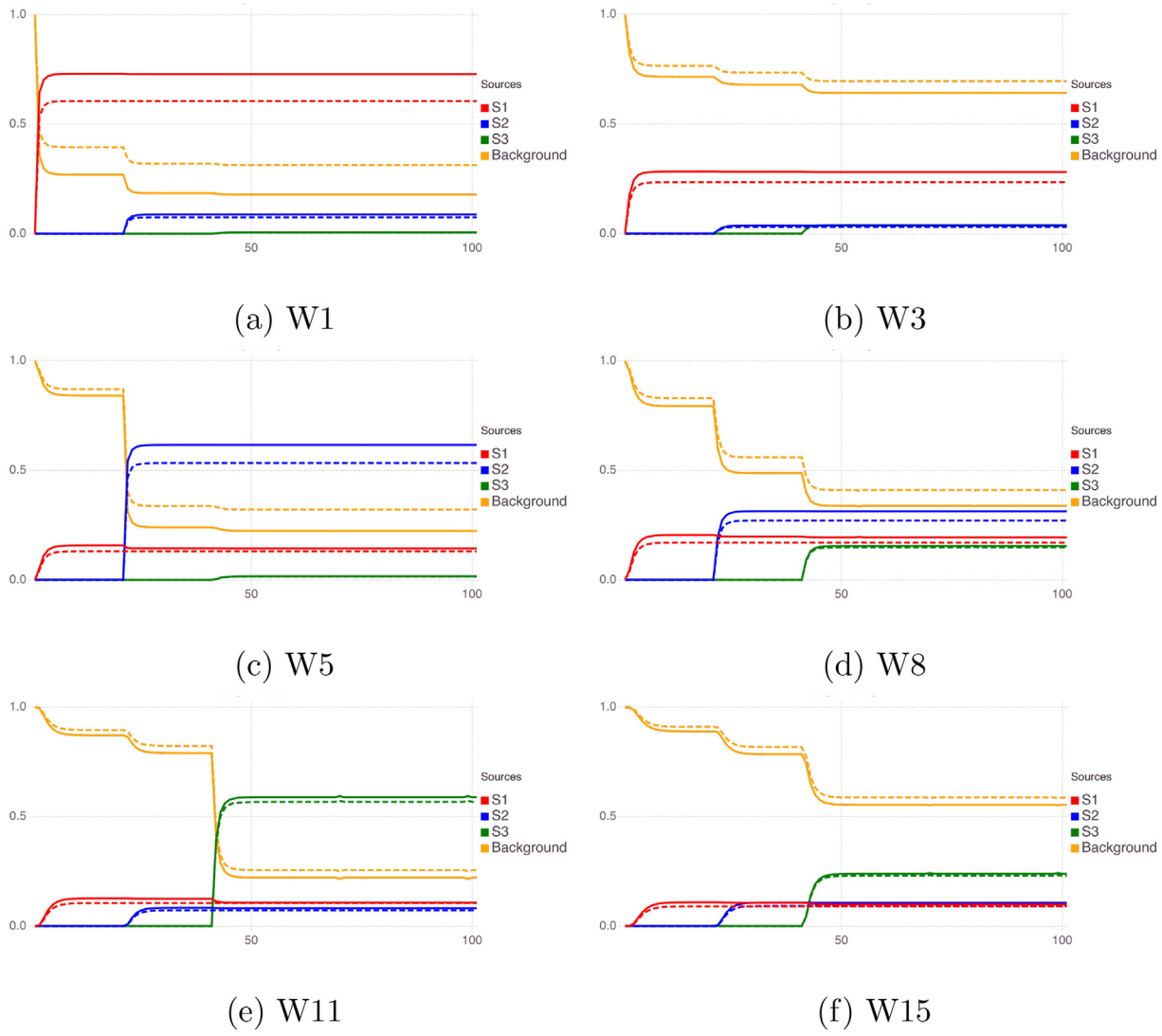


Fig. 4. Transients in the "true" (dashed lines) and estimated (solid lines) mixing coefficients of the four contaminant sources (groundwater types) at six of the monitoring wells. The vertical axis present dimensionless mixing rations between 0 and 1 and the horizontal axis is time in years. Note that these are the actual "true" mixing coefficients without a measurement noise.

3.1.1. Example #1: 3 sources, 4 geochemical constituents, 5 wells and 5 time frames

We consider an example randomly generated to represent three unknown synthetic sources (groundwater types). The "true" concentrations of four geochemical constituents (A, B, C & D) representing the three synthetic sources are presented in Table 1; this is the "true" matrix W (Eq. (7)). These sources are mixed at each well using random mixing coefficient representing the core tensor G (Eq. (7)). The "true" W and G are applied to estimate the "true" concentrations C (Table 2) of four geochemical constituents (A, B, C & D) at five monitoring wells and five different time frames. Here the measurement errors are assumed to be zero. When we apply NTFk, W and G are unknown; the number of sources (groundwater types) is also unknown. The only information provided to NTFk is data tensor C.

Here and in the examples presented below, the source concentrations and well mixing coefficients are generated using standard pseudo random number generation capabilities provided in Julia (Bezanson et al., 2014); the random numbers have uniform distribution between 0 and 1. For convenience and without loss of generality, the source concentrations are scaled so that the maximum concentration at the sources for each species is 1. The random mixing coefficients are also scaled so that the mixing ratios for each well/time frame add up to 1. As discussed above, this requirement comes from the problem setup; the groundwater concentrations at each well are expected to be defined by mixing of all the sources.

We used the data tensor C in NTFk to estimate the number of sources and reconstruct the unknown source concentrations and mixing coefficients at the wells over time. To identity the number of sources, the algorithm performs analyses where the number of sources, k, is equal to 2, 3, and 4. For each of these 3 cases, NTFk processes the reconstruction quality Q, silhouette width S, and AIC. The results are presented in Table 3. Based on Table 3, the number of sources is three. This is estimated based on the behavior of the robustness (silhouette width S) and AIC criteria. The silhouette width S is close to 1 for the cases of 2 and 3 sources; however, it drops substantially for 4 sources. This suggests that the solution for 4 sources is not stable and non-unique. Therefore, the solution for 3 sources should be preferred. Similarly, AIC shows a substantial drop between cases of 2 and 3 sources; this also suggests that the solution with 3 sources should be selected. The same conclusion can be also drawn here by the reconstruction quality. Clearly the solution for 3 sources produces a much better fit to the data than the solution for 2 sources. The solution for 4 sources produces a slightly better match than the solution for 3 sources but using far more model parameters (i.e., more degrees of freedom). In this

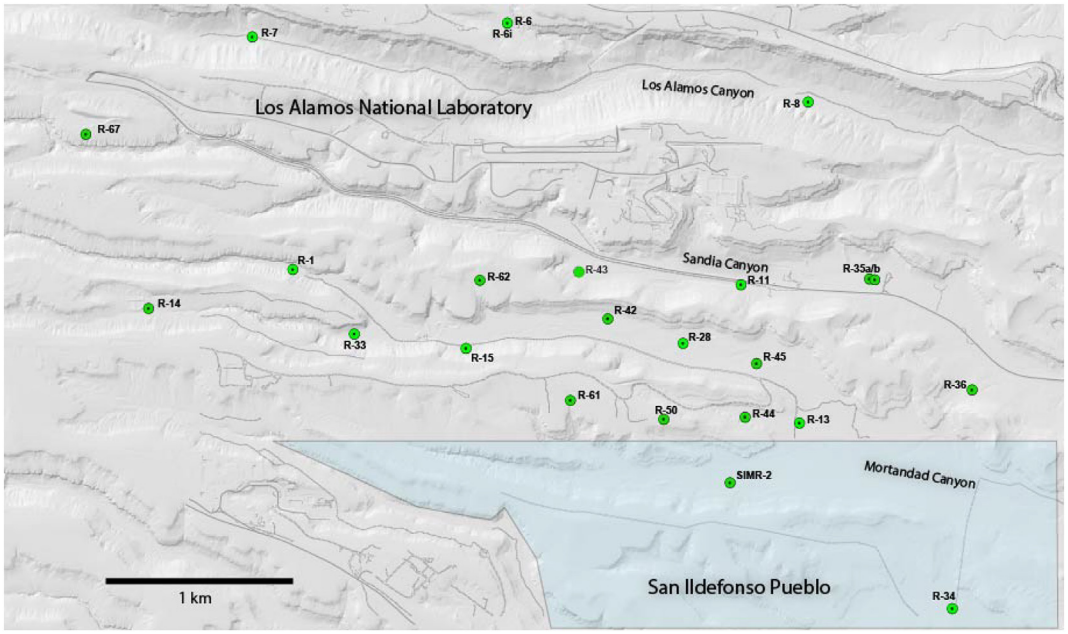


Fig. 5. LANL site map showing locations of some of the monitoring wells.

case, the 3-source solution has 62 adjustable model parameters ($5 \times 5 \times 2 + 3 \times 4$) while the 4-source solutions has 91 adjustable model parameters ($5 \times 5 \times 3 + 4 \times 4$). There are only 100 observations ($5 \times 5 \times 4$) in all cases.

The NTF*k* estimated concentrations of the four geochemical constituents (A, B, C & D) representing three synthetic sources are presented in Table 1. As can be seen, the algorithm accurately estimates the geochemical signatures of the sources. It is also capable of accurately reproducing the observed concentrations (Table 2).

The same synthetic problem was executed 1000 times with different randomly generated concentrations (using different randomly generated mixing coefficients and species concentrations). In all 1000 cases, the algorithm correctly identified the true number of sources. The minimum silhouette width S from the 1000 runs for $k = 3$ was 0.951. The maximum silhouette width S for $k = 4$ was -0.573. This demonstrates that the gap in the minimum silhouette width S between the optimal solution ($k = 3$) and the next solution with one extra source ($k = 4$) is sufficiently large and this criteria is adequate by itself to select the optimal number of sources in all the 1000 test cases.

The same synthetic problem was also rerun 1000 times adding random noise to the concentrations in the data tensor C; the applied noise is normally distributed (mean equal to zero and standard deviation equal to 0.01) representing random measurement errors. Again, the algorithm correctly identified the true number of sources in all test cases. The minimum silhouette width S from all the 1000 runs for $k = 3$ was 0.593. The maximum silhouette width S for $k = 4$ was -0.168. Again, in all the 1000 cases the solution for $k = 3$ will be selected based on the silhouette width S.

3.1.2. Example #2: 5 sources, 8 geochemical constituents, 12 wells and 12 time frames

As a second test, we consider an example randomly generated to represent five unknown synthetic sources (groundwater types) observed at 12 observation points over 12 time frames. Each source is represented by varying concentrations of 8 geochemical species. The number of wells (observation points), time frames, and geochemical species is consistent with the real problem presented in Section 3.2. The random concentrations are generated following the procedure outlined in the previous Section 3.1.2 The concentration data are perturbed by adding random noise from a normal distribution (mean equal to zero and standard deviation equal to 0.01) representing measurement errors. The concentration tensor C is provided to NTF*k* to estimate the number of sources and spatial/temporal dynamics of the contaminant mixing.

Based on NTF*k* results listed in Table 4, the number of sources is five. This is estimated by the behavior of the average silhouette width S and *AIC* criteria as a function of the number of sources k. The average silhouette width S is close to 1 for the cases when $k \leq 5$. S drops slightly for $k = 5$ but it is still close to 1. A substantial drop for S occurs for $k > 5$. This suggests that the solution for more than 5 sources is non-unique and depends strongly on the random initial guesses for the unknown components of matrix W and tensor G. *AIC* shows a substantial drop between cases of 4 and 5 sources; this also suggests that the solution with 5 sources should be selected.

The same conclusion can be also drawn here by the reconstruction quality Q. Clearly, the solution for 5 sources produces a much better fit to the data than the solution for 4 sources. The solution for 6 sources also produces a good match but based on the parsimony principal (also captured by *AIC*), it should be rejected because it is using far more model adjustable parameters. In this case, the 5 source NTF*k* solution has 616 adjustable parameters ($12 \times 12 \times 4 + 5 \times 8$) while the 6 source solution has 768 adjustable parameters ($12 \times 12 \times 5 + 6 \times 8$). In all cases, there are only 1152 observations ($12 \times 12 \times 8$).

The same synthetic problem was rerun 1000 times with different randomly-generated "true" concentrations C. All the runs are performed adding random noise from a normal distribution (mean equal to zero and standard deviation equal to 0.01). In all the 1000 cases, the algorithm correctly identified the true number of sources. The minimum silhouette width S from all the 1000 runs for $k = 5$ was 0.870. The maximum silhouette width S for $k = 6$ was 0.162. Based on this, in all the 1000 cases, the solution for $k = 5$ will be selected.

3.1.3. Example #3: 3 sources, 4 geochemical constituents, 15 wells and 101 time frames

NTF*k* is applied to analyze a synthetic groundwater contamination problem generated using a model simulating advective-dispersive transport. A map showing locations of monitoring wells providing data to characterize three point contaminant sources is presented in Fig. 2. The concentration of geochemical species released from the source are estimated using an analytical solution of three-dimensional advective-dispersive contaminant transport (Wexler and Wexler, 1992; Park and

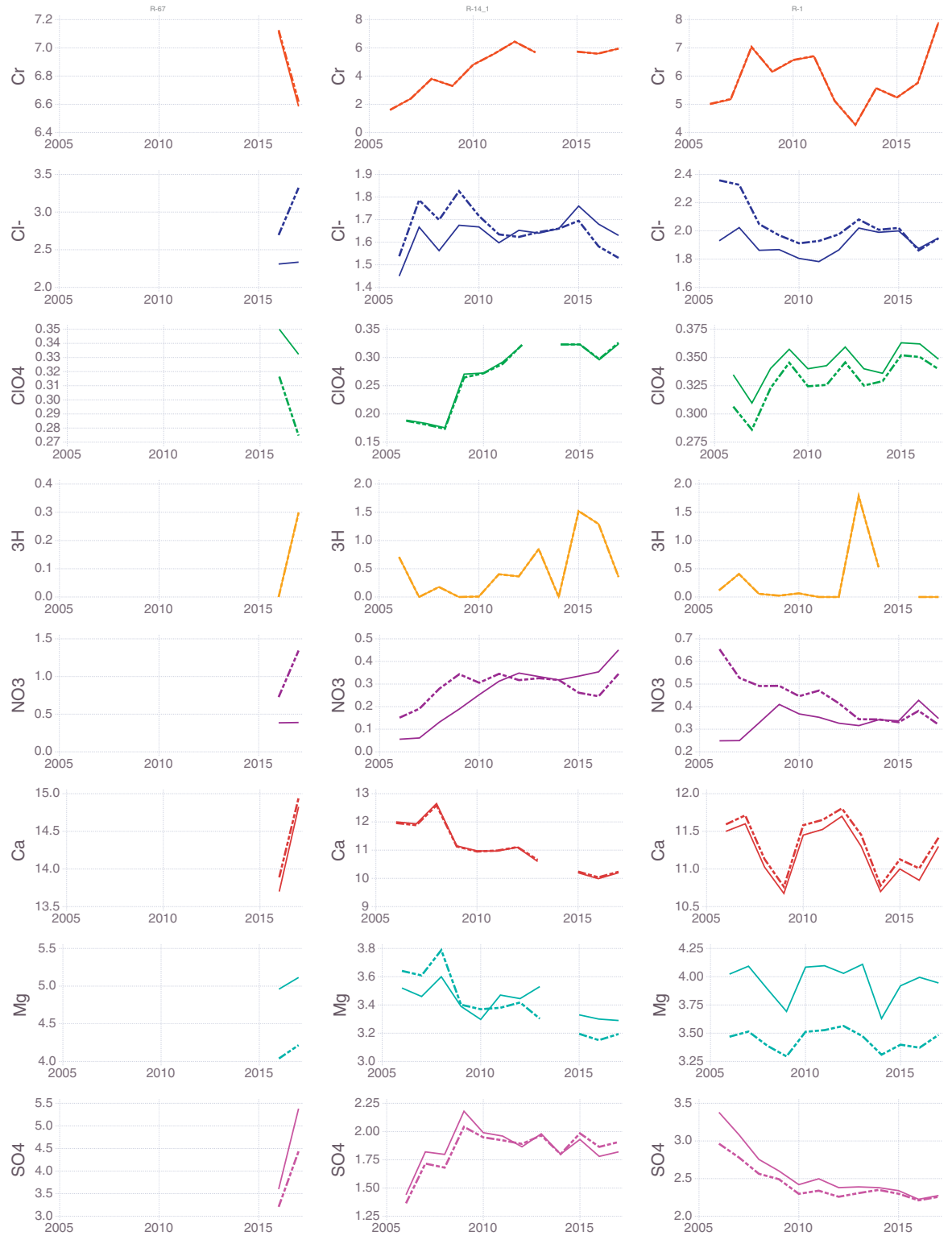
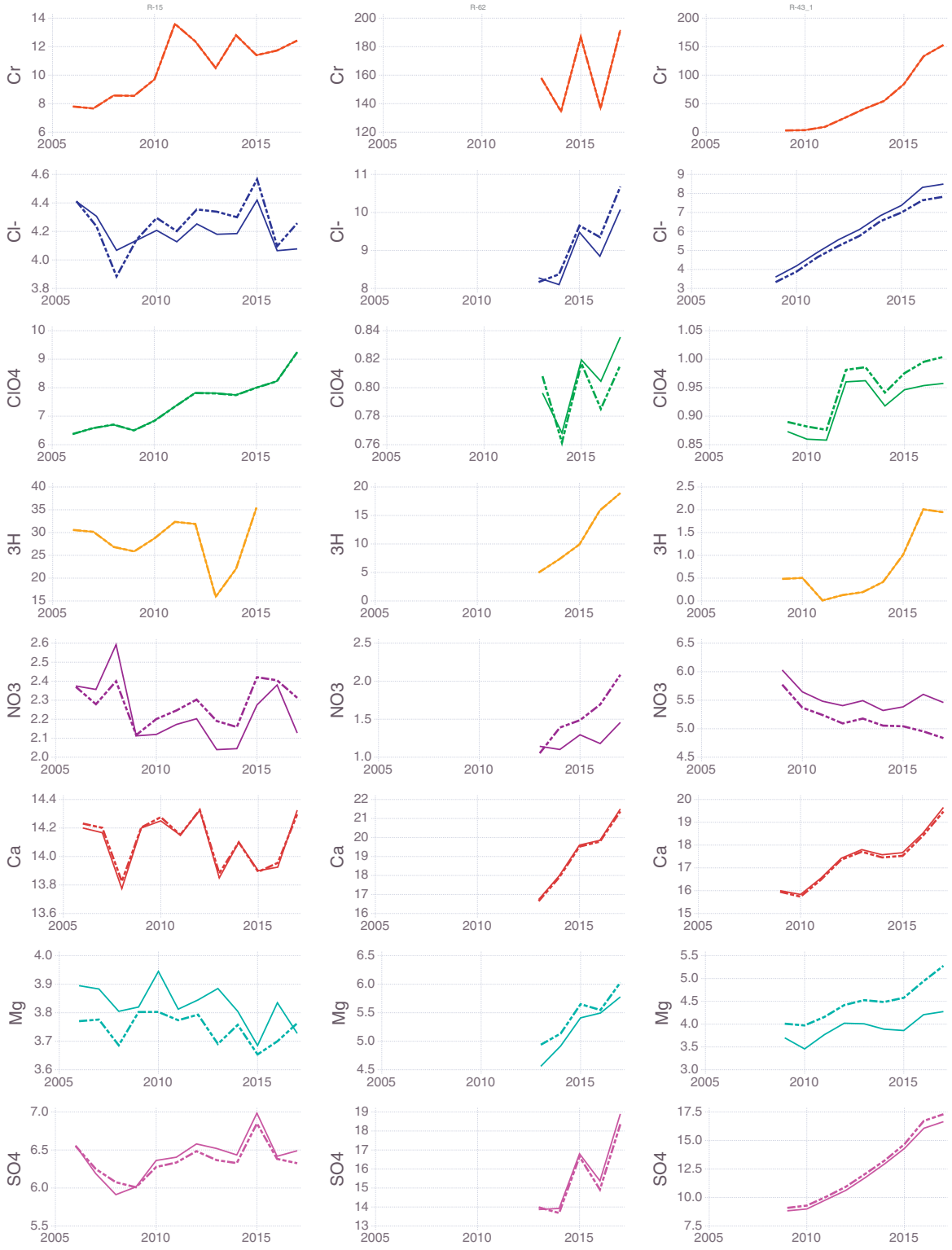


Fig. 6. Observed (dashed lines) and NTF k-predicted (solid lines) concentrations at the monitoring wells; note that for some of the wells/species the two lines overlap.

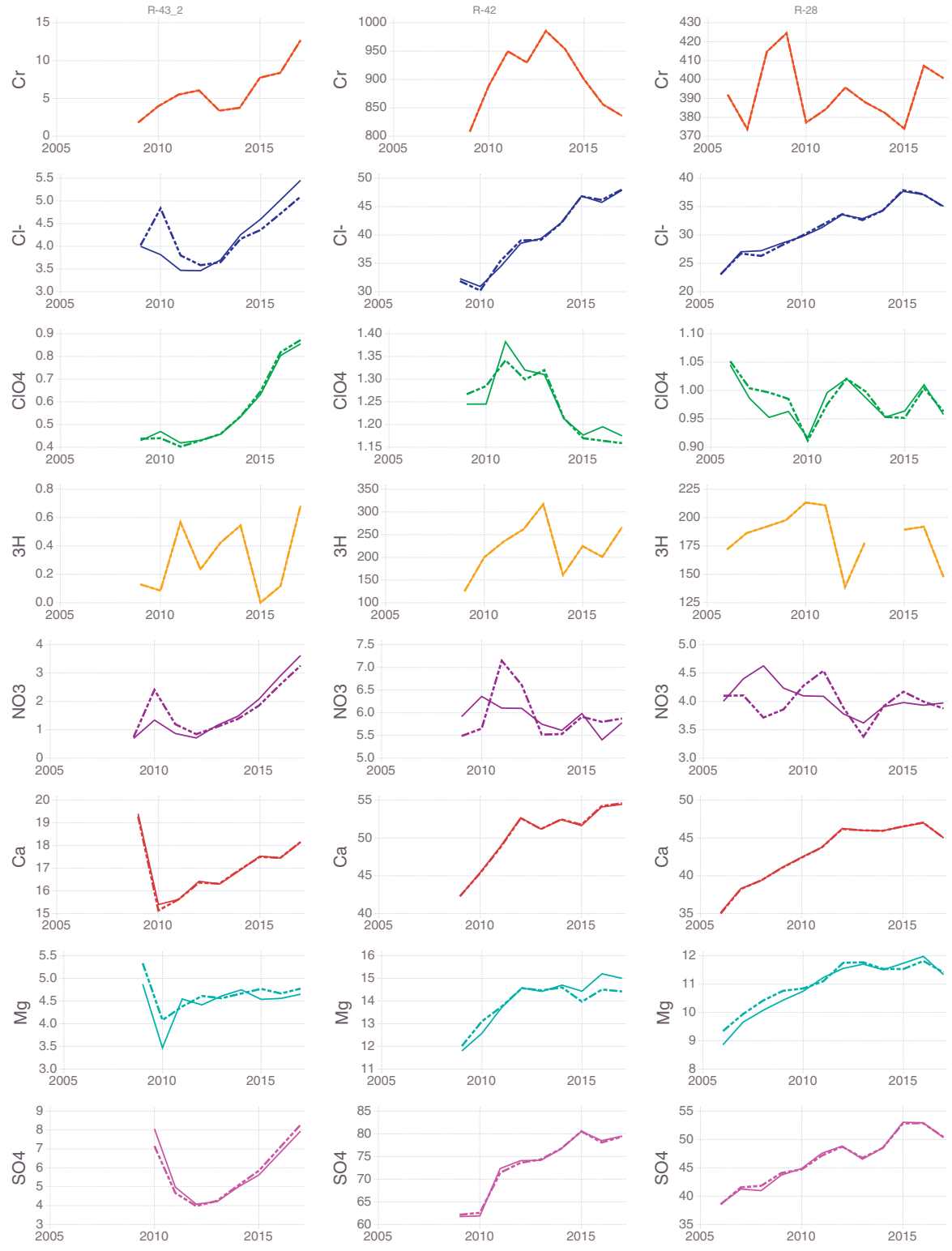
Zhan, 2001). The concentrations are computed using open-source codes Anasol.jl (O'Malley and Vesselinov, 2018) and Mads.jl (Vesselinov and O'Malley, 2016a; Vesselinov and O'Malley, 2016b) written in Julia (Bezanson et al., 2014). In these computations, it is assumed that each of the three sources is characterized by four geochemical species (A, B, C, and D). In addition, there is also an unknown source representing the background concentrations of these species. The "true" unknown concentrations (in *ppm*) of the geochemical species A, B, C, and D are shown in Table 5. The concentrations of the four geochemical species at the 15 monitoring wells are computed for 101 annual time frames from



(d) R-15 (e) R-62 (f) R-43#1

Fig. 6. (*continued*)

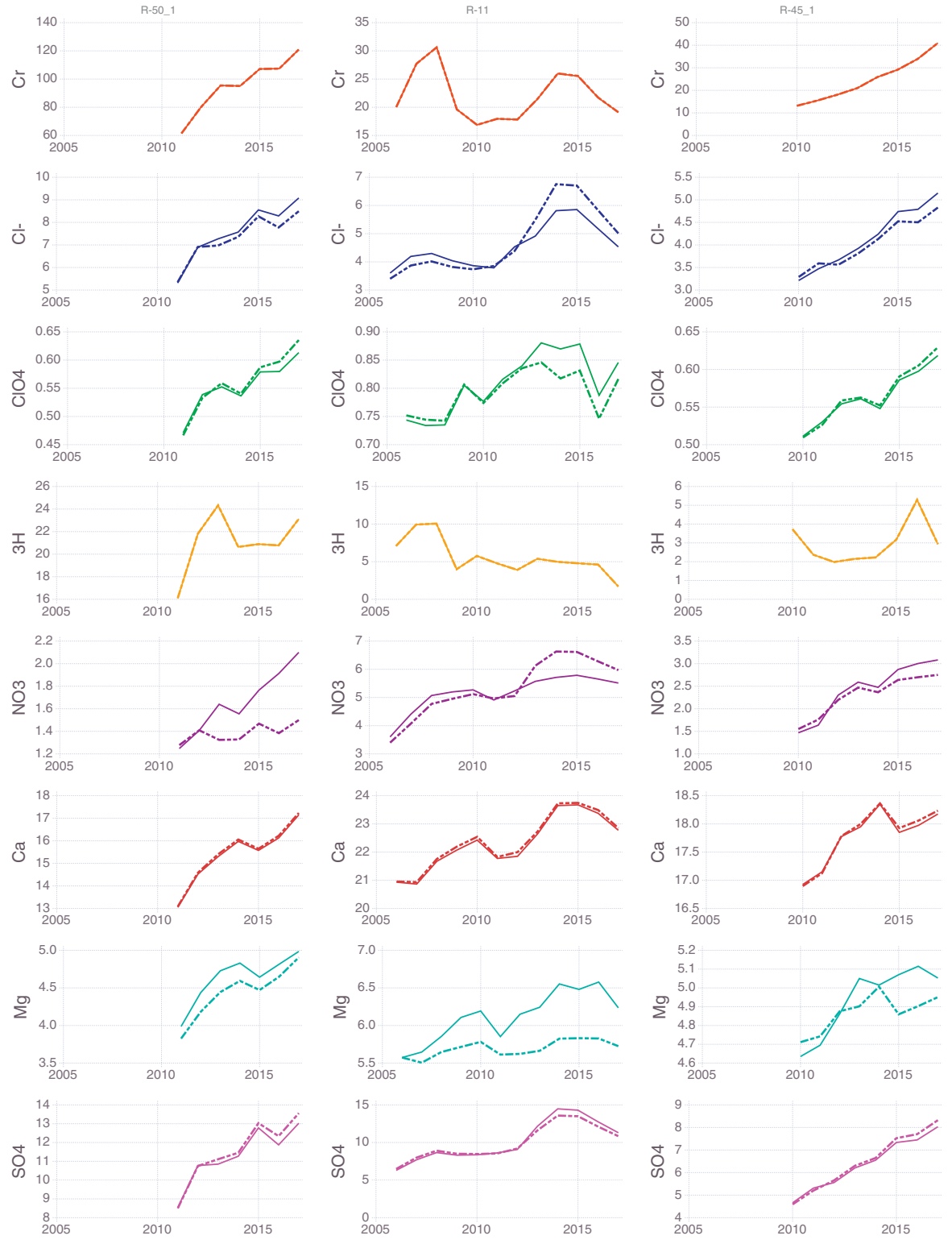
0 to 100 years. Concentration curves for six of the monitoring wells are presented in Fig. 3. Random uniform measurement errors of 10% have been added to the concentration data. The flow and transport parameters applied to compute the contraction transient are: advective (linear) pore velocity = 10 m/yr, logitudinal dispersivity = 70 m, transverse horizontal dispersivity = 15 m, transverse vertical dispersivity = 0.3 m, porosity = 0.1, contaminant flux = 50 kg/yr (constant at each point source). The first, second and third sources are activated at times equal to 0, 20 and 40 years. The three-dimensional data tensor with size (15 × 4 × 101) representing concentrations in 15 wells



(g) R-43#2 (h) R-42 (i) R-28

Fig. 6. (*continued*)

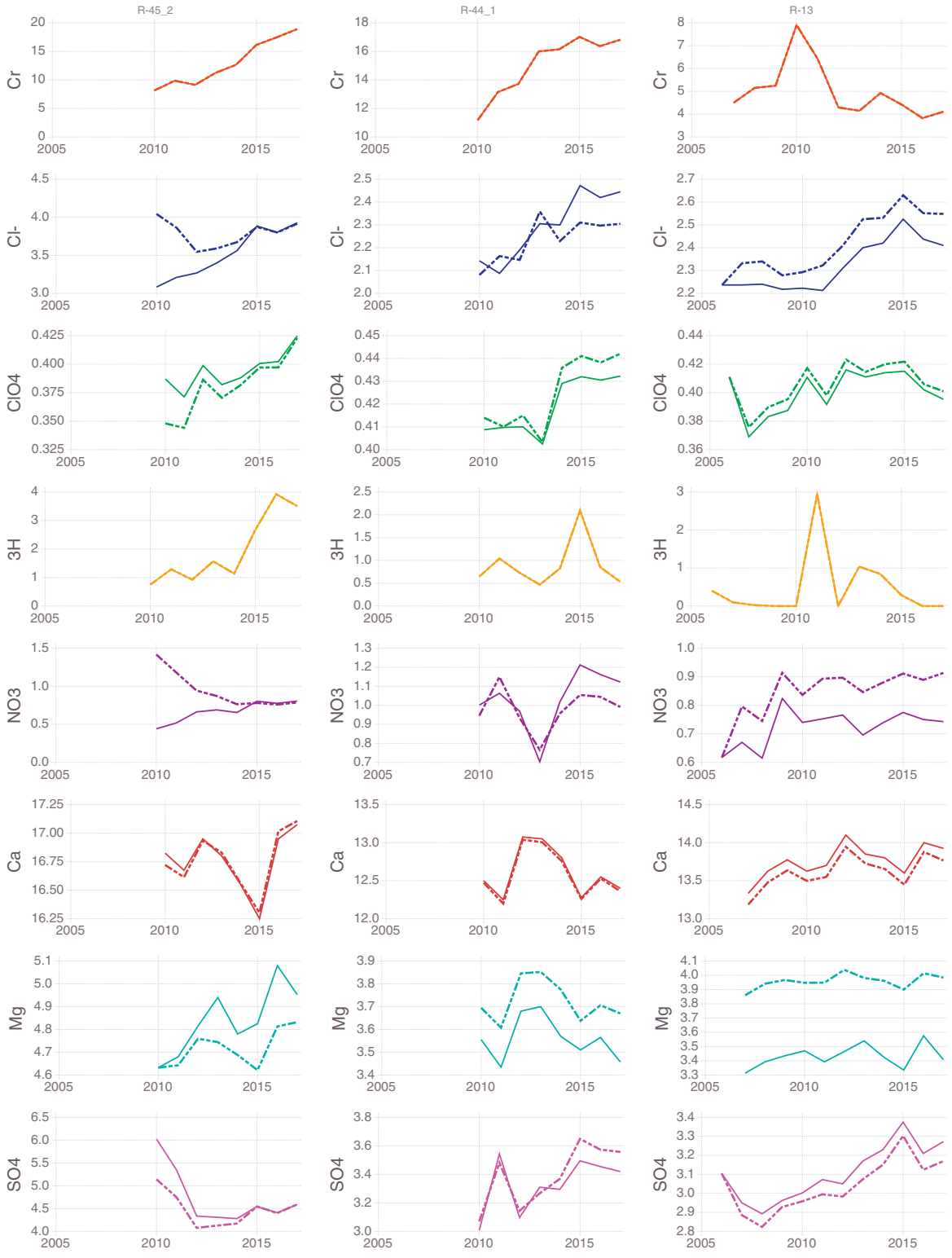
of 4 geochemical constituents for 101 time frames is analyzed using NTF*k*. Again, this is the only information provided to the algorithm. NTF*k* automatically identifies the number of sources (4) and the concentrations of geochemical species A, B, C, and D at the 4 sources. The results for parameters applied to estimate the number of sources are listed in Table 6. Clearly, the *AIC* drops substantially once the solution reaches 4 sources due to substantial decrease of the reconstruction quality *O*. *AIC* and *O* do not substantially improve for 5 sources. The silhouette width *S* also declines below 1 for 5 sources which also indicated that the solution with 4 sources is the correct one. The NTF*k*



(j) R-50#1 (k) R-11 (l) R-45#1

Fig. 6. (*continued*)

estimates of the source concentrations are shown in Table 5. A comparison between the "true" and NTF*k* estimated concentrations for six of the wells are presented in Fig. 3. Similarly, the "true" and estimated mixing coefficients are presented in Fig. 4. The NTF*k* estimates provide a very good representation of the mixing coefficients at each well over time. This demonstrates the capability of NTF*k* to predict the spatial and temporal dynamics of contaminant mixing. The results for the other monitoring wells (not shown in the figures) are consistent with the results presented in Figs. 3 and 4.

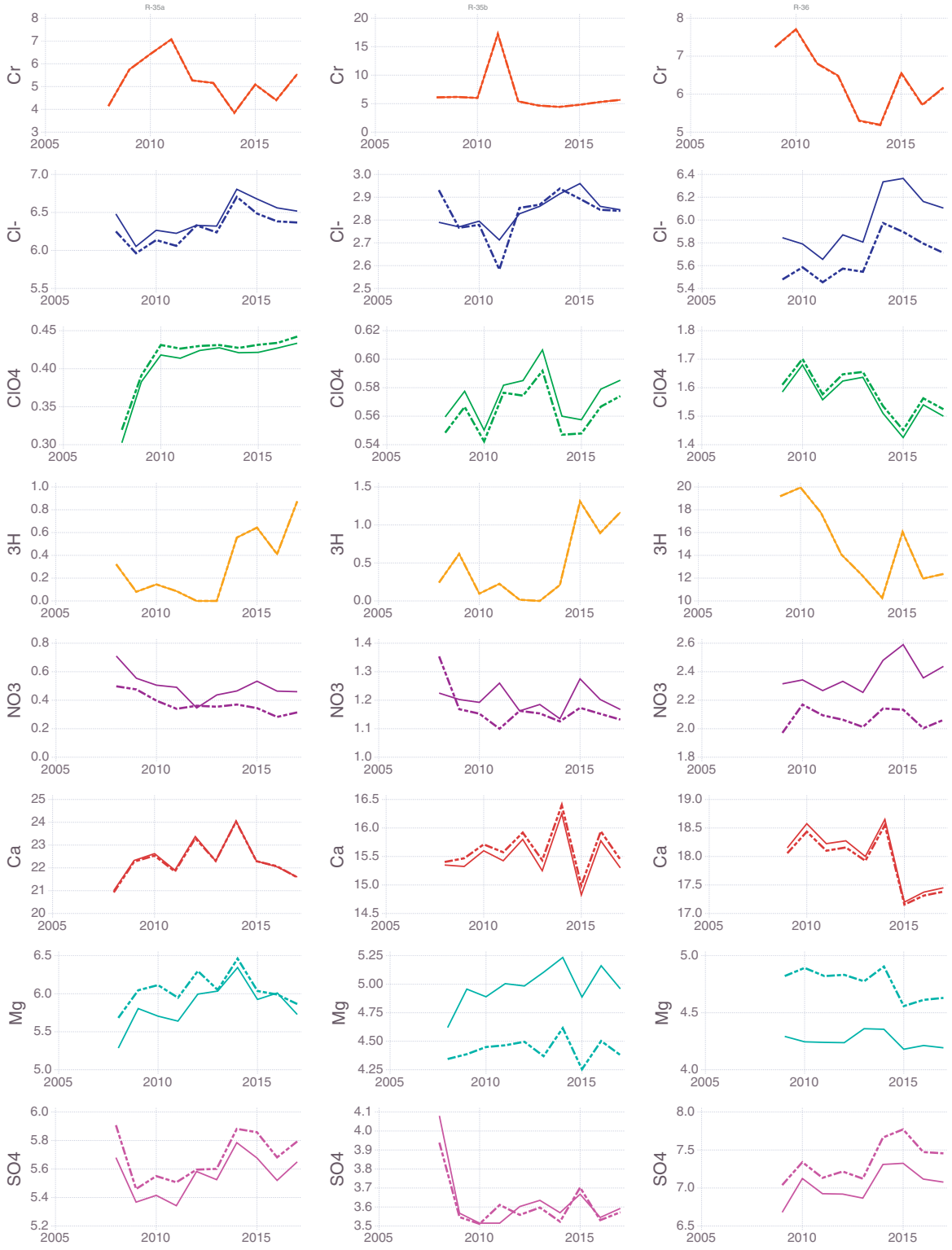


(m) R-45#2 (n) R-44#2 (o) R-13

Fig. 6. (*continued*)

3.2. NTFk analysis of site data

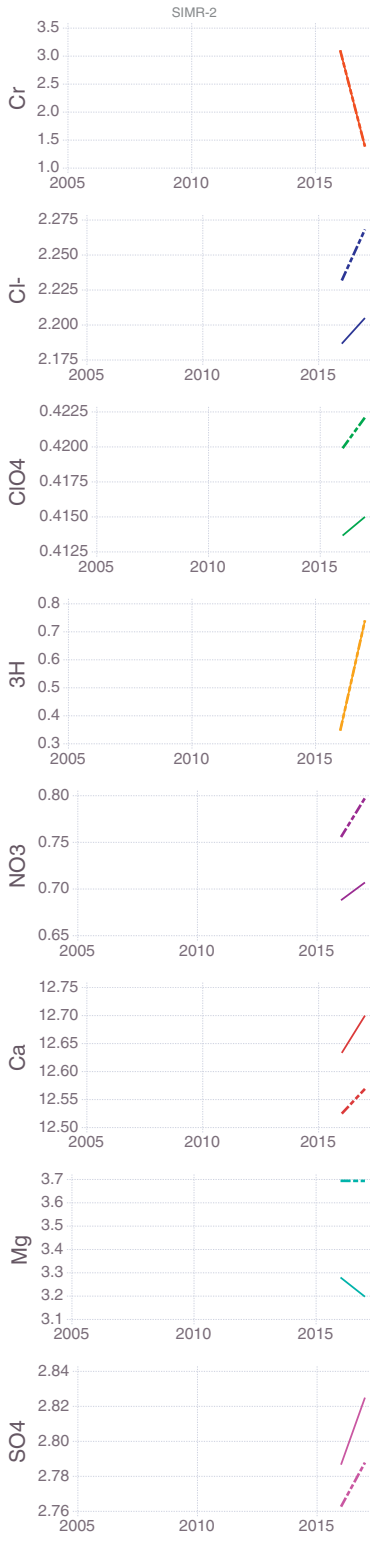
NTF*k* is applied to analyze the groundwater geochemistry data observed in the regional aquifer beneath the Los Alamos National Laboratory (LANL). The aquifer is contaminated with chromium (Cr^{6+}) and there are several contaminant sources that might have contributed to the contaminant plume beneath the LANL site near Sandia and Mortandad Canyons (Fig. 5). The investigation of the contaminant plume is ongoing (Vesselinov et al., 2013; Vesselinov et al., n.d.-b; LANL, 2009; LANL, 2012; LANL, 2018a). The site conceptual model



(p) R-35a (q) R-35b (r) R-36

Fig. 6. (*continued*)

describing the physical and biogeochemical processes controlling the movement of groundwater and contaminants in the environment is presented in detail in (Vesselinov et al., n.d.-a; LANL, 2009; LANL, 2012; LANL, 2018a; LANL, 2018b). It is important to note that due to site complexities, it is unknown how many different contaminant sources (groundwater types) are mixed in the regional aquifer. The geochemical signatures associated with these sources are also unknown. In addition, to chromium, some of the contaminant sources are expected to have elevated tritium (^3H), nitrate (NO_3^-), chloride (Cl^-), and perchlorate (ClO_4). Different contaminant sources are expected to



(s) SIMR-2

Fig. 6. (*continued*)

have different geochemical signatures representing a mixture of different contaminants. The contaminants have been released along Los Alamos, Sandia and Mortandad Canyons (Fig. 5). However, due to complexities of the three-dimensional flow in vadose zone (including perching horizons), the contaminants have been mixed before they entered the regional aquifer in the general area between Sandia and Mortandad Canyons (Fig. 5). Furthermore, contaminant releases along the same canyon are expected to have different signatures over time due to transients in the infiltration and contaminant-mass fluxes (LANL, 2018a).

Table 7
NTF*k* results for the LANL site problem; the reconstruction quality *O*, silhouette width *S*, and *AIC* are estimated for number of sources $k = 2, \ldots, 8$.

k	*O*	*S*	*AIC*
2	$9.185 \cdot 10^{5}$	1.000	989.225
3	$9.054 \cdot 10^{3}$	1.000	538.317
4	$2.026 \cdot 10^{2}$	0.997	175.943
5	$2.566 \cdot 10^{1}$	1.000	0.767
6	0.823	0.999	−322.662
7	0.009	0.759	−758.968
8	$1.054 \cdot 10^{-14}$	−0.383	−3681.497
9	$1.312 \cdot 10^{-14}$	−0.222	−3609.816

A subset of the geochemical data collected at the site is applied for the NTF*k* analysis and is presented in Fig. 6. The data comes from 19 monitoring well screens: R-67, R-14#1, R-1, R-15, R-62, R-43#1, R-43#2, R-42, R-28, R-50#1, R-11, R-45#1, R-45#2, R-44#1, R-13, R-35a, R-35b, R-36, and SIMR-2 (the number after # represent screen number within multi-screen wells ordered vertically from top to bottom; the screen names without # indicate single-screen wells). The well order approximately follows the direction of the groundwater flow which is from west to east. The data include representative measurements for eight geochemical species: chromium (*Cr*), chloride (Cl^{-}), perchlorate (ClO_4), tritium (^{3}H), nitrate (NO_3), calcium (*Ca*), magnesium (*Mg*), and sulfate (SO_4). Other geochemical species have been measured at these wells; however, prior geochemical analyses of the site data (LANL, 2018a) have identified the above subset of geochemical species to be the most representative of the site conditions. The analyzed data represents annual averages for each year between 2005 and 2016 (Fig. 6). Annual averages are processed not due to methodological or computational limitations but due to irregularities in the sampling events. The geochemical data are collected on a quarterly basis (but sampling events are on different dates for different wells). In addition, there are numerous irregular sampling events. Due to general irregularity of the sampling events, we computed yearly averages. The final dataset includes 12 geochemical time snapshots in total. Note that there are gaps in the processed dataset (Fig. 6). The dataset was analyzed using NTF*k* to define the potential groundwater sources (groundwater types) that are represented as geochemical mixtures in the monitoring data over time. The NTF*k* analysis accounts for the mixing of different groundwater types, where some of the types might be associated with background groundwater types and others might be caused by the contamination sources.

The NTF*k* results are presented in Table 7. NTF*k* identifies 7 original groundwater sources with different geochemical composition are mixed in the aquifer. This estimate is based on the silhouette width *S* values. Note that $S \approx 1$ for $k \leq 7$. The *AIC* values suggest the existence of 7 sources as well.

The NTF*k*-estimated concentrations of each of the eight geochemical species prior to mixing with regional aquifer water for the identified seven sources (groundwater types) are presented in Table 8. These are the concentrations of the groundwater types that are mixed to reproduce the observed concentrations at the wells over time.

Fig. 6 shows the observed versus estimated geochemical concentrations at monitoring wells over time. NTF*k* accurately reproduces the geochemical transients observed at all the site monitoring wells.

Note that all the identified sources (groundwater types) have distinct geochemistry (Table 8). Sources 1, 2, 4, 5 and 6 are clearly associated with contaminant sources because of the presence of constituent concentrations above background. Source 1 has elevated values

Table 8
NTF*k* estimated concentrations of the 7 groundwater types (contaminant sources) mixed at each observation well.

Sources	*Cr*	Cl^-	ClO_4	^3H	NO_3	*Ca*	*Mg*	SO_4
	(μg/L)	(mg/L)	(μg/L)	(pCi/L)	(mg/L)	(mg/L)	(mg/L)	(mg/L)
S1	2970.22	63.06	0.00	0.00	13.94	73.37	24.74	171.02
S2	0.79	0.35	13.87	0.00	0.49	5.27	1.71	0.61
S3	0.24	3.62	0.00	0.00	0.01	40.77	10.90	0.06
S4	0.48	0.14	0.00	0.00	10.49	21.09	5.00	10.18
S5	20.53	50.57	0.00	949.53	2.39	66.54	14.89	49.63
S6	1.46	64.24	0.00	0.00	2.81	50.92	10.43	68.08
S7	0.10	0.03	0.00	0.00	0.01	0.43	0.78	0.88

for *Cr*, Cl^-, NO_3, *Ca*, *Mg*, and SO_4. This is the main source of chromium that constitutes the majority of the plume footprint. This source is a combination of releases on the ground surface along Sandia Canyon. The estimated chromium concentration of about 3000 *ppb* is corroborated by the source concentrations estimated using hydrogeologic data and techniques (LANL, 2018a). Source 2 has elevated ClO_4, which is a known contaminant released on the ground surface in Mortandad Canyon (LANL, 2009; LANL, 2012). Source 4 has increased NO_3 and potentially comes from Los Alamos, Sandia or Mortandad Canyons. Source 5 has elevated ^3H; *Cr*, Cl^-, *Ca*, *Mg*, and SO_4 are also high. This is potentially a mixed source where contaminants originating along Los Alamos, Sandia and Mortandad Canyons are mixing in perched groundwater horizons in the vadose zone before their arrival at the regional aquifer (LANL, 2009; LANL, 2012). Source 6 has elevated Cl^-, *Ca*, *Mg*, and SO_4 and *Cr*. This might be groundwater originating from the same source as earlier chromium-contaminated water released in Sandia Canyon, since it has very low concentrations of ^3H. The above interpretations of the NTF*k* estimated source (groundwater types) are consistent with the site conceptual model but provide new insights about the contaminant fate and transport at the site.

Sources 3 and 7 (Table 8) represent the non-contaminated groundwater signature in the plume area ("background" groundwater types). The variations in the mixing of these two background groundwater types represent variability in the background compositions, potentially as a result of some mixing with contaminated groundwater sources. The temporal dynamics of sources 3 and 7 may represent geochemical reactions occurring in the aquifer due to the mixing of groundwater with different geochemistry or heterogeneity in the aquifer materials causing changes in the groundwater geochemistry.

NTF*k* also provides estimates of the mixing dynamics of the sources over time. The estimated temporal evolution of how the seven groundwater types are represented and mixed at each monitoring well is presented in Fig. 7.

All seven groundwater types are observed at appreciable levels in only two of the monitoring wells: R-42 and R-28 (Fig. 7h and i). These wells have the highest chromium concentrations (Fig. 6h and i) and are located in the center of the chromium plume (Fig. 5). At both wells, the mixing ratios for background source 7 are decreasing over time, while the mixing ratios for contaminant source 6 are increasing over time. Source 4 at R-28 seems to be decreasing as well. The R-42 transients may suggest a peak mixing ratio for sources 1 and 5 at R-42 in 2013.

Results for wells R-43, R-62 and R-67 upgradient from R-42 and R-28 (Fig. 5) potentially represent recent arrival of contaminants in this area. The mixing transients observed in the upper and lower screens in R-43 (R-43#1 and R-43#2, respectively) are very different. R-43#1 is dominated by source 4, although the contribution seems to be diminishing in time while the contribution of contaminant source 1 appears to be sharply increasing (Fig. 7f). R-43#2 is observing increasing contributions of sources 4 and 6 (Fig. 7g) which might be caused by slow vertical groundwater flow and transport from shallow portions into deeper portions of the aquifer. At R-62, the background source 7 is decreasing, but the sources 1, 2, 4 and 5 are increasing (Fig. 7e). R-67 is dominated by background sources 3 and 7, but their contribution is decreasing (Fig. 7a) while the impact of source 4 is increasing which suggests a contaminant arrival (Fig. 6a). The further upgradient wells (R-14 and R-1; Fig. 5) are dominated by background sources 3 and 7 (Fig. 7b and c) and there are no contaminant sources detected at these wells.

The near-field downgradient wells from R-42 and R-28 (R-50#1. R-11. R-45#1, R-45#2, R-44#1, R-13, and SIMR-2) also show transients that represent arrival of contaminated groundwater. At R-50#1, the contributions of the background groundwater types are changing over time (Fig. 7j) and there is an increase in contaminated source 1. R-11 mixing ratios show increasing contributions of contaminated sources 4 and 6 (Fig. 7k); source 2 is going up while the sources 1 and 5 might be slightly going down in time. At the upper screen of R-45 (R-45#1), the contaminated sources 1, 2, 4 and 5 are increasing. In contrast, at the lower screen (R-45#2) source 4 is decreasing over time. The difference in the behavior of source 4 in R-45#1 (increasing) and R-45#2 (decreasing) potentially suggests complex groundwater flow/transport conditions and/or differences in the geochemical processes associated with different rock types within the regional aquifer. R-44#1 is dominated by sources 3, 4, and 7 (Fig. 7n); the contribution of sources 1 and 2 is slightly increasing. R-13 shows (Fig. 7o) a slight increase of source 4. SIMR-2 is affected by low proportions of contaminant sources 2, 4, and 6 (Fig. 7s).

R-35b and R-35a are shallower and deeper wells screened at different depths next to an existing water-supply well. R-35b is completed close to the regional water table and contaminant sources 2, 4 and 6 appear to be are present (Fig. 7q). The vertical location of the R-35a screen matches the top of supply-wells louvers. R-36b is dominated by background sources 3 and 7; however, contaminant sources 2 and 6 are potentially present at this well in low proportions (Fig. 7p).

R-36 is anomalous and very different from all the other wells. All groundwater types are present here except source 1 (Fig. 7r). This is extremely surprising, considering the well location and the mixing ratios observed at the nearby wells. Groundwater screened at R-36 may represent an area of infiltration with geochemical composition very different from all the other wells. However, a more probable explanation is that R-36 might be tapping groundwater in a perched saturated horizon in the vadose zone which is above the regional-aquifer water table and detached from the regional aquifer sometime in the past. In this case, the water observed at R-36 may represent old aquifer groundwater which was flowing in the aquifer in the past before the perched zone was detached from the regional aquifer. This interpretation is also corroborated by the water-level data observed at R-36 (LANL, 2018a). As a result, R-36 is probably not representative of the aquifer conditions. It is quite possible that very different contaminant conditions might exist within the regional aquifer at the location of R-36.

R-15 is also very different from the other wells, it is predominantly influenced by source 2 (ClO_4), which appears to show an increasing contribution over time (Fig. 7d). Source 2 has been also detected at R-50#1 and SIMR-2.

The analyses also suggest anomalous behavior in 2010 at R-43#2 (Fig. 7g) and in 2014 at R-28 (Fig. 7i). This might be caused by

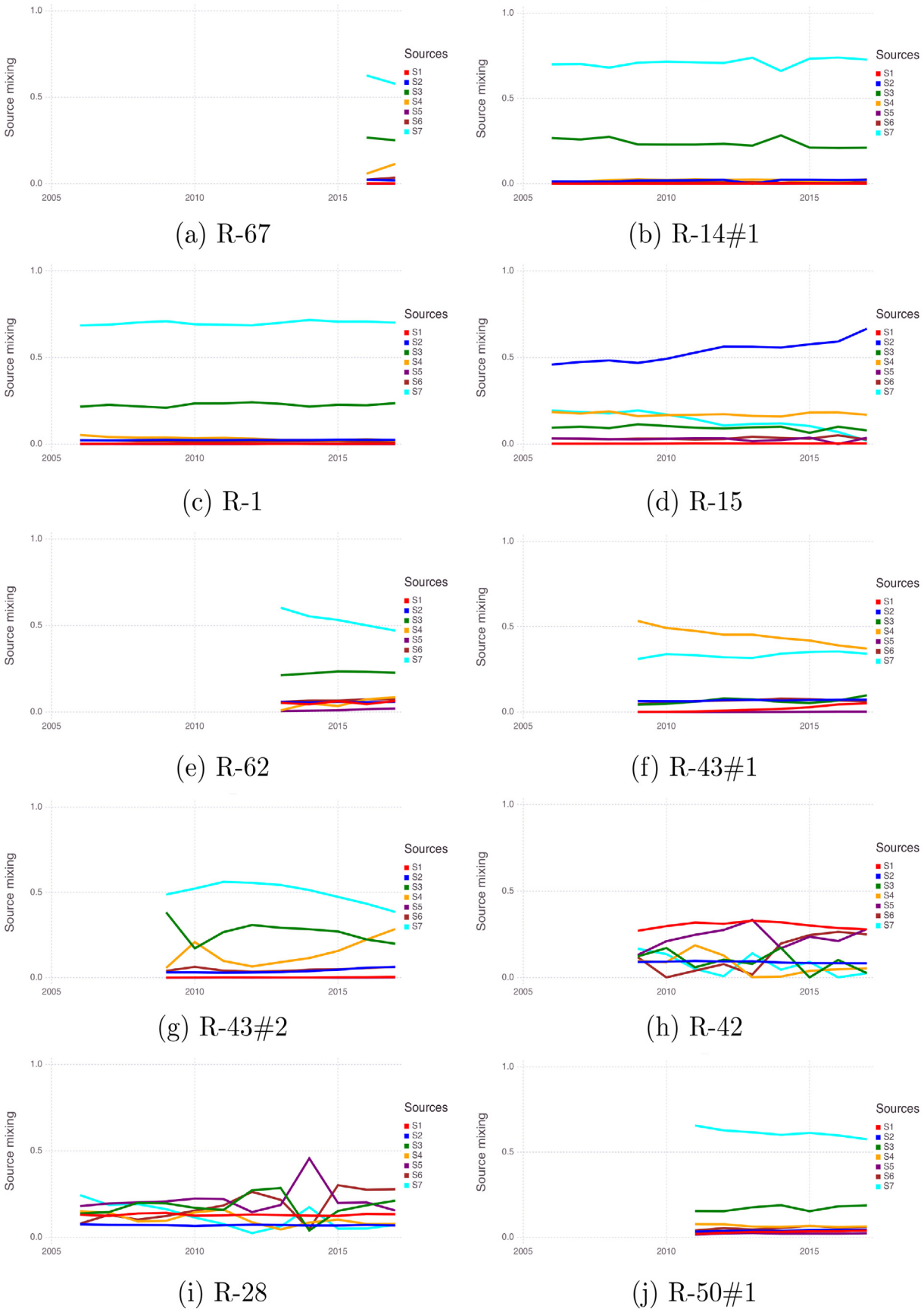


Fig. 7. NTF*k* estimated transient mixing ratios of the seven groundwater types at the site monitoring wells. Note that the mixing ratios for each period of record for each well add to 1.

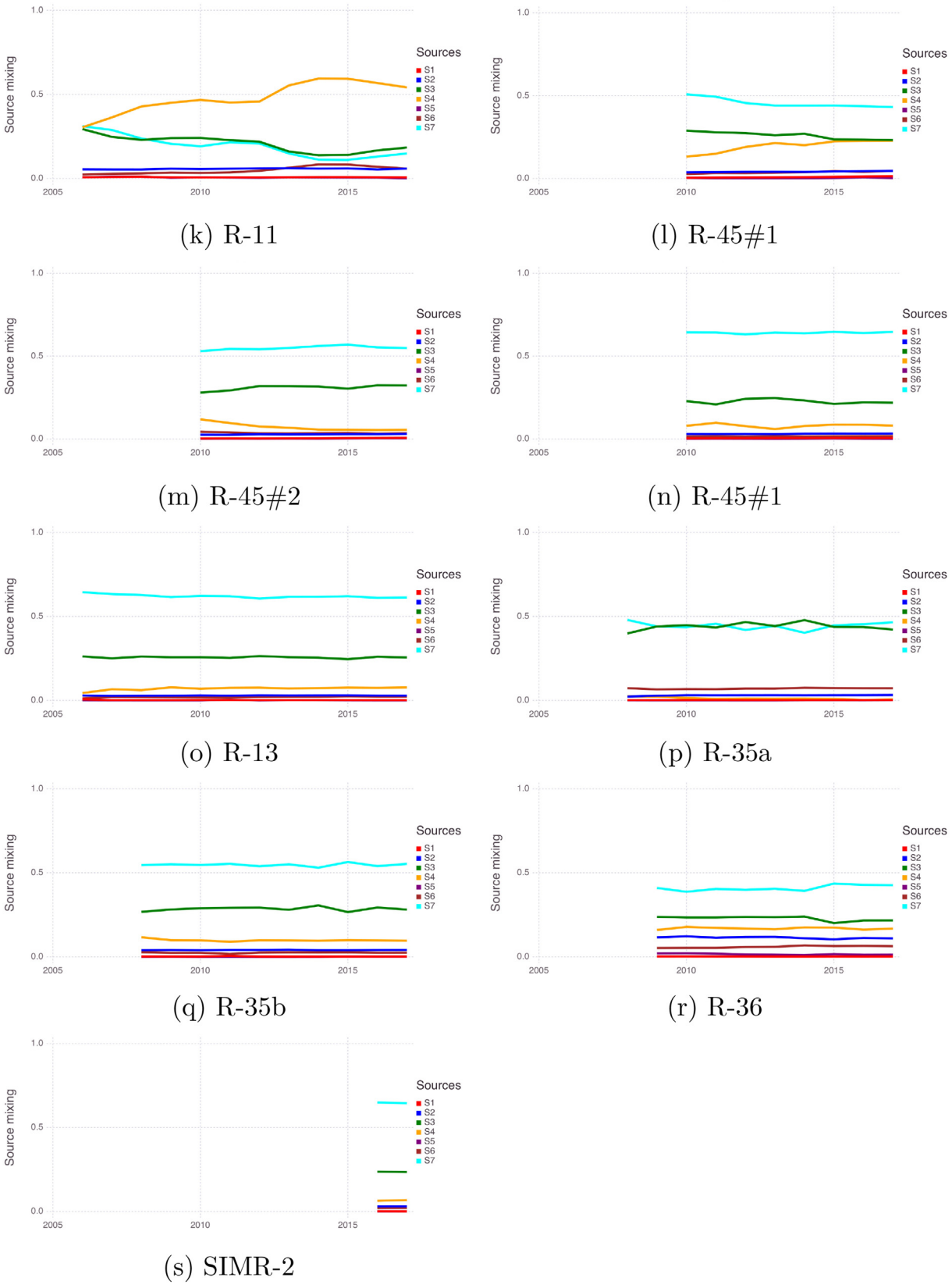


Fig. 7. (*continued*)

systematic errors associated with the field sampling of these two wells; for example, issues with bore-hole water sampling systems; (systematic errors caused by laboratory sample analyses can be ruled out because most of the well samples are processed simultaneously in batches (LANL, 2018a)). Alternative explanation is that these anomalies might represent the effects of field activities conducted at these wells (LANL, 2018a).

It is important to note that even though limited data are available

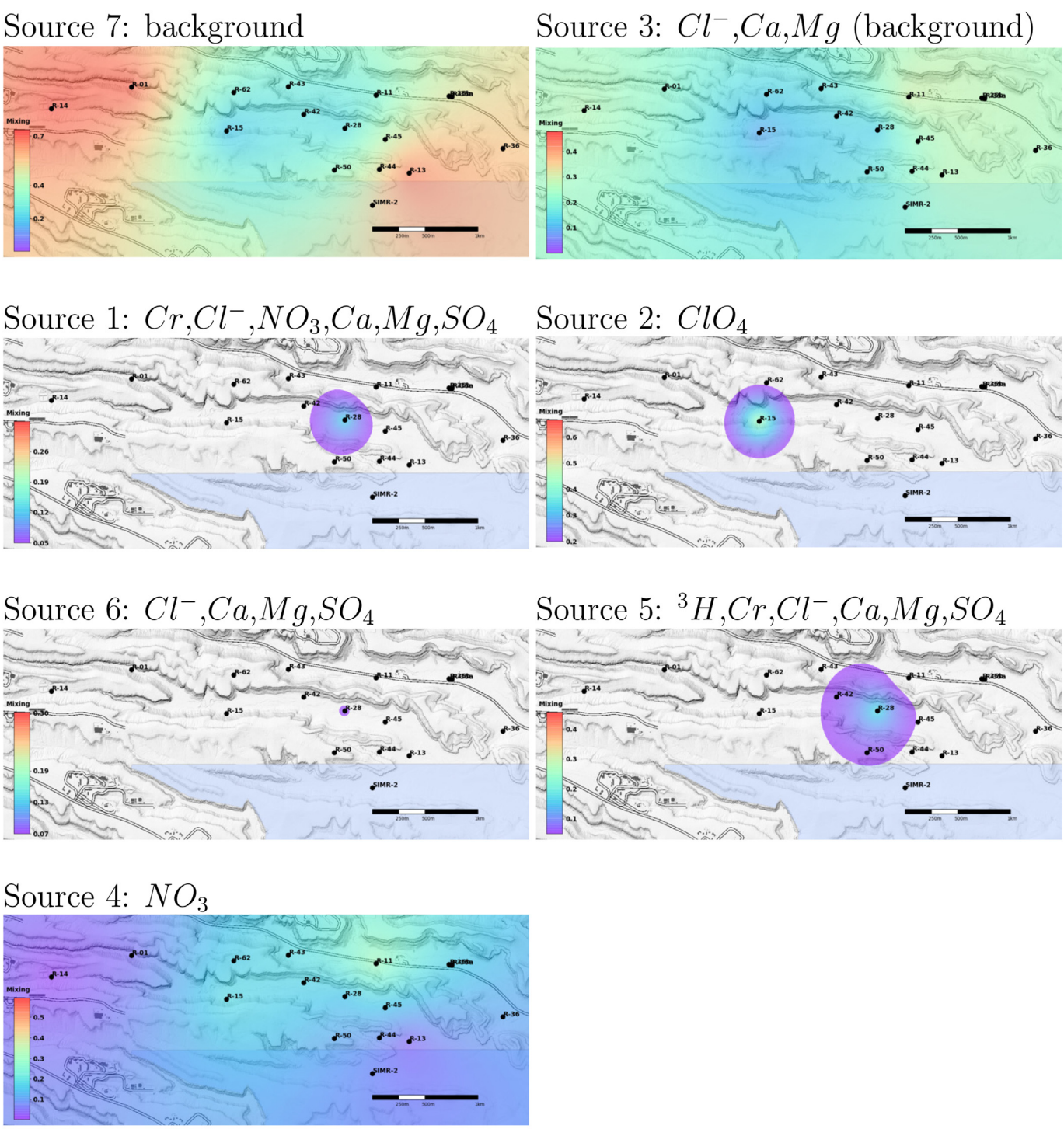
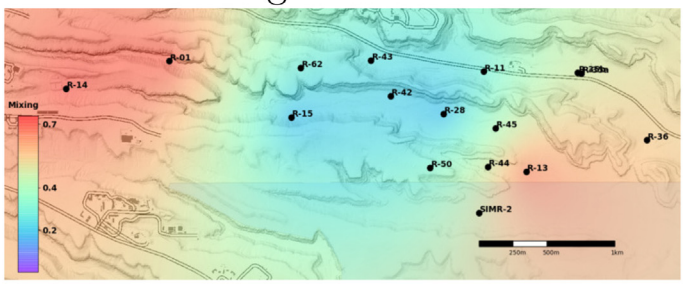


Fig. 8. NTF*k* estimated transients in the spatial footprint of the seven sources (ground-water types) at the LANL site.

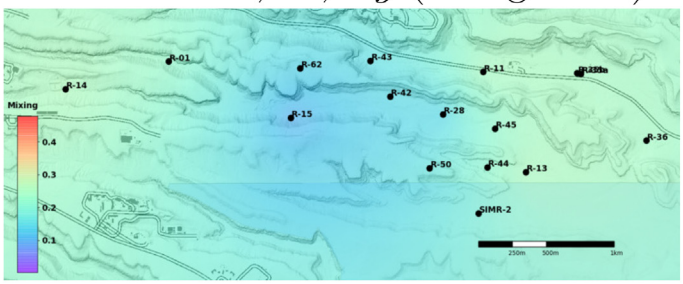
for some of the wells (e.g R-67 (Fig. 6a) and SIMR-2 (Fig. 6s), the NTF*k* analyses are capable to extract meaningful information about the geochemical mixing at these wells.

The mixing information presented in Fig. 7 is also shown as spatial maps in Fig. 8. The maps depict the transient mixing ratios of the seven groundwater types (sources) identified as present at the site monitoring wells. The maps show the mixing ratios of each source (groundwater type). The mixing ratios are estimated at the wells and interpolated in space between the wells using Kriging. The interpolation is performed for each temporal time frame separately. An exponential variogram is applied with an integration scaling coefficient equal to 1000 m. The maps represent 12 temporal snapshots of mixing different sources

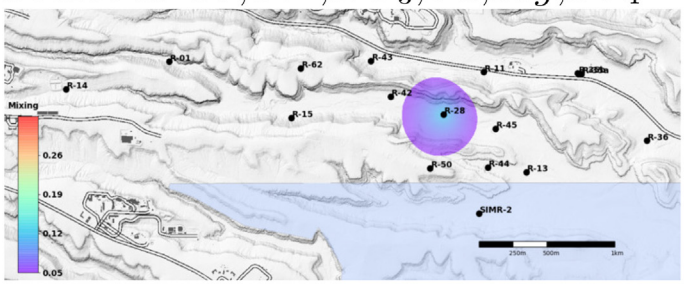
Source 7: background



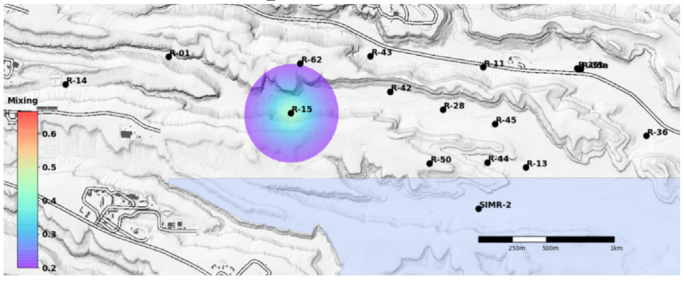
Source 3: Cl^{-},Ca,Mg (background)



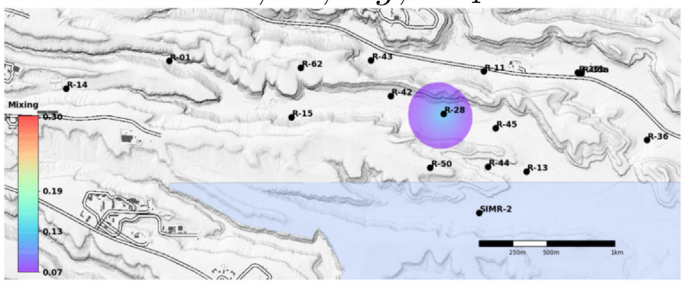
Source 1: Cr,Cl^{-},NO_3,Ca,Mg,SO_4



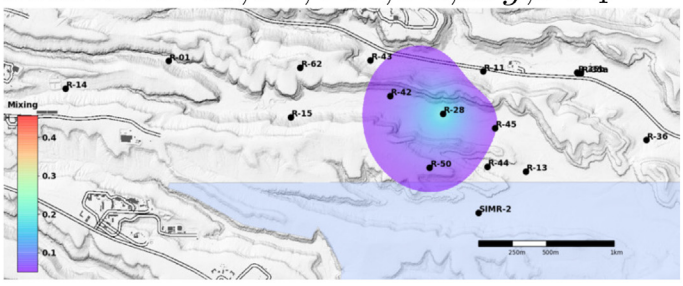
Source 2: ClO_4



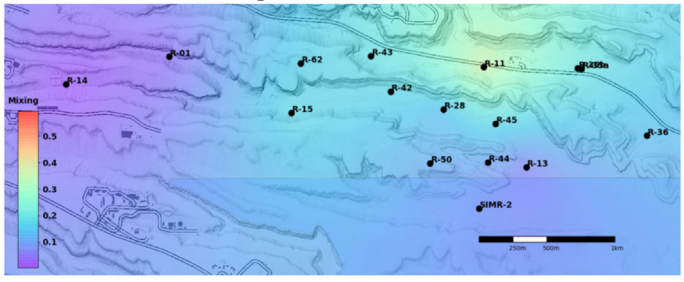
Source 6: Cl^{-},Ca,Mg,SO_4



Source 5: ^{3}H,Cr,Cl^{-},Ca,Mg,SO_4



Source 4: NO_3



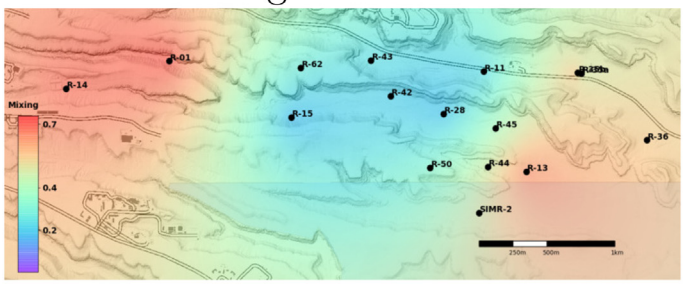
(b) January - December 2006

Fig. 8. (*continued*)

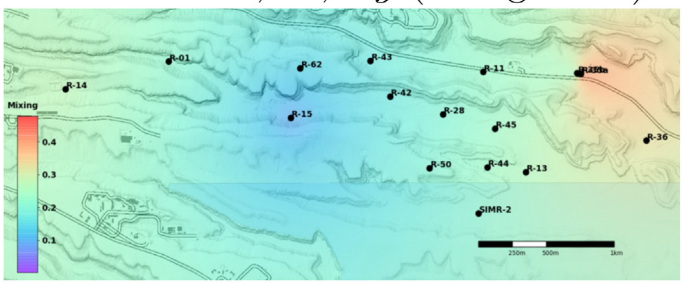
(groundwater types) from 2005 to 2016 based on the averaged geochemical data (Fig. 6).

Based on the maps in Fig. 8, contaminant sources 1, 5 and 6 are centered in the area of R-28 and R-42. The changes in the shape of the estimated spatial extent of these sources (groundwater types) are predominantly driven by the addition of new monitoring wells over the years (see Fig. 6). The major difference between sources 1, 5 and 6 is that source 5 is not dominant in R-62 and R-43. However, sources 1 and 6 are present at R-62 and R-43. Source 2 is centered in the area of R-15. Source 4 is in the area of R-43 and R-11; it has been also observed in R-62 and R-15. However, its impact seems to be diminishing at the R-62/R-15 area and increasing at R-11 in recent years. The transients in the mixing ratios between 2008 and 2013 (snapshots for source 4 in Fig. 8) may suggest impacts of lateral plume migration or shifts in the

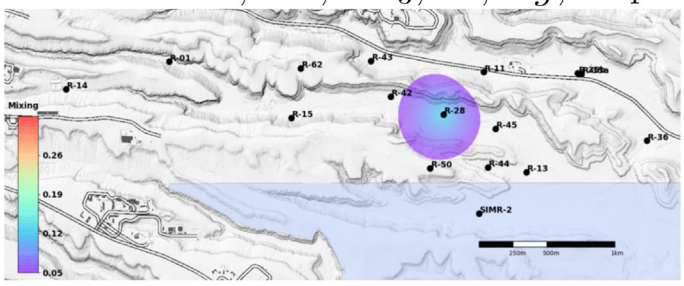
Source 7: background



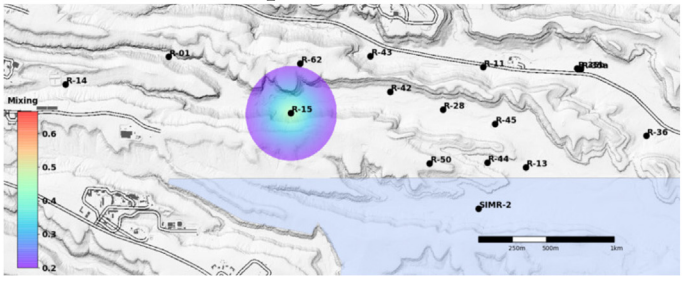
Source 3: Cl^{-},Ca,Mg (background)



Source 1: Cr,Cl^{-},NO_3,Ca,Mg,SO_4

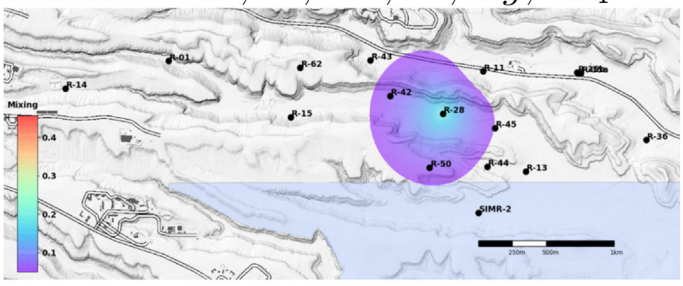


Source 2: ClO_4

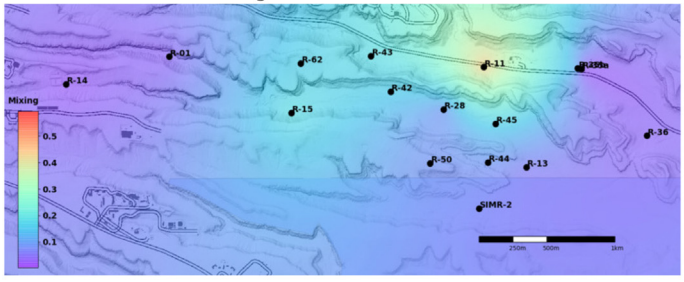


Source 6: Cl^{-},Ca,Mg,SO_4

Source 5: ^{3}H,Cr,Cl^{-},Ca,Mg,SO_4



Source 4: NO_3



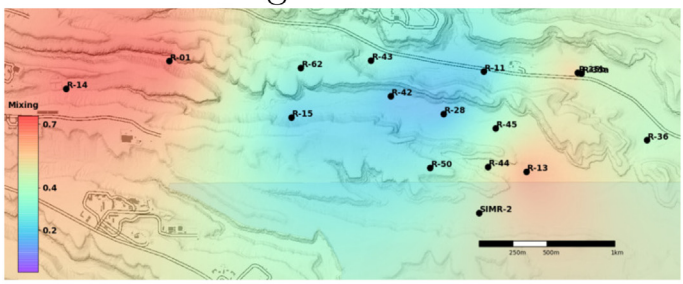
(c) January - December 2007

Fig. 8. (*continued*)

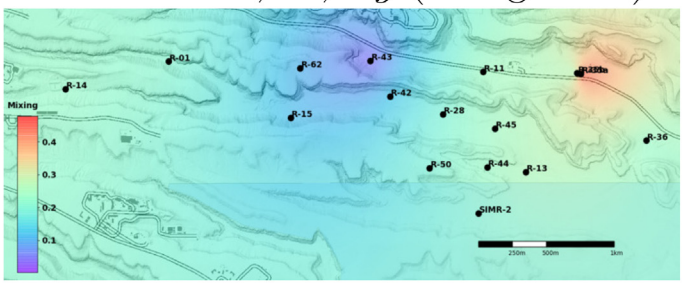
infiltration pathways within the vadose zone. The background groundwater types captured as source 3 represents the temporal and spatial dynamics of the *Cl*, *Ca*, and *Mg* geochemical species. Background source 7 represents the SO_4 dynamics. Diminished background mixing ratios are shown in the center of the chromium plume (area of the wells detecting sources 1, 5 and 6). The temporal dynamics of the mixing ratios in the area of chromium plume represent shifts in the mixing of background and contaminated groundwater. These dynamics can also represent geochemical processes occurring between water infiltrated from the vadose zone (e.g., sources 1, 5 and 6) interacting with the background regional groundwater.

The maps presented in Fig. 8 are unrealistic because the spatial distribution of the groundwater types (contaminant plumes) are expected to have much more complex shapes due to aquifer heterogeneity

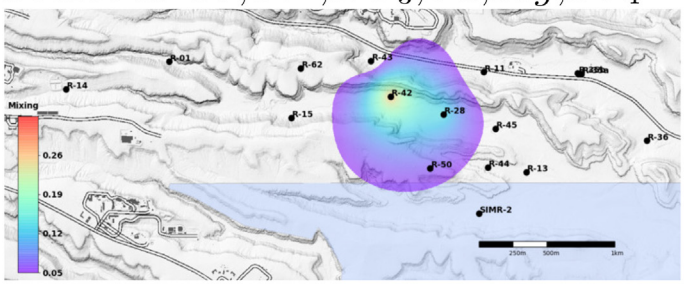
Source 7: background



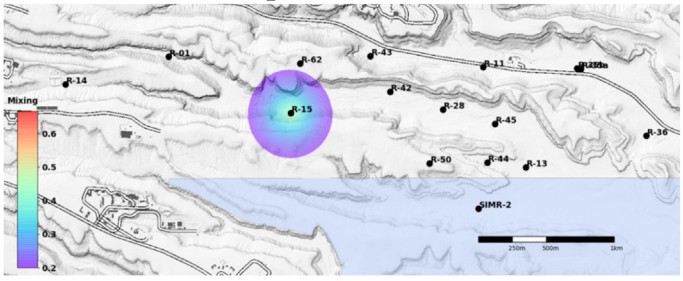
Source 3: Cl^{-},Ca,Mg (background)



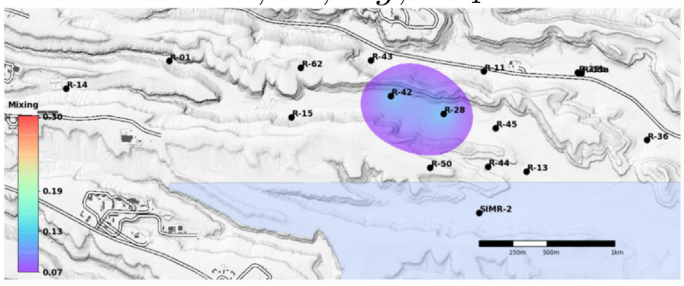
Source 1: Cr,Cl^{-},NO_3,Ca,Mg,SO_4



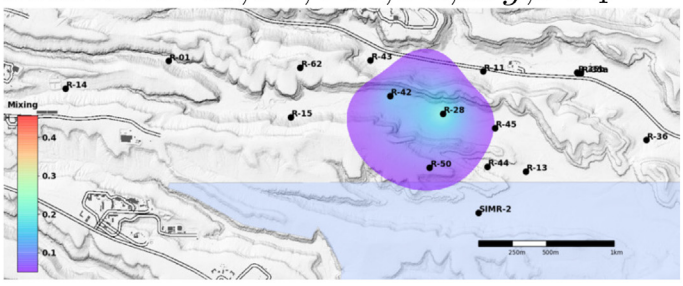
Source 2: ClO_4



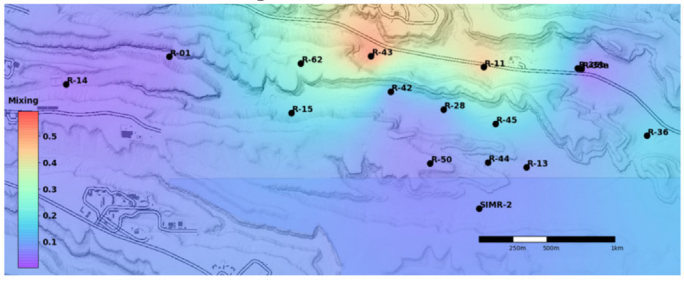
Source 6: Cl^{-},Ca,Mg,SO_4



Source 5: ^{3}H,Cr,Cl^{-},Ca,Mg,SO_4



Source 4: NO_3

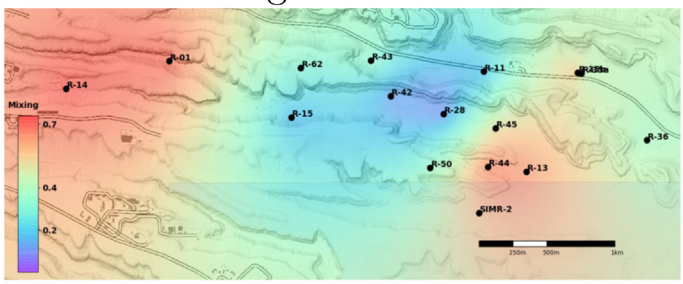


(d) January - December 2008

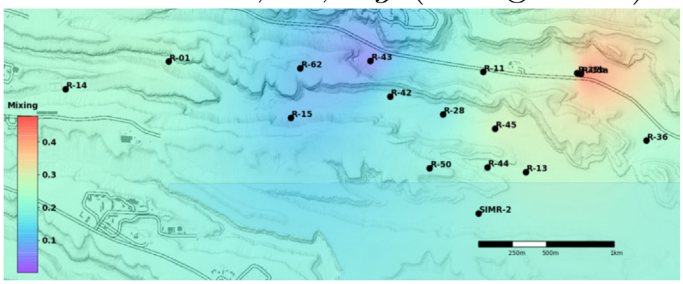
Fig. 8. (*continued*)

and complexity of the physical and geochemical processes impacting contaminant transport. However, even with this limitation, they provide a visual representation of the potential plume shapes. It is important to note that the NTF*k* analyses are very fast; the results presented here take minutes to generate. Based on our site modeling experience (LANL, 2018a), similar geochemical inverse-model analyses of all the data presented here will take months of computational work (the development and testing of the simulation models will make this process even longer). For example, the current LANL site model is open source and available at GitLab (Vesselinov et al., 2018); the model includes more than 140,000 calibration targets of pressure and concentration transients and more than 200 adjustable model parameters estimated during the model calibration. Currently, only the chromium transients are applied in the inversion process, and the model

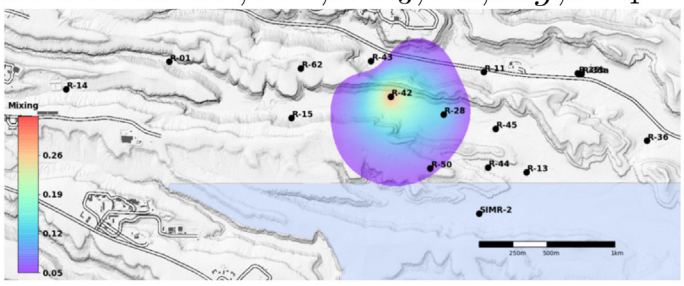
Source 7: background



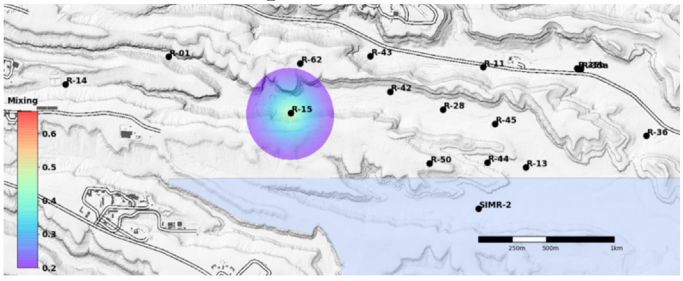
Source 3: Cl^{-},Ca,Mg (background)



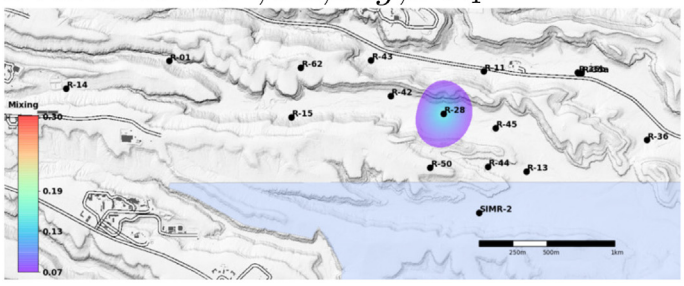
Source 1: Cr,Cl^{-},NO_3,Ca,Mg,SO_4



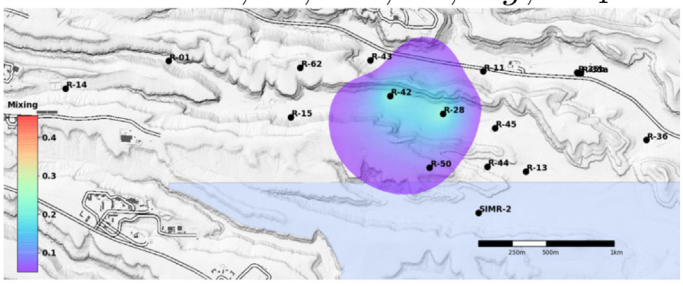
Source 2: ClO_4



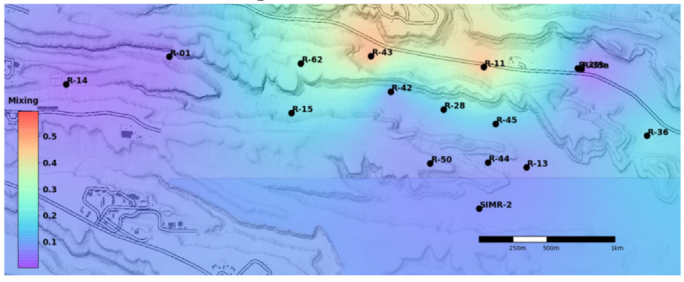
Source 6: Cl^{-},Ca,Mg,SO_4



Source 5: ^{3}H,Cr,Cl^{-},Ca,Mg,SO_4



Source 4: NO_3



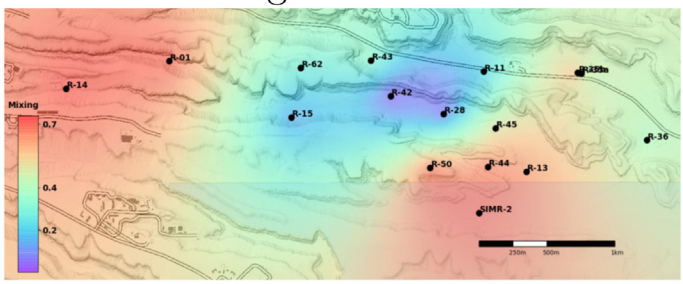
(e) January - December 2009

Fig. 8. (*continued*)

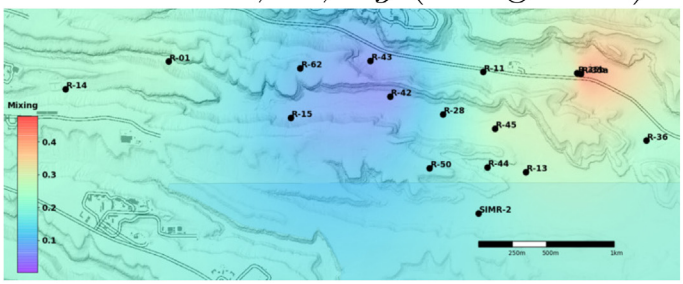
calibration takes about a month in parallel utilizing up to 640 processors.

The NTF*k* results presented and discussed above are consistent with more complicated inverse analyses using numerical models applied to solve this problem (Vesselinov et al., 2013; Vesselinov et al., n.d.-b; LANL, 2012; LANL, 2018a). The results are also generally consistent with machine-learning analyses obtained using a matrix-based technique (NMF*k*; (Vesselinov et al., n.d.-a)). In the future, the NTF*k* results will be applied as input to inverse analyses of site numerical models. In this way, instead of calibrating against all the geochemical data, the numerical models would be calibrated against the NTF*k* predicted geochemical mixtures. The anomalies detected at R-36, R-43#2 and R-28 through our unsupervised ML algorithm demonstrate its power for exploratory data analyses.

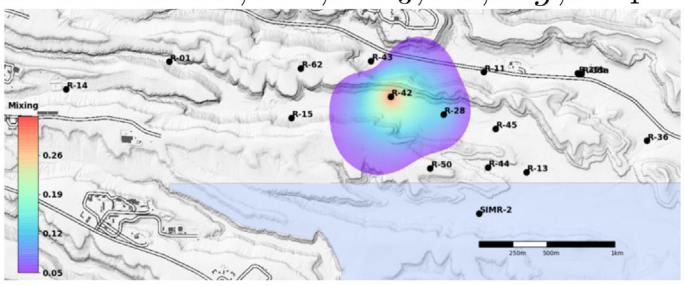
Source 7: background



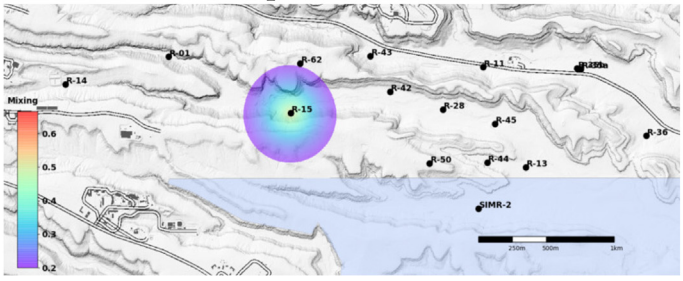
Source 3: Cl^{-},Ca,Mg (background)



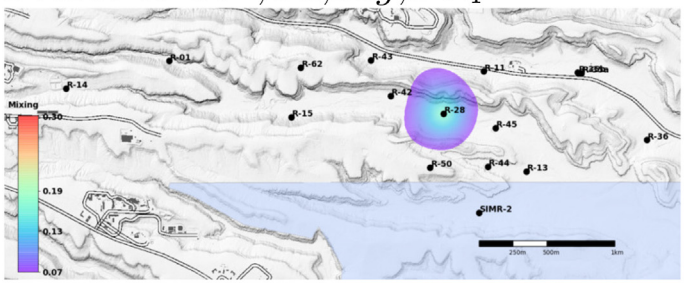
Source 1: Cr,Cl^{-},NO_3,Ca,Mg,SO_4



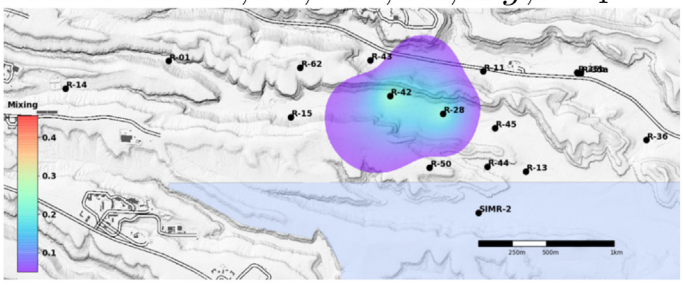
Source 2: ClO_4



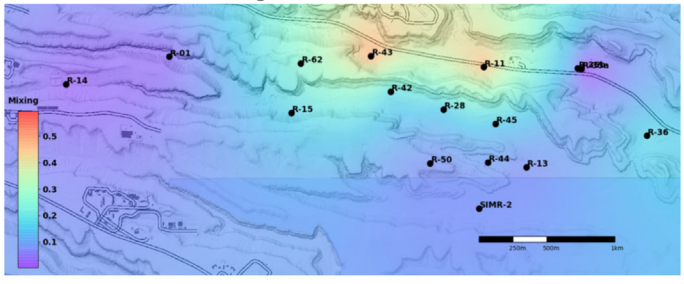
Source 6: Cl^{-},Ca,Mg,SO_4



Source 5: ^{3}H,Cr,Cl^{-},Ca,Mg,SO_4



Source 4: NO_3



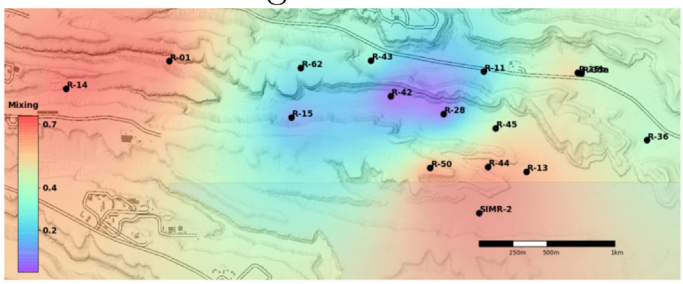
(f) January - December 2010

Fig. 8. (*continued*)

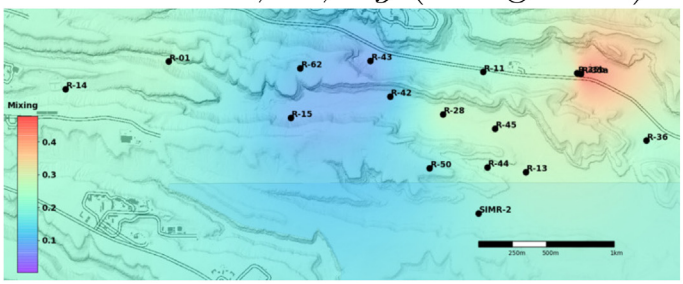
4. Conclusions

We have developed a novel unsupervised Machine Learning (ML) method based on Nonnegative Tensor Factorization (NTF) combined with a custom *k*-means clustering called NTF*k*. Our work demonstrates the applicability of our NTF*k* algorithm for Blind Source Separation (BSS). NTF*k* has been applied to identify contaminant sources based on high-dimensional (tensor) datasets representing spatial and temporal variation of observed geochemical species. The NTF*k* approach is an extension of our matrix-based machine learning methods presented in (Vesselinov et al., n.d.-a; Alexandrov and Vesselinov, 2014). Our results demonstrate that NTF*k* can be applied to identify (1) the unknown number of groundwater types (contaminant sources) present in an aquifer, (2) the original geochemical concentrations (signatures) of

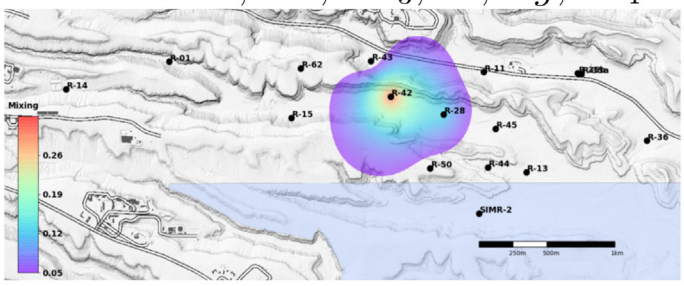
Source 7: background



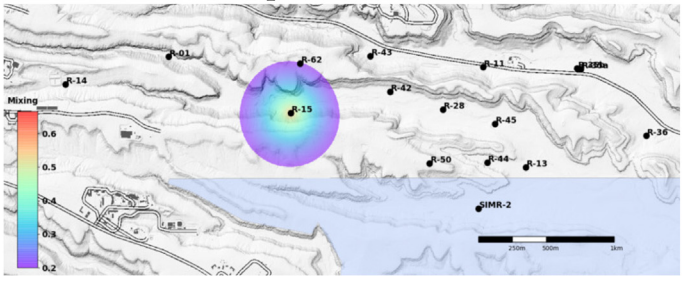
Source 3: Cl^{-},Ca,Mg (background)



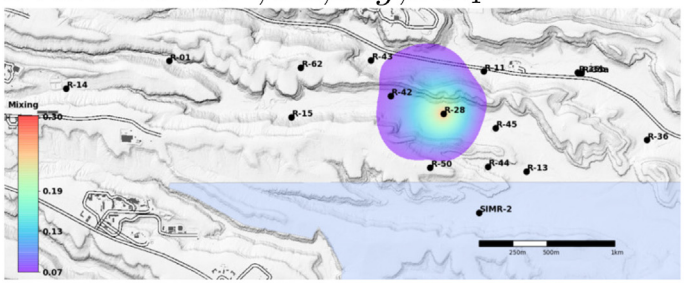
Source 1: Cr,Cl^{-},NO_3,Ca,Mg,SO_4



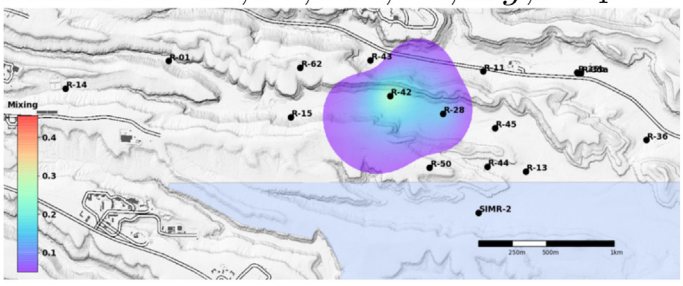
Source 2: ClO_4



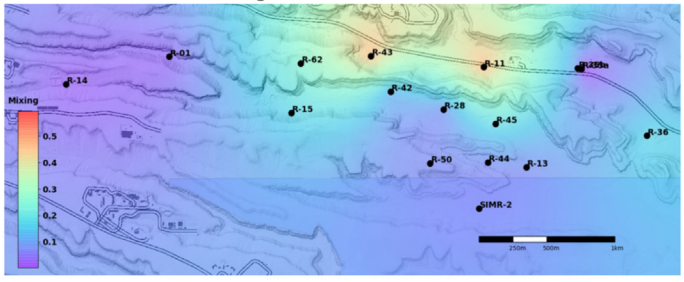
Source 6: Cl^{-},Ca,Mg,SO_4



Source 5: ^{3}H,Cr,Cl^{-},Ca,Mg,SO_4



Source 4: NO_3



(g) January - December 2011

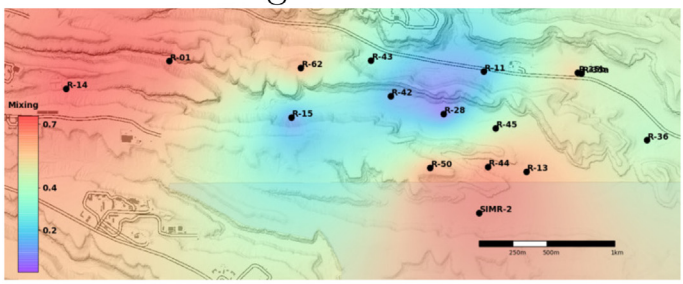
Fig. 8. (*continued*)

these groundwater types before their mixing in the aquifer, and (3) spatial and temporal dynamics in the mixing of these groundwater types.

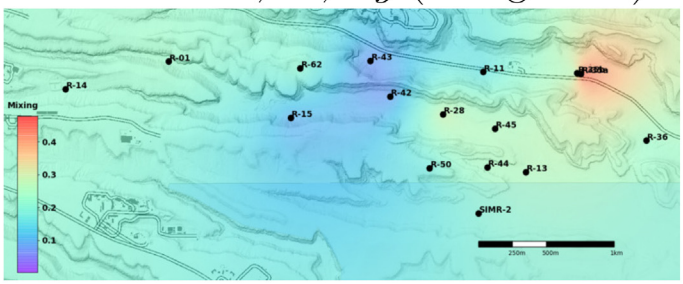
The inverse problem solved in the NTF*k* analysis is under-determined. To address this, the NTF*k* algorithm thoroughly explores the plausible inverse solutions, and seeks to narrow the set of possible solutions by estimating the number of contaminant source signals needed to robustly and accurately reconstruct the observed data.

In the synthetic tests, we generated datasets representing unknown contaminant sources detected as a set of mixed signals (groundwater types / contamination sources) at a series of monitoring wells (detectors / sensors) and for a series of time frames (snapshots). Using only the synthetic datasets representing the observed concentrations at the monitoring wells, NTF*k* correctly identified the number of contaminant

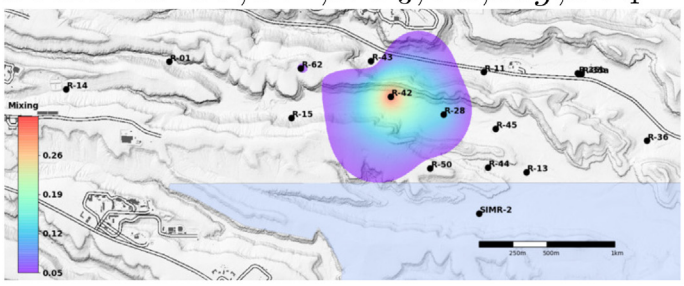
Source 7: background



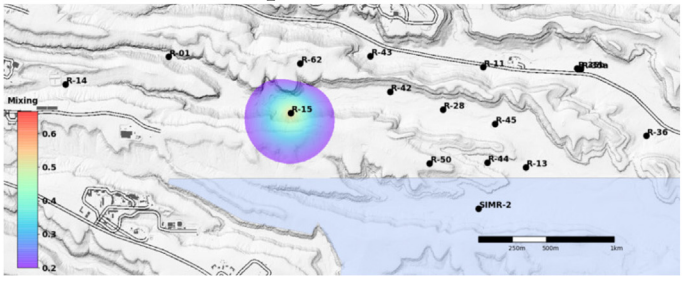
Source 3: Cl^{-},Ca,Mg (background)



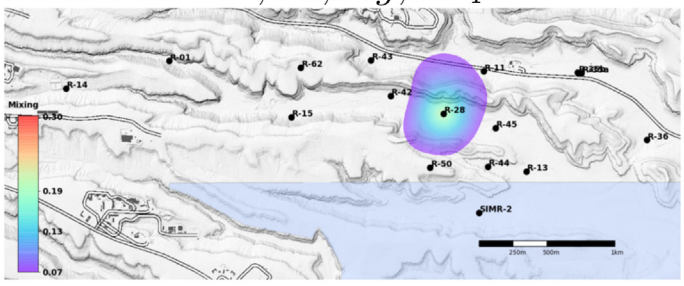
Source 1: Cr,Cl^{-},NO_3,Ca,Mg,SO_4



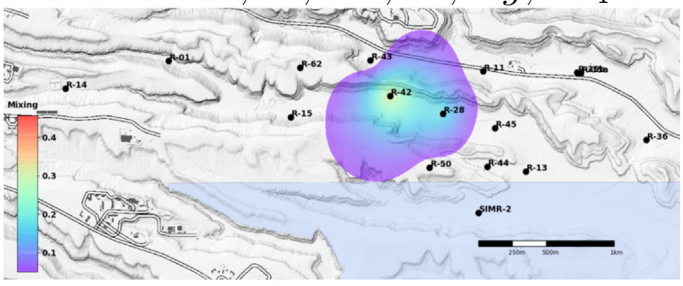
Source 2: ClO_4



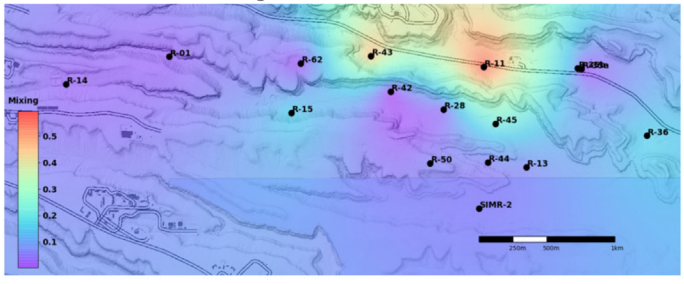
Source 6: Cl^{-},Ca,Mg,SO_4



Source 5: ^{3}H,Cr,Cl^{-},Ca,Mg,SO_4



Source 4: NO_3



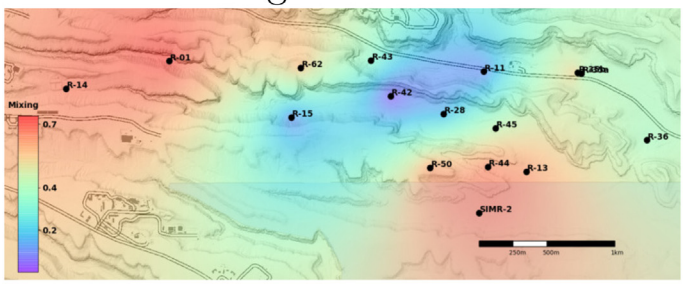
(h) January - December 2012

Fig. 8. (*continued*)

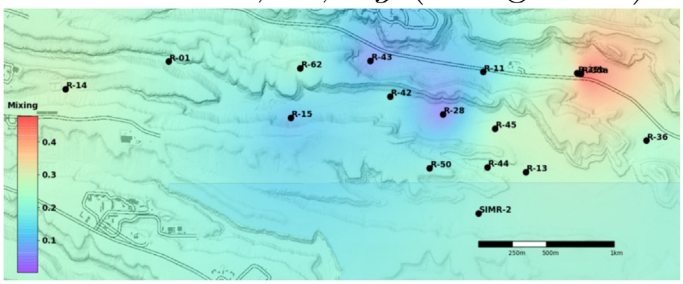
sources, the geochemical signatures of the original groundwater types before being mixed, and mixing coefficients at the wells over time.

We also applied NTF*k* on a real-world dataset related to the LANL chromium contamination site. The results of this analysis are consistent with previous data and model analyses conducted at the site (Vesselinov et al., n.d.-a; Vesselinov et al., 2013; Vesselinov et al., n.d.-b; LANL, 2012; LANL, 2018a), and provide additional insights. We highlight two insights in particular. The first is that the differences observed at the upper and lower screens R-43 suggest a late arriving contaminant source that has not had an opportunity to penetrate the deeper portion of the aquifer. The lack of ^{3}H in R-43#1 indicates that this late arriving source in the regional aquifer may be associated with an early contaminant release that took a long time to move through the vadose zone. The second is that the anomalous mixing ratios at R-36

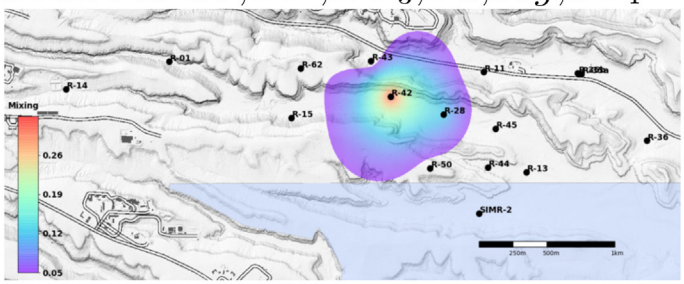
Source 7: background



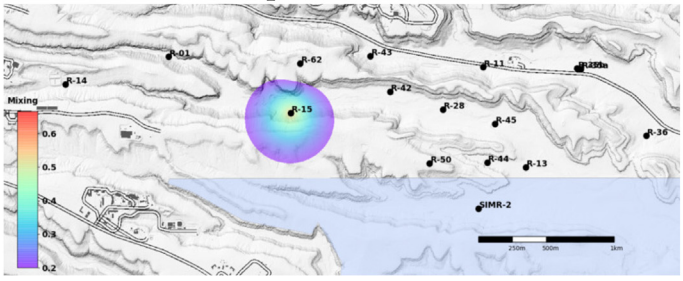
Source 3: Cl^{-},Ca,Mg (background)



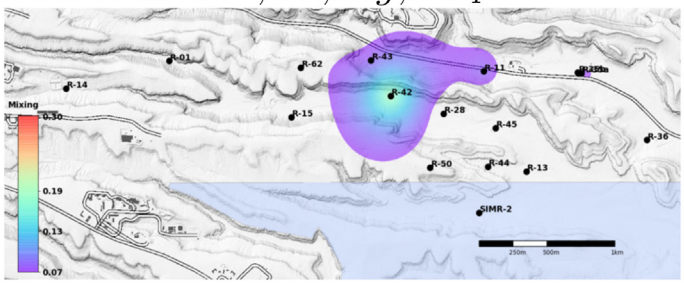
Source 1: Cr,Cl^{-},NO_3,Ca,Mg,SO_4



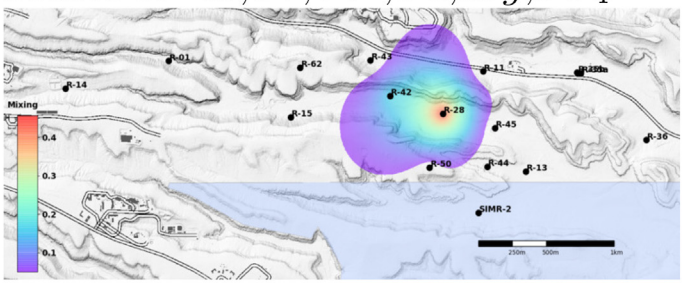
Source 2: ClO_4



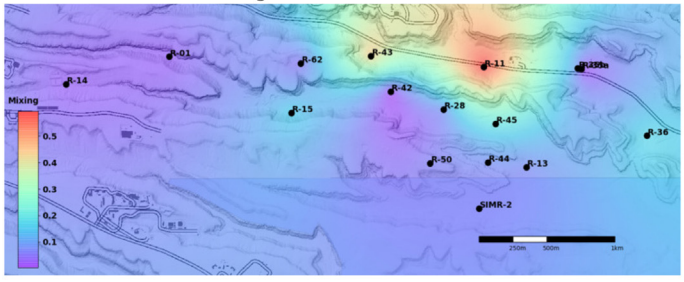
Source 6: Cl^{-},Ca,Mg,SO_4



Source 5: ^{3}H,Cr,Cl^{-},Ca,Mg,SO_4



Source 4: NO_3



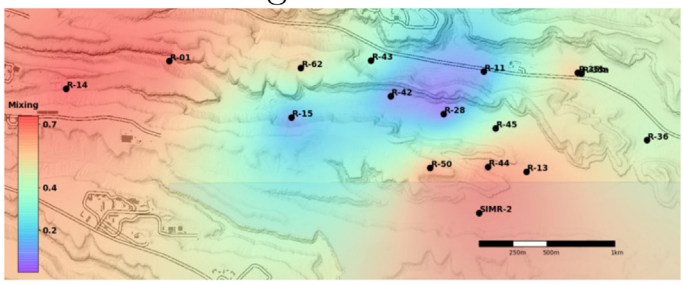
(i) January - December 2013

Fig. 8. (*continued*)

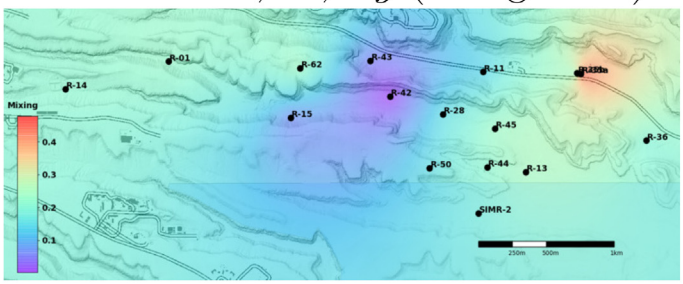
indicate that it may be in a perched zone that is disconnected from the regional aquifer. If so, it may be beneficial to add a monitoring well near R-36 that goes deeper into the subsurface. The anomalous hydrologic data at this location further corroborates the hypothesis that R-36 is disconnected from the aquifer (LANL, 2018a). In addition to these insights, the NTF*k* algorithm demonstrated its capabilities to identify systematic errors and anomalies in LANL site data.

NTF*k* allows the contaminant fields observed at a series of the detectors to be "unmixed" into a series of independent plumes with different geochemical signatures. This results from the NTF*k* analyses can be applied to guide the conceptualization of the site conditions and the design of numerical models that are developed to represent these conditions. In some cases, decoupled model analyses might be applied to independently analyze the groundwater transport of each

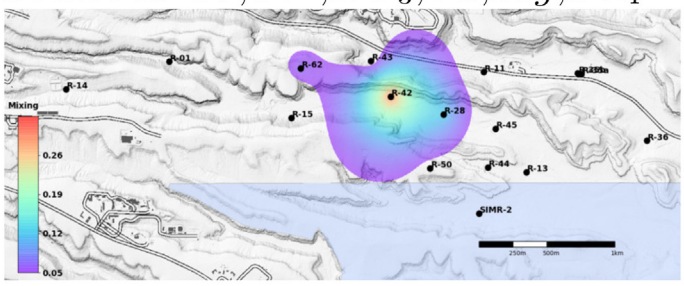
Source 7: background



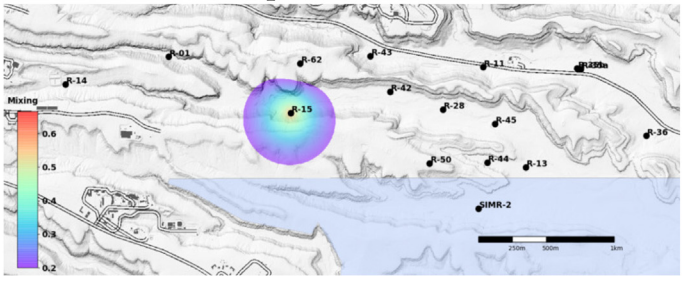
Source 3: Cl^{-},Ca,Mg (background)



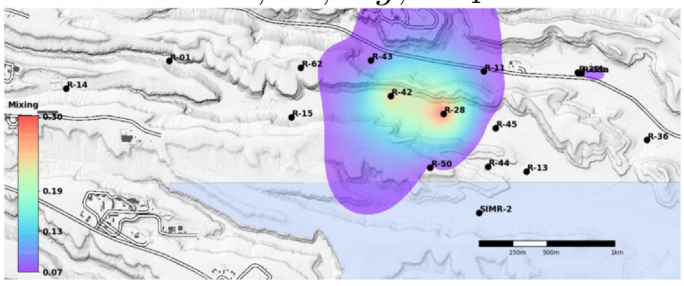
Source 1: Cr,Cl^{-},NO_3,Ca,Mg,SO_4



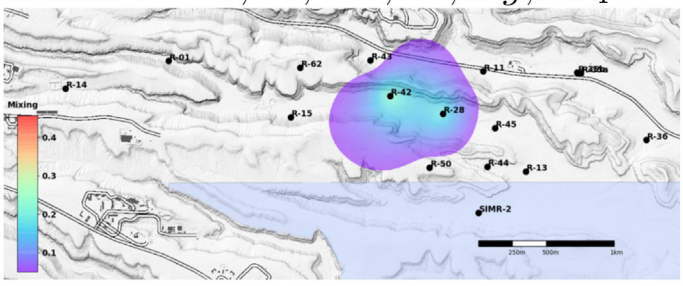
Source 2: ClO_4



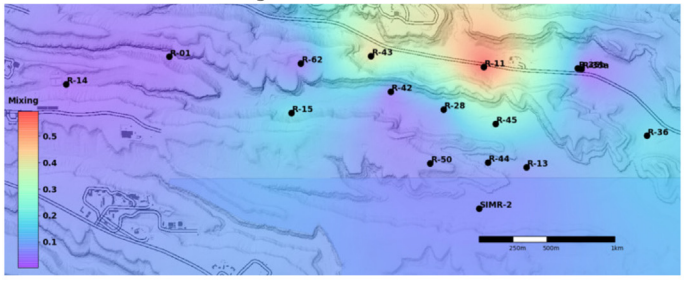
Source 6: Cl^{-},Ca,Mg,SO_4



Source 5: ^{3}H,Cr,Cl^{-},Ca,Mg,SO_4



Source 4: NO_3



(j) January - December 2014

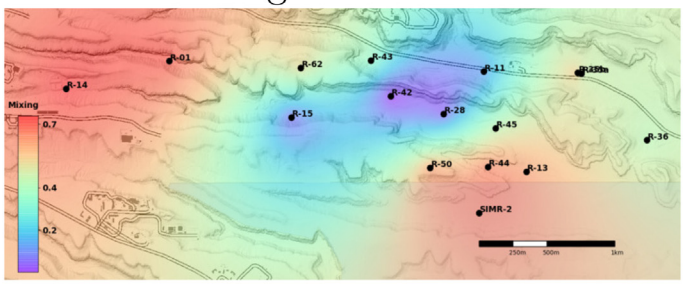
Fig. 8. (*continued*)

contaminant source which can be computationally much more efficient. NTF*k* results coupled with model analysis can yield crucial information needed to (1) development of site contaminant fate and transport conceptual models, (2) make predictions of contaminant behavior, (3) assess contamination risks, and (4) guide remediation strategies.

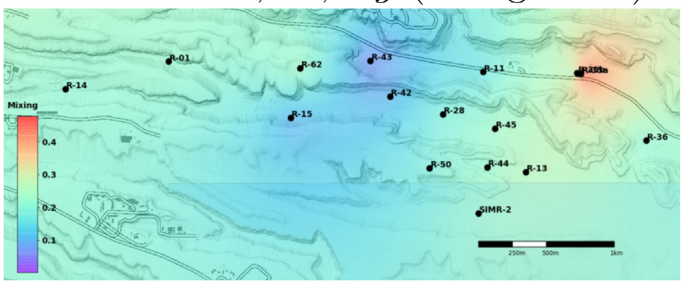
The NTF*k* analyses are fast and relatively easy to implement. An open-source code written in Julia (Bezanson et al., 2014) is in development and will be released soon. All the analyses presented in the paper take several minutes to execute in serial. Since most of the computations are independent, the algorithm can be performed also in parallel which further increases its computational efficiency and scalability.

It is important to note that the presented NTF analyses are following the classical BSS formulation assuming a linear mixing problem.

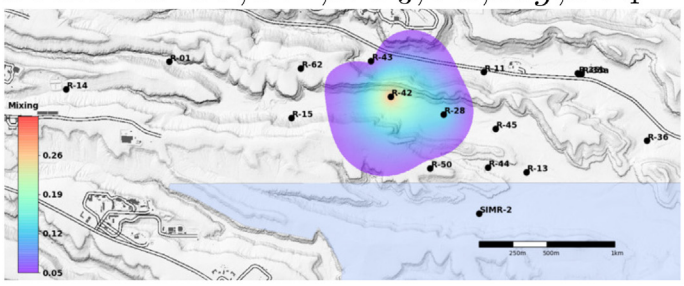
Source 7: background



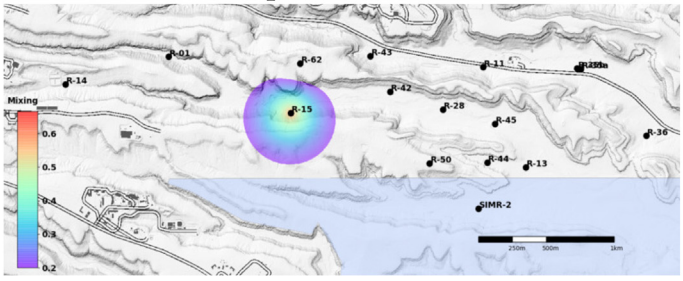
Source 3: Cl^{-},Ca,Mg (background)



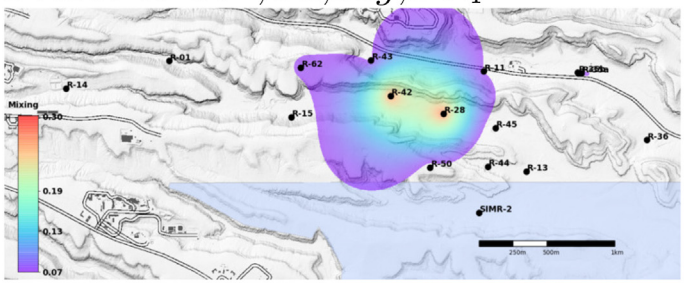
Source 1: Cr,Cl^{-},NO_3,Ca,Mg,SO_4



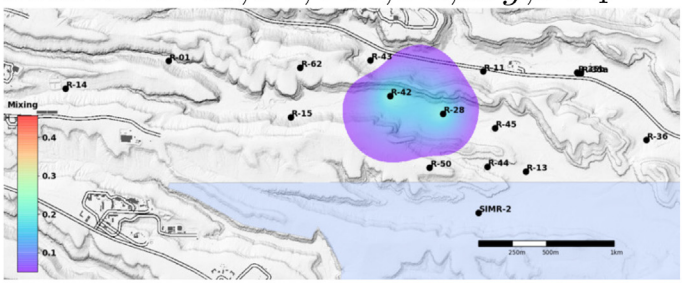
Source 2: ClO_4



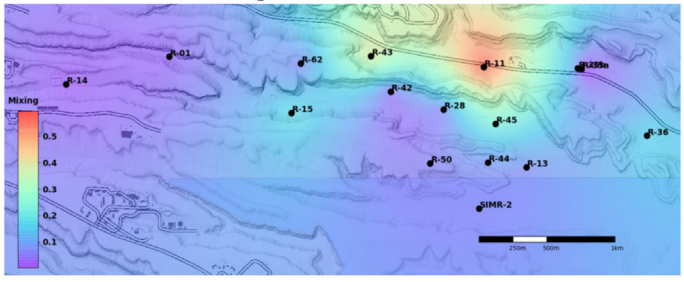
Source 6: Cl^{-},Ca,Mg,SO_4



Source 5: ^{3}H,Cr,Cl^{-},Ca,Mg,SO_4



Source 4: NO_3



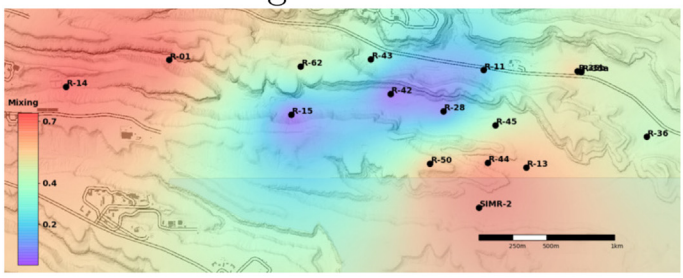
(k) January - December 2015

Fig. 8. (*continued*)

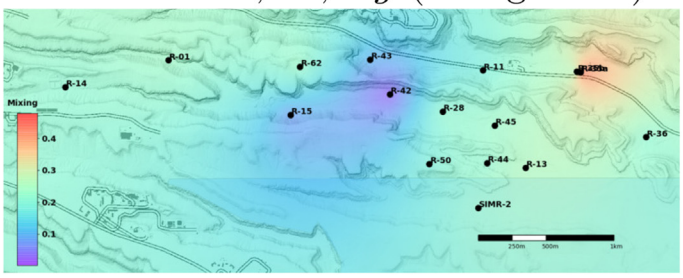
However, since the NTF problem is solved using nonlinear minimization procedure as discussed in Section 2.1, the BSS problem can be expanded to account for nonlinear mixing or geochemical processes. This will increase the number of unknowns as well as the computational complexity but as long as data are available to represent nonlinear processes, the BSS problem can be solved. We plan to extend our ML analyses to account for nonlinear processes in the future.

In summary, the major pros of the proposed methodology are that it is fast, scalable and unbiased. It can be applied to tackle large high-dimensional site datasets without prior site knowledge and assumptions. The cons are that it assumes linearity (in its current form) and requires informative monitoring data. If the latter is an issue, our ML methodology can be applied to evaluate information content of the data and guide additional data collection strategies that can provide

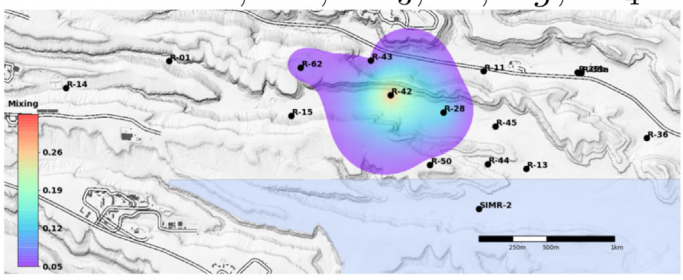
Source 7: background



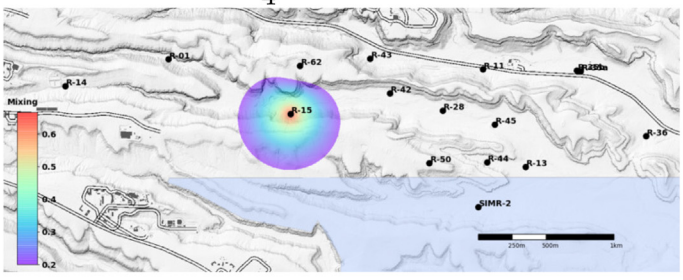
Source 3: Cl^{-},Ca,Mg (background)



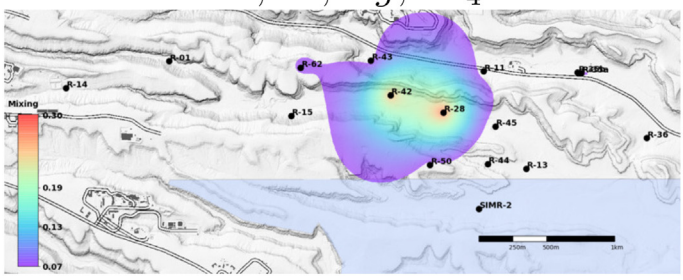
Source 1: Cr,Cl^{-},NO_3,Ca,Mg,SO_4



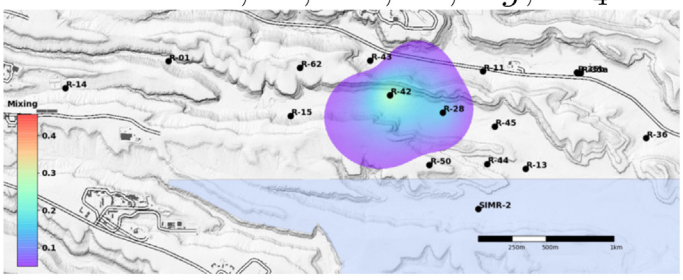
Source 2: ClO_4



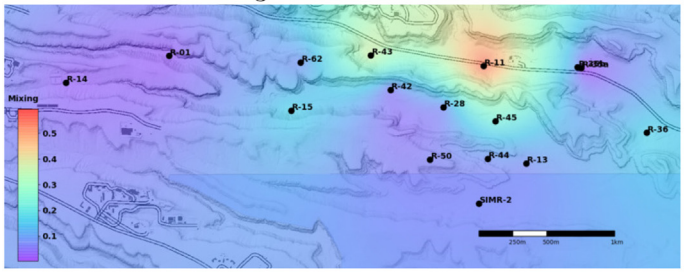
Source 6: Cl^{-},Ca,Mg,SO_4



Source 5: ^{3}H,Cr,Cl^{-},Ca,Mg,SO_4



Source 4: NO_3



(l) January - December 2016

Fig. 8. (*continued*)

informative data.

The possible applications of the NTF*k* approach are not limited to groundwater contamination problems. NTF*k* can be readily used to identify contaminant sources based on soil and air pollution data. NTF*k* can be applied to analyze any mixture of ingredients. In this case, our constrained NTF*k* algorithm can be applied to identify the ingredients of the sources that are mixed to produce observed mixtures.

NTF*k* is applicable for unsupervised ML analyses for solving various types of data analytics problems including feature extraction and exploratory analyses. NTF*k* can also be applied to large high-dimensional datasets with constraints only related to physical memory of the computational resources that are used.

Acknowledgments

This research was funded by the Environmental Programs

Directorate and LDRD grants 20180060DR and 20190020DR of the Los Alamos National Laboratory. In addition, Velimir V. Vesselinov and Daniel O'Malley were supported by the DiaMonD project (An Integrated Multifaceted Approach to Mathematics at the Interfaces of Data, Models, and Decisions, U.S. Department of Energy, Office of Science, Grant #11145687). This research used resources provided by the Los Alamos National Laboratory Institutional Computing Program, which is supported by the U.S. Department of Energy, National Nuclear Security Administration. The authors also want to acknowledge the excellent comments provided by the anonymous reviewers that substantially improved the manuscript.

References

Akaike, H., 2011. Akaike's Information Criterion. In: Lovric, M. (Ed.), International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_110.

Alexandrov, B.S., Vesselinov, V.V., 2014. Blind source separation for groundwater pressure analysis based on nonnegative matrix factorization. Water Resour. Res. 50 (9), 7332–7347.

Amari, S.-i., Cichocki, A., Yang, H.H., 1996. A new learning algorithm for blind signal separation. Adv. Neural Inf. Proces. Syst. 8, 757–763.

Andersson, C.A., Bro, R., 2000. The N-way toolbox for MATLAB. Chemom. Intell. Lab. Syst. 52 (1), 1–4.

Atmadja, J., Bagtzoglou, A.C., 2001. Pollution source identification in heterogeneous porous media. Water Resour. Res. 37 (8), 2113–2125.

Ayvaz, M.T., 2010. A linked simulation–optimization model for solving the unknown groundwater pollution source identification problems. J. Contam. Hydrol. 117 (1–4), 46–59.

Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., Moulines, E., 1997. A blind source separation technique using second-order statistics. IEEE Trans. Signal Process. 45 (2), 434–444.

Bezanson, J., Edelman, A., Karpinski, S., Shah, V.B., 2014. Julia: a Fresh Approach to Numerical Computing. SIAM Rev. 59 (1), 65–98 (ISSN 0036-1445).

J. Bezanson, S. Karpinski, V. B. Shah, A. Edelman, Julia: A fast dynamic language for technical computing

Böhlke, J., Denver, J., 1995. Combined use of groundwater dating, chemical, and isotopic analyses to resolve the history and fate of nitrate contamination in two agricultural watersheds, Atlantic coastal plain, Maryland. Water Resour. Res. 31 (9), 2319–2339.

Borukhov, V., Zayats, G., 2015. Identification of a time-dependent source term in nonlinear hyperbolic or parabolic heat equation. Int. J. Heat Mass Transf. 91, 1106–1113.

Cervone, G., Franzese, P., Keesee, A.P., 2010. Algorithm quasi-optimal (AQ) learning. Wiley Interdisc. Rev. 2 (2), 218–236.

Chan, C.W., Huang, G.H., 2003. Artificial intelligence for management and control of pollution minimization and mitigation processes. Eng. Appl. Artif. Intell. 16 (2), 75–90.

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.

De Lathauwer, L., De Moor, B., Vandewalle, J., 2000. A multilinear singular value decomposition. SIAM J. Matrix Anal. Appl. 21 (4), 1253–1278.

Deutsch, W.J., Siegel, R., 1997. Groundwater Geochemistry: Fundamentals and Applications to Contamination. CRC press.

Diday, E., Simon, J., 1980. Clustering analysis. In: Digital Pattern Recognition. Springer, pp. 47–94.

Drucker, H., Wu, D., Vapnik, V.N., 1999. Support vector machines for spam categorization. IEEE Trans. Neural Netw. 10 (5), 1048–1054.

I. Dunning, J. Huchette, M. Lubin, JuMP: A modeling language for mathematical optimization

Fetter, C.W., Fetter, C., 1999. Contaminant Hydrogeology. Prentice Hall, New Jersey.

Gelhar, L.W., 1993. Stochastic Subsurface Hydrology. Prentice-Hall.

Guan, J., Aral, M.M., Maslia, M.L., Grayman, W.M., 2006. Identification of contaminant sources in water distribution systems using simulation–optimization method: case study. J. Water Resour. Plan. Manag. 132 (4), 252–262.

Gzyl, G., Zanini, A., Fraczek, R., Kura, K., 2014. Contaminant source and release history identification in groundwater: a multi-step approach. J. Contam. Hydrol. 157, 59–72.

Hamdi, A., Mahfoudhi, I., 2013. Inverse source problem in a one-dimensional evolution linear transport equation with spatially varying coefficients: application to surface water pollution. Inverse Problems Sci. Eng. 21 (6), 1007–1031.

Hammond, G.E., Lichtner, P.C., Mills, R., 2014. Evaluating the performance of parallel subsurface simulators: an illustrative example with PFLOTRAN. Water Resour. Res. 50 (1), 208–228.

Hansen, S.K., Pandey, S., Karra, S., Vesselinov, V.V., 2017. CHROTRAN 1.0: a mathematical and computational model for in situ heavy metal remediation in heterogeneous aquifers. Geosci. Model Dev. 10 (12), 4525–4538.

Harman, H.H., 1976. Modern Factor Analysis. University of Chicago Press.

Harshman, R.A., Lundy, M.E., 1994. PARAFAC: parallel factor analysis. Comput. Stat. Data Anal. 18 (1), 39–72.

Haykin, S., Chen, Z., 2005. The cocktail party problem. Neural Comput. 17 (9), 1875–1902.

Helena, B., Pardo, R., Vega, M., Barrado, E., Fernandez, J.M., Fernandez, L., 2000. Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. Water Res. 34 (3), 807–816.

Hitchcock, F.L., 1927. The expression of a tensor or a polyadic as a sum of products. Stud. Appl. Math. 6 (1–4), 164–189.

Illman, W.A., Berg, S.J., Liu, X., Massi, A., 2010. Hydraulic/partitioning tracer tomography for DNAPL source zone characterization: Small-scale sandbox experiments. Environ. Sci. Technol. 44 (22), 8609–8614.

James, A.I., Graham, W.D., Hatfield, K., Rao, P., Annable, M.D., 2000. Estimation of spatially variable residual nonaqueous phase liquid saturations in nonuniform flow fields using partitioning tracer data. Water Resour. Res. 36 (4), 999–1012.

Jin, M., Delshad, M., Dwarakanath, V., McKinney, D.C., Pope, G.A., Sepehrnoori, K., Tilburg, C.E., Jackson, R.E., 1995. Partitioning tracer test for detection, estimation, and remediation performance assessment of subsurface nonaqueous phase liquids. Water Resour. Res. 31 (5), 1201–1211.

Jolliffe, I.T., 1986. Principal component analysis and factor analysis. In: Principal Component Analysis, Springer Series in Statistics, Springer Science + Business Media, LLC, New York, USA, pp. 115–128.

Jolliffe, I., 2002. Principal Component Analysis. Wiley Online Library.

A. Khalil, M. N. Almasri, M. McKee, J. J. Kaluarachchi, Applicability of statistical learning algorithms in groundwater quality modeling, Water Resour. Res. 41 (5).

Knudson, E.J., Duewer, D.L., Christian, G.D., Larson, T.V., 1977. Application of factor analysis to the study of rain chemistry in the Puget Sound region. In: Chemometric: Theory and Application. ACS Symposium Series, Washington, DC, pp. 80–116.

Kolda, T.G., Bader, B.W., 2009. Tensor decompositions and applications. SIAM Rev. 51 (3), 455–500.

LANL, 2009. Investigation Report for Sandia Canyon. LANL.

LANL, 2012. Phase II Investigation Report for Sandia Canyon. LANL.

LANL, 2018a. Compendium of Technical Reports Related to Chroimum Contaminated Groundwater at Los Alamos National Laboratory. LANL.

LANL, 2018b. Evaluation of Potential Source Areas for the Chromium Plume Using Machine Learning Data Analyses of Geochemical Data. LANL.

Lapworth, D., Baran, N., Stuart, M., Ward, R., 2012. Emerging organic contaminants in groundwater: a review of sources, fate and occurrence. Environ. Pollut. 163, 287–303.

Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401 (6755), 788–791.

Mamonov, A.V., Tsai, Y.R., 2013. Point source identification in nonlinear advection–diffusion–reaction systems. Inverse Problems 29 (3), 035009.

A. M. Michalak, P. K. Kitanidis, Estimation of historical groundwater contaminant distribution using the adjoint state method applied to geostatistical inverse modeling, Water Resour. Res. 40 (8).

Mørup, M., Hansen, L.K., Arnfred, S.M., 2008. Algorithms for sparse nonnegative Tucker decompositions. Neural Comput. 20 (8), 2112–2131.

Murray-Bruce, J., Dragotti, P.L., 2014. Spatio-temporal sampling and reconstruction of diffusion fields induced by point sources. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 31–35.

Neupauer, R.M., Borchers, B., Wilson, J.L., et al., 2000. Comparison of inverse methods for reconstructing the release history of a groundwater contamination source. Water Resour. Res. 36 (9), 2469–2475.

O'Malley, D., Vesselinov, V.V., 2018. Analytical Solutions for Groundwater Contaminant Transport in Julia. URL. https://github.com/madsjulia/Anasol.jl.

Paatero, P., Tapper, U., 1994. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5 (2), 111–126.

Pang-Ning, T., Steinbach, M., Kumar, V., 2006. Introduction to Data Mining. Addison-Wesley.

Park, E., Zhan, H., 2001. Analytical solutions of contaminant transport from finite one-, two-, and three-dimensional sources in a finite-thickness aquifer. J. Contam. Hydrol. 53 (1–2), 41–61.

Rasekh, A., Brumbelow, K., 2012. Machine learning approach for contamination source identification in water distribution systems. In: World Environmental and Water Resources Congress, Palm Springs, CA.

Ross, D.A., Zemel, R.S., 2006. Learning parts-based representations of data. J. Mach. Learn. Res. 7, 2369–2397.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.

Scholkopft, B., Mullert, K.-R., 1999. Fisher discriminant analysis with kernels. Neural Networks Signal Process. 1 (1), 1.

Shrestha, S., Kazama, F., 2007. Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan. Environ. Model Softw. 22 (4), 464–475.

Sun, A.Y., Painter, S.L., Wittmeyer, G.W., 2006. A robust approach for iterative contaminant source location and release history recovery. J. Contam. Hydrol. 88 (3–4), 181–196.

Tariq, S.R., Shah, M.H., Shaheen, N., Jaffar, M., Khalique, A., 2008. Statistical source identification of metals in groundwater exposed to industrial contamination. Environ. Monit. Assess. 138 (1–3), 159–165.

Throckmorton, H.M., Newman, B.D., Heikoop, J.M., Perkins, G.B., Feng, X., Graham, D.E., O'Malley, D., Vesselinov, V.V., Young, J., Wullschleger, S.D., et al., 2016. Active layer hydrology in an arctic tundra ecosystem: quantifying water sources and cycling using water stable isotopes. Hydrol. Process. 30 (26), 4972–4986.

Tipping, M.E., 2001. Sparse Bayesian learning and the relevance vector machine. J. Mach. Learn. Res. 1 (Jun), 211–244.

Tucker, L.R., 1966. Some mathematical notes on three-mode factor analysis. Psychometrika 31 (3), 279–311.

Vengosh, A., Jackson, R.B., Warner, N., Darrah, T.H., Kondash, A., 2014. A critical review

of the risks to water resources from unconventional shale gas development and hydraulic fracturing in the United States. Environ. Sci. Technol. 48 (15), 8334–8348.
Vesselinov, V.V., O'Malley, D., 2016a. Model analysis of complex systems behavior using MADS. In: AGU Fall Meeting, San Francisco, CA.
Vesselinov, V.V., O'Malley, D., 2016b. Model Analyses and Decision Support in Julia. URL. http://mads.lanl.gov/.
Vesselinov, V.V., Broxton, D., Birdsell, K., Reneau, S., Harp, D.R., Mishra, P.K., Katzman, D., Goering, T., Vaniman, D., Longmire, P., Fabryka-Martin, J., Heikoop, J., Ding, M., Hickmott, D., Jacobs, E., 2013. Data and model-driven decision support for environ. manag. of a chromium plume at Los Alamos National Laboratory. In: WMSYM2013, Phoenix, Arizona, USA.
Vesselinov, V.V., O'Malley, D., Hyman, J., 2018. Waffle2017: Model of Groundwater Ow and Transport at the LANL Chromium Site. URL. https://gitlab.com/LANL-EM/waffle2017.
V. V. Vesselinov, B. S. Alexandrov, D. O'Malley, Contaminant source identification using semi-supervised machine learning, J. Contam. Hydrol. .
V. V. Vesselinov, D. 0'Malley, D. Katzman, Model-Assisted Decision analyses Related to a Chromium Plume at Los Alamos National Laboratory, in: WMSYM2015, Phoenix, Arizona, USA, 2015.
Vijayakumar, S., Schaal, S., 2000. Locally weighted projection regression: Incremental real time learning in high dimensional space. In: Proceedings of the Seventeenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc, pp. 1079–1086.
Wächter, A., 2002. An Interior Point Algorithm for Large-Scale Nonlinear Optimization with Applications in Process Engineering. Ph.D. Thesis. Carnegie Mellon University, Pittsburgh, PA, USA.
Wächter, A., Biegler, L.T., 2005. Line search filter methods for nonlinear programming: Motivation and global convergence. SIAM J. Optim. 16 (1), 1–31.
Wächter, A., Biegler, L.T., 2006. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. Math. Program. 106 (1), 25–57 (ISSN 1436-4646).
Wagner, B.J., 1992. Simultaneous parameter estimation and contaminant source characterization for coupled groundwater flow and contaminant transport modelling. J. Hydrol. 135 (1), 275–303.
Wexler, E., Wexler, B.E.J., 1992. Analytical Solutions for One-, Two-, and Three-Dimensional Solute Transport in Ground-Water Flow Systems with Uniform Flow.
Yegnanarayana, B., 2009. Artificial Neural Networks. PHI Learning Pvt. Ltd.
T.-C. J. Yeh, J. Zhu, Hydraulic/partitioning tracer tomography for characterization of dense nonaqueous phase liquid source zones, Water Resour. Res. 43 (6).
Zhang, Y., Graham, W.D., 2001. Spatial characterization of a hydrogeochemically heterogeneous aquifer using partitioning tracers: Optimal estimation of aquifer parameters. Water Resour. Res. 37 (8), 2049–2063.