

Optimized Dynamic Mode Decomposition for Reconstruction and Forecasting of Atmospheric Chemistry Data

Meghana Velegar¹, Christoph Keller^{2,3} and J. Nathan Kutz^{1,4}

¹ *Department of Applied Mathematics, University of Washington, Seattle, WA*

² *Morgan State University, Baltimore, MD, USA*

³ *NASA Global Modelling and Assimilation Office,*

Goddard Space Flight Center, Greenbelt, MD, USA and

⁴ *Department of Electrical and Computer Engineering, University of Washington, Seattle, WA*

We introduce the optimized dynamic mode decomposition algorithm for constructing an adaptive and computationally efficient reduced order model and forecasting tool for 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. Forecasting is also achieved with a linear model that uses a linear superposition of the dominant spatio-temporal features. The DMD method is demonstrated on three months of global chemistry dynamics data, showing its significant performance in computational speed and interpretability. We show that the presented decomposition method successfully extracts known major features of atmospheric chemistry, such as summertime surface pollution and biomass burning activities. Moreover, the DMD algorithm allows for rapid reconstruction of the underlying linear model, which can then easily accommodate non-stationary data and changes in the dynamics.

I. 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 [1]. 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 years. 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 that can become massive as the resolution of each dimension is increased. Dimensionality reduction is a critically enabling aspect of machine learning and data science [2] that can be leveraged to approximate the monitoring and forecasting capabilities of global chemistry with more readily tractable computational algorithms [3]. *Dynamic mode decomposition* (DMD) is a data-driven regression architecture for adaptively learning linear dynamics models over snapshots of temporal data, specifically in a low-dimensional subspace. DMD has been broadly used in the scientific community due to its ease of use, interpretability and adaptive nature [4]. When applied to the spatio-temporal dynamics of atmospheric chemistry, we demonstrate that the method provides an effective and computationally 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) [6] versions of the DMD algorithm are critical for characterizing the complexities of the chemical interaction dynamics and uncertainties.

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 the underlying 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 scheme in such a spatio-temporal system. Building models from sensor data directly is no different: sensors must be 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 an attractive alternative to large scale computing. ROMs provide a mathematical architecture for reducing the computational complexity of mathematical models in numerical simulations [7–10]. 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 [11], 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 $t + \Delta t$ [12, 13]. In-

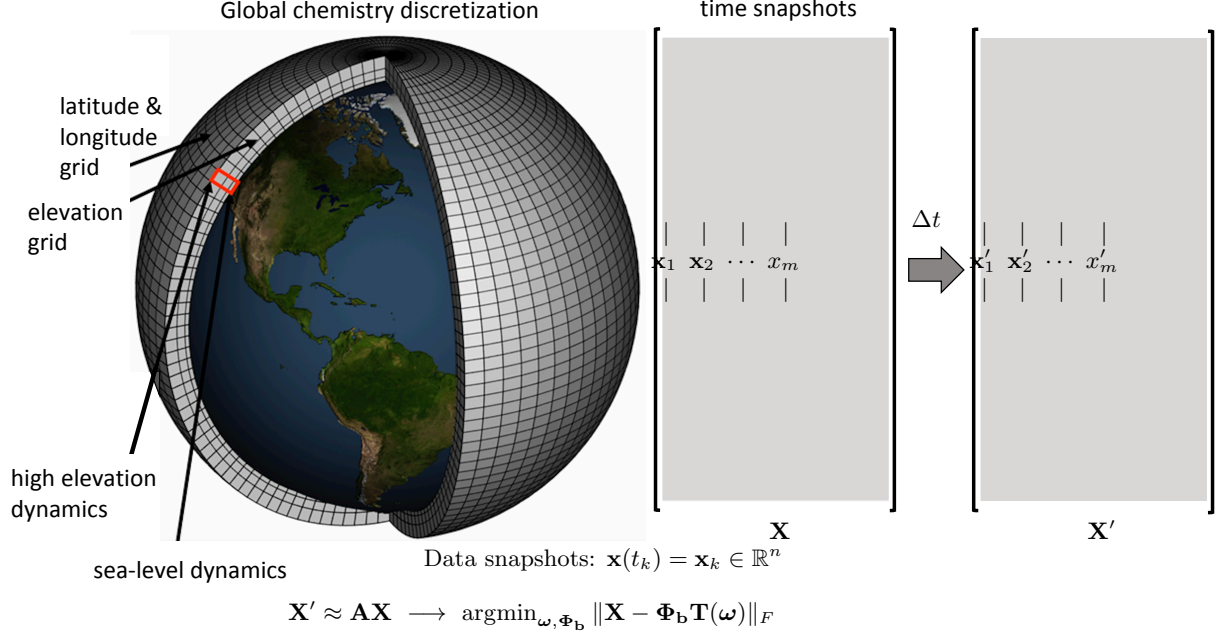


FIG. 1. The spatial grid for atmospheric chemistry data sets on the left panel. The data $\mathbf{x}(t_k)$ is collected into snapshot matrices $\mathbf{X}$ which are used to regress to the best exponential (linear) solution $\operatorname{argmin}_{\omega, \Phi_b} \|\mathbf{X} - \Phi_b \mathbf{T}(\omega)\|_F$, where Φ_b are the weighted DMD modes and $\mathbf{T}$ is a matrix of exponentials for fitting the data (6).

deed, 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, [14] and [15] developed a suite of neural network based methods for learning time-stepping models for tropospheric bromine chemistry and cardiovascular dynamics, respectively. Moreover, [14] 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 in order for the results to be valid for withheld test sets. 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 an algorithm by [16] and has rapidly become a commonly used data-driven analysis tool. It is the leading approximation method for the Koopman (linear) op-

erator from data [17]. DMD by construction provides a method for identifying spatio-temporal coherent structures in high-dimensional time-series data. DMD analysis offers a dynamic version of standard dimensionality reduction methods such as the *proper orthogonal decomposition* (POD), which highlighted low-rank features in spatio-temporal data [18]. 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

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

where $\mathbf{x}(t)$ is an r -rank approximation to a collection of state space measurements $\mathbf{x}_k = \mathbf{x}(t_k)$ ($k = 1, 2, \dots, n$). 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 [19]. 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 [4, 20]. DMD is modular due to its simple formulation in terms of linear algebra, resulting in innovations related to control [21, 22], compression [23, 24], reduced-order modeling [25], and multi-resolution analysis [26, 27], among others.

II. ATMOSPHERIC CHEMISTRY DATA SETS, DATA PRE-PROCESSING, AND METHODS

A. Atmospheric chemistry model

Understanding the composition of the atmosphere is critical for a wide range of applications, including air quality, stratospheric ozone loss, and environmental degradation [1]. Chemical transport models (CTM) are used to simulate the evolution of atmospheric constituents in space and time [28]. A CTM solves the system of coupled continuity equations for an ensemble of m species with number density vector $\mathbf{n} = (n_1, \dots, n_m)^T$ via operator splitting of transport and local processes:

$$\frac{\partial n_i}{\partial t} = -\nabla \cdot (n_i \mathbf{U}) + (P_i - L_i)(\mathbf{n}) + E_i - D_i \quad i \in [1, m] \quad (2)$$

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,

$$\frac{\partial n_i}{\partial t} = -\nabla \cdot (n_i \mathbf{U}) \quad i \in [1, m] \quad (3)$$

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

$$\frac{dn_i}{dt} = (P_i - L_i)(\mathbf{n}) + E_i - D_i \quad i \in [1, m] \quad (4)$$

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

Chemistry models repeatedly solve equations (3) and (4), 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, species) can become massive, which makes it impractical to output them at high temporal frequency and refined spatial resolution. 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. 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 sub-sampling makes it possible to represent the majority 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.

1. Global Atmospheric Chemistry Simulations

The reference simulation of atmospheric composition was generated using the GEOS-Chem model, as described in [3]. GEOS-Chem (<https://geoschem.github.io>) is an open-source global model of atmospheric chemistry used for a wide range of applications. The model can be run in offline mode as a chemical transport model (CTM) [29, 30] or as an online component within the NASA Goddard Earth System Model (GEOS) [31, 32]. The dataset used here was produced using 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). Model chemistry includes detailed HOx-NOx-VOC-ozone-BrOx tropospheric chemistry as originally described by [29], with addition of BrOx chemistry by [33] and updates to isoprene oxidation as described by [34]. Stratospheric chemistry is simulated using a linearized mechanism as described by [35].

The model output covers one year (July 2013 - June 2014) at $4^\circ \times 5^\circ$ horizontal resolution, providing a comprehensive set of atmospheric chemistry model diagnostics. For every chemistry time step of 20 minutes, the concentrations of all 143 chemical constituents were archived immediately before and after chemistry in units of molecules/cm³. The difference between these concentration pairs are the species tendencies due to chemistry (expressed in units of molecules/cm³/s). Since the solution of chemical kinetics is sensitive to 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 [36] as implemented in GEOS-Chem by [37] and [38]. At every time step, the data set thus consists of $n_{\text{features}} = 143 + 91 + 3 + 143 = 380$ data points at every grid location. We restrict our analysis to the lowest 30 model levels to avoid influence from the stratosphere. The resulting data set has dimensions $n_{\text{lon}} \times n_{\text{lat}} \times n_{\text{lev}} \times n_{\text{times}} \times n_{\text{features}} = 72 \times 46 \times 30 \times 26280 \times 380 = 9.9 \times 10^{11}$.

B. Data Pre-Processing

Many dimensionality reduction techniques rely on an underlying singular value decomposition of the data that extracts correlated patterns in the data. A fundamental weakness of such SVD-based approaches is the inability

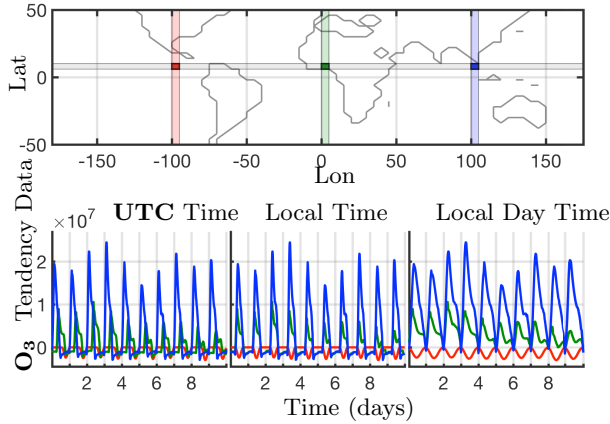


FIG. 2. Shifting the data for each cell in time to align the local time zones across a latitude to the prime meridian (Lon = 0°) local time, shown here for O_3 tendency data for Lat = 30°. The bottom left panel is the raw data for the 3 highlighted cells, the bottom center panel is this data shifted in time, and the bottom right panel shows isolated day time values only.

to efficiently handle invariances in the data. Specifically, translational and/or rotational invariances of low-rank features in the data are not well captured [2–4, 18]. 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 day time, 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 O_3 (Ozone) is plotted with respect to UTC time for one latitude = 30°/elevation = 1 and three different longitudes = [−100°, 0°, 100°] on bottom left in Fig. 2, highlighting the translational invariance in the absolute concentration data. Any SVD-based approach will be unable to capture this translational invariance and correlate across snapshots in time, producing an artificially high dimensionality, i.e., higher number of modes would be needed to characterize the dynamics due to translation [2, 18]. To overcome this issue the time series for each grid point are shifted to align with the GMT time, as shown on bottom middle in Fig. 2. With the local times for each grid point aligned SVD-based dimensionality reduction techniques can now identify and isolate coherent low-dimensional features in the data. Similarly, the current season dictates length of days and nights. Latitudes where the days are very short, i.e., the ‘turn-on’ times are very short, the chemistry exhibits “spiky” patterns. SVD-based approaches would again need an artificially high number of modes to capture the low-rank features in the data. To work around this issue the day time chemistry can be isolated and analysis performed on the isolated day times, especially if there is total ‘turn-off’ of dynamics during night

For NO_{CONC} data

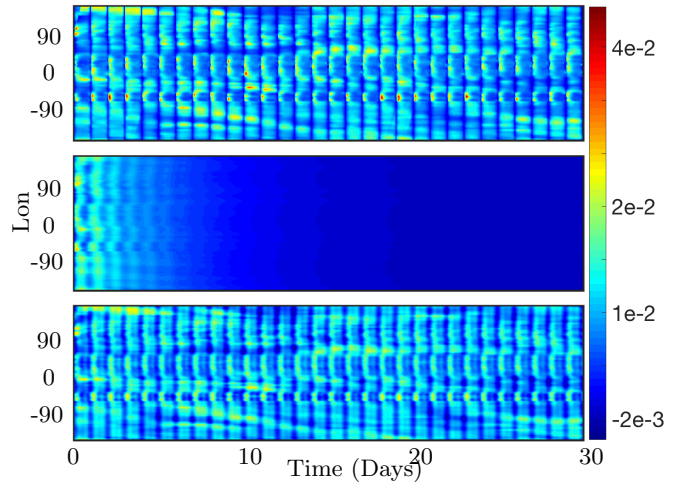


FIG. 3. Comparing 30 day reconstruction results for Classical and Optimized DMD at the surface of NO preprocessed data at Lat = 30°. The results are for absolute concentration or $CONC$ data; the top panel shows the preprocessed data, the middle panel shows the reconstruction from the Classical DMD, 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.

times. On the bottom right the day time chemistry is isolated showing only the non-zero data during daytime.

C. Optimized Dynamic Mode Decomposition (DMD)

The DMD algorithm schematic is shown in the right panel of Fig. 1. The DMD algorithm seeks the leading spectral decomposition of the best fit linear operator $\mathbf{A}$ [2] that approximately advances the snapshot measurements of the state of a system $\mathbf{x} \in \mathbb{R}^n$ forward in time by stepsize Δt :

$$\mathbf{X}' \approx \mathbf{A}\mathbf{X} \quad (5)$$

which leads to the mathematical definition of operator $\mathbf{A}$ as the best fit one-step operator [39].

However, the DMD formulated by this regression 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 (5) is extremely sensitive to noise in the data, causing a bias in the computed DMD modes and eigenvalues [40–42]. The *optimized DMD* algorithm of Askham and Kutz [5], which uses a variable projection method [43] 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

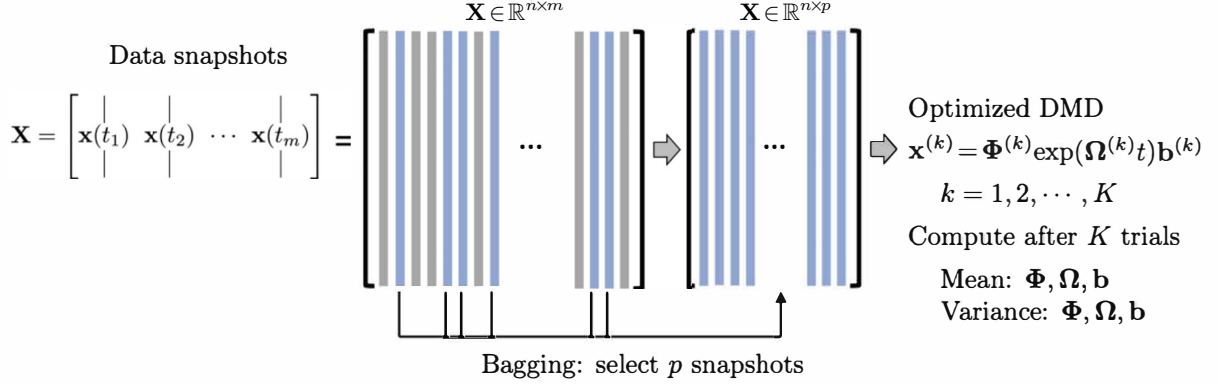


FIG. 4. Summary of the BOP-DMD architecture reproduced with permission from [6]. The data snapshots $\mathbf{x}(t_k)$ are collected over m snapshots into the matrix $\mathbf{X}$. Columns of $\mathbf{X}$ are randomly sub-selected into the matrix $\mathbf{X}^{(k)}$ to build an optimized DMD model. Each DMD model $\mathbf{x}^{(k)} = \Phi^{(k)} \exp(\Omega^{(k)} t) \mathbf{b}^{(k)}$ is used to compute the statistics (mean and variance) of the DMD parametrizations $\Phi, \Omega, \mathbf{b}$ which are used in building a the BOP-DMD ensemble solution with Uncertainty Quantification (UQ).

performance is mathematically guaranteed by the exponential fitting procedure of Askham and Kutz [5]. The exponential fitting is given by

$$\operatorname{argmin}_{\omega_k, \phi_k, b_k} \left\| \mathbf{X} - \sum_{k=1}^r b_k \phi_k \exp(\omega_k t) \right\| \quad (6)$$

where a rank r approximation is estimated. As noted, optimized DMD iterates to a solution of this non-convex problem by using variable projection [44]. This has been shown to provide a superior decomposition due to its ability to optimally suppress noise bias and handle snapshots collected at arbitrary times. Fig. 3 shows a comparison of surface nitrogen oxide (NO) as produced by GEOS-Chem (top panel), reconstructed using classical or exact DMD (middle panel), and using optDMD (bottom panel). The classical DMD reconstruction dies out within a few days, failing in the task of even reconstructing the time-series data, let alone forecasting, as it was originally regressed to. In contrast, the optDMD is able to capture, sustain and faithfully reconstruct the original time series.

We can also introduce constraints to the optDMD algorithm, including constraining all the DMD eigenvalues in (6) to (i) The imaginary axis:

$$\text{subject to } \Re(\omega_k) = 0 \quad (7)$$

(ii) The closed left-half plane:

$$\text{subject to } \Re(\omega_k) \leq 0 \quad (8)$$

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 through variable projection [43], often which can at times fail to converge.

D. Bagging OPTimized Dynamic Mode Decomposition (BOP-DMD)

BOP-DMD [6] leverages Breiman's statistical bagging sampling strategy [45] in partnership with the optimized DMD algorithm. The BOP-DMD architecture is presented in Fig. 4. 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 [46]. Further innovations include stabilizing the variable projection technique used by optDMD so that it converges consistently to an optimal solution [6]. 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 Φ_0 can be used to seed the BOP-DMD. p snapshots are randomly selected from the full data matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, to form a subset data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$. optDMD produces the model for this subset data, and we save the resulting model parameters. The process is repeated for K trials producing an ensemble of optDMD models. The mean $\{\langle \Phi \rangle, \langle \Omega \rangle, \langle \mathbf{b} \rangle\}$ and variance $\{\langle \Phi^2 \rangle, \langle \Omega^2 \rangle, \langle \mathbf{b}^2 \rangle\}$ of the model parameters $\Phi, \Omega, \mathbf{b}$ can now be computed. Hence, in addition to producing the DMD model itself,

Algorithm1 BOP-DMD

Input: Input $(\mathbf{X}, p, K)$
procedure BOPDMD($\mathbf{X}, p, K$)
 Compute $\Phi_0, \Omega_0, \mathbf{b}_0$ $\triangleright$ optDMD regression
 for $k \in \{1, 2, \dots, K\}$ **do** $\triangleright$ Compute K optDMD models.
 Choose p of m snapshots ($p < m$) $\triangleright$ Bagging
 optDMD $\Phi_k, \Omega_k, \mathbf{b}_k$ $\triangleright$ Initialize with Ω_0
 Update $\Phi, \Omega, \mathbf{b}$ $\triangleright$ Add $\Phi_k, \Omega_k, \mathbf{b}_k$ to $\Phi, \Omega, \mathbf{b}$
 end for
 Compute mean $\mu = \{\langle \Phi \rangle, \langle \Omega \rangle, \langle \mathbf{b} \rangle\}$
 Compute variance $\sigma = \{\langle \Phi^2 \rangle, \langle \Omega^2 \rangle, \langle \mathbf{b}^2 \rangle\}$
 return μ, σ $\triangleright$ Return optDMD parameters.
end procedure

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.

III. RESULTS

The analysis is performed for preprocessed or time-shifted raw data for 60 days, from July, 2ND - August, 30TH. This time period is characterized by very active photo-chemistry in the Northern Hemisphere. The photolysis rate dictates a different kinetic environment for many key species of interest. To simplify interpretation, the analysis is performed on surface data (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 daylight chemistry period is much shorter as compared to the nighttime chemistry period. Thus, the data exhibits a spiky pattern that needs much higher modes to accurately reconstruct it; and/or we would need to isolate the day time values only when there are active chemical kinetics present. Hence, we are picking latitude = 30°N for the analysis, which has the longest day times for the latitudes considered. The first 40 days of data is used as 'training' data, and the DMD diagnostics below 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 six different chemical species of interest [3]: Nitrous Oxide **NO**, Ozone **O₃**, Nitrous dioxide **NO₂**, Hydroxyl radical **OH**, Isoprene **ISOP**, and Carbon Monoxide **CO**. For each species, we have **CONC** or absolute concentration data (expressed in units of molecules/cm³) and **TEND** or tendency/rate of change data (expressed in units of molecules/cm³/s). Using the diagnostics from the 40 day training period (July 2 - August 10), we then forecast the chemical evolution for the following 20 days (August 11 - 30).

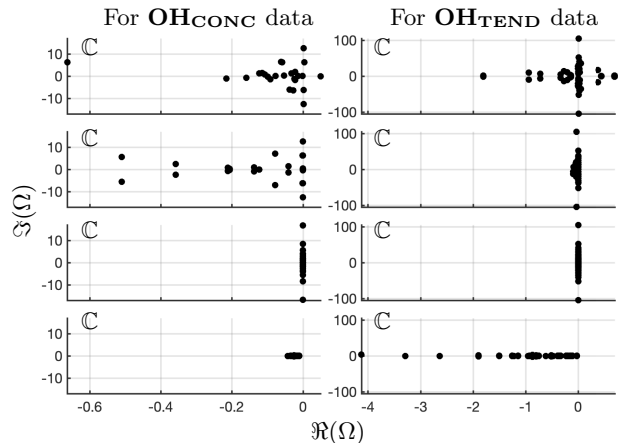


FIG. 5. Comparing the spectrum for 40 day reconstruction results for Classical and Optimized DMD at the surface of **OH** preprocessed data. On the left 4 panels are the eigenvalues of **OH_{CONC}** data; on the right 4 panels are the eigenvalues of **OH_{TEND}** at Lat = 30°. The top panels show the 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.

A. DMD Diagnostics

The optDMD decomposes data into time dynamics represented by the spectrum of eigenvalues Ω and the corresponding spatial modes Φ . We will be presenting diagnostics from four different DMD approaches: (i) optDMD without constraining the eigenvalues; (ii) optDMD with eigenvalues constrained to the left-half plane; (iii) optDMD with eigenvalues constrained to the imaginary axis; and finally (iv) exact DMD. This is to examine which decomposition is best suited for reconstruction and forecasting of the chemistry dynamics. The diagnostics are presented for the 40-day time series of the hydroxyl radical species (**OH**). The results are consistent for all chemical species of interest. We have used a hard rank threshold truncation of $r = 25$ for the **CONC** data and $r = 50$ for the **TEND** data. Truncating the rank for the DMD models is described below. The diagnostics are presented for both absolute concentration of the chemical species, or **OH_{CONC}** data, on the left panels and rate of change of concentrations/tendencies due to chemistry, or **OH_{TEND}** data, on the right panels in Fig. 5 and Fig. 6. Four different spectra of the DMD eigenvalues are presented in Fig. 5, and the corresponding reconstruction of data is shown in panels 2-5 of Fig. 6. The top two panels in Fig. 6 are the actual **OH_{CONC}** data on the left and actual **OH_{TEND}** data on the right, presented for comparison.

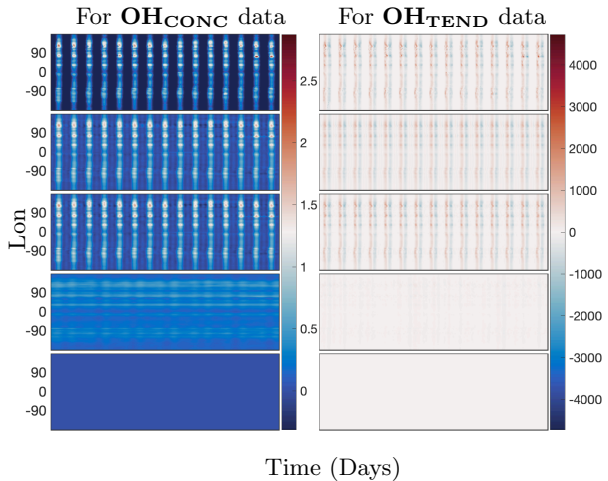


FIG. 6. Comparing 40 day reconstruction results for Classical, optimized DMD, and optimized DMD with no constraints at the surface of **OH** preprocessed data at $\text{Lat} = 30^\circ$. The left panel is for absolute concentration or **CONC** data and the right panel is for Tendency data; the top panels 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.

1. The spectrum for optDMD with no constraints on the eigenvalues for **OH_{CONC}** data is presented on the top left panel, and for **OH_{TEND}** data is presented on the top right panel of Fig. 5. 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 Fig. 6. 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.
2. 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 Fig. 5. The corresponding reconstruction of data is presented in the third two panels of Fig. 6. optDMD with these constraints not only faithfully reconstructs the data, but the forecasting results are also accurate, as presented in the following section.

3. 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 Fig. 5. The corresponding reconstruction of data is presented in the fourth two panels of Fig. 6. optDMD with these constraints is not able to capture the data dynamics, and will not be used henceforth.

4. Finally, results from Exact DMD for both data sets are presented in the bottom two panels of Fig. 5 and Fig. 6. 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.

Thus, we will 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**, **O₃**, **NO₂**, **OH**, **ISOP**, **CO**, **CONC** and **TEND** data, we now compute the relative error 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 Fig. 7. A larger number of modes is needed to reconstruct the **TEND** data as compared to the **CONC** data. Based on the results, we use 20-30 modes for optimal diagnostics of **CONC** data, depending on the chemical species. For the **TEND** data we pick between 30-50 modes.

Finally, we present the global spatial modes for **CO** and **NO** computed at 12° latitudes -14° through 30° in Fig. 8 and Fig. 9 respectively. The 12 latitudes are selected for having consistent day lengths across all longitudes and at least 4 snapshots during day time. As described above, the optDMD is performed for one latitude at a time to have consistent day time 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.

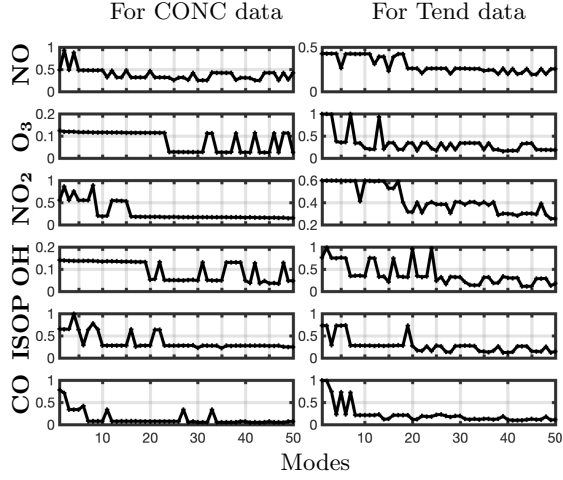


FIG. 7. Relative Error plotted against number of modes used for Optimized **DMD** with eigenvalues constrained to the left-half plane; for 6 different chemical species and **CONC** and **TEND** data at Latitude=30°

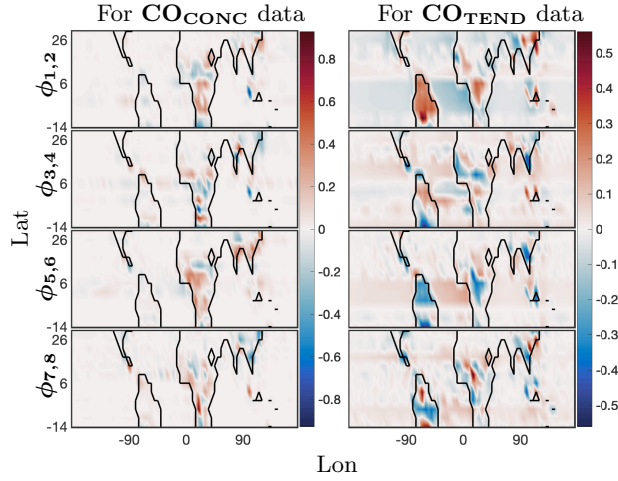


FIG. 8. 40 day reconstruction results for Optimized **DMD** at the surface of **CO** preprocessed data. The analysis was computed for 12 latitudes -14° through 30°. The left panel show the dominant four spatial modes for **CONC** data; and the right panel show four of the corresponding spatial modes for the **TEND** data.

B. 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 40-day training window and a given elevation/latitude. We now forecast the time series data for future times beyond the training window. Using (1), with amplitudes $\mathbf{b}$ /modes Φ /eigenvalues Ω computed by optDMD during the training window, we forecast time series for the subsequent 20 days. The results for **CONC** and **TEND** data for two chemical species **OH** and **NO** are presented for 6 longitudes, and latitude 30° at the sur-

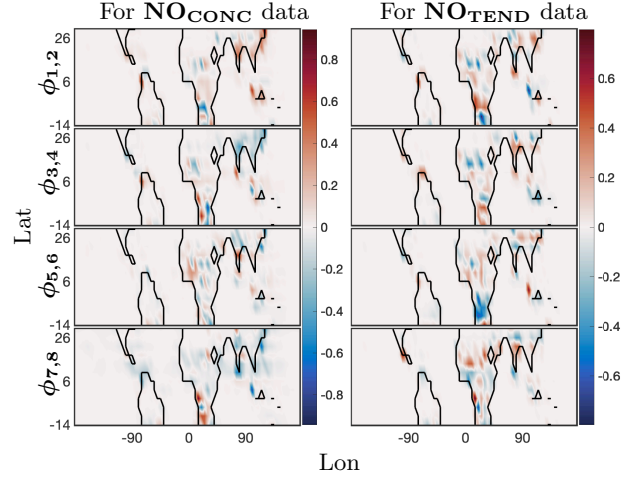


FIG. 9. 40 day reconstruction results for Optimized **DMD** at the surface of **NO** preprocessed data. The analysis was computed for 12 latitudes -14° through 30°. The left panel show four spatial modes for **CONC** data; and the right panel show four of the corresponding spatial modes for the **TEND** data.

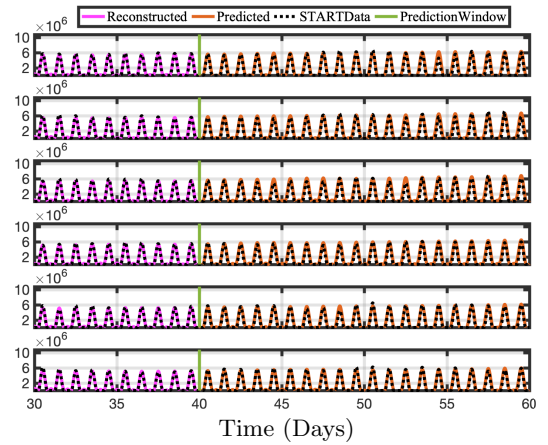


FIG. 10. Time series of reconstructed and predicted results with **OH_{CONC}** data at Lat 30° and 6 longitudes -180°:5°:-155°. Both the reconstructed data, shown here for 10 days; and the forecasted time series, shown here for the 20 day testing period, faithfully reconstruct and forecast the actual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

face(elevation=1) in Figures 10, 11, 12, and 13.

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

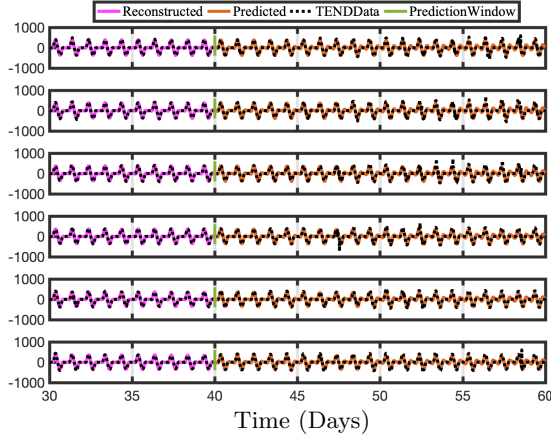


FIG. 11. Time series of reconstructed and predicted results with OH_{TEND} data at Lat 30° and 6 longitudes $-180^\circ:5^\circ:-155^\circ$. Again, both the reconstructed data, shown here for 10 days; and the forecasted time series, shown here for the 20 day testing period, faithfully reconstruct and forecast the actual data for OH_{TEND} .

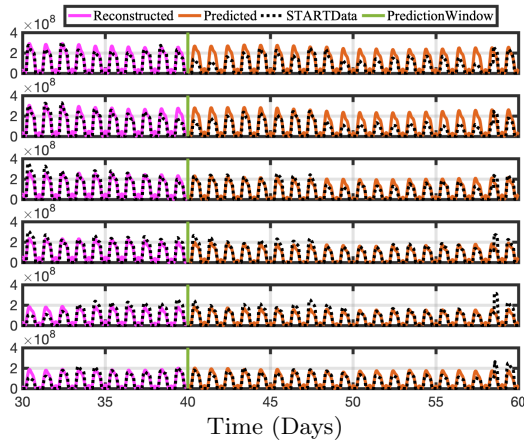


FIG. 12. Time series of reconstructed and predicted results with NO_{CONC} data at Lat 30° and 6 longitudes $-180^\circ:5^\circ:-155^\circ$. Both the reconstructed data, shown here for 10 days; and the forecasted time series, shown here for the 20 day testing period, reproduce the actual data for NO_{CONC} well.

sometimes not reproduced and we see a sinusoidal best fit time series instead. The NO_{TEND} results in Fig. 13 demonstrates this.

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 **CONC** and **TEND** data in Fig. 14 in color red. The 95-percentile confidence intervals

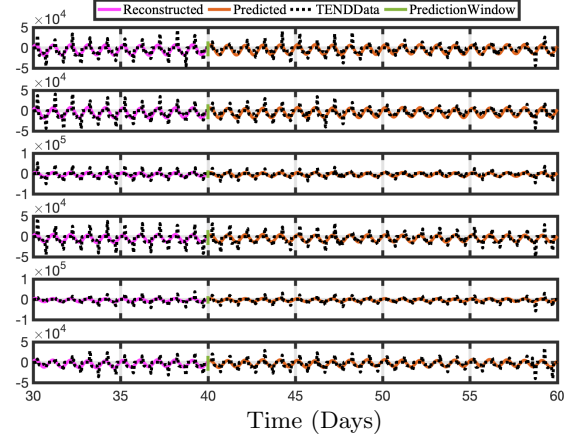


FIG. 13. Time series of reconstructed and predicted results with NO_{TEND} data at Lat 30° and 6 longitudes $-180^\circ:5^\circ:-155^\circ$. Both the reconstructed data, shown here for 10 days; and the forecasted time series, shown here for the 20 day testing period, do not capture the spikes in the actual data for NO_{TEND} . Since we are using only 20-30 modes for reconstruction, we get a sinusoidal best fit.

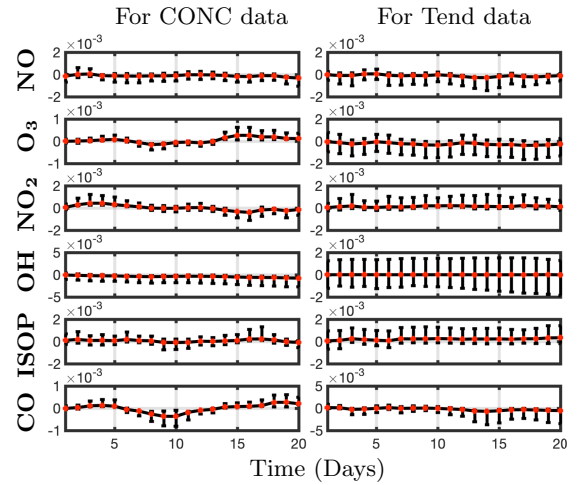


FIG. 14. Mean relative error with 95-percentile confidence intervals forecasting **CONC** and **TEND** data at Lat 30° for a prediction window of 20 days; and for 6 different chemical species. The relative error stays nearly the same or changes only slightly as the number of days we are forecasting out to increase. **optDMD** does better at forecasting **CONC** data as compared to the **TEND** data.

for each day is 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, 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-

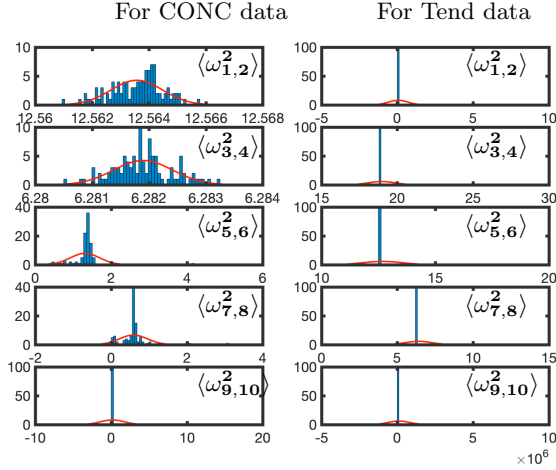


FIG. 15. Temporal uncertainty quantification for absolute of eigenvalues for **OH**_{CONC} and **OH**_{TEND} data at Lat 30°. The red lines represent a least-square fit of a normal distribution. 60 days of training data was used with a sample size of 3 days and 100 cycles.

series, while the underlying dynamics are forecast faithfully. Considering that the underlying dynamics represent a moving state with time, the constrained optDMD minimizes model bias with the variable projection optimization, thus leading to stable forecasting capabilities. The performance is slightly worse in forecasting the **TEND** data as compared to the **CONC** data, which is due to the intrinsic rank of the **TEND** data being higher. Increasing the truncation rank of the projection would lead to improvement in forecasting of the **TEND** data.

The optDMD performs worst in forecasting the chemical species **OH**. **OH** has a very short tropospheric life-time of less than a second and exhibits rapid chemical cycling during the daytime. Consequently, this chemical species needs the highest number of modes to capture its dynamics (Fig. 7).

C. Temporal Uncertainty Quantification

We now present the results from BOP-DMD in partnership with the optimized DMD algorithm to produce ensemble models and compute temporal uncertainty for the eigenvalue spectrum of both **CONC** 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, 2ND - August, 30TH) to create an initial seed $\Phi_0, \Omega_0, \mathbf{b}_0$ 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 Fig. 4. optDMD now computes the eigenvalues of various subsets using the aforementioned initial conditions.

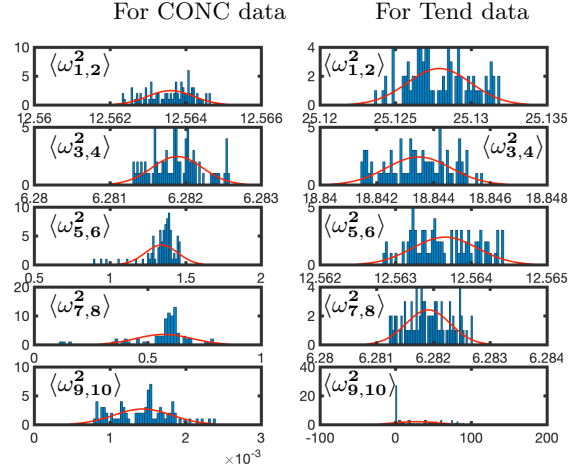


FIG. 16. Temporal uncertainty quantification for absolute of trimmed eigenvalues for with **OH**_{CONC} and **OH**_{TEND} data at Lat 30°. The data has been trimmed to remove outliers below 10 percentile and above 90 percentile. The red lines represent a least-square fit of a normal distribution.

The $K = 100$ ensemble models' eigenvalues are used to produce the temporal UQ metrics.

Fig. 15 shows the BOP-DMD distributions of the absolute value of the first five eigenvalues for each of the subsets of data for **OH**_{CONC} and **OH**_{TEND} data at Lat 30°. The BOP-DMD quantifies the temporal uncertainty 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 **OH**_{CONC} data and for all modes of **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. Fig. 16 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 **OH**_{TEND}.

IV. DISCUSSION

Based on the results presented in this work, we conclude that 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 producing a reasonable forecast. This is due to the significant bias in the model from energetic localized convective phenomena present in the atmospheric simulation data. The optDMD algorithm casts the regression problem as a nonlinear optimization enabled by

variable projection techniques [5], 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 **CONC** 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 accurate 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 **CONC** 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. In particular, the current

study did not look at the ability of optDMD to faithfully reproduce yearly patterns in the chemistry data, and accurately forecast seasonal variations. The BOP-DMD can be leveraged 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.

The here presented approaches have the potential to produce reliable estimates of ‘business-as-usual’ patterns of global atmospheric composition in real-time and at very low computational cost. They are not designed to capture unusual events such as air pollution due to wildfires or sudden pollutant emissions changes (as e.g. experienced in the wake of the COVID-19 outbreak). However, when combined with actual atmospheric observations, the presented method can be used to identify and quantify air pollution anomalies.

ACKNOWLEDGEMENTS

The authors acknowledge support from the National Science Foundation AI Institute in Dynamic Systems (grant number 2112085). JNK further acknowledges support from the Air Force Office of Scientific Research (FA9550-19-1-0011).

-
- [1] Daniel J Jacob. *Introduction to atmospheric chemistry*. Princeton university press, 1999.
 - [2] Steven L. Brunton and J. Nathan Kutz. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2019.
 - [3] 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.
 - [4] 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.
 - [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] Diya Sashidhar and J Nathan Kutz. Bagging, optimized dynamic mode decomposition for robust, stable forecast with spatial and temporal uncertainty quantification. *Philosophical Transactions of the Royal Society A*, 380(2229):20210199, 2022.
 - [7] 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.
 - [8] Athanasios C. Antoulas. *Approximation of Large-Scale Dynamical Systems*. Society for Industrial and Applied Mathematics, 2005.
 - [9] Alfio Quarteroni, Andrea Manzoni, and Federico Negri. *Reduced basis methods for partial differential equations: An introduction*. 01 2015.
 - [10] Jan Hesthaven, Gianluigi Rozza, and Benjamin Stamm. *Certified Reduced Basis Methods for Parametrized Partial Differential Equations*. 01 2016.
 - [11] 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.
 - [12] 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.
 - [13] 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.
- [14] 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.
 - [15] 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.
 - [16] Peter J. Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28, 2010.
 - [17] 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.
 - [18] J Nathan Kutz. *Data-driven modeling & scientific computation: methods for complex systems & big data*. Oxford University Press, 2013.
 - [19] 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.
 - [20] 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.
 - [21] 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.
 - [22] 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.
 - [23] N Benjamin Erichson, Sergey Voronin, Steven L Brunton, and J Nathan Kutz. Randomized matrix decompositions using r. *arXiv preprint arXiv:1608.02148*, 2016.
 - [24] Steven L. Brunton, Joshua L. Proctor, Jonathan H. Tu, and J. Nathan Kutz. Compressed sensing and dynamic mode decomposition. *Journal of Computational Dynamics*, 2(2):165–191, 2015.
 - [25] Alessandro Alla and J. Nathan Kutz. Nonlinear model order reduction via dynamic mode decomposition. *SIAM Journal on Scientific Computing*, 39(5):B778–B796, 2017.
 - [26] J. Nathan Kutz, Xing Fu, and Steven L. Brunton. Multiresolution dynamic mode decomposition. *SIAM Journal on Applied Dynamical Systems*, 15(2):713–735, 2016.
 - [27] Yuying Liu, Colin Ponce, Steven L. Brunton, and J. Nathan Kutz. Multiresolution convolutional autoencoders. *Journal of Computational Physics*, 474:111801, 2023.
 - [28] Guy P Brasseur and Daniel J Jacob. *Modeling of Atmospheric Chemistry*. Cambridge University Press, 2017.
 - [29] Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation. *Journal of Geophysical Research: Atmospheres*, 106(D19):23073–23095, 2001.
 - [30] 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.
 - [31] 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.
 - [32] 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.
 - [33] 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 Roozendaal. Tropospheric bromine chemistry: implications for present and pre-industrial ozone and mercury. *Atmospheric Chemistry and Physics*, 12(15):6723–6740, 2012.
 - [34] 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.
 - [35] 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.
 - [36] 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.
 - [37] 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.
 - [38] Sebastian D. Eastham, Debra K. Weisenstein, and Steven R.H. Barrett. Development and evaluation of the unified tropospheric-stratospheric chemistry extension (ucx) for the global chemistry-transport model geos-chem. *Atmospheric Environment*, 89:52 – 63, 2014.
 - [39] Jonathan H. Tu, Clarence W. Rowley, Dirk M. Luchtburg, Steven L. Brunton, and J. Nathan Kutz. On dynamic mode decomposition: Theory and applications.

- Journal of Computational Dynamics*, 1(2):391–421, 2014.
- [40] Shervin Bagheri. Effects of weak noise on oscillating flows: Linking quality factor, Floquet modes, and Koopman spectrum. *Physics of Fluids*, 26(9), 09 2014.
 - [41] 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.
 - [42] Maziar S. Hemati, Clarence W. Rowley, Eric A. Deem, and Louis N. Cattafesta. De-biasing the dynamic mode decomposition for applied koopman spectral analysis of noisy datasets. *Theoretical and Computational Fluid Dynamics*, 31(4):349–368, 2017.
 - [43] Gene Golub and Victor Pereyra. Separable nonlinear least squares: the variable projection method and its applications. *Inverse problems*, 19(2):R1, 2003.
 - [44] Gene H Golub and Victor Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on numerical analysis*, 10(2):413–432, 1973.
 - [45] Charles J. Stone R.A. Olshen Leo Breiman, Jerome Friedman. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.
 - [46] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. 12 2020.