Reduced Order Model for Global Atmospheric Chemistry Data

Meghana Velegar

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

J. Nathan Kutz, Chair

Steven L. Brunton

Matthew Lorig

Program Authorized to Offer Degree:
Applied Mathematics

University of Washington

Abstract

Reduced Order Model for Global Atmospheric Chemistry Data

Meghana Velegar

Chair of the Supervisory Committee:
J. Nathan Kutz
Department of Applied Mathematics; Department of Electrical and Computer Engineering

Global 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 spatio-temporal evolution of these chemical constituents need to keep track of each chemical species on a global scale (longitude, latitude, elevation) and at each point in time. This data can be exceptionally high-dimensional so as to be not computationally tractable. Thus computationally scalable methods are required for the analysis, reproduction and forecasting of atmospheric chemistry dynamics. First, we introduce a new set of algorithmic tools capable of producing scalable, low-rank decompositions of global spatio-temporal 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 with improved interpretability. Importantly, our proposed algorithms scale with the intrinsic rank of the global chemistry space rather than measurement space, thus allowing for efficient representation and compression of the data. Next, we introduce the optimized dynamic mode decomposition algorithm for constructing an adaptive and computationally efficient reduced order model of global atmospheric chemistry dynamics. Forecasting is also achieved with a low-rank linear model that uses a linear superposition of the dominant spatio-temporal features. Bagging OPtimized DMD or BOP-DMD produces an ensemble of DMD models, thereby quantifying uncertainty, reducing model variance and suppressing over-fitting by

design. We compute the temporal uncertainty metrics for the optDMD forecasts using the BOP-DMD architecture. Lastly, we explore a data-driven scalable sparse sensor placement architecture for monitoring and reproduction of global atmospheric chemistry dynamics. By combining 1) machine learning, i.e. the POD dimensionality reduction technique, which learns and extracts a set of tailored library of features in the training data to produce low-dimensional representations of the full state, and 2) sparse sampling, i.e. designing highly specialized optimal sensors using the tailored features and QR pivoting, we reconstruct the full signal in the POD basis from a small subset of sensor or point measurements instantaneously. We also discover correlation between different chemical species, indicating that the chemical space can also be compressed.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I'm immensely grateful to have such a supportive and truly amazing advisor, Nathan. Nathan has been applauded by his students for his mentorship (both academically and on the personal work-life balance front), enthusiasm, wisdom, the relationships he cultivates with them, his vast knowledge of the field and his work ethics to name a few. Along with all of these, I am also indebted to Nathan for the grace he had with me as I navigated a high risk pregnancy, new motherhood, postpartum mental health issues, caring for a child with special needs, and finding time for productive research. Mine was definitely not a straightforward path towards my degree, and I know for a fact that I would not been here without Nathan as my advisor. I am truly blessed that he gave me a fair shake and took me on as his graduate student, advised and mentored me, had an unshakable faith in me and advocated for me always. I am also very thankful that he gave me the opportunity to work and play with such an interesting data set. This data presented me with several research avenues to explore and exciting data-driven analysis methods to learn and apply.

I would also like to thank Dr. Christoph A. Keller for making the global atmospheric chemistry simulation data available for my research, and for his invaluable contributions and feedback on my research. His vast knowledge of atmospheric chemistry research field and attention to detail was integral to the success of my work. A sincere thanks for my committee members Drs. Steve Brunton, Matthew Lorig and Marine Denolle for their thoughtful inputs and their time. I extend my thanks to all my professors in the Applied Math Department, they all enabled me to build a solid foundation that has helped me succeed. I would also like to acknowledge Lauren Lederer and Sarah Riley for their excellence as Graduate Program Advisors.

I am indebted to my dearest friend, mentor, and mother, Devender, for everything that I am and hope to be. A huge thank you to my father for cultivating in me an undying

curiosity and a pursuit of excellence in all of my endeavors. Many, many thanks to my awesome husband Sharat. He has shown me infallible love, support and patience while I navigated all the ups and many downs of my graduate school journey for the past decade plus. Much love and gratitude to my son Param, being your mom is one of the biggest blessings of my life. I am also very grateful to have the warmth and support of my aunt and uncle in-laws Mala and Shama, ever since I have been married into their family.

I am also very grateful to have the support of my therapist extraordinaire, Catie Heath, throughout my recent arduous journey back to work and finishing my degree. I would also like to thank my prescribing health caregivers, and especially my obstetrician Dr. Carol Salerno, for taking the time to navigate possibilities to identify my care needs. My warmest thanks to Blue Stiley, my good friend and coach, for his undying enthusiasm and verve. Heartfelt gratitude to Reen Stiley, one of my dearest friends, for her outstanding love and support for me, always. I have also been supported by my friend Harini Ramaprasad, and my network of mom friends at Shorenorth Coop preschool, who have cheered me on and have always been my phone-a-friend. Last but not least, thank you to my fur babies Albus, Rubeus and Ada, my little Patronous charms.

DEDICATION

For my family.

In loving memory of my father, Shyam Velegar.

Chapter 1

INTRODUCTION

The monitoring and forecasting of global atmospheric chemistry is critical for understanding the effects of air quality, chemistry-climate interactions, and global biogeochemical cycling. The dynamics of atmospheric chemistry is characterized by complex interactions among hundreds of chemical species which can produce kinetics across temporal scales spanning many orders of magnitude, from microseconds to minutes. Accurate monitoring and prediction requires full knowledge of the chemical state of the atmosphere at all locations and times, resulting in a 5-dimensional data set for longitude, latitude, elevation, species, relevant variables and time that can become massive as the resolution of each dimension is increased. Well resolved simulations generate massive data sets that are often not amenable to diagnostic analysis. For example, a single snapshot of the chemical state of an atmospheric chemistry model at $25 \times 25\,\mathrm{km}^2$ horizontal resolution requires 60 GB of storage space. Thus, this data can be exceptionally high-dimensional so as to be not computationally tractable. Efficient and computationally scalable methods are required for the analysis of atmospheric chemistry dynamics.

This thesis explores data-driven scalable reduced order modeling based methods for analysis, compression, accurate reproductions and stable forecasting of atmospheric chemistry simulation data. The goal of this work is to synthesize the global atmospheric chemistry simulation data into an accurate and computationally tractable Reduced Order Model (ROM) to approximate the high-dimensional nonlinear dynamical system and predict future states. We will be applying data-driven techniques to global atmospheric chemistry simulation data to build a Reduced Order Model (ROM) for the evolution of chemical species in the atmosphere. Our methods extract leading-order features of chemical concentrations for diagnostics, reconstruction and future state predictions.

1.1 Organization

An overview of the GEOS-Chem model simulation data and data preprocessing steps are presented in Chapter 2.

In Chapter 3 we introduce a suite of new set of algorithmic tools based on emerging *randomized linear algebra* algorithms capable of producing scalable, low-rank decompositions of global spatio-temporal atmospheric chemistry data. These decompositions 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 spatio-temporal measurement space, thus allowing for efficient representation and compression of the data. In addition to scalability, two additional innovations are proposed for improved interpretability: (i) a non-negative 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 thresholds 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 its significant improvement in computational speed and interpretability. We show that the here presented decomposition methods successfully extract known major features of atmospheric chemistry, such as summertime surface pollution and biomass burning activities.

In Chapter 4 we introduce the optimized dynamic mode decomposition algorithm for constructing an adaptive and computationally efficient reduced order model of global atmospheric chemistry dynamics. By exploiting a low-dimensional set of global spatio-temporal modes, interpretable characterizations of the underlying spatial and temporal scales can be computed. The dimensionality reduction DMD methods are demonstrated on three months of global chemistry dynamics data. We show that the presented decomposition methods successfully extract known major features of atmospheric chemistry, such as summertime surface pollution. Stable forecasting capabilities are also achieved with the low-rank linear DMD model. Temporal uncertainty in the forecasts is quantified by using the statistical bagging BOP-DMD architecture. Moreover, the DMD algorithm allows for

rapid reconstruction of the DMD model, which can then easily accommodate non-stationary data and changes in the dynamics.

Finally, we conclude with exploring optimized sensor placement based on tailored low-rank libraries of features extracted from training data in Chapter 5. The POD dimensionality reduction technique is applied to high-dimensional surface chemical species training data to construct a library of POD basis modes, that are tailored to produce low-dimensional representations of the data. We then apply the extremely efficient QR pivoting algorithm to the library of tailored POD basis modes that enables optimized placement of sparse spatial sensors. A small subset of point measurements from these sparse sensors is able to reconstruct the full state of absolute concentrations of chemical species. For the rate of change of absolute concentration data, however, incremental updates to the POD basis library were shown to be helpful to accurately reconstruct the full state. We also discovered that compression in chemical space can be achieved, i.e., the same set of sensor locations can be used for point measurements of different chemical species to reproduce their full states accurately.

1.2 Published work

Part of the work presented in this thesis has resulted in the following peer-reviewed publication. The majority of textual and visual content in Chapters 2 and 3 is reproduced with permission from the following:

Meghana Velegar, N Benjamin Erichson, Christoph A Keller, and J Nathan Kutz. Scalable diagnostics for global atmospheric chemistry using ristretto library (version 1.0). Geoscientific Model Development, 12(4):1525-1539, 2019.

Chapter 2

GLOBAL ATMOSPHERIC CHEMISTRY SIMULATION DATA

The GEOS-Chem global atmospheric chemistry model simulates atmospheric chemistry by solving 3-D continuity equations for the concentrations of chemical species including the effects of emissions, transport, chemistry, and deposition. This is commonly done with chemical transport models (CTMs) driven by input meteorological data and surface boundary conditions. Earth system models (ESMs) calculate the ensemble of processes affecting the Earth system prognostically. An Earth System Modeling Framework (ESMF) interface can couple the atmospheric chemistry from CTMs to atmospheric dynamics provided by ESMs. The GEOS-Chem global chemical transport model (CTM) has been re-engineered to serve as an atmospheric chemistry module for Earth system models (ESMs) by Long et. al. [42]. The ESMF capability was deployed in the NASA Goddard Earth Observing System (GEOS) developed at NASA's Global Modeling and Assimilation Office (GMAO). We will be working with the global atmospheric chemistry simulation data generated by this coupled GEOS-Chem CTM and ESM implementation.

The GEOS-chem atmospheric chemistry models simulates data on a global discretized mesh of longitude, latitude and elevation, as shown in Figure 2.1. In this work, we are specifically concerned with time-series measurements of the concentrations (**START** data) and rates of change of concentrations (**TEND** data) of chemical species, for which a new snapshot is recorded every 20 minutes for each of the grid cells. Section 2.1 presents an overview of this global atmospheric chemistry model used to simulate the dynamics. Section 2.2 describes the simulation engine used to produce the data of interest. To keep the development of the ROM tractable, we focus on a few chemical species of interest, enumerated in Section 2.3. We used efficient data slicing and organizing to extract the time series data of these chemicals for the absolute concentrations and rates of change of concentrations, resulting in the original data matrix shown in Figure 2.1. The last

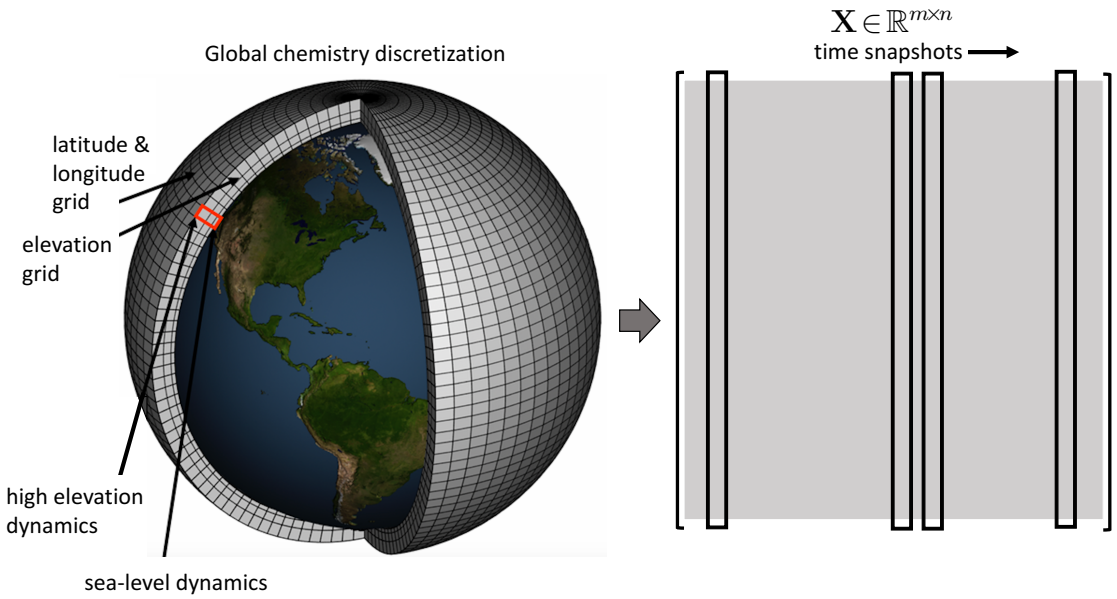


Figure 2.1: *Atmospheric chemistry simulation on a global mesh with discretized longitude, latitude and elevation (left panel modified from NOAA). Each illustrated grid cell contains time-series data for the atmospheric chemistry dynamics. Shown in the right panel is the original data matrix* $\mathbf{X} \in \mathbb{R}^{m \times n}$*, where* m *is the number of grid points and* n *is the number of snapshots.*

Section 2.4 describes a critical preprocessing step needed to successfully apply the scalable dimensionality reduction techniques in this work.

2.1 Atmospheric chemistry model GEOS-Chem

Chemical transport models (CTM) are used to simulate the evolution of atmospheric constituents in space and time [20]. A CTM solves the system of coupled continuity equations for an ensemble of m species with number density vector $\mathbf{n} = (n_1, \ldots, n_m)^T$ via operator splitting of transport and local processes:

$$\frac{\partial n_i}{\partial t} = -\nabla \cdot (n_i \mathbf{U}) + (P_i - L_i)(\mathbf{n}) + E_i - D_i \qquad i \in [1, m] \tag{2.1}$$

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

$$\frac{\partial n_i}{\partial t} = -\nabla \cdot (n_i \mathbf{U}) \qquad i \in [1, m] \tag{2.2}$$

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

$$\frac{dn_i}{dt} = (P_i - L_i)(\mathbf{n}) + E_i - D_i \qquad i \in [1, m] \tag{2.3}$$

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

Chemistry models repeatedly solve equations 2.2 and 2.3, which requires full knowledge of the chemical state of the atmosphere at all locations and times. The resulting 4-dimensional data sets (longitude, latitude, levels, and species) can become massive, which makes it unpractical to output them at high temporal frequency. As a consequence, model output is generally restricted to a few selected species of interest (e.g. ozone), while the full model state is only output very infrequently, e.g. to archive the information for future model restarts ('restart file'). We show here 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 here on identifying low-ranked features across the spatio-temporal dimension (i.e., for each species separately) the presented methods could similarly (and independently) be applied across the species domain. The life times of the involved chemical species range from milliseconds to years, which makes the ODE system computationally stiff. Numerical integration of this set of stiff ODEs requires implicit integration schemes such as Rosenbrock solvers to ensure optimal numerical stability [108, 107].

Since the model integration must be performed repeatedly at every spatial grid point and for all time steps, atmospheric chemistry models are computationally expensive. The major computational costs are related to (a) transport and (b) numerical integration of the chemically coupled ODEs. The chemical operator is highly parallelizable since it does

not involve spatial coupling [79, 42]. In contrast, the relative computational cost for transport increases with increasing horizontal resolution and becomes at least as important as chemistry for high-resolution (≤ 25 km^2 model applications) [61].

As we show in this work, 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. Since low-ranked features can be identified across the spatio-temporal dimension as well as the species domain, this method offers the potential to address both the transport and the chemical term of Equation (2.1).

2.2 Global Atmospheric Chemistry Simulations

The reference simulation of atmospheric chemistry was generated using the GEOS-Chem model. GEOS-Chem (`http://geos-chem.org`) 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`). GEOS-Chem can be run in offline mode as a chemical transport model (CTM) [1, 42] or as an online component within the NASA Goddard Earth System Model (GEOS) [79, 61]. We use here the offline 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 HOx-NOx-VOC-ozone-BrOx tropospheric chemistry as originally described by [1] and with addition of BrOx chemistry by [96] and updates to isoprene oxidation as described by [89]. Dynamic and chemical time steps are 30 and 20 minutes, respectively. Stratospheric chemistry is modelled using a linearized mechanism as described by [91].

We performed a one-year simulation of GEOS-Chem (July 2013 - June 2014) at $4^\circ \times 5^\circ$ 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} (**START** data). The difference between these concentration pairs are the species tendencies

due to chemistry (expressed in units of molecules $\mathrm{cm}^{-3}\,\mathrm{s}^{-1}$. This is the **TEND** data).

Since 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 [17] as implemented in GEOS-Chem by [88] and [43]. At every time step, the data set thus consists of n 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 dimensions nlon × nlat × nlev × ntimes × nfeatures = $72 \times 46 \times 30 \times 26280 \times 380 = 9.9 \times 10^{11}$.

2.3 Six chemical species studied in this work

In particular, these six chemical species' dynamics are of interest to Atmospheric chemists:

1. **Ozone (O3)** is the elephant in the room. It is critical in the upper atmosphere because it shields earth from UV radiation, but toxic near the surface. Surface concentrations are highly regulated. Studies show that reducing average surface ozone by just one ppb (2-4%) increases economic output by roughly 1 billion US dollar thanks to higher crop yields and reduced human health impacts – in the US alone.

2. **Nitrogen oxides (NO and NO2)**. These two are basically twins because they are tightly coupled chemically. They are one of the critical component of ozone production and the one that can be controlled because some of the main sources of NO and NO2 are emissions from power plants and road traffic.

3. The **hydroxyl radical OH** is the holy grail. It is the cleansing agent of the atmosphere that can react with almost everything, but it's extremely short-lived and almost impossible to measure. True **OH** concentration dynamics are unknown, all the reactions that determine formation/destruction of **OH** are also unknown.

4. **Isoprene (ISOP)**: primarily emitted from trees and an important precursor to the formation of ozone. It's reaction mechanisms are very complicated and a big part

of the total chemistry mechanism is centered around isoprene chemistry. However, isoprene lifetime is short, hence all these reactions may be irrelevant in many parts of the atmosphere.

5. **Carbon monoxide (CO)**: a presumably easy chemical compound to compute. It has a few, relatively well known emission sources, only one sink (reaction with the **OH** radical). However, all chemistry models are having problems with determining **CO** concentrations, and it's not quite clear why.

A detailed analysis is done for these six chemical species. 2. Dynamic Mode Decomposition (DMD): Applying DMD algorithms to capitalize on the low-order dynamics for future state predictions. 3. POD/DMD analysis in chemical space: Identifying key chemicals in the dynamics, and correlations between chemical species of interest to isolate their dynamics and gain better understanding of the underlying chemistry. 4. POD/DMD analysis in physical space: Identifying key geographic locations that influence the dynamics globally.

2.4 Data preprocessing

In this work, we are specifically concerned with time-series measurements of the concentration of chemical species collected from spatial locations in the atmosphere. As discussed above, the year-long simulation data is exceptionally high-dimensional and massive, hence applying the ROM models required careful considerations of preprocessing the data. First, an efficient slicing and reorganization of the data was carried out, isolating the data sets of interest.

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 the *singular value decomposition* (SVD) which allows one to exploit covariances manifest in the data [34].

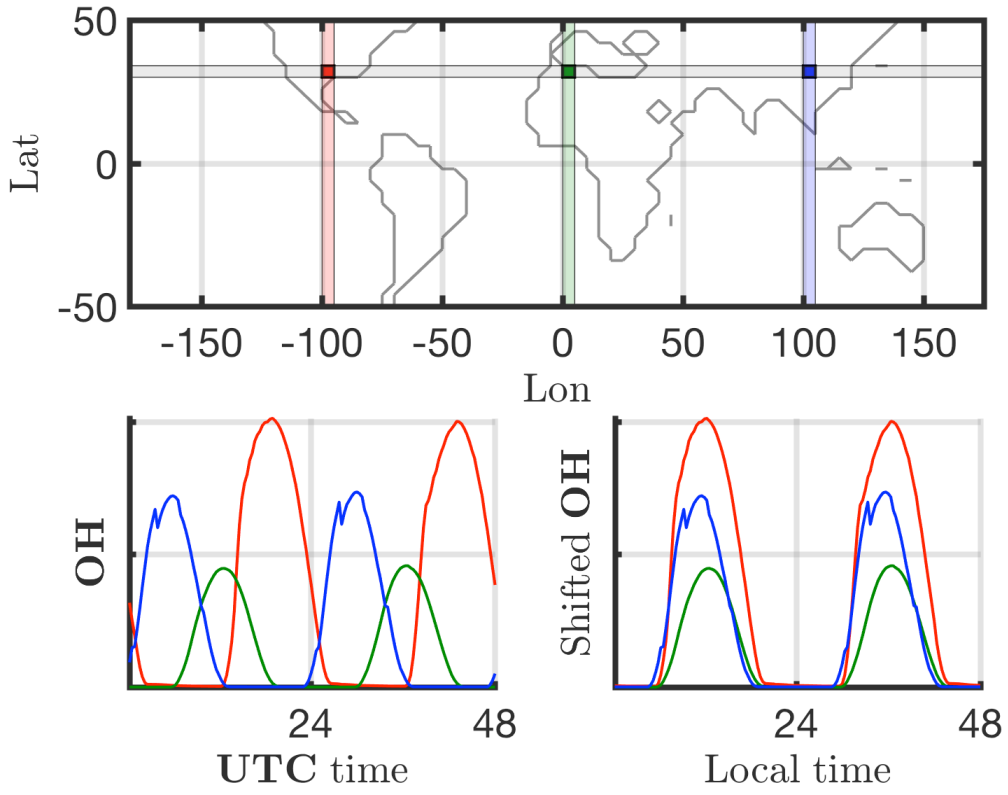


Figure 2.2: *Shifting the data for each cell in time to align the local time zones across a latitude to the prime meridian*(Lon = 0) *local time, shown here for* **OH** *absolute concentration for* Lat = 30

A fundamental weakness of such SVD-based approaches is the inability to efficiently handle invariances in the data. Specifically, translational and/or rotational invariances of low-rank features in the data are not well captured [67, 68, 24]. One of the key environmental variables driving the chemistry is photolysis rate, the absolute concentrations of many chemicals of interest accordingly 'turn on' and are non-zero 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 the bottom left in Figure. 2.2, 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., higher number of modes would be needed to characterize the dynamics due to translation [67]. To overcome this issue, we need to factor out or remove the translational invariance. The time series for each grid point are thus shifted to align with the GMT time, as shown on the bottom right in Figure. 2.2. With the local times for each grid point aligned, the data is aligned and SVD-based dimensionality reduction techniques can now identify and isolate coherent low-dimensional features in the data, as we show in the subsequent chapters.

Chapter 3

SCALABLE DIAGNOSTICS FOR GLOBAL ATMOSPHERIC CHEMISTRY USING RISTRETTO LIBRARY (VERSION 1.0)

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 datasets has severely challenged our ability to analyze data using traditional matrix algorithms, especially for ever-increasing refinements of computational models. 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 [55, 80, 38, 49, 46]here to extract the dominant, low-rank mode structures from a full three-dimensional data set of atmospheric chemistry. 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. They further offer an alternative approach for 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, [53] can compute the SVD of a 2.2 TB terabyte data-set in about 60 seconds, given a super computer with many nodes. However, if super computing is not available, the randomized method offers an attractive alternative which does not require expensive compute hours on a cluster.

The chapter is outlined as follows: Section. 3.1 highlights the various decomposition methods that can be produced using randomized linear algebra techniques. Section 3.2 presents the results of the dimensionality reduction procedures, highlighting the effectiveness of each technique. Section 3.3 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 3.4 provides concluding remarks and a brief outlook for data sciences applied to atmospheric dynamics and global chemistry analysis.

3.1 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 spatio-temporal 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 spatio-temporal 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 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 since 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 application intended. Regardless, the suite of methods allows for a more nuanced view of the data.

3.1.1 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

$$\begin{array}{ccccc} \mathbf{X} & \approx & \mathbf{E} & \mathbf{F} \\ m \times n & & m \times r & r \times n \end{array} \tag{3.1}$$

where the columns of $\mathbf{E}$ spans the column space of $\mathbf{X}$, and the rows of $\mathbf{F}$ spans 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 two 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 [55]. 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

$$\mathbf{X} \approx \mathbf{Q}\mathbf{Q}^{\mathsf{T}}\mathbf{X} \tag{3.2}$$

  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, constructed by first drawing a set of k independent random vectors $\{\boldsymbol{\omega}_i\}_{i=1}^{k}$, for instance, from the standard normal distribution; then mapping $\mathbf{X}$ 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 from the sampling matrix $\mathbf{Y} \in \mathbb{R}^{m \times k}$ are given by

$$\mathbf{Y} := \mathbf{X}\boldsymbol{\Omega} \tag{3.3}$$

$\mathbf{Y}$ 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 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:

- *Oversampling*: For target rank k, for most data matrices we may have non-zero 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 [90].
- *Power iteration scheme*: The quality of $\mathbf{Q}$ can be improved by the concept of power sampling iterations [55, 104]. 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 $\mathbf{B}$

$$\mathbf{B} := \mathbf{Q}^{\mathsf{T}}\mathbf{X} \in \mathbb{R}^{l \times n} \tag{3.4}$$

i.e., 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

$$\begin{array}{ccccc} \mathbf{X} & \approx & \mathbf{Q} & \mathbf{B} \\ m \times n & & m \times l & l \times n \end{array} \tag{3.5}$$

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 [123].

3.1.2 Randomized Singular Value Decomposition

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

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}} \tag{3.6}$$

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 $\mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{I}$. 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 non-negative 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$\{\sigma_i : i \geq r + 1\}$ are zero. The corresponding singular vectors span the left and right null spaces of the matrix. In practical applications, the data matrix is 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

$$\begin{aligned}\mathbf{X}_k :&= \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k \\ &= [\mathbf{u}_1, \ldots, \mathbf{u}_k] \ \mathrm{diag}\left(\sigma_1, \ldots, \sigma_k\right) \ [\mathbf{v}_1, \ldots, \mathbf{v}_k]^{\mathsf{T}}\end{aligned} \tag{3.7}$$

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. On the other hand, if a very low dimensional representation of dominant features is desired, then k might be chosen to be much smaller. The Eckart-Young theorem [44] states that the low-rank SVD provides the optimal rank-k reconstruction of a matrix in the least-squares sense

$$\mathbf{X}_k := \underset{\mathrm{rank}(\mathbf{X}_k')}{\mathrm{argmin}} \left\| \mathbf{X} - \mathbf{X}_k' \right\| \tag{3.8}$$

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

$$\left\| \mathbf{X} - \mathbf{X}_k \right\|_2 = \sigma_{k+1}\left(\mathbf{X}\right) \tag{3.9}$$

and

$$\left\| \mathbf{X} - \mathbf{X}_k \right\|_F = \sqrt{\sum_{j=k+1}^{\min(m,n)} \sigma_j^2\left(\mathbf{X}\right)} \tag{3.10}$$

For massive datasets, 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 computational efficient and 'surprisingly' reliable, these techniques can be used to obtain an approximate rank-k SVD at a substantially more efficient cost of $O\left(mnk\right)$.

The probabilistic framework is used to obtain a near-optimal low rank approximation $\mathbf{B} \in \mathbb{R}^{l \times n}$, where $l \ll \min\left(m, n\right)$. This can now be used instead of the data matrix $\mathbf{X}$, and a full SVD of $\mathbf{B}$ is computed

$$\mathbf{B} = \tilde{\mathbf{U}} \mathbf{\Sigma} \mathbf{V}^{\mathsf{T}} \tag{3.11}$$

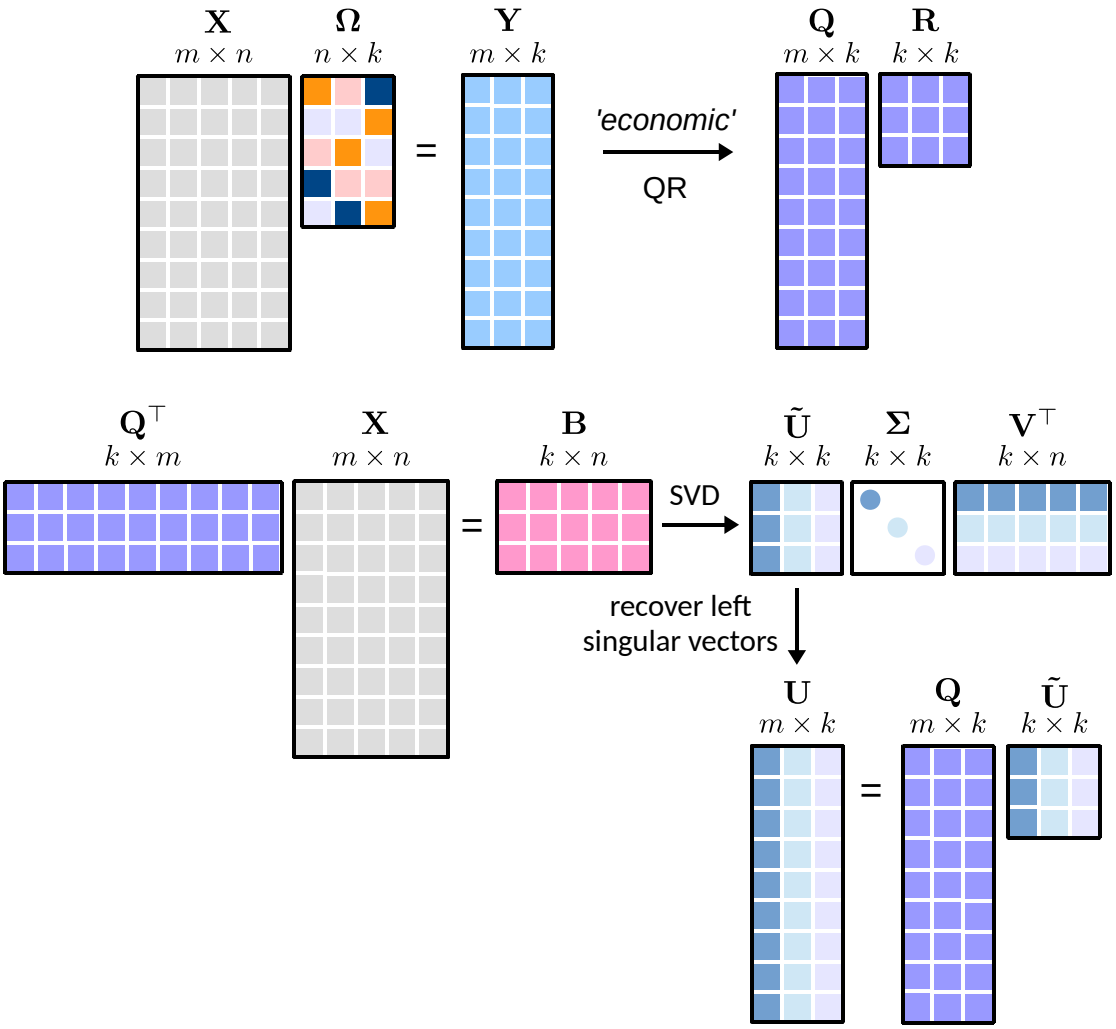


Figure 3.1: *Illustration of the randomized matrix decomposition technique. The random sampling matrix* $\mathbf{\Omega}$ *is used to produce a new matrix* $\mathbf{Y}$ *which can be decomposed using a QR decomposition. This leads to the construction of the matrix,* $\mathbf{B}$ *which is used for approximating the left and right singular vectors.*

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

$$\mathbf{U} \approx \mathbf{Q}\tilde{\mathbf{U}} \tag{3.12}$$

For the absolute concentration data matrix, note that the right singular vectors $\mathbf{V}$ are temporal, and the left singular vectors $\mathbf{U}$ 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:

$$\frac{\sum_{i=1}^{j} \sigma_i^2}{\text{Total Energy in the Data}} \tag{3.13}$$

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

The algorithm architecture is conceptually outlined in Figure 3.1. This shows the basic architecture and the structure, which allows for a rapid approximation of the left and right singular values and eigenvectors.

3.1.3 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 for the absolute concentration of chemical species data, 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 [71, 93], this paradigm is the nonnegative matrix factorization (NMF). NMF has emerged as a powerful dimension reduction tool that allows computation of 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

$$\begin{array}{ccccc} \mathbf{X} & \approx & \mathbf{W} & & \mathbf{H} \\ m \times n & & m \times k & & k \times n \end{array} \tag{3.14}$$

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 on the other hand gives an additive parts-based representation of the data that preserves useful properties such as sparsity and nonnegativity by imposing additional nonnegativity constraints: $\mathbf{W} \geq \mathbf{0}$ and $\mathbf{H} \geq \mathbf{0}$. The sparse parts-based features have an intuitive interpretation, which have been exploited in environmental modeling [93]. In environmental data, the error estimates of data can be widely varying and non-negativity is often an essential feature of the underlying models [64, 72, 97, 124]. For example, the NMF components of a face image dataset reveal individual features such as the nose and the mouth, whereas PCA components yield holistic features known as *eigenfaces*.

Traditionally, the NMF problem is formulated as the following optimization problem:

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{W}, \mathbf{H}) \quad = \quad \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 \\ &\text{subject to} \quad \mathbf{W} \geq \mathbf{0} \quad \text{and} \quad \mathbf{H} \geq \mathbf{0} \end{aligned} \tag{3.15}$$

This optimization problem is nonconvex and ill-posed. Since no convexification exists to simplify the optimization, no exact or unique solution is guaranteed [52]. Different NMF algorithms, therefore, can produce distinct decompositions that minimize the objective function. Since the problem is nonconvex with respect to both factors $\mathbf{W}$ and $\mathbf{H}$, most NMF algorithms divide the problem into simpler subproblems which have closed form solutions. The convex subproblem is solved by keeping one factor fixed while updating the other, alternating and iterating until convergence. The Hierarchical Alternating Least Squares (HALS) is one variant of this method, proved to be highly efficient [32], 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 Equation 3.15 with respect to the remaining $k-1$ components

$$\text{minimize } J_j\left(\mathbf{W}_{(:,j)}, \mathbf{H}_{(j,:)}\right) = \left\|\mathbf{R}^{(j)} - \mathbf{W}_{(:,j)}\mathbf{H}_{(j,:)}\right\|_F^2, \tag{3.16}$$

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

$$\mathbf{R}^{(j)} := \mathbf{X} - \sum_{i \neq j}^{k} \mathbf{W}_{(:,i)} \mathbf{H}_{(i,:)} \tag{3.17}$$

Gradients are derived to find the stationary points for both components, for details see [32].

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 $\mathbf{\Omega}$ are drawn independently from the uniform distribution with support $\omega \in [0, 1]$. We now have the following optimization problem:

$$\begin{array}{lllll} \text{minimize} & \tilde{f}\left(\tilde{\mathbf{W}}, \mathbf{H}\right) & = & \left\| \mathbf{B} - \tilde{\mathbf{W}}\mathbf{H} \right\|_F^2 \\ \text{subject to} & \mathbf{Q}\tilde{\mathbf{W}} \geq \mathbf{0} & \text{and} & \mathbf{H} \geq \mathbf{0} \end{array} \tag{3.18}$$

where the nonnegativity constraints need apply to the high dimensional factor matrix $\mathbf{W}$, but not necessarily to $\tilde{\mathbf{W}}$, since $\tilde{\mathbf{W}}$ can be rotated back to high dimensional space using the approximate relation $\mathbf{W} \approx \mathbf{Q}\tilde{\mathbf{W}}$. Since $\mathbf{Q}\mathbf{Q}^{\mathsf{T}} \neq \mathbf{I}$, Equation 3.18 can only be solved approximately. The randomized HALS algorithm is formulated as

$$\text{minimize } J_j\left(\tilde{\mathbf{W}}_{(:,j)}, \mathbf{H}_{(j,:)}\right) = \left\| \tilde{\mathbf{R}}^{(j)} - \tilde{\mathbf{W}}_{(:,j)} \mathbf{H}_{(j,:)} \right\|_F^2, \tag{3.19}$$

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

$$\tilde{\mathbf{R}}^{(j)} := \mathbf{B} - \sum_{i \neq j}^{k} \tilde{\mathbf{W}}_{(:,i)} \mathbf{H}_{(i,:)} \tag{3.20}$$

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 [48].

For the absolute chemistry concentration data matrix, the columns of the factor $\mathbf{W}$ are the spatial modes while those of the factor $\mathbf{H}$ are the temporal modes. The randomized NMF algorithm starts with an initial guess derived from an 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, 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 Equation 3.13.

3.1.4 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

$$\mathbf{z}_i = \mathbf{X}\mathbf{w}_i \tag{3.21}$$

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

$$\mathbf{Z} = \mathbf{X}\mathbf{W} \tag{3.22}$$

with $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{m \times n}$ and $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_n] \in \mathbb{R}^{n \times n}$. 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

$$\begin{aligned} \underset{\mathbf{W}}{\text{minimize}} f(\mathbf{W}) &= \tfrac{1}{2} \left\| \mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{W}^{\mathsf{T}} \right\|_{\mathrm{F}}^2 \\ \text{subject to } \mathbf{W}^{\mathsf{T}}\mathbf{W} &= \mathbf{I} \end{aligned} \tag{3.23}$$

The classic PCA approach outlined above generates global PCs as a linear combination of all n variables, hence tends to often mix or blend various spatio-temporal 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 $\mathbf{w}_i$ have only a few non-zero values, hence 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 Equation 3.23 as

$$\begin{aligned} \underset{\mathbf{A},\mathbf{W}}{\text{minimize}} f(\mathbf{A},\mathbf{W}) &= \tfrac{1}{2}\left\|\mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{A}^{\mathsf{T}}\right\|_{\mathrm{F}}^{2} + \psi(\mathbf{W}) \\ \text{subject to } \mathbf{A}^{\mathsf{T}}\mathbf{A} &= \mathbf{I} \end{aligned} \tag{3.24}$$

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 $\widetilde{\mathbf{X}} = \mathbf{Z}\mathbf{A}^{\mathsf{T}}$, where $\mathbf{Z}$ is the PC matrix given by Equation 3.22. In Equation 3.24, ψ is a sparsity inducing regularizer such as

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

$$\psi_0(\mathbf{x}) = \|\mathbf{x}\|_0 \tag{3.25}$$

- ℓ_1 norm, in this case the regularization problem is also known as LASSO (Least Absolute Shrinkage and Selection Operator) [120]

$$\psi_1(\mathbf{x}) = \alpha\|\mathbf{x}\|_1 \tag{3.26}$$

  where α controls the degree of sparsity

- The elastic net [127] which is a combination of the ℓ_1 norm and quadratic penalty

$$\psi_{\mathrm{E}}(\mathbf{x}) = \alpha\|\mathbf{x}\|_1 + \beta\|\mathbf{x}\|_2^2 \tag{3.27}$$

  where α, β control the degree of sparsity

Note that the optimization problem in Equation 3.24 is nonconvex and is solved similar to the NMF optimization problem by keeping one factor fixed while updating the other, alternating and iterating till convergence. The SPCA formulation defined in Equation 3.24 can be generalized to

$$\underset{\mathbf{A},\mathbf{W}}{\text{minimize}} f\left(\mathbf{A},\mathbf{W}\right)=\rho\left(\mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{A}^{\mathsf{T}}\right)+\psi\left(\mathbf{W}\right)+\varphi\left(\mathbf{A}\right) \tag{3.28}$$

where ρ is a measure of fit (for example, the least-squares problem with the Frobenius norm), ψ is a sparsity regularizer and φ is a constraint, such as the orthogonality constraint. The measure of fit can be refined to be other robust measures (for example, the Huber norm). For further details, refer to [50].

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 Equation 3.3 so that we can reformulate Equation 3.24 as

$$\begin{aligned} &\underset{\mathbf{A},\mathbf{W}}{\text{minimize}} f\left(\mathbf{A},\mathbf{W}\right)=\tfrac{1}{2}\left\|\widetilde{\mathbf{X}}-\widetilde{\mathbf{X}}\mathbf{W}\mathbf{A}^{\mathsf{T}}\right\|_{\mathrm{F}}^{2}+\psi\left(\mathbf{W}\right) \\ &\text{subject to } \mathbf{A}^{\mathsf{T}}\mathbf{A} = \mathbf{I} \end{aligned} \tag{3.29}$$

The absolute concentration data matrix is first scaled to have mean 0. The spatial modes are the columns of the matrix $\mathbf{W}$. The temporal modes or the PCs are the columns of $\mathbf{Z}$ computed from $\mathbf{X}=\mathbf{Z}\mathbf{A}^{\mathsf{T}}$. The minimization algorithm also formulates the problem as an eigen value problem, and returns the eigen values λ_j associated with the j^{th} 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:

$$\frac{\sum_{i=1}^{j}\lambda_i\times(n-1)}{\text{Total Energy in the scaled Data}} \tag{3.30}$$

where n is the total number of snapshots in time.

3.2 Data diagnostics

In this section we illustrate results from the decomposition of the GEOS-Chem model output using absolute concentration i.e. **START** data of ozone ($\mathbf{O_3}$) as an example. The additional

five chemical species, including Nitrous oxide **NO**, Nitrous dioxide $\mathbf{NO_2}$, Hydroxyl ion **OH**, Isoprene **ISOP** and Carbon monoxide **CO**, are known to be equally important to ozone. For succinctness of the manuscript we only present ozone here, the supplementary materials for [122] provide diagnostics for five additional chemicals known to dominate the global atmospheric chemistry dynamics. Overall, there are close to two hundred chemicals that are interacting dynamically. Each chemical of interest can be diagnostic in a similar fashion to ozone, in order to determine its dominant global variability. It remains an open research question how the interactions across the entire chemical space ultimately drive the observed variability. The scalable diagnostics advocated here provides a computational architecture allowing scientists to explore this further by providing global diagnostics for all chemicals in a computationally tractable manner.

Ozone is a key oxidant of the atmosphere, and high surface concentrations of $\mathbf{O_3}$ are harmful to human health and vegetation [6, 115]. Ozone production involves the photochemical oxidation of volatile organic compounds (VOCs) and carbon monoxide (**CO**) in the presence of nitrogen oxide radicals ($\mathbf{NOx \equiv NO + NO_2}$). The chemistry of ozone is highly complex, involving hundreds of chemical species. This makes ozone a challenging compound for chemistry models, e.g. [117, 114, 87]. We find that despite the underlying complexity of the chemistry, the ozone 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 $= 26208$ for the year-long data (one snapshot every 20 minutes).

3.2.1 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, absolute concentration values of nitric oxide (**NO**) are several orders of magnitude higher over China and eastern Russia, as compared to those over oceans and less populated regions in the world. Correspondingly, the dominant spatial modes are very localized as exhibited in the top

panel of Figure 3.2, 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 $\mathbf{X}_{\log} = \log(\mathbf{X} + 1)$. The second most dominant mode of the logarithm of the data as shown in the bottom panel of Figure 3.2 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 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, the large spike in the data is so large that the rest of the data is like noise comparatively. By normalizing with the logarithm, a more balanced global view of the chemistry dynamics can be extracted from the modal structures.

3.2.2 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 ozone ($\mathbf{O_3}$). These are presented in the top panel of Figure 3.3. The amount of energy explained by the most dominant singular values gives a good indication about the low-rank nature of the underlying data. The top panel of Figure 3.3 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%

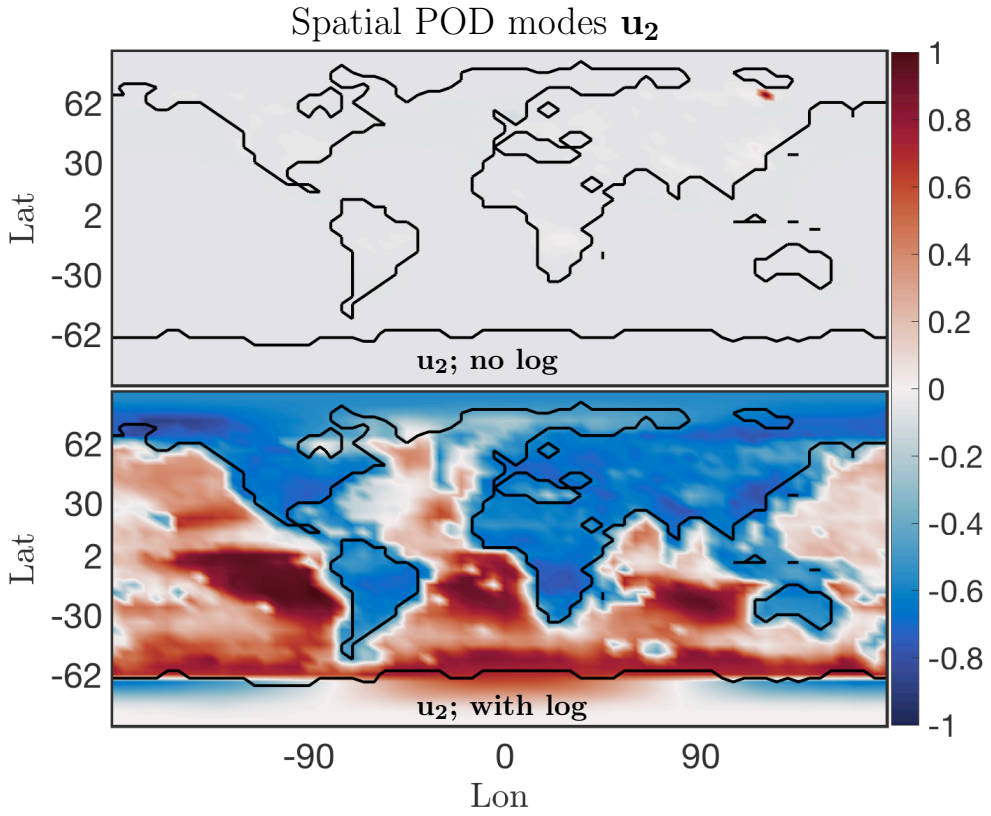


Figure 3.2: *Dominant spatial mode 2 at surface for* **NO** *absolute concentration preprocessed data before and after taking a logarithm of the preprocessed data. Taking a logarithm scales the preprocessed data so that the corresponding spatial modes exhibit the global low dimensional features, instead of only picking up on the dominant chemistry in one localized region.*

of the total energy. Thus, it is possible to explain 99% of the spatio-temporal structure of the highly complex ozone field with just 20 modes. These modes reveal many of the dominant features of atmospheric ozone. The bottom panel of Figure 3.3 illustrates the structure of the 4 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 ozone. Modes 2 and 3 (red and yellow, respectively) both exhibit a frequency of 1 year, capturing features occurring on an annual basis. The 4th most dominant mode (purple) has a frequency of 6 months. Geophysical interpretation of these modes is easiest

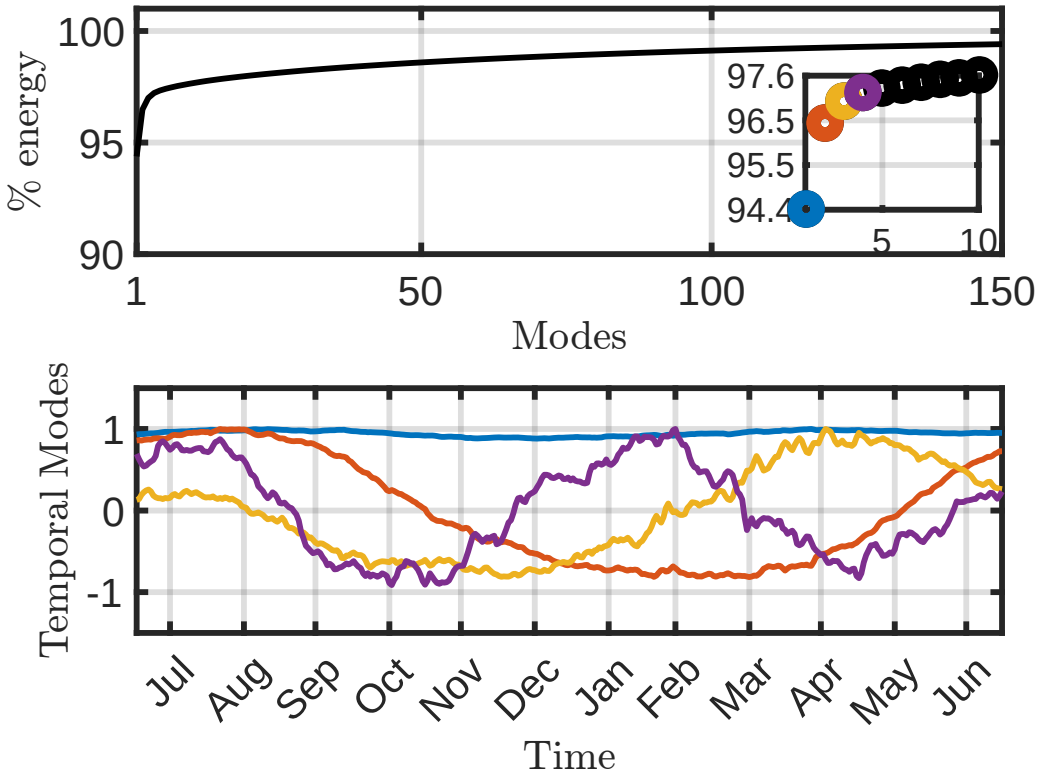


Figure 3.3: *Cumulative energy spectrum (and inset detail) of the Singular Value Decomposition (top) and the corresponding 4 dominant temporal modes (bottom) for* $\mathbf{O}_3$ *absolute concentration preprocessed data.*

when combining the temporal pattern with the corresponding spatial features, the latter of which are shown in Figure 3.4. Shown are the spatial pattern of the 8 most dominant modes for the surface. It should be emphasized that the spatial patterns change with altitude, as illustrated in the supplemental material in [122].

Surface ozone exhibits distinct seasonal patterns, which are captured by the first four modes: the first mode (top left panel in Figure 3.4) resembles the annual average surface concentration of ozone. It can be interpreted as time-invariant 'average ozone' field, from which all other modes add or subtract to describe the spatio-temporal variability of ozone

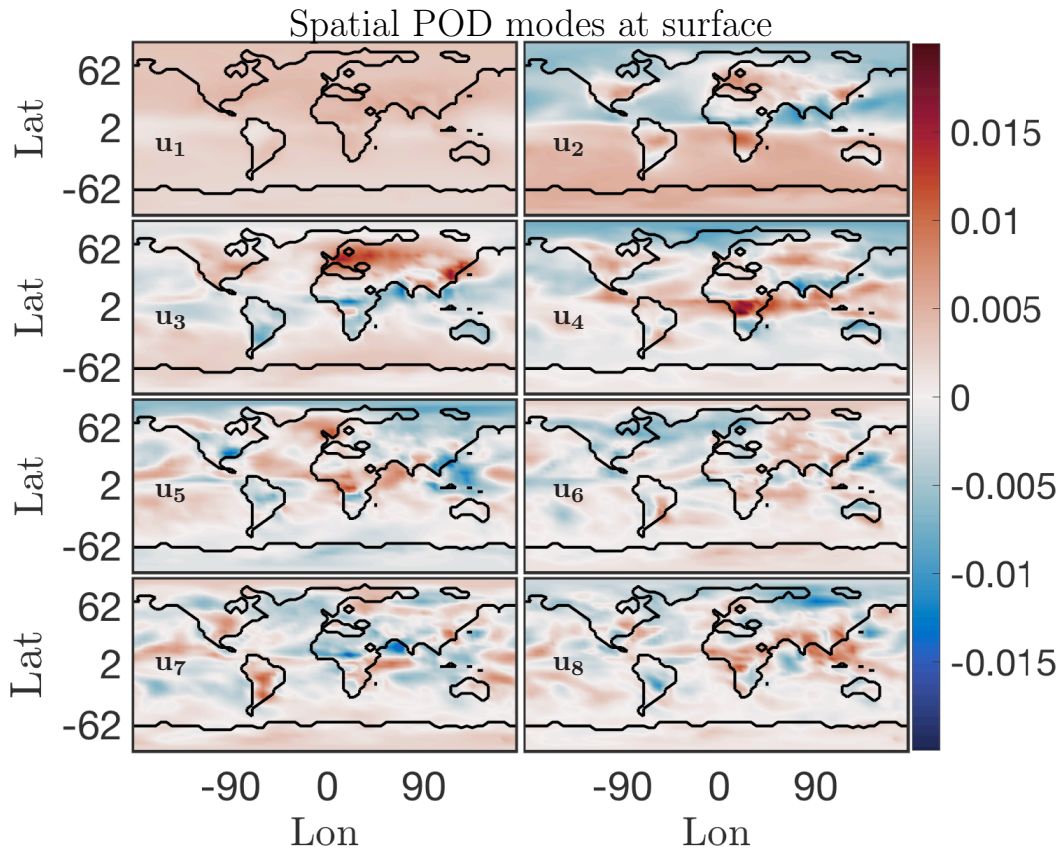


Figure 3.4: *First 8 dominant spatial modes at surface for* $\mathbf{O}_3$ *absolute concentration preprocessed data. Mode 1 is the constant or mean value mode, its corresponding temporal behavior is the blue trend in bottom panel of 3.3. Global low dimensional spatial features for this chemical species are exhibited in order of dominance in Modes 2 through 8.*

in greater detail. The second singular vector (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 Figure 3.3) broadly follows observed ozone burdens in the Southern Hemisphere (SH) [33], and ozone 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 ozone: it shows dominant features over polluted areas (Europe, 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 a biomass burning signal. This is supported by the frequency pattern of this mode, which shows two peaks in Jan/Feb and Jul/Aug, which is in agreement with the two biomass burning seasons over Africa [103].

To summarize, inspection of the spatial and temporal patterns of the dominant modes of ozone shows that randomized SVD successfully reveals prominent features of tropospheric ozone chemistry, such as elevated summertime ozone 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.

3.2.3 Modes from Randomized NMF

A drawback of the SVD solution presented in Section 3.2.2 is that it accepts both negative and positive solutions, which can result in physically unrealistic negative species concentrations. As discussed in Section 3.1.3, positive solutions can be enforced using NMF. The results from NMF of the ozone absolute concentration data are presented in Figures 3.5, 3.6. The cumulative energy spectrum exhibited in the top panel of Figure 3.5 shows a much slower decay as compared to 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 as compared to the SVD. The four dominant temporal modes are presented in the bottom panel of Figure 3.5. These now capture approximately 20% of the total energy spectrum, as compared to 97% for the SVD. This is in large parts because the positivity-constraint prevents the NMF to create a mode for annual mean ozone that can explain most of the energy spectrum - akin to mode one for SVD - but that requires both

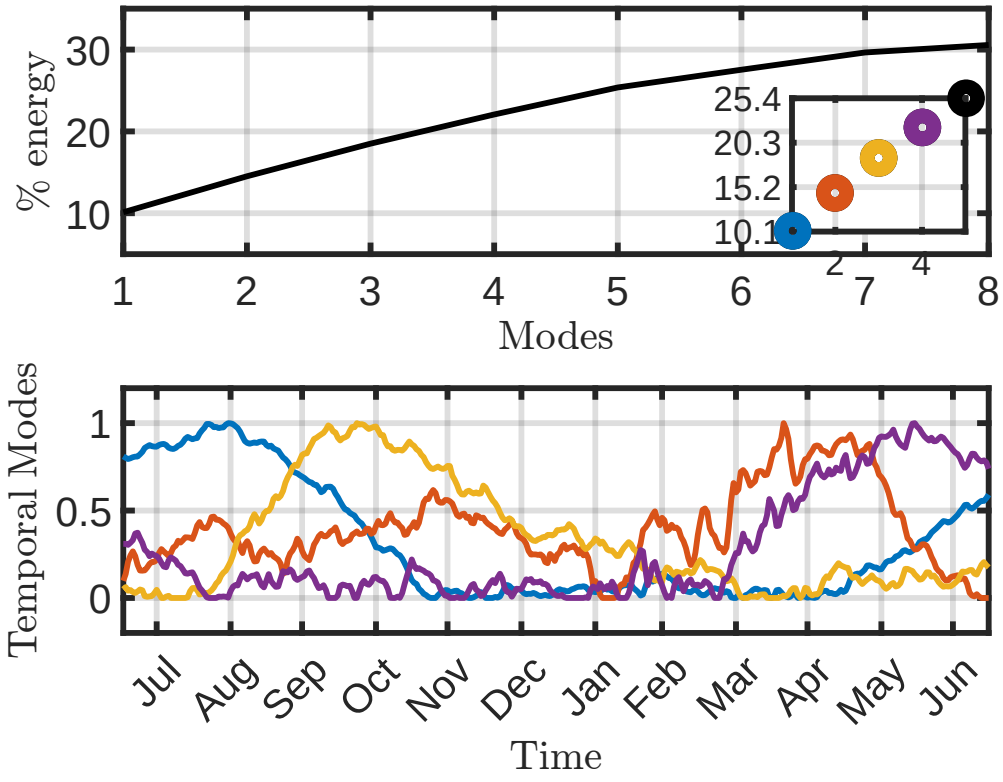


Figure 3.5: *Cumulative energy spectrum from the Nonnegative Matrix Factorization (top) and the corresponding first four columns of the ordered* **H** *temporal factor for* $\mathbf{O}_3$ *absolute concentration preprocessed data (bottom).*

additions and subtractions from this mean field to describe ozone variations in more detail. As a result, none of the NMF modes reflects a distinct representation of the global average ozone field. This is supported by the lack of a time-invariant mode (see Figure 3.5) and also becomes apparent from the corresponding spatial patterns shown in Figure 3.6. None of those resemble the average mean ozone concentration field, as e.g. SVD mode one (see Figure 3.4). Still, the first four spatial and temporal modes of NMF reflect some well known features of ozone chemistry, albeit less obvious than for SVD. The most dominant NMF mode

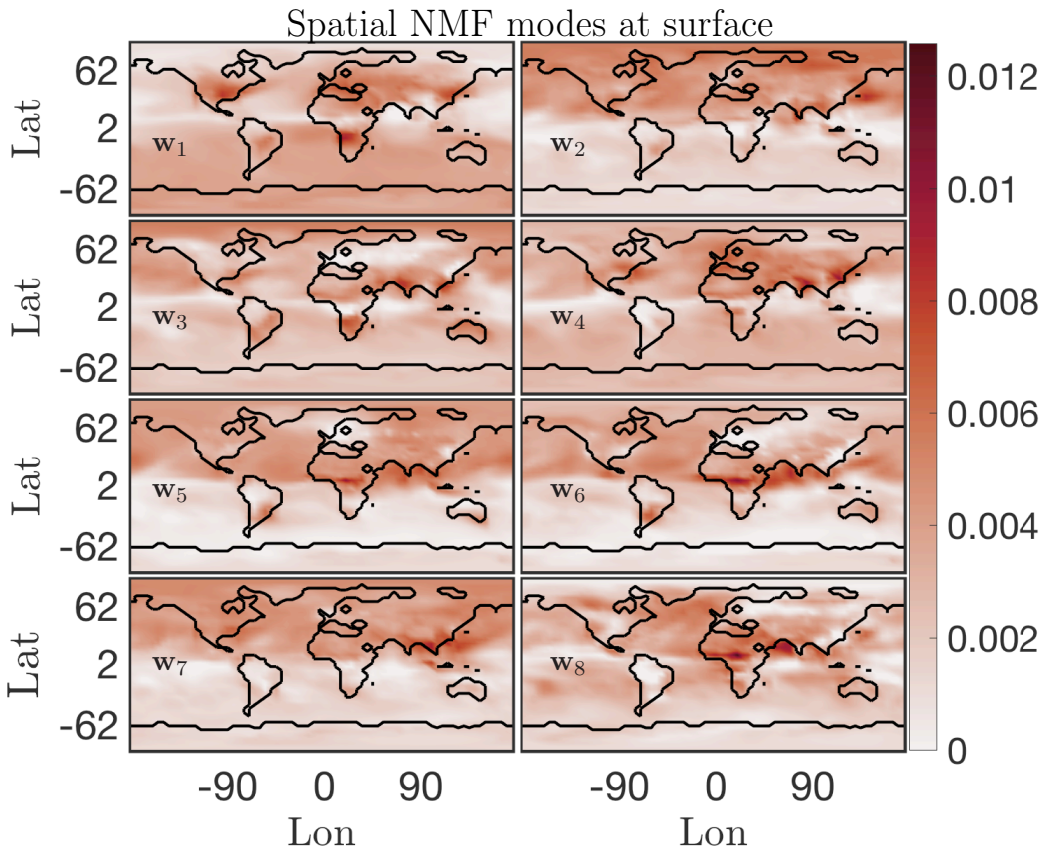


Figure 3.6: *First 8 columns of ordered* $\mathbf{W}$ *spatial factor from NMF at surface for* $\mathbf{O}_3$ *absolute concentration preprocessed data. These modes lend themselves to easy interpretation, the most dominant mode* $\mathbf{w}_1$ *indicates that* $\boldsymbol{O}_3$ *absolute concentration is most active near eastern coastal urban China, North America and western coastal African continent around the region of Congo.*

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 Mar-May, but that also contributes during most other months except Jan. Mode three peaks during Sep/Oct 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 ozone concentrations due to fire activities (e.g. South America burning season Aug/Sep/Oct, India Oct/Nov). Mode four is similar to mode three of the SVD, with strong signals over Europe and Eastern China that peak during boreal spring.

Similar to SVD, the spatio-temporal modes of surface ozone derived from NMF reveal many of the characteristics of ozone chemistry, such as increased ozone concentrations over urban 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 compared to SVD, and significantly more modes are needed to reproduce the spatio-temporal pattern of ozone in detail. This makes SVD better suited for offline pattern discovery applications. However, for practical employment of reduced-order modeling techniques within an Earth System Model, we consider NMF superior since it still realistically captures ozone patterns with relatively few (10's) of modes, but its concentrations are guaranteed to be positive.

3.2.4 Modes from Randomized SPCA

Spatial modes computed from the randomized SPCA are shown in Figure 3.8. Note the localized features isolated by SPCA in these dominant spatial modes as compared to the modes computed by the full SVD. We impose the sparsity regularizer given by Equation 3.27 with $\alpha = 1e-4, \beta = 1e-12$. Reducing the value of α gives a less sparse decomposition. The cumulative energy spectrum in the top panel of Figure 3.7 again demonstrates the much slower decay as compared to the SVD, and more modes are needed to capture the same amount of energy due to the sparsity constraint. This is again as expected, the SPCA computes sparse, localized components at the cost of requiring more modes to capture the same amount of energy as compared to the SVD. In terms of energy explained and interpretability of the modes, the SPCA results for ozone sit in between the results for SVD and NMF discussed above. The first four SPCA modes capture more than 50% of the total energy (Figure 3.7), more than NMF but significantly less than SVD. As for NMF, the lower amount of energy compared to the SVD can be attributed to the fact that the SPCA does not compute a dominant mode for the mean annual ozone concentration. This

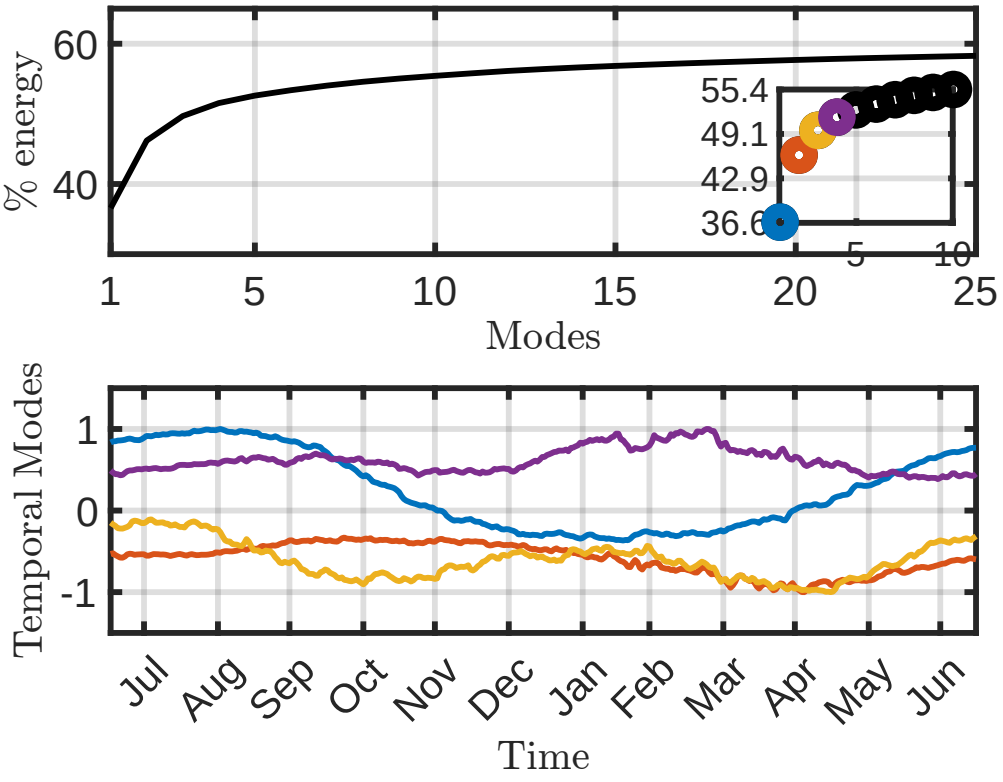


Figure 3.7: *Cumulative energy spectrum from the Sparse Principal Component Analysis (top) and the corresponding 4 dominant temporal modes (bottom) for* $\mathbf{O}_3$ *absolute concentration preprocessed data.*

is expected since SPCA is designed to capture spatially distinct features, rather than broad-based patterns. It thus 'assembles' total ozone concentrations from a series of modes that all show distinct spatial features. Of the dominant four modes shown here, the fourth one 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 ozone over Europe and Eastern China, but at a muted rate during

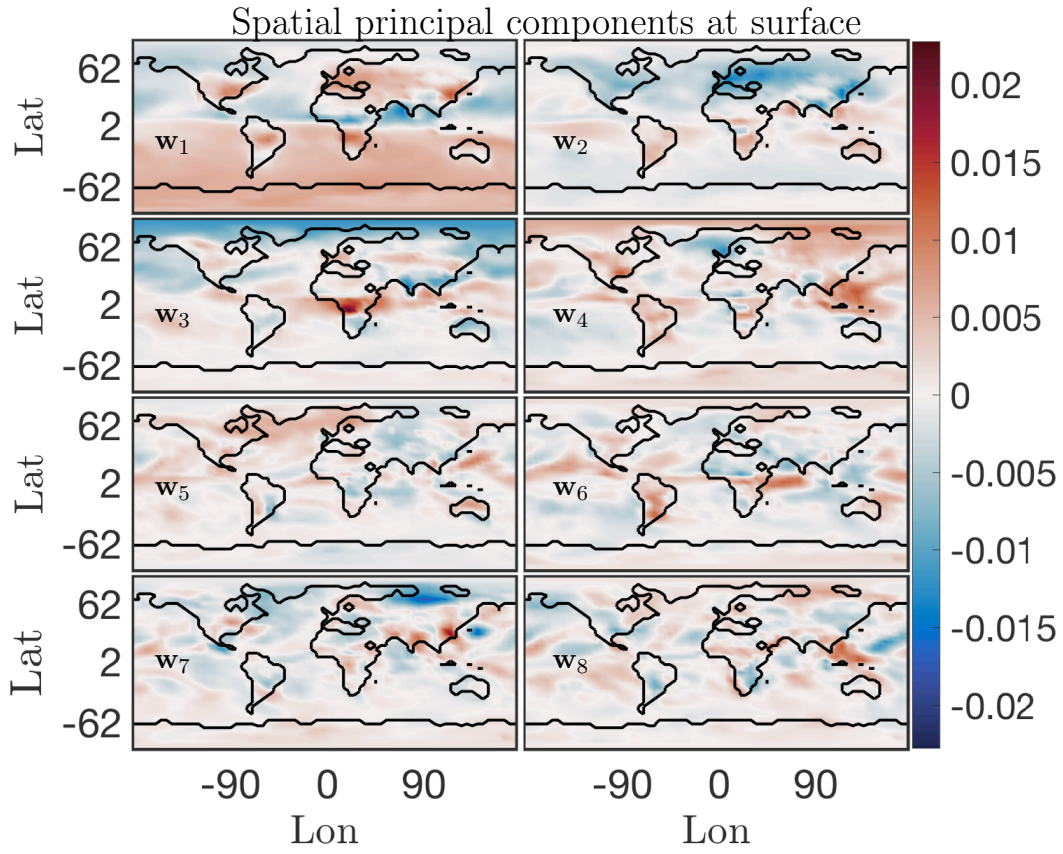


Figure 3.8: *First 8 principal components from SPCA at surface for* $\mathbf{O}_3$ *absolute concentration preprocessed data. With the sparsity constraint, these spatial modes exhibit only localized low dimensional features as compared to those from the SVD of the data. Compare the SVD mean value Mode 1* $\mathbf{u}_1$ *from 3.4 which exhibits a more or less constant field as the dominant low dimensional global feature, to 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.*

Mar-May and also Jul/Aug. It thus has a similar effect as mode 3 of the SVD, but with opposite sign. Mode 3 can be interpreted as a biomass burning signal, with its distinct hot spot over Africa and the two seasonal peaks.

3.3 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 at 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 in a basis which is amenable to a sparse representation [67]. For instance, images can be massively compressed by using wavelet or Fourier basis elements, since natural images are sparse in these basis elements. Compression formats such as JPEG2000 are critically enabling for the electronics industry and allowing for our electronic devices to hold an exceptionally large number of video, audio and picture 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 here simply require a fraction of the data to be stored in the $\mathbf{Q}$ matrix and the rank-r embedding columns of $\tilde{\mathbf{U}}$, $\mathbf{\Sigma}$ and $\mathbf{V}$. For images and video, compression allows almost perfect reconstruction of the original data while storing only a few percent of the wavelet coefficients. It is expected that similar compression performance can be achieved with the basis elements of the scalable SVD.

As an illustrative example, Figure 3.9(a) shows a reconstruction of the absolute concentration of surface $\mathbf{O}_3$ 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 require only storing 0.025%, 0.25% and 0.5% of data, respectively, as opposed to 87 million data points of the original annual surface ozone data (See Figure. 3.9(b)). The reconstruction with as few as 5 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 $\mathbf{U}$ and $\mathbf{V}$ matrices, along with the first r diagonal terms of $\mathbf{\Sigma}$. This gives a

data compression ratio of $(m \times n)/(m \times r + r \times n + r)$ (See Figure 3.1). 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 since 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) [14]. 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 a fraction of the computational cost. Indeed, Monte-Carlo computations of many high-dimensional problems of interest are currently intractable even with super computers, thus highlighting the need for proxy models that can be computed at reduced cost. In future work, we will aim to develop ROMs that exploit the low-rank embeddings computed with our scalable algorithms.

3.4 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 of 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 work, 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 looking at only 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 [47, 13] 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. However, 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.

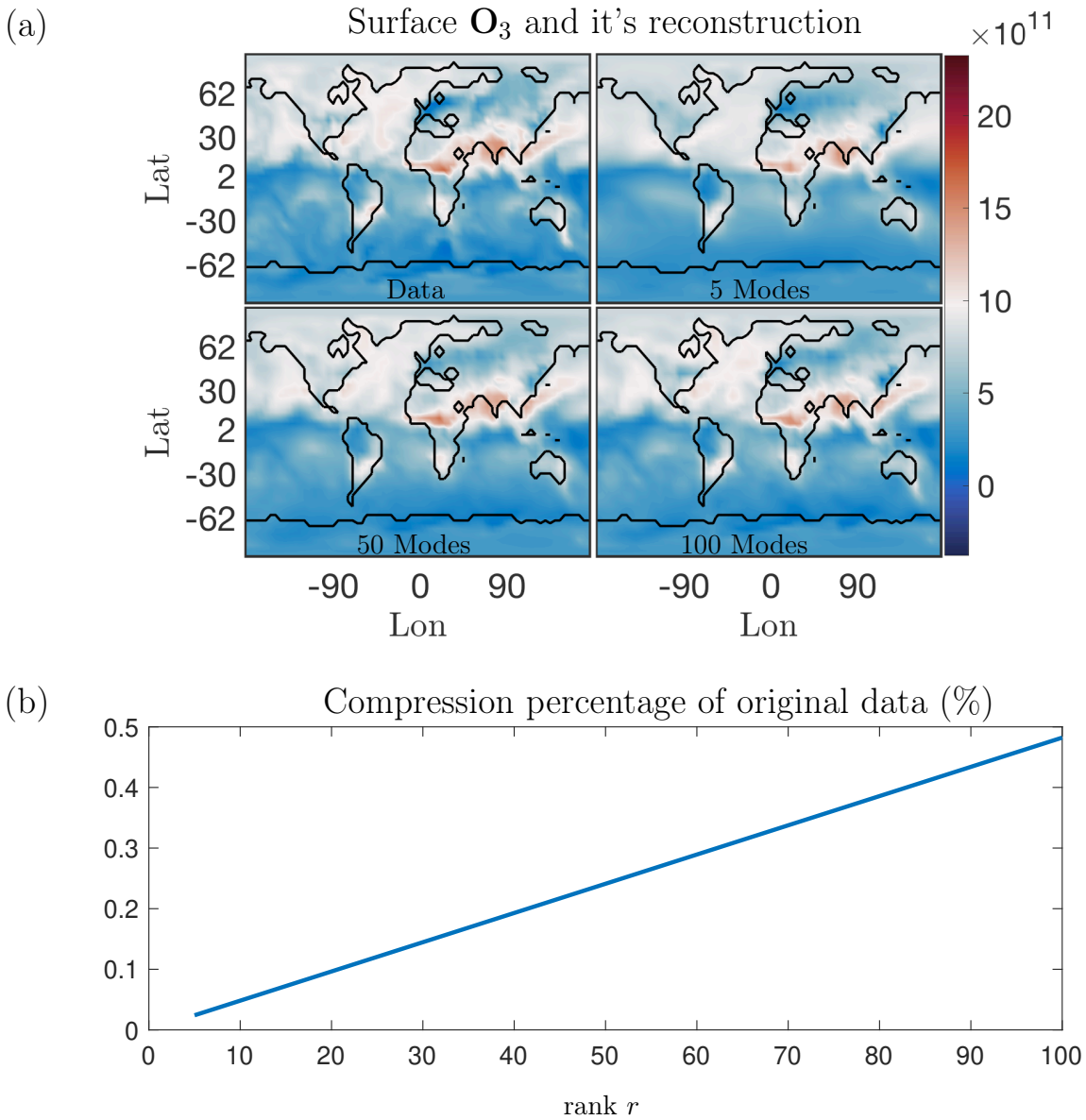


Figure 3.9: (a) *One-time snapshot of surface* $\mathbf{O}_3$ *absolute concentration reference data (top left) and its reconstruction using* **5, 50** *and* **100** SVD *modes, respectively. Using 5 modes, only the most dominant features are reconstructed successfully, but as the number of modes used 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.*

Chapter 4

OPTIMIZED DYNAMIC MODE DECOMPOSITION FOR RECONSTRUCTION AND FORECASTING OF ATMOSPHERIC CHEMISTRY DATA

Dimensionality reduction is a critically enabling aspect of machine learning and data science that can be leveraged to approximate the monitoring and forecasting capabilities of global chemistry with more readily tractable computational algorithms. *Dynamic mode decomposition* (DMD) is a leading data-driven regression architecture for adaptively learning linear dynamics models over snapshots of temporal data in a low-dimensional subspace. DMD has been broadly used in the scientific community due to its ease of use, interpretability and adaptive nature [68]. When applied to the spatio-temporal dynamics of atmospheric chemistry, we demonstrate that the method provides an effective and computational efficient *reduced order modeling* strategy that can be used for characterization, monitoring and forecasting of global chemical concentrations with either computational or sensor data. Moreover, we show that the optimized DMD algorithm [5] and bagging optimized DMD (BOP-DMD) [109] versions of the DMD algorithm are critical for characterizing the complexities of the chemical interaction dynamics.

The characterization of multiscale phenomenon, such as that embodied by global atmospheric chemistry, remains challenging due to the need to resolve spatial and temporal scales that are separated by many orders of magnitude. Computational methods, which are typically based upon partial differential equations that model the governing dynamics, easily become intractable due to the need to resolve the finest space scales and the fastest time scales. Thus numerical stiffness is automatically imposed upon a numerical schemes in such a spatio-temporal system. Building models from sensor data directly is no different: sensors must placed densely in space in order to resolve spatial features. This also places significant limits on practicality as sensors are not only prohibitively expensive, but also require completely impractical global coverage. Computations and sensors, however, are

typically used in combination and provide the critical data infrastructure for modeling the multiscale physics of atmospheric chemistry. So despite the limitations and cost, many advances have been made in our ability to characterize, predict and monitor global chemistry.

Reduced order models (ROMs) provide and attractive alternative to large scale computing. ROMs provide a mathematical architecture for reducing the computational complexity of mathematical models in numerical simulations [15, 4, 101, 59]. Fundamental to rendering simulations computationally tractable is the construction of a low-dimensional subspace on which the dynamics can be approximately embedded. Unfortunately, projective-based ROM construction often produces a low-rank model for the dynamics that can be unstable [27], i.e. the models produced generate solutions that rapidly go to infinity in time. Machine learning techniques offer a diversity of alternative methods for computing the time-dynamics in the low-rank subspace, with a diversity of neural networks showing how to advance solutions, or learn the flow map from time t to $t+\Delta t$ [100, 76]. Indeed, deep learning algorithms provide a flexible framework for constructing a mapping between successive time steps. The typical ROM architecture constrains the dynamics to a subspace spanned by POD (proper orthogonal decomposition), thus in the new POD coordinate system, time evolution can be used to construct a time-stepping model using neural networks. Recently, Parish and Carlberg [95] and Regazzoni et al. [102] developed a suite of neural network based methods for learning time-stepping models for tropospheric bromine chemistry and cardiovascular dynamics, respectively. Moreover, Parish and Carlberg provide extensive comparisons between different neural network architectures along with traditional techniques for time-series modeling.

Projective ROMs are often unstable and ill-suited for massive multiscale systems, while deep learning models require significant time and data for training and also assume stationarity of the data. Both of these limitations make their use in global atmospheric modeling problematic. However, a computationally efficient and adaptive ROM approach is embodied by DMD. DMD was introduced as as algorithm by Schmid [112] and has rapidly become a commonly used data-driven analysis tool. It is the leading approximation method for the Koopman (linear) operator from data [105]. DMD by construction provides a method for identifying spatio-temporal coherent structures in high-dimensional time-series

data. DMD analysis offers an dynamic version of standard dimensionality reduction methods such as the *proper orthogonal decomposition* (POD), which highlighted low-rank features in spatio-temporal data [67]. However, DMD not only provides a low-rank subspace, but each mode is associated with linear (exponential) behavior in time, often given by oscillations at a fixed frequency with growth or decay. Thus, DMD is a regression to solutions of the form

$$\mathbf{x}(t) = \sum_{j=1}^{r} \phi_j e^{\omega_j t} b_j = \mathbf{\Phi} \exp(\mathbf{\Omega} t)\mathbf{b}, \tag{4.1}$$

where $\mathbf{x}(t)$ is an r-rank approximation to a collection of state space measurements $\mathbf{x}_k = \mathbf{x}(t_k)$ ($k = 1, 2, \cdots, n$). The algorithm regresses to values of the DMD eigenvalues ω_j, DMD modes ϕ_j and their loadings b_j. The ω_j determines the temporal behavior of the system associated with a modal structure ϕ_j. Such a regression can also be learned from time-series data [70]. DMD may be thought of as a combination of SVD/POD in space with the Fourier transform in time, combining the strengths of each approach [31, 68]. DMD is modular due to its simple formulation in terms of linear algebra, resulting in innovations related to control [99, 36], compression [25, 49], reduced-order modeling [2], and multi-resolution analysis [69, 77], among others.

4.1 Optimized Dynamic Mode Decomposition (DMD)

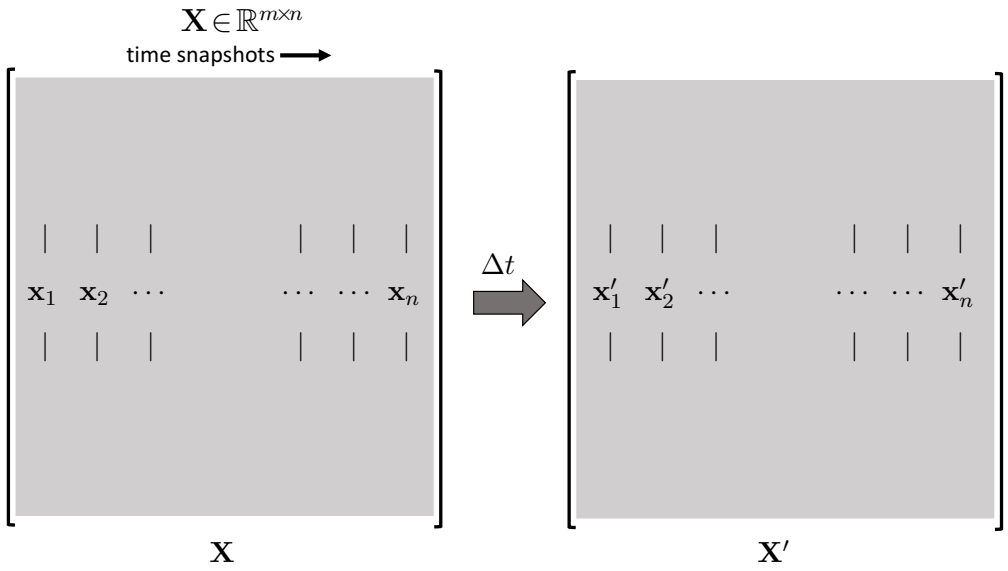
The DMD algorithm schematic is shown in Figure 4.1. The DMD algorithm seeks the leading spectral decomposition of the best fit linear operator $\mathbf{A}$ [24] that approximately advances the snapshot measurements of the state of a system $\mathbf{x} \in \mathbb{R}^m$ forward in time by step size Δt:

$$\mathbf{X}' \approx \mathbf{A}\mathbf{X} \tag{4.2}$$

which leads to the mathematical definition of operator $\mathbf{A}$ as:

$$\mathbf{A} = \arg\min_{\mathbf{A}} \|\mathbf{X}' - \mathbf{A}\mathbf{X}\|_F = \mathbf{X}'\mathbf{X}^\dagger \tag{4.3}$$

where $\|\cdot\|_F$ is the Frobenius norm and † denotes the pseudo-inverse. This is computed by an SVD of $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$, where * denotes the complex-conjugate transpose. The matrices $\mathbf{U} \in \mathbb{R}^{\mathbf{m}\times\mathbf{m}}$ and $\mathbf{V} \in \mathbb{R}^{\mathbf{n}\times\mathbf{n}}$ are unitary with $\mathbf{U}^*\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^*\mathbf{V} = \mathbf{I}$. The columns of $\mathbf{U}$ are the left-singular values, also known as the POD modes. Now $\mathbf{X}^\dagger = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^*$, and we can



Data snapshots: $\mathbf{x}(t_k) = \mathbf{x}_k \in \mathbb{R}^m$

$$\mathbf{X}' \approx \mathbf{A}\mathbf{X} \longrightarrow \underset{\boldsymbol{\omega}, \boldsymbol{\Phi}_\mathbf{b}}{\arg\min} \|\mathbf{X} - \boldsymbol{\Phi}_\mathbf{b}\mathbf{T}(\boldsymbol{\omega})\|_F$$

Figure 4.1: *The data* $\mathbf{x}(t_k)$ *is collected into snapshot matrices* $\mathbf{X}$ *which are used to regress to the best exponential (linear) solution* $\underset{\boldsymbol{\omega}, \boldsymbol{\Phi}_\mathbf{b}}{\arg\min} \|\mathbf{X} - \boldsymbol{\Phi}_\mathbf{b}\mathbf{T}(\boldsymbol{\omega})\|_F$*, where* $\boldsymbol{\Phi}_\mathbf{b}$ *are the weighted DMD modes and* $\mathbf{T}$ *is a matrix of exponentials for fitting the data (4.9).*

compute the operator $\mathbf{A} = \mathbf{X}'\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^*$. However, for a high-dimensional state vector, $\mathbf{x} \in \mathbb{R}^m$ it may be intractable to compute the spectral decomposition of $\mathbf{A}$. Instead, exact DMD leverages the low-dimensional structure of the state by projecting $\mathbf{A}$ into the $r-$leading singular vectors, resulting in a reduced matrix

$$\tilde{\mathbf{A}} = \tilde{\mathbf{U}}_\mathbf{r}{}^* \mathbf{A} \tilde{\mathbf{U}}_\mathbf{r} = \tilde{\mathbf{U}}_\mathbf{r}{}^* \mathbf{X}' \tilde{\mathbf{V}}_\mathbf{r} \boldsymbol{\Sigma}_\mathbf{r}^{-1} \tag{4.4}$$

Schmid [112] designed a procedure to approximate the high-dimensional DMD modes or eigenvectors of $\mathbf{A}$ from this reduced matrix $\tilde{\mathbf{A}}$ and the data snapshot matrix $\mathbf{X}$. Tu et al. [121] proved that these approximate modes are the exact eigenvectors of full $\mathbf{A}$ under certain conditions. The leading spectral decomposition of $\mathbf{A}$ is approximated by $\tilde{\mathbf{A}}\mathbf{W} = \mathbf{W}\boldsymbol{\Lambda}$, with DMD eigenvalues being entries of a diagonal matrix $\boldsymbol{\Lambda}$, which correspond to the eigenvalues of the full matrix $\mathbf{A}$. The columns of $\mathbf{W}$ are the corresponding eigenvectors. The high-dimensional DMD modes can be reconstructed using the time-shifted data matrix as:

$$\boldsymbol{\Phi} = \mathbf{X}' \tilde{\mathbf{V}}_\mathbf{r} \boldsymbol{\Sigma}_\mathbf{r}^{-1} \mathbf{W} \tag{4.5}$$

We can now reconstruct the solution $\mathbf{x}(\mathbf{t})$ as shown in Equation. 4.1, where the amplitudes or loadings of each mode is given by:

$$\mathbf{b} = \mathbf{\Phi}^{\dagger}\mathbf{x_1} \tag{4.6}$$

The full data matrix $\mathbf{X}$ is thus approximated by:

$$\mathbf{X} \approx \mathbf{\Phi}(\mathbf{b})\mathbf{T}(\omega) \tag{4.7}$$

$$= \begin{bmatrix} | & | & | \\ \phi_1 & \cdots & \phi_r \\ | & | & | \end{bmatrix} \begin{bmatrix} b_1 & & \\ & \ddots & \\ & & b_r \end{bmatrix} \begin{bmatrix} e^{\omega_1 t_1} & \cdots & e^{\omega_1 t_n} \\ \vdots & \ddots & \vdots \\ e^{\omega_r t_1} & \cdots & e^{\omega_r t_n} \end{bmatrix} \tag{4.8}$$

where $\omega_i = \log(\lambda_i)$, the DMD eigenvalues.

However, the DMD is rarely used for forecasting and/or reconstruction of time-series data, except in cases with noise-free or nearly noise-free data. This is because the exact DMD is extremely sensitive to noise in the data, causing a bias in the computed DMD modes and eigenvalues [8, 35, 58].

The *optimized DMD* algorithm of Askham and Kutz [5], which uses a variable projection method for nonlinear least squares to compute the DMD for unevenly timed samples, provides the best and most optimal performance of any algorithm currently available. Indeed, this optimal performance is mathematically guaranteed by the exponential fitting procedure of Askham and Kutz [5]. The optimized DMD algorithm solves the exponential fitting problem:

$$\underset{\boldsymbol{\omega},\mathbf{\Phi_b}}{\arg\min} \|\mathbf{X} - \mathbf{\Phi_b}\mathbf{T}(\boldsymbol{\omega})\|_{\mathbf{F}}, \tag{4.9}$$

where $\mathbf{\Phi_b} = \mathbf{\Phi}\mathbf{diag}(\mathbf{b})$. This has been shown to provide a superior decomposition due to its ability to optimally suppress bias and handle snapshots collected at arbitrary times. Figure. 4.2 compares the $\mathbf{NO_{START}}$ 30-day preprocessed data (top panel) with the classical or exact DMD reconstruction results (middle panel), and the optDMD reconstruction results (bottom panel). The classical DMD reconstruction dies out within a few days, failing in the task of even reconstructing the time-series data it was originally regressed to, whereas the optDMD is able to capture, sustain and faithfully reconstruct the original time series.

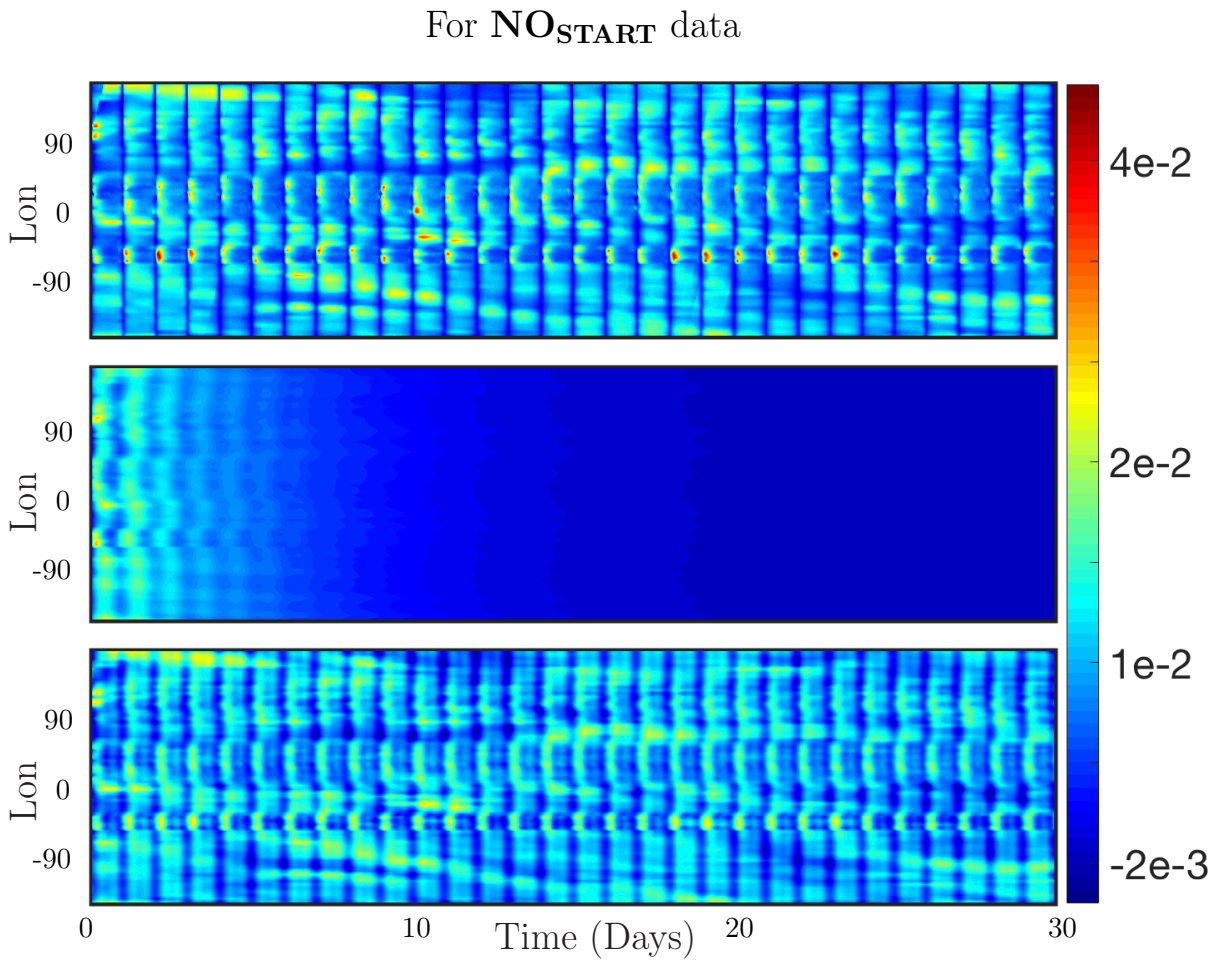


Figure 4.2: *Comparing* 30 *day reconstruction results for Classical and Optimized DMD at the surface of* **NO** *preprocessed data at* Lat $= 30°$. *The results are for absolute concentration or* Start *data; the top panel shows the preprocessed data, the middle panel shows the reconstruction from the Classical* DMD, *and the bottom panel shows the reconstruction from Optimized* DMD. *The Classical* DMD *is unable to capture the dynamics for the absolute concentration data and it decays down to zero. The Optimized* DMD *reconstructs the data and resolves the dynamics accurately.*

Since the optimized DMD solves a nonlinear optimization problem, we can also introduce certain constraints. The optDMD algorithm can be constrained to produce eigenvalues

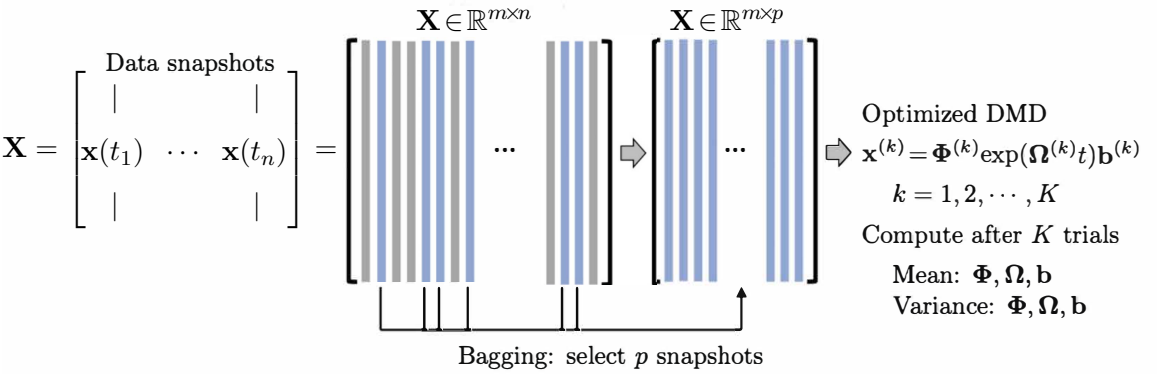


Figure 4.3: *Summary of the BOP-DMD architecture, reproduced with permission from [109]. The data snapshots* $\mathbf{x}(t_k)$ *are collected over* m *snapshots into the matrix* $\mathbf{X}$*. Columns of* $\mathbf{X}$ *are randomly sub-selected into the matrix* $\mathbf{X}^{(k)}$ *to build an optimized DMD model. Each DMD model* $\mathbf{x}^{(k)} = \mathbf{\Phi}^{(k)} \exp(\mathbf{\Omega}^{(k)} t) \mathbf{b}^{(k)}$ *is used to compute the statistics (mean and variance) of the DMD parametrizations* $\mathbf{\Phi}$, $\mathbf{\Omega}$, $\mathbf{b}$ *which are used in building the BOP-DMD ensemble solution with Uncertainty Quantification (UQ).*

constrained to (i) The imaginary axis:

$$\arg\min_{\boldsymbol{\omega}, \mathbf{\Phi_b}} \|\mathbf{X} - \mathbf{\Phi_b} \mathbf{T}(\boldsymbol{\omega})\|_{\mathbf{F}}, \text{ subject to } \Re(\boldsymbol{\omega} = \mathbf{0}) \tag{4.10}$$

(ii) The closed left-half plane:

$$\arg\min_{\boldsymbol{\omega}, \mathbf{\Phi_b}} \|\mathbf{X} - \mathbf{\Phi_b} \mathbf{T}(\boldsymbol{\omega})\|_{\mathbf{F}}, \text{ subject to } \Re(\boldsymbol{\omega} \leq \mathbf{0}) \tag{4.11}$$

As discussed below, these constraints further stabilize and make robust reproduction and forecast of the time series data. The disadvantage of optimized DMD is that one must solve a nonlinear optimization problem, which can at times fail to converge.

4.2 Bagging OPtimized Dynamic Mode Decomposition (BOP-DMD)

BOP-DMD [109] leverages Breiman's statistical bagging sampling strategy [74] in partnership with the optimized DMD algorithm. The BOP-DMD architect is presented in Figure. 4.3. Bagging is designed to produce an ensemble of models, thereby reducing model variance and suppressing over-fitting by design. Not only does ensembling improve DMD, it also is effective in deep neural network regressions [3]. Further innovations include stabilizing the variable projection technique used by optDMD so that it converges consistently to an optimal solution [109]. Specifically, the optimized DMD (opt-DMD) algorithm relies on the nonlinear optimization of variable projection. Its ability to converge is often dependent upon a suitable initial guess for the DMD eigenvalues and eigenvectors.

The BOP-DMD algorithm accounts for the initialization process and further provides the optimal solutions to linear models by using optDMD as the regression architecture. Algorithm 1 shows the algorithmic structure of BOP-DMD, highlighting the bagging, initialization and ensembling of the DMD models to produce an ensemble, probabilistic DMD model. The initialization of DMD is accomplished by first constructing an optDMD model approximation, whose eigenvalues and eigenvectors $\mathbf{\Phi}_0$ can be used to seed the BOP-DMD. p snapshots are randomly selected from the full data matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, to form a subset data matrix $\mathbf{X} \in \mathbb{R}^{m \times p}$. optDMD produces the model for this subset of data, and we save the resulting model parameters. The process is repeated for K trials producing an ensemble of optDMD models. The mean $\{\langle \mathbf{\Phi} \rangle, \langle \mathbf{\Omega} \rangle, \langle \mathbf{b} \rangle\}$ and variance $\{\langle \mathbf{\Phi}^2 \rangle, \langle \mathbf{\Omega}^2 \rangle, \langle \mathbf{b^2} \rangle\}$ of the model parameters $\mathbf{\Phi},\ \mathbf{\Omega},\ \mathbf{b}$ can now be computed. Hence, in addition to producing the DMD model itself, the output of algorithm 1 generates both spatial and temporal uncertainty quantification metrics or UQ metrics. In this work, we primarily focus on the temporal UQ metrics for forecasting.

4.3 Analysis Data

The analysis is performed for preprocessed or time-shifted raw data for two months or 60 days July, 2^{ND} - August, 30^{TH}. At this time of the year there are shorter days in the Southern Hemisphere and longer days in the Northern Hemisphere, and each latitude has a different length of the day. The photolysis rate dictates a different 'turn on' duration for the dynamics of many key chemicals of interest. Hence, the analysis is performed on data from

Algorithm 1 BOP-DMD

Input: Input ($\mathbf{X}$, p, K)

procedure BOPDMD($\mathbf{X}$, p, K)
  Compute $\mathbf{\Phi_0}, \mathbf{\Omega_0}, \mathbf{b_0}$ ▷ optDMD regression
  for $k \in \{1, 2, \cdots, K\}$ **do** ▷ Compute K optDMD models.
    Choose p of n snapshots ($p < n$) ▷ Bagging
    optDMD $\mathbf{\Phi_k}, \mathbf{\Omega_k}, \mathbf{b_k}$ ▷ Initialize with $\mathbf{\Omega_0}$
    Update $\mathbf{\Phi}, \mathbf{\Omega}, \mathbf{b}$ ▷ Add $\mathbf{\Phi_k}, \mathbf{\Omega_k}, \mathbf{b_k}$ to $\mathbf{\Phi}, \mathbf{\Omega}, \mathbf{b}$
  end for
  Compute mean $\boldsymbol{\mu} = \{\langle\mathbf{\Phi}\rangle, \langle\mathbf{\Omega}\rangle, \langle\mathbf{b}\rangle\}$
  Compute variance $\boldsymbol{\sigma} = \{\langle\mathbf{\Phi}^2\rangle, \langle\mathbf{\Omega}^2\rangle, \langle\mathbf{b^2}\rangle\}$
  return $\boldsymbol{\mu}, \boldsymbol{\sigma}$ ▷ Return optDMD parameters.
end procedure

the elevation = 1 and one latitude at a time, and for all 72 longitudes with data shifted in time as described above.

In most of the latitudes in the Southern Hemisphere the days are much shorter than the nights, and accordingly the 'turn on' period is much shorter as compared to the nighttimes 'turn off' period. Hence, the data exhibits a spiky pattern and will need much higher modes to accurately reconstruct it; and/or we would need to isolate the day time values only when there are active dynamics present. Hence, we are picking latitude = 30° for the analysis, which has the longest day times. The first 40 days of data is the 'training' data, the DMD diagnostics are presented for this time period and for latitude = 30°. With 72 snapshots per day, we have a data matrix of $72(lon) \times 2880(time)$ for each latitude. The optDMD is performed for this data matrix. We perform the analysis for the six different chemical species of interest: Nitrous Oxide **NO**, Ozone $\mathbf{O_3}$, Nitrous dioxide $\mathbf{NO_2}$, Hydroxyl **OH**, Isoprene **ISOP**, and Carbon Monoxide **CO**. For each species, we have **START** or absolute concentration data (expressed in units of molecules/cm^3) and **TEND** or tendency/rate of change data (expressed in units of molecules/cm^3/s). Using the resulting diagnostics, the

dynamics for the next 20 days are forecast.

4.4 DMD Diagnostics

The optDMD decomposes data into time dynamics represented by the spectrum of eigenvalues $\mathbf{\Omega}$ and the corresponding spatial modes $\mathbf{\Phi}$. We will be presenting results from Exact DMD and optDMD with and without constraining the eigenvalues. The diagnostics are presented for the chemical species **OH** (hydroxyl radical) data for 40 days, with a hard threshold truncation of $r = 25$ for the **START** data and $r = 50$ for the **TEND** data. Truncating the rank for the models is described below. The diagnostics are presented for both absolute concentration of the chemical species or $\mathbf{OH_{START}}$ data on the left panels and rate of change of concentrations/tendencies of the chemical species due to chemistry or $\mathbf{OH_{TEND}}$ data on the right panels in Figure 4.4 and Figure 4.5. Four different spectra of the DMD eigenvalues are presented in Figure 4.4, and the corresponding reconstruction of data in second through fifth two panels of Figure 4.5.

- The top two panels in Figure4.5 are the actual $\mathbf{OH_{START}}$ data on the left and actual $\mathbf{OH_{TEND}}$ data on the right, presented for comparison.

- The spectrum for optDMD with no constraints on the eigenvalues for $\mathbf{OH_{START}}$ data is presented on the top left panel, and for $\mathbf{OH_{TEND}}$ data is presented on the top right panel of Figure 4.4. For both data sets, some eigenvalues fall on the right-half plane with positive real parts, causing the corresponding modes to grow in time. The corresponding reconstruction of data is presented in the second two panels of Figure 4.5. optDMD with no constraints does a faithful reconstruction of data, but the forecasting results are poor, with the time series growing exponentially as a result of some eigenvalues on the right-half plane. This approach is not used henceforth.

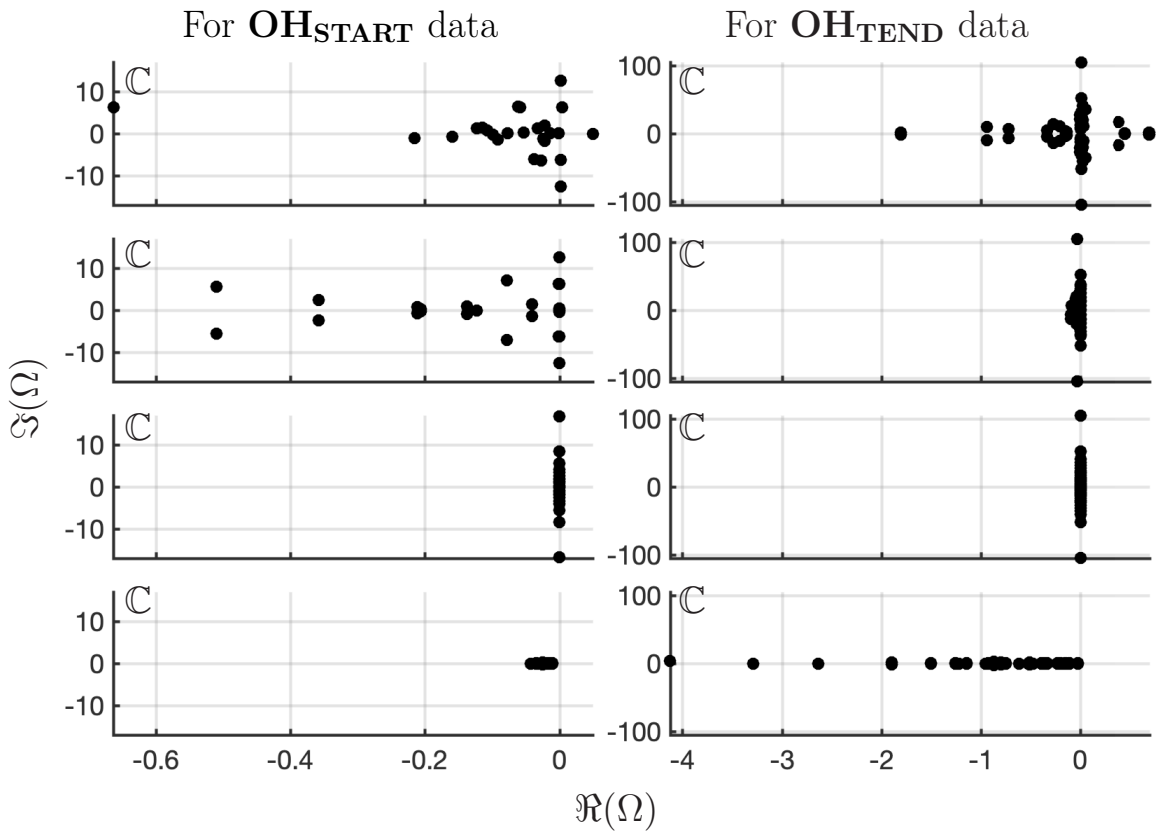


Figure 4.4: *Comparing the spectrum for* 40 *day reconstruction results for Classical and Optimized DMD at the surface of* **OH** *preprocessed data. On the left 4 panels are the eigenvalues of* $\mathbf{OH_{START}}$ *data; on the right 4 panels are the eigenvalues of* $\mathbf{OH_{TEND}}$ *at* Lat $= 30°$. *The top panels show the spectrum from Optimized* DMD *with no constraints, the second set of panels show the spectrum from Optimized* DMD *with linearized constraints that the eigenvalues be on the left-half plane, the third set of panels show the spectrum from optimized* DMD *with linearized constraints that the eigenvalues be imaginary, and the bottom panels show the spectrum from Classical or Exact* DMD.

- The optDMD is then constrained to produce only eigenvalues with negative or zero real parts, i.e. eigenvalues on the closed left-half plane ($\Re(\omega_i \leq 0)$). The resulting spectrum for the two data sets is presented on the second two panels in Figure 4.4. The

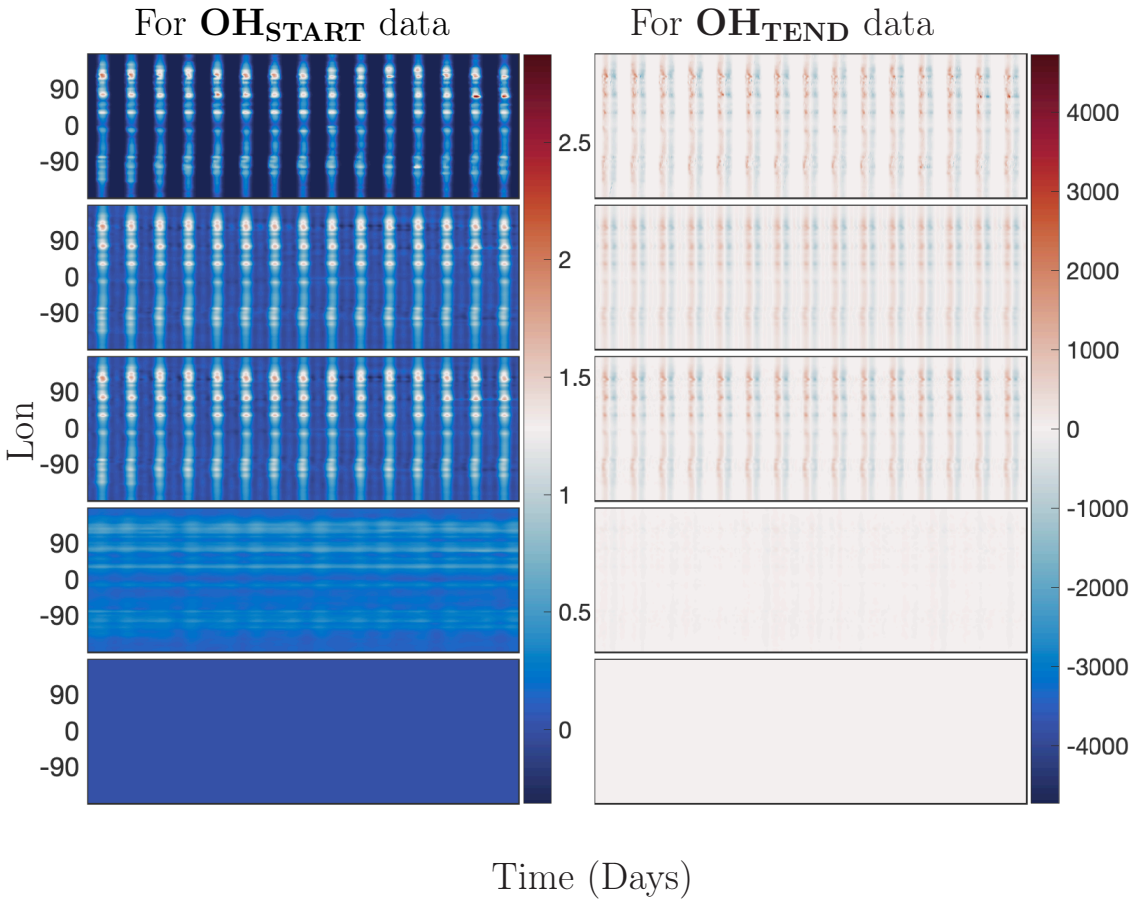


Figure 4.5: *Comparing* 40 *day reconstruction results for Classical, Optimized DMD, and Optimized DMD with no constraints at the surface of* **OH** *preprocessed data at* Lat = 30°. *The left panel is for absolute concentration or* Start *data and the right panel is for* Tendency *data; the top panels show the preprocessed data, the second panels show the reconstruction from Optimized* DMD, *the third panels show the reconstruction from Optimized* DMD *with eigenvalues constrained to the Left half-plane, the fourth panels show the reconstruction from Optimized* DMD *with eigenvalues constrained to the Imaginary axis, and the bottom panels show the reconstruction from the Classic* DMD. *The Classical* DMD *is unable to reconstruct the dynamics for the absolute concentration and tendency data.*

corresponding reconstruction of data is presented in the third two panels of Figure 4.5. optDMD with these constraints not only faithfully reconstructs the data, but the

forecasting results are also accurate, as presented in the following section.

- The optDMD is then constrained to produce only imaginary eigenvalues with zero real parts ($\Re(\omega_i = 0)$). The resulting spectrum for the two data sets is presented on the third two panels in Figure 4.4. The corresponding reconstruction of data is presented in the fourth two panels of Figure 4.5. optDMD with these constraints is not able to capture the data dynamics, and will not be used henceforth.

- Finally, results from Exact DMD for both data sets are presented in the bottom two panels of Figure 4.4 and Figure 4.5. The resulting spectrum for the two data sets have most eigenvalues on the negative real axis, implying decaying modes. The corresponding reconstruction of data also decays out with no dynamics from the data captured or represented faithfully. This approach is not used henceforth.

We use optDMD with eigenvalues constrained on the closed left-half plane $\Re(\omega_i \leq 0)$. When computing the optDMD, we truncate the number of modes to avoid fitting dynamics to the lowest energy modes, which may cause over-fitting and may be corrupted by noise. We would be truncating using *hard-thresholding* at a rank r at which the relative error in reconstruction has an 'elbow', i.e. the error graph flattens out without further decrease. Focusing on six key chemicals of interest: **NO**, $\mathbf{O_3}$, $\mathbf{NO_2}$, **OH**, **ISOP**, **CO START** and **TEND** data, we now compute the relative error in reconstruction as we increase the number of modes from 1 to 50. The results for the two data sets and the six chemical species is presented in Figure 4.6. A larger number of modes is needed to reconstruct the **TEND** data as compared to the **START** data. Based on the results, we use 20-30 modes for optimal diagnostics of **START** data, depending on the chemical species. For the **TEND** data, we pick between 30-50 modes.

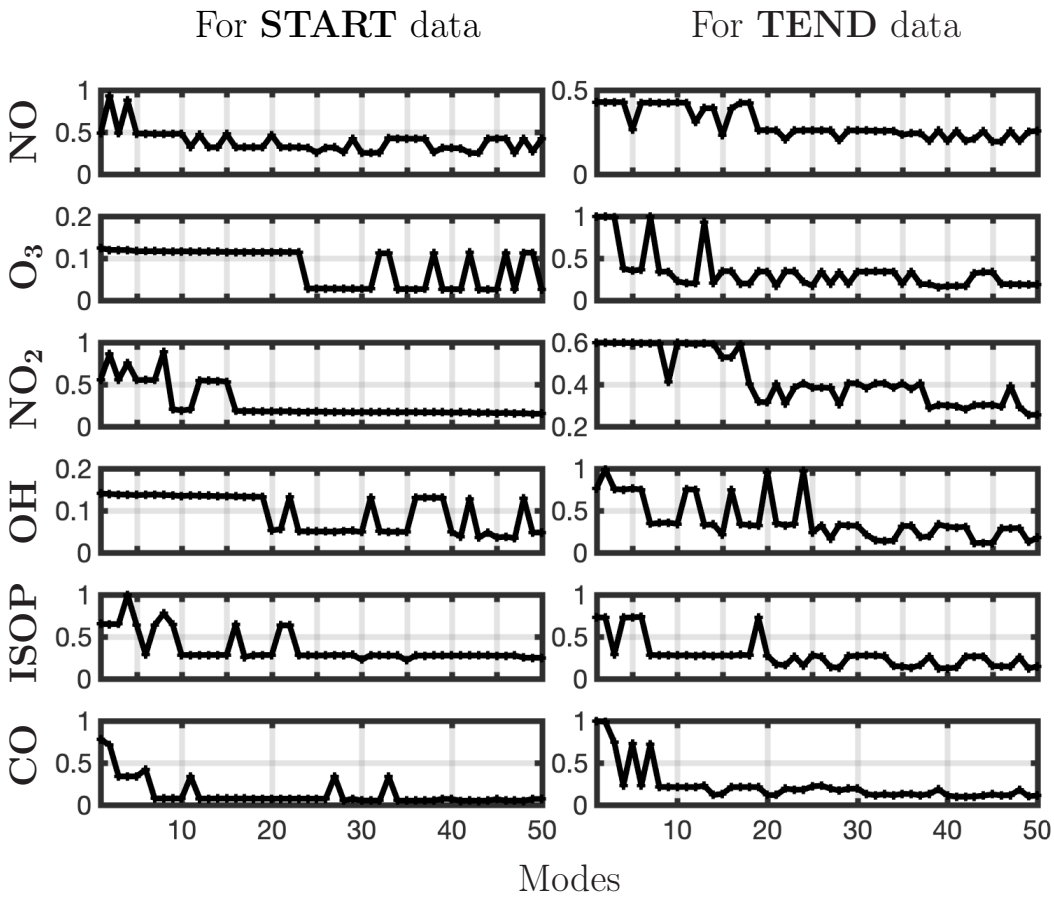


Figure 4.6: *Relative reconstruction error plotted against number of modes used for Optimized DMD with eigenvalues constrained to the left-half plane; for 6 different chemical species and Start and Tend data at Latitude=30°*

Finally, we present the global spatial modes for **CO** and **NO** computed at 12° latitudes -14° through 30° in Figure 5.5 and Figure 4.8 respectively. The 12 latitudes are selected for having consistent day lengths across all longitudes and at least 4 snapshots during daytime. As described above, the optDMD is performed for one latitude at a time to have consistent daytime lengths across all the time series, and the resulting spatial modes are pieced together to present a global picture. The underlying spatial features of the data sets are resolved well by the constrained optDMD diagnostics. The high-variance features at the coastlines and within hot spots in the land for the chemical species are represented clearly.

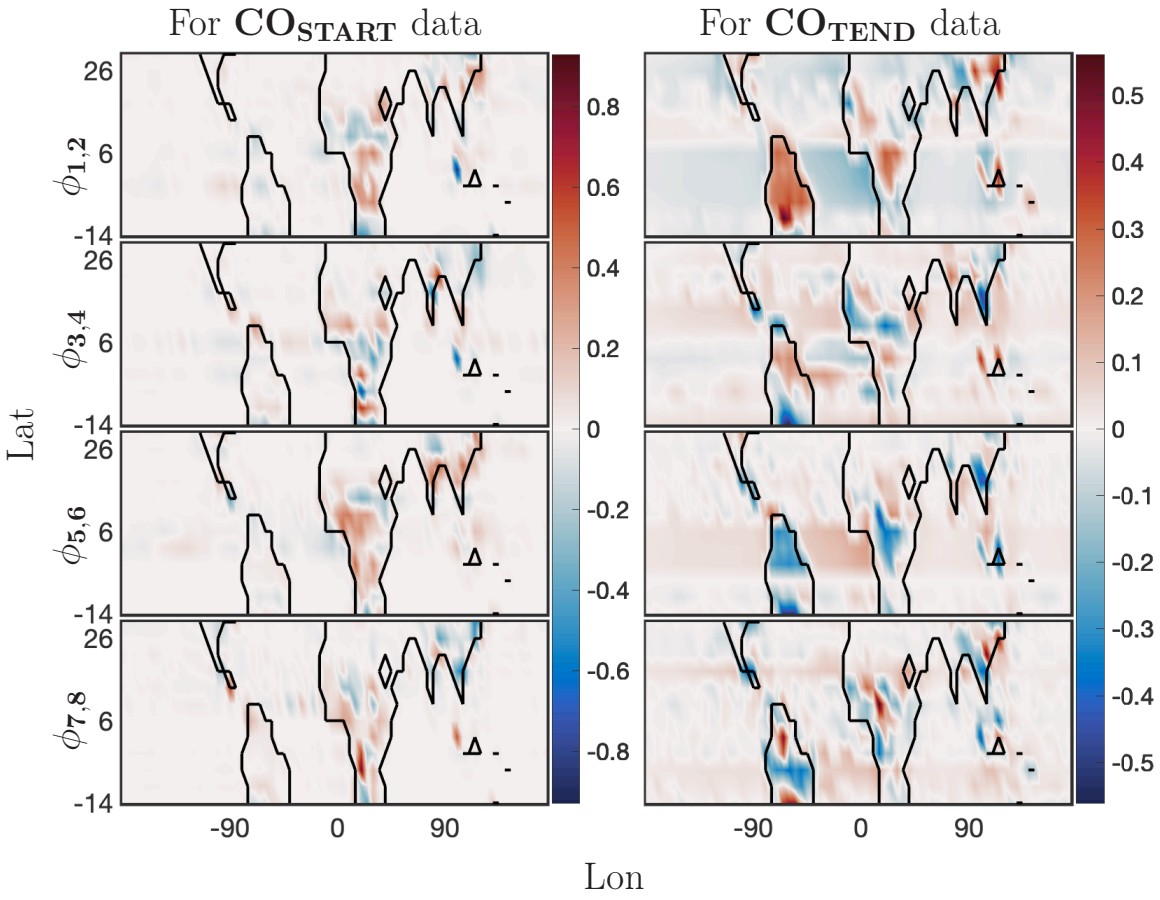


Figure 4.7: 40 *day reconstruction results for Optimized DMD at the surface of* **CO** *preprocessed data. The analysis was computed for 12 latitudes -14° through 30°. The left panel show the dominant four spatial modes for START data; and the right panel show four of the corresponding spatial modes for the TEND data.*

4.5 Forecasting

As described above, using an appropriate rank truncation, the optDMD with eigenvalues constrained to the closed left-half plane faithfully reconstructs the time series data for a 40-day training window and a given elevation/latitude. We now forecast the time series data for future times beyond the training window. Using (4.1), with amplitudes $\mathbf{b}$/modes $\mathbf{\Phi}$/eigenvalues $\mathbf{\Omega}$ computed by optDMD during the training window, we forecast time series

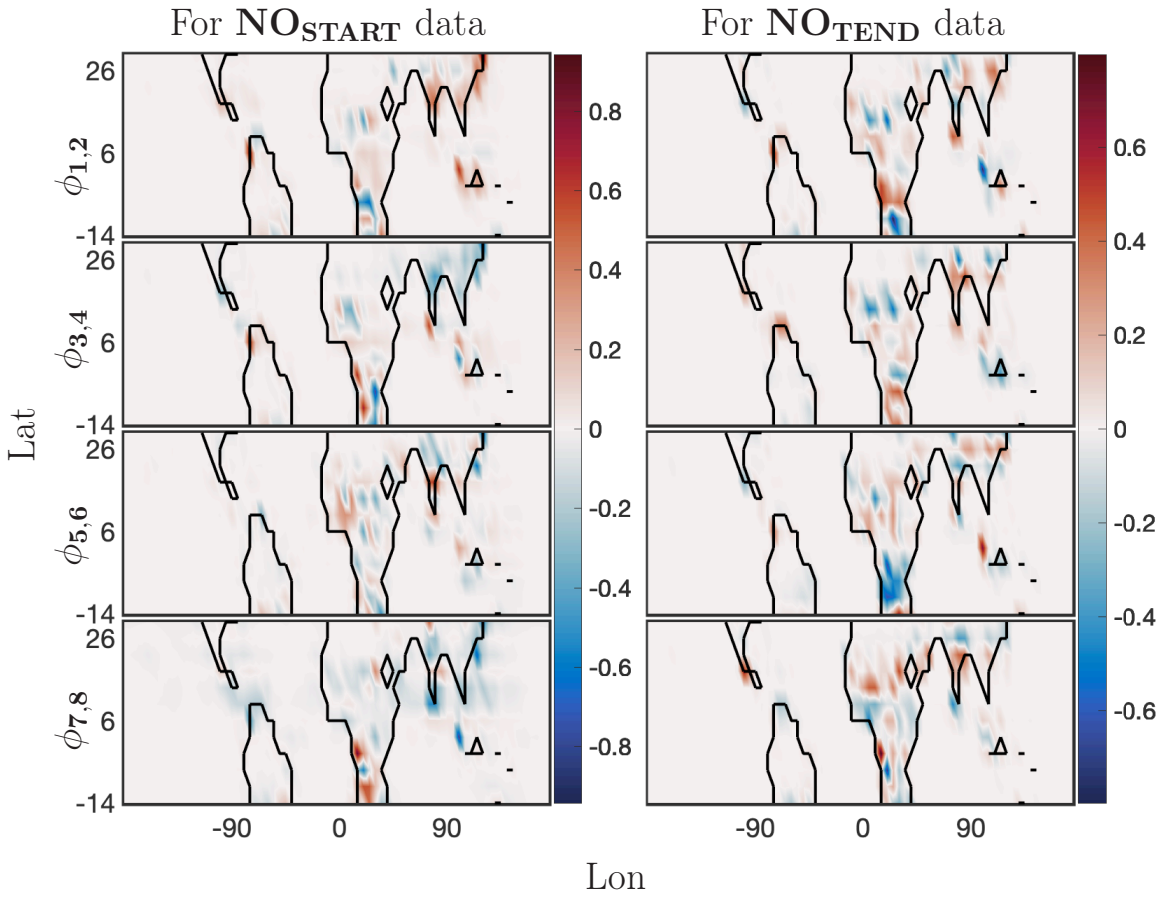


Figure 4.8: 40 *day reconstruction results for Optimized DMD at the surface of* **NO** *preprocessed data. The analysis was computed for 12 latitudes -14° through 30°. The left panel show four spatial modes for START data; and the right panel show four of the corresponding spatial modes for the TEND data.*

for the future 20 days. The results for **START** and **TEND** data for two chemical species **OH** and **NO** are presented for 6 longitudes, and latitude 30° at the surface(elevation=1) in Figures 4.9, 4.10, 4.11, and 4.12.

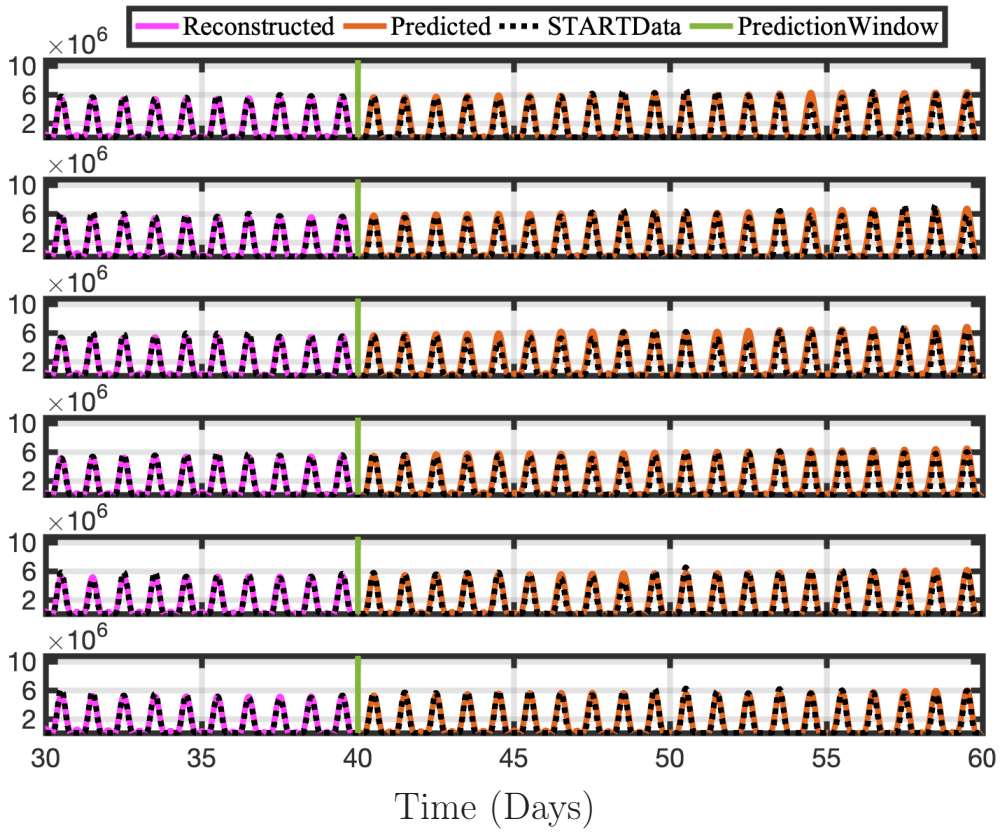


Figure 4.9: *Time series of reconstructed and predicted results with* $\mathbf{OH_{START}}$ *data at Lat 30° and 6 longitudes -180°:5°:-155°. Both the reconstructed data, shown here for 10 days; and the forecasted time series, shown here for the 20-day testing period, faithfully reconstruct and forecast the actual data for* $\mathbf{OH_{START}}$.

Constrained optDMD faithfully reconstructs and forecasts the time series for the 20 days tested. Since we use the fewest modes possible, spikes in actual data are sometimes not reproduced, and we see a sinusoidal best fit time series instead. The $\mathbf{NO_{TEND}}$ results in Figure 4.12 demonstrates this.

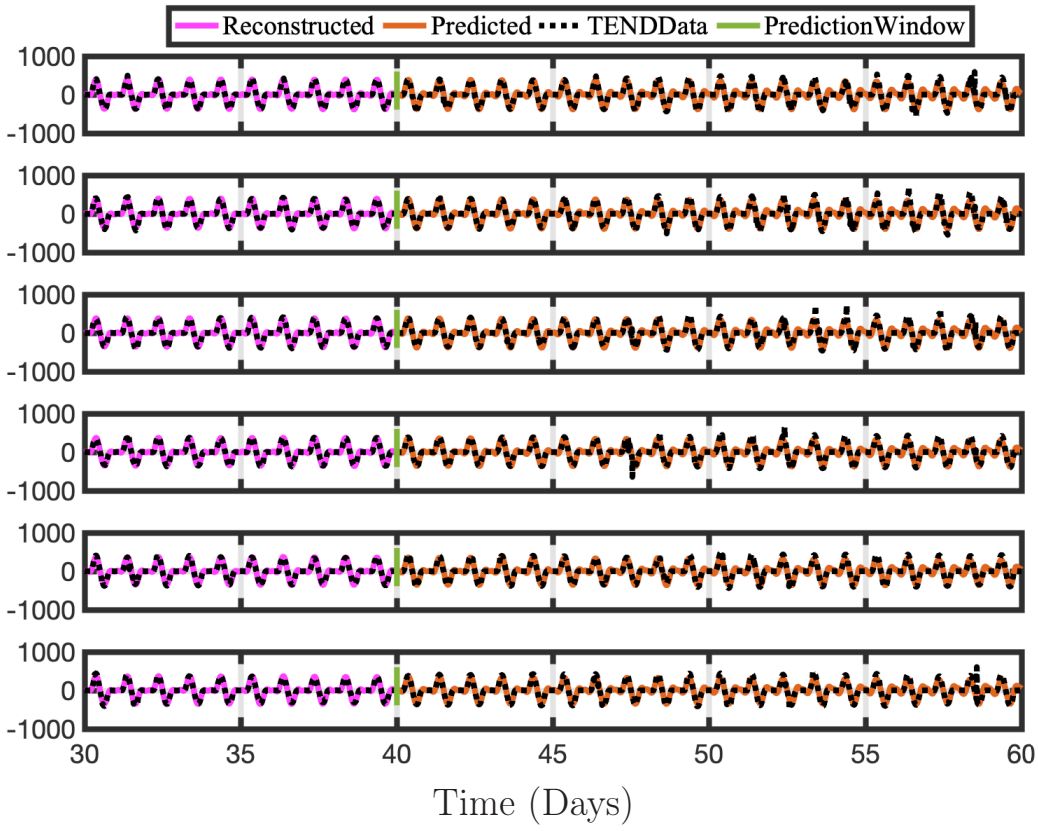


Figure 4.10: *Time series of reconstructed and predicted results with* $\mathbf{OH_{TEND}}$ *data at Lat 30° and 6 longitudes -180°:5°:-155°. Again, both the reconstructed data, shown here for 10 days; and the forecasted time series, shown here for the 20-day testing period, faithfully reconstruct and forecast the actual data for* $\mathbf{OH_{TEND}}$.

We have snapshots of the data every 20-minutes, hence 72 snapshots per day. We compute the relative error for all longitudes for each day, and average across space and snapshots for each day. The resulting mean relative errors are presented for all 6 chemical species of interest and for both **START** and **TEND** data in Figure 4.13 in color red. The 95-percentile confidence intervals for each day are presented as black bars, indicating the variance for the mean relative errors. Constrained optDMD does an excellent job in forecasting the immediate future snapshots and does consistently well during the entire 20-day data tested,

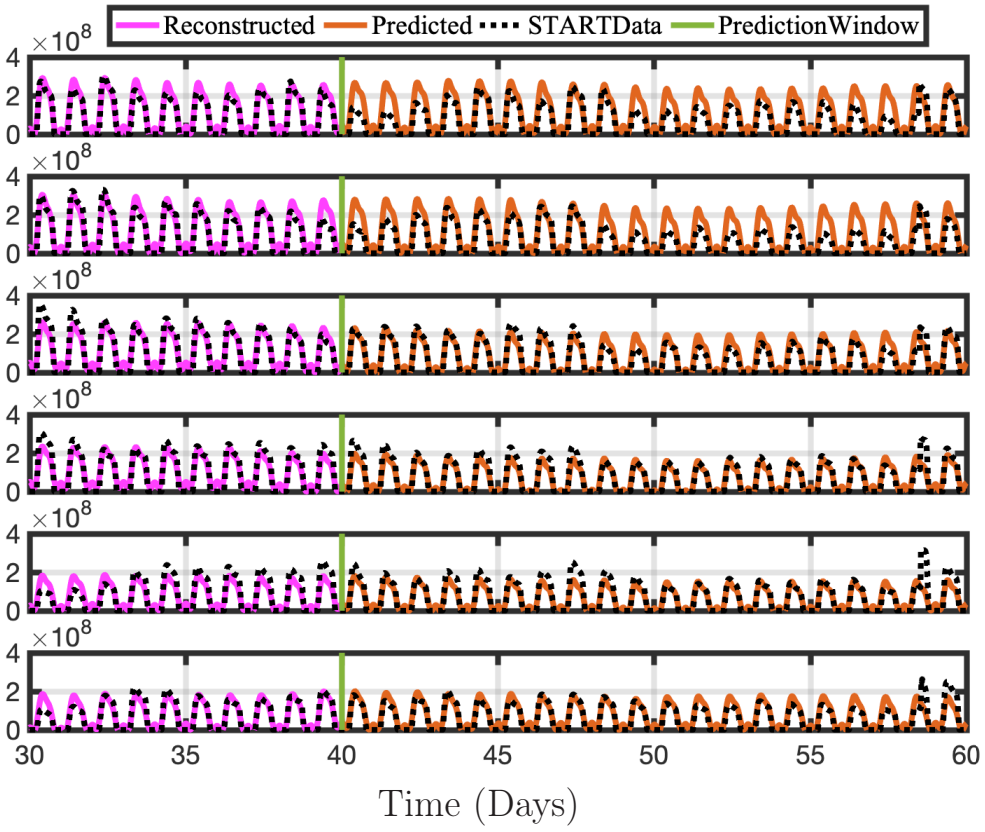


Figure 4.11: *Time series of reconstructed and predicted results with* $\mathbf{NO_{START}}$ *data at Lat 30° and 6 longitudes -180°:5°:-155°. Both the reconstructed data, shown here for 10 days; and the forecasted time series, shown here for the 20-day testing period, reproduce the actual data for* $\mathbf{NO_{START}}$ *well.*

with mean errors/uncertainty in forecasting increasing only slightly for some chemical species as the number of prediction days increases away from the last snapshot used from training. No exponential growth/decay is observed in the forecast time-series, while the underlying dynamics are forecast faithfully. The performance is slightly worse in forecasting the **TEND** data as compared to the **START** data, which is due to the intrinsic rank of the **TEND** data being higher. **TEND** data also exhibits many energetic localized convective phenomena on top of the global slower dynamics. Increasing the truncation rank of the projection will

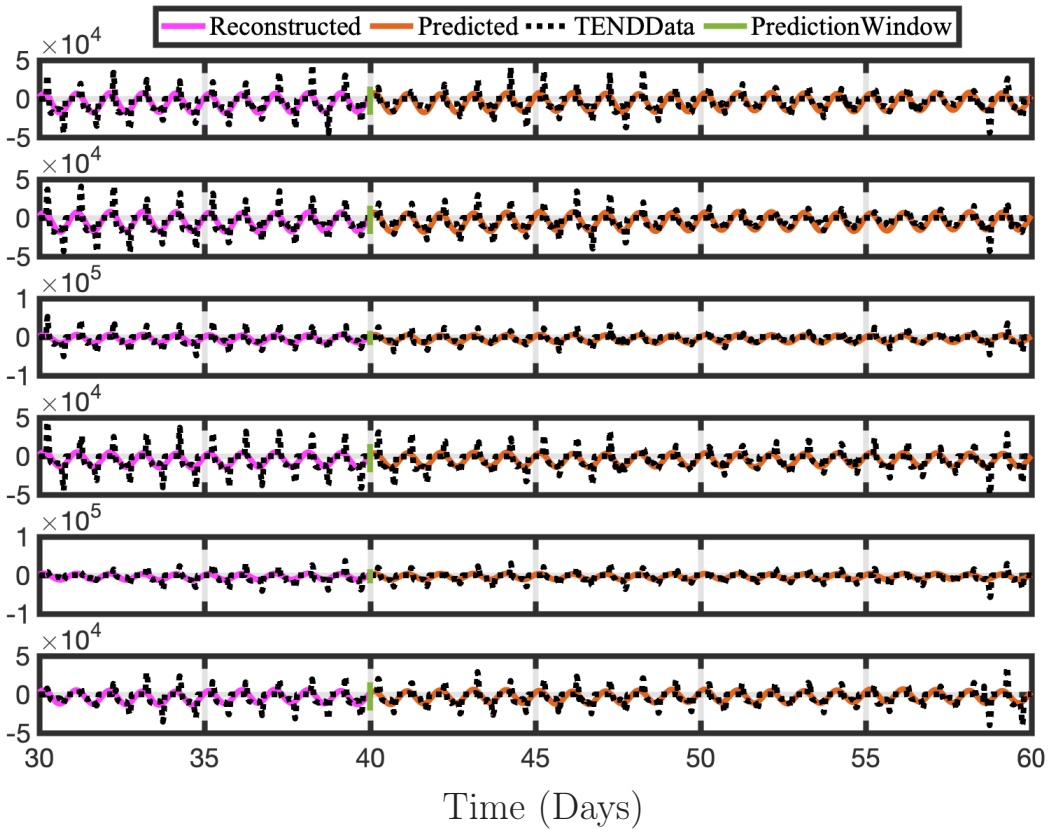


Figure 4.12: *Time series of reconstructed and predicted results with* $\mathbf{NO_{TEND}}$ *data at Lat 30° and 6 longitudes -180°:5°:-155°. Both the reconstructed data, shown here for 10 days; and the forecasted time series, shown here for the 20-day testing period, do not capture the spikes in the actual data for* $\mathbf{NO_{TEND}}$*. Since we are using only 20-30 modes for reconstruction, we get a sinusoidal best fit.*

further refine these convective phenomena, leading to improvement in forecasting. The optDMD performs worst in forecasting the chemical species **OH** (Hydroxyl ion). As can be seen in Figure 4.6 this chemical species needs the highest number of modes to capture the significant energy in **OH** dynamics and have a low relative error in reconstruction as well. Considering that the underlying dynamics represent a moving state with time, the constrained optDMD minimizes model bias with the variable projection optimization, thus

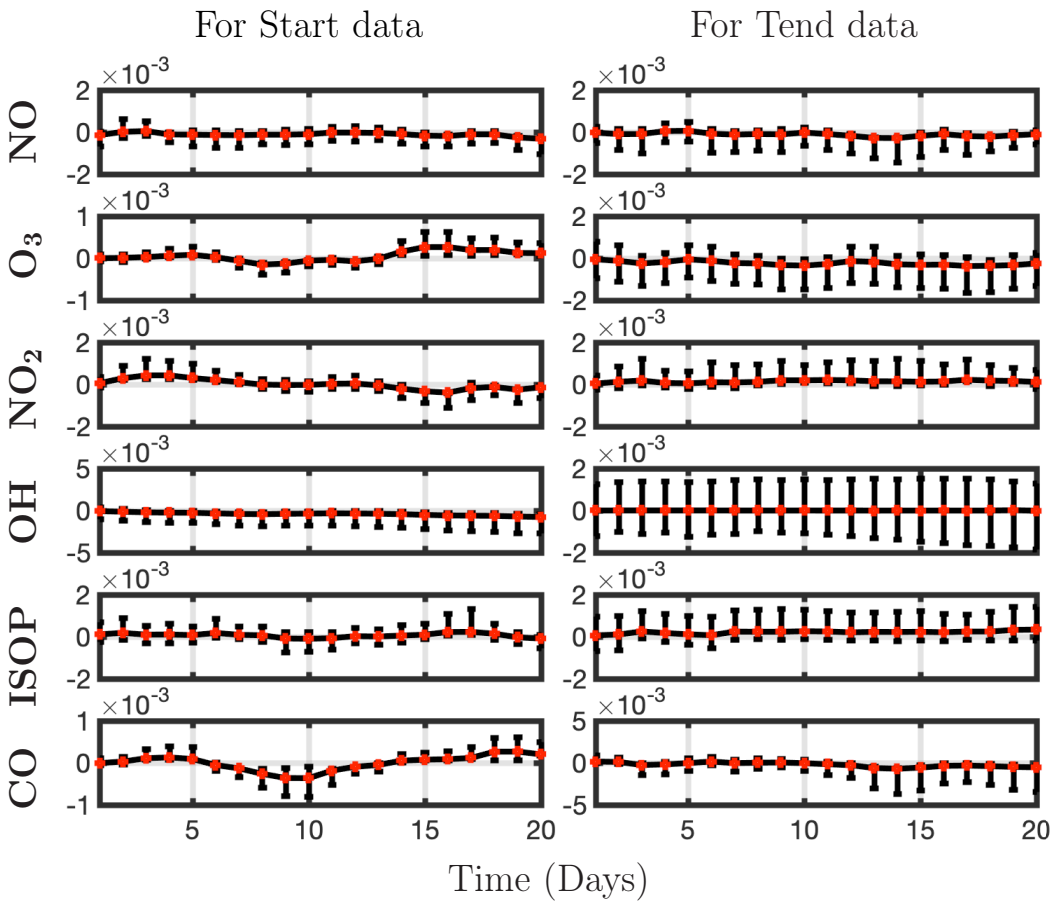


Figure 4.13: *Mean relative error with 95-percentile confidence intervals forecasting* **START** *and* **TEND** *data at Lat 30° for a prediction window of 20 days; and for 6 different chemical species. The relative error stays nearly the same or changes only slightly as the number of days we are forecasting out to increase.* **optDMD** *does better at forecasting* **START** *data as compared to the* **TEND** *data.*

leading to stable forecasting capabilities.

4.6 Temporal Uncertainty Quantification

We now present the results from BOP-DMD in partnership with the optimized DMD

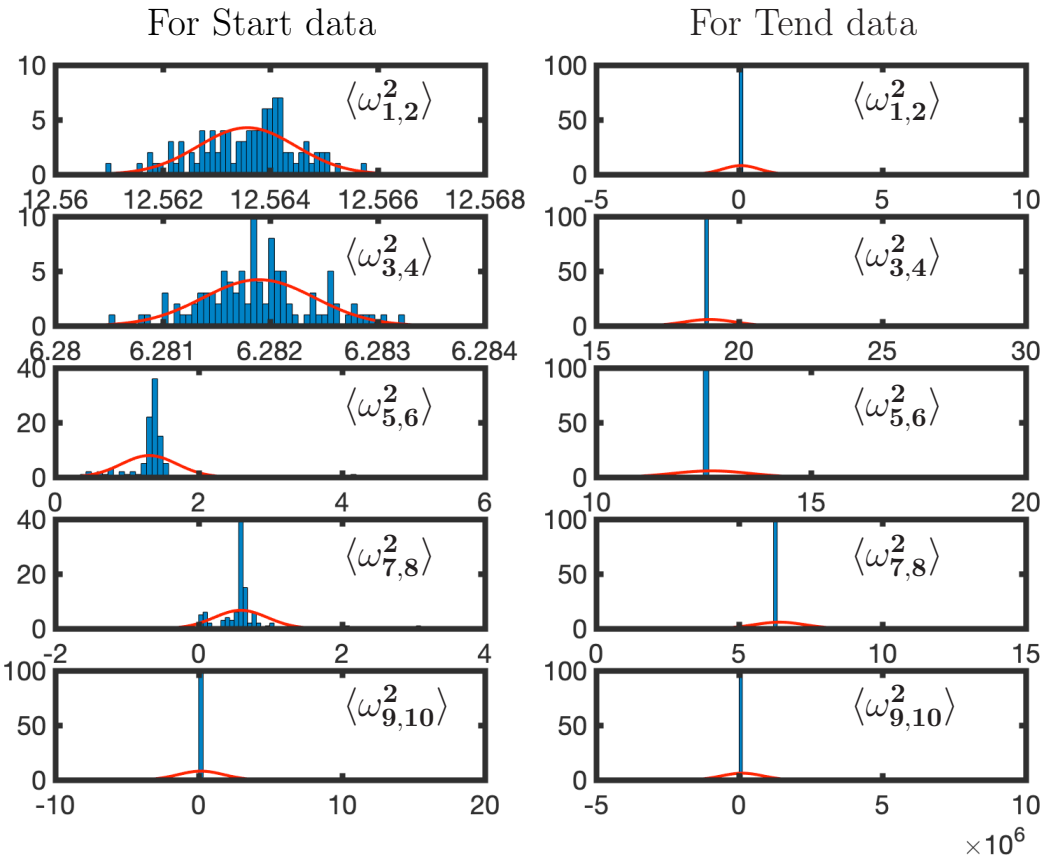


Figure 4.14: *Temporal uncertainty quantification for absolute of eigenvalues for* $\mathbf{OH_{START}}$ *and* $\mathbf{OH_{TEND}}$ *data at Lat 30°. The red lines represent a least-square fit of a normal distribution. 60 days of training data was used with a sample size of 3 days and 100 cycles.*

algorithm to produce ensemble models and compute temporal uncertainty for the eigenvalue spectrum of both **START** and **TEND** data for the six chemical species of interest at Lat 30°. We use the constrained optDMD as described above on a full training data set of 60 days July, 2^{ND} - August, 30^{TH} to create an initial seed $\mathbf{\Phi_0}, \mathbf{\Omega_0}, \mathbf{b_0}$ for the BOP-DMD algorithm. For $K = 100$ trials, we randomly select $p = 216$ snapshots/columns i.e. data for 3 days out of the 60 days to create our subset of data, as shown in Figure 4.3. optDMD now computes the eigenvalues of various subsets using the aforementioned initial conditions. The $K = 100$ ensemble models' eigenvalues are used to produce the temporal UQ metrics.

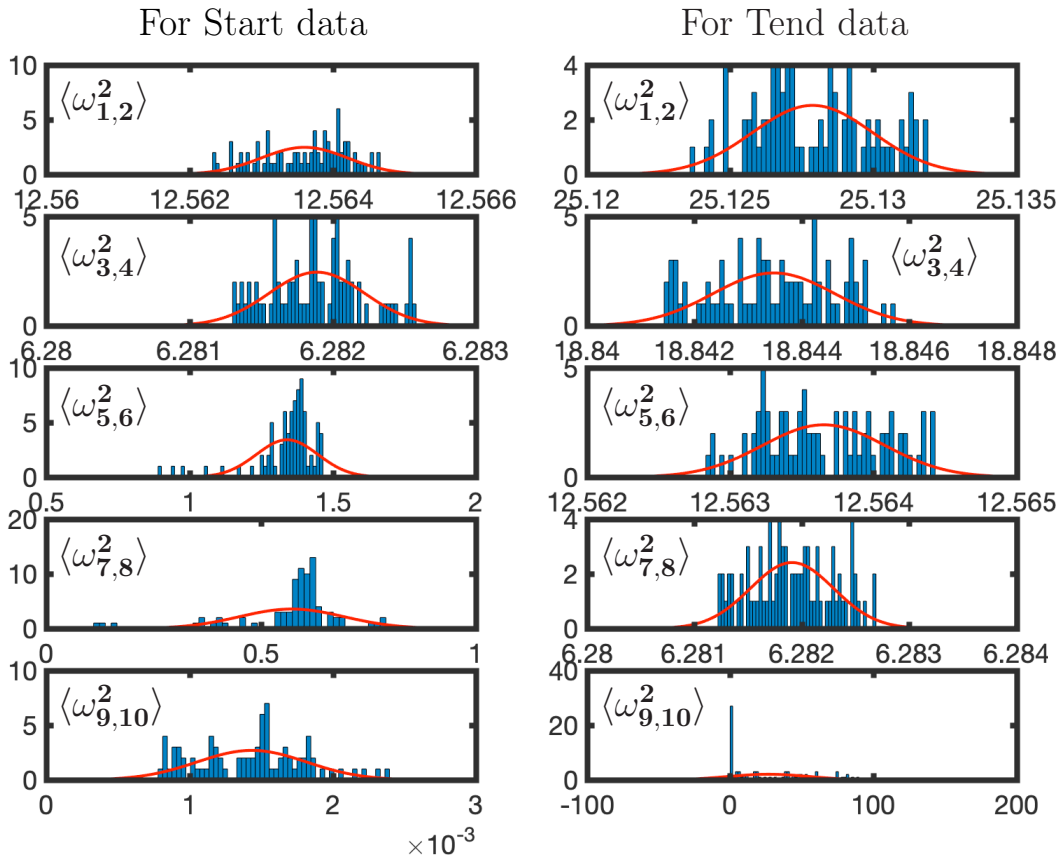


Figure 4.15: *Temporal uncertainty quantification for absolute value of trimmed eigenvalues for with* $\mathbf{OH_{START}}$ *and* $\mathbf{OH_{TEND}}$ *data at Lat 30°. The data has been trimmed to remove outliers below 10 percentile and above 90 percentile. The red lines represent a least-square fit of a normal distribution.*

Figure 4.14 shows the BOP-DMD distributions of the absolute value of the first five eigenvalues for each of the subsets of data for $\mathbf{OH_{START}}$ and $\mathbf{OH_{TEND}}$ data at Lat 30°. The BOP-DMD quantifies the temporal uncertainly by allowing for a Gaussian fit, shown in red. For both of the data sets, we see a high temporal uncertainty in eigenvalues, with outliers skewing the distributions. The temporal uncertainty gets worse for the higher modes

in the $\mathbf{OH_{START}}$ data and for all modes of $\mathbf{OH_{TEND}}$ data. Then we trim the eigenvalue distribution data to exclude the outliers below 10-*percentile* and above 90-textitpercentile to improve the UQ metrics. Figure 4.15 shows the distributions of the trimmed absolute eigenvalues, and the Gaussian fit is better with lower variances, and only 1 distribution with outliers. Still, we see that there is significant temporal variability, especially for higher modes for $\mathbf{OH_{TEND}}$.

4.7 Conclusion

Based on the results presented in this work, we believe the constrained optDMD is the DMD algorithm of choice for the reconstruction and forecasting of global atmospheric data. Exact DMD fails in the task of reconstructing the chemistry time-series it is regressed to, let alone produce a reasonable forecast. This is due to the significant bias in the model from energetic localized convective phenomena present in the atmospheric simulation data. [5] optDMD algorithm casts the regression problem as a nonlinear optimization enabled by variable projection techniques, hence providing an optimal de-biasing for the atmospheric chemistry dynamics. The optDMD is thus better able to capture hidden dynamics, showing an order of magnitude improvement in the reconstruction error. optDMD also produces modes which more accurately describe the localized energetic convective phenomena in the **START** and especially the **TEND** chemistry dynamics. The nonlinear optimization problem in the optDMD also allows for constraints. By adding a constraint $\Re(\omega_i \leq 0)$ to the optDMD minimization, we obtain superior eigenvalues that are able to produce high-fidelity, stable and robust forecasts. For the entire testing time window, the forecasts remain accurate as we increase time away from the training time window, not displaying any growth, decay or loss of accuracy. However, computing the optDMD requires solution of a nonlinear, nonconvex optimization problem, which often fails to converge to a solution. The computational cost of the optDMD is higher, as we increase the number of snapshots, the cost increase becomes more significant. The solutions obtained here nevertheless represent significant improvements. Partnering the optDMD algorithm with the statistical bagging and ensembling of the BOP-DMD produces temporal UQ metrics, and highlights the high temporal variance in the eigenvalues produced by optDMD. This temporal variance gets

worse for higher modes of the **START** data; eigenvalues for the **TEND** data have quite high temporal variance.

An interesting further direction would be to apply the optDMD to an entire year's worth of data, a still computationally tractable problem. It would be interesting to see if the optDMD can faithfully reproduce yearly patterns in the chemistry data, and accurately forecast seasonal variations. We can leverage the BOP-DMD further to produce spatial UQ metrics, illustrating the spatial patterns where optDMD is most uncertain in its ability to provide accurate representations. optDMD can be further empowered by partnering with the BOP-DMD by (i) an initialization procedure to stabilize its convergence, improving the robustness and accuracy of the regression, (ii) leveraging statistical bagging to produce a stable model with reduced variance in the model parameters, and (iii) leveraging this stable model to forecast future states of spatio-temporal atmospheric chemistry system, with Monte Carlo simulations to produce UQ for future states.

Chapter 5

OPTIMAL SPARSE SENSOR PLACEMENT FOR RECONSTRUCTION OF GLOBAL ATMOSPHERIC CHEMISTRY DATA

Optimal sensor placement determines placement of sensors in complex, evolving systems to optimize several downstream objectives, such as full state reconstruction and prediction from point measurements. Sensors provide local state measurements from which global properties of the state may be inferred. The sensor locations have to be determined from a massive set of possible locations, typically amounting to a brute-force search among combinatorial possibilities. For small scale problems, this approach has been successful [30]. For moderate sized search spaces there are well-known model based solutions using optimal experiment design [18, 63], and information theoretic and Bayesian criteria [28, 54, 75, 113, 94].

Scalable optimization of sensor location for high-dimensional nonlinear dynamical systems is still a challenging problem. One promising indicator of making this problem tractable is the fact that most high-dimensional systems, such as are found in fluids, epidemiology, neuroscience, atmospheric and earth systems etc. typically exhibit dominant coherent structures that evolve in a low-dimensional subspace. These low-dimensional patterns are often identified using dimensionality reduction techniques [67]. Dimensionality reduction is a critically enabling aspect of machine learning and data science that can be leveraged to identify and exploit low-dimensional patterns and features in high-dimensional systems. Proper orthogonal decomposition (POD) is one of the most important and widely used data-driven dimensionality reduction technique available to analyze high-dimensional complex spatio-temporal systems such as turbulent fluid flows [116, 16, 60], structural mechanics and vibrational analysis [66, 7, 56, 73], neuroscience [10, 65, 119], atmospheric sciences [125, 126], where it is called empirical orthogonal functions (EOFs), to name a few.

Moreover, key innovations in signal recovery are exploiting the *sparsity* or compressible

nature of signals, i.e. the idea that most natural signals have only a few active or nonzero components when expressed in a generic basis. For e.g., the theory of *compressed sensing* [26, 37, 11] leverages the geometry of sparse vectors in high dimensional space to provide convex algorithms to solve the combinatorial sparse signal reconstruction problem. In [21] the theory of compressed sensing was applied to design optimal sparse sensor locations for classification decisions based on high-dimensional data. Compressed sensing sparsity-promoting algorithms such as the *lasso* regression [118] in combination with machine learning [57, 62] have been widely applied to characterize and control dynamical systems [111, 92, 23, 98, 22, 81, 78, 106, 41, 110], including modeling high-dimensional fluid systems using POD [9].

This compressed sensing strategy is ideal for the recovery of a high-dimensional signal of 'unknown' content using random measurements in a universal basis. However, if information is available about the type of signal (such as the signal is a turbulent velocity field), it is possible to design optimized sensors that are tailored for the particular signals of interest. Recently, [82] explored optimized sensor placement for signal reconstruction based on a *tailored* library of dominant features extracted from training data consisting of representative examples of the system using singular value decomposition based methods such as POD. *Empirical interpolation methods* (EIMs) seek the best interpolation points for such a a given basis of POD features mined from patterns in the data to facilitate the discovery of sparse optimal sensors using QR pivoting. Drastic reductions in the number of sensors required with improved reconstruction results were observed in examples ranging from facial reconstruction to fluid vorticity fields. This architecture has also been successfully applied to sparse approximation for insect flight dynamics [83], sensor placement for predictive manufacturing [84], optimized sampling for multiscale dynamics [85], and sensor and actuator placement for optimal control [86].

In this work we design data-driven scalable optimized sensor placements for global atmospheric chemistry signal reconstruction based on this data-driven, scalable, sparse sensor placement architecture. The dynamics of atmospheric chemistry is characterized by complex interactions among hundreds of chemical species which can produce kinetics across temporal scales spanning many orders of magnitude, from microseconds to minutes.

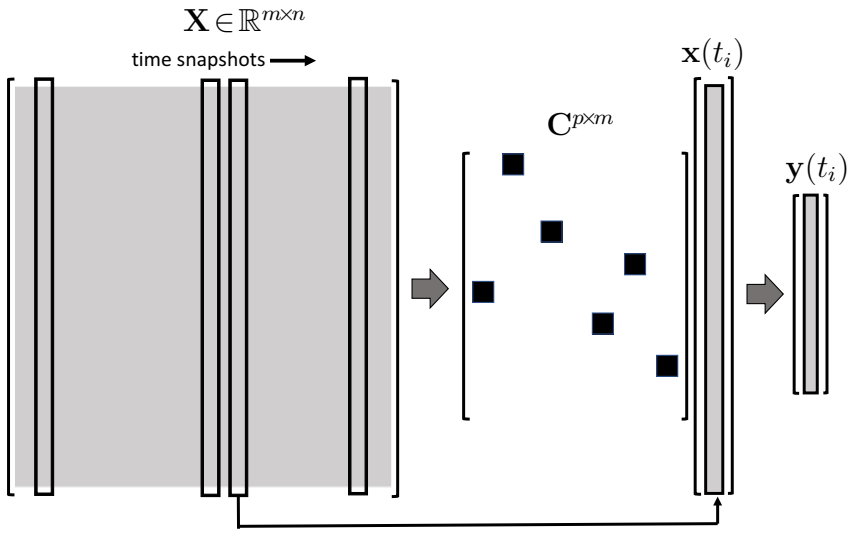


Figure 5.1: *Atmospheric chemistry simulation on a global mesh with discretized longitude, latitude and elevation generates the original training data matrix* $\mathbf{X} \in \mathbb{R}^{n \times m}$*, where* m *is the number of grid points and* n *is the number of snapshots. The goal of this work is to design an optimal measurement matrix* $\mathbf{C}$ *that sample* p *measurements of the full state, represented in a low dimensional POD subspace, for sparse point measurements* $\mathbf{y_i}$*, such that we can accurately reconstruct the full state* $\mathbf{x}$ *from* $\mathbf{y}$*.*

Accurate monitoring and prediction requires full knowledge of the chemical state of the atmosphere at all locations and times, resulting in a 5-dimensional data set for longitude, latitude, elevation, species and time. Well resolved simulations generate massive data sets that are often not amenable to diagnostic analysis. The proposed algorithm offers an alternative scalable architecture for monitoring and prediction of global atmospheric chemistry. First, the POD learns from the surface atmospheric chemistry training data sets to extract dominant coherent low-dimensional patterns, and builds a tailored library of POD basis functions providing compact representation of the high-dimensional chemistry states. Next, QR pivoting is applied to leverage the POD basis to pick the most optimized sparse

sensors that can estimate the global atmospheric chemistry system with surprisingly few point measurements.

5.1 Methods for optimal sparse sensing of atmospheric chemistry data

The instantaneous reproduction of atmospheric chemistry data in this work is based on combining two key techniques: 1) *machine learning*, which exploits patterns in the data for low-dimensional representation in a tailored POD basis, and 2) *sparse sampling*, where we reconstruct the full state from a small subset of measurements from optimized sensor locations. This combination is *synergistic*, in that underlying low-rank representations mined from patterns in the chemistry data are leveraged to facilitate the design of specialized sensors.

POD low-rank representations have already been used in Reduced Order Modeling or ROM community to approximate high-dimensional spatio-temporal systems by low dimensional subspaces that produce nearly identical characteristics. *Empirical interpolation methods* seek the best interpolation points for a given POD basis to speed up the evaluation of high-order nonlinear terms in the high-dimensional, parametrized systems [24]. However, these resulting interpolation points correspond to measurements in the state space, hence they can also be used for data-driven sensor location selection. We will be present this formulation of sensor selection and explore a sparse, convex, and greedy optimization method to solve it based on [82].

5.2 Proper Orthogonal Decomposition

POD is the singular value decomposition (SVD) algorithm applied to partial differential equations that expresses high-dimensional states $\mathbf{x} \in \mathbb{R}^m$ as a small linear combination of orthonormal spatial eigenmodes, i.e. POD modes $\boldsymbol{\Psi}(x)$ that define the low-dimensional embedding space. A low dimensional representation of the state $\mathbf{x}$ can be lifted to the full state by a linear combination of the POD modes and their corresponding weights or POD coefficients $\mathbf{a}$:

$$\mathbf{x}_i \approx \sum_{j=1}^{r} a_j(t_i)\boldsymbol{\psi}_j(x) \tag{5.1}$$

where for time-series data $\mathbf{x}_i$, the POD coefficients $a_j(t_i)$ are time-varying and the POD modes $\boldsymbol{\psi}_j(x)$ are spatial without any time dependence, resulting in a space-time separation of variables.

Building this low-rank embedding space requires training data to *tailor* the POD modes to the specific problem at hand. The POD eigenmodes $\boldsymbol{\psi}_j(x)$ and the POD coefficients $a_j(t_i)$ are easily obtained from the SVD of the training data matrix. The dynamics of the atmospheric chemistry system are sampled at fixed time intervals of 20 minutes to generate our training data set $\mathbf{X} \in \mathbb{R}^{m \times n}$, as shown in Figure 5.1, where m is the number of grid points and n is the total number of snapshots of data. SVD provides a unique matrix decomposition for this data matrix as:

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathsf{T}} \tag{5.2}$$

where $\mathbf{U} \in \mathbb{R}^{\mathbf{m} \times \mathbf{m}}$ and $\mathbf{V} \in \mathbb{R}^{\mathbf{n} \times \mathbf{n}}$ are *unitary* matrices and $\boldsymbol{\Sigma} \in \mathbb{R}^{\mathbf{m} \times \mathbf{n}}$ is the diagonal matrix of non-negative singular values. Here $^{\mathsf{T}}$ denotes the transpose. The columns of $\mathbf{U}$ are the left singular vectors of $\mathbf{X}$, guaranteed to provide the best set of modes to approximate $\mathbf{X}$ in an ℓ_2 sense. We seek a minimal number of modes r necessary to accurately represent the dynamics of atmospheric chemistry data, i.e. a rank-r approximation to the true dynamics, where typically $r \ll m, n$. This is the low-rank SVD given by:

$$\tilde{\mathbf{X}} = \tilde{\mathbf{U}}\tilde{\boldsymbol{\Sigma}}\tilde{\mathbf{V}}^{\mathsf{T}} \tag{5.3}$$

where $\tilde{\mathbf{U}} \in \mathbb{R}^{m \times r}$, $\tilde{\mathbf{V}} \in \mathbb{R}^{n \times r}$, and $\tilde{\boldsymbol{\Sigma}} \in \mathbb{R}^{r \times r}$. Thus, we are keeping only the first r singular values, or *truncating* at rank$= r$. The optimal POD basis modes are given by:

$$\tilde{\mathbf{U}} = \boldsymbol{\Psi}_{\mathbf{r}} = \begin{bmatrix} | & | & | \\ \psi_1 & \cdots & \psi_{\mathbf{r}} \\ | & | & | \end{bmatrix} \tag{5.4}$$

The SVD is the optimal least squares approximation to the data for a given rank r as proved by Eckart-Young theorem [45]:

$$\tilde{\mathbf{X}} = \arg\min_{\hat{\mathbf{X}}} \|\mathbf{X} - \hat{\mathbf{X}}\|_F \quad \text{s.t. } \operatorname{rank}(\hat{\mathbf{X}}) = r \tag{5.5}$$

The low-dimensional vector of POD coefficients for a given state $\mathbf{x}$ is obtained from the orthogonal projection $\mathbf{a} = \boldsymbol{\Psi}_r^{\mathsf{T}}\mathbf{x}$.

Choosing the target low rank r that adequately captures most of the dominant low-rank dynamical features in the high-dimensional data without magnifying the noise in the data is a difficult task [24]. There are many factors that can determine where to truncate, like a desired rank of the system, distribution of the singular values, the decay rate of the singular value spectrum, the magnitude of noise in the system. Often the rank r is chosen by truncating the singular values to capture a predetermined amount of variance or energy in the data, or identifying 'elbows' or 'knees' in the singular value distribution and truncating at these points. Truncation is a hard threshold on the singular values. Recent work by [51] provides an optimal hard threshold based on singular value distribution as well as the aspect ratio of the data matrix, assuming additive Gaussian noise of unknown variance. This hard optimal threshold τ has been effective in practice, even when the noise is likely not Gaussian. We would be using a hybrid of both approaches here.

5.3 Sparse sensor placement with QR pivoting

The ability of sparse sensor measurements to accurately reconstruct the full state of the system is critically dependent on the placement of the sensor location [24]. Here is presented the framework to optimize the sensor locations specifically to reconstruct high-dimensional states from sparse point measurements, given a tailored basis [82]. The low rank-r representation of the state $\mathbf{x} \in \mathbb{R}^m$ in the tailored basis can be expressed as:

$$\mathbf{x} = \mathbf{\Psi}_r \mathbf{a} \tag{5.6}$$

where $\mathbf{a} \in \mathbb{R}^r$ is a sparse vector indicating which modes of $\mathbf{\Psi}_r$ are active. The goal now is to design a point measurement sampling matrix $\mathbf{C} \in \mathbb{R}^{p \times m}$, where $p \ll m$ is a few optimized point measurements:

$$\mathbf{y} = \mathbf{C}\mathbf{x} \tag{5.7}$$

where $\mathbf{y} \in \mathbb{R}^p$. $\mathbf{y}$ should enable accurate reconstruction of $\mathbf{a}$, and hence $\mathbf{x}$. Combining Equations 5.6 and 5.7 gives our *sparse sensing* problem:

$$\mathbf{y} = (\mathbf{C}\mathbf{\Psi}_r)\mathbf{a} = \mathbf{\Theta}\mathbf{a} \tag{5.8}$$

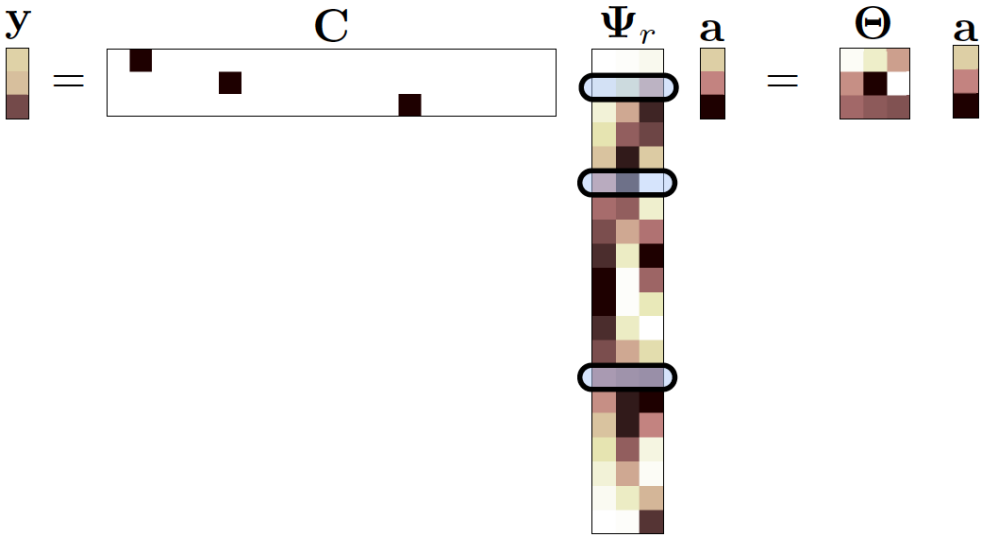


Figure 5.2: *Reproduced with permission from [82]. Full state reconstruction of* $\mathbf{x}$ *from point observations* $\mathbf{y}$ *is accomplished using least squares estimation of the POD coefficients, i.e.* $\mathbf{a} = \boldsymbol{\Theta}^{\dagger}\mathbf{y}$

where *effective* measurements $\mathbf{C}$ given the tailored basis $\boldsymbol{\Psi}_r$ are chosen such that the operator $\boldsymbol{\Theta}$ is well-conditioned for full-state reconstruction. The sparse sensing problem schematic is presented from [82] in Figure 5.2.

From Equation 5.1 the full states are also represented as an unknown linear combination of the tailored basis, rewriting it in the coordinate form:

$$x_i = \sum_{j=1}^{r} \Psi_{kj} a_j \tag{5.9}$$

Hence, we need to design a point measurement sampling matrix $\mathbf{C} \in \mathbb{R}^{p \times m}$ that is optimized to recover the modal coefficients $\mathbf{a}$ from the sensor measurements $\mathbf{y}$. Point measurements structure the measurement matrix $\mathbf{C}$ as:

$$\mathbf{C} = \begin{bmatrix} \mathbf{e}_{\gamma_1} & \mathbf{e}_{\gamma_2} & \cdots & \mathbf{e}_{\gamma_p} \end{bmatrix}^{\mathsf{T}} \tag{5.10}$$

where $\mathbf{e}_j$ are the canonical basis vectors for $\mathbb{R}^m$. This results in the linear system written in the coordinate form:

$$\mathbf{y}_i = \sum_{j=1}^{m} C_{ij} x_j = \sum_{j=1}^{m} C_{ij} \sum_{k=1}^{r} \Psi_{jk} a_k \tag{5.11}$$

Hence the observations in $\mathbf{y}$ consists of p elements (sparse measurements) of $\mathbf{x}$:

$$\mathbf{y} = \mathbf{C}\mathbf{x} = \begin{bmatrix} x_{\gamma_1} & x_{\gamma_2} & \cdots & x_{\gamma_p} \end{bmatrix}^{\mathsf{T}} \tag{5.12}$$

An accurate reconstruction of the full state $\hat{\mathbf{x}}$ is now obtained by combining Equations 5.8 and 5.11:

$$\hat{\mathbf{x}} = \mathbf{\Psi}_r \hat{\mathbf{a}}, \text{ where } \hat{\mathbf{a}} = \begin{cases} \mathbf{\Theta^{-1}}\mathbf{y} = (\mathbf{C}\mathbf{\Psi}_r)^{-1}\mathbf{y}, & p = r, \\ \mathbf{\Theta}^{\dagger}\mathbf{y} = (\mathbf{C}\mathbf{\Psi}_r)^{\dagger}\mathbf{y}, & p > r \end{cases} \tag{5.13}$$

where † is the Moore-Penrose pseudoinverse. Denote the matrix to be inverted as $\mathbf{M}_\gamma = \mathbf{\Theta}^T\mathbf{\Theta}$ (with $\mathbf{M}_\gamma = \mathbf{\Theta}$ if $p = r$). The optimal sensor locations are the ones that would enable the most accurate, the best possible reconstruction $\hat{\mathbf{x}}$, i.e. select the rows of $\mathbf{\Psi}_r$ that optimally condition the inversion of matrix $\mathbf{M}_\gamma$. Here the dependency on γ is due to the fact that it determines the sensor locations, hence the condition number for $\mathbf{M}_\gamma$. This leads to the formulation of the following optimization problem, find γ_* such that:

$$\gamma_* = \underset{\gamma, |\gamma|=p}{\arg\max} |\det \mathbf{M}_\gamma| = \underset{\gamma, |\gamma|=p}{\arg\max} \prod_i \sigma_i(\mathbf{M}_\gamma) \tag{5.14}$$

where $|\gamma|$ is the ℓ_1 norm (sum of the absolute values of the vector, in this case equal to the number of nonzero entries). *Empirical interpolation methods* provide near optimal sampling of a system to compute interpolation points that enable accurate reconstructions of high-order nonlinear terms in ROMs [24, 12, 29, 39]. We build upon one of the variants, Q-DEIM [39] which leverages a matrix QR factorization with column pivoting of $\mathbf{\Psi}_r^{\mathsf{T}}$ to compute the best measurement points for our sparse sensing problem.

The reduced matrix QR factorization with column pivoting decomposes a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ into an orthogonal $\mathbf{Q} \in \mathbb{R}^{m\times n}$, an upper triangular matrix $\mathbf{R} \in \mathbb{R}^{n\times n}$, and a column permutation matrix $\mathbf{C}$ such that $\mathbf{A}\mathbf{C}^{\mathsf{T}} = \mathbf{Q}\mathbf{R}$. Recall that the determinant of a matrix, when expressed as a product of an orthogonal factor and an upper-triangular factor, is the product of the diagonal entries in the upper-triangular factor:

$$|\det \mathbf{A}\mathbf{C}^{\mathsf{T}}| = |\det \mathbf{Q}||\det \mathbf{R}| = \prod_i |r_{ii}| \tag{5.15}$$

The pivoting algorithm thus provides an approximate greedy solution to the optimization problem in Equation 5.14, since it permutes the matrix $\mathbf{\Psi}_r^{\mathsf{T}}$ with $\mathbf{C}^{\mathsf{T}}$ to enforce a diagonally

dominant structure for $\mathbf{R}$, maximizing the product of the singular values:

$$\sigma_i^2 = |r_{ii}|^2 \geq \sum_{j=i}^{k} |r_{jk}|^2; \quad 1 \leq i \leq k \leq m \tag{5.16}$$

The QR factorization with column pivoting thus gives the best p point sensors or pivots that best sample or characterize the dominant r basis dynamical modes $\mathbf{\Psi}_r$ by:

$$\mathbf{\Psi}_r^{\mathsf{T}} \mathbf{C}^{\mathsf{T}} = \mathbf{QR}, \quad \text{for} \quad p = r \tag{5.17}$$

$$(\mathbf{\Psi}_r \mathbf{\Psi}_r^{\mathsf{T}}) \mathbf{C}^{\mathsf{T}} = \mathbf{QR}, \quad \text{for} \quad p > r \tag{5.18}$$

The $p = r$ case was developed in [39] and is referred to as the Q-DEIM case. A major contribution of [82] was extending this to the *over sampled* case with $p > r$.

5.4 Incremental SVD updates for updating the library of POD modes

The time-series measurements of the concentrations (**START** data) and rates of change of concentrations (**TEND** data) of chemical species record a new snapshot every 20 minutes. As discussed below, from 89 total days of preprocessed data, we select a training window of the first 60 days of data, and the rest of the 29 days are the test or validation snapshots. Sensors and POD basis modes are hence only trained on the first 60 days worth of data and are tested on their reproduction abilities against the test snapshots.

These data sets, the **TEND** data set in particular, have intermittent spatially localized moderately energetic or low energy temporal convective features that also have to be tracked by the sensors. However, modal separation of intermittent convective phenomenon is difficult from a time-invariant POD analysis. Separating isolated, low-energy temporal events cannot be done by a variance characterizing SVD based method such as POD. To test whether updating the library of basis modes $\mathbf{\Psi}_r$ can enable better tracking of these intermittent phenomena, we implement the incremental SVD updates that continually update $\mathbf{\Psi}_r$ as each new snapshot comes in. Since we only have point measurements of data, we implement the incremental SVD algorithm by [19].

As defined in Equation 5.3, the reduced rank-r SVD has been computed, with the optimal POD basis modes given by the columns of $\tilde{\mathbf{U}}$, i.e., $\tilde{\mathbf{X}} = \mathbf{\Psi}_r \tilde{\mathbf{\Sigma}} \tilde{\mathbf{V}}^{\mathsf{T}}$. Consider new data,

i.e. a new full state measurement column $\mathbf{z} \in \mathbb{R}^m$, is now to be added to the low rank representation of the original data matrix as $\tilde{\mathbf{X}}' = [\tilde{\mathbf{X}} \quad \mathbf{z}]$. We now need to update the library $\mathbf{\Psi}_r$ by computing an incremental SVD update. Consider the projections of the new measurement $\mathbf{z}$ within and orthogonal to the subspace spanned by $\mathbf{\Psi}_r$:

$$\mathbf{z}_{\|} = \mathbf{\Psi}_r^{\mathsf{T}} \mathbf{z} \tag{5.19}$$

$$\mathbf{z}_{\perp} = (\mathbf{I} - \mathbf{\Psi}_r \mathbf{\Psi}_r^{\mathsf{T}}) \mathbf{z} \tag{5.20}$$

The parallel component $\mathbf{z}_{\|}$ causes the singular values and the space $\mathbf{\Psi}_r$ to be rotated, while the orthogonal component will effectively increase the rank of the SVD. Define vector $\mathbf{u} = \mathbf{z}_{\perp} / \|\mathbf{z}_{\perp}\|$. Consider the identity:

$$\begin{bmatrix} \mathbf{\Psi}_r & \mathbf{u} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{\Sigma}} & \mathbf{\Psi}_r^{\mathsf{T}} \mathbf{z} \\ 0 & \|\mathbf{z}_{\perp}\| \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{V}} & 0 \\ 0 & \mathbf{I} \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} \mathbf{\Psi}_r \tilde{\mathbf{\Sigma}} \tilde{\mathbf{V}}^{\mathsf{T}} & \mathbf{z} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{X}} & \mathbf{z} \end{bmatrix} = \tilde{\mathbf{X}}' \tag{5.21}$$

As in a SVD, the left and right matrices in the product are unitary and orthogonal. The middle matrix is diagonal with a one-column border. To compute the updated SVD:

$$\tilde{\mathbf{X}}' = \mathbf{\Psi}_r' \tilde{\mathbf{\Sigma}}' \tilde{\mathbf{V}}'^{\mathsf{T}} \tag{5.22}$$

first the middle matrix is diagonalized:

$$\begin{bmatrix} \tilde{\mathbf{\Sigma}} & \mathbf{\Psi}_r^{\mathsf{T}} \mathbf{z} \\ 0 & \|\mathbf{z}_{\perp}\| \end{bmatrix} = \mathbf{U}'' \mathbf{\Sigma}'' \mathbf{V}''^{\mathsf{T}} \tag{5.23}$$

and then the updated SVD factors are computed as:

$$\mathbf{\Psi}_r' = \begin{bmatrix} \mathbf{\Psi}_r & \mathbf{u} \end{bmatrix} \mathbf{U}'' \tag{5.24}$$

$$\tilde{\mathbf{\Sigma}}' = \mathbf{\Sigma}'' \tag{5.25}$$

$$\tilde{\mathbf{V}}' = \begin{bmatrix} \tilde{\mathbf{V}} & 0 \\ 0 & \mathbf{I} \end{bmatrix} \mathbf{V}'' \tag{5.26}$$

The numerical error is contained by reorthogonalizing columns of the left and right singular vectors using Gram-Schmidt procedure.

However, we only have the state measurements $\mathbf{y} \in \mathbb{R}^p$ at the p sensor locations, thus we have $m - p$ missing values in the new column $\mathbf{z}$ to be added. To handle the missing values,

instead of imputing the missing values to update the entire library as in [19], we update only the rows of $\mathbf{\Psi}_r$ that correspond to the known measurement points $\mathbf{y} \in \mathbb{R}^p$. Thus, with $p = r$ point measurements of data, only the corresponding r rows of the POD basis $\mathbf{\Psi}_r$ are updated.

5.5 Analysis Data

The time-series measurements of the surface concentrations and rates of change of concentrations of chemical species are collected from surface spatial locations in the atmosphere, illustrated in Figure 5.1. The analysis is performed for preprocessed or time-shifted raw data for three months or 89 days July, 2^{ND} - July, 31^{ST}, August, 2^{ND} - August, 31^{ST}, and September 2^{ND} - September, 30^{TH}. The southern and northern poles' grid cells that are covered in ice and have only day times or nighttimes have very different chemical dynamics that skew the analysis for the rest of the globe if included. Hence, we exclude the poles and only consider 33 latitudes $-58°$ through $70°$. We perform the analysis for the six different chemical species of interest: Nitrous Oxide **NO**, Ozone $\mathbf{O_3}$, Nitrous dioxide $\mathbf{NO_2}$, Hydroxyl **OH**, Isoprene **ISOP**, and Carbon Monoxide **CO**. For each species, we have **START** or absolute concentration data (expressed in units of molecules/cm^3) and **TEND** or tendency/rate of change data (expressed in units of molecules/cm^3/s). For some chemical species, the absolute concentration values in a small localized region dominate over the values in the rest of the grid cells. Correspondingly, the dominant spatial modes are very localized [122]. The 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 **START** data matrix, which has only non-negative values, now is $\mathbf{X}_{\mathbf{START}_{\log}} = \log(\mathbf{X}_{\mathbf{START}} + 1)$. The **TEND** data which also have negative rates of change values on the other hand are scaled as $\mathbf{X}_{\mathbf{TEND}_{\log}} = \text{sign}(\mathbf{X}_{\mathbf{TEND}}) * \log(|\mathbf{X}_{\mathbf{TEND}}| + 1)$. 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 hot-spots.

The first 60 days of data are the "training" data snapshots, and the rest of the 29 days are

the "testing" or validation data snapshots. With 72 snapshots per day, we have a training data matrix $\mathbf{X}$ of the size $(72(lon) \times 33(lat) = 2376) \times 4320(time)$ for each data set. The proper orthogonal decomposition or POD is computed by a singular value decomposition or SVD of each of these training data matrices, to build the library of tailored basis $\mathbf{\Psi}$, and compute the corresponding singular values $\mathbf{\Sigma}$. The results are presented in the following subsections.

5.6 Proper Orthogonal Decomposition of the data: The cumulative energy spectrum

The cumulative energy spectra for the **START** and **TEND** data for the six chemical species are presented on the left in Figure 5.3 and Figure 5.4 respectively. Also shown is the Gavish and Donoho [51] optimal hard threshold τ for both the data sets. On the right of Figure 5.3 and Figure 5.4, the cumulative energy spectra for the first 200 modes are presented for **START** data and for the first 500 modes for the **TEND** data, respectively. The exponential growth rate and plateauing of the cumulative energy spectra indicate that the global chemistry has an intrinsic low-rank structure. However, the growth rate is slow for the **START** data, and even slower for the **TEND** data, indicating noisy or hidden low energy dynamics in the state. Hence, it would be difficult to characterize the dynamics of these data sets using a minimal number of modes $r \ll m$.

For the **START** data, we need fewer modes to capture the significant energy as compared to the **TEND** data. As presented in Figure 5.3, for $\mathbf{NO_{START}}$ only about 62% energy is captured by the first 200 modes, while for $\mathbf{O_{3START}}$ and $\mathbf{CO_{START}}$ almost all the significant energy is captured by the first 200 modes. The rest of the **START** chemical species have about 80%-90% of the significant energy captured by the first 200 modes. We will thus be truncating at 200 modes for all the **START** data sets, using a threshold for singular values that captures 80% or above of the variance or energy in the data for almost all chemical species except for $\mathbf{NO_{START}}$. The effect of this hard threshold will be apparent when we reconstruct snapshots based on sensor measurements with truncated modes. For the

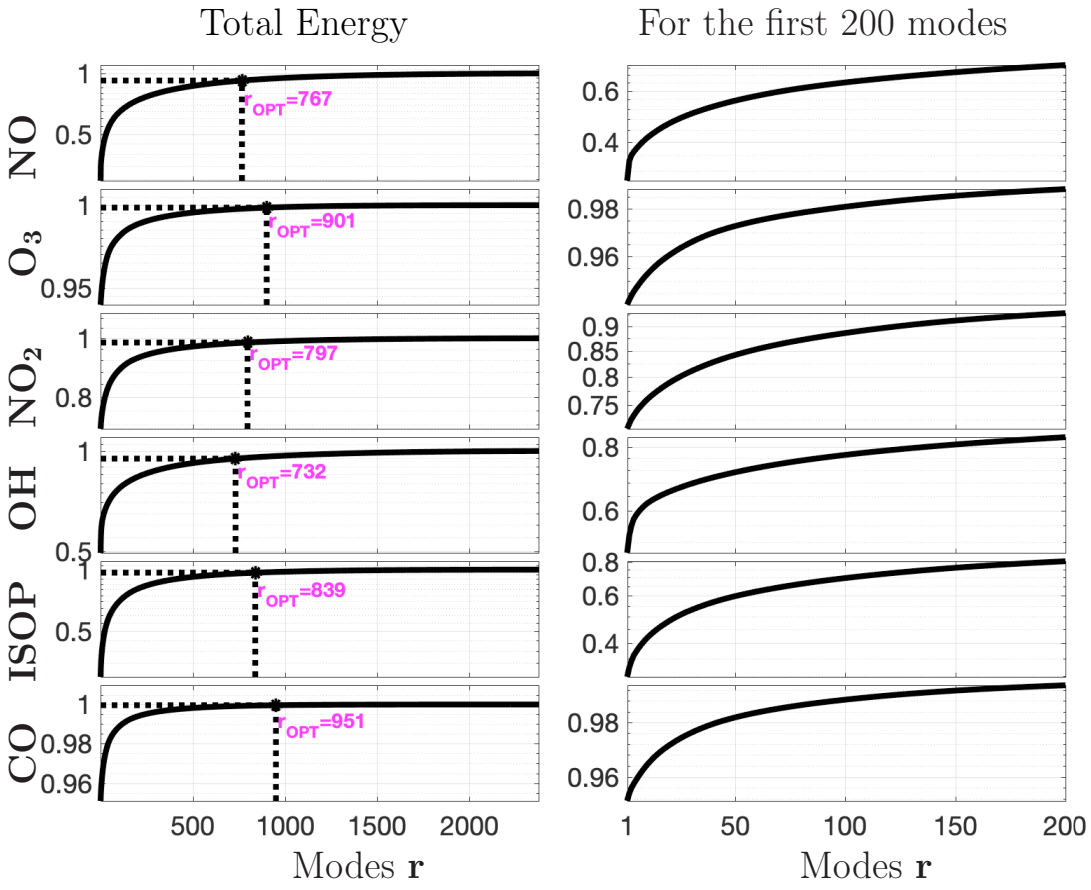


Figure 5.3: *Cumulative energy captured in the first* **r** *modes for all 6 chemical species of interest for* **START** *data on a logarithmic scale. On the left panels, the complete energy spectrum is shown along with the Gavish and Donoho [51] optimal hard threshold for truncation in pink as* $\mathbf{r_{OPT}}$*. The right panels show the energy spectrum for just the first 200 modes. For* $\mathbf{NO_{START}}$ *only about 60% of the energy is captured by the first 200 modes, while for* $\mathbf{O_{3START}}$ *and* $\mathbf{CO_{START}}$ *100% of the significant energy is captured.*

relative error in reconstruction versus number of modes/sensors though, we will be testing the relative error till the [51] optimal threshold for both **START** and **TEND** data sets.

The **TEND** data on the other hand exhibits a much higher intrinsic rank, as presented

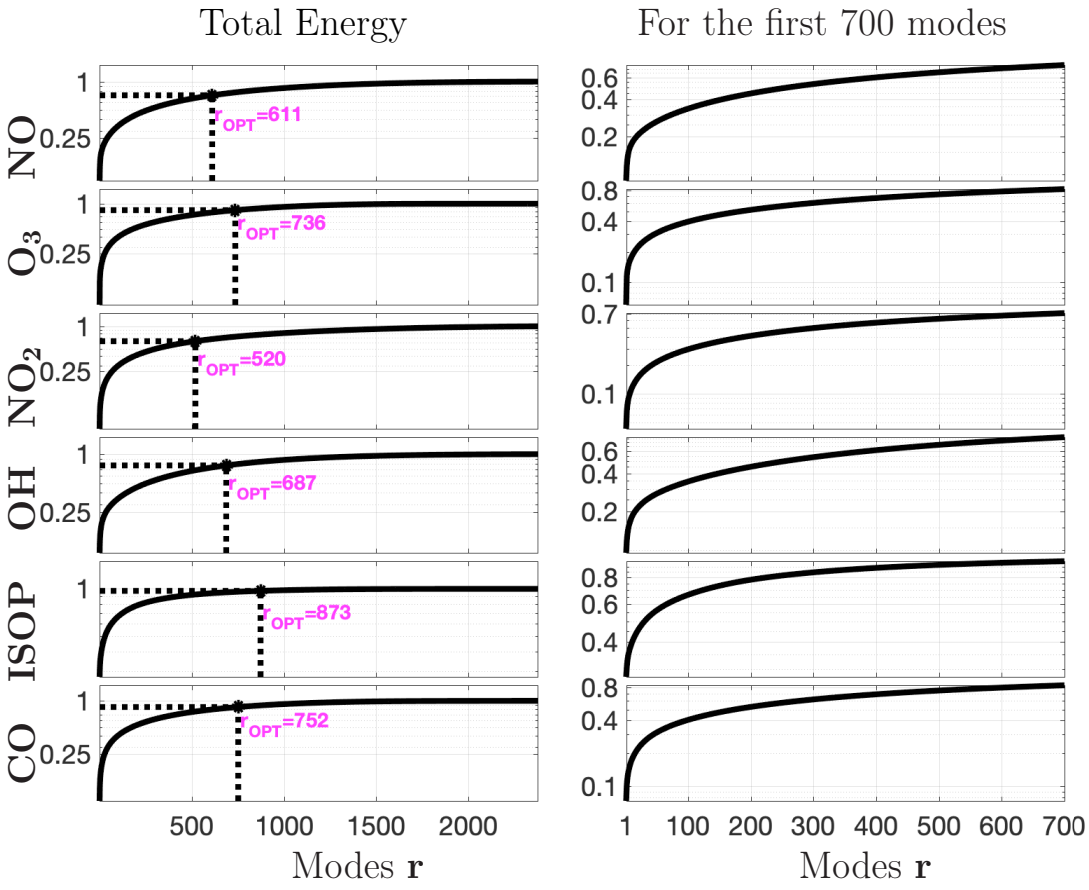


Figure 5.4: *Cumulative energy captured in the first* **r** *modes for all 6 chemical species of interest for* **TEND** *data on a logarithmic scale. On the left panels, the complete energy spectrum is shown along with the Gavish and Donoho [51] optimal hard threshold for truncation in pink as* $\mathbf{r_{OPT}}$*. The right panels show the energy spectrum for just the first 500 modes. For* $\mathbf{NO_{TEND}}$ *only about 60% of the energy is captured by first 700 modes, while for* $\mathbf{ISOP_{TEND}}$ *80% of the significant energy is captured.* **TEND** *data requires significantly more modes to capture about 60%-80% energy vs* **START** *data.*

in Figure 5.4. Even though the optimal hard threshold values as computed by the Gavish and Donoho [51] algorithm show a lower $\mathbf{r_{OPT}}$(indicating the 'elbow' of the distribution), there are still relevant meaningful low energy dynamics in the state. We need the first 700 modes to capture 60% of the most significant energy for $\mathbf{NO_{TEND}}$, $\mathbf{NO_{2TEND}}$, and

$\mathbf{OH_{TEND}}$. For $\mathbf{ISOP_{TEND}}$ we capture 80% of the most significant energy within the first 700 modes, while for $\mathbf{O_{3TEND}}$ and $\mathbf{CO_{TEND}}$ we capture 70% of the most significant energy. We will thus be truncating at 700 modes for all the **TEND** data sets, using a threshold for singular values that captures 70% or above of the variance or energy in the data for almost all chemical species except for $\mathbf{NO_{TEND}}$, $\mathbf{OH_{TEND}}$. The effect of this hard threshold will again be apparent when we reconstruct snapshots based on sensor measurements with truncated modes.

5.7 Proper Orthogonal Decomposition of the data: The dominant spatial modes Ψ

We now present the most dominant spatial modes 2-5 for both **START** and **TEND** data and two chemical species **CO**, **ISOP** in Figure 5.5 and Figure 5.6 respectively. Mode 1 for both data sets and chemical species is a baseline constant that does not display any significant features, hence is excluded here. Mode 2 peak features indicate the spatial regions of the second-highest spatial variability in the dynamics. Mode 3 shows the regions of the third-highest spatial variability and so on. For **CO** chemical species, regions of highest spatial variability are on land, as compared to **ISOP** chemical species where the spatial modes have the highest variability along the coastlines.

5.8 Sparse sensor placement for Reconstruction

As described in Section. 5.3, we now implement the Q-DEIM [40] for sensor placement (Case $\mathbf{p = r}$) and it's extension [82] for oversampled sensor placement (Case $\mathbf{p > r}$) to compute optimal sensor locations. The results are presented for $\mathbf{CO_{START}}$, $\mathbf{ISOP_{START}}$, $\mathbf{CO_{TEND}}$, and $\mathbf{ISOP_{TEND}}$ data sets as the dominant modes for these data sets are presented above. For the **START** data sets, we truncate at $\mathbf{r = 50}$ modes, and compute sensor placements for $\mathbf{r = p = 50}$ sensors and $\mathbf{r < p = 200}$ sensors.

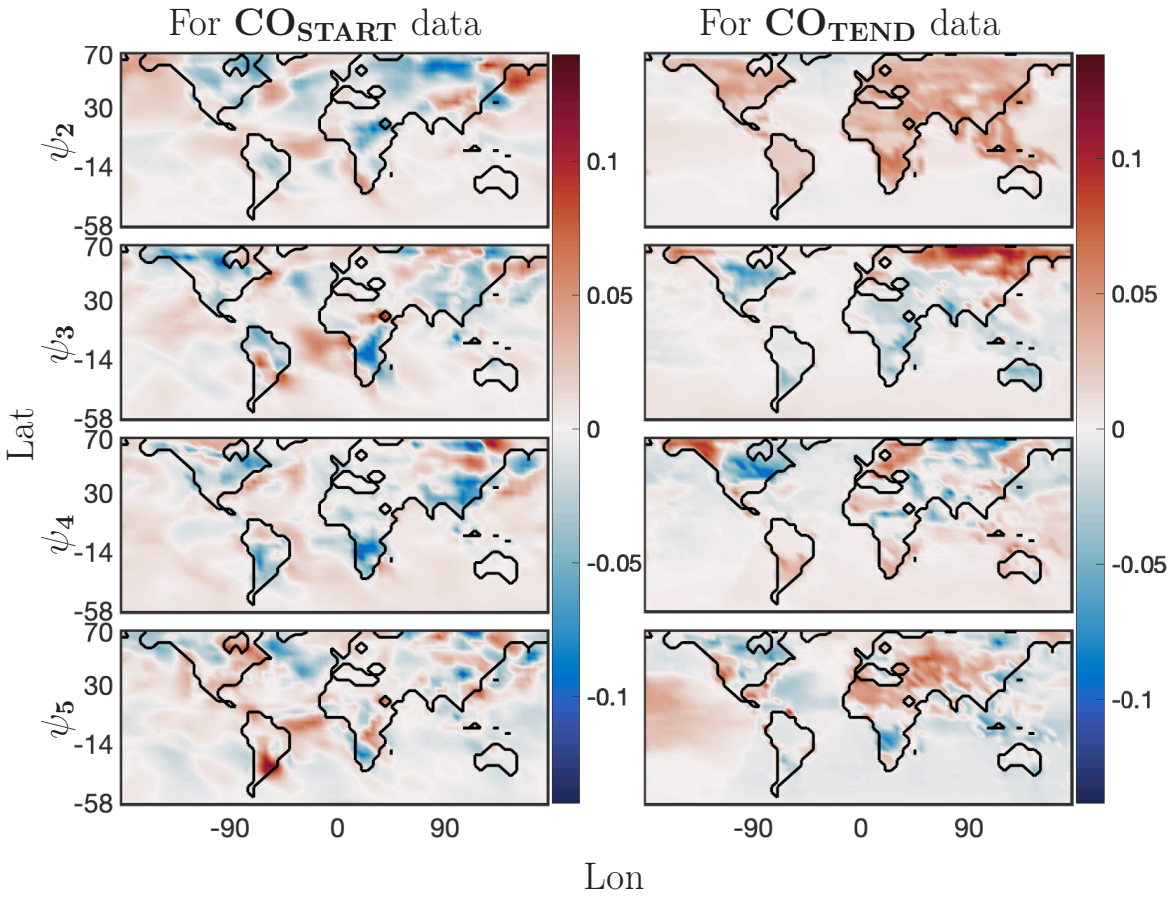


Figure 5.5: *Dominant spatial modes for* 60 *days of training data for surface* **CO** *preprocessed data. The analysis was computed for latitudes -58° through 70°. The left panel show four spatial modes for START data; and the right panel show four of the corresponding spatial modes for the TEND data.*

We pick an unknown or validation snapshot to test the reconstruction with these sensor locations.

Figure 5.7 shows these snapshots on the top panels, and the corresponding $\mathbf{r = 50}$ mode truncated POD projections on the second set of panels. The reconstructions of this snapshot using $\mathbf{r = p = 50}$ sensors are shown in the third set of panels, along with the sensor locations as pink dots. The reconstructions using sparse sensors are as good as the dense projections into the POD modes for the data sets. The sensors are clustered along the hot

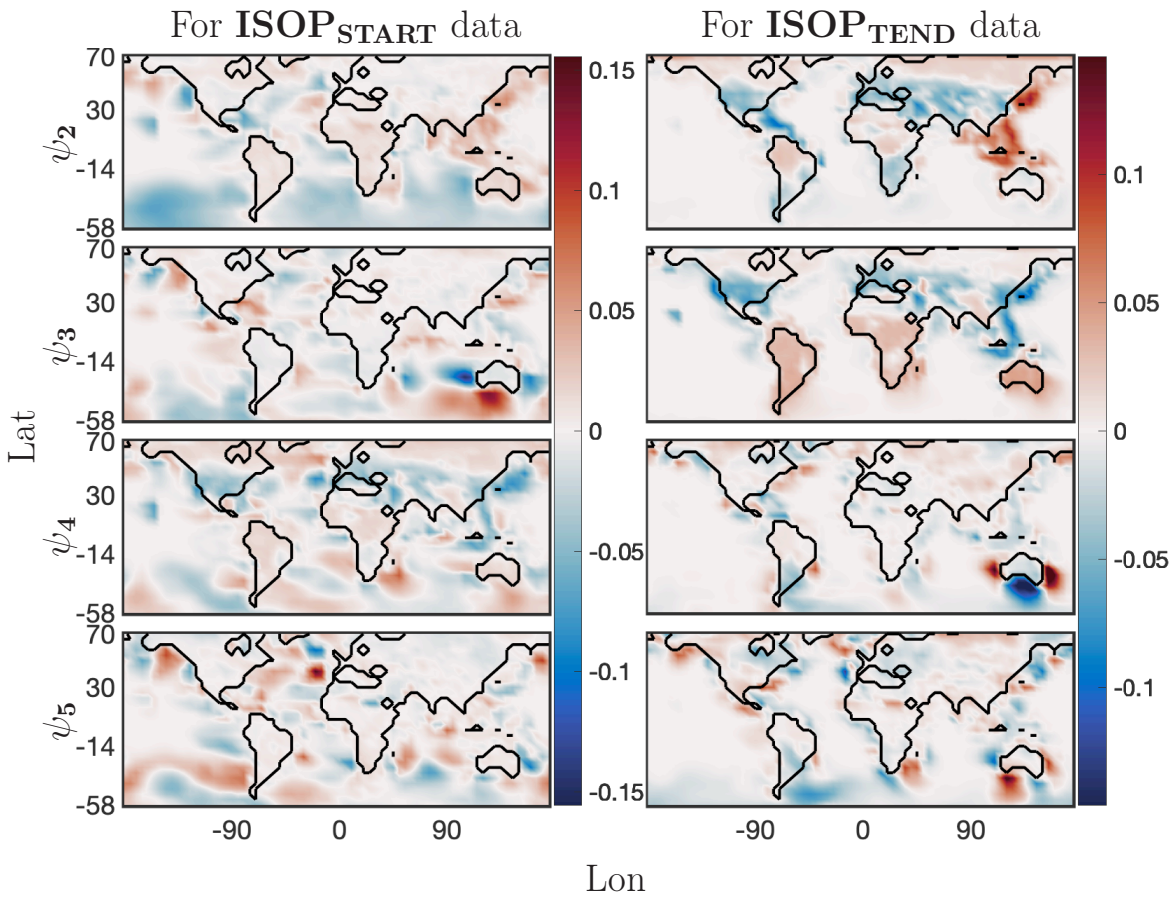


Figure 5.6: *Dominant spatial modes for* 60 *days of training data for surface* **ISOP** *preprocessed data. The analysis was computed for latitudes -58° through 70°. The left panel show four spatial modes for START data; and the right panel show four of the corresponding spatial modes for the TEND data.*

spots in Figure 5.5 and Figure 5.6 **START** spatial modes, picking up on localized regions with maximal variance in the data sets and placing optimal sensors in these regions. For **ISOP$_{START}$** these regions are the coastlines, whereas for **CO$_{START}$** these are localized on land. The resulting reconstructions represent most of the dominant features in the test data, and are as good as the snapshot's approximation in the same truncated POD basis. As presented in the fourth set of panels, when we over sample with $\mathbf{r < p = 200}$ sensors the reconstruction refines the reconstruction and picks up on most of the features in the

test data, since we have more sensors to leverage and pick up coherent patterns with lower variance signatures. The results here seen for two of the chemical species are consistent for all six chemicals tested. Based on these results, it seems possible to approximate the full state of the **START** data sets using a minimal number of sensors and modes, and we can truncate even lower than $r = 200$ and still have most of the dominant dynamics represented. Increasing the number of modes/sensors will pick up on the lower energy but uninformative features, and may also cause overfitting in the ℓ_2 reconstructions.

On the other hand, since **TEND** data sets exhibit an intrinsically higher rank, we truncate at $\mathbf{r = 200}$ modes, and compute sensor placements for $\mathbf{r = p = 200}$ sensors and $\mathbf{r < p = 700}$ sensors. The results are presented in Figure 5.8 with the same placement of the test data represented on the top panels, POD projection with $\mathbf{r = 200}$ mode truncation on the second set of panels, and the two cases of reconstruction with $\mathbf{r = p = 200}$ sensors and $\mathbf{r < p = 700}$ sensors on the bottom third set and fourth set of panels respectively. Q-DEIM again picks up the hot spots or spatial regions of maximal variance in the dominant modes from the training data, and places optimal sensors in these locations that are customized for each of the different chemical species data. The reconstructions using sparse sensors are again nearly as good as the dense projections into the POD modes for the data sets. The sensors are placed along the coastlines for $\mathbf{ISOP_{TEND}}$ data and inside landed regions for $\mathbf{CO_{TEND}}$. In this **TEND** data case, the $\mathbf{r = p = 200}$ sensor reconstruction misses out on picking out some meaningful lower energy features in the test data, hence $\mathbf{p = 200}$ are not adequate to capture enough of significant energy of the underlying dynamics. But the reconstructed snapshot when we oversample with $\mathbf{r < p = 700}$ refines the reconstruction further and now represents most of the features in the test data accurately, since now we have an adequate number of sensors to leverage and pick up most of the meaningful significant energy in the features of the test data. The reconstructed snapshots for the over sampled case are thus the same as the projection in the truncated POD basis.

In all four data sets tested, oversampling definitely improves the accuracy of the reconstructed snapshot, with the results being more obvious in the **TEND** data. However,

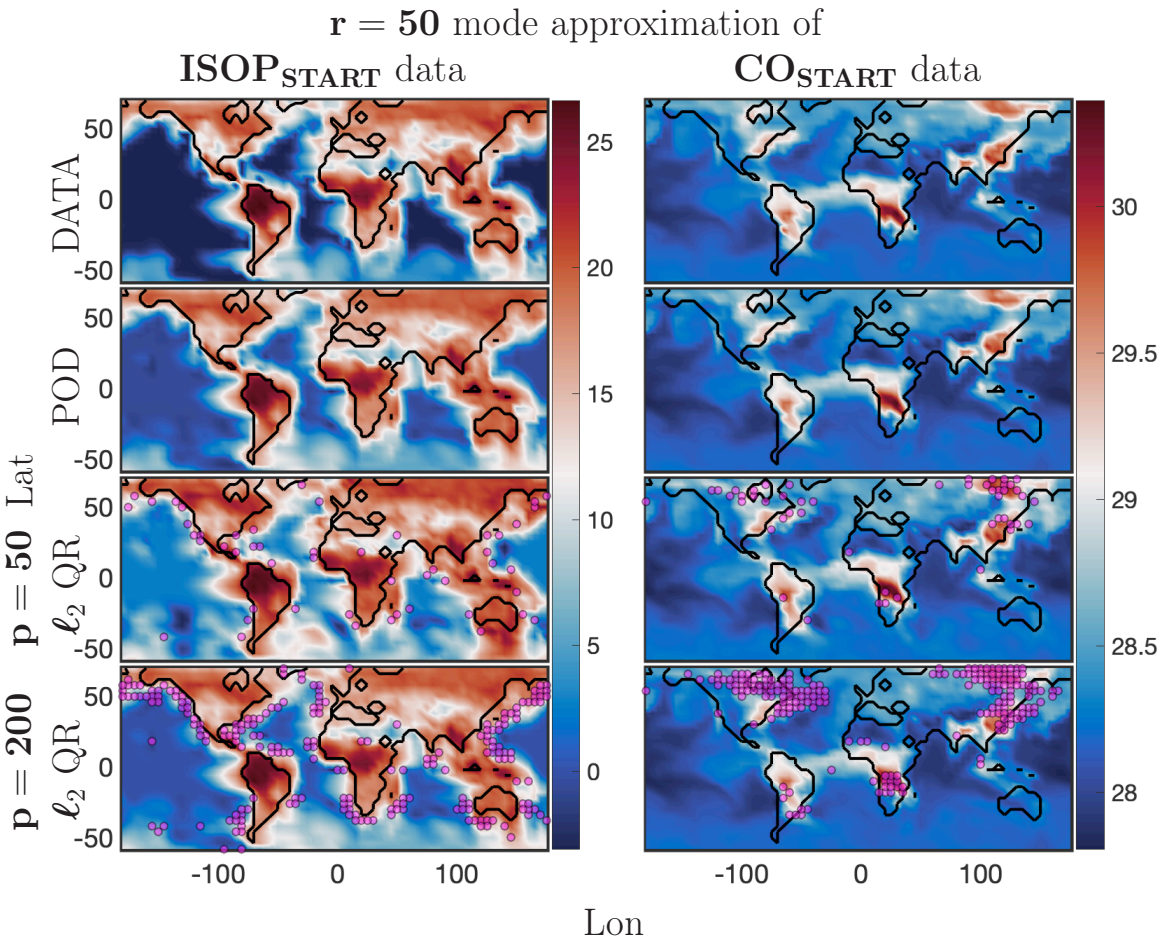


Figure 5.7: *Reconstruction of a single test snapshot from the* **POD** *projection and Least Squares reconstruction with* **QR** *sensors for* $\mathbf{ISOP_{START}}$ *and* $\mathbf{CO_{START}}$ *preprocessed data. The analysis was computed for latitudes -58° through 70°. The top panels are the* **START** *data; the second panels are the* $\mathbf{r = 50}$ *mode truncation* **POD** *projection; the third panels are the least squares reconstruction from* $\mathbf{p = r = 50}$ **QR** *sensors; and finally the fourth panels are the least squares reconstruction from oversampling with* $\mathbf{r < p = 200}$ **QR** *sensors. The sensor locations are indicated by the pink dots.*

the over sampled case requires an expensive QR factorization of a full state dimension $m \times m$ matrix, with the storage requirements also scaling quadratically with the state dimension. It is also a trade-off between increasing reconstruction accuracy and decreasing the cost associated with acquiring, placing and maintaining sensors.

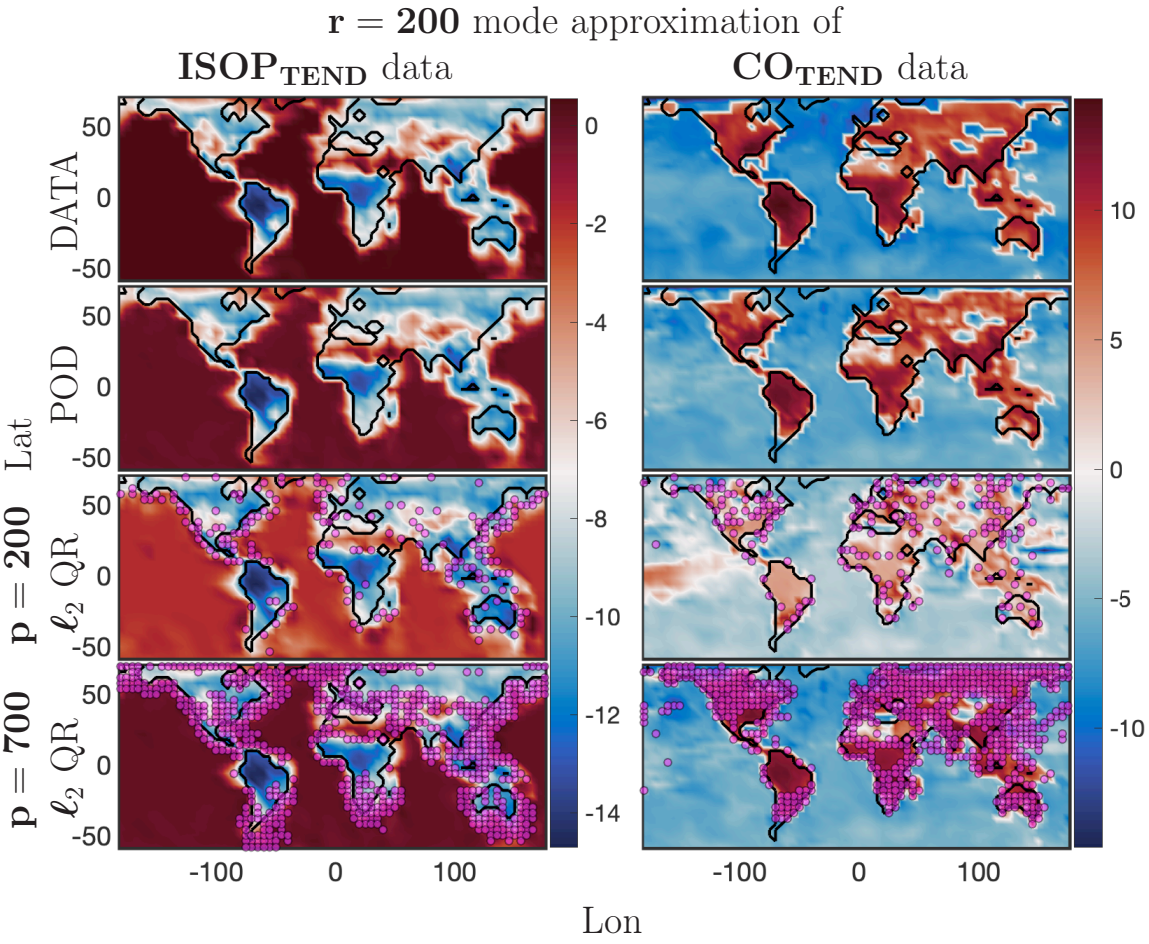


Figure 5.8: *Reconstruction of a single test snapshot from the* **POD** *projection and Least Squares reconstruction with* **QR** *sensors for* $\mathbf{ISOP_{TEND}}$ *and* $\mathbf{CO_{TEND}}$ *preprocessed data. The analysis was computed for latitudes -58° through 70°. The top panels are the* **TEND** *data; the second panels are the* $\mathbf{r = 200}$ *mode truncation* **POD** *projection; the third panels are the least squares reconstruction from* $\mathbf{p = r = 200}$ **QR** *sensors; and finally the fourth panels are the least squares reconstruction from oversampling with* $\mathbf{r < p = 700}$ **QR** *sensors. The sensor locations are indicated by the pink dots.*

5.9 Relative error for reconstruction

For the results presented in this section, we only use the Q-DEIM $\mathbf{r} = \mathbf{p}$ case. 500 snapshots are randomly selected from the training (already seen or Reconstruction snapshots) and validation (future or test snapshots, not seen before) **START** and **TEND** data sets for the six chemical species. Relative error is computed for each of these snapshots while the number of modes/sensors are increased in steps of 10 from $\mathbf{r} = \mathbf{p} = \mathbf{10}$ through the optimal rank truncation $\mathbf{r} = \mathbf{p} = \mathbf{r_{OPT}}$, computed using the Gavish and Donoho [51] hard threshold in Sec. 5.7. We then take an average for each $\mathbf{r}$, the resulting mean relative errors are presented in Figure 5.9. The mean relative error for **START** data are on the left panels, while the **TEND** data results are on the right panels.

We begin the discussion with the **START** data results. The mean relative errors for reconstruction or "known" snapshots decrease exponentially as the number of modes/sensors $\mathbf{r} = \mathbf{p}$ increase. For $\mathbf{NO_{START}}$ data set, the error decay rate is the slowest, which is a reflection of its singular value spectrum decay rate being the slowest. On the other hand, for chemical species $\mathbf{O_{3START}}$ and $\mathbf{CO_{START}}$ most of the significant energy is captured by the first few modes, and their errors are the lowest with the fastest decay rates. The mean relative errors for the test or "unknown" snapshots stay more or less the same or decrease only slightly as the number of modes/sensors $\mathbf{r} = \mathbf{p}$ increase. This is investigated further below.

The **TEND** data has the slowest decay rates in singular values for all the chemical species tested here. The relative mean errors for the reconstruction or "known" snapshots reflect this, showing a slight decay only as we start nearing the $\mathbf{r} = \mathbf{p} = \mathbf{r_{OPT}}$ optimal rank truncation limit. We would need a lot more modes/sensors or have to use oversampling to observe the error to start decreasing exponentially. The chemical species $\mathbf{ISOP_{TEND}}$ is an exception, with the fastest decay rates in both the singular values and the mean relative errors. The relative mean errors for the test or "unknown" snapshots for the **TEND** are worse as well, growing instead of decaying in some cases as we increase $\mathbf{r} = \mathbf{p}$. Some of

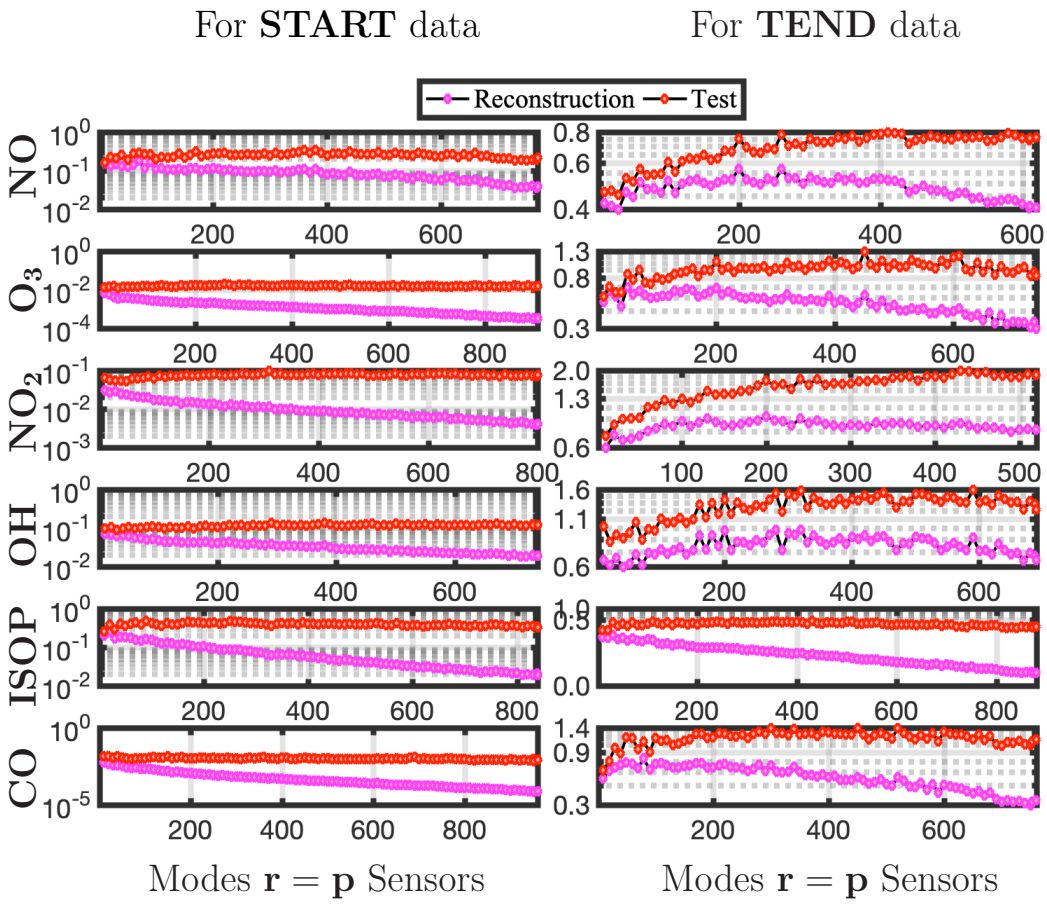


Figure 5.9: *Mean relative error for reconstruction and test or prediction versus number of modes* $\mathbf{r}$ *for* **START** *and* **TEND** *data for 6 different chemical species. 500 random reconstruction and test snapshots are used to compute the relative error, which is then averaged. The number of sensors* $\mathbf{p} = \mathbf{r}$ *for all the results shown here. The number of modes/sensors is increased from* $\mathbf{r} = \mathbf{p} = \mathbf{1} : \mathbf{10} : \mathbf{r_{OPT}}$ *for the chemical species. The mean error for reconstructing training snapshots flattens out within* $\mathbf{p} = \mathbf{r} = \mathbf{200}$ *for* **START** *data as compared to the* **TEND** *data, where we need* $\mathbf{p} = \mathbf{r} = \mathbf{500}$ *sensors for the mean reconstruction error to flatten out for most chemical species. In contrast, the errors for predicting future test snapshots remains more or less the same as number of modes/sensors increase.*

the error in reconstruction should be resolved by increasing $\mathbf{r} = \mathbf{p}$ beyond the optimal rank truncation, since the **TEND** data is an intrinsically high dimensional data set. The other possible issue with reconstructing validation snapshots is investigated below.

Atmospheric chemistry is a continuously evolving or moving model, with several intermittent localized convective phenomena with energetic contributions to the global scale evolution. This is especially true of the **TEND** data as compared to the **START** data. A 200-mode truncation for the **START** data set adequately recovers the large-scale dynamics, while increasing the number of modes/sensors will refine the localized convective phenomena. We will need much more modes for the **TEND** data, as discussed above.

5.10 Time dependency of relative errors

We will now investigate what effect, if any, time has on the error in reconstruction of test snapshots. Instead of randomly selecting test snapshots, we pick time-ordered test snapshots for the 29 test days. With 72 snapshots per day, we have a total of $72 \times 29 = 2088$ validation snapshots. We compute the relative error for each of these snapshots, and average the relative errors for each day. The number of modes/sensors is held constant at $\mathbf{r} = \mathbf{p} = \mathbf{200}$ for **START** test data and $\mathbf{r} = \mathbf{p} = \mathbf{700}$ for **TEND** test data. Results are presented in Figure 5.10. For the chemicals $\mathbf{O_{3START}}$, $\mathbf{OH_{START}}$, $\mathbf{CO_{START}}$, the error increases for the first 2-3 days and then relatively stays constant. On the other hand, for chemicals $\mathbf{NO_{START}}$, $\mathbf{NO_{2START}}$, $\mathbf{ISOP_{START}}$ the error increases sharply for the first three days and then does not follow a distinguishable trend. The initial increase in mean relative error is sharper for the **TEND** chemicals. Except for $\mathbf{ISOP_{TEND}}$ the mean error in reconstructing goes above 100% for the **TEND** data, and it tends to increase as we increase the number of days and go further away from the training window.

We conclude from the above that (i) Q-DEIM sparse sensor placement works better for the **START** data as compared to the **TEND** data; (ii) increasing the number of modes/sensors exponentially decreases the relative errors in reconstructing known/seen snapshots, but for reconstructing randomly selected unknown test snapshots the error

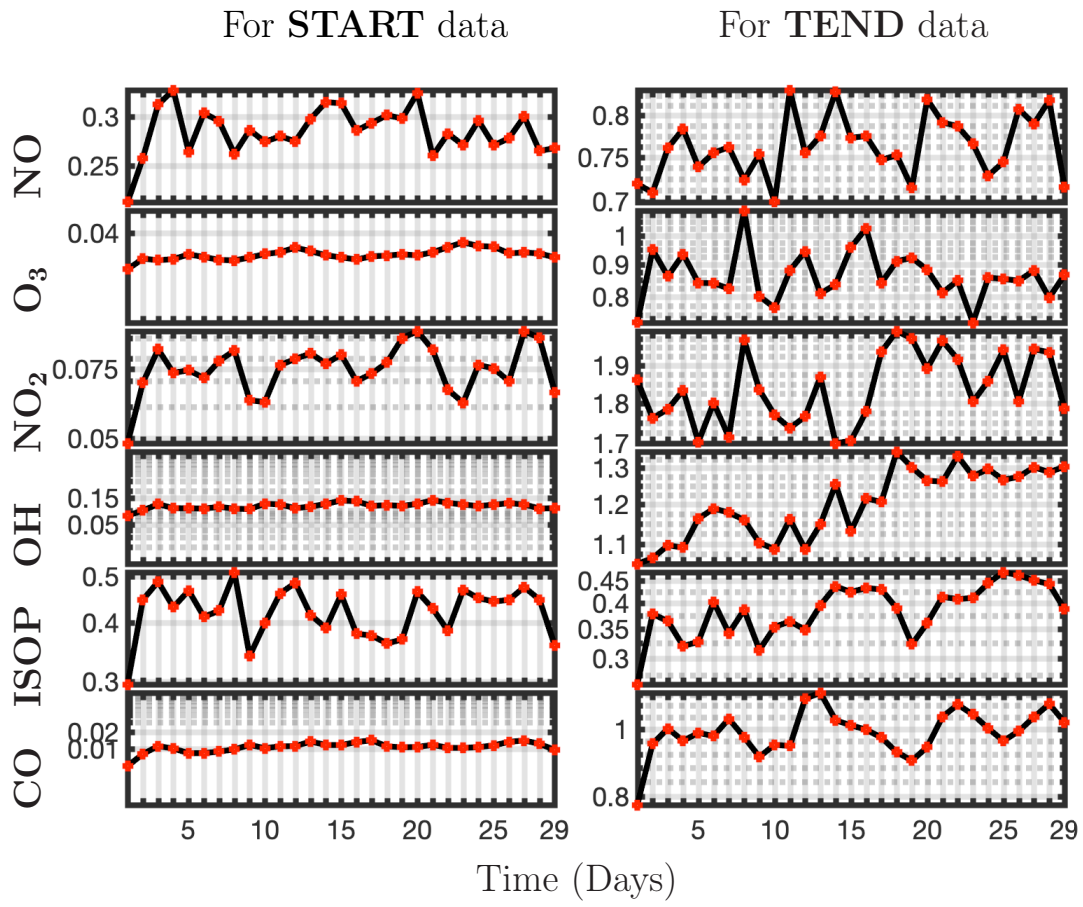


Figure 5.10: *Mean relative error predicting future* **START** *and* **TEND** *snapshots at the surface for a prediction window of 29 days and for 6 different chemical species. Relative errors are computed for 29 days and averaged for each day. The* **START** *snapshots use* $\mathbf{p = r = 200}$ *modes/sensors and the* **TEND** *snapshots use* $\mathbf{p = r = 700}$ *modes/sensors. The mean relative error increases sharply for the first 2â3 days and from then on stays nearly the same; for* **TEND** *data it tends to increase further as the number of days we are predicting increase.* **START** *data has better relative error as compared to the* **TEND** *data. For the latter four, chemical species have errors greater than 100%.*

does not decrease; and (iii) increasing time after the last training snapshot increases reconstruction errors in test data sharply for the first 2â3 days and from then on the error only does worse.

5.11 Effect of adding incremental updates to the POD basis on relative errors

Updating the library of modes $\mathbf{\Psi_r}$ with the 'new', i.e. from the unseen test window, sparse measurements of the state might help resolve the issue of errors increasing with time. To explore this further, we add an *incremental-SVD* update as described in Section. 5.4. The first test snapshot is reproduced as above with a projection of the measurements from sensor locations on the current $\mathbf{\Psi_r}$ space. The measurements from sensor locations are also used to now update the library of modes $\mathbf{\Psi_r}$ with an incremental SVD from [19]. The next test snapshot is reproduced using the updated library $\mathbf{\Psi'_r}$, and we continue this process with a running update of the library of modes $\mathbf{\Psi'_r}$. With a low-rank truncation $r \ll n, m$ and sparse measurements $p = r$.

The resulting relative mean errors in reproduction of test snapshots for the **TEND** after including the incremental SVD update are presented in Fig. 5.11. As is demonstrated, this approach leads to a significant improvement in the mean relative errors for reproducing test snapshots. The errors for test snapshots are now on the same scale as the errors for reproducing known or 'seen' snapshots as shown in Fig. 5.9, with none of the errors exceeding a 100%. Therefore, updating the library of modes captures some of the meaningful fast-changing energetic convective features in the **TEND** data sets, significantly improving the quality of the reproduced test snapshots. Also, note that for chemical species $\mathbf{O_{3TEND}}$, $\mathbf{OH_{TEND}}$, $\mathbf{ISOP_{TEND}}$ there is a clear increase in the errors as time increases away from the last training snapshot.

5.12 Correlation between dynamics of chemical species

With most of the chemical species having an intrinsic low-rank structure, we now investigate whether there exists any correlation between the spatial and temporal dynamics of different chemical species. We have placed sensors customized for one chemical species data, we will now measure the other chemical species at these customized sensor locations, and reconstruct snapshots based on these measurements. To check if there is any correlation

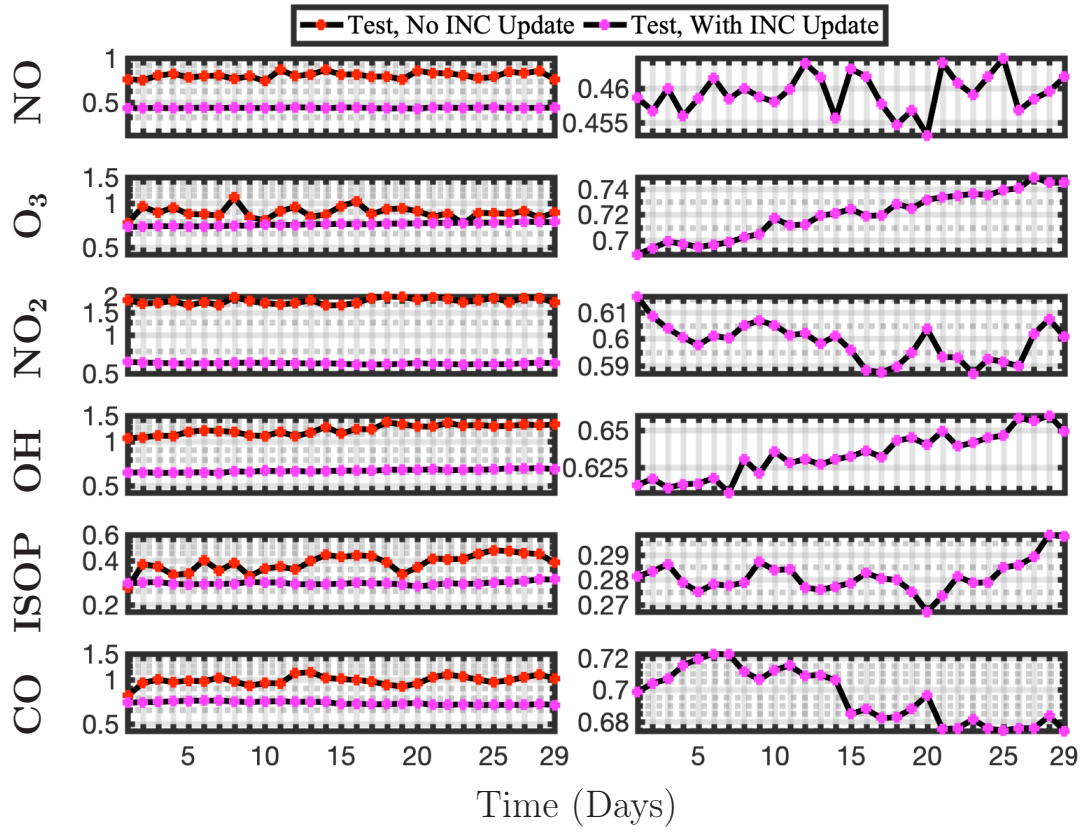


Figure 5.11: *Mean relative error predicting future* **TEND** *snapshots at the surface for a prediction window of 29 days and for 6 different chemical species. An incremental SVD is used to update the library of modes after each new sparse measurement, the reproduced snapshots are computed by projecting into this updated projection space. Relative errors are computed for 29 days and averaged for each day. The* **TEND** *snapshots use* $\mathbf{p = r = 700}$ *modes/sensors. For* **TEND** *data, the performance shows a definite improvement in reconstruction errors of test snapshots. For some chemical species, there is a clear trend of error increasing further as the number of days we are predicting increase.* **TEND** *data has better relative error with the incremental updates to the projection space data. None of the chemical species have errors greater than 100% after the incremental updates are added.*

between the **START** and **TEND** data of the same chemical species, we will also measure the **TEND** data of a chemical species at the sensors customized for its **START** data, and vice versa. We carry this experiment for all the six chemical species of interest. 500 randomly selected known snapshots are measured at sensor locations and the mean relative error is computed for the reconstructed snapshot. The relative errors along with their 97.5 percentile confidence intervals for $\mathbf{p = r = 200}$ sensors/modes are presented in Figure 5.12 for **START** data sets and in Figure 5.14 with $\mathbf{p = r = 700}$ sensors/modes for the **TEND** data sets. The first panel in the Figure 5.12 shows the mean relative error in pink with the confidence intervals in black bars to reconstruct $\mathbf{O_{3START}}$, $\mathbf{NO_{2START}}$, $\mathbf{OH_{START}}$, $\mathbf{ISOP_{START}}$, $\mathbf{CO_{START}}$, and $\mathbf{NO_{TEND}}$ data measured at the sensor locations customized for $\mathbf{NO_{START}}$ data. The second panel shows the mean relative error/confidence intervals to reconstruct $\mathbf{NO_{START}}$, $\mathbf{NO_{2START}}$, $\mathbf{OH_{START}}$, $\mathbf{ISOP_{START}}$, $\mathbf{CO_{START}}$, and $\mathbf{O_{3TEND}}$ measured at sensor locations customized for $\mathbf{O_{3START}}$ data; and so on. The layout is the same for the **TEND** data sensor locations presented in Figure 5.14.

Figure 5.12 shows that the chemical species $\mathbf{O_{3START}}$ and $\mathbf{CO_{START}}$ consistently have low errors when measured at customized sensor locations of the other **START** chemical species. We also note that the **TEND** chemical species data cannot be reconstructed accurately with measurements from sensor locations customized for their respective **START** data. Part of the reason is that we need many more sensors to capture the significant coherent low-order structures in the **TEND** data. The results from reconstructing a single snapshot of $\mathbf{CO_{START}}$ using sensors customized for the $\mathbf{O_{3START}}$ is presented in Figure 5.13. It can be seen that even though the sensors were customized for $\mathbf{O_{3START}}$, they are able to pick up on relevant spatial features in $\mathbf{CO_{START}}$ data and reconstruct the test snapshot faithfully.

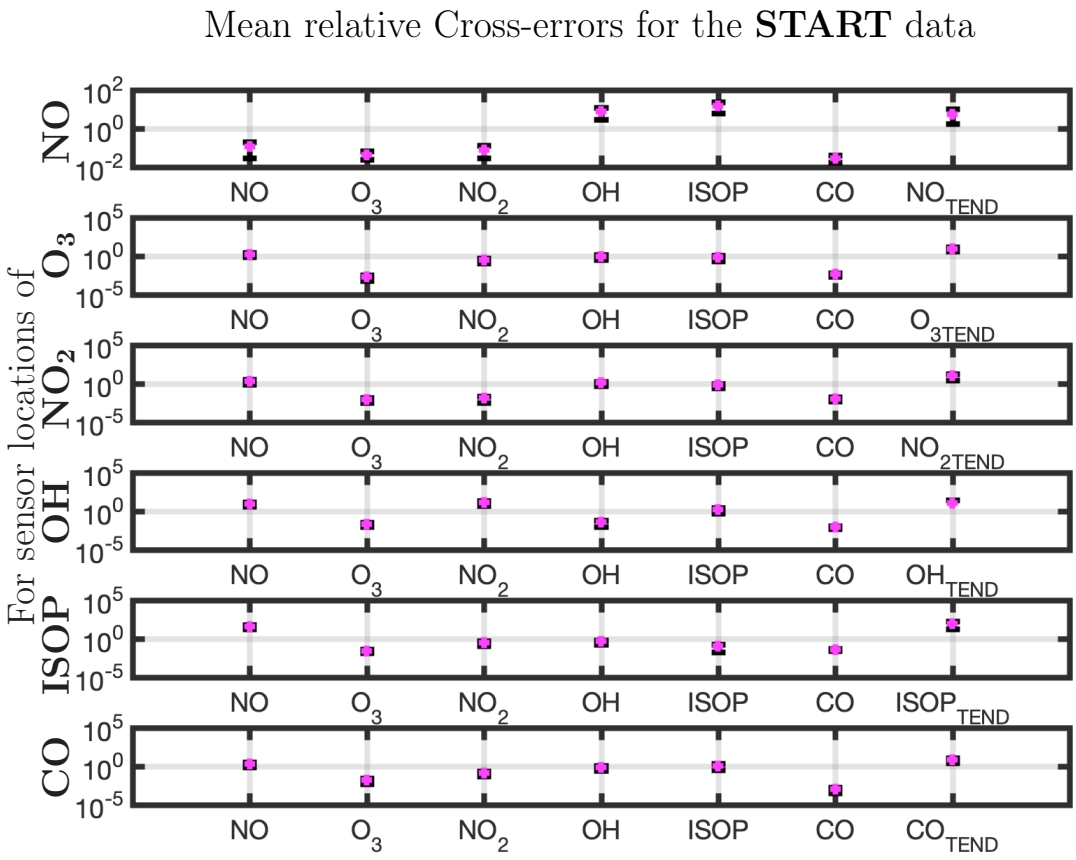


Figure 5.12: *Mean relative cross-errors along with the 95-percentile confidence intervals for* **START** *snapshots for six different chemical species at the surface. The first panel shows relative errors that are computed for measuring* $\mathbf{O_{3START}}$, $\mathbf{NO_{2START}}$, $\mathbf{OH_{START}}$, $\mathbf{ISOP_{START}}$, $\mathbf{CO_{START}}$, *and* $\mathbf{NO_{TEND}}$ *snapshots at* $\mathbf{NO_{START}}$ *sensors; and so on.* $\mathbf{p = r = 200}$ *modes/sensors are used. It is consistently shown that the chemicals* $\mathbf{O_{3START}}$ and $\mathbf{CO_{START}}$ *can be measured at customized sensor locations of other chemical species and reconstructed accurately.*

Figure 5.14 shows that the **START** chemical species consistently have low reconstruction errors when measured at customized sensor locations of their respective **TEND** chemical species. Again, since we have more sensors than we need for capturing significant features in the **START** data, we have lower relative errors in reconstructing these snapshots. The

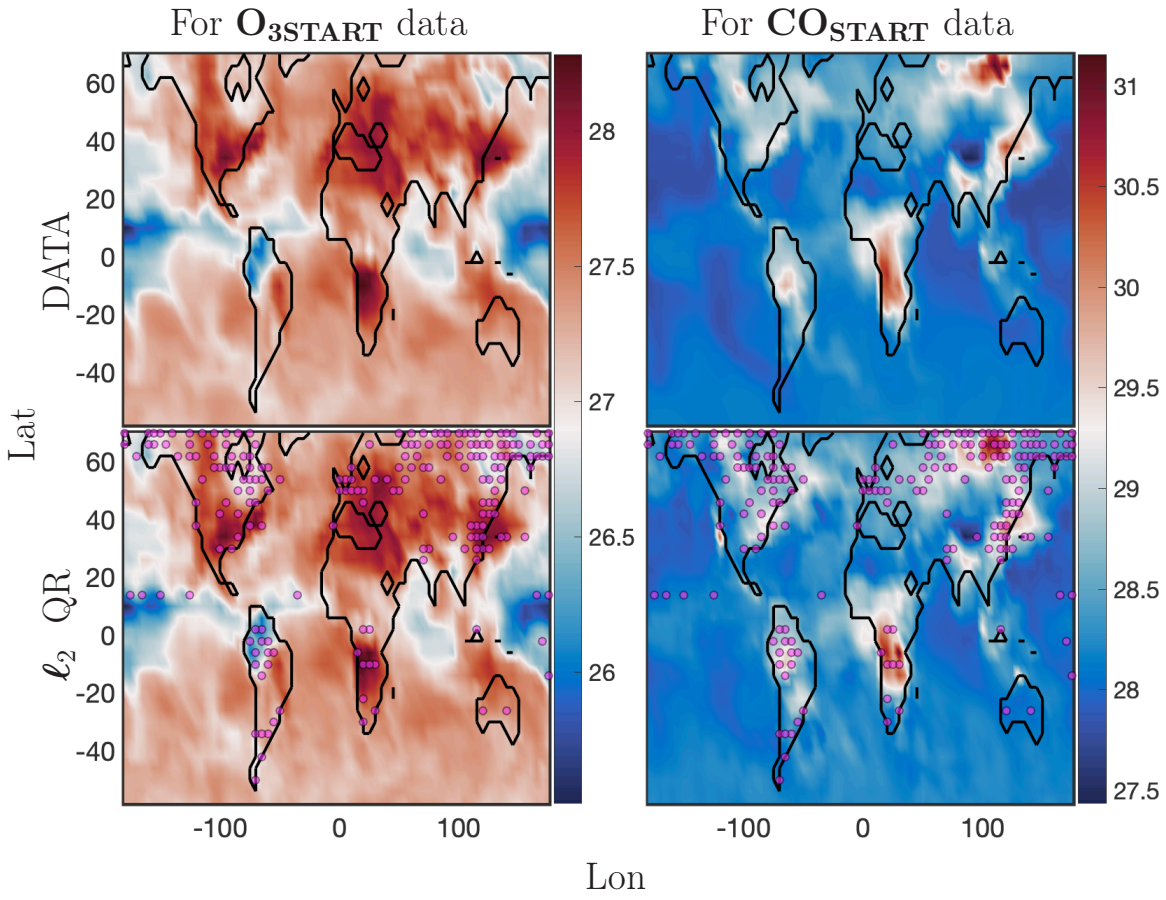


Figure 5.13: *Results from reconstructing single test snapshot for surface concentrations of* $\mathbf{O_{3START}}$ *and* $\mathbf{CO_{START}}$ *preprocessed data. Sensors were customized for the* $\mathbf{O_{3START}}$ *data with* $\mathbf{r = p = 200}$ *modes/sensors. The top left panel shows a randomly selected* $\mathbf{O_{3START}}$ *test snapshot, and the bottom left panel shows the resulting reconstruction of the test snapshot with sparse measurements from the customized sensor locations. The right top panel shows a randomly selected test snapshot of* $\mathbf{CO_{START}}$*, and the bottom right panel shows the resulting reconstruction of this snapshot when the* $\mathbf{CO_{START}}$ *chemical species is measured at the customized sensor locations for* $\mathbf{O_{3START}}$*. Even though the sensors were customized for* $\mathbf{O_{3START}}$*, they are able to pick up on relevant spatial features in* $\mathbf{CO_{START}}$ *data and reconstruct the test snapshot faithfully.*

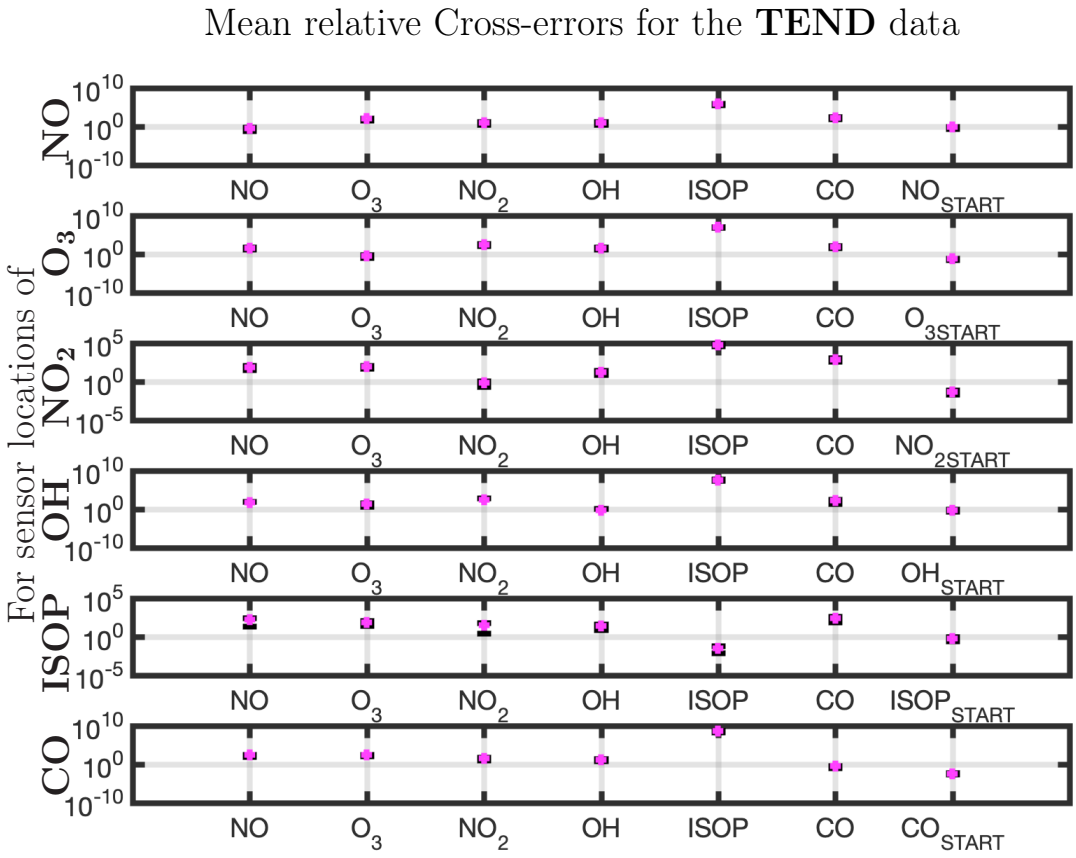


Figure 5.14: *Mean relative cross-errors along with the 95-percentile confidence intervals for* **START** *snapshots for six different chemical species at the surface. The first panel shows relative errors that are computed for measuring* $\mathbf{O_{3TEND}}$, $\mathbf{NO_{2TEND}}$, $\mathbf{OH_{TEND}}$, $\mathbf{ISOP_{TEND}}$, $\mathbf{CO_{TEND}}$, *and* $\mathbf{NO_{START}}$ *snapshots at* $\mathbf{NO_{TEND}}$ *sensors; and so on.* $\mathbf{p = r = 700}$ *modes/sensors are used. It is consistently shown that the* ***START*** *data can be measured at customized sensor locations of the corresponding* ***TEND*** *data and reconstructed accurately.*

results from reconstructing a single snapshot of $\mathbf{CO_{START}}$ using sensors customized for the $\mathbf{CO_{TEND}}$ is presented in Figure 5.15. It can be seen that even though the sensors were customized for $\mathbf{CO_{TEND}}$, they are able to pick up on relevant spatial features in $\mathbf{CO_{START}}$ data and reconstruct the test snapshot faithfully.

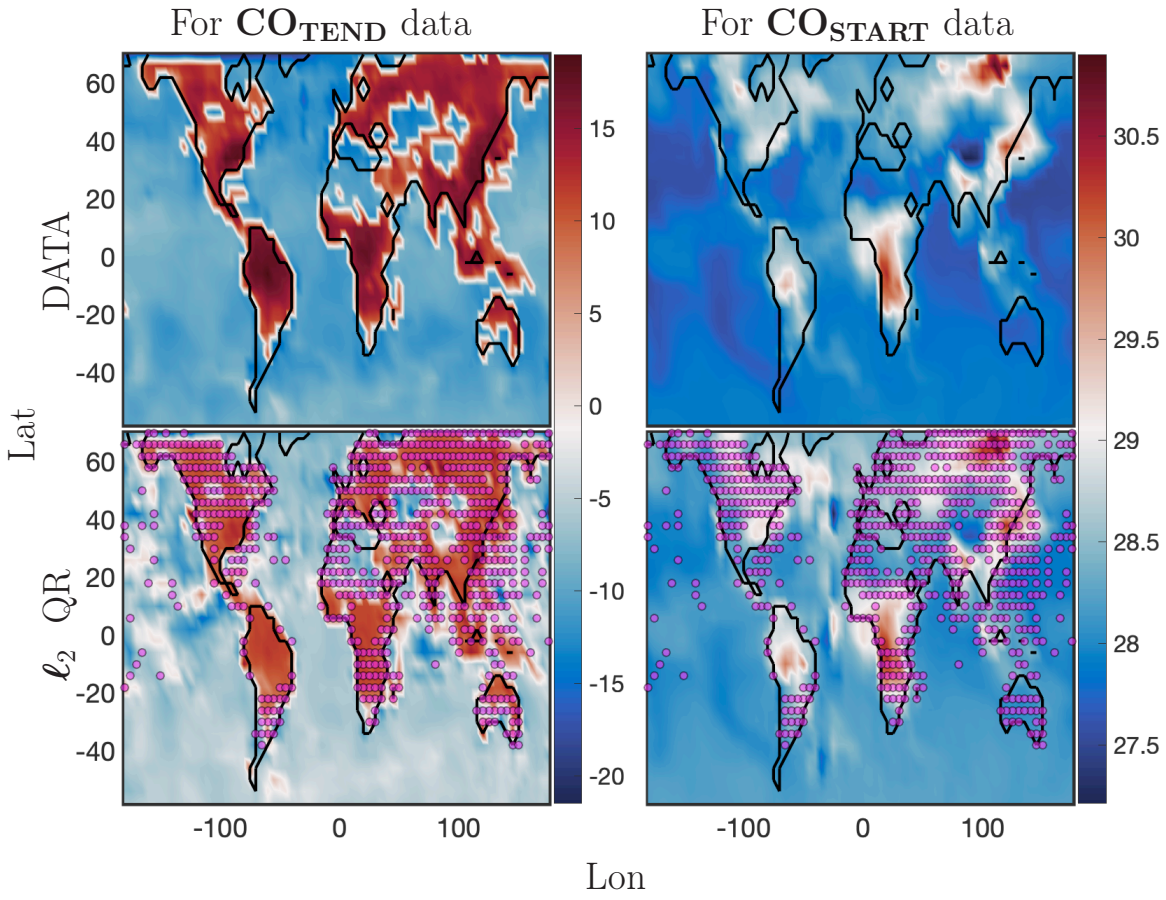


Figure 5.15: *Results from reconstructing single test snapshot for surface concentrations of* $\mathbf{CO_{TEND}}$ *and* $\mathbf{CO_{START}}$ *preprocessed data. Sensors were customized for the* $\mathbf{CO_{TEND}}$ *data with* $r = p = 700$ *modes/sensors. The top left panel shows a randomly selected* $\mathbf{CO_{TEND}}$ *test snapshot, and the bottom left panel shows the resulting reconstruction of the test snapshot with sparse measurements from the customized sensor locations. The right top panel shows a randomly selected test snapshot of* $\mathbf{CO_{START}}$*, and the bottom right panel shows the resulting reconstruction of this snapshot when the* $\mathbf{CO_{START}}$ *chemical species is measured at the customized sensor locations for* $\mathbf{CO_{TEND}}$*. Even though the sensors were customized for* $\mathbf{CO_{TEND}}$*, they are able to pick up on relevant spatial features in* $\mathbf{CO_{START}}$ *data and reconstruct the test snapshot faithfully.*

5.13 Conclusions

In this work we demonstrate the ability of data-driven sensor optimization to drastically reduce the number of measurements required to accurately reconstruct and predict full state of global atmospheric chemistry data. We combine 1) machine learning, i.e. the POD dimensionality reduction technique, which learns and exploits patterns in the data to produce low-dimensional representations of the full state, and 2) sparse sampling, i.e. designing highly specialized optimal sensors that reconstruct the full signal in a tailored POD basis from a small subset of sensor measurements instantaneously.

The **START** data sets, i.e. the absolute concentration of the chemical species, exhibit a low intrinsic rank as is observed from the cumulative energy spectra of their SVD singular values. The dynamics of these data sets also render themselves more amenable for the optimal sparse sensing problem. As few as 50 sensors/modes (2% of the full state measurements) can reproduce the full state of the **START** data accurately for most chemical species. For $\mathbf{NO_{START}}$ the intrinsic rank of the data set is higher, so more sensors will be needed. When we work with 200 sensors/modes (8% of the full state measurements) we further refine the reconstruction results. For $\mathbf{O_{3START}}$ and $\mathbf{CO_{START}}$ the relative reconstruction error for test snapshots is time independent and stays nearly the same. For all other chemical species the relative error in reconstruction however is time-dependent, sharply increasing for the first couple hundred test times after the training window and then staying approximately the same. The error in full state reproduction is low for four of the six chemical species tested, with the worst relative errors observed for $\mathbf{NO_{START}}$ at 40% and $\mathbf{ISOP_{START}}$ at 50%. Hence, there are two different classes of chemicals, one where the sparse sensor measurements excel at full state reconstruction of test snapshots, and the other with dynamics not easily reproduced using the sparse sensor architecture. The results are still more promising for the **START** data set.

The **TEND** data sets, i.e. the rates of change of absolute concentration of the chemical species, exhibit a much higher intrinsic rank as is observed from their cumulative energy spectra. The dynamics of these data sets exhibit many intermittent, spatially localized, and low-moderate energy convective phenomena. Hence, the decay rate of the singular values

is very slow and there are a lot of informative low-energy features that would be missed out when we truncate to a target low rank. Accordingly, the reconstruction errors for test snapshots is both higher and more time-dependent. Sensors are tasked with tracking these intermittent, convective phenomenon, and they fail, resulting in reconstruction errors that are higher than 100% for most of the chemical species tested here. This is because the modal separation of intermittent convective phenomenon is difficult from a time-invariant POD analysis. Separating isolated, low-energy temporal events cannot be done by a variance characterizing SVD based method such as POD. In this case, however, continually updating the library of POD modes using incremental SVD updates significantly improve the reconstruction errors. However, the reconstruction errors are still high, leading to the conclusion that the **TEND** data dynamics are not reproduced accurately by sparse sensing. In this case we used 700 sensors/modes, so about 30% of the total number of spatial measurements with not much alleviation in errors.

We also discovered correlation between the spatial and temporal dynamics of different chemical species by measuring a set of the chemical species concentrations (**START** data) at the sensor locations customized for a different chemical species. With a broader set of chemical species tested, we can possibly group the chemical species with strong correlations and measure them at a common set of sensor locations, reducing the number of measurements needed in the chemical space as well as spatial dimension.

An interesting future direction will be to apply the optimal sparse sensing architecture to a full year's worth of data, and including several elevations as well, making it a 4-dimensional data set for each of the chemical species. In this work, only one season's worth of data, i.e. three months of data, was used. Especially for the **START** data species it would be of interest to see if the POD can capture yearly patterns in the data, and how the sparse sensing problem would scale. Since the time-invariant POD analysis failed to reconstruct the **TEND** data dynamics accurately, it would be interesting if a tensor decomposition, or temporal-frequency analyses such as multi-resolution dynamic mode decomposition can succeed. However, sensor selection using non-normal modes arising from such decompositions is also an open problem. Clustering of chemical species that evolve with similar dynamics across the chemical space is another interesting topic of future research.

Chapter 6

CONCLUSION

In this thesis we developed data-driven, scalable ROMs that enable (i) analysis of the global atmospheric chemistry dynamics to extract low-rank features of the high-dimensional states, (ii) efficient and accurate reconstructions of the full state, (iii) stable forecasts that predict the future states of the system, and (iv) sparse optimized sensor placement using measurements derived from matrix pivots of the low-rank embeddings.

First, we introduced a suite of diagnostic algorithmic tools based on *randomized linear algebra*. Three distinct matrix decompositions, i.e. Randomized SVD, Randomized NMF and Randomized PCA algorithms, exploit the fact that the data itself has low-rank features to produced scalable, low-rank decompositions of global spatio-temporal atmospheric chemistry data. The resulting diagnostics were easy to interpret and successfully extracted known major features of atmospheric chemistry, such as summertime surface pollution and biomass burning activities. Full state reproductions were computed in real time using a small subset of basis libraries, thus allowing for efficient representation and compression of the data. Since these algorithms scale with the intrinsic rank of the global chemistry space rather than the spatio-temporal measurement space, our methods were demonstrated on a full year of global chemistry dynamics data, showing its significant improvement in computational speed and interpretability. 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 spatio-temporal activity for individual chemicals, a collection of chemicals, or the entire chemical space. The spatio-temporal features extracted give 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 understanding of the complex processes driving the global chemistry profile.

Next, we applied the data-driven regression architecture optDMD to construct an

adaptive and computationally efficient reduced order model of atmospheric chemistry dynamics. The model was used for characterization, monitoring and forecasting of chemical concentrations. However, for successful application of the DMD careful considerations of the input data were needed. We had to work with one latitude at a time with consistent day lengths, so that the time series have the same fixed periods when the chemistry 'turns on' during day times, and 'turns off' during nighttimes. Also in latitudes where the day lengths are very brief, the optDMD performance suffered, and isolating day times may be a better alternative. We also limited the analysis for one season or three months of data. Scaling the regression problem to the global three-dimensional grid, and to a longer period of training data, is the focus of future work.

Computing the optDMD required solution of a nonlinear, nonconvex optimization problem, which often fails to converge to a solution. The BOP-DMD version of the optDMD algorithm added bagging, initialization and ensembling of DMD models to produce an ensemble, probabilistic DMD model, reducing model variance and suppressing over-fitting by design. Not only did ensembling improve DMD, further innovations included stabilizing the variable projection technique used by optDMD so that it converged consistently to an optimal solution. The BOP-DMD also produced uncertainty metrics, and highlighted the high temporal variance in the eigenvalues produced by optDMD. Hence, BOP-DMD versions of the DMD algorithm are critical for characterizing the complexities of the chemical interaction dynamics. Focus of future work would be to use the BOP-DMD model for characterization, monitoring and forecasting of global chemical concentrations with either computational or sensor data.

Lastly, we explored the sparse sensing problem for global atmospheric chemistry, i.e., how to optimally place sparse sensors such that we can reproduce the full state of the chemical concentrations from a few sensor point measurements. The POD dimensionality reduction ROM learns and extracts low-rank embeddings from the high dimensional chemistry simulation data. QR pivoting of these low-rank features derive the optimal point measurements, i.e. the sensor placements, for instantaneous reproduction of the full chemistry state. The compression of data under low-rank embeddings (POD) enables the reduction of high-dimensional signals with $\mathcal{O}(10^6)$ points to sets of 200-700 optimal

sensors. The **START** data can be accurately reproduced and predicted using only 2%-8% of the available measurements of the state. However, errors in the prediction of **TEND** are significant. We needed to use a much higher number of modes to represent the low-dimensional features in the data, and correspondingly more sensors were needed (about 30% of the total measurement points). Oversampling the data helped significantly to pick up the lower energy relevant hidden dynamics. We added an incremental SVD update as well, which used the point measurements of data to also continually refresh the library of POD modes, which helped reduce the reconstruction errors significantly.

Not only do the **TEND** data sets have higher intrinsic rank, they also exhibit relevant low-energy, intermittent and spatially localized convective phenomena. Hence, these data sets were not amenable to variance characterizing time invariant POD. Adding the incremental SVD updates were one measure taken to improve the reconstruction. Tensor decompositions and temporal-frequency analysis such as multiresolution DMDs have also succeeded at capturing, separating and identifying isolated, low energy temporal events. However, sensor placement using non-normal basis modes is an open problem. This opens future directions for exploring the sparse sensing problem further for global atmospheric data.

BIBLIOGRAPHY

[1] Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation. *Journal of Geophysical Research: Atmospheres*, 106(D19):23073–23095, 2001.

[2] Alessandro Alla and J. Nathan Kutz. Nonlinear model order reduction via dynamic mode decomposition. *SIAM Journal on Scientific Computing*, 39(5):B778–B796, 2017.

[3] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning, 12 2020.

[4] Athanasios C. Antoulas. *Approximation of Large-Scale Dynamical Systems*. Society for Industrial and Applied Mathematics, 2005.

[5] Travis Askham and J Nathan Kutz. Variable projection methods for an optimized dynamic mode decomposition. *SIAM Journal on Applied Dynamical Systems*, 17(1):380–416, 2018.

[6] Shiri Avnery, Denise L. Mauzerall, Junfeng Liu, and Larry W. Horowitz. Global crop yield reductions due to surface ozone exposure: 1. year 2000 crop production losses and economic damage. *Atmospheric Environment*, 45(13):2284 – 2296, 2011.

[7] M.F.A. Azeez and A.F. Vakakis. Proper orthogonal decomposition (pod) of a class of vibroimpact oscillations. journal of sound and vibration, 240 (5): 859-889. *Journal of Sound and Vibration*, 240:859–889, 03 2001.

[8] Shervin Bagheri. Effects of weak noise on oscillating flows: Linking quality factor, Floquet modes, and Koopman spectrum. *Physics of Fluids*, 26(9), 09 2014.

[9] Zhe Bai, Thakshila Wimalajeewa, Zachary Berger, Wang Guannan, Mark Glauser, and P.K. Varshney. Low-dimensional approach for reconstruction of airfoil data via compressive sensing. *AIAA Journal*, 53:920–933, 04 2015.

[10] Madhusudhanan Balasubramanian, Stanislav Zabic, Christopher Bowd, Hilary Thompson, Peter Wolenski, Sundararaj Iyengar, Binita Karki, and Linda Zangwill. A framework for detecting glaucomatous progression in the optic nerve head of an eye using proper orthogonal decomposition. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, 13:781–93, 05 2009.

[11] Richard G. Baraniuk. Compressive sensing [lecture notes]. *IEEE Signal Processing Magazine*, 24(4):118–121, 2007.

[12] Maxime Barrault, Yvon Maday, Ngoc Nguyen, and Anthony Patera. An âempirical interpolationâ method: Application to efficient reduced-basis discretization of partial differential equations. *Comptes Rendus Mathematique*, 339:667â672, 11 2004.

[13] Casey Battaglino, Grey Ballard, and Tamara G Kolda. A practical randomized cp tensor decomposition. *SIAM Journal on Matrix Analysis and Applications*, 39(2):876–901, 2018.

[14] Peter Benner, Serkan Gugercin, and Karen Willcox. A survey of projection-based model reduction methods for parametric dynamical systems. *SIAM review*, 57(4):483–531, 2015.

[15] Peter Benner, Serkan Gugercin, and Karen Willcox. A survey of projection-based model reduction methods for parametric dynamical systems. *SIAM Review*, 57:483–531, 06 2015.

[16] G Berkooz, PJ Holmes, and John Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Annual Review of Fluid Mechanics*, 25:539–575, 11 2003.

[17] Huisheng Bian and Michael J. Prather. Fast-j2: Accurate simulation of stratospheric photolysis in global chemical models. *Journal of Atmospheric Chemistry*, 41(3):281–296, Mar 2002.

[18] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[19] Matthew Brand. Incremental singular value decomposition of uncertain data with missing values. In Anders Heyden, Gunnar Sparr, Mads Nielsen, and Peter Johansen, editors, *Computer Vision — ECCV 2002*, pages 707–720, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[20] Guy P Brasseur and Daniel J Jacob. *Modeling of Atmospheric Chemistry*. Cambridge University Press, 2017.

[21] Bingni Brunton, Steven Brunton, J.L. Proctor, and J.N. Kutz. Sparse sensor placement optimization for classification. *SIAM Journal on Applied Mathematics*, 76:2099–2122, 01 2016.

[22] Steven Brunton, Joshua Proctor, and J. Kutz. Discovering governing equations from data: Sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113:3932â3937, 09 2015.

[23] Steven Brunton, Jonathan Tu, Ido Bright, and J. Kutz. Compressive sensing and low-rank libraries for classification of bifurcation regimes in nonlinear dynamical systems. *SIAM Journal on Applied Dynamical Systems*, 13, 12 2013.

[24] Steven L. Brunton and J. Nathan Kutz. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2019.

[25] Steven L. Brunton, Joshua L. Proctor, Jonathan H. Tu, and J. Nathan Kutz. Compressed sensing and dynamic mode decomposition, 2015.

[26] Emmanuel CandÃšs, Justin Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59, 08 2006.

[27] K. Carlberg, M. Barone, and H. Antil. Galerkin v. least-squares Petrov–Galerkin projection in nonlinear model reduction. *Journal of Computational Physics*, 330:693–734, 2017.

[28] W.F. Caselton and J.V. Zidek. Optimal monitoring network designs. *Statistics Probability Letters*, 2(4):223–227, 1984.

[29] Saifon Chaturantabut and Danny Sorensen. Nonlinear model reduction via discrete empirical interpolation. *SIAM J. Scientific Computing*, 32:2737–2764, 01 2010.

[30] Kevin K. Chen and Clarence W. Rowley. H2 optimal actuator and sensor placement in the linearised complex ginzburgâlandau system. *Journal of Fluid Mechanics*, 681:241 – 260, 2010.

[31] Kevin K. Chen, Jonathan H. Tu, and Clarence W. Rowley. Variants of dynamic mode decomposition: Boundary condition, koopman, and fourier analyses. *Journal of Nonlinear Science*, 22(6):887–915, 2012.

[32] Andrzej Cichocki and Anh Huy Phan. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE Transactions*, 92-A(3):708–721, 2009.

[33] Matthew Cooper, Randall V. Martin, Catherine Wespes, Pierre-Francois Coheur, Cathy Clerbaux, and Lee T. Murray. Tropospheric nitric acid columns from the iasi satellite instrument interpreted with a chemical transport model: Implications for parameterizations of nitric oxide production by lightning. *Journal of Geophysical Research: Atmospheres*, 119(16):10068–10079, 2014. 2014JD021907.

[34] John P Cunningham and Zoubin Ghahramani. Linear dimensionality reduction: survey, insights, and generalizations. *Journal of Machine Learning Research*, 16(1):2859–2900, 2015.

[35] Scott T. M. Dawson, Maziar S. Hemati, Matthew O. Williams, and Clarence W. Rowley. Characterizing and correcting for the effect of sensor noise in the dynamic mode decomposition. *Experiments in Fluids*, 57(3):42, 2016.

[36] Eric A. Deem, Louis N. Cattafesta, Maziar S. Hemati, Hao Zhang, Clarence Rowley, and Rajat Mittal. Adaptive separation control of a laminar boundary layer using online dynamic mode decomposition. *Journal of Fluid Mechanics*, 903:A21, 2020.

[37] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[38] Petros Drineas and Michael W Mahoney. Randnla: randomized numerical linear algebra. *Communications of the ACM*, 59(6):80–90, 2016.

[39] Zlatko Drmac and Serkan Gugercin. A new selection operator for the discrete empirical interpolation method—improved a priori error bound and extensions. *SIAM Journal on Scientific Computing*, 38, 05 2015.

[40] Zlatko Drmac and Serkan Gugercin. A new selection operator for the discrete empirical interpolation method—improved a priori error bound and extensions. *SIAM Journal on Scientific Computing*, 38, 05 2015.

[41] Thomas Duriez, Steven Brunton, and Bernd Noack. *Machine Learning Control â Taming Nonlinear Dynamics and Turbulence*, volume 116. 11 2016.

[42] S. D. Eastham, M. S. Long, C. A. Keller, E. Lundgren, R. M. Yantosca, J. Zhuang, C. Li, C. J. Lee, M. Yannetti, B. M. Auer, T. L. Clune, J. Kouatchou, W. M. Putman, M. A. Thompson, A. L. Trayanov, A. M. Molod, R. V. Martin, and D. J. Jacob. Geos-chem high performance (gchp): A next-generation implementation of the geos-chem chemical transport model for massively parallel applications. *Geoscientific Model Development Discussions*, 2018:1–18, 2018.

[43] Sebastian D. Eastham, Debra K. Weisenstein, and Steven R.H. Barrett. Development and evaluation of the unified troposphericâstratospheric chemistry extension (ucx) for the global chemistry-transport model geos-chem. *Atmospheric Environment*, 89:52 – 63, 2014.

[44] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[45] Carl Eckart and G. Marion Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936.

[46] N Benjamin Erichson, Steven L Brunton, and J Nathan Kutz. Compressed singular value decomposition for image and video processing. In *Computer Vision Workshop (ICCVW), 2017 IEEE International Conference on*, pages 1880–1888. IEEE, 2017.

[47] N Benjamin Erichson, Krithika Manohar, Steven L Brunton, and J Nathan Kutz. Randomized cp tensor decomposition. *arXiv preprint arXiv:1703.09074*, 2017.

[48] N Benjamin Erichson, Ariana Mendible, Sophie Wihlborn, and J Nathan Kutz. Randomized nonnegative matrix factorization. *Pattern Recognition Letters*, 104:1–7, 2018.

[49] N Benjamin Erichson, Sergey Voronin, Steven L Brunton, and J Nathan Kutz. Randomized matrix decompositions using r. *arXiv preprint arXiv:1608.02148*, 2016.

[50] N Benjamin Erichson, Peng Zeng, Krithika Manohar, Steven L Brunton, J Nathan Kutz, and Aleksandr Y Aravkin. Sparse principal component analysis via variable projection. *arXiv preprint arXiv:1804.00341*, 2018.

[51] Matan Gavish and David L. Donoho. The optimal hard threshold for singular values is 4/sqrt(3), 2014.

[52] Nicolas Gillis. Introduction to nonnegative matrix factorization. 2017.

[53] Alex Gittens, Kai Rothauge, Shusen Wang, Michael W Mahoney, Lisa Gerhardt, Jey Kottalam, Michael Ringenburg, Kristyn Maschhoff, et al. Accelerating large-scale data analysis by offloading to high-performance computing libraries using alchemist. *arXiv preprint arXiv:1805.11800*, 2018.

[54] Carlos Guestrin, Andreas Krause, and Ajit Paul Singh. Near-optimal sensor placements in gaussian processes. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 265â272, New York, NY, USA, 2005. Association for Computing Machinery.

[55] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[56] S. Han and Brian Feeny. Application of proper orthogonal decomposition to structural vibration analysis. mechanical systems and signal processing, 17 (5): 989-1001. *Mechanical Systems and Signal Processing - MECH SYST SIGNAL PROCESS*, 17:989–1001, 09 2003.

[57] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[58] Maziar S. Hemati, Clarence W. Rowley, Eric A. Deem, and Louis N. Cattafesta. Debiasing the dynamic mode decomposition for applied koopman spectral analysis of noisy datasets. *Theoretical and Computational Fluid Dynamics*, 31(4):349–368, 2017.

[59] Jan Hesthaven, Gianluigi Rozza, and Benjamin Stamm. *Certified Reduced Basis Methods for Parametrized Partial Differential Equations*. 01 2016.

[60] Philip Holmes, John Lumley, Gahl Berkooz, and Clarence Rowley. Turbulence, coherent structures, dynamical systems and symmetry. *Turbulence, Coherent Structures, Dynamical Systems and Symmetry, by Philip Holmes , John L. Lumley , Gahl Berkooz , Clarence W. Rowley, Cambridge, UK: Cambridge University Press, 2012*, 02 2012.

[61] L. Hu, C. A. Keller, M. S. Long, T. Sherwen, B. Auer, A. Da Silva, J. E. Nielsen, S. Pawson, M. A. Thompson, A. L. Trayanov, K. R. Travis, S. K. Grange, M. J. Evans, and D. J. Jacob. 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). *Geoscientific Model Development Discussions*, 2018:1–32, 2018.

[62] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R.* Springer, 2013.

[63] Siddharth Joshi and Stephen Boyd. Sensor selection via convex optimization. *IEEE Transactions on Signal Processing*, 57(2):451–462, 2009.

[64] Sirkka Juntto and Pentti Paatero. Analysis of daily precipitation data by positive matrix factorization. *Environmetrics*, 5(2):127–144, 1994.

[65] Anthony Kellems, Saifon Chaturantabut, Danny Sorensen, and Steven Cox. Morphologically accurate reduced order modeling of spiking neurons. *Journal of computational neuroscience*, 28:477–94, 03 2010.

[66] GaÃ«tan Kerschen, J.-C Golinval, ALEXANDER VAKAKIS, and LAWRENCE BERGMAN. The method of proper orthogonal decomposition for dynamical characterization and order reduction of mechanical systems: An overview. *Nonlinear Dynamics*, 41:147–169, 08 2005.

[67] J Nathan Kutz. *Data-driven modeling & scientific computation: methods for complex systems & big data.* Oxford University Press, 2013.

[68] J. Nathan Kutz, Steven L. Brunton, Bingni W. Brunton, and Joshua L. Proctor. *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems.* SIAM-Society for Industrial and Applied Mathematics, USA, 2016.

[69] J. Nathan Kutz, Xing Fu, and Steven L. Brunton. Multiresolution dynamic mode decomposition. *SIAM Journal on Applied Dynamical Systems*, 15(2):713–735, 2016.

[70] Henning Lange, Steven L. Brunton, and J. Nathan Kutz. From fourier to koopman: Spectral methods for long-term time series prediction. *CoRR*, abs/2004.00574, 2020.

[71] D D Lee and S H Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[72] Eddie Lee, Chak K Chan, and Pentti Paatero. Application of positive matrix factorization in source apportionment of particulate pollutants in hong kong. *Atmospheric Environment*, 33(19):3201–3212, 1999.

[73] V. Lenaerts, GaÃ«tan Kerschen, and J.-C Golinval. Proper orthogonal decomposition for model updating of non-linear mechanical systems. *Mechanical Systems and Signal Processing*, 15:31–43, 01 2001.

[74] Charles J. Stone R.A. Olshen Leo Breiman, Jerome Friedman. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.

[75] David Lindley. On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, 27:986–1005, 1956.

[76] Yang Liu, Wissam Sid-Lakhdar, Elizaveta Rebrova, Pieter Ghysels, and Xiaoye Sherry Li. A parallel hierarchical blocked adaptive cross approximation algorithm. *The International Journal of High Performance Computing Applications*, 34(4):394–408, 2020.

[77] Yuying Liu, Colin Ponce, Steven L. Brunton, and J. Nathan Kutz. Multiresolution convolutional autoencoders. *Journal of Computational Physics*, 474:111801, 2023.

[78] Jean-Christophe Loiseau and Steven Brunton. Constrained sparse galerkin regression. *Journal of Fluid Mechanics*, 838, 11 2016.

[79] M. S. Long, R. Yantosca, J. E. Nielsen, C. A. Keller, A. da Silva, M. P. Sulprizio, S. Pawson, and D. J. Jacob. Development of a grid-independent geos-chem chemical transport model (v9-02) as an atmospheric chemistry module for earth system models. *Geoscientific Model Development*, 8(3):595–602, 2015.

[80] Michael W Mahoney et al. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

[81] Niall Mangan, Steven Brunton, Joshua Proctor, and J. Kutz. Inferring biological

networks by sparse identification of nonlinear dynamics. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2, 05 2016.

[82] Krithika Manohar, Bingni Brunton, J. Kutz, and Steven Brunton. Data-driven sparse sensor placement for reconstruction. *IEEE control systems*, 38:63–, 05 2018.

[83] Krithika Manohar, Steven L. Brunton, and J. Nathan Kutz. Environment identification in flight using sparse approximation of wing strain. *Journal of Fluids and Structures*, 70:162–180, 2017.

[84] Krithika Manohar, Thomas Hogan, Jim Buttrick, Ashis G. Banerjee, J. Nathan Kutz, and Steven L. Brunton. Predicting shim gaps in aircraft assembly with machine learning and sparse sensing. *Journal of Manufacturing Systems*, 48:87–95, 2018.

[85] Krithika Manohar, Eurika Kaiser, Steven L. Brunton, and J. Nathan Kutz. Optimized sampling for multiscale dynamics. *Multiscale Modeling & Simulation*, 17(1):117–136, 2019.

[86] Krithika Manohar, J. Kutz, and Steven Brunton. Optimal sensor and actuator selection using balanced model reduction. *IEEE Transactions on Automatic Control*, PP:1–1, 01 2021.

[87] J. Mao, A. Carlton, R. C. Cohen, W. H. Brune, S. S. Brown, G. M. Wolfe, J. L. Jimenez, H. O. T. Pye, N. Lee Ng, L. Xu, V. F. McNeill, K. Tsigaridis, B. C. McDonald, C. Warneke, A. Guenther, M. J. Alvarado, J. de Gouw, L. J. Mickley, E. M. Leibensperger, R. Mathur, C. G. Nolte, R. W. Portmann, N. Unger, M. Tosca, and L. W. Horowitz. Southeast atmosphere studies: learning from model-observation syntheses. *Atmospheric Chemistry and Physics*, 18(4):2615–2651, 2018.

[88] J. Mao, D. J. Jacob, M. J. Evans, J. R. Olson, X. Ren, W. H. Brune, J. M. St. Clair, J. D. Crounse, K. M. Spencer, M. R. Beaver, P. O. Wennberg, M. J. Cubison, J. L. Jimenez, A. Fried, P. Weibring, J. G. Walega, S. R. Hall, A. J. Weinheimer, R. C. Cohen, G. Chen, J. H. Crawford, C. McNaughton, A. D. Clarke, L. Jaeglé, J. A. Fisher, R. M. Yantosca, P. Le Sager, and C. Carouge. Chemistry of hydrogen oxide

radicals (ho_x) in the arctic troposphere in spring. *Atmospheric Chemistry and Physics*, 10(13):5823–5838, 2010.

[89] Jingqiu Mao, Fabien Paulot, Daniel J. Jacob, Ronald C. Cohen, John D. Crounse, Paul O. Wennberg, Christoph A. Keller, Rynda C. Hudman, Michael P. Barkley, and Larry W. Horowitz. Ozone and organic nitrates over the eastern united states: Sensitivity to isoprene chemistry. *Journal of Geophysical Research: Atmospheres*, 118(19):11,256–11,268, 2013.

[90] Per-Gunnar Martinsson. Randomized methods for matrix computations. *arXiv preprint arXiv:1607.01649*, 2016.

[91] Lee T. Murray, Daniel J. Jacob, Jennifer A. Logan, Rynda C. Hudman, and William J. Koshak. Optimized regional and interannual variability of lightning in a global chemical transport model constrained by lis/otd satellite data. *Journal of Geophysical Research: Atmospheres*, 117(D20):n/a–n/a, 2012. D20307.

[92] Vidvuds Ozolins, Rongjie Lai, Russel Caflisch, and Stanley Osher. Compressed modes for variational problems in mathematics and physics. *Proceedings of the National Academy of Sciences of the United States of America*, 110:18368–18373, 10 2013.

[93] Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[94] Liam Paninski. Asymptotic theory of information-theoretic experimental design. *Neural computation*, 17:1480–507, 08 2005.

[95] E. Parish and K. Carlberg. Time-series machine-learning error models for approximate solutions to parameterized dynamical systems. *Computer Methods in Applied Mechanics and Engineering*, 365:112990, 2020.

[96] J. P. Parrella, D. J. Jacob, Q. Liang, Y. Zhang, L. J. Mickley, B. Miller, M. J. Evans, X. Yang, J. A. Pyle, N. Theys, and M. Van Roozendael. Tropospheric bromine

chemistry: implications for present and pre-industrial ozone and mercury. *Atmospheric Chemistry and Physics*, 12(15):6723–6740, 2012.

[97] Kurtis G Paterson, Jessica L Sagady, Dianne L Hooper, Steve B Bertman, Mary Anne Carroll, and Paul B Shepson. Analysis of air quality data using positive matrix factorization. *Environmental Science & Technology*, 33(4):635–641, 1999.

[98] J.L. Proctor, Steven Brunton, Bingni Brunton, and J.N. Kutz. Exploiting sparsity and equation-free architectures in complex systems. *European Physical Journal: Special Topics*, 223:2665–2684, 12 2014.

[99] Joshua L. Proctor, Steven L. Brunton, and J. Nathan Kutz. Dynamic mode decomposition with control. *SIAM Journal on Applied Dynamical Systems*, 15(1):142–161, 2016.

[100] Tong Qin, Kailiang Wu, and Dongbin Xiu. Data driven governing equations approximation using deep neural networks. *Journal of Computational Physics*, 395:620–635, oct 2019.

[101] Alfio Quarteroni, Andrea Manzoni, and Federico Negri. *Reduced basis methods for partial differential equations: An introduction*. 01 2015.

[102] Francesco Regazzoni, Dominique Chapelle, and Philippe Moireau. Combining data assimilation and machine learning to build data-driven models for unknown long time dynamics—applications in cardiovascular modeling. *International Journal for Numerical Methods in Biomedical Engineering*, 37(7):e3471, 2021.

[103] G. Roberts, M. J. Wooster, and E. Lagoudakis. Annual and diurnal african biomass burning temporal dynamics. *Biogeosciences*, 6(5):849–866, 2009.

[104] V. Rokhlin, A. Szlam, and M. Tygert. A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1100–1124, 2010.

[105] Clarence Rowley, Igor Mezic, SHERVIN BAGHERI, Philipp Schlatter, and DAN

HENNINGSON. Spectral analysis of nonlinear flows. *Journal of Fluid Mechanics*, 641:115 – 127, 12 2009.

[106] Samuel Rudy, Steven Brunton, Joshua Proctor, and J. Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3, 09 2016.

[107] A. Sandu, J.G. Verwer, J.G. Blom, E.J. Spee, G.R. Carmichael, and F.A. Potra. Benchmarking stiff ode solvers for atmospheric chemistry problems ii: Rosenbrock solvers. *Atmospheric Environment*, 31(20):3459 – 3472, 1997.

[108] A. Sandu, J.G. Verwer, M. Van Loon, G.R. Carmichael, F.A. Potra, D. Dabdub, and J.H. Seinfeld. Benchmarking stiff ode solvers for atmospheric chemistry problems-i. implicit vs explicit. *Atmospheric Environment*, 31(19):3151 – 3166, 1997. EUMAC: European Modelling of Atmospheric Constituents.

[109] Diya Sashidhar and J Nathan Kutz. Bagging, optimized dynamic mode decomposition for robust, stable forecasting with spatial and temporal uncertainty quantification. *Philosophical Transactions of the Royal Society A*, 380(2229):20210199, 2022.

[110] Hayden Schaeffer. Learning partial differential equations via data discovery and sparse optimization. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 473:20160446, 01 2017.

[111] Hayden Schaeffer, Russel Caflisch, Cory Hauck, and Stanley Osher. Sparse dynamics for partial differential equations. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 03 2013.

[112] Peter J. Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5â28, 2010.

[113] Paola Sebastiani and Henry P. Wynn. Maximum Entropy Sampling and Optimal Bayesian Experimental Design. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 62(1):145–157, 01 2002.

[114] T. Sherwen, M. J. Evans, R. Sommariva, L. D. J. Hollis, S. M. Ball, P. S. Monks, C. Reed, L. J. Carpenter, J. D. Lee, G. Forster, B. Bandy, C. E. Reeves, and W. J. Bloss. Effects of halogens on european air-quality. *Faraday Discuss.*, 200:75–100, 2017.

[115] Raquel A Silva, J Jason West, Yuqiang Zhang, Susan C Anenberg, Jean-FranÃ§ois Lamarque, Drew T Shindell, William J Collins, Stig Dalsoren, Greg Faluvegi, Gerd Folberth, Larry W Horowitz, Tatsuya Nagashima, Vaishali Naik, Steven Rumbold, Ragnhild Skeie, Kengo Sudo, Toshihiko Takemura, Daniel Bergmann, Philip Cameron-Smith, Irene Cionni, Ruth M Doherty, Veronika Eyring, Beatrice Josse, I A MacKenzie, David Plummer, Mattia Righi, David S Stevenson, Sarah Strode, Sophie Szopa, and Guang Zeng. Global premature mortality due to anthropogenic outdoor air pollution and the contribution of past climate change. *Environmental Research Letters*, 8(3):034005, 2013.

[116] Lawrence Sirovich. Turbulence and the dynamics of coherent structures. i - coherent structures. ii - symmetries and transformations. iii - dynamics and scaling. *Quarterly of Applied Mathematics - QUART APPL MATH*, 45, 10 1987.

[117] D. S. Stevenson, F. J. Dentener, M. G. Schultz, K. Ellingsen, T. P. C. van Noije, O. Wild, G. Zeng, M. Amann, C. S. Atherton, N. Bell, D. J. Bergmann, I. Bey, T. Butler, J. Cofala, W. J. Collins, R. G. Derwent, R. M. Doherty, J. Drevet, H. J. Eskes, A. M. Fiore, M. Gauss, D. A. Hauglustaine, L. W. Horowitz, I. S. A. Isaksen, M. C. Krol, J.-F. Lamarque, M. G. Lawrence, V. Montanaro, J.-F. MÃŒller, G. Pitari, M. J. Prather, J. A. Pyle, S. Rast, J. M. Rodriguez, M. G. Sanderson, N. H. Savage, D. T. Shindell, S. E. Strahan, K. Sudo, and S. Szopa. Multimodel ensemble simulations of present-day and near-future tropospheric ozone. *Journal of Geophysical Research: Atmospheres*, 111(D8), 2006.

[118] Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 12 2018.

[119] Zhou Ting and Jiang Hui. Eeg signal processing based on proper orthogonal decomposition. pages 636–640, 07 2012.

[120] N. Trendafilov, I. T. Jolliffe, and M. Uddin. A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, 12:531–â547, 2003.

[121] Jonathan H. Tu, Clarence W. Rowley, Dirk M. Luchtenburg, Steven L. Brunton, and J. Nathan Kutz. On dynamic mode decomposition: Theory and applications, 2014.

[122] Meghana Velegar, N Benjamin Erichson, Christoph A Keller, and J Nathan Kutz. Scalable diagnostics for global atmospheric chemistry using ristretto library (version 1.0). *Geoscientific Model Development*, 12(4):1525–1539, 2019.

[123] Sergey Voronin and Per-Gunnar Martinsson. 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.

[124] Yu-Long Xie, Philip K Hopke, Pentti Paatero, Leonard A Barrie, and Shao-Meng Li. Identification of source nature and seasonal variations of arctic aerosol bypositive matrix factorization. *Journal of the Atmospheric Sciences*, 56(2):249–260, 1999.

[125] Cao Yanhua, Jiang Zhu, and Ionel Navon. Reduced-order modeling of the upper tropical pacific ocean model using proper orthogonal decomposition. *Computers Mathematics with Applications*, 52:1373–1386, 10 2006.

[126] Cao Yanhua, Jiang Zhu, and Ionel Navon. A reducedâorder approach to fourâdimensional variational data assimilation using proper orthogonal decomposition. *International Journal for Numerical Methods in Fluids*, 53:1571 – 1583, 04 2007.

[127] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2003.