# Identification of anomalous diffusion sources by unsupervised learning

Raviteja Vangara [1,2] Kim Ø. Rasmussen [1] Dimiter N. Petsev [2,3] Golan Bel,[4,5,*] and Boian S. Alexandrov [1,†]

[1]*Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87544, USA*
[2]*Department of Chemical and Biological Engineering, University of New Mexico, Albuquerque, New Mexico 87106, USA*
[3]*Center for Microengineered Materials, University of New Mexico, Albuquerque, New Mexico 87131, USA*
[4]*Department of Solar Energy and Environmental Physics, BIDR, Ben-Gurion University of the Negev, Sede Boqer Campus 8499000, Israel*
[5]*Center of Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*

Fractional Brownian motion (fBm) is a ubiquitous diffusion process in which the memory effects of the stochastic transport result in the mean-squared particle displacement following a power law $\langle \Delta r^2 \rangle \sim t^\alpha$, where the diffusion exponent $\alpha$ characterizes whether the transport is subdiffusive ($\alpha < 1$), diffusive ($\alpha = 1$), or superdiffusive ($\alpha > 1$). Due to the abundance of fBm processes in nature, significant efforts have been devoted to the identification and characterization of fBm sources in various phenomena. In practice, the identification of the fBm sources often relies on solving a complex and ill-posed inverse problem based on limited observed data. In the general case, the detected signals are formed by an unknown number of release sources, located at different locations and with different strengths, that act simultaneously. This means that the observed data are composed of mixtures of releases from an unknown number of sources, which makes the traditional inverse modeling approaches unreliable. Here, we report an unsupervised learning method, based on non-negative matrix factorization, that enables the identification of the unknown number of release sources as well the anomalous diffusion characteristics based on limited observed data and the general form of the corresponding fBm Green's function. We show that our method performs accurately for different types of sources and configurations with a predetermined number of sources with specific characteristics and introduced noise.

## I. INTRODUCTION

Anomalous diffusion has been observed in numerous systems, and a variety of underlying mechanisms have been discussed [1,2]. Anomalous diffusion is associated with the nonlinear dependence of the mean-square displacement on time, $\langle \Delta \mathbf{r}(t) \rangle \sim t^\alpha$ with $\alpha \neq 1$. Different mechanisms can lead to the same asymptotic dependence of the mean-square displacement on time, or alternatively, to the same diffusion exponent. However, such processes can differ in their propagator, probability distribution function [1,2], aging [3], and ergodic properties [4]. Fractional Brownian motion (fBm) is a common model for anomalous diffusion which stems from long-range correlations, stationarity, and scaling of the increments [5]. The fBm has applications in many fields including finance, climate, solar activity, hydrology, turbulence, and many others [6]. The fBm diffusion exponent $\alpha$ determines the diffusion regime. When $\alpha < 1$, the process is subdiffusive; when $\alpha > 1$, the process is superdiffusive; and when $\alpha = 1$, we have the normal Brownian diffusion [1]. Examples of subdiffusion are the kinetics of passive molecular tracers in lipid bilayers [7], hopping transport in disordered systems [8], propagation of nonlinear waves in quasiperiodic potentials [9], evolution of index prices in financial systems [10], and many others. Superdiffusion is typically associated with active processes, and has been observed in living cells [11], in radiative transport [12], in intracellular particle motion [13], in transport processes in porous media [14–16], and in many other cases.

Due to the non-Markovian nature of fBm, much effort has been devoted to the development of inference methods for the diffusion characteristics from observed data. Most of this research has focused on parameter inference using inverse modeling [17–26]. Inverse modeling typically requires identification of the characteristics of the sources as well as of the medium. For fractional diffusion, there is the additional complexity of the propagator describing the process, which has focused the inverse modeling efforts on one-dimensional problems [18,19,21,22,24,25,27], while less attention has been devoted to two- and three-dimensional problems [23,28]. Furthermore, most research in this field assumes that the measurements exist at high spatial and temporal resolutions, a situation that is rarely encountered. Usually, the identification of release sources relies on solving a complex ill-posed inverse model against limited amounts of observed data. Importantly, in most of the naturally occurring fBm phenomena, the number of the release sources (and their locations and strengths) is unknown, which severely limits the reliability of the classical inverse modeling.
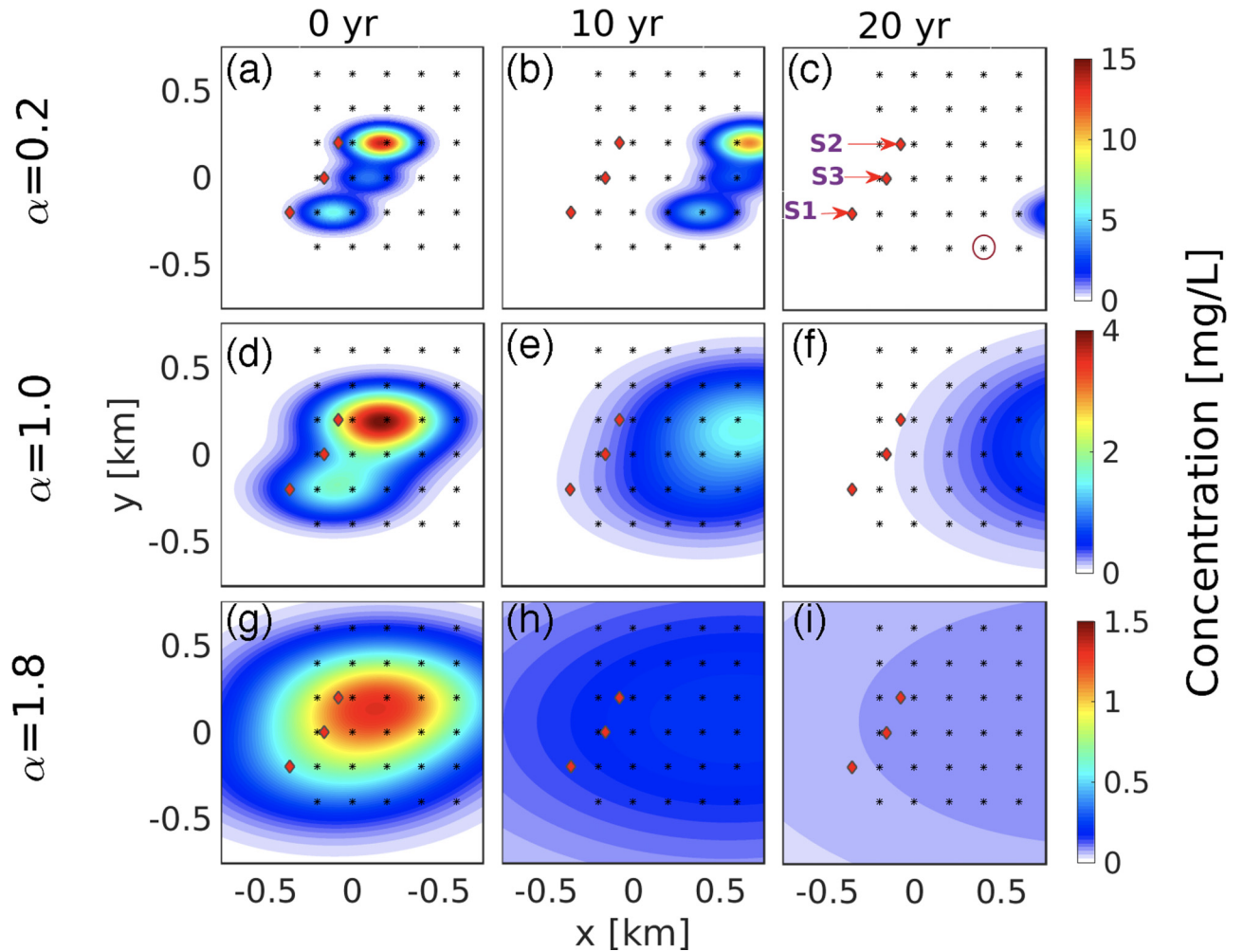
*bel@bgu.ac.il
†boian@lanl.gov

FIG. 1. The configuration of point sources (red diamonds) and the array of detectors (black points). The panels in each row correspond to different values of the diffusion exponent $\alpha$ (as specified), while the columns correspond to the denoted times. All the sources emitted at time $t_0 = -5$ years. The source strengths are $q_{1,2,3} = 0.25, 0.6, 0.15 \times 10^{12}$ mg/km, respectively, and their positions are (in km) $[-0.36, -0.2]$, $[-0.08, 0.2]$, and $[-0.16, 0]$, respectively. The color denotes the concentration at the different locations and times for each type of diffusion.

Recently, a hybrid method based on matrix factorization (NMF) combined with Brownian diffusion Green's function, called hNMF, was proposed for identifying the properties of unknown number of emission sources releasing simultaneously [29].

Here, we report a generalization of hNMF that enables identification of fBm diffusion processes (subdiffusion, superdiffusion, or normal diffusion) based on limited observed data and fBm Green's functions. To demonstrate the performance of our generalized method, we generate examples that include anisotropic two-dimensional fBm with drift and a predetermined number of release sources. We show that our method determines accurately the unknown number, locations, and properties of the release sources used to generate the data (including point sources as well as spatially and temporally extended sources). We also demonstrate that the generalized hNMF correctly determines the generalized diffusion coefficients, the advection velocity, and the diffusion exponent $\alpha$, and that it accurately estimates the spatial and temporal extension of the emission in the presence of noise.

## II. PROBLEM FORMULATION

We focus on the problem of identifying the anomalous diffusion process characteristics based on a limited number of concentration time series recorded by spatially fixed detectors. This is fundamentally an inverse problem, but unlike many previous works, our method is based on the recording of the concentration at specific locations rather than on single-particle tracking. The general problem is complex due to the large number of parameters and the need to identify several different characteristics. The fundamental step is to be able to identify the unknown number of sources and their locations and strengths. This is not a simple task due to the fact that the concentrations, recorded by each detector, potentially contain contributions from all sources. Further, each source may have a finite extent in space and may have a temporally varying amplitude of emission. In addition to the number of sources $\tilde{N}_s$, their locations $x_s$, $y_s$, and strengths $q_s$ (where $s = 1, 2, \ldots, \tilde{N}_s$) the medium characteristics must also be determined. The medium, in which the fBm process takes place, has the following characteristics: (1) the type of diffusion,
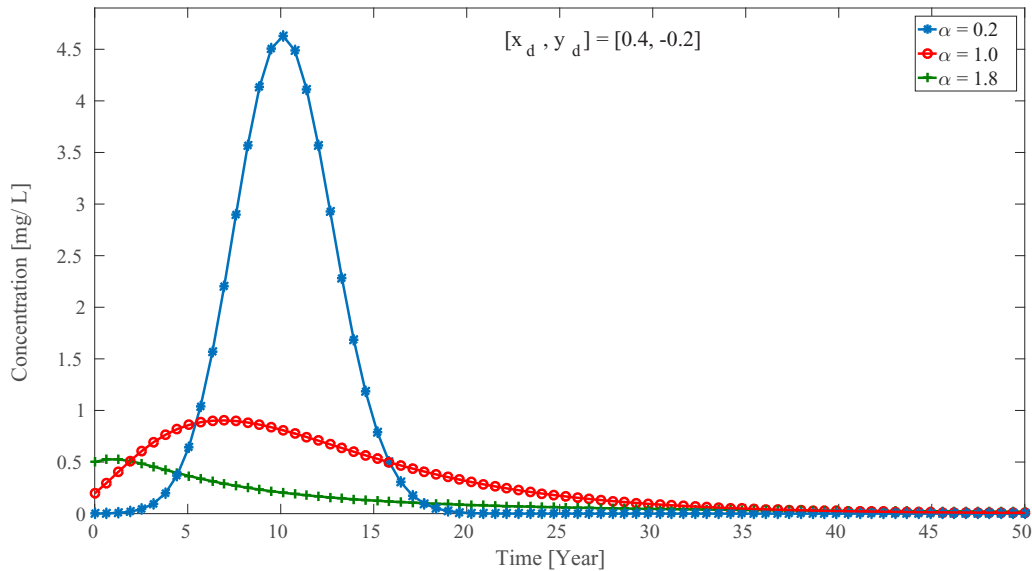
FIG. 2. The concentrations recorded by a detector for three different types of diffusion as denoted by the values of the diffusion exponent $\alpha$. The shown time series corresponds to concentration at the circled detector in the top rightmost panel of Fig. 1 at position $(x_d, y_d) = (0.4 \text{ km}, -0.2 \text{ km})$. All three sources ($S_1$, $S_2$, and $S_3$ marked by red diamonds in Fig. 1) release at $t_0 = -5$ years.

i.e., superdiffusion, subdiffusion, or normal diffusion, which is estimated by the value of the diffusion exponent $\alpha$ (see Sec. III A); (2) the generalized diffusion coefficients $D_x$ and $D_y$, which for the general anisotropic case, can be different in the two ($x$ and $y$) directions; and (3) the drift velocity $\vec{u}$, which for transport in porous media, affects the diffusion.

To illustrate the complexity of the problem, we present in Fig. 1 temporal snapshots of concentration profiles for subdiffusive ($\alpha = 0.2$), normal diffusive ($\alpha = 1$), and superdiffusive ($\alpha = 1.8$) transport (to generate the synthetic data, we used parameters that are typical of the transport and dispersion of contaminants in an aquifer [29]). In each case, emission originates instantaneously at $t_0 = -5$ years from three point sources $S_1$, $S_2$, and $S_3$, marked by red diamonds. In all three cases, the flow is subjected to a drift velocity $\mathbf{u} = (u_x, 0)$, where $u_x = 0.05$ km/year in the $x$ direction. In the subdiffusive case, the contribution from each source remains relatively distinct throughout the illustrated time period of 20 years. However, for the normal diffusive and superdiffusive cases, the contributions from the three sources increasingly mix and become indistinguishable. In reality, complete spatiotemporal concentration profiles as shown in Fig. 1 are unavailable. Concentration information is only available at a limited number of locations, where the detectors are positioned, shown as black points in Fig. 1. Examples of concentration times series from a single detector located at $(x_d, y_d) = (0.4 \text{ km}, -0.2 \text{ km})$ (marked by a red circle in the upper right panel of Fig. 1) are given in Fig. 2 for the three different types of diffusion discussed in Fig. 1. Again, as a result of the slow diffusion in the subdiffusive case $\alpha = 0.2$, the signal arrives at the detector in the form of a narrow and concentrated peak. However, for normal diffusion $\alpha = 1$ and superdiffusion $\alpha = 1.8$, the signal is less concentrated and spread over a much longer time period as a result of the faster dispersion processes and the mixing of the source contributions. The extension of hNMF, we report here, correctly estimates the unknown

number of point sources and evaluates the sources' locations and strengths, as well as the medium properties (diffusion exponent, anisotropic generalized diffusion coefficients, and drift), from a limited number of time series similar to those shown in Fig. 2. More complicated scenarios, such as temporally varying emission rates and spatially extended sources, were also considered.

## III. METHODS

### A. Fractional Brownian motion Green's function

There are various mechanisms that lead to anomalous diffusion [1,2]. The simplest mechanism is fractional Brownian motion (fBm). fBm relies on a memory effect that can be introduced into the Langevin equation describing the dynamics of a particle [30]. The mechanism was suggested to be relevant to many processes [6]. Despite the memory effects, fBm is a Gaussian process, in which, unlike classical Brownian motion, the increments of the fBm are not independent. The two-dimensional (2D) Fokker-Planck equation describing Galilei invariant fBM is

$$\frac{\partial C}{\partial t} = \alpha t^{\alpha-1}\left(D_x \frac{\partial^2 C}{\partial x^2} + D_y \frac{\partial^2 C}{\partial y^2}\right) - u_x \frac{\partial C}{\partial x} + Q, \quad (1)$$

where $C$ is the concentration, $\alpha$ is the diffusion exponent ($0 < \alpha < 2$), $D_x$, $D_y$ are the generalized diffusion coefficients in the $x$ and $y$ directions, respectively, $u_x$ is the drift velocity in the $x$ direction (the coordinate system is chosen such that $u_y \equiv 0$), and $Q(\vec{r}, t)$ is the source function.

The Green's function (propagator) of the above Fokker-Planck equation, with a physical boundary condition requesting the concentration to vanish at an infinite distance from the source, is given by (see [31] for the 1D propagator)

$$G_{\text{fBm}}(\vec{r}, t) = \frac{1}{4\pi \sqrt{D_x D_y} t^\alpha} e^{-\frac{(x-u_x t)^2}{4D_x t^\alpha}} e^{-\frac{y^2}{4D_y t^\alpha}}. \quad (2)$$

For $\tilde{N}_s$ instantaneous point sources [i.e., $Q(\vec{r}, t) = \sum_{s=1}^{\tilde{N}_s} q_s \delta(\vec{r} - \vec{r}_s) \delta(t - t_s)$], the concentration is given by the sum of the contributions of the different sources [the contribution of each point source is give by Eq. (2) multiplied by the source strength $q_s$]:

$$C(\vec{r}, t) = \sum_{s=1}^{\tilde{N}_s} \frac{q_s}{4\pi\sqrt{D_x D_y}(t - t_s)^\alpha}$$
$$\times e^{-\frac{[x - u_x(t - t_s) - x_s]^2}{4D_x(t - t_s)^\alpha}} e^{-\frac{(y - y_s)^2}{4D_y(t - t_s)^\alpha}}. \quad (3)$$

For release sources whose emission is finite in duration, the source function $Q_T(\vec{r}, t)$ takes the form

$$Q_T(\vec{r}, t) = \sum_{s=1}^{\tilde{N}_s} q_s \delta(\vec{r} - \vec{r}_s) \Theta(t - t_s^{on}) \Theta(t_s^f - t), \quad (4)$$

where $\Theta(t) = 0$, if $t < 0$, and $\Theta(t) = 1$, if $t \geqslant 0$, while $t_s^{on}, t_s^f$ are the beginning and end times of emission by the *s*th source, respectively. In this case, for $t > t_s^f > t_s^{on}$ for $s \in [1, \tilde{N}_s]$, the

concentration is given by

$$C_T(\vec{r}, t) = \int_{-\infty}^{t} G_{fBm}(\vec{r} - \vec{r}', t - t') Q_T(\vec{r}', t') dt'$$
$$= \sum_{s=1}^{\tilde{N}_s} q_s \int_{t_s^{on}}^{t_s^f} G_{fBm}(\vec{r} - \vec{r}_s, t - t') dt'. \quad (5)$$

Here, the limits of integration in Eq. (5) arise from the finite emission duration of the sources.

For spatially extended rectangular sources, the source function $Q_S(\vec{r}, t)$ takes the form

$$Q_S(\vec{r}, t) = \sum_{s=1}^{\tilde{N}_s} q_s \Theta(x - x_s^l) \Theta(x_s^r - x)$$
$$\times \Theta(y - y_s^l) \Theta(y_s^r - y) \delta(t - t_s). \quad (6)$$

Here, $x_s^l$, $x_s^r$ are the left and right boundaries of the *s*th source and similarly for the *y* coordinate to define rectangular sources. $t_s$ is time at which the *s*th source emitted. The concentration in this case is given by

$$C(\vec{r}, t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G_{fBm}(\vec{r} - \vec{r}', t) Q_S(\vec{r}', t) d\vec{r}' = \sum_{s=1}^{\tilde{N}_s} q_s \int_{x_s^l}^{x_s^r} \int_{y_s^l}^{y_s^r} dx' dy' G_{fBm}(\vec{r} - \vec{r}', t - t_i)$$
$$= \sum_{s=1}^{\tilde{N}_s} \frac{q_s}{4} \left[ \text{erf}\left(\frac{u_x(t - t_s) + x_s^r - x}{\sqrt{4D_x(t - t_s)^\alpha}}\right) - \text{erf}\left(\frac{u_x(t - t_s) + x_s^l - x}{\sqrt{4D_x(t - t_s)^\alpha}}\right) \right] \left[ \text{erf}\left(\frac{y_s^r - y}{\sqrt{4D_y(t - t_s)^\alpha}}\right) - \text{erf}\left(\frac{y_s^l - y}{\sqrt{4D_y(t - t_s)^\alpha}}\right) \right]. \quad (7)$$

### B. Hybrid non-negative matrix factorization (hNMF) method

The hybrid non-negative matrix factorization (hNMF) method reported in Ref. [29] combines the following: (i) Green's function of the Fokker-Planck equation (FPe) that describes normal diffusion transport with (ii) a nonlinear iterative minimization procedure and (iii) a customized clustering, introduced previously as a method for estimating the number of the latent features in NMF [32]. It was shown that the hNMF can successfully identify the unknown number of release sources, as well as the parameters of the FPe. To do this, the hNMF's algorithm explores the space of plausible solutions and narrows the set of possibilities by estimating the optimal number of release sources needed to reconstruct the observed data in a robust manner. Here, we generalize the hNMF to be adequate for fBm processes and show that the extended method can accurately determine the number of release sources and their locations and strengths, as well as the diffusion exponent and the transport properties of the medium ($D_x, D_y, u_x$). We also demonstrate the importance of this extension for the correct identification of the unknown number of release sources. In what follows, we assume point sources with the Green's function given by Eq. (2); for the other source types, it has to be replaced by the appropriate Green's function.

#### 1. Non-negative matrix factorization (NMF)

The usual interpretation of NMF is as a method for a low-rank matrix approximation of the observed data matrix

$C$, whose size is $T \times n$, by two unknown matrices $W$ and $H$, $C \approx WH$, both containing one small dimension $N_s$ (which is the estimate for the actual number of sources $\tilde{N}_s$). This approximation (that is, matrix decomposition) is performed through nonconvex minimization with a given distance metric $|| \ldots ||_{dist}$: $\min ||C_{ij} - \sum_{s=1}^{N_s} W_{is} H_{sj}||_{dist}$ constrained by the non-negativity of $W$ and $H$, $W_{is} \geqslant 0$; $H_{sj} \geqslant 0$. NMF has proven very useful for face recognition, text recognition, dimension reduction, unsupervised learning, anomaly detection, blind source separation (BSS), etc. [33]. NMF is underpinned by a well-defined statistical model of superimposed components that, when the distance metrics $|| \ldots ||_{dist}$ is the Euclidean distance, can be treated as a Gaussian mixture model [34]. In this probabilistic interpretation, NMF is equivalent to the expectation-maximization (EM) algorithm [35]. EM is developed to find the maximum likelihood estimates of parameters in statistical models, when the model depends on unobserved, i.e., latent or hidden, variables. In this probabilistic interpretation of NMF, the observables $c_1, c_2, \ldots, c_n$ ($c_i$ is a column vector with $T$ elements), which are the columns of the data, $C$, are generated by $N_s$ latent variables $h_1, h_2, \ldots, h_{N_s}$. Specifically, each observable $c_i$ is generated from a probability distribution with mean, $\langle c_i \rangle = \sum_{s=1}^{N_s} W_{is} h_s$, where $N_s$ is the (unknown) number of latent variables. The influence of $h_s$ on $c_i$ is through the basis patterns and features of the considered phenomenon $w_{:s}$ represented by the columns of $W$ [36]. It is known that the probabilistic interpretation of NMF is particularly valuable when dealing with stochastic signals [34].

In our case, the observed data $C$ are formed by the mixing of $\tilde{N}_s$ signals, at the locations of each one of the $n$ detectors, $\vec{r}_d \equiv (x_d, y_d)$, $d = 1, \ldots, n$, at the times of the records $t_m$, $m = 1, 2, \ldots, T$. Each one of the $n$ detectors is situated at a different location, $\vec{r}_d = (x_d, y_d)$, and measures a mixture of the signals originating from $\tilde{N}_s$ point sources located at the positions $\vec{r}_s = (x_s, y_s)$, $s = 1, 2, \ldots, \tilde{N}_s$, with respective strengths $q_s$. The fractional diffusion, characterized by the diffusion exponent $\alpha$, occurs in a medium with generalized diffusion coefficient $(D_x, D_y)$ and drift $u_x$. The transient signal recorded by the $d$th detector is assumed to be generated from a normal probability distribution with mean

$$\langle c_{\vec{r}_d, t_m} \rangle = \sum_{s=1}^{N_s} W_{s,(\vec{r}_d, t_m)} H_s, \tag{8}$$

where the index $\vec{r}_d$ denotes the position of the recording detector, and the index $t_m$ denotes the time point when the observed data have been recorded. Further, $W_{s,(\vec{r}_d, t_m)}$ are the fBm Green's functions [Eq. (2)] of each source. $W_{s,(\vec{r}_d, t_m)}$ in Eq. (8) can be considered as a specific kernel [37] that depends on space-time. $H$ contains $N_s$ hidden variables, independent of time and space, that generate the observables $c(\vec{r}_d, t_m)$ at points $(\vec{r}_d, t_m)$ via the weights $W_{s,(\vec{r}_d, t_m)}$. Note that both $W_{s,(\vec{r}_d, t_m)}$ and $H_s \equiv q_s$ are non-negative.

For given detectors' locations, time points, and the functional form of the fBm Green's function [Eq. (2)], our aim is to estimate the (unknown ) number of sources $\tilde{N}_s$ and to determine the source parameters $x_s$, $y_s$, $q_s$ ($s \in [1, \tilde{N}_s]$) and the medium transport characteristics $u_x$, $D_x$, $D_y$, as well as the diffusion exponent $\alpha$. In this interpretation and from Eq. (8), the minimization can be performed by nonlinear least squares on the objective function $O$,

$$O = \sum_{d=1}^{n} \sum_{m=1}^{T} \left( C_{\vec{r}_d, t_m} - \sum_{s=1}^{N_s} W_{s,(\vec{r}_d, t_m)} H_s \right)^2, \tag{9}$$

where $d$ marks the detectors, and $m$ marks the time points of the records.

The minimization of the objective function $O$ assumes that each measurement at a given space-time point $(\vec{r}_d, t_m)$ is an independent Gaussian-distributed random variable. If each detector has its own distinct (possibly time-dependent) measurement error, the objective function in (9) should be replaced by a weighted sum of least squares, where each deviation from the corresponding observation is weighted by the inverse square of its measurement error.

### 2. Custom clustering and NMFk

NMF is sufficient to carry a constrained optimization problem to extract desired parameters when the number of sources $\tilde{N}_s$ is known; however, we rely on the signals measured by the detectors, which record mixtures arising from an unknown number of sources. To determine this unknown number of sources (that is, the number of the latent features), we utilize an approach called NMFk, introduced in earlier works [29,32,38].

NMFk explores the possible number of sources $\mathcal{N}_s$, starting from $\mathcal{N}_s = 1, 2, \ldots, P$ [$P$ is less than $\min(n, T)$]. For each explored number of sources $\mathcal{N}_s$, a set $U_{\mathcal{N}_s}$ (we call it a run)

of $M \sim 200$ minimizations are computed, each minimization generated from random initial guesses for the unknown parameters (within specific physical bounds). The set of $M$ solutions in the set $U_{\mathcal{N}_s}$, for $\mathcal{N}_s$ sources, is given by

$$\begin{aligned} U_{\mathcal{N}_s} = ([(X, Y, Q)_1, (u_x, D_x, D_y, \alpha)_1], \\ \ldots, [(X, Y, Q)_M, (u_x, D_x, D_y, \alpha)_M]), \end{aligned} \tag{10}$$

where $(X, Y, Q)_i$ denotes the unknown coordinates and strengths of the $\mathcal{N}_s$ sources; $[(x_1, y_1, q_1), \ldots, (x_{\mathcal{N}_s}, y_{N_s}, q_{\mathcal{N}_s})]_i$ in the $i$th NMF minimization with advection velocity, generalized diffusion coefficients and diffusion exponent $(u_x, D_x, D_y, \alpha)_i$. After we build the set of $M$ NMF solutions (for $\mathcal{N}_s$ sources), the parameters pertaining to sources $[(x_1, y_1, q_1), \ldots, (x_{\mathcal{N}_s}, y_{\mathcal{N}_s}, q_{\mathcal{N}_s})]_i$ are subjected to customized clustering into $\mathcal{N}_s$ number of clusters. For a single NMF run with $\mathcal{N}_s$ sources, a total of $M \sim 200$ simulations are carried out, which gives $M$ tuples $[(X, Y, Q)_i, (u_x, D_x, D_y, \alpha)_i]$. Each of the tuples $[(X, Y, Q)_i, (u_x, D_x, D_y, \alpha)_i]$ represents a distinct solution for nominally equivalent NMF minimizations, where the difference arises as a result of the random initial guesses. Next, we perform customized clustering, assigning the parameters of each $W_{s,(\vec{r}_d, t_m)}$ of all $M$ solutions to one of $\mathcal{N}_s$ clusters. This customized clustering is similar to $k$-means clustering but with an additional constraint which constrains the number of elements in each of the clusters to be equal to the number of solutions $M$. For example, with $M = 200$, each one of the $\mathcal{N}_s$ identified clusters must contain exactly 200 solutions. This condition has to be enforced since each NMF minimization (producing a given $[(X, Y, Q)_i, (u_x, D_x, D_y, \alpha)_i]$ tuple) contributes only one solution, and accordingly has to supply exactly one element to each cluster. During the clustering, similarity between the elements is measured by cosine similarity, which calculates the cosine of the angle between two vectors and is a natural choice for similarity between non-negative vectors [39].

Further, to identify the optimum number of sources, NMFk calculates the clusters' stability for each explored number of sources $\mathcal{N}_s$. The optimum number of sources $N_s$, which is used to estimate the true $\tilde{N}_s$, is the number of sources for which the corresponding clusters are relatively stable and separable and whose centroids result in a small reconstruction error (see below).

To quantify the stability and separability of the clustering for a given number of sources, NMFk utilizes the silhouette statistics $S$ [40], which is developed to measure the similarity between an element and the elements of its own cluster compared to the centroids of other clusters. The $S$ values are between $[-1, 1]$, and $S$ measures how well each element has been classified by the clustering. The main idea of NMFk is to use the cohesion of the clusters (how compact they are) and the separation between them as a measure of the stability of the solutions of the minimization with different random initial guesses and hence the quality of a particular choice of $\mathcal{N}_s$. In previous works, NMFk even "shuffled" randomly the observed data $C$ to increase the effect of robustness of the average solutions [41–43].

The intuitive reasoning behind the idea of stability is as follows. In the case of underfitting, i.e., for solutions with a
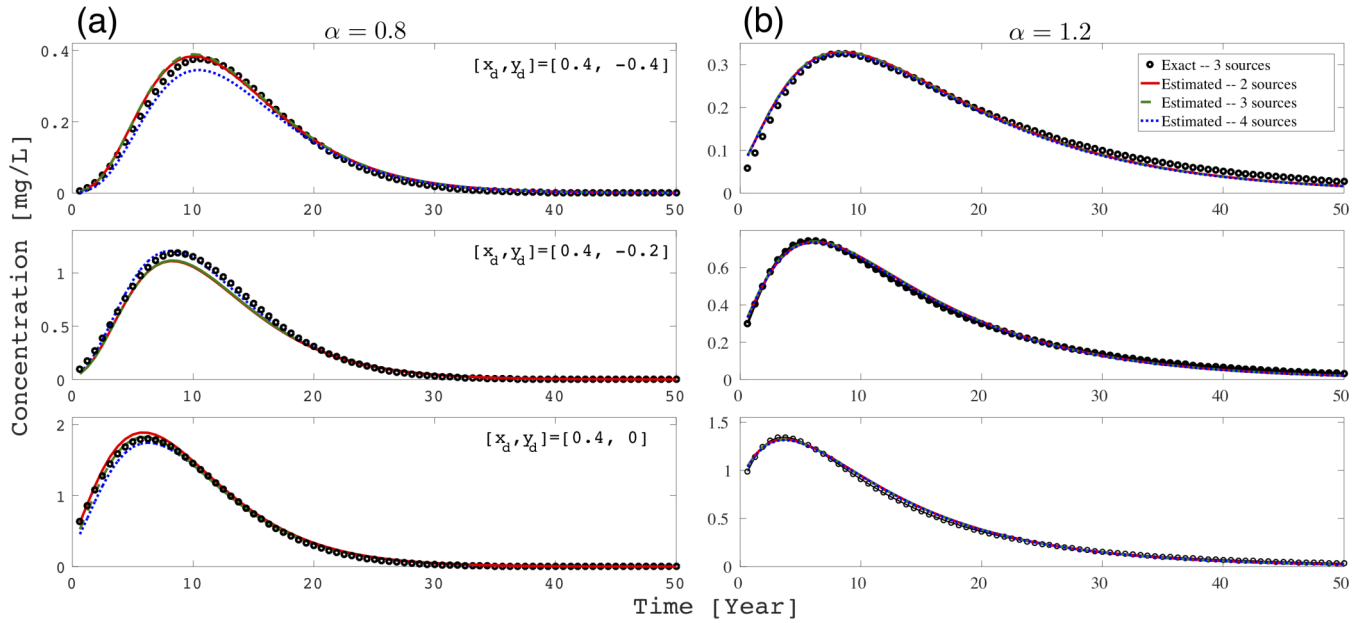
FIG. 3. Measured (black lines) and estimated signals (green, red, and blue markers) at three detectors. The left column (a) shows the reconstructed signal for two, three, or four sources and $\alpha = 0.8$, and the right column (b) shows the same information for $\alpha = 1.2$. The different rows correspond to different detectors whose coordinates are specified. The estimated signals are all for normal diffusion $\alpha = 1$.

number of sources less than the actual number of sources, the clustering could be good; for example, several of the sources could be combined to produce a "supercluster." However, the clustering will break down significantly in the case of overfitting, when the estimated number of sources exceeds the true number of sources. Indeed, in this case, we do not expect the solutions to be well clustered since there is no unique way to reconstruct the solutions with number of clusters $> \tilde{N}_s$, and at least some of the clusters will be artificial, rather than real entities.

The other metrics that NMFk utilizes is the relative reconstruction error $R = ||C - WH||/||C||$, which measures the relative deviation of the obtained solution $WH$ from the original data $C$. The reconstruction error $R$ evaluates the accuracy with which the average solution, that is, the solution constructed with the parameters taken from the centroids of the clusters, reproduce the observed data $C$. In general, the solution accuracy increases (while the stability of the clusters decreases) with the increase in the number of the sources (increasing the number of the parameters in the minimization).

In summary, NMFk calculates the average silhouette width $S$ and the average reconstruction error $R$ for each choice of the unknown number of sources, in order to estimate the true number of release sources $\tilde{N}_s$. NMFk determines $\tilde{N}_s$ to be equal to the number of sources that accurately reconstruct the observations (i.e., their relative reconstruction error $R$ is small enough) and the clustering of the sets of solutions corresponding to $N_s$, obtained with random initial conditions, to be sufficiently robust (i.e., the average silhouette width $S$ to be close to 1).

## IV. GENERATION OF SYNTHETIC DATA

In order to validate the generalization of the hNMF method, we generated synthetic data sets using Eq. (3)

for different types of diffusion (dictated by the values of $\alpha$). The parameters that were used here to generate the synthetic data are typical for groundwater contamination [29]. The parameters correspond to the case of normal diffusion, and for the case of anomalous diffusion, the values of the Brownian dispersion coefficients were used as generalized diffusion coefficients. We used $D_x = 0.005$ km$^2$/year$^\alpha$ and $D_y = 0.00125$ km$^2$/year$^\alpha$, $t_0 = -5$ years and $\alpha \in [0, 2]$ ($\alpha < 1$ corresponds to subdiffusion and $\alpha > 1$ corresponds to superdiffusion). The choices of source and detector positions that were used for the results in Figs. 3 and 4 are shown in Fig. 1. In Fig. 5, two different configurations are shown. Each synthetic data set contains 30 time series for 0 years $< t <$ 50 years with equal sampling intervals such that each series includes a total of 80 concentration values (see Figs. 2 and 5 for examples).

## V. NUMERICAL EXPERIMENTS

### A. Generalized hNMF is needed for the accurate identification of fBm sources

In order to illustrate the importance of the generalized hNMF method, we considered synthetic data generated using $\alpha = 0.8$ (subdiffusion) and $\alpha = 1.2$ (superdiffusion), the parameters listed in Sec. IV, and the configuration of sources and detectors shown in Fig. 1. Both values of the diffusion exponent are close to normal diffusion $\alpha = 1$. We ran the hNMF with $\alpha$ set to 1 (i.e., removing the degree of freedom of the diffusion exponent). The reconstructed signals at the locations of three different detectors are depicted in Fig. 3. The synthetic data corresponding to $\alpha = 0.8$ and 1.2 are presented by solid black lines in the left (right) panels, and the circles represent the reconstructed signals at these locations using 2, 3, or 4 point sources. Figure 3 demonstrates that in all
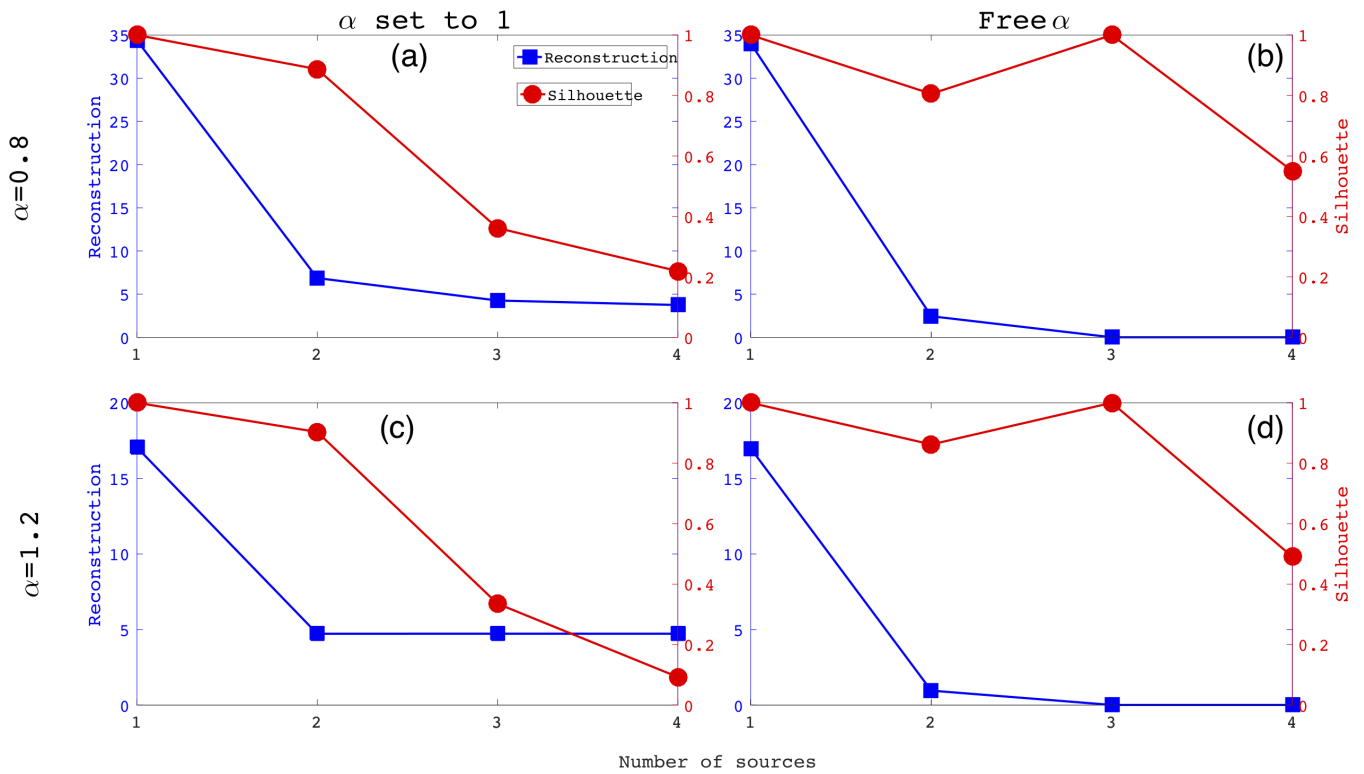
FIG. 4. The reconstruction and silhouette measures for different numbers of sources. The upper row (a), (b) corresponds to $\alpha = 0.8$ and the lower row (c), (d) to $\alpha = 1.2$. The left column corresponds to the method with $\alpha$ set equal to one (i.e., looking for the best normal diffusion parameters to describe the measurements), and the right column corresponds to the full method with $\alpha$ being extracted by the method.

cases, the detector observations are well approximated despite the wrong propagator used and the wrong number of sources. It is, therefore, clear that the ability to approximate detector observations alone is not sufficient to determine the number of sources and their locations or the diffusion characteristics.

If the $\alpha$ is not forced, we see slightly better reconstructions as generalized hNMF recognizes the correct number of sources with the appropriate diffusion characteristics.

In Fig. 4, we present the reconstruction metrics $R$ and average silhouette $S$ for the different numbers of sources ($\mathcal{N}_s$)
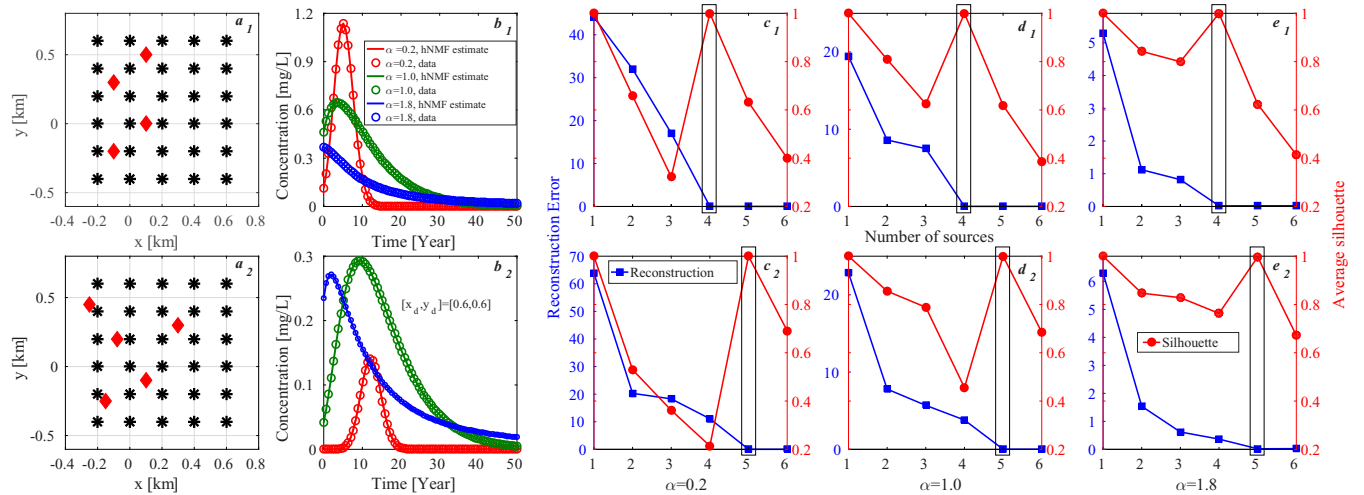


FIG. 5. Generalized hNMF benchmarks for two spatial configurations with four (upper panel 1) and five (bottom panel 2) release sources. Subfigures, noted by (a$_i$), (b$_i$), (c$_i$), and (e$_i$), $i = 1, 2$, represent (a$_i$) spatial configuration of sources and detectors, (b$_i$) time series of the concentration measured by a detector at $\vec{r}_d = (0.6, 0.6)$ with the reconstructed signal; solid line represents recorded signals, and reconstructed signals are plotted with red markers ($\alpha = 0.2$, subdiffusion), green markers ($\alpha = 1.0$, normal diffusion), and blue markers ($\alpha = 1.8$, superdiffusion). Subfigures (c$_i$), (d$_i$), and (e$_i$) represent the reconstruction error $R$ (left, blue) and the average silhouette width $S$ (right, red) of the solutions; the shadowed box depicts the true number of sources $\tilde{N}_s$, which is equal to the optimal number of sources determined by the hNMF.
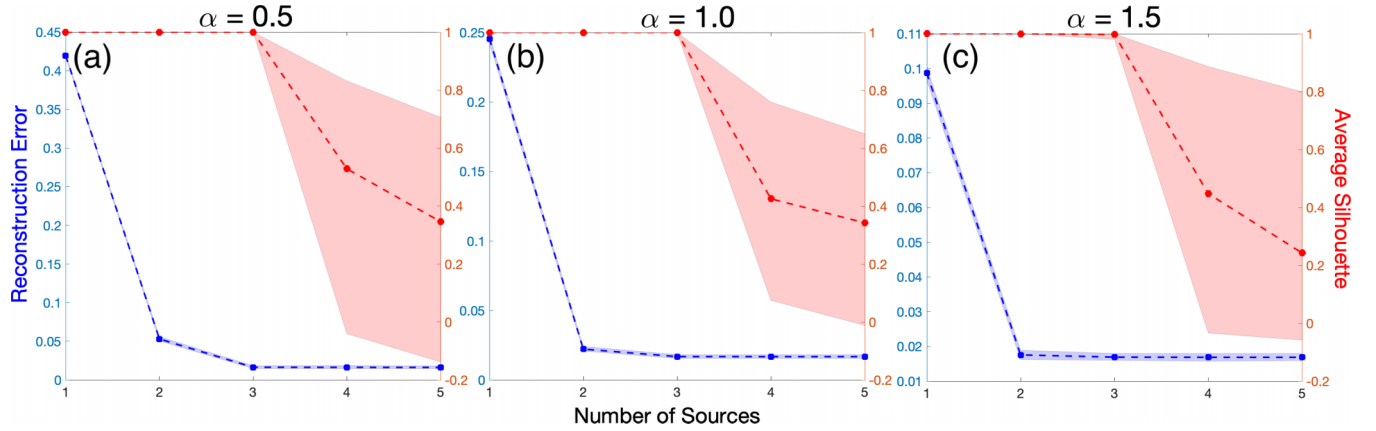
FIG. 6. Reconstruction error and average silhouette for noisy data. The three panels correspond to the denoted values of $\alpha$. The configuration is the same as shown in Fig. 1 and the noise amplitude is $\mu = 0.03$ [see Eq. (11)]. The shaded area represents the range of values obtained from 50 realizations of the noise.

for $\alpha = 0.8$ (subdiffusion) (upper row) and $\alpha = 1.2$ (superdiffusion) (lower row). The left column corresponds to a forced $\alpha = 1$, that is, to the old hNMF. The right column corresponds to the results of the generalized hNMF with the diffusion exponent derived by the minimization described in Sec. III B. From Fig. 4, it can be concluded that with $\alpha$ set to one (left column), both the reconstruction $R$ and the average silhouette $S$ decrease as the number of sources increases, for both subdiffusion and superdiffusion. This observation implies that, despite the better fit of the data by the estimated parameters, the increase in the number of sources worsens the clustering quality (a reduced silhouette). The reduction of the average silhouette stems from the fact that the old hNMF model used to describe the data does not account for anomalous diffusion. In the right column, we show that when the generalized hNMF is used, with the diffusion exponent extracted by the minimization, for both subdiffusion and superdiffusion, the maximum value of the silhouette and the minimal value of the reconstruction are obtained for the correct number of sources (three) that was used to generate the synthetic data.

### B. Generalized hNMF accurately identifies fBm sources and properties of the medium

We applied the generalized hNMF method to different configurations of point sources and detectors and found that the method provides a correct estimate of the number of sources, their strengths and positions, and medium characteristics (drift velocity, generalized diffusion coefficients, and the diffusion exponent). Figure 5 demonstrates the results of our extended hNMF method, for two different numbers of release point sources and spatial configurations, and for three values of the diffusion exponent corresponding to subdiffusion, normal diffusion, and superdiffusion. Figures 5(b$_1$) and 5(b$_2$) show the reconstruction of the original signal measured by the detector located at (0.6,0.6). The solid line represents the original signal recorded by the detector, and the markers are the solution estimates by the extended hNMF method. It is clear that the signals are well reconstructed for all cases of the diffusion exponent $\alpha$. Figures 5(c)–5(e) demonstrate that the extended hNMF identified the correct number of release sources.

Figures 5(c)–5(e) of the two panels show the average reconstruction errors (left, blue) and average silhouette widths (right, red) used to determine the unknown number of release sources present in the system. It is clear from these figures that our extended method correctly estimates the number of sources, their characteristics, and the transport characteristics for all the cases of the diffusion exponent considered here.

We also tested the performance of the method with configurations consisting of temporally extended and spatially extended sources. In all these cases, the generalized hNMF method provides accurate estimates of the duration of emission and the spatial extension of the sources. We also verified that for the synthetic data that was generated using point sources, when the method was implemented using the propagator for spatially extended sources, the estimated size of the sources was smaller than any spatial scale of the configuration (the dispersion radius between sequential measurements, the distance between detectors, the distance between sources, and the drift distance between sequential measurements).

### C. Noisy data

Real data often include noise. The noise can be due to the measuring device(s) or due to other effects which are not accounted for in the Green's function describing the process. In many realistic cases, the noise is proportional to the signal. In order to test the applicability and robustness of the hNMF, we considered noisy data. The data were generated in the same way described in Sec. IV for the configuration shown in Fig. 1. In order to add the noise, the precisely calculated concentration (at each time and position) was multiplied by a random number drawn from a uniform distribution according to

$$\mathcal{C}(\vec{r}, t) = C(\vec{r}, t)[1 + \mu z(\vec{r}, t)]. \tag{11}$$

The probability density function of $z(\vec{r}, t)$ is simply

$$p(z(\vec{r}, t)) = \begin{cases} 1/2, & -1 \leqslant z \leqslant 1 \\ 0, & \text{else.} \end{cases} \tag{12}$$

In Fig. 6, we present the reconstruction error and the average silhouette for different numbers of sources. The three
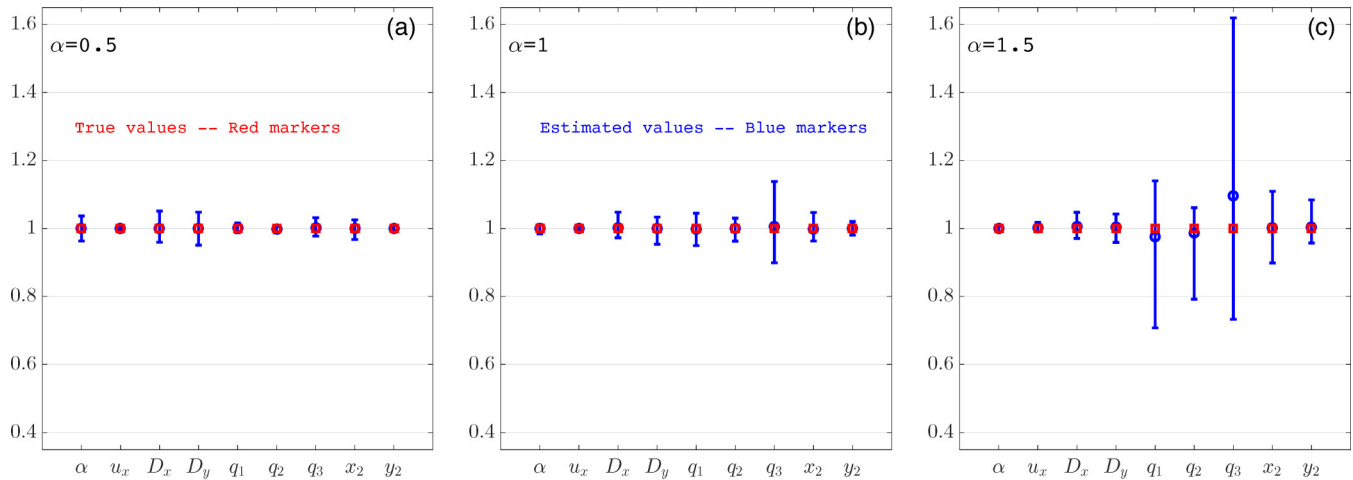
FIG. 7. The normalized hNMF estimated parameters and the associated uncertainties. The three panels correspond to the denoted values of $\alpha$. The configuration is the same as shown in Fig. 1 and the noise amplitude is $\mu = 0.03$ [see Eq. (11)]. The error bars represent the range of values obtained from 50 realizations of the noise.

panels correspond to the specified values of the diffusion exponent $\alpha$ (corresponding to subdiffusion, normal diffusion, and superdiffusion).

For the cases of subdiffusion ($\alpha = 0.5$) and normal diffusion ($\alpha = 1$), the reconstruction error and the average silhouette show a similar behavior to that found for data in the absence of noise, and the correct number of sources is identified. For the superdiffusive case, where the mixing of the signals from the three sources is stronger, for some realizations of the noise, the average silhouette does not show the same sharp decrease, and the identification of the number of sources is more challenging. For lower values of the noise, the hNMF shows the same performance as for the exact data.

The noise also introduces uncertainty in the estimated values of the parameters. In Fig. 7, we present the estimated values of the parameters for the same diffusion exponents shown in Fig. 6. Each parameter was normalized by the true value in order to present the relative error and uncertainty. The uncertainty range is larger for larger diffusion exponents where the mixing of the signals from the different sources is larger. For the normal diffusion and the superdiffusion, the uncertainty in the estimation of the amplitude of the third source is the largest. The noise is proportional to the total signal, and the third source has the smallest amplitude. Therefore, the noise relative to the signal from this source is much larger and it affects the uncertainty associated with its estimated amplitude.

## VI. CONCLUSIONS

Inverse modeling of anomalous diffusion is of great interest in many fields, including material science, physics, finances, biology, and many others. The value of an inverse modeling approach lies in its ability to correctly estimate the unknown number of release sources and their characteristics, as well

the medium characteristics and the diffusion exponent. We demonstrated that the generalized hNMF introduced here, which is based on the integration of unsupervised learning and Green's function of the fBm governing equation, is capable of providing accurate estimates of the number of sources, their properties, and the medium characteristics that result in subdiffusion, Brownian diffusion, or superdiffusion. The method was shown to work with different types of sources, including point sources, temporally extended sources, and spatially extended ones with various configurations. The generalized hNMF method also provides information regarding the validity of the type of the propagator that is used. If the propagator is not appropriate (e.g., if it is the propagator of normal diffusion), the method showed that it is impossible to have maximal silhouette and minimal reconstruction metrics for any number of sources. When the adequate propagator is used, the optima that were found corresponded to the configurations that were used to generate the synthetic data. The results presented here and previous successful applications of the hNMF for inverse problems suggest that this method might be further extended to account for different mechanisms of anomalous diffusion (such as the continuous-time random walk and quenched disorder). Of particular interest are generalizations of the hNMF that would enable the extraction of the scaling relations based on limited observations.

[1] J.-P. Bouchaud and A. Georges, Anomalous diffusion in disordered media: Statistical mechanisms, models and physical applications, Phys. Rep. **195**, 127 (1990).

[2] R. Metzler and J. Klafter, The random walk's guide to anomalous diffusion: a fractional dynamics approach, Phys. Rep. **339**, 1 (2000).

[3] E. Barkai and Y.-C. Cheng, Aging continuous time random walks, J. Chem. Phys. **118**, 6167 (2003).

[4] G. Bel and E. Barkai, Weak Ergodicity Breaking in the Continuous-Time Random Walk, Phys. Rev. Lett. **94**, 240602 (2005).

[5] J. L. McCauley, G. H. Gunaratne, and K. E. Bassler, Hurst exponents, markov processes, and fractional brownian motion, Phys. A (Amsterdam) **379**, 1 (2007).

[6] F. Biagini, Y. Hu, B. Øksendal, and T. Zhang, *Stochastic Calculus for Fractional Brownian Motion and Applications* (Springer, London, 2008).

[7] G. R. Kneller, K. Baczynski, and M. Pasenkiewicz-Gierula, Communication: Consistent picture of lateral subdiffusion in lipid bilayers: Molecular dynamics simulation and exact results, J. Chem. Phys. **135**, 141105 (2011).

[8] E. Lucioni, B. Deissler, L. Tanzi, G. Roati, M. Zaccanti, M. Modugno, M. Larcher, F. Dalfovo, M. Inguscio, and G. Modugno, Observation of Subdiffusion in a Disordered Interacting System, Phys. Rev. Lett. **106**, 230403 (2011).

[9] M. Larcher, T. V. Laptyeva, J. D. Bodyfelt, F. Dalfovo, M. Modugno, and S. Flach, Subdiffusion of nonlinear waves in quasiperiodic potentials, New J. Phys. **14**, 103036 (2012).

[10] K. Staliunas, Anticorrelations and subdiffusion in financial systems, Adv. Complex Syst. **6**, 251 (2003).

[11] A. Caspi, R. Granek, and M. Elbaum, Enhanced Diffusion in Active Intracellular Transport, Phys. Rev. Lett. **85**, 5655 (2000).

[12] E. Pereira, J. M. G. Martinho, and M. N. Berberan-Santos, Photon Trajectories in Incoherent Atomic Radiation Trapping as Lévy Flights, Phys. Rev. Lett. **93**, 120201 (2004).

[13] J. F. Reverey, J.-H. Jeon, H. Bao, M. Leippe, R. Metzler, and C. Selhuber-Unkel, Superdiffusion dominates intracellular particle motion in the supercrowded cytoplasm of pathogenic acanthamoeba castellanii, Sci. Rep. **5**, 11690 (2015).

[14] F. J. Molz and G. K. Boman, A fractal-based stochastic interpolation scheme in subsurface hydrology, Water Resour. Res. **29**, 3769 (1993).

[15] F. Molz, H. Liu, and J. Szulga, Fractional brownian motion and fractional gaussian noise in subsurface hydrology: A review, presentation of fundamental properties, and extensions, Water Resour. Res. **33**, 2273 (1997).

[16] A. Cortis and B. Berkowitz, Anomalous transport in classical soil and sand columns, Soil Sci. Soc. Am. J. **68**, 1539 (2004).

[17] J.-F. Coeurjolly, Estimating the parameters of a fractional brownian motion by discrete variations of its sample paths, Statist. Inference Stochastic Process. **4**, 199 (2001).

[18] J. Cheng, J. Nakagawa, M. Yamamoto, and T. Yamazaki, Uniqueness in an inverse problem for a one-dimensional fractional diffusion equation, Inverse Probl. **25**, 115002 (2009).

[19] H. Wei, W. Chen, H. Sun, and X. Li, A coupled method for inverse source problem of spatial fractional anomalous diffusion equations, Inverse Probl. Sci. Eng. **18**, 945 (2010).

[20] H. Yaozhong, N. David, X. Weilin, and Z. Weiguo, Exact maximum likelihood estimator for drift fractional brownian motion at discrete observation, Acta Math. Sci. **31**, 1851 (2011).

[21] K. M. Furati, O. S. Iyiola, and M. Kirane, An inverse problem for a generalized fractional diffusion, Appl. Math. Comput. **249**, 24 (2014).

[22] B. Jin and W. Rundell, A tutorial on inverse problems for anomalous diffusion processes, Inverse Probl. **31**, 035003 (2015).

[23] L. Ning, C. Westin, and Y. Rathi, Estimating diffusion propagator and its moments using directional radial basis functions, IEEE Trans. Med. Imaging **34**, 2058 (2015).

[24] J. Janno and N. Kinash, Reconstruction of an order of derivative and a source term in a fractional diffusion equation from final measurements, Inverse Probl. **34**, 025007 (2018).

[25] Y.-X. Zhang, J. Jia, and L. Yan, Bayesian approach to a nonlinear inverse problem for a time-space fractional diffusion equation, Inverse Probl. **34**, 125002 (2018).

[26] C. Cai, X. Cheng, W. Xiao, and X. Wu, Parameter identification for mixed fractional brownian motions with the drift parameter, Phys. A: Statistical Mechanics and its Applications **536**, 120942 (2019).

[27] C. Liu and W. P. Ball, Application of inverse methods to contaminant source identification from aquitard diffusion profiles at dover afb, delaware, Water Resour. Res. **35**, 1975 (1999).

[28] T. F. Zhang and Q. Chen, Identification of contaminant sources in enclosed environments by inverse cfd modeling, Indoor Air **17**, 167 (2007).

[29] V. G. Stanev, F. L. Iliev, S. Hansen, V. V. Vesselinov, and B. S. Alexandrov, Identification of release sources in advection–diffusion system by machine learning combined with green's function inverse method, Appl. Math. Modell. **60**, 64 (2018).

[30] B. B. Mandelbrot and J. W. Van Ness, Fractional brownian motions, fractional noises and applications, SIAM Rev. **10**, 422 (1968).

[31] R. Metzler and J. Klafter, The restaurant at the end of the random walk: recent developments in the description of anomalous transport by fractional dynamics, J. Phys. A: Math. Gen. **37**, R161 (2004).

[32] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale *et al.*, Signatures of mutational processes in human cancer, Nature (London) **500**, 415 (2013).

[33] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation* (Wiley, Hoboken, NJ, 2009).

[34] C. Févotte and A. T. Cemgil, Nonnegative matrix factorizations as probabilistic inference in composite models, in *2009 17th European Signal Processing Conference* (IEEE, Piscataway, NJ, 2009), pp. 1913–1917.

[35] A. P. Dempster, F. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. R. Stat. Soc.: Ser. B (Methodological) **39**, 1 (1977).

[36] D. D. Lee and H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature (London) **401**, 788 (1999).

[37] D. Zhang, Z.-H. Zhou, and S. Chen, Non-negative matrix factorization on kernels, in *Pacific Rim International Conference on Artificial Intelligence* (Springer, Berlin, 2006), pp. 404–412.

[38] B. S. Alexandrov and V. V. Vesselinov, Blind source separation for groundwater pressure analysis based on nonnegative matrix factorization, Water Resour. Res. **50**, 7332 (2014).

[39] T. Korenius, J. Laurikkala, and M. Juhola, On principal component analysis, cosine and euclidean measures in information retrieval, Inf. Sci. **177**, 4893 (2007).

[40] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. **20**, 53 (1987).

[41] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton, Deciphering signatures of mutational processes operative in human cancer, Cell Rep. **3**, 246 (2013).

[42] G. Chennupati, R. Vangara, E. Skau, H. Djidjev, and B. Alexandrov, Distributed non-negative matrix factorization with determination of the number of latent features, J. Supercomput. 1 (2020).

[43] V. Stanev, V. V. Vesselinov, A. G. Kusne, G. Antoszewski, I. Takeuchi, and B. S. Alexandrov, Unsupervised phase mapping of x-ray diffraction data by nonnegative matrix factorization integrated with custom clustering, npj Comput. Mater. **4**, 43 (2018).