Geosci. Model Dev., 18, 4667–4684, 2025
https://doi.org/10.5194/gmd-18-4667-2025

Geoscientific
Model Development
Open Access

Development and technical paper

Optimized dynamic mode decomposition for reconstruction and forecasting of atmospheric chemistry data

Meghana Velagar[1], Christoph Keller[2,3], and J. Nathan Kutz[1]

[1]Department of Applied Mathematics, University of Washington, Seattle, WA 98195, USA
[2]NASA Global Modeling and Assimilation Office, Goddard Space Flight Center, Greenbelt, MD, USA
[3]GESTAR II, Morgan State University, Baltimore, MD, USA

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

Received: 13 April 2024 – Discussion started: 1 October 2024
Revised: 24 March 2025 – Accepted: 25 March 2025 – Published: 30 July 2025

Abstract. We introduce the optimized dynamic mode decomposition (DMD) 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 3 months of global chemistry dynamics data, showing its significant performance in terms of computational speed and interpretability. We show that the presented decomposition method successfully extracts and forecasts chemical patterns for leading chemical indicators, including nitric oxide, ozone, nitrogen dioxide, hydroxyl radical, isoprene, and carbon monoxide. Moreover, the DMD algorithm allows for rapid reconstruction of the underlying linear model, which can then easily accommodate nonstationary data and changes in the dynamics.

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 (Jacob, 1999). 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 require full knowledge of the chemical state of the atmosphere at all locations and times, resulting in a 4-dimensional data set for longitude, latitude, elevation, and time for each chemical species 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 (Brunton and Kutz, 2019) that can be leveraged to approximate the monitoring and forecasting capabilities of global chemistry with more readily tractable computational algorithms (Velegar et al., 2019). *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 (Kutz et al., 2016a). 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 the characterization, monitoring, and forecasting of global chemical concentrations with either computational or sensor data. Moreover, we show that the optimized DMD algorithm (Askham and Kutz, 2018) and bagging optimized DMD (BOP-DMD) (Sashidhar and Kutz, 2022) versions of the DMD algorithm are critical for characterizing the complexities of the chemical interaction dynamics and their uncertainties.

Characterization of the 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.

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

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. Hence, 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 (Benner et al., 2015; Antoulas, 2005; Quarteroni et al., 2015; Hesthaven et al., 2016). 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 (Carlberg et al., 2017.), 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 $t + \Delta t$ (Qin et al., 2019; Liu et al., 2020; Gin et al., 2021). 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 proper orthogonal decomposition (POD); 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 (2020) and Regazzoni et al. (2021) 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 (2020) provided 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 chemistry modeling problematic. Certainly, the landscape of models is growing rapidly, with machine learning techniques especially proving useful in weather and temperature forecasting. These methods are driven by leading tech companies that, at scale, are training such models with many GPUs over long periods of time to achieve their exceptional performance. However, a computationally efficient and adaptive ROM approach is embodied by DMD, which is a simple regression requiring no training, cross-validation, and hyper-parameter tuning. It is a straight regression, much like a line fit. DMD was introduced as an algorithm by (Schmid, 2010) and has rapidly become a commonly used data-driven analysis tool. It is the leading approximation method for the Koopman (linear) operator from data (Rowley et al., 2009). 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 highlights low-rank features in spatio-temporal data (Kutz, 2013). 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

$$x(t) = \sum_{j=1}^{r} \boldsymbol{\phi}_j e^{\omega_j t} b_j = \Phi \exp(\Omega t) b, \qquad (1)$$

where $x(t)$ is an r-rank approximation to a collection of state-space measurements $x_k = x(t_k)$ $(k = 1, 2, \cdots, n)$. The algorithm regresses to values of the DMD eigenvalues ω_j, DMD modes $\boldsymbol{\phi}_j$, and their loadings b_j. ω_j determines the temporal behavior of the system associated with a modal structure $\boldsymbol{\phi}_j$. Such a regression can also be learned from time-series data (Lange et al., 2020). DMD may be thought of as a combination of singular value decomposition (SVD)/POD in space with the Fourier transform in time, combining the strengths of each approach (Chen et al., 2012; Kutz et al., 2016a). DMD is modular due to its simple formulation in terms of linear algebra, resulting in innovations related to control (Proctor et al., 2016; Deem et al., 2020), compression (Erichson et al., 2019; Brunton et al., 2015), reduced-order modeling (Alla and Kutz, 2017), and multiresolution analysis (Kutz et al., 2016b; Liu et al., 2023; Lapo et al., 2024), among others. The SVD/DMD can even be done on terabytes of data in seconds (Eiximeno et al., 2025).

2 Atmospheric chemistry data sets, data preprocessing, and methods

2.1 Atmospheric chemistry model

Many of the dominant spatio-temporal features of atmospheric chemistry are well-understood through extensive simulation and data collection (Jacob, 1999; Brasseur and Jacob, 2017). This will not be the focus of this work but rather a robust, computationally efficient, and accurate reduced-order model for reconstructing and forecasting the dynamics. Chemical transport models (CTMs) are used to simulate the evolution of atmospheric constituents in space and time (Brasseur and Jacob, 2017). A CTM solves the system of coupled continuity equations for an ensemble of m species with number density vector $\boldsymbol{n} = (n_1, \ldots, n_m)^T$ via operator splitting of transport and local processes:

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

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

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

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

$$\frac{\mathrm{d}n_i}{\mathrm{d}t} = (P_i - L_i)(\boldsymbol{n}) + E_i - D_i \quad i \in [1, m], \quad (4)$$

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

Chemistry models repeatedly solve Eqs. (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, and 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 that 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.

2.1.1 Global atmospheric chemistry simulations

The reference simulation of the atmospheric composition was generated using the GEOS-Chem model, as described in Velegar et al. (2019). GEOS-Chem (https://geoschem.github.io, last access: 12 June 2025) 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) (Bey et al., 2001; Eastham et al., 2018) or as an online component within the NASA Goddard Earth Observing System (GEOS) model (Long et al., 2015; Hu et al., 2018). The data set 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 HO_x-NO_x-VOC-ozone-BrO_x tropospheric chemistry as originally described by Bey et al. (2001), with the addition of BrO_x chemistry by Parrella et al. (2012) and updates to isoprene oxidation as described by Mao et al. (2013). Stratospheric chemistry is simulated using a linearized mechanism as described by Murray et al. (2012).

The model output covers 1 year (July 2013–June 2014) at 4° × 5° horizontal resolution, providing a comprehensive set of atmospheric chemistry model diagnostics. For every chemistry time step of 20 min, the concentrations of all 143 chemical constituents were archived immediately before and after chemistry in units of molec. cm^{-3}. The difference between these concentration pairs is the species tendencies due to chemistry (expressed in units of molec. $\mathrm{cm}^{-3}\,\mathrm{s}^{-1}$). Because 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 Bian and Prather (2002) as implemented in GEOS-Chem by Mao et al. (2010) and Eastham et al. (2014). At every time step, the data set thus consists of 143 chemical concentrations 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 $\mathrm{nlon} \times \mathrm{nlat} \times \mathrm{nlev} \times \mathrm{ntimes} \times \mathrm{nfeatures} = 72 \times 46 \times 30 \times 26\,280 \times 380 = 9.9 \times 10^{11}$. The value of 380 in the feature space breaks down as $143 + 91 + 3 + 143 = 380$, which refers to the chemical species concentration before integration, the photolysis rates, the three meteorological variables, and the tendencies (rate of change) of all species due to chemistry, as specified in the GEOS-Chem simulations (https://geoschem.github.io, last access: 12 June 2025).

2.2 Data preprocessing

Many dimensionality reduction techniques rely on an underlying singular value decomposition of the data that extracts correlated patterns in the data. A fundamental weakness of such SVD-based approaches is the inability to efficiently handle invariances in the data. Specifically, translational and/or rotational invariances of low-rank features in the data are not well captured (Kutz, 2013; Kutz et al., 2016a; Brunton and Kutz, 2019; Velegar et al., 2019). One of the key environmental variables driving the chemistry is the 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. Thus, sunlight activates many of the chemical reactions in the atmospheric chemistry dynamics network. 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 the bottom left in Fig. 2, highlighting the translational invariance in the absolute concentration data. Any SVD-based approach will be unable to capture this translational invariance and correlate across snapshots in time, producing an artificially high dimensionality, i.e., a higher number of modes would be needed to characterize the dynamics due to translation (Kutz, 2013; Brunton and Kutz, 2019). To overcome this issue, the time series for each grid point are shifted to align with GMT time, as shown on the 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 the length of days and nights. For 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 daytime chemistry can be isolated and analysis performed on the isolated daytimes, especially if there is total turn-off of the dynamics at night. The daytime chemistry is isolated, showing only the non-zero data during daytime. We further note that out of the large number of latitude, longitude, and elevation settings, we highlight surface dynamics (elevation $= 1$), as this elevation is not only rich dynamically but also the elevation at which humans are exposed to the atmospheric chemistry dynamics. As will discussed in what follows, we have made judicious choices to demonstrate the dynamics present.

2.3 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 $\mathbf{A}$ (Brunton and Kutz, 2019) that approximately advances the snapshot measurements of the state of a system $x \in \mathbb{R}^n$ forward in time by step size Δt:

$$\mathbf{X}' \approx \mathbf{A}\mathbf{X}, \tag{5}$$

which leads to the mathematical definition of operator $\mathbf{A}$ as the best-fit one-step operator (Tu et al., 2014).

However, the DMD formulated by this regression is rarely used for the 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 (Eq. 5) is extremely sensitive to noise in the data, causing a bias in the computed DMD modes and eigenvalues (Bagheri, 2014; Dawson et al., 2016; Hemati et al., 2017). The *optimized DMD* (optDMD) algorithm of Askham and Kutz (Askham and Kutz, 2018), which uses a variable projection method (Golub and Pereyra, 2003) 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 (Askham and Kutz, 2018). The exponential fitting is given by

$$\operatorname{argmin}_{\omega_k,\phi_k,b_k} \|\mathbf{X} - \sum_{k=1}^{r} b_k \boldsymbol{\phi}_k \exp(\omega_k t)\|_2^2, \tag{6}$$

where a rank r approximation is estimated. As noted, optimized DMD iterates to a solution of this non-convex problem by using variable projection (Golub and Pereyra, 2003). This approach has been shown to provide a superior decomposition due to its ability to optimally suppress noise bias and handle snapshots collected at arbitrary times. Figure 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 generated 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, 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 Eq. (6) to (i) the imaginary axis:

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

or (ii) the closed left half-plane:

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

As discussed below, these constraints further stabilize and make robust the 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 (Golub and Pereyra, 2003), which can often fail to converge.

2.4 Bagging OPtimized Dynamic Mode Decomposition (BOP-DMD)

BOP-DMD (Sashidhar and Kutz, 2022) leverages Breiman's statistical bagging sampling strategy (Breiman et al., 1984) 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 overfitting by design. Not only does ensembling improve DMD, but it is also effective in deep neural network regressions (Allen-Zhu and Li, 2020). Further innovations include stabilizing the variable projection technique used by optDMD so that it converges consistently to an optimal solution (Sashidhar and Kutz, 2022). Its ability to converge is often dependent upon a suitable initial guess for the DMD eigenvalues and eigenvectors.

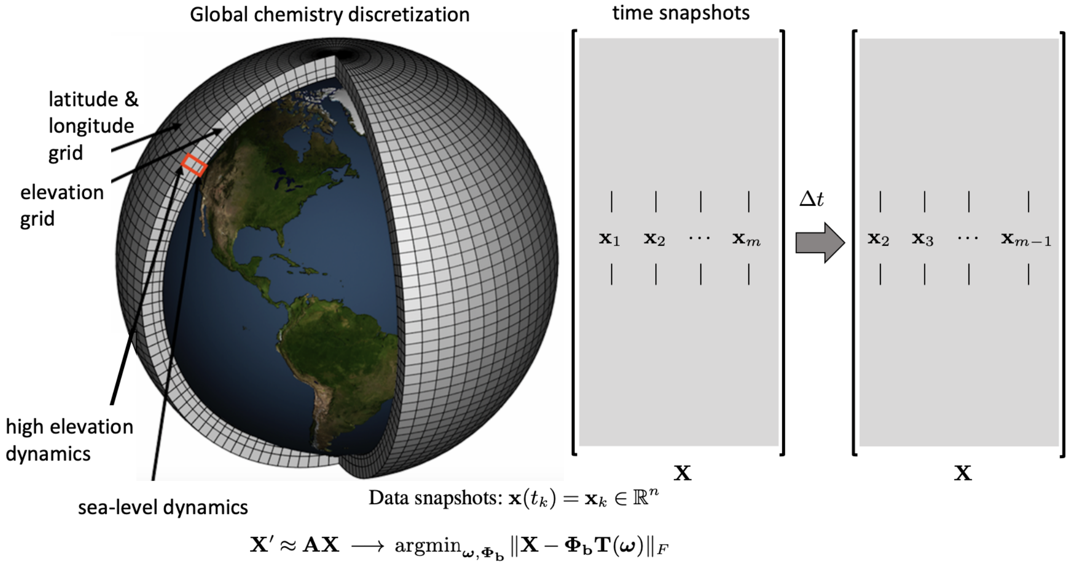
Global chemistry discretization
latitude & longitude grid
elevation grid
high elevation dynamics
sea-level dynamics
time snapshots
$\mathbf{x}_1$ $\mathbf{x}_2$ $\cdots$ $\mathbf{x}_m$
Δt
$\mathbf{x}_2$ $\mathbf{x}_3$ $\cdots$ $\mathbf{x}_{m-1}$
$\mathbf{X}$
$\mathbf{X}$
Data snapshots: $\mathbf{x}(t_k) = \mathbf{x}_k \in \mathbb{R}^n$
$\mathbf{X}' \approx \mathbf{AX} \longrightarrow \text{argmin}_{\boldsymbol{\omega},\boldsymbol{\Phi}_\mathbf{b}} \|\mathbf{X} - \boldsymbol{\Phi}_\mathbf{b}\mathbf{T}(\boldsymbol{\omega})\|_F$

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

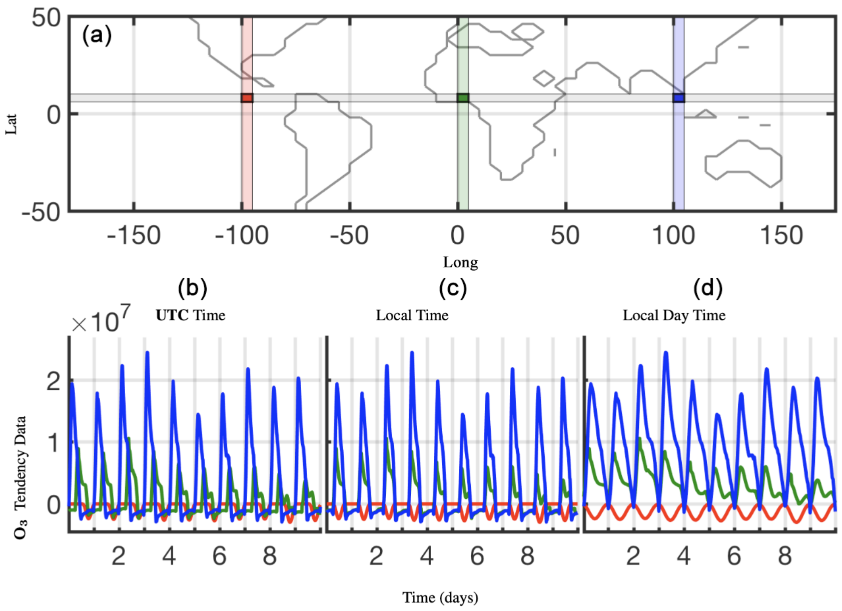

Figure 2. Shifting the data for each cell in time to align the local time zones across a latitude to the prime meridian (long = 0°) local time, shown here for O_3 tendency data for lat = 30°. **(b)** is the raw data for the three highlighted cells, **(c)** is those data shifted in time, and **(d)** shows isolated daytime values only.

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 $\{\langle\boldsymbol{\Phi}\rangle, \langle\boldsymbol{\Omega}\rangle, \langle\boldsymbol{b}\rangle\}$ and variance $\{\langle\Phi^2\rangle, \langle\boldsymbol{\Omega}^2\rangle, \langle\boldsymbol{b}^2\rangle\}$ of the model parameters Φ, $\boldsymbol{\Omega}$, $\boldsymbol{b}$ can now

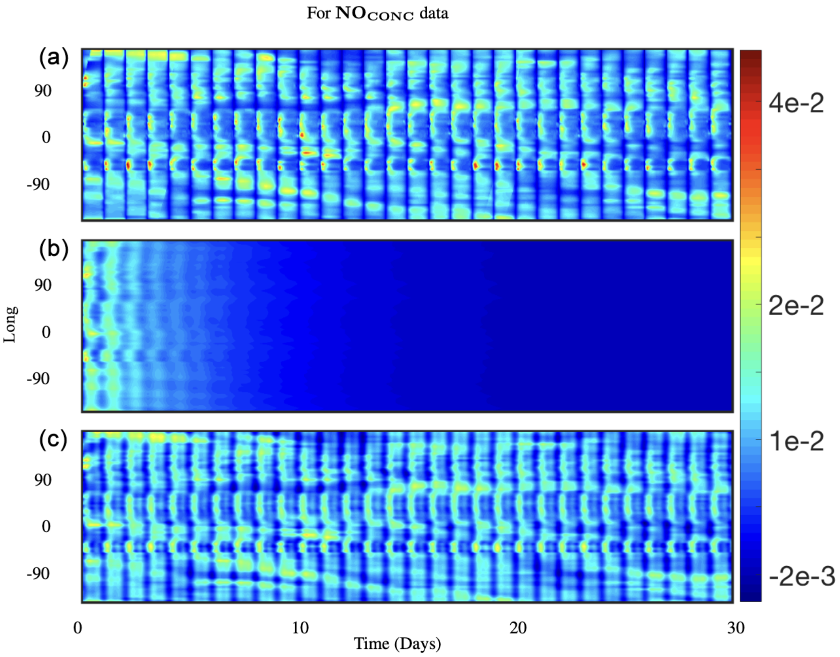

Figure 3. Comparing 30 d reconstruction results for classical and optimized DMD at the surface of NO preprocessed data at lat = 30°. The results are for absolute concentration or CONC data; **(a)** shows the preprocessed data, **(b)** shows the reconstruction from classical DMD, and **(c)** shows the reconstruction from optimized DMD. Classical DMD is unable to capture the dynamics for the absolute concentration data, and it decays down to zero. Optimized DMD reconstructs the data and resolves the dynamics accurately.

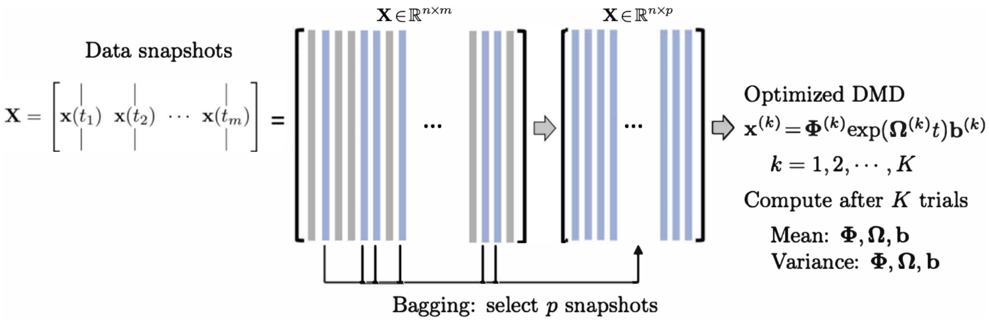
Data snapshots
$\mathbf{X} = \begin{bmatrix} | & | & & | \\ \mathbf{x}(t_1) & \mathbf{x}(t_2) & \cdots & \mathbf{x}(t_m) \\ | & | & & | \end{bmatrix} =$
$\mathbf{X} \in \mathbb{R}^{n\times m}$
$\mathbf{X} \in \mathbb{R}^{n\times p}$
Optimized DMD
$\mathbf{x}^{(k)} = \mathbf{\Phi}^{(k)}\exp(\mathbf{\Omega}^{(k)}t)\mathbf{b}^{(k)}$
$k = 1, 2, \cdots, K$
Compute after K trials
Mean: $\mathbf{\Phi}, \mathbf{\Omega}, \mathbf{b}$
Variance: $\mathbf{\Phi}, \mathbf{\Omega}, \mathbf{b}$
Bagging: select p snapshots

Figure 4. Summary of the BOP-DMD architecture, reproduced with permission from Sashidhar and Kutz (2022). The data snapshots $\boldsymbol{x}(t_k)$ 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 $\boldsymbol{x}^{(k)} = \boldsymbol{\Phi}^{(k)}\exp(\boldsymbol{\Omega}^{(k)}t)\boldsymbol{b}^{(k)}$ is used to compute the statistics (mean and variance) of the DMD parameterizations $\boldsymbol{\Phi}$, $\boldsymbol{\Omega}$, $\boldsymbol{b}$, which are used to build the BOP-DMD ensemble solution with uncertainty quantification (UQ).

be computed. Hence, in addition to producing the DMD model itself, the output of algorithm 1 generates both spatial and temporal uncertainty quantification (UQ) metrics. In this work, we primarily focus on the temporal UQ metrics for forecasting.

3 Results

The analysis is performed for preprocessed or time-shifted raw data for 60 d from 2 July–30 August. 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 the interpretation, the analysis is performed on surface data (elevation = 1), 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; accordingly, the daylight chemistry period is much shorter than the nighttime chemistry period. Thus, the data exhibit a spiky pattern that needs much higher modes to accurately reconstruct it, and/or we would need to isolate the daytime values only when there are active chemical kinetics present. Hence, we choose latitude = 30° N for the analysis, which has the longest daytimes

Algorithm 1 BOP-DMD.

```
Input: Input $(\mathbf{X}, p, K)$
  Procedure: BOPDMD $(\mathbf{X}, p, K)$
  Compute $\mathbf{\Phi}_0, \mathbf{\Omega}_0, \mathbf{b}_0$
  For $k \in \{1, 2, \cdots, K\}$
          Choose $p$ of $m$ snapshots $(p < m)$
          optDMD $\mathbf{\Phi}_k, \mathbf{\Omega}_k, \mathbf{b}_k$ and Initialize with $\mathbf{\Omega}_0$
          Update $\mathbf{\Phi}, \mathbf{\Omega}, \mathbf{b}$ by adding $\mathbf{\Phi}_k, \mathbf{\Omega}_k, \mathbf{b}_k$ to $\mathbf{\Phi}, \mathbf{\Omega}, \mathbf{b}$
  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: $\boldsymbol{\mu}, \boldsymbol{\sigma}$ which are optDMD parameters.
```

for the latitudes considered. The first 40 d of data are 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(long) × 2880(time) for each latitude. optDMD is performed for this data matrix. We perform the analysis for six different chemical species of interest (Velegar et al., 2019): nitric oxide NO, ozone O_3, nitrogen dioxide NO_2, hydroxyl radical OH, isoprene ISOP, and carbon monoxide CO. For each species, we have CONC or absolute concentration data (expressed in units of molec. cm^{-3}) and TEND or tendency/rate of change data (expressed in units of molec. $cm^{-3}\,s^{-1}$). Using the diagnostics from the 40 d training period (2 July–10 August), we then forecast the chemical evolution for the following 20 d (11–30 August). The number of days used for fitting (40 d) is one of two hyper-parameters for the DMD regression, the other being the number of modes (rank) used. A sliding window approach for sampling for DMD has been shown to be quite effective for reconstruction and forecasting (Kutz et al., 2016b; Lapo et al., 2024). Typically, a shorter sampling window helps in forecasting, as the data is often non-stationary and long time histories compromise the DMD model. Thus, we use a fairly representative model history of 40 d for forecasting, which also makes the model smaller and thus easier to manage. In general, this is also in keeping with the DMD philosophy of a model that can be simply run again due to its small computational footprint. Although there are hundreds of chemicals whose dynamics can be demonstrated, the six selected are chemicals commonly associated with atmospheric diagnostics, including pollution and environmental health. Similarly, out of the large number of latitude, longitude, and elevation settings, we highlighted surface dynamics, as these are often some of the richest and most relevant for understanding the role of atmospheric chemistry affecting humans. It is an intractable task to show all chemicals at all locations. Thus, the judicious choices represent those of greatest impact and which are commonly considered by experts in practice. The code provided allows one to consider any chemical at any location desired. There are, of course, limitations to the methodology presented, especially when considering chemical dynamics that are highly intermittent and which lack any periodic or quasi-periodic behavior. Ozone is an example of a chemical that is intermittently active in its dynamics, thus compromising the ability of an algorithm like DMD to produce quality reconstructions and forecasts. Such chemicals have been excluded from consideration, as methods for such time-series behavior are currently lacking.

3.1 DMD diagnostics

The optDMD algorithm decomposes data into time dynamics represented by the spectrum of eigenvalues Ω and the corresponding spatial modes Φ. Here, we present 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 approach is taken to examine which decomposition is best suited for the reconstruction and forecasting of the chemistry dynamics. The constraints are important in practice, especially for forecasting the atmospheric chemistry. Without constraints, and often due to noise, the data can generate eigenvalues that have positive real parts. Even moderate-length forecasts will blow up artificially due to the real part being positive. The optDMD algorithm allows us to remove this unbounded artificial exponential growth. Growth of the solution is still accommodated by modeling it as the first part of an oscillatory solution (which looks like it is growing but is, in reality, an oscillating mode). Similarly, it has already been noted that noise can also artificially bias the eigenvalues towards the left half-plane, which makes solutions decay to zero. Thus, a forecast will exponentially die away to zero. The constraint of the eigenvalue on the imaginary axis guarantees a stable long-term forecast that neither grows nor decays. Of course, this is a pure regression problem, which induces its own limitations, but in regards to forecasting, it has the important and desirable property of stability for long-term forecasting. There is an additional inherent assumption with constraining the eigenvalues to the imaginary axis: conservation of mass of that chemical species. The diagnostics are presented for the 40 d time series of the hydroxyl radical species (OH). The results are consistent for all chemical species of interest. Specifically, the forecasting performance and error are agnostic to the specific chemical species considered, thus suggesting the DMD behavior is independent of the specific chemistry being modeled. 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. These specific target ranks are chosen through hyper-parameter tuning of their forecasting performance. Too few modes compromises the DMD model because there are not enough features to accurately reconstruct and forecast the data. Too many modes causes overfitting to occur on the training data. Thus, although arbitrary, these specific values show generically strong performance across chemical species for the task of forecasting. The diagnostics are presented for both absolute concentration of the chemical species, or OH_{CONC} data, on the left panels and the rate of change of concentrations/-tendencies due to chemistry, or OH_{TEND} data, on the right panels in Figs. 5 and 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 give the actual OH_{CONC} data (left) and actual OH_{TEND} data (right), presented for comparison.

i. The spectrum for optDMD with no constraints on the eigenvalues for OH_{CONC} data is presented in the top-left panel and for OH_{TEND} data is presented in 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.

ii. 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 in the second two panels in Fig. 5. The corresponding reconstruction of data is presented in the third two panels of Fig. 6. Not only does optDMD with these constraints faithfully reconstruct the data, but the forecasting results are also accurate, as presented in the following section.

iii. 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 in 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.

iv. Finally, results from exact DMD for both data sets are presented in the bottom two panels of Figs. 5 and 6. The resulting spectra 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 optDMD, we truncate the number of modes to avoid fitting dynamics to the lowest energy modes, which may cause overfitting and may be corrupted by noise. We would be truncating using *hard-thresholding* at a rank r at which the relative error in the reconstruction has an elbow, i.e., the error graph flattens out without further decrease. Focusing on six key chemicals of interest – NO, O_3, NO_2, OH, ISOP, CO – and the 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 are presented in Fig. 7. A larger number of modes is needed to reconstruct the TEND data compared to the CONC data. Based on the results, we use 20–30 modes for optimal diagnostics of the CONC data, depending on the chemical species. For the TEND data, we pick between 30 and 50 modes.

Finally, we present the global spatial modes for CO and NO, computed at 12 latitudes from $-14°$ through $30°$, in Figs. 8 and 9, respectively. The 12 latitudes are selected for having consistent day lengths across all longitudes and at least four snapshots during daytime. As described above, optDMD is performed 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 inland for the chemical species are represented clearly (Jacob, 1999; Brasseur and Jacob, 2017).

3.2 Forecasting

As described above, using an appropriate rank truncation, the optDMD algorithm with eigenvalues constrained to the closed left half-plane faithfully reconstructs the time-series data for a 40 d training window and a given elevation/latitude. We now forecast the time-series data for future times beyond the training window. Using Eq. (1), with amplitudes b, modes Φ, and eigenvalues Ω computed by optDMD during the training window, we forecast time series for the subsequent 20 d. The results for the CONC and TEND data for two chemical species OH and NO are presented for six longitudes and latitude 30° at the surface (elevation = 1) in Figs. 10, 11, 12, and 13.

Constrained optDMD faithfully reconstructs and forecasts the time series for the 20 d tested. Because we use the fewest modes possible, spikes in the actual data are sometimes not reproduced, and we see a sinusoidal best-fit time series instead. The NO_{TEND} results in Fig. 13 demonstrate this.

We have snapshots of the data every 20 min, 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 six chemical species of interest and for both CONC

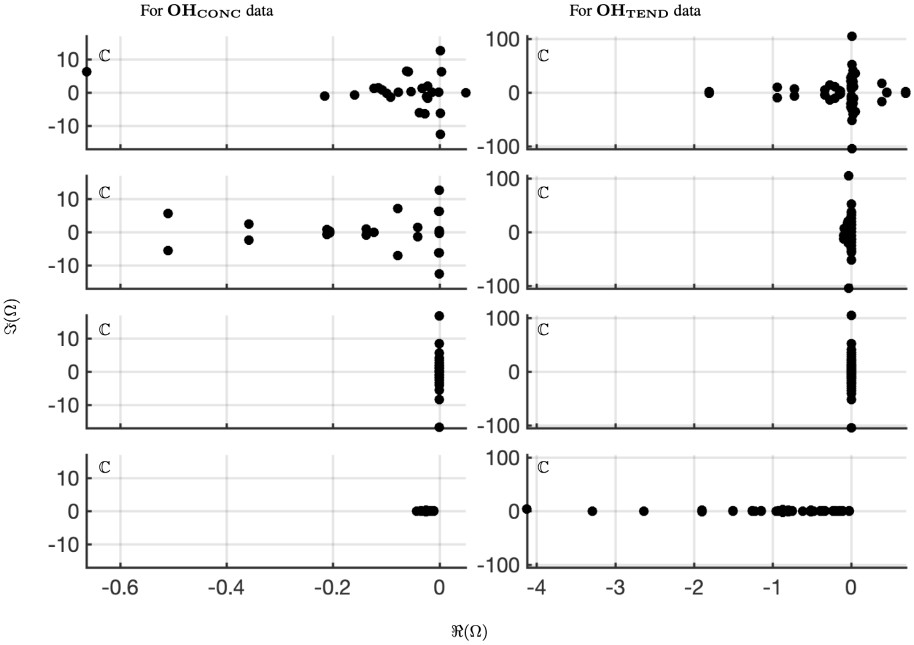

Figure 5. Comparing the spectrum for 40 d reconstruction results for classical and optimized DMD at the surface of OH preprocessed data. On the left four panels are the eigenvalues of OH_{CONC} data; on the right four panels are the eigenvalues of OH_{TEND} at lat = 30°. The top panels show the spectrum from optimized DMD with no constraints, the second set of panels show the spectrum from optimized DMD with linearized constraints requiring that the eigenvalues be on the left half-plane, the third set of panels show the spectrum from optimized DMD with linearized constraints requiring that the eigenvalues be imaginary, and the bottom panels show the spectrum from classical or exact DMD. Note that a hard rank threshold truncation of $r = 25$ for the CONC data and $r = 50$ for the TEND data has been used.

and TEND data in Fig. 14, colored in 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 d 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 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 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 an improvement in forecasting of the TEND data.

The optDMD algorithm performs worst in forecasting the chemical species OH. OH has a very short tropospheric lifetime 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).

3.3 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 the 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 d (2 July–30 August) to create an initial seed Φ_0, Ω_0, b_0 for the BOP-DMD algorithm. For $K = 100$ trials, we randomly select $p = 216$ snapshots/columns, i.e., data for 3 d out of the 60 d, to create our subset of data, as shown in Fig. 4. optDMD now computes the eigenvalues of various subsets using the aforementioned initial conditions. The eigenvalues for the $K = 100$ ensemble models are used to produce the temporal UQ metrics. The UQ metrics are critical for understanding the ability of the BOP-DMD algorithm to perform long-term forecasting. Specifically, BOP-DMD is a low-cost computational tool, as opposed to Monte Carlo simulations, for evaluating the divergence of future state predictions from an ensemble of predictions, specifically drawn from the BOP-DMD eigenvalue distribution.

Figure 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 algorithm quantifies the temporal uncertainty by

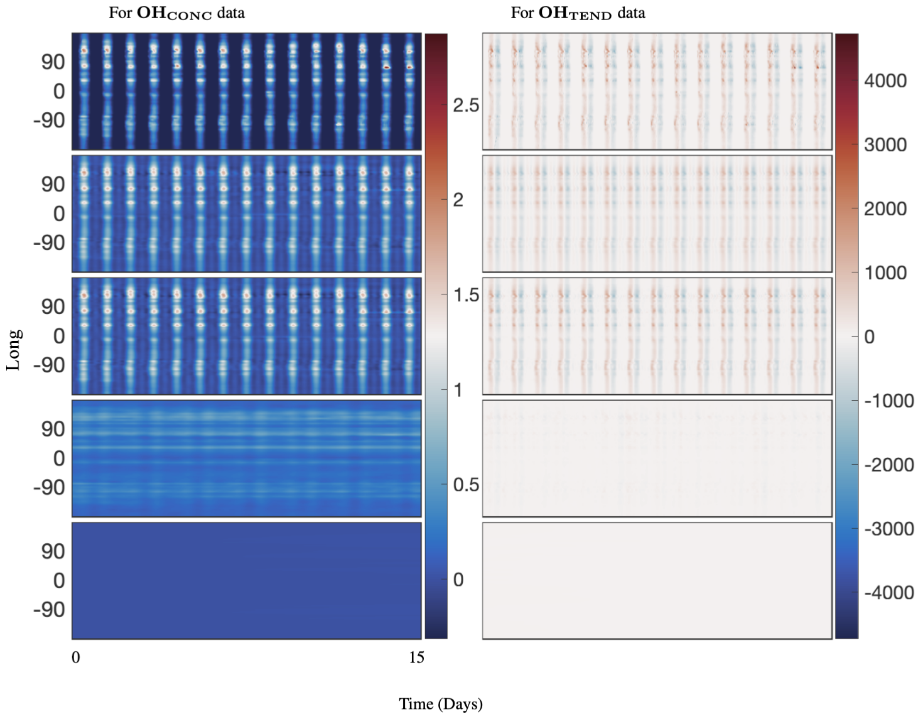

Figure 6. Comparing 40 d reconstruction results for classical DMD, optimized DMD, and optimized DMD with no constraints at the surface of OH preprocessed data at lat = 30°. The left panel is for absolute concentration or CONC data, and the right panel is for Tendency data. The top panels show the preprocessed 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 classic DMD. Classic DMD is unable to reconstruct the dynamics for the absolute concentration and tendency data. Note that a hard rank threshold truncation of $r = 25$ for the CONC data and $r = 50$ for the TEND data has been used.

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 $\mathrm{OH}_{\mathrm{CONC}}$ data and for all modes of the $\mathrm{OH}_{\mathrm{TEND}}$ data. Then, we trim the eigenvalue distribution data to exclude the outliers below the 10th percentile and above the 90th percentile to improve the UQ metrics. Figure 16 shows the distributions of the trimmed absolute eigenvalues, where the Gaussian fit is clearly better with lower variances, and only 1 distribution with outliers. Still, we see that there is significant temporal variability, especially for higher modes for $\mathrm{OH}_{\mathrm{TEND}}$.

4 Discussion

Based on the results presented in this work, we conclude that the constrained optDMD algorithm 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 (Askham and Kutz, 2018), hence providing an optimal de-biasing for the atmospheric chemistry dynamics. optDMD is thus better able to capture hidden dynamics, showing an order of magnitude improvement in the reconstruction error. optDMD also produces modes that more accurately describe the localized energetic convective phenomena in the CONC and especially the TEND chemistry dynamics. The nonlinear optimization problem in 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 optDMD requires solving a nonlinear, non-convex optimization problem, which often fails to converge to a solution. The computational cost of 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 BOP-DMD

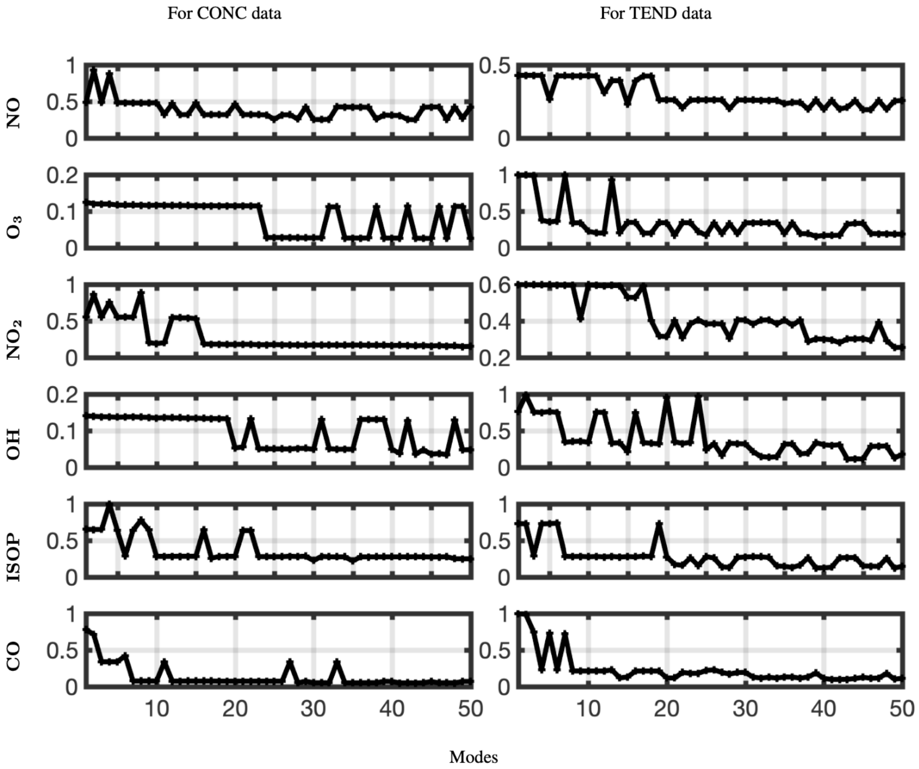

Figure 7. Relative error plotted against the number of modes used for optimized DMD with the eigenvalues constrained to the left half-plane for six different chemical species and the CONC and TEND data at latitude = 30°.

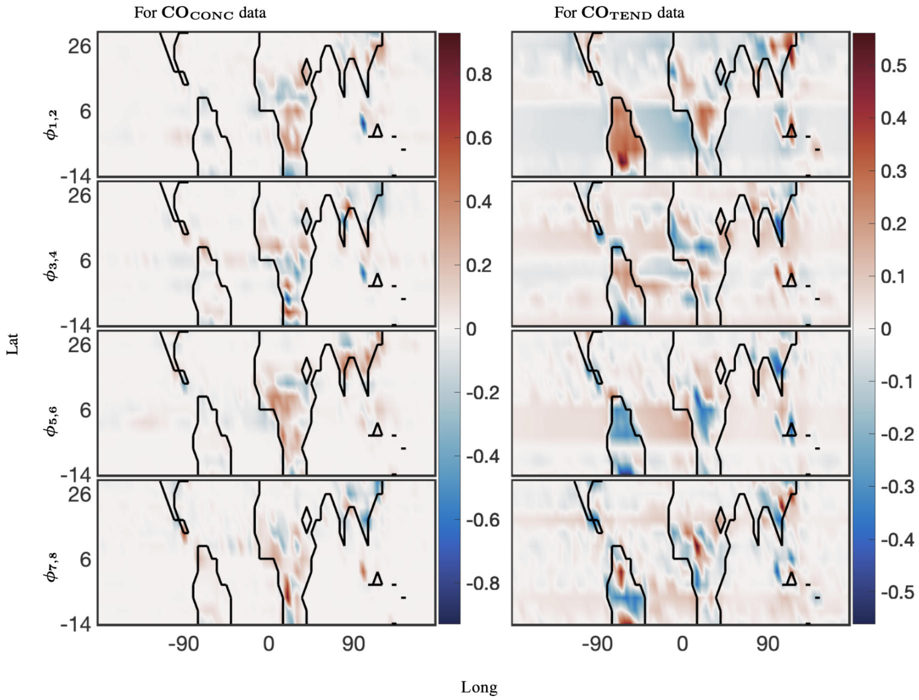

Figure 8. The 40 d reconstruction results for optimized DMD at the surface of CO preprocessed data. The analysis was computed for 12 latitudes from −14° through 30°. The left panel shows the dominant four spatial modes for the CONC data, and the right panel shows four of the corresponding spatial modes for the TEND data. The complex conjugate pair of DMD modes are denoted by $\phi_{i,j}$, where, for the pairing, $j = i + 1$. Thus, ω_1 and ω_2 are the complex conjugate pairs whose real parts are identical.

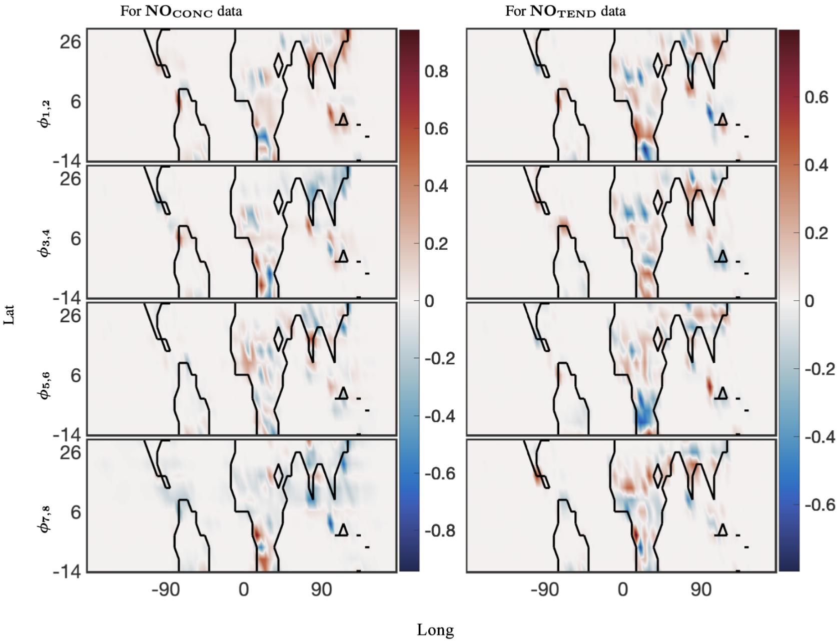

Figure 9. The 40 d reconstruction results for optimized DMD at the surface of NO preprocessed data. The analysis was computed for 12 latitudes from −14° through 30°. The left panel shows four spatial modes for the CONC data, and the right panel shows four of the corresponding spatial modes for the TEND data. The complex conjugate pair of DMD modes is denoted by $\phi_{i,j}$, where, for the pairing, $j = i + 1$. Thus, ω_1 and ω_2 are the complex conjugate pairs whose real parts are identical.

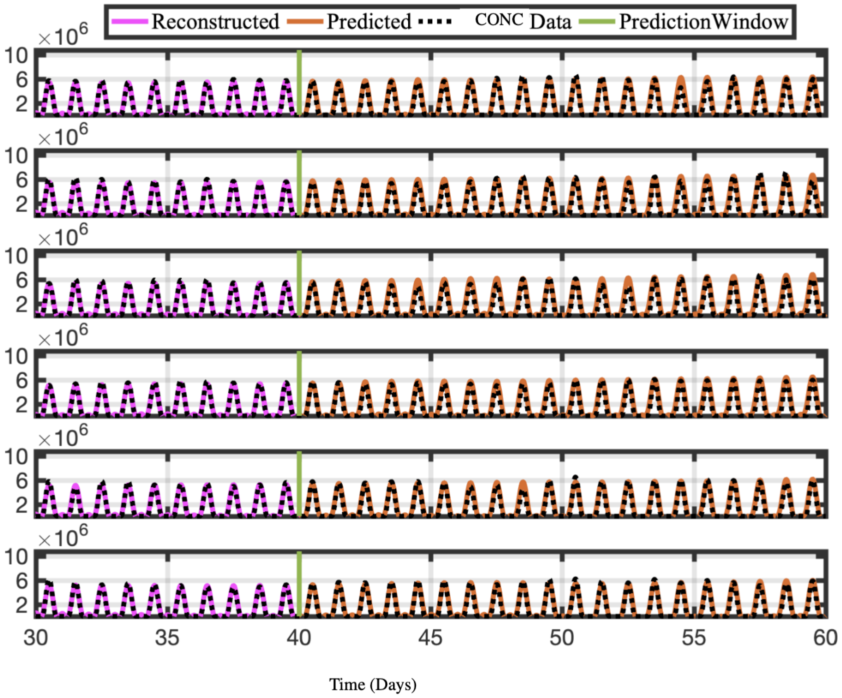

Figure 10. Time series of reconstructed and predicted results with OH_{CONC} data at lat = 30° and six longitudes $-180° : 5° : -155°$. Both the reconstructed data, shown here for 10 d, and the forecasted time series, shown here for the 20 d testing period, faithfully reconstruct and forecast the actual data for OH_{CONC}.

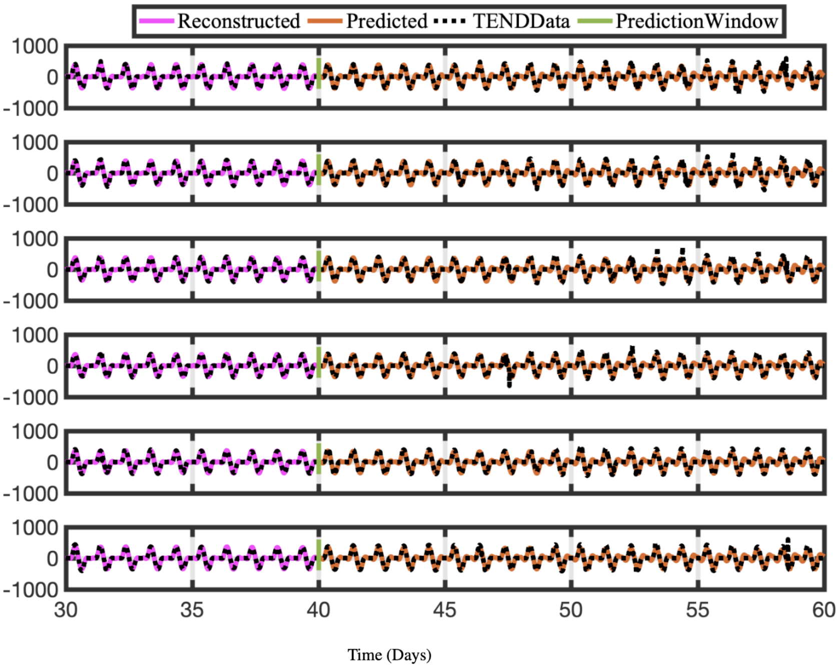

Figure 11. Time series of reconstructed and predicted results with OH_{TEND} data at lat = 30° and six longitudes −180° : 5° : −155°. Again, both the reconstructed data, shown here for 10 d, and the forecasted time series, shown here for the 20 d testing period, faithfully reconstruct and forecast the actual data for OH_{TEND}.

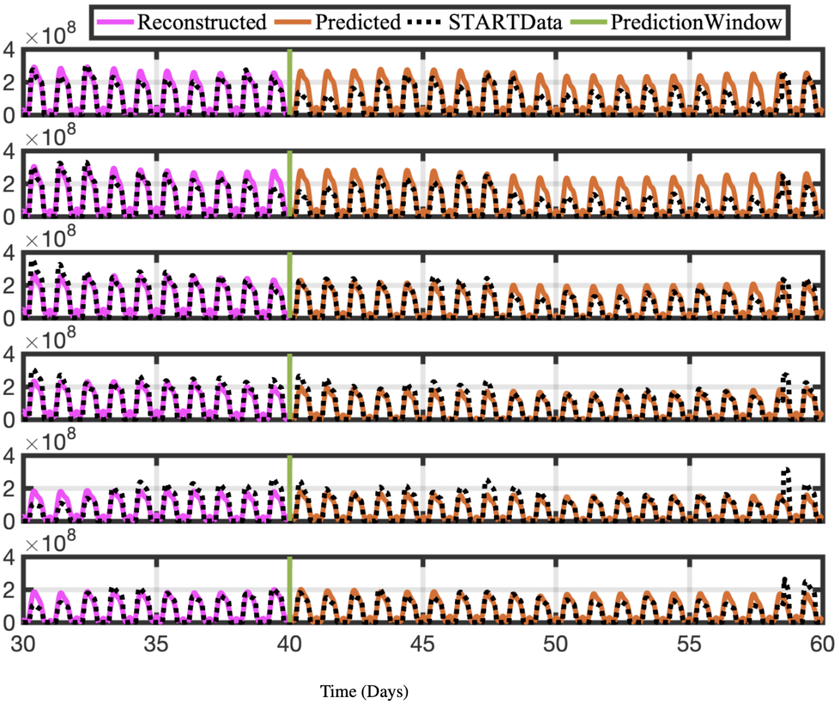

Figure 12. Time series of reconstructed and predicted results with NO_{CONC} data at lat = 30° and six longitudes −180° : 5° : −155°. Both the reconstructed data, shown here for 10 d, and the forecasted time series, shown here for the 20 d testing period, reproduce the actual data for NO_{CONC} well.

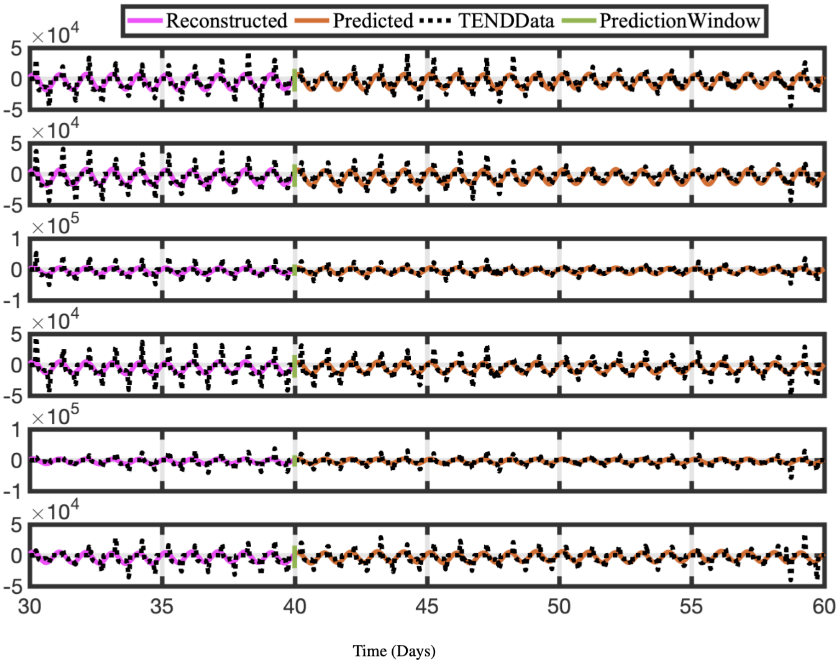

Figure 13. Time series of reconstructed and predicted results with NO_{TEND} data at lat = 30° and six longitudes −180° : 5° : −155°. Neither the reconstructed data, shown here for 10 d, nor the forecasted time series, shown here for the 20 d testing period, capture the spikes in the actual data for NO_{TEND}. Because we are using only 20–30 modes for reconstruction, we get a sinusoidal best fit. In general, spikes in time-series data are difficult to capture and forecast with any method, including DMD. Although more modes can provide a better reconstruction, it often is then overfit on training data for forecasting purposes.

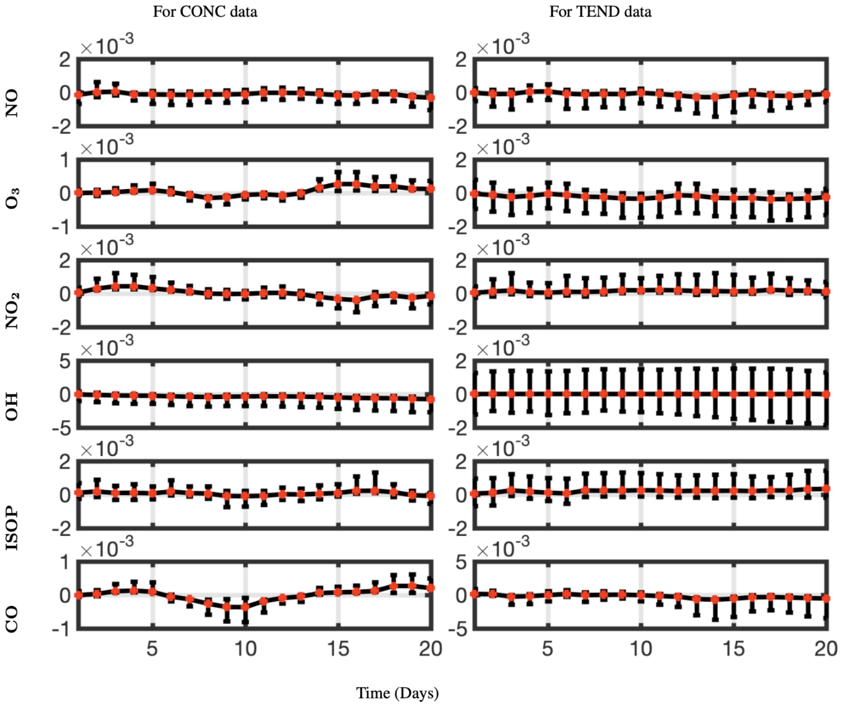

Figure 14. Mean relative error with 95-percentile confidence intervals forecasting CONC *and* TEND data at lat = 30° for a prediction window of 20 d and for six different chemical species. The relative error stays nearly the same or changes only slightly as the number of days we are forecasting out to increases. optDMD does better at forecasting the CONC data than for the TEND data.

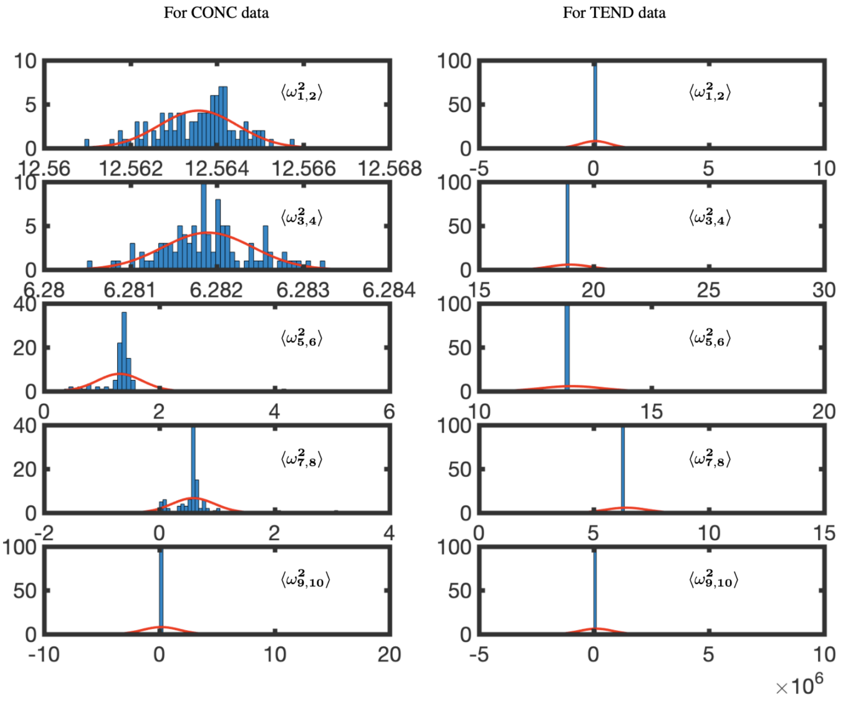

Figure 15. Temporal uncertainty quantification for absolute eigenvalues for the OH_{CONC} and OH_{TEND} data at lat = 30°. The red lines represent a least-squares fit of a normal distribution; 60 d of training data was used, with a sample size of 3 d and 100 cycles. The complex conjugate pair frequencies are denoted by $\langle \omega^2_{i,j} \rangle$, where, for the pairing, $j = i + 1$. Thus, ω_1 and ω_2 are the complex conjugate pairs whose variance is evaluated jointly.

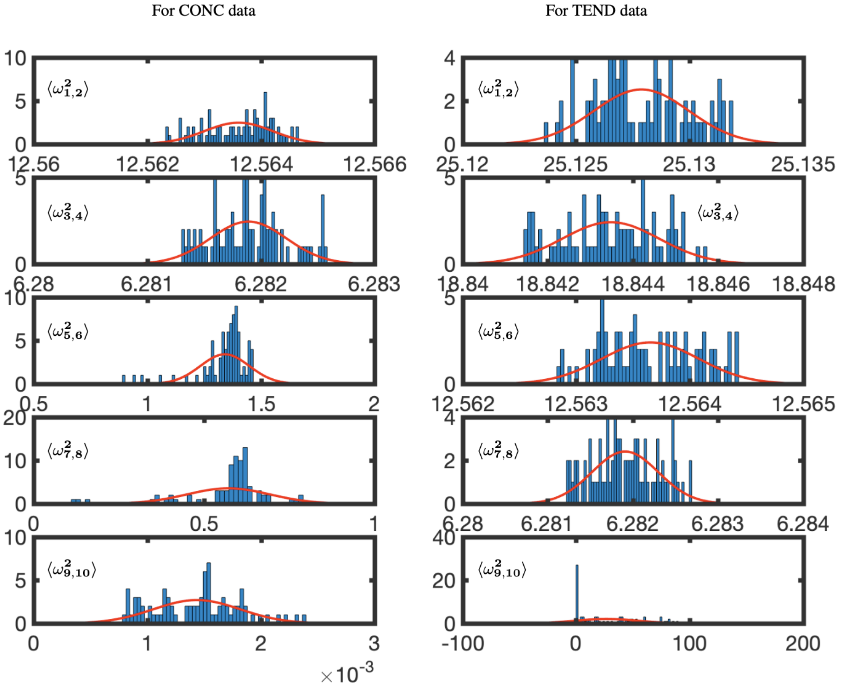

Figure 16. Temporal uncertainty quantification for absolute trimmed eigenvalues for the OH_{CONC} and OH_{TEND} data at lat = 30°. The data have been trimmed to remove outliers below the 10th percentile and above the 90th percentile. The red lines represent a least-squares fit of a normal distribution. The complex conjugate pair frequencies are denoted by $\langle \omega^2_{i,j} \rangle$, where, for the pairing, $j = i + 1$. Thus, ω_1 and ω_2 are the complex conjugate pairs whose variance is evaluated jointly.

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 optDMD to an entire years 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 algorithm 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 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 a spatio-temporal atmospheric chemistry system, with Monte Carlo simulations to produce UQ for future states.

The 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 emission 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.

Code and data availability. The code is openly available at the following GitHub link: https://github.com/mvelegar/DMDPaper (last access: 13 November 2023). The code and data are available on Zenodo: https://doi.org/10.5281/zenodo.12754943 (Velaghar and Kutz, 2024).

Author contributions. Conceptualization, JNK and MV; methodology, MV and JNK; software, MV; validation, MV, CK, and JNK; formal analysis, MV, CK, and JNK; resources, CK and JNK; data curation, CK and MV; writing – original draft preparation, MV, CK, and JNK; writing – review and editing, MV, CK, and JNK; visualization, MV; supervision, JNK and CK; funding acquisition, JNK. All authors have read and agreed to the published version of the paper.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Financial support. The authors received support from the National Science Foundation AI Institute in Dynamic Systems (grant no. 2112085). J. Nathan Kutz further received support from the Air Force Office of Scientific Research (grant no. FA9550-19-1-0011).

Review statement. This paper was edited by Patrick Jöckel and reviewed by Narendra Ojha and one anonymous referee.

References

Alla, A. and Kutz, J. N.: Nonlinear Model Order Reduction via Dynamic Mode Decomposition, SIAM J. Sci. Comput., 39, B778–B796, https://doi.org/10.1137/16M1059308, 2017.

Allen-Zhu, Z. and Li, Y.: Towards Understanding Ensemble, Knowledge Distillation and Self-Distillation in Deep Learning, arXiv [preprint], https://doi.org/10.48550/arXiv.2012.09816, 2020.

Antoulas, A. C.: Approximation of Large-Scale Dynamical Systems, Society for Industrial and Applied Mathematics, https://doi.org/10.1137/1.9780898718713, 2005.

Askham, T. and Kutz, J. N.: Variable projection methods for an optimized dynamic mode decomposition, SIAM J. Appl. Dyn. Syst., 17, 380–416, 2018.

Bagheri, S.: Effects of weak noise on oscillating flows: Linking quality factor, Floquet modes, and Koopman spectrum, Phys. Fluids, 26, 094104, https://doi.org/10.1063/1.4895898, 2014.

Benner, P., Gugercin, S., and Willcox, K.: A Survey of Projection-Based Model Reduction Methods for Parametric Dynamical Systems, SIAM Rev., 57, 483–531, https://doi.org/10.1137/130932715, 2015.

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

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

Brasseur, G. P. and Jacob, D. J.: Modeling of Atmospheric Chemistry, Cambridge University Press, ISBN 9781107146969, 2017.

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J.: Classification and Regression Trees, Chapman and Hall/CRC, https://doi.org/10.1201/9781315139470, 1984.

Brunton, S. L. and Kutz, J. N.: Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control, Cambridge University Press, https://doi.org/10.1017/9781108380690, 2019.

Brunton, S. L., Proctor, J. L., Tu, J. H., and Kutz, J. N.: Compressed sensing and dynamic mode decomposition, Journal of Computational Dynamics, 2, 165–191, https://doi.org/10.3934/jcd.2015002, 2015.

Carlberg, K., Barone, M., and Antil, H.: Galerkin v. least-squares Petrov–Galerkin projection in nonlinear model reduction, J. Comput. Phys., 330, 693–734, 2017.

Chen, K. K., Tu, J. H., and Rowley, C. W.: Variants of Dynamic Mode Decomposition: Boundary Condition, Koopman, and Fourier Analyses, J. Nonlinear Sci., 22, 887–915, 2012.

Dawson, S. T. M., Hemati, M. S., Williams, M. O., and Rowley, C. W.: Characterizing and correcting for the effect of sensor noise in the dynamic mode decomposition, Exp. Fluids, 57, 42, https://doi.org/10.1007/s00348-016-2127-7, 2016.

Deem, E. A., Cattafesta, L. N., Hemati, M. S., Zhang, H., Rowley, C., and Mittal, R.: Adaptive separation control of a laminar boundary layer using online dynamic mode decomposition, J. Fluid Mech., 2020, 903, https://doi.org/10.1017/jfm.2020.546, 2020.

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

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

Eiximeno, B., Miró, A., Begiashvili, B., Valero, E., Rodriguez, I., and Lehmkhul, O.: pyLOM: A HPC open source reduced order model suite for fluid dynamics applications, Comput. Phys. Commun., 308, 109459, https://doi.org/10.1016/j.cpc.2024.109459, 2025.

Erichson, N. B., Voronin, S., Brunton, S. L., and Kutz, J. N.: Randomized matrix decompositions using R, J. Stat. Softw., 89, 1–48, 2019.

Gin, C., Lusch, B., Brunton, S. L., and Kutz, J. N.: Deep learning models for global coordinate transformations that linearise PDEs, Eur. J. Appl. Math., 32, 515–539, 2021.

Golub, G. and Pereyra, V.: Separable nonlinear least squares: the variable projection method and its applications, Inverse Probl., 19, 2, https://doi.org/10.1088/0266-5611/19/2/201, 2003.

Hemati, M. S., Rowley, C. W., Deem, E. A., and Cattafesta, L. N.: De-biasing the dynamic mode decomposition for applied Koopman spectral analysis of noisy datasets, Theor. Comp. Fluid Dyn., 31, 349–368, 2017.

Hesthaven, J., Rozza, G., and Stamm, B.: Certified Reduced Basis Methods for Parametrized Partial Differential Equations, ISBN 978-3-319-22470-1, https://doi.org/10.1007/978-3-319-22470-1, 2016.

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

Jacob, D. J.: Introduction to atmospheric chemistry, Princeton University Press, ISBN 0691001855, 1999.

Kutz, J. N.: Data-driven modeling & scientific computation: methods for complex systems & big data, Oxford University Press, ISBN 0199660344, 2013.

Kutz, J. N., Brunton, S. L., Brunton, B. W., and Proctor, J. L.: Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems, SIAM – Society for Industrial and Applied Mathematics, USA, ISBN 9781611974492, 2016a.

Kutz, J. N., Fu, X., and Brunton, S. L.: Multiresolution Dynamic Mode Decomposition, SIAM J. Appl. Dyn. Syst., 15, 713–735, https://doi.org/10.1137/15M1023543, 2016b.

Lange, H., Brunton, S. L., and Kutz, J. N.: From Fourier to Koopman: Spectral Methods for Long-term Time Series Prediction, CoRR, arXiv [preprint], https://doi.org/10.48550/arXiv.2004.00574, 2020.

Lapo, K., Ichinaga, S. M., and Kutz J. N.: A method for unsupervised learning of coherent spatiotemporal patterns in multiscale data, P. Natl. Acad. Sci. USA, 122, e2415786122, https://doi.org/10.1073/pnas.2415786122, 2024.

Liu, Y., Sid-Lakhdar, W., Rebrova, E., Ghysels, P., and Li, X. S.: A parallel hierarchical blocked adaptive cross approximation algorithm, Int. J. High Perform. C., 34, 394–408, 2020.

Liu, Y., Ponce, C., Brunton, S. L., and Kutz, J. N.: Multiresolution convolutional autoencoders, J. Comput. Phys., 474, 111801, https://doi.org/10.1016/j.jcp.2022.111801, 2023.

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

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

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

Murray, L. T., Jacob, D. J., Logan, J. A., Hudman, R. C., and Koshak, W. J.: Optimized regional and interannual variability of lightning in a global chemical transport model constrained by LIS/OTD satellite data, J. Geophys. Res.-Atmos., 117, D20307, https://doi.org/10.1029/2012JD017934, d20307, 2012.

Parish, E. and Carlberg, K.: Time-series machine-learning error models for approximate solutions to parameterized dynamical systems, Comput. Method. Appl. M., 365, 112990, https://doi.org/10.1016/j.cma.2020.112990, 2020.

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

Proctor, J. L., Brunton, S. L., and Kutz, J. N.: Dynamic Mode Decomposition with Control, SIAM J. Appl. Dyn. Syst., 15, 142–161, https://doi.org/10.1137/15M1013857, 2016.

Qin, T., Wu, K., and Xiu, D.: Data driven governing equations approximation using deep neural networks, J. Comput. Phys., 395, 620–635, https://doi.org/10.1016/j.jcp.2019.06.042, 2019.

Quarteroni, A., Manzoni, A., and Negri, F.: Reduced basis methods for partial differential equations: An introduction, Springer, ISBN 978-3-319-15430-5, https://doi.org/10.1007/978-3-319-15431-2, 2015.

Regazzoni, F., Chapelle, D., and Moireau, P.: Combining data assimilation and machine learning to build data-driven models for unknown long time dynamics – Applications in cardiovascular modeling, Int. J. Numer. Meth. Bio., 37, e3471, https://doi.org/10.1002/cnm.3471, 2021.

Rowley, C., Mezic, I., Bagheri, S., Schlatter, P., and Henningson, D.: Spectral analysis of nonlinear flows, J. Fluid Mech., 641, 115–127, https://doi.org/10.1017/S0022112009992059, 2009.

Sashidhar, D. and Kutz, J. N.: Bagging, optimized dynamic mode decomposition for robust, stable forecasting with spatial and temporal uncertainty quantification, Philos. T. R. Soc. A, 380, 20210199, https://doi.org/10.1098/rsta.2021.0199, 2022.

Schmid, P. J.: Dynamic mode decomposition of numerical and experimental data, J. Fluid Mech., 656, 5–28, https://doi.org/10.1017/S0022112010001217, 2010.

Tu, J. H., Rowley, C. W., Luchtenburg, D. M., Brunton, S. L., and Kutz, J. N.: On dynamic mode decomposition: Theory and applications, Journal of Computational Dynamics, 1, 391–421, https://doi.org/10.3934/jcd.2014.1.391, 2014.

Velaghar, M. and Kutz, J. N.: Dynamic Mode Decomposition Data and Code for Atmospheric Chemistry, Zenodo [data set], https://doi.org/10.5281/zenodo.12754943, 2024.

Velegar, M., Erichson, N. B., Keller, C. A., and Kutz, J. N.: Scalable diagnostics for global atmospheric chemistry using Ristretto library (version 1.0), Geosci. Model Dev., 12, 1525–1539, https://doi.org/10.5194/gmd-12-1525-2019, 2019.