Geosci. Model Dev., 12, 1525–1539, 2019
https://doi.org/10.5194/gmd-12-1525-2019

Geoscientific
Model Development
Open Access

Scalable diagnostics for global atmospheric chemistry using Ristretto library (version 1.0)

Meghana Velegar[1], **N. Benjamin Erichson**[1], **Christoph A. Keller**[2,3], **and J. Nathan Kutz**[1]

[1]Department of Applied Mathematics, University of Washington, Seattle, WA 98195, USA
[2]NASA Global Modeling and Assimilation Office, Goddard Space Flight Center, Greenbelt, MD 20771, USA
[3]Universities Space Research Association, Columbia, MD 21046, USA

Correspondence: J. Nathan Kutz (kutz@uw.edu)

Received: 4 December 2018 – Discussion started: 19 December 2018
Revised: 27 March 2019 – Accepted: 29 March 2019 – Published: 18 April 2019

Abstract. We introduce a new set of algorithmic tools capable of producing scalable, low-rank decompositions of global spatiotemporal atmospheric chemistry data. By exploiting emerging *randomized linear algebra* algorithms, a suite of decompositions are proposed that extract the dominant features from *big data* sets (i.e., global atmospheric chemistry at longitude, latitude, and elevation) with improved interpretability. Importantly, our proposed algorithms scale with the intrinsic rank of the global chemistry space rather than the ever increasing spatiotemporal measurement space, thus allowing for the efficient representation and compression of the data. In addition to scalability, two additional innovations are proposed for improved interpretability: (i) a nonnegative decomposition of the data for improved interpretability by constraining the chemical space to have only positive expression values (unlike PCA analysis); and (ii) sparse matrix decompositions, which threshold small weights to zero, thus highlighting the dominant, localized spatial activity (again unlike PCA analysis). Our methods are demonstrated on a full year of global chemistry dynamics data, showing the significant improvement in computational speed and interpretability. We show that the decomposition methods presented here successfully extract known major features of atmospheric chemistry, such as summertime surface pollution and biomass burning activities.

1 Introduction

Dimensionality reduction is a critically enabling aspect of machine learning and data science in the era of *big data*. Specifically, extracting the dominant low-rank features from a high-dimensional data matrix $\mathbf{X}$ allows one to efficiently perform tasks associated with clustering, classification, reconstruction, and prediction (forecasting). Commonly used *linear* dimensionality reduction methods are typically based upon *singular value decomposition* (SVD) which allows one to exploit covariances manifest in the data (Cunningham and Ghahramani, 2015). Thus, the analysis of big data, such as the atmospheric chemistry data considered here, relies on a variety of matrix decomposition methods which seek to exploit low-rank features exhibited by the high-dimensional data. Despite our ever-increasing computational power, the emergence of large-scale data sets has severely challenged our ability to analyze data using traditional matrix algorithms, especially for ever increasing refinements of computational models.

In this work, we are specifically concerned with time-series measurements of the concentration of chemical species collected from spatial locations in the atmosphere, illustrated in Fig. 1. On a global scale (longitude, latitude, and elevation), this data can be exceptionally high-dimensional so as to be not computationally tractable. Thus, computationally scalable methods are required for the analysis of atmospheric chemistry dynamics. Indeed, atmospheric chemistry is an exceptionally high-dimensional problem as it involves hundreds of chemical species that are coupled with each other via a set of ordinary differential equations. Models of atmospheric chemistry that are used to simulate the spatiotem-

Published by Copernicus Publications on behalf of the European Geosciences Union.

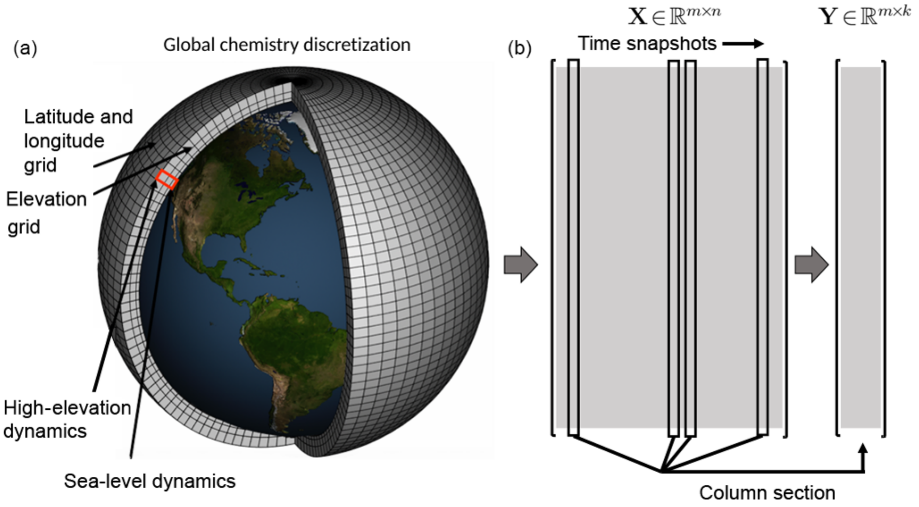
(a)
Global chemistry discretization
Latitude and longitude grid
Elevation grid
High-elevation dynamics
Sea-level dynamics
(b)
$\mathbf{X} \in \mathbb{R}^{m \times n}$
Time snapshots
$\mathbf{Y} \in \mathbb{R}^{m \times k}$
Column section

Figure 1. Atmospheric chemistry simulation on a global mesh with discretized longitude, latitude, and elevation (panel **a** modified from NOAA). Each illustrated grid cell contains time-series data for the atmospheric chemistry dynamics. Well resolved simulations generate massive data sets that are often not amenable to diagnostic analysis. Our proposed algorithms offer a scalable architecture for the analysis of global spatiotemporal data. As shown in the two right panels **(b)**, the original data matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, where m is the number of grid points and n is the number of snapshots, can be downsampled (here via random column sampling) to form the matrix $\mathbf{Y} \in \mathbb{R}^{m \times k}$, where $k \ll n$. Although random column selection is shown, we can also use a random measurement matrix to sample the data as shown in Sect. 3.

poral evolution of these chemical constituents need to keep track of each chemical species on a global scale (longitude, latitude, and elevation) and at each point in time. The resulting data sets – used for scientific analysis or required for subsequent restarts of the model – quickly become massive, especially as the horizontal model resolution steadily increases. For example, a single snapshot of the chemical state of an atmospheric chemistry model at 25 km × 25 km horizontal resolution requires 60 GB of storage space.

To tackle this challenge, we present a variety of emerging matrix decomposition methods that can be used for scalable diagnostics of global atmospheric chemistry dynamics. Specifically, we use randomized linear algebra methods (Halko et al., 2011; Mahoney, 2011; Drineas and Mahoney, 2016; Erichson et al., 2016, 2017a) to extract the dominant, low-rank mode structures from a full three-dimensional atmospheric chemistry data set. These methods are highly scalable and can thus be used on emerging big data sets describing global chemistry dynamics, providing a useful tool for scientific discovery and analysis. Furthermore, they offer an alternative approach for the storage of large-scale atmospheric chemistry data. Importantly, randomized methods are an efficient alternative to distributed computing if these computational resources are not available. For instance, Gittens et al. (2018) can compute the SVD of a 2.2 TB (terabyte) data set in about 60 s, given a supercomputer with many nodes. However, if supercomputing is not available, randomized methods offer an attractive alternative which does not require expensive computation hours on a cluster.

The paper is outlined as follows: Sect. 2 gives an overview of the global chemistry simulation engine used to produce the data of interest. Section 3 highlights the various decomposition methods that can be produced using randomized linear algebra techniques. Section 4 shows the results of the dimensionality reduction procedures, highlighting the effectiveness of each technique. Section 5 shows how such techniques can be used for data compression and reduced order models, enabling compact representations of the data for a variety of broader scientific studies. Section 6 provides concluding remarks and a brief outlook for data sciences applied to atmospheric dynamics and global chemistry analysis.

2 Atmospheric chemistry model and data

Understanding the composition of the atmosphere is critical for a wide range of applications, including air quality, chemistry–climate interactions, and global biogeochemical cycling. Chemical transport models (CTMs) are used to simulate the evolution of atmospheric constituents in space and time (Brasseur and Jacob, 2017). A CTM solves the system of coupled continuity equations for an ensemble of m species with number density vector $\boldsymbol{n} = (n_1, \ldots, n_m)^{\mathrm{T}}$ via operator splitting of transport and local processes:

$$\frac{\partial n_i}{\partial t} = -\nabla \cdot (n_i \boldsymbol{U}) + (P_i - L_i)(\boldsymbol{n}) + E_i - D_i \quad i \in [1, m], \tag{1}$$

where $\boldsymbol{U}$ is the wind vector, $(P_i - L_i)(\boldsymbol{n})$ is the (local) chemical production and loss terms, E_i is the emission rate, and D_i is the deposition rate of species i. The transport operator

$$\frac{\partial n_i}{\partial t} = -\nabla \cdot (n_i \boldsymbol{U}) \quad i \in [1, m] \tag{2}$$

involves spatial coupling across the model domain but no coupling between chemical species, whereas the chemical operator

$$\frac{\mathrm{d}n_i}{\mathrm{d}t} = (P_i - L_i)(\boldsymbol{n}) + E_i - D_i \quad i \in [1, m] \tag{3}$$

includes no spatial coupling but the species are chemically linked through a system of ordinary differential equations (ODEs).

Chemistry models repeatedly solve Eqs. (2) and (3), which require full knowledge of the chemical state of the atmosphere at all locations and times. The resulting four-dimensional data sets (longitude, latitude, levels, and species) can become massive, which makes it unpractical to output them at a high temporal frequency. As a consequence, model output is generally restricted to a few selected species of interest (e.g., ozone), whereas the full model state is only output very infrequently, e.g., to archive the information for future model restarts ("restart file"). Here we show that the chemical state of a CTM, such as GEOS-Chem, has distinct low-ranked features and exploiting these properties using modern diagnostic tools such as variable reduction or subsampling makes it possible to represent the same amount of information in a computationally more efficient manner. While we focus on identifying low-ranked features across the spatiotemporal dimension (i.e., for each species separately), the methods presented could similarly (and independently) be applied across the species domain.

2.1 Global atmospheric chemistry simulations

The reference simulation of atmospheric chemistry was generated using the GEOS-Chem model. GEOS-Chem (http://geos-chem.org, last access: 11 April 2019) is an open-source global model of atmospheric chemistry that is used by over a hundred active research groups in 25 countries around the world for a wide range of applications. The code is freely available through an open license (http://acmg.seas.harvard.edu/geos/geos_licensing.html, last access: 11 April 2019). GEOS-Chem can be run in off-line mode as a chemical transport model (CTM) (Bey et al., 2001; Eastham et al., 2018) or as an online component within the NASA Goddard Earth system model (GEOS) (Long et al., 2015; Hu et al., 2018). In this study we use the off-line version of GEOS-Chem v11-01, driven by archives of assimilated meteorological data from the GEOS Forward Processing (GEOS-FP) data stream of the NASA Global Modeling and Assimilation Office (GMAO). The model chemistry scheme includes detailed HO_x–NO_x–VOC–ozone–BrO_x tropospheric chemistry as originally described by Bey et al. (2001), with addition of BrO_x chemistry from Parrella et al. (2012), and updates to isoprene oxidation as described by Mao et al. (2013). Dynamic and chemical time steps are 30 and 20 min, respectively. Stratospheric chemistry is modeled using a linearized mechanism as described by Murray et al. (2012).

We performed a 1-year simulation of GEOS-Chem (July 2013–June 2014) at $4° \times 5°$ horizontal resolution to generate a comprehensive set of atmospheric chemistry model diagnostics. For every chemistry time step, the concentrations of all 143 chemical constituents were archived immediately before and after chemistry (in units of molecules cm^{-3}). The difference between these concentration pairs is the species tendencies due to chemistry (expressed in units of molecules $\mathrm{cm}^{-3}\ \mathrm{s}^{-1}$). As the solution of chemical kinetics is also a function of the environment, we further output key environmental variables such as temperature, pressure, water vapor, and photolysis rates. The latter are computed online by GEOS-Chem using the Fast-JX code of Bian and Prather (2002) as implemented in GEOS-Chem by Mao et al. (2010) and Eastham et al. (2014). Thus, at every time step, the data set consists of $n\text{features} = 143 + 91 + 3 + 143 = 380$ data points at every grid location. We restrict our analysis to the lowest 30 model levels to avoid influence from the stratosphere. The resulting data set has the following dimensions: $n\text{long} \times n\text{lat} \times n\text{lev} \times n\text{times} \times n\text{features} = 72 \times 46 \times 30 \times 26\,280 \times 380 = 9.9 \times 10^{11}$.

2.2 Data preprocessing

Many dimensionality reduction techniques rely on an underlying singular value decomposition of the data that extracts correlated patterns in the data. A fundamental weakness of such SVD-based approaches is the inability to efficiently handle invariance in the data. Specifically, translational and/or rotational invariance of low-rank features in the data are not well captured (Kutz, 2013; Kutz et al., 2016). One of the key environmental variables driving the chemistry is photolysis rate; therefore, the absolute concentrations of many chemicals of interest accordingly "turn on" and are nonzero during daytime, and "turn off" or go to zero during the night. The time series of absolute chemical concentrations exhibit a translating wave traversing the globe from east to west with constant velocity. The time series for the chemical species OH (hydroxyl radical) is plotted with respect to UTC time for one latitude/elevation and three different longitudes on bottom left in Fig. 2, highlighting the translational invariance in the absolute concentration data. Any SVD-based approach will be unable to capture this translational invariance and correlate across snapshots in time, producing an artificially high dimensionality, i.e., a higher number of modes would be needed to characterize the dynamics due to translation (Kutz, 2013). To overcome this issue the time series for each grid point are shifted to align with the local GMT time, as shown on bottom right in Fig. 2. With the local times for each grid point aligned, SVD-based dimensionality reduction techniques can now identify and isolate coherent low-dimensional features in the data.

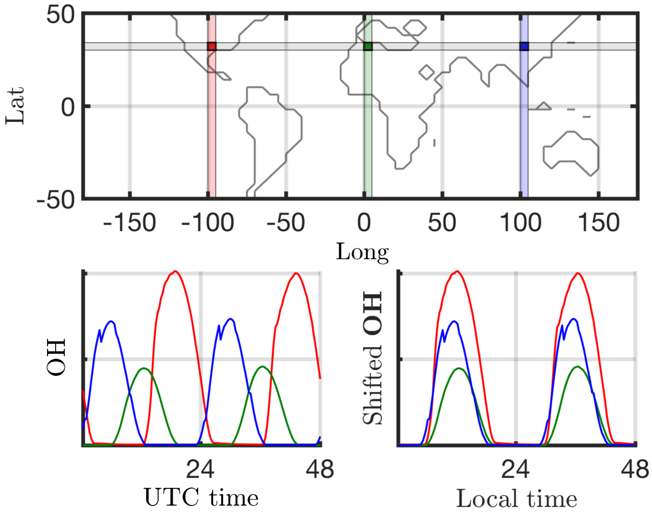

Figure 2. Shifting the data for each cell in time to align the local time zones across a latitude to the prime meridian (Long = 0) LT, shown here for the OH absolute concentration for Lat = 30.

3 Scalable matrix decompositions for diagnostics

The following subsections detail a probabilistic framework for matrix decompositions that includes a nonnegative matrix factorization as well as a sparsity-promoting technique. The mathematical architectures proposed provide scalable computational tools for the analysis of global chemistry dynamics. Moreover, by providing three different dimensionality architectures, a more nuanced objective analysis of the dominant spatiotemporal patterns that emerge in the global chemistry dynamics is achieved. The standard analysis would be a simple randomized SVD decomposition, whereby the dominant correlated structures are computed. A more refined approach to computing the dominant correlated structures involves restricting the dominant spatiotemporal structures to reasonable physical considerations. Specifically, the nonnegative matrix factorization restricts all chemicals to positive concentrations, a restriction which is physically motivated and especially important for diagnostics when physical interpretation is required. The randomized SVD will generally produce a negative concentration of chemicals in individual modes, but the overall concentration is positive when the modes are summed together. Likewise, the sparse PCA analysis zeroes out very small concentrations so that the modes extracted highlight only nonzero contributions to the dynamics. This is an important modification of the randomized SVD, as it generally produces all nonzero entries in the modal structures, regardless if it is physical. This is due to the least-square nature of the SVD algorithm. Again, a sparsification penalty produces modes where only the dominant coefficients are nonzero. What one chooses to use may depend strongly on the intended application. Regardless, the suite of methods allows for a more nuanced view of the data.

3.1 Probabilistic framework for low-rank approximations

Assume that the data matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ has rank r, where $r \leq \min\{m,n\}$. The objective of a low-rank matrix approximation to the input data matrix $\mathbf{X}$ is to find two smaller matrices

$$\underset{m\times n}{\mathbf{X}} \approx \underset{m\times r}{\mathbf{E}} \quad \underset{r\times n}{\mathbf{F}}, \tag{4}$$

where the columns of $\mathbf{E}$ span the column space of $\mathbf{X}$, and the rows of $\mathbf{F}$ span the row space of $\mathbf{X}$. These factors can be stored much more efficiently, and can be used to approximate the massive input data matrix and summarize the interesting low-dimensional features which are often interpretable. Probabilistic algorithms have been established over the past 2 decades to compute such computationally tractable smaller matrix approximations. We seek a near-optimal low-dimensional approximation of the input data matrix $\mathbf{X}$ using a probabilistic framework as formulated by Halko et al. (2011). Conceptually, the probabilistic framework splits the task of computing a near-optimal low-rank approximation into two logical stages:

- *Stage A*. Compute a low-dimensional subspace that approximates the column space of $\mathbf{X}$. We aim to find a near-optimal basis $\mathbf{Q} \in \mathbb{R}^{m\times k}$ with orthonormal columns such that

  $$\mathbf{X} \approx \mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{X} \tag{5}$$

  is satisfied, where k is the desired target rank. Random projections are used to sample the column space of the input matrix $\mathbf{X}$. Random projections are data agnostic, and are constructed by first drawing a set of k independent random vectors $\{\boldsymbol{\omega}_i\}_{i=1}^{k}$, for instance, from the standard normal distribution; $\mathbf{X}$ is then mapped to the low-dimensional space to obtain the random sample projections $\mathbf{y}_i := \mathbf{X}\boldsymbol{\omega}_i$ for $i = 1, \ldots, k$. Define a random test matrix $\boldsymbol{\Omega} = [\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_k] \in \mathbb{R}^{n\times k}$, where the sample random projections form the sampling matrix $\mathbf{Y} \in \mathbb{R}^{m\times k}$ are given by

  $$\mathbf{Y} := \mathbf{X}\boldsymbol{\Omega} \tag{6}$$

  $\mathbf{Y}$ is denoted as the *sketch matrix*. The columns of $\mathbf{Y}$are now orthonormalized using the QR-decomposition $\mathbf{Y} = \mathbf{QR}$, where $\mathbf{Q}$ is the near-optimal low-dimensional basis that approximates the column space of the input data matrix. For most real-world data matrices with a gradually decaying singular value spectrum, this basis matrix $\mathbf{Q}$ does not provide a good approximation for the column space of the input data matrix. A much better approximation is obtained by the following:

  - *Oversampling*. For target rank k, for most data matrices we may have nonzero singular values

$\{\sigma_i\}_{i=k+1}^{\min(m,n)}$. As a consequence, the sketch $\mathbf{Y}$ obtained above does not exactly span the column space of the input data matrix. Oversampling, i.e., using $l = k + p$ random projections to form the sketch overcomes this issue, and a small number of additional projections $p = \{5, 10\}$ is often sufficient to obtain a good basis comparable to the best possible basis (Martinsson, 2016).

  - *Power iteration scheme.* The quality of $\mathbf{Q}$ can be improved by the concept of power sampling iterations (Halko et al., 2011; Rokhlin et al., 2010). An improved sketch is defined under this concept as $\mathbf{Y} := \mathbf{X}^{(q)}\mathbf{\Omega}$, where q is an integer specifying the number of power iterations. This process enforces a more rapid decrease of the singular values, enabling the algorithm to sample the relevant information related to the dominant singular values while the unwanted information is suppressed. As few as $q = \{1, 2, 3\}$ power iterations can considerably improve the accuracy of the approximation. Orthogonalizing the sketch between each iteration further improves the numerical stability of the algorithm.

- *Stage B.* At this stage, we form a smaller matrix $\mathbf{B}$

$$\mathbf{B} := \mathbf{Q}^{\mathrm{T}}\mathbf{X} \in \mathbb{R}^{l \times n} \tag{7}$$

In other words, we restrict the high-dimensional input matrix to the low-dimensional space spanned by the near-optimal basis $\mathbf{Q}$ obtained in Stage A. Geometrically, this is a projection which takes points in a high-dimensional measurement space to a low-dimensional space while maintaining the structure in a Euclidean sense.

The probabilistic framework detailed above is referred to as the QB decomposition of the input data matrix $\mathbf{X}$, and yields the following low-rank approximation:

$$\underset{m \times n}{\mathbf{X}} \approx \underset{m \times l}{\mathbf{Q}} \; \underset{l \times n}{\mathbf{B}} \tag{8}$$

Note that the randomized algorithm outlined here requires two passes over the entire data matrix to construct the basis matrix $\mathbf{Q}$. The near-optimal low-rank approximation $\mathbf{B} \in \mathbb{R}^{l \times n}$, where $l \ll \min(m, n)$, can now be used instead of the data matrix $\mathbf{X}$ to compute traditional deterministic matrix decompositions for data analysis. The QB decomposition can also be extended to distributed and parallel computing, see Voronin and Martinsson (2015).

3.2 Randomized singular value decomposition

The data matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ has a singular value decomposition (SVD) of the form

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} \tag{9}$$

with unitary matrices $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_m] \in \mathbb{R}^{m \times m}$ and $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_n] \in \mathbb{R}^{n \times n}$ orthonormal such that $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$. The left singular vectors in $\mathbf{U}$ provide a basis for the range (column space), and the right singular vectors in $\mathbf{V}$ provide a basis for the domain (row space) of the data matrix $\mathbf{X}$. The rectangular diagonal matrix $\mathbf{\Sigma} \in \mathbb{R}^{m \times n}$ has the corresponding nonnegative singular values $\sigma_1 \geq \ldots \geq \sigma_n \geq 0$, which describe the spectrum of the data. Low-rank matrices have rank r that is much smaller than the dimension of the measurement space, i.e., $r \ll m, n$ and the singular values $\{\sigma_i : \geq r + 1\}$ are zero. The corresponding singular vectors span the left and right null spaces of the matrix. In practical applications the data matrix are often contaminated by errors making its effective rank smaller than the exact rank r. In such cases the matrix can be well approximated by only those singular vectors which correspond to the singular values of a significant magnitude, and a reduced version of the SVD is computed:

$$\begin{aligned}\mathbf{X}_k &:= \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k \\ &= [\mathbf{u}_1, \ldots, \mathbf{u}_k]\,\mathrm{diag}(\sigma_1, \ldots, \sigma_k)\,[\mathbf{v}_1, \ldots, \mathbf{v}_k]^{\mathrm{T}},\end{aligned} \tag{10}$$

where k denotes the desired target rank of the approximation. Choosing an optimal k is highly dependent on the task. If a highly accurate reconstruction of the original data is desired, then k should be chosen closer to the effective rank of the data matrix. Conversely, if a very low-dimensional representation of dominant features is desired, then k might be chosen to be much smaller. The Eckart–Young theorem (Eckart and Young, 1936) states that the low-rank SVD provides the optimal rank-k reconstruction of a matrix in the least-squares sense

$$\mathbf{X}_k := \underset{\mathrm{rank}(\mathbf{X}'_k)}{\mathrm{argmin}} \left\|\mathbf{X} - \mathbf{X}'_k\right\| \tag{11}$$

with the reconstruction error in the spectral and Frobenius norm given by

$$\|\mathbf{X} - \mathbf{X}_k\|_2 = \sigma_{k+1}(\mathbf{X}) \tag{12}$$

and

$$\|\mathbf{X} - \mathbf{X}_k\|_{\mathrm{F}} = \sqrt{\sum_{j=k+1}^{\min(m,n)} \sigma_j^2(\mathbf{X})} \tag{13}$$

For massive data sets, however, the cost of computing the full SVD of the data matrix $\mathbf{X}$ is order $O\left(mn^2\right)$, from which the first k components can then be extracted to form $\mathbf{X}_k$. Randomized algorithms are computationally efficient and ‘surprisingly’ reliable; these techniques can be used to obtain an approximate rank-k SVD at a substantially more efficient cost of $O(mnk)$.

The probabilistic framework is used to obtain a near-optimal low-rank approximation $\mathbf{B} \in \mathbb{R}^{l \times n}$, where $l \ll$

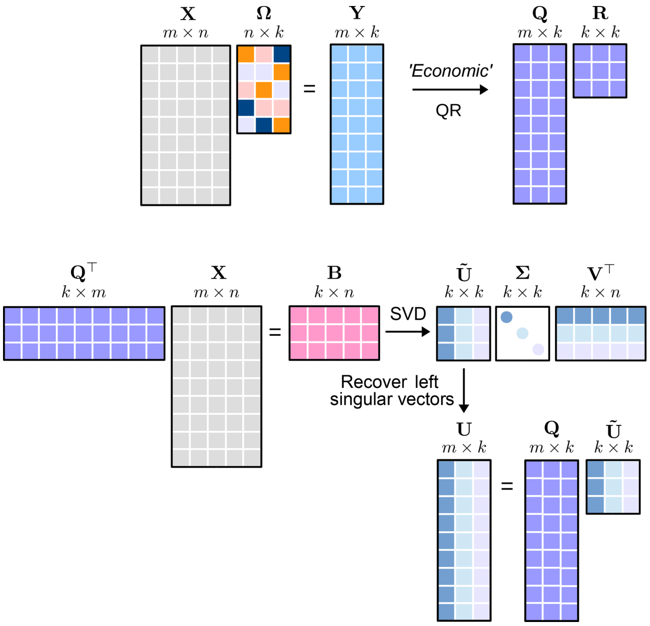

Figure 3. Illustration of the randomized matrix decomposition technique. The random sampling matrix $\mathbf{\Omega}$ is used to produce a new matrix $\mathbf{Y}$ which can be decomposed using a QR decomposition. This leads to the construction of the matrix $\mathbf{B}$ which is used for approximating the left and right singular vectors.

$\min(m,n)$. This can now be used instead of the data matrix $\mathbf{X}$, and a full SVD of $\mathbf{B}$ is computed

$$\mathbf{B} = \widetilde{\boldsymbol{U}}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \tag{14}$$

to give the first l right singular vectors $\boldsymbol{V} \in \mathbb{R}^{n\times l}$ and the corresponding singular values $\boldsymbol{\Sigma} \in \mathbb{R}^{l\times l}$. The left singular vectors $\boldsymbol{U} \in \mathbb{R}^{m\times l}$ are recovered from the approximate left singular vectors $\widetilde{\boldsymbol{U}} \in \mathbb{R}^{l\times l}$ using the near-optimal basis matrix $\mathbf{Q}$

$$\boldsymbol{U} \approx \mathbf{Q}\widetilde{\boldsymbol{U}} \tag{15}$$

For the absolute concentration data matrix, note that the right singular vectors $\boldsymbol{V}$ are temporal and the left singular vectors $\boldsymbol{U}$ are the spatial dominant features of the system. We also compute a cumulative energy spectrum from the singular values, the energy in the first j dominant modes is given by

$$\frac{\sum_{i=1}^{j}\sigma_i^2}{\text{Total energy in the data}}, \tag{16}$$

where the total energy in the data is computed using the Frobenius norm as $\|\mathbf{X}\|_{\mathrm{F}}^2$.

The algorithm architecture is conceptually outlined in Fig. 3. This shows the basic architecture and the structure which allows for a rapid approximation of the left and right singular values and eigenvectors.

3.3 Randomized nonnegative matrix factorization

A significant drawback of commonly used dimensionality reduction techniques, such as SVD based principal component analysis (PCA), is that they permit both positive and negative terms in their components. In many data applications, such as in the absolute concentration, negative terms fail to be interpretable in a physically meaningful sense, i.e., chemical concentrations are not negative. To address this problem the set of basis vectors are constrained to nonnegative terms (Lee and Seung, 1999; Paatero and Tapper, 1994) – this paradigm is the nonnegative matrix factorization (NMF). NMF has emerged as a powerful dimension reduction tool that allows the computation of a sparse, parts-based representation of physically meaningful additive factors that describe coherent structures within the data. Given the data matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$, the NMF has to find two matrices of a much lower rank

$$\underset{m\times n}{\mathbf{X}} \approx \underset{m\times k}{\mathbf{W}}\ \underset{k\times n}{\mathbf{H}}, \tag{17}$$

where k is the target rank. The SVD finds an exact solution of this problem in the least-squares sense, as detailed in the previous section, but the resulting factors are not guaranteed to be physically meaningful, i.e., positive values. NMF, in comparison, gives an additive parts-based representation of the data that preserves useful properties such as sparsity and nonnegativity by imposing additional nonnegativity constraints: $\mathbf{W} \geq 0$ and $\mathbf{H} \geq 0$. The sparse parts-based features have an intuitive interpretation which has been exploited in environmental modeling (Paatero and Tapper, 1994). In environmental data, the error estimates of data can be widely varying and nonnegativity is often an essential feature of the underlying models (Juntto and Paatero, 1994; Lee et al., 1999; Paterson et al., 1999; Xie et al., 1999). Traditionally, the NMF problem is formulated as the following optimization problem:

$$\begin{array}{lccc} \text{minimize} & f(\mathbf{W},\mathbf{H}) & = & \|\mathbf{X}-\mathbf{WH}\|_{\mathrm{F}}^2 \\ \text{subject to} & \mathbf{W} \geq 0 & \text{and} & \mathbf{H} \geq 0 \end{array} \tag{18}$$

This optimization problem is nonconvex and ill-posed. As no convexification exists to simplify the optimization, no exact or unique solution is guaranteed (Gillis, 2017). Therefore, different NMF algorithms can produce distinct decompositions that minimize the objective function. As the problem is nonconvex with respect to both factors $\mathbf{W}$ and $\mathbf{H}$, most NMF algorithms divide the problem into simpler subproblems that have closed form solutions. The convex subproblem is solved by keeping one factor fixed while updating the other, and alternating and iterating until convergence. The hierarchical alternating least squares (HALS) is one variant of this method, proved to be highly efficient (Cichocki and Phan, 2009), and this is the algorithm employed here for computing the NMF.

Block coordinate descent (BCD) iterative methods fix a block of components and optimize with respect to the remaining components. The factors $\mathbf{W}$ and $\mathbf{H}$ are initialized and updated by fixing most terms except for the block comprised of the jth column $\mathbf{W}_{(:,j)}$ and the jth row $\mathbf{H}_{(j,:)}$. HALS approximately minimizes the cost function in Eq. (18) with respect to the remaining $k-1$ components

$$\text{minimize } J_j\left(\mathbf{W}_{(:,j)}, \mathbf{H}_{(j,:)}\right) = \left\|\mathbf{R}^{(j)} - \mathbf{W}_{(:,j)}\mathbf{H}_{(j,:)}\right\|_{\mathrm{F}}^2, \tag{19}$$

where $\mathbf{R}^{(j)}$ is the jth residual

$$\mathbf{R}^{(j)} := \mathbf{X} - \sum_{i \neq j}^{k} \mathbf{W}_{(:,i)}\mathbf{H}_{(i,:)} \tag{20}$$

Gradients are derived to find the stationary points for both components, for details see Cichocki and Phan (2009).

For massive data sets randomness is again employed to replace the high-dimensional input data matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ by its near-optimal low-rank approximation $\mathbf{B} \in \mathbb{R}^{l \times n}$, where $l \ll \min(m,n)$, with the exception that the entries of $\boldsymbol{\Omega}$ are drawn independently from the uniform distribution with support $\omega \in [0,1]$. We now have the following optimization problem:

$$\begin{array}{llll} \text{minimize} & \widetilde{f}\left(\widetilde{\mathbf{W}}, \mathbf{H}\right) & = & \left\|\mathbf{B} - \widetilde{\mathbf{W}}\mathbf{H}\right\|_{\mathrm{F}}^2 \\ \text{subject to} & \mathbf{Q}\widetilde{\mathbf{W}} \geq 0 & \text{and} & \mathbf{H} \geq 0, \end{array} \tag{21}$$

where the nonnegativity constraints need apply to the high-dimensional factor matrix $\mathbf{W}$, but not necessarily to $\widetilde{\mathbf{W}}$, as $\widetilde{\mathbf{W}}$ can be rotated back to high-dimensional space using the approximate relation $\mathbf{W} \approx \mathbf{Q}\widetilde{\mathbf{W}}$. As $\mathbf{Q}\mathbf{Q}^{\mathrm{T}} \neq \mathbf{I}$, Eq. (21) can only be solved approximately. The randomized HALS algorithm is formulated as

$$\text{minimize } J_j\left(\widetilde{\mathbf{W}}_{(:,j)}, \mathbf{H}_{(j,:)}\right) = \left\|\widetilde{\mathbf{R}}^{(j)} - \widetilde{\mathbf{W}}_{(:,j)}\mathbf{H}_{(j,:)}\right\|_{\mathrm{F}}^2, \tag{22}$$

where $\mathbf{R}^{(j)}$ is the jth compressed residual

$$\widetilde{\mathbf{R}}^{(j)} := \mathbf{B} - \sum_{i \neq j}^{k} \widetilde{\mathbf{W}}_{(:,i)}\mathbf{H}_{(i,:)} \tag{23}$$

The components are updated again by deriving the gradients. For further details, such as initialization techniques, stopping criterion, and variants of randomized HALS we refer to Erichson et al. (2018a).

For the absolute chemistry concentration data matrix, the columns of the factor $\mathbf{W}$ are the spatial modes while those of factor $\mathbf{H}$ are the temporal modes. The randomized NMF algorithm starts with an initial guess derived from a SVD of the data matrix, and returns the $\mathbf{W}$, $\mathbf{H}$ factors with columns that are not ordered. The 2-norm of the columns is computed, and the columns are normalized and ordered. A product of the ordered column-wise 2-norms gives the "spectrum" for the decomposition. From this spectrum a cumulative energy spectrum is computed similar to Eq. (16).

3.4 Sparse randomized principal component analysis

Principal component analysis is a prevalent technique for dimensionality reduction, it exploits relationships among points in high-dimensional space to construct a new set of uncorrelated low-dimensional variables or principal components (PCs). The first PC explains most of the variation in the data, the second PC accounts for the second greatest variance in the data, and so on. For the data matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, which has now been centered with zero mean, with m being the number of observations and n being the number of variables, the PCs $\mathbf{z}_i \in \mathbb{R}^m$ are constructed as a weighted linear combination of the original variables

$$\mathbf{z}_i = \mathbf{X}\boldsymbol{w}_i, \tag{24}$$

where $\boldsymbol{w}_i \in \mathbb{R}^n$ is a vector of the corresponding weights, also denoted as modes or basis functions. Expressed concisely,

$$\mathbf{Z} = \mathbf{X}\mathbf{W} \tag{25}$$

with $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{m \times n}$ and $\mathbf{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n] \in \mathbb{R}^{n \times n}$. In most dimensionality reduction applications only the first k PCs will be of interest to visualize the data in a low-dimensional space, and as the relevant features used for data clustering, classification and regression. The problem of finding the PCs can be formulated as a variance maximization problem or as a least-squares problem, i.e., minimizing the sum of squared residual errors with orthogonality constraints on the weight matrix as

$$\begin{array}{lll} \underset{\mathbf{W}}{\text{minimize}} f(\mathbf{W}) & = & \frac{1}{2}\left\|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{W}^{\mathrm{T}}\right\|_{\mathrm{F}}^2 \\ \text{subject to } \mathbf{W}^T\mathbf{W} & = & \mathbf{I} \end{array} \tag{26}$$

The classic PCA approach outlined above generates global PCs as a linear combination of all n variables; hence, this approach tends to mix or blend various spatiotemporal scales and fails to identify and isolate underlying governing dynamics acting at each scale. Sparse principal component analysis (SPCA) is a variant which provides interpretable PCs with localized spatial support, providing a "parsimonious" decomposition through sparsity promoting regularizers on the weights $\mathbf{W}$. Each of the sparse weight vectors $\boldsymbol{w}_i$ have only a few nonzero values; therefore, we get a linear combination of only a few of the original variables. The SPCA is mathematically formulated as a variant of PCA outlined in Eq. (26) as

$$\begin{array}{lll} \underset{\mathbf{A},\,\mathbf{W}}{\text{minimize}} f(\mathbf{A},\,\mathbf{W}) & = & \frac{1}{2}\left\|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{A}^{\mathrm{T}}\right\|_{\mathrm{F}}^2 + \psi(\mathbf{W}) \\ \text{subject to } \mathbf{A}^{\mathrm{T}}\mathbf{A} & = & \mathbf{I}, \end{array} \tag{27}$$

where $\mathbf{W}$ is now a sparse weight matrix, and $\mathbf{A}$ is an orthonormal inverse transform matrix, i.e., the data can be approximately constructed as $\widetilde{\mathbf{X}} = \mathbf{Z}\mathbf{A}^{\mathrm{T}}$, where $\mathbf{Z}$ is the PC matrix given by Eq. (25). In Eq. (27), ψ is a sparsity inducing regularizer such as

- ℓ_0 norm defined as the number of nonzero elements in a vector $\boldsymbol{x}$, which is constrained to be $\ll n$

  $$\psi_0(\boldsymbol{x}) = \|\boldsymbol{x}\|_0. \tag{28}$$

- ℓ_1 norm, in this case the regularization problem is also known as LASSO (least absolute shrinkage and selection operator) (Trendafilov et al., 2003)

  $$\psi_1(\boldsymbol{x}) = \alpha\|\boldsymbol{x}\|_1, \tag{29}$$

  where α controls the degree of sparsity.

- The elastic net (Zou and Hastie, 2003) which is a combination of the ℓ_1 norm and quadratic penalty

  $$\psi_\mathrm{E}(\boldsymbol{x}) = \alpha\|\boldsymbol{x}\|_1 + \beta\|\mathbf{x}\|_2^2, \tag{30}$$

  where α and β control the degree of sparsity.

Note that the optimization problem in Eq. (27) is nonconvex and is solved similar to the NMF optimization problem by keeping one factor fixed while updating the other, and alternating and iterating until convergence. For further details refer to Erichson et al. (2018b).

For massive data sets, randomization using the probabilistic framework is employed again, where the original input data matrix $\mathbf{X}$ is projected to the range of $\mathbf{Y}$ defined in Eq. (6) so that we can reformulate Eq. (27) as

$$\begin{aligned} \underset{\mathbf{A},\,\mathbf{W}}{\text{minimize}} f(\mathbf{A},\mathbf{W}) \quad &= \quad \tfrac{1}{2}\|\widetilde{\mathbf{X}} - \widetilde{\mathbf{X}}\mathbf{W}\mathbf{A}^\mathrm{T}\|_\mathrm{F}^2 + \psi(\mathbf{W}) \\ \text{subject to } \mathbf{A}^\mathrm{T}\mathbf{A} \quad & \qquad\qquad\qquad \mathbf{I} \end{aligned} \tag{31}$$

The absolute concentration data matrix is first scaled to have mean zero. The spatial modes are the columns of matrix $\mathbf{W}$. The temporal modes or the PCs are the columns of $\mathbf{Z}$ computed from $\mathbf{X} = \mathbf{Z}\mathbf{A}^\mathrm{T}$. The minimization algorithm also formulates the problem as an eigenvalue problem, and returns the eigenvalues λ_j associated with the jth mode of the decomposition, which help compute the energy spectrum of the decomposition. The energy captured by the first j modes of the decomposition is computed as

$$\frac{\sum_{i=1}^{j}\boldsymbol{\lambda}_i \times (n-1)}{\text{Total energy in the scaled data}}, \tag{32}$$

where n is the total number of snapshots in time.

4 Data diagnostics

In this section we illustrate results from the decomposition of the GEOS-Chem model output using absolute concentration of ozone (O_3) as an example. The Supplement provides diagnostics for five additional chemicals known to dominate the global atmospheric chemistry dynamics. The additional five chemical species, including NO, NO_2, OH, isoprene (ISOP), and CO, are known to be equally important to O_3. To keep the paper succinct, we only present O_3 here, and the other species are presented in the Supplement. Overall, there are close to 2 hundred chemicals that interact dynamically. Each chemical of interest can be diagnostic in a similar fashion to O_3 in order to determine its dominant global variability. However, how the interactions across the entire chemical space ultimately drive the observed variability remains an open research question. The scalable diagnostics advocated here provide a computational architecture allowing scientists to explore this further by providing global diagnostics for all chemicals in a computationally tractable manner.

O_3 is a key oxidant of the atmosphere, and high surface concentrations of this species are harmful to human health and vegetation (Avnery et al., 2011; Silva et al., 2013). O_3 production involves the photochemical oxidation of volatile organic compounds (VOCs) and carbon monoxide (CO) in the presence of nitrogen oxide radicals ($NO_x \equiv NO + NO_2$). The chemistry of O_3 is highly complex, involving hundreds of chemical species. This makes O_3 a challenging compound for chemistry models (e.g., Stevenson et al., 2006; Sherwen et al., 2017; Mao et al., 2018). We find that despite the underlying complexity of the chemistry, the O_3 concentration fields produced by GEOS-Chem exhibit prominent, low-ranked features.

For a given chemical species of interest the absolute concentration data matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ has dimensions $m = \text{nlon} \times \text{nlat} \times \text{nlev} = 72 \times 46 \times 30$ spatial cells, and $n =$ number of time snapshots $= 26\,208$ for the yearlong data (one snapshot every 20 min).

4.1 Taking a logarithm of the data

For some chemical species the absolute concentration values in a small localized region dominate over the values in the rest of the grid cells. For instance, the absolute concentration values of nitric oxide (NO) are several orders of magnitude higher over China and eastern Russia compared with those over oceans and less populated regions if the world. Correspondingly the dominant spatial modes are very localized as exhibited in the top panel of Fig. 4, with only one nonzero peak over eastern Russia for the second most dominant spatial mode. SVD is unable to resolve the underlying global low-order spatial features. To resolve this issue a logarithm of the data values is used instead, to bring all the concentration values to the same scale and prevent smaller signals from being damped out. The data matrix now is $\mathbf{X}_\mathrm{log} = \log(\mathbf{X}+1)$. The second most dominant mode of the logarithm of the data, as shown in the bottom panel of Fig. 4, now exhibits global low-order features of the data. Thus, the SVD and other matrix decomposition techniques will be able to identify and isolate global dominant low-order structure in the system for chemical species exhibiting localized dominant values.

Normalization of data is a common practice in data science. Indeed, the ubiquitous PCA analysis requires that each

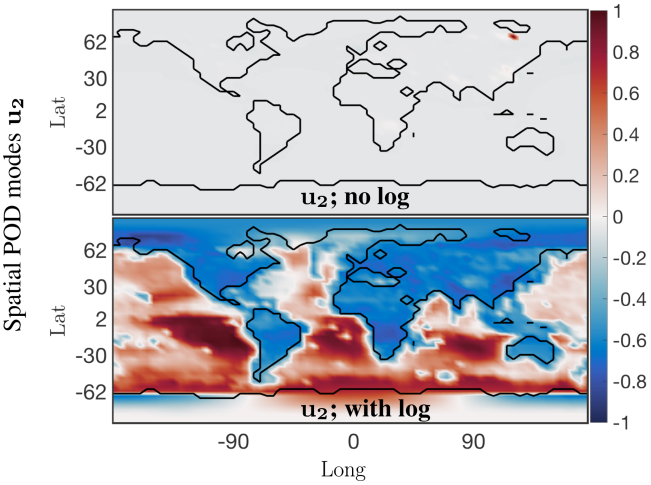

Figure 4. Dominant spatial mode 2 at the surface for the NO absolute concentration preprocessed data before and after taking a logarithm of the preprocessed data. Taking a logarithm scales the preprocessed data so that the corresponding spatial modes exhibit the global low-dimensional features, instead of only picking up on the dominant chemistry in one localized region.

measurement type in the data have mean zero and unit variance. If this is not enforced, then those signals that are measured with large numbers will simply drown out the signals measured in small numbers. Thus, the units of the different measurements are neutralized by requiring a mean zero and unit variance. Similarly here, the large spike in the data is so large that the rest of the data are like noise comparatively. By normalizing with the logarithm, a more balanced global view of the chemistry dynamics can be extracted from the modal structures.

4.2 Modes from randomized SVD

We begin by considering the singular value spectrum and the dominant four temporal modes from the randomized SVD of the absolute concentration of O_3 (O_3). These are presented in the top panel of Fig. 5. The amount of energy explained by the most dominant singular values gives a good indication regarding the low-rank nature of the underlying data. Figure 5a shows the cumulative energy explained by the 150 most dominant singular values, as derived from randomized SVD. If all 2.7×10^{11} model output data points were perfectly independent, each singular value would represent $1.0/2.7 \times 10^{11} = 3.7 \times 10^{-10}$ % of the total energy. Instead, we find that the first 4 singular values combined explain 97 % of the total field energy, and the first 150 singular values capture almost 100 % of the total energy. Thus, it is possible to explain 99 % of the spatiotemporal structure of the highly complex O_3 field with just 20 modes. These modes reveal many of the dominant features of atmospheric O_3. Figure 5b illustrates the structure of the four dominant temporal modes. The most dominant mode (blue line) has a flat temporal structure, i.e., its importance is independent of the time of the year. The next three dominant modes all have distinct temporal patterns, i.e., they capture periodical features of atmospheric O_3. Modes 2 and 3 (red and yellow, respectively) both exhibit a frequency of 1 year, capturing features occurring on an annual basis. The fourth most dominant mode (purple) has a frequency of 6 months. Geophysical interpretation of these modes is easiest when combining the temporal pattern with the corresponding spatial features, the latter of which are shown in Fig. 6. Shown are the spatial pattern of the eight most dominant modes for the surface. It should be emphasized that the spatial patterns change with altitude, as illustrated in the Supplement. Surface O_3 exhibits distinct seasonal patterns, which are captured by the first four modes: the first mode (top left panel in Fig. 6) resembles the annual average surface concentration of O_3. It can be interpreted as the time-invariant “average O_3” field from which all other modes add or subtract to describe the spatiotemporal variability of O_3 in greater detail. The second singular value (top right panel) shows a strong gradient at the Equator as well as a distinct urban pattern over the Northern Hemisphere (NH). The seasonal variability of this mode (peaking in August, see Fig. 5) broadly follows observed O_3 burdens in the Southern Hemisphere (SH) (Cooper et al., 2014), and O_3 is known to increase during summertime in urban areas in the NH as a result of increased photochemical activity. Singular mode 3 can be seen as an additional “forcing” to this seasonality for NH O_3: it shows dominant features over polluted areas (Europe and East China) and its seasonal amplitude complements that of singular mode 2. The most distinct feature of mode 4 is the strong pattern over Africa. We interpret this as biomass burning signal. This is supported by the frequency pattern of this mode, which shows two peaks in January–February and July–August, which is in agreement with the two biomass burning seasons over Africa (Roberts et al., 2009).

To summarize, inspection of the spatial and temporal patterns of the dominant modes of O_3 shows that randomized SVD successfully reveals prominent features of tropospheric O_3 chemistry, such as elevated summertime O_3 over polluted urban areas or the two biomass burning seasons over Africa. While the data set used in this study is too short to generalize the findings, these results demonstrate the potential of randomized SVD for pattern discovery of atmospheric chemistry model output. In particular, the extent and temporal variability of the singular values can help identify highly correlated “chemical domains” within the model, which has practical applications for model reduction considerations.

4.3 Modes from randomized NMF

A drawback of the SVD solution presented in Sect. 4.2 is that it accepts both negative and positive solutions, which can result in physically unrealistic negative species concentrations. As discussed in Sect. 3.3, positive solutions can be enforced

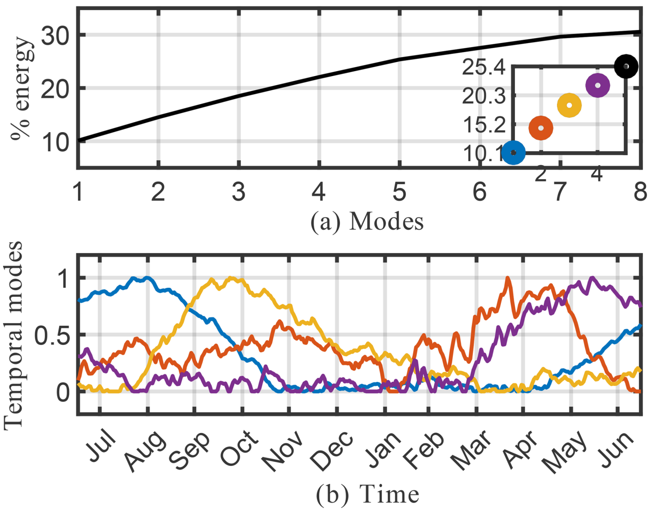

Figure 5. Cumulative energy spectrum (and inset detail) of the singular value decomposition **(a)** and the corresponding four dominant temporal modes **(b)** for the O_3 absolute concentration preprocessed data.

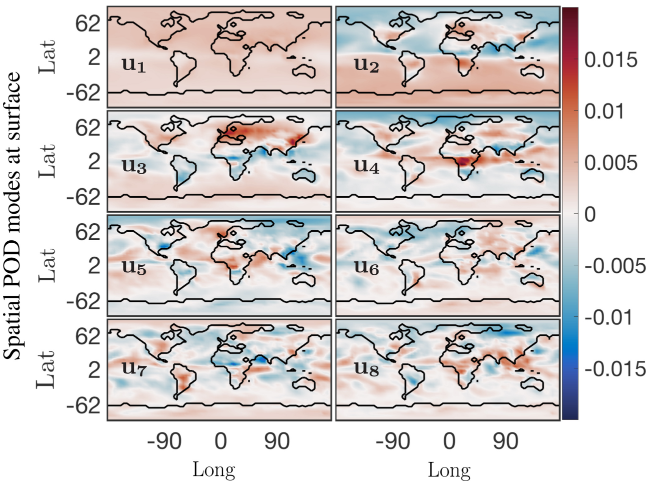

Figure 6. First eight dominant spatial modes at the surface for the O_3 absolute concentration preprocessed data. Mode 1 is the constant or mean value mode, its corresponding temporal behavior is the blue trend in Fig. 5b. Global low-dimensional spatial features for this chemical species are exhibited in order of dominance in modes 2 through 8.

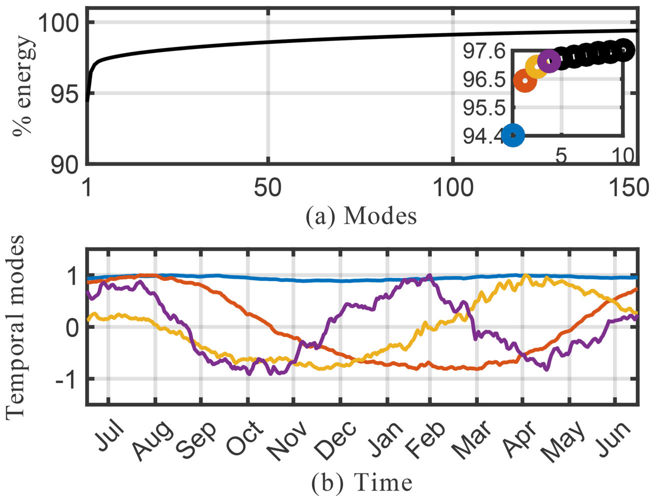

Figure 7. Cumulative energy spectrum from the nonnegative matrix factorization **(a)** and the corresponding first four columns of the ordered **H** temporal factor for the O_3 absolute concentration preprocessed data **(b)**.

using NMF. The results from NMF for the O_3 absolute concentration data are presented in Figs. 7 and 8. The cumulative energy spectrum exhibited in Fig. 7a shows a much slower decay than the spectrum from the SVD decomposition. This is to be expected, as NMF computes an additive parts-based representation of the low-order features in the data, which preserves sparsity in the data but requires more modes to capture the same level of energy compared with the SVD. The four dominant temporal modes are presented in Fig. 7b. These now capture approximately 20 % of the total energy spectrum, compared with 97 % for the SVD. This is, in large part, due to the fact that the positivity-constraint prevents the NMF from creating a mode for annual mean O_3 that can explain most of the energy spectrum – akin to mode 1 for SVD – but that requires both additions and subtractions from this mean field to describe O_3 variations in more detail. As a result, none of the NMF modes reflects a distinct representation of the global average O_3 field. This is supported by the lack of a time-invariant mode (see Fig. 7) and also becomes apparent from the corresponding spatial patterns shown in Fig. 8. None of those resemble the average mean O_3 concentration field as, e.g., SVD mode 1 (see Fig. 6). Still, the first four spatial and temporal modes of NMF reflect some well known features of O_3 chemistry, albeit less obvious than for SVD. The most dominant NMF mode shows a pattern comparable to the second mode of SVD, and also has an almost identical temporal structure with a distinct peak in July–August. The second mode is almost a mirror image of the first mode, with a strong, broad-based signal in the NH that is most dominant during March–May but that also contributes during most other months except January. Mode 3 peaks during September–October but contributes meaningfully until February. Its spatial pattern is strongest over South America, India, eastern China, and southern Africa, and thus captures some of the increased O_3 concentrations due to fire activities (e.g., the South American burning season in August–September–October and the Indian burning season in October–November). Mode 4 is similar to mode 3 of the SVD, with strong signals over Europe and eastern China that peak during boreal spring.

Similar to SVD, the spatiotemporal modes of surface O_3 derived from NMF reveal many of the characteristics of O_3 chemistry, such as increased O_3 concentrations over urban

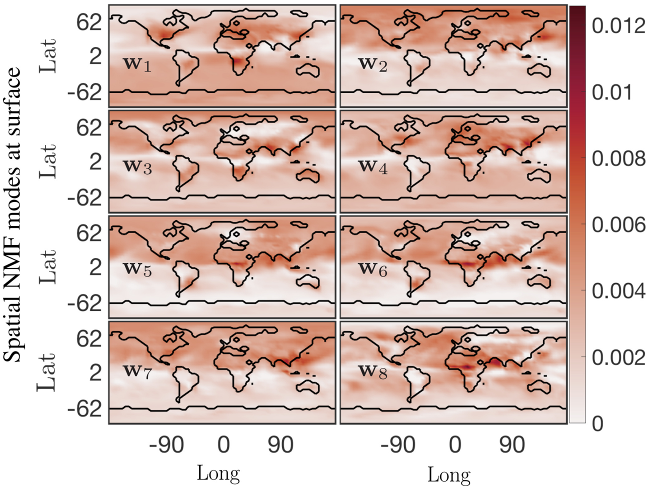

Figure 8. First eight columns of the ordered **W** spatial factor from NMF at the surface for the O_3 absolute concentration preprocessed data. These modes lend themselves to easy interpretation; the most dominant mode $\mathbf{w}_1$ indicates that the O_3 absolute concentration is most active near eastern coastal urban China, North America, and the western coastal African continent around the region of Congo.

areas and biomass burning regions, as well as the seasonality of these events. Due to the strict positiveness of the solution, the signal is more muted than SVD, and significantly more modes are needed to reproduce the spatiotemporal pattern of O_3 in detail. This makes SVD better suited for off-line pattern discovery applications. However, for the practical employment of reduced-order modeling techniques within an Earth system model, we consider NMF to be superior as it still realistically captures O_3 patterns with relatively few (tens of) modes but its concentrations are guaranteed to be positive.

4.4 Modes from randomized SPCA

Spatial modes computed from the randomized SPCA are shown in Fig. 10. Note the localized features isolated by SPCA in these dominant spatial modes compared with the modes computed by the full SVD. We impose the sparsity regularizer given by Eq. (30) with $\alpha = 1 \times 10^{-4}$ and $\beta = 1 \times 10^{-12}$. Reducing the value of α gives a less sparse decomposition. The cumulative energy spectrum in the top panel of Fig. 9 again demonstrates much slower decay than the SVD and more modes are needed to capture the same amount of energy due to the sparsity constraint. In terms of energy explained and interpretability of the modes, the SPCA results for O_3 sit in between the results for SVD and NMF discussed above. The first four SPCA modes capture more than 50 % of the total energy (Fig. 9), which is more than NMF but significantly less than SVD. As for NMF, the lower amount of energy compared with the SVD can be attributed to the fact that the SPCA does not compute a dominant mode for the mean annual O_3 concentration. This is expected as SPCA is designed to capture spatially distinct features, rather than broad-based patterns. Thus, it “assembles” total O_3 concentrations from a series of modes that all show distinct spatial features. Of the dominant four modes shown here, the fourth mode most closely resembles a generic mean concentration field that contributes to the signal throughout the year (even though the signal is stronger during boreal winter). The SPCA reveals many features that are also apparent in the SVD and NMF results. The SPCA mode 1 is almost identical to mode 2 of SVD, both in spatial extent and its temporal variability. Mode 2 acts to lower O_3 over Europe and eastern China, but at a muted rate during March–May and also July–August. Therefore, it has a similar effect to mode 3 of the SVD, but with the opposite sign. Mode 3 can be interpreted as a biomass burning signal, with its distinct hot spot over Africa and the two seasonal peaks.

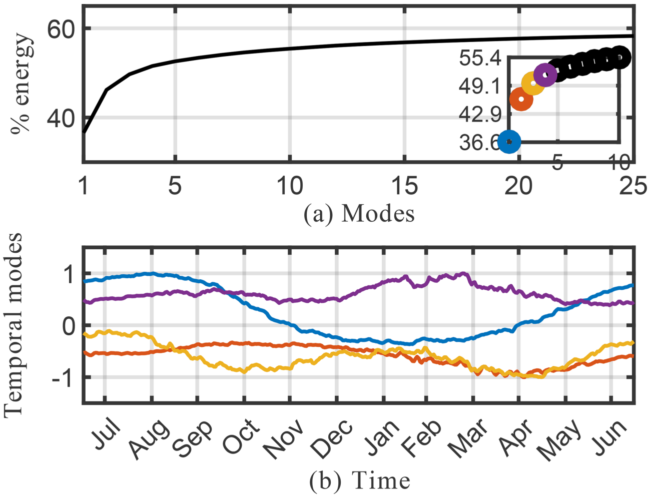

Figure 9. Cumulative energy spectrum from the sparse principal component analysis **(a)** and the corresponding four dominant temporal modes **(b)** for the O_3 absolute concentration preprocessed data.

5 Data compression and reduced order modeling

Scalable diagnostic analysis is only one critically enabling aspect of the randomized decomposition methods. Indeed, the various randomized algorithms can be used to compute low-rank embeddings of the data that can be used for data compression. Thus, an accurate approximation of the data can be stored with a fraction of the memory requirements of the full, high-fidelity simulation. Compression is exploited in most portable electronic formats (e.g., smart phones) by representing the data using a basis which is amenable to a sparse representation (Kutz, 2013). For instance, images can be massively compressed using wavelet or Fourier basis elements as natural images are sparse in these basis elements.

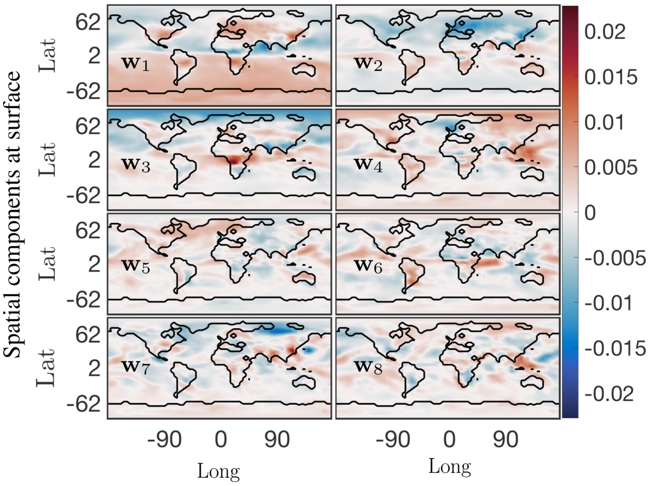

Figure 10. First eight principal components from SPCA at the surface for the O_3 absolute concentration preprocessed data. With the sparsity constraint these spatial modes exhibit only localized low-dimensional features compared with those from the SVD of the data. Compare the SVD mean value mode 1 $\mathbf{u}_1$ from Fig. 6, which exhibits a more or less constant field as the dominant low-dimensional global feature, with SPCA mode 1 $\mathbf{w}_1$ here, which picks up on localized dominant features in the data. The corresponding temporal SPCA mode 1 also exhibits a seasonal variation.

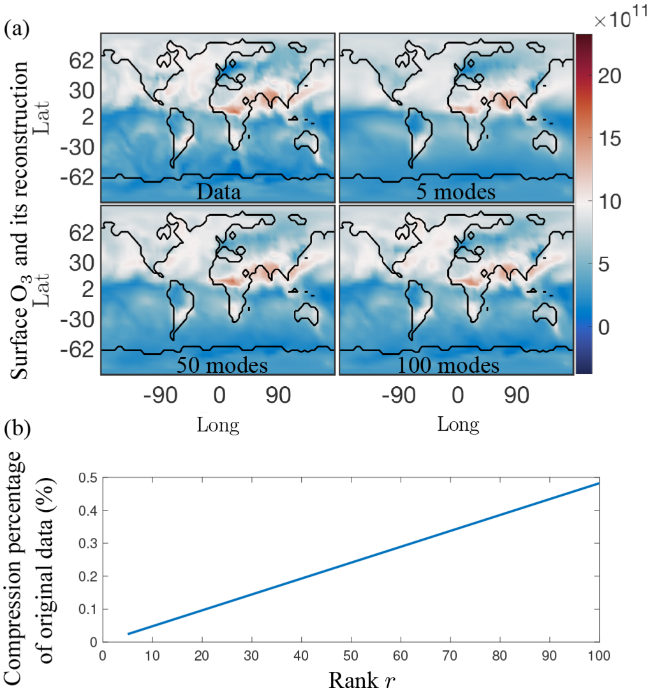

Figure 11. (a) Single snapshot of the surface O_3 absolute concentration reference data (top left) and its reconstruction using 5, 50, and 100 SVD modes, respectively. Using five modes, only the most dominant features are reconstructed successfully, but as the number of modes used for reconstruction increases more of the finer local features in the original data are picked up. Similar results hold for both SPCA and NMF. **(b)** Compression percentage of the original data (%) as a function of the rank of the modes retained. For the 5, 50, and 100 modes illustrated in **(a)**, the data can be compressed into as little as 0.025 % for five modes, and 0.5 % for 100 modes.

Compression formats such as JPEG2000 are critically enabling for the electronics industry and allow for electronic devices to hold an exceptionally large number of video, audio, and image files.

Specifically, the compression advocated here is achieved by producing a low-rank representation for constructing the high-dimensional data, i.e., it should not be confused with standard data compression algorithms. The scalable decomposition methods advocated in this paper simply require a fraction of the data to be stored in the $\mathbf{Q}$ matrix and the rank-r embedding columns of $\widetilde{\boldsymbol{U}}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{V}$.

As an illustrative example, Fig. 11a shows a reconstruction of the absolute concentration of surface O_3 at a randomly selected time using the first 5, 50, and 100 of the SVD modes, respectively, as computed from the randomized algorithm. These reconstructions only require the storage of 0.025 %, 0.25 %, and 0.5 % of the data, respectively, as opposed to the 87 million data points of the original annual surface O_3 data (see Fig. 11b). The reconstruction with as few as five modes already shows that the dominant features are readily captured. It is also noted that there is virtually no difference between using 50 and 100 modes. The compression of the data with r modes can be computed from the first r columns of the $\boldsymbol{U}$ and $\boldsymbol{V}$ matrices along with the first r diagonal terms of $\boldsymbol{\Sigma}$. This gives a data compression ratio of $(m \times n)/(m \times r + r \times n + r)$ (see Fig. 3). The compression ratio is over 4000 for 5 modes, and approximately 200 for 100 modes.

This simple example shows that the compression of modes using our randomized architecture can serve as a critically enabling tool for the storage of numerical simulations and atmospheric chemistry data, with compression rates of up to a thousand fold. This allows the real-time analysis of simulations and data sets to be performed on laptop level computing platforms. Moreover, data can be much more easily shared for collaborative purposes as file sizes can be compressed from a terabyte to only a few hundred megabytes (5 modes) to a few gigabytes (100 modes). Such compression allows the data to be easily stored and shared on USB thumb drives.

In addition to data storage and diagnostics, the low-rank embedding spaces computed in our scalable algorithms can be used for projection-based reduced order models (ROMs) (Benner et al., 2015). ROMs are an important emerging computational framework for solving high-fidelity, complex systems in computationally tractable ways. ROMs are especially useful for enabling Monte Carlo simulations of high-dimensional systems that have stochastic variability, such as turbulent flows. The ROMs enable computation of statistical quantities like lift and drag in turbulent flows at fraction of the computational cost. Indeed, Monte Carlo computations of many high-dimensional problems of interest are currently intractable even with supercomputers; this highlights the need for proxy models that can be computed at a reduced cost. In future work, we will aim to develop ROMs that exploit the low-rank embeddings computed with our scalable algorithms.

6 Conclusions

Global environmental monitoring is becoming realizable through modern sensor technologies and emerging diagnostic algorithms. Despite tremendous advances and innovations, the data collection process can quickly produce volumes of data that cannot be analyzed and diagnosed in real-time, especially for applications like global atmospheric chemistry modeling which must integrate knowledge on hundreds of chemical species across a global longitude, latitude, and elevation grid. This emerging big data era requires diagnostic tools that can scale to meet the rapidly increasing information acquired from new monitoring technologies which are producing more fine-scale spatial and temporal measurements. We demonstrate a new set of diagnostic tools that are capable of extracting the dominant global features of global atmospheric chemistry dynamics. Not only are the methods scalable for both current and future sensor networks, they also have critical innovations allowing for improved interpretability, feature extraction, and data compression.

As demonstrated in this paper, emerging randomized linear algebra algorithms are critically enabling for scalable big data applications. The randomized algorithms exploit the fact that the data itself has low-rank features. Indeed, the method scales with the intrinsic rank of the dynamics rather than the dimension of the measurements/sensor space. Analysis of global atmospheric chemistry data shows that low-rank features indeed dominate the data. Thus, full spatial mode structures can be extracted (longitude, latitude, and elevation). This is in contrast to standard PCA reductions which do not scale well with the data size so that one is forced, due to computational constraints, to only analyze the data at fixed spatial features, such as only looking at a certain elevation. Alternatively, one can think of the scalable methods as being critically enabling for producing real-time analysis of emerging, streaming big data sets from the atmospheric chemistry community. Moreover, the dominant features of the data can be used for an efficient compression of the data for storage or reduced order modeling applications. Randomized tensor decompositions (Erichson et al., 2017b; Battaglino et al., 2018) are also viable for producing scalable diagnostic features of the global chemistry data. However, for the specific data considered here, little or no improvement was achieved. Nevertheless, in future work, we will consider such tensor decompositions across space, time, and chemicals where the randomized tensor decomposition is ideally suited for extracting higher-dimensional features.

An important aspect of this work is that simulation data, through the GEOS-Chem model, can be used to approximate the dominant global patterns of spatiotemporal activity for individual chemicals, a collection of chemicals, or the entire chemical space. The spatiotemporal features extracted provide new possibilities for understanding the interaction dynamics and relevant spatial regions where various chemical dynamics are important. This gives new possibilities for scientific discovery and the understanding of the complex processes driving the global chemistry profile.

Code and data availability. The code to run and generate the results presented in the paper, plot all the figures in the paper, and produce the three randomized matrix decompositions for O_3 is linked to the github account of Meghana Velegar: https://github.com/mvelegar/ScalableDiagnostics (last access: 1 April 2019; Velagar, 2019). The randomized algorithms used can be found on the github account of Benjamin Erichson: https://github.com/erichson/ristretto (last access: 1 April 2019; Erichson, 2019).

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/gmd-12-1525-2019-supplement.

Author contributions. MV, CK, and JNK designed the numerical and data extraction experiment, and MV carried them out. CK developed the GEOS-Chem model code and performed the simulations to generate the data, whereas MV and BE ran the algorithms and preprocessing steps required for the analysis. All authors contributed to the preparation of the paper.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. J. Nathan Kutz acknowledges support from the Air Force Office of Scientific Research (grant no. AFOSR FA9550-17-1-0329). Christoph A. Keller acknowledges support by the NASA Modeling, Analysis, and Prediction (MAP) program.

Review statement. This paper was edited by Patrick Jöckel and reviewed by two anonymous referees.

References

Avnery, S., Mauzerall, D. L., Liu, J., and Horowitz, L. W.: Global crop yield reductions due to surface ozone exposure: 1. Year 2000 crop production losses and economic damage, Atmos. Environ., 45, 2284–2296, https://doi.org/10.1016/j.atmosenv.2010.11.045, 2011.

Bey, I., Jacob, D. J., Yantosca, R. M., Logan, J. A., Field, B. D., Fiore, A. M., Li, Q., Liu, H. Y., Mickley, L. J., and Schultz, M. G. Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation, J. Geophys. Res., 106, 23073–23095, https://doi.org/10.1029/2001JD000807, 2001.

Battaglino, C., Ballard, G., and Kolda, T. G.: A practical randomized CP tensor decomposition, SIAM J. Matrix Anal. A., 39, 876–901, 2018.

Benner, P., Gugercin, S., and Willcox, K.: A survey of projection-based model reduction methods for parametric dynamical systems, SIAM Rev., 57, 483–531, 2015.

Bian, H. and Prather, M. J.: Fast-J2: Accurate Simulation of Stratospheric Photolysis in Global Chemical Models, J. Atmos. Chem., 41, 281–296, https://doi.org/10.1023/A:1014980619462, 2002.

Brasseur, G. P. and Jacob, D. J.: Modeling of Atmospheric Chemistry, Cambridge University Press, Cambridge, UK, 2017.

Cichocki, A. and Phan, A. H.: Fast Local Algorithms for Large Scale Nonnegative Matrix and Tensor Factorizations, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, E92.A, 708–721, 2009.

Cooper, M., Martin, R. V., Wespes, C., Coheur, P.-F., Clerbaux, C., and Murray, L. T.: Tropospheric nitric acid columns from the IASI satellite instrument interpreted with a chemical transport model: Implications for parameterizations of nitric oxide production by lightning, J. Geophys. Res.-Atmos., 119, 10068–10079, https://doi.org/10.1002/2014JD021907, 2014.

Cunningham, J. P. and Ghahramani, Z.: Linear dimensionality reduction: survey, insights, and generalizations, J. Mach. Learn. Res., 16, 2859–2900, 2015.

Drineas, P. and Mahoney, M. W.: RandNLA: randomized numerical linear algebra, Commun. ACM, 59, 80–90, 2016.

Eastham, S. D., Weisenstein, D. K., and Barrett, S. R.: Development and evaluation of the unified tropospheric–stratospheric chemistry extension (UCX) for the global chemistry-transport model GEOS-Chem, Atmos. Environ., 89, 52–63, https://doi.org/10.1016/j.atmosenv.2014.02.001, 2014.

Eastham, S. D., Long, M. S., Keller, C. A., Lundgren, E., Yantosca, R. M., Zhuang, J., Li, C., Lee, C. J., Yannetti, M., Auer, B. M., Clune, T. L., Kouatchou, J., Putman, W. M., Thompson, M. A., Trayanov, A. L., Molod, A. M., Martin, R. V., and Jacob, D. J.: GEOS-Chem High Performance (GCHP v11-02c): a next-generation implementation of the GEOS-Chem chemical transport model for massively parallel applications, Geosci. Model Dev., 11, 2941–2953, https://doi.org/10.5194/gmd-11-2941-2018, 2018.

Eckart, C. and Young, G.: The approximation of one matrix by another of lower rank, Psychometrika, 1, 211–218, 1936.

Erichson, N. B., Voronin, S., Brunton, S. L., and Kutz, J. N.: Randomized matrix decompositions using R, arXiv preprint, arXiv:1608.02148, 2016.

Erichson, N. B., Brunton, S. L., and Kutz, J. N.: Compressed singular value decomposition for image and video processing, in: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017, IEEE, 1880–1888, 2017a.

Erichson, N. B., Manohar, K., Brunton, S. L., and Kutz, J. N.: Randomized CP tensor decomposition, arXiv preprint, arXiv:1703.09074, 2017b.

Erichson, N. B., Mendible, A., Wihlborn, S., and Kutz, J. N.: Randomized Nonnegative Matrix Factorization, Pattern Recogn. Lett., 104, 1–7, 2018a.

Erichson, N. B., Zeng, P., Manohar, K., Brunton, S. L., Kutz, J. N., and Aravkin, A. Y.: Sparse Principal Component Analysis via Variable Projection, arXiv preprint, arXiv:1804.00341, 2018b.

Erichson, N. B.: Ristretto, available at: https://github.com/erichson/ristretto, last access: 15 April 2019.

Gillis, N.: Introduction to nonnegative matrix factorization, arXiv preprint arXiv: 1703.00663, 2017.

Gittens, A., Rothauge, K., Wang, S., Mahoney, M. W., Gerhardt, L., Kottalam, J., Ringenburg, M., and Maschhoff, K.: Accelerating Large-Scale Data Analysis by Offloading to High-Performance Computing Libraries using Alchemist, arXiv preprint, arXiv:1805.11800, 2018.

Halko, N., Martinsson, P.-G., and Tropp, J. A.: Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, SIAM Rev., 53, 217–288, 2011.

Hu, L., Keller, C. A., Long, M. S., Sherwen, T., Auer, B., Da Silva, A., Nielsen, J. E., Pawson, S., Thompson, M. A., Trayanov, A. L., Travis, K. R., Grange, S. K., Evans, M. J., and Jacob, D. J.: Global simulation of tropospheric chemistry at 12.5 km resolution: performance and evaluation of the GEOS-Chem chemical module (v10-1) within the NASA GEOS Earth system model (GEOS-5 ESM), Geosci. Model Dev., 11, 4603–4620, https://doi.org/10.5194/gmd-11-4603-2018, 2018.

Juntto, S. and Paatero, P.: Analysis of daily precipitation data by positive matrix factorization, Environmetrics, 5, 127–144, 1994.

Kutz, J. N.: Data-driven modeling & scientific computation: methods for complex systems & big data, Oxford University Press, Oxford, UK, 2013.

Kutz, J. N., Brunton, S. L., Brunton, B. W., and Proctor, J. L.: Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems, SIAM-Society for Industrial and Applied Mathematics, USA, 2016.

Lee, D. D. and Seung, S. H.: Learning the parts of objects by nonnegative matrix factorization, Nature, 401, 788–791, 1999.

Lee, E., Chan, C. K., and Paatero, P.: Application of positive matrix factorization in source apportionment of particulate pollutants in Hong Kong, Atmos. Environ., 33, 3201–3212, 1999.

Long, M. S., Yantosca, R., Nielsen, J. E., Keller, C. A., da Silva, A., Sulprizio, M. P., Pawson, S., and Jacob, D. J.: Development of a grid-independent GEOS-Chem chemical transport model (v9-02) as an atmospheric chemistry module for Earth system models, Geosci. Model Dev., 8, 595–602, https://doi.org/10.5194/gmd-8-595-2015, 2015.

Mahoney, M. W.: Randomized algorithms for matrices and data, Foundations and Trends in Machine Learning, 3, 123–224, 2011.

Mao, J., Jacob, D. J., Evans, M. J., Olson, J. R., Ren, X., Brune, W. H., Clair, J. M. St., Crounse, J. D., Spencer, K. M., Beaver, M. R., Wennberg, P. O., Cubison, M. J., Jimenez, J. L., Fried, A., Weibring, P., Walega, J. G., Hall, S. R., Weinheimer, A. J., Cohen, R. C., Chen, G., Crawford, J. H., McNaughton, C., Clarke, A. D., Jaeglé, L., Fisher, J. A., Yantosca, R. M., Le Sager, P., and Carouge, C.: Chemistry of hydrogen oxide radicals (HOx) in the Arctic troposphere in spring, Atmos. Chem. Phys., 10, 5823–5838, https://doi.org/10.5194/acp-10-5823-2010, 2010.

Mao, J., Paulot, F., Jacob, D. J., Cohen, R. C., Crounse, J. D., Wennberg, P. O., Keller, C. A., Hudman, R. C., Barkley, M. P., and Horowitz, L. W.: Ozone and organic nitrates over the eastern United States: Sensitivity to isoprene chemistry, J. Geophys. Res.-Atmos., 118, 11256–11268, https://doi.org/10.1002/jgrd.50817, 2013.

Mao, J., Carlton, A., Cohen, R. C., Brune, W. H., Brown, S. S., Wolfe, G. M., Jimenez, J. L., Pye, H. O. T., Lee Ng, N., Xu, L., McNeill, V. F., Tsigaridis, K., McDonald, B. C., Warneke, C., Guenther, A., Alvarado, M. J., de Gouw, J., Mickley, L. J., Leibensperger, E. M., Mathur, R., Nolte, C. G., Portmann, R. W., Unger, N., Tosca, M., and Horowitz, L. W.: Southeast Atmosphere Studies: learning from model-observation syntheses, Atmos. Chem. Phys., 18, 2615–2651, https://doi.org/10.5194/acp-18-2615-2018, 2018.

Martinsson, P.-G.: Randomized methods for matrix computations, arXiv preprint, arXiv:1607.01649, 2016.

Murray, L. T., Jacob, D. J., Logan, J. A., Hudman, R. C., and Koshak, W. J.: Optimized regional and interannual variability of lightning in a global chemical transport model constrained by LIS/OTD satellite data, J. Geophys. Res.-Atmos., 117, D20307, https://doi.org/10.1029/2012JD017934, 2012.

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

Parrella, J. P., Jacob, D. J., Liang, Q., Zhang, Y., Mickley, L. J., Miller, B., Evans, M. J., Yang, X., Pyle, J. A., Theys, N., and Van Roozendael, M.: Tropospheric bromine chemistry: implications for present and pre-industrial ozone and mercury, Atmos. Chem. Phys., 12, 6723–6740, https://doi.org/10.5194/acp-12-6723-2012, 2012.

Paterson, K. G., Sagady, J. L., Hooper, D. L., Bertman, S. B., Carroll, M. A., and Shepson, P. B.: Analysis of air quality data using positive matrix factorization, Environ. Sci. Technol., 33, 635–641, 1999.

Roberts, G., Wooster, M. J., and Lagoudakis, E.: Annual and diurnal african biomass burning temporal dynamics, Biogeosciences, 6, 849–866, https://doi.org/10.5194/bg-6-849-2009, 2009.

Rokhlin, V., Szlam, A., and Tygert, M.: A Randomized Algorithm for Principal Component Analysis, SIAM J. Matrix Anal. A., 31, 1100–1124, 2010.

Sherwen, T., Evans, M. J., Sommariva, R., Hollis, L. D. J., Ball, S. M., Monks, P. S., Reed, C., Carpenter, L. J., Lee, J. D., Forster, G., Bandy, B., Reeves, C. E., and Bloss, W. J.: Effects of halogens on European air-quality, Faraday Discuss., 200, 75–100, https://doi.org/10.1039/C7FD00026J, 2017.

Silva, R. A., West, J. J., Zhang, Y., Anenberg, S. C., Lamarque, J.-F., Shindell, D. T., Collins, W. J., Dalsoren, S., Faluvegi, G., Folberth, G., Horowitz, L. W., Nagashima, T., Naik, V., Rumbold, S., Skeie, R., Sudo, K., Takemura, T., Bergmann, D., Cameron-Smith, P., Cionni, I., Doherty, R. M., Eyring, V., Josse, B., MacKenzie, I. A., Plummer, D., Righi, M., Stevenson, D. S., Strode, S., Szopa, S., and Zeng, G.: Global premature mortality due to anthropogenic outdoor air pollution and the contribution of past climate change, Environ. Res. Lett., 8, 034005, https://doi.org/10.1088/1748-9326/8/3/034005, 2013.

Stevenson, D. S., Dentener, F. J., Schultz, M. G., Ellingsen, K., van Noije, T. P. C., Wild, O., Zeng, G., Amann, M., Atherton, C. S., Bell, N., Bergmann, D. J., Bey, I., Butler, T., Cofala, J., Collins, W. J., Derwent, R. G., Doherty, R. M., Drevet, J., Eskes, H. J., Fiore, A. M., Gauss, M., Hauglustaine, D. A., Horowitz, L. W., Isaksen, I. S. A., Krol, M. C., Lamarque, J.-F., Lawrence, M. G., Montanaro, V., Müller, J.-F., Pitari, G., Prather, M. J., Pyle, J. A., Rast, S., Rodriguez, J. M., Sanderson, M. G., Savage, N. H., Shindell, D. T., Strahan, S. E., Sudo, K., and Szopa, S.: Multimodel ensemble simulations of present-day and near-future tropospheric ozone, J. Geophys. Res., 111, D08301, https://doi.org/10.1029/2005JD006338, 2006.

Trendafilov, N., Jolliffe, I. T., and Uddin, M.: A modified principal component technique based on the LASSO, J. Comput. Graph. Stat., 12, 531–547, 2003.

Velagar, M.: Scalable Diagnostics, available at: https://github.com/mvelegar/ScalableDiagnostics, last access: 15 April 2019.

Voronin, S. and Martinsson, P.-G.: RSVDPACK: An implementation of randomized algorithms for computing the singular value, interpolative, and CUR decompositions of matrices on multi-core and GPU architectures, arXiv preprint, arXiv:1502.05366, 2015.

Xie, Y.-L., Hopke, P. K., Paatero, P., Barrie, L. A., and Li, S.-M.: Identification of Source Nature and Seasonal Variations of Arctic Aerosol bypositive matrix factorization, J. Atmos. Sci., 56, 249–260, 1999.

Zou, H. and Hastie, T.: Regularization and Variable Selection via the Elastic Net, J. R. Stat. Soc. B, 67, 301–320, 2003.