Linear Inverse Modeling of Large-Scale Atmospheric Flow Using Optimal Mode Decomposition

FRANK KWASNIOK[a]

[a] *Department of Mathematics, University of Exeter, Exeter, United Kingdom*

(Manuscript received 14 July 2021, in final form 14 April 2022)

ABSTRACT: Linear inverse modeling or principal oscillation pattern (POP) analysis is a widely applied tool in climate science for extracting from data dominant spatial patterns together with their dynamics as approximated by a linear Markov model. The system is projected onto a principal linear subspace and the system matrix is estimated from data. The eigenmodes of the system matrix are the POPs, with the eigenvalues providing their decay time scales and oscillation frequencies. Usually, the subspace is spanned by the leading principal components (PCs) and empirical orthogonal functions (EOFs). Outside of climate science, this procedure is now more commonly referred to as dynamic mode decomposition (DMD). Here, we use optimal mode decomposition (OMD) to address the full linear inverse modeling problem of simultaneous optimization of the principal subspace and the linear operator. The method is illustrated on two pedagogical examples and then applied to a three-level quasigeostrophic atmospheric model with realistic mean state and variability. The OMD models significantly outperform the EOF/DMD models in predicting the time evolution of the large-scale flow modes. The advantage of the OMD models stems from finding more persistent modes as well as from better capturing the nonnormality of the linear operator and the associated nonmodal growth. The dynamics of the large-scale flow modes turn out to be markedly non-Markovian and the OMD modes are superior to the EOF/DMD modes also in a modeling setting with a higher-order vector autoregressive process. The OMD modes could also be used as basis functions for a nonlinear dynamical model although they are not optimized for that purpose. Potential applications of the OMD method in weather and climate science include ENSO or MJO prediction, reduced-rank data assimilation, and generation of initial perturbations for ensemble prediction.

KEYWORDS: Atmospheric circulation; Optimization; Pattern detection; Statistical techniques; Time series; Statistical forecasting

1. Introduction

Despite the obvious nonlinearity of the underlying governing equations of atmospheric, oceanic, and climate dynamics, the time evolution of anomalies, that is, departures from a background or reference state, is often well described by linear dynamics. Classically, this refers to the evolution of small perturbations in normal mode analysis or small-amplitude errors in weather forecasting. But also for full-amplitude anomalies around the time mean state, possibly coarse-grained via spatial and/or temporal filtering or averaging, the framework of stochastically forced linear dynamics is often a surprisingly good approximation. In linear inverse modeling (LIM) or principal oscillation pattern (POP) analysis or empirical normal mode (ENM) analysis or dynamic mode decomposition (DMD) (Hasselmann 1988; Penland 1989; von Storch et al. 1995; Schmid 2010) a high-dimensional spatiotemporal dataset is projected onto a principal subspace and the system matrix or linear operator is inferred from the data. The system matrix then does not only correspond to the linearized equations of motion but also contains linear parameterizations in terms of the coarse-grained anomalies of (i) nonlinear interactions among the resolved anomalies and (ii) fluxes due to scales and processes not resolved by the reduced model. The eigenvectors and eigenvalues of the linear operator yield the dominant spatial structures of the system with their characteristic damping time scales and oscillation periods. Also the covariance matrix of the noise, capturing its amplitude and spatial structure can be estimated from the data. Linear inverse modeling has been successfully applied to, for example, tropical sea surface temperatures (Xu and von Storch 1990; Penland and Magorian 1993; Penland and Sardeshmukh 1995), and the extratropical large-scale atmospheric circulation (Penland and Ghil 1993; DelSole 1996; Winkler et al. 2001).

POP analysis was extended to a cyclostationary setting by including the annual cycle in the system matrix (Blumenthal 1991) and to a nonstationary setting by introducing trend functions in the system matrix in order to detect, anticipate and predict critical transitions from data (Kwasniok 2018). A linear model driven by additive Gaussian noise is limited to Gaussian statistics; when considering multiplicative noise and in particular correlated additive and multiplicative (CAM) noise the realism of the model is greatly enhanced and also symmetric and asymmetric non-Gaussian statistics can be captured (Sura et al. 2005; Sardeshmukh and Sura 2009).

The principal subspace used for linear inverse modeling is usually spanned by the leading principal components (PCs) associated with the leading empirical orthogonal functions (EOFs) (Jolliffe 2002). This is a canonical choice due to the simplicity of computation and the immediate interpretability of the patterns in terms of their explained variance. There is also a theoretical justification: it follows from the fluctuation–dissipation theorem (Kubo 1966) that under uniform or unbiased stochastic forcing the least damped, that is, the most predictable modes are those with the largest variance. If there are nearly

Corresponding author: Frank Kwasniok, f.kwasniok@exeter.ac.uk

DOI: 10.1175/JAS-D-21-0193.1

neutral modes in the system this relationship is very strong and those modes dominate the variance even in the presence of moderate biases in the stochastic forcing and/or observational error in the data.

The situation may be different if interest is in more than just the most predictable modes or in a system with generally low predictability. For modes with significant damping the space of the leading EOFs becomes entrained with unpredictable noise. A linear inverse model based on EOF truncation can then be expected to underestimate the predictability in the system and also to exhibit errors in the identification of the spatial patterns of the eigenmodes. The issue becomes even more severe in the presence of nonuniform stochastic forcing.

Another setting in which an EOF truncation of the linear dynamics is inefficient occurs if the linear operator of the system is strongly nonnormal. This is generically the case in shear turbulence (Farrell and Ioannou 1996; Schmid 2007). Then significant transient nonmodal growth accounting for most of the variance of the system may be observed which is characterized by the singular values of the finite-time propagator matrix and pairs of patterns, the optimal excitation patterns or stochastic optimals given by the right singular vectors, and the optimal response patterns given by the left singular vectors (Farrell 1989; Farrell and Ioannou 1993; Penland and Sardeshmukh 1995; Farrell and Ioannou 1996; Trefethen and Embree 2005; Schmid 2007). While the optimal response patterns are well captured in an EOF truncation the optimal excitation patterns are typically small-scale, low-variance patterns, structurally very different from the optimal response patterns.

A method for truncating a linear operator which includes both the excitation and response modes in a well-defined manner and also comes with an estimate of the truncation error is offered by balanced truncation (Moore 1981; Glover 1984). However, there are the very restrictive assumptions that the dynamics of the full system are purely linear and that the full system matrix is known. Balanced truncation was implemented by Farrell and Ioannou (2001a) for a prototypical fluid shear instability under uniform stochastic forcing and shown to be far superior to an EOF truncation. The algorithm can readily be extended to a noise with arbitrary covariance structure but then the noise covariance matrix would also need to be known a priori.

The actual importance of nonmodal linear growth in a nonlinear system is somewhat unclear. It may happen that certain optimal excitation patterns do not lie on the attractor of the nonlinear dynamics and thus the associated transient growth is never realized. A study by DelSole (2007), though, on quasigeostrophic turbulence using the linear operator obtained from linearizing the governing equations around the time mean state showed that the optimal excitation patterns occur sufficiently strongly and frequently in the nonlinear model to account for the energy-containing eddies and that only a small number of evolved singular vectors are needed to explain the dominant eddy structure of the model.

Two similar approaches to deriving an optimized subspace for linear inverse modeling called principal dynamical components (PDCs) (de la Iglesia and Tabak 2013) and optimal mode decomposition (OMD) (Wynn et al. 2013) were proposed. They minimize a one-step prediction error under a linear model simultaneously with respect to the basis functions and the system matrix. Moreover, the linear dynamical mode (LDM) decomposition (Gavrilov et al. 2019) maximizes a likelihood under a diagonal linear model in a Bayesian framework. The LDMs were subsequently used in a nonlinear modeling setting with artificial neural networks for ENSO prediction. The LDM technique is a special case of a more general nonlinear dynamical dimension reduction methodology (Gavrilov et al. 2016). In principal interaction pattern (PIP) analysis (Hasselmann 1988; Kwasniok 1996, 1997, 2004, 2007), a subspace is optimized for a nonlinear reduced model. The model may either be given by a projection of known governing equations onto the subspace or may be data driven or a mixture of both.

The present paper investigates the use of OMD, here extended beyond one-step prediction, for linear inverse modeling of large-scale atmospheric flow in a quasigeostrophic three-level model with realistic mean state and variability. The study focuses on the choice of basis functions, that is, EOF/DMD versus OMD models, on Markovian versus non-Markovian dynamics, and on modal versus nonmodal predictability.

The remainder of the paper is organized as follows: In section 2, the methodology of linear inverse modeling is briefly recapitulated and the OMD is introduced. The advantages of the OMD over the traditional EOF/DMD truncation are illustrated in section 3 on two simple pedagogical examples. Section 4 discusses the main application of the OMD to an intermediate-complexity atmospheric model. Some conclusions are drawn in section 5.

2. Methodology

Linear inverse modeling or POP analysis (Hasselmann 1988; Penland 1989; von Storch et al. 1995) is described and OMD is introduced as an extension of the standard theory.

a. *LIM/POP analysis*

The system under consideration is described by a D-dimensional state vector $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$, which may reflect only partial observation of the system as well as spatial and/or temporal filtering or averaging. An equally sampled dataset of length N, $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, with sampling interval δt is given where $\mathbf{x}_n = \mathbf{x}(t_n)$; $n = 1, \ldots, N$. We here use anomalies

$$\mathbf{y} = \mathbf{x} - \langle \mathbf{x} \rangle, \tag{1}$$

with $\langle \mathbf{x} \rangle$ being the time mean state of the system given by

$$\langle \mathbf{x} \rangle = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n. \tag{2}$$

The anomaly dataset is $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ with $\mathbf{y}_n = \mathbf{x}_n - \langle \mathbf{x} \rangle$; $n = 1, \ldots, N$. The covariance matrix of the system is given as

$$\mathbf{C} = \langle \mathbf{y}\mathbf{y}^{\mathrm{T}} \rangle = \frac{1}{N} \sum_{n=1}^{N} \mathbf{y}_n \mathbf{y}_n^{\mathrm{T}}. \tag{3}$$

We consider a J-dimensional linear subspace ($J < D$) spanned by the orthonormal set of vectors $\{\mathbf{q}_j\}_{j=1}^{J}$ which are arranged as

column vectors in the $D \times J$ matrix $\mathbf{Q}$ with $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$. The system anomalies are projected onto the reduced subspace as

$$\mathbf{z} = \mathbf{Q}^{\mathrm{T}}\mathbf{y}. \tag{4}$$

The reconstructed state vector in the full state space is

$$\hat{\mathbf{y}} = \mathbf{Q}\mathbf{z} = \mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{y}. \tag{5}$$

This projection minimizes the representation error $|\mathbf{y} - \mathbf{Q}\mathbf{z}|^2$ with respect to $\mathbf{z}$. Here and in the following, $|\cdot|$ denotes the Euclidean vector norm.

The dynamics in the subspace are assumed to be governed by a system of linear stochastic differential equations,

$$d\mathbf{z} = \mathbf{A}\mathbf{z}\,dt + \mathbf{\Sigma}^{1/2}\,d\mathbf{W}, \tag{6}$$

where $\mathbf{A}$ is the $J \times J$ system matrix or linear operator, $\mathbf{W}$ is a column vector of length J of independent standard Wiener processes and $\mathbf{\Sigma}$ is the symmetric, positive semidefinite noise covariance matrix. Here, only additive and stationary noise is considered, that is, $\mathbf{\Sigma}$ is independent of the system state $\mathbf{z}$ and time. The covariance matrix $\mathbf{\Gamma}$ of the reduced model of Eq. (6) is linked to the system and noise covariance matrices by a fluctuation–dissipation relationship (Gardiner 2010) given by the continuous Lyapunov equation

$$\mathbf{A}\mathbf{\Gamma} + \mathbf{\Gamma}\mathbf{A}^{\mathrm{T}} + \mathbf{\Sigma} = 0. \tag{7}$$

The reduced model is invariant under linear transformations. Let $\mathbf{N}$ be any invertible complex $J \times J$ matrix such that the columns are either real or come as complex conjugate pairs. When transforming the set of patterns as $\mathbf{P} = \mathbf{Q}\mathbf{N}$ and at the same time transforming $\mathbf{z} \to \mathbf{N}^{-1}\mathbf{z}$, $\mathbf{A} \to \mathbf{N}^{-1}\mathbf{A}\mathbf{N}$ and $\mathbf{\Sigma} \to \mathbf{N}^{-1}\mathbf{\Sigma}(\mathbf{N}^{-1})^{\mathrm{T}}$ the reduced dynamics and the state lifted back to the full space are invariant. The projection is then nonorthogonal and the representation error $|\mathbf{y} - \mathbf{P}\mathbf{z}|^2$ is minimized by $\mathbf{z} = \mathbf{P}^{+}\mathbf{y}$ with the Moore–Penrose inverse $\mathbf{P}^{+} = (\mathbf{P}^{\mathrm{H}}\mathbf{P})^{-1}\mathbf{P}^{\mathrm{H}}$, where the superscript H denotes the complex conjugate transpose.

We remark in passing that the projection with respect to the Euclidean scalar product and norm could be generalized to an arbitrary symmetric positive-definite metric $\mathbf{M}$. One would then minimize $(\mathbf{y} - \mathbf{Q}\mathbf{z})^{\mathrm{T}}\mathbf{M}(\mathbf{y} - \mathbf{Q}\mathbf{z}) = |\mathbf{M}^{1/2}(\mathbf{y} - \mathbf{Q}\mathbf{z})|^2$ or $(\mathbf{y} - \mathbf{P}\mathbf{z})^{\mathrm{T}}\mathbf{M}(\mathbf{y} - \mathbf{P}\mathbf{z}) = |\mathbf{M}^{1/2}(\mathbf{y} - \mathbf{P}\mathbf{z})|^2$, respectively. The special case $\mathbf{M} = \mathbf{I}$ corresponds to the Euclidean metric. This is easiest implemented by applying the transformation $\mathbf{y}' = \mathbf{M}^{1/2}\mathbf{y}$ and doing the analysis on $\mathbf{y}'$ with the Euclidean metric. The patterns $\mathbf{Q}'$ and $\mathbf{P}'$ associated with $\mathbf{y}'$ are related to the original patterns $\mathbf{Q}$ and $\mathbf{P}$ by $\mathbf{Q}' = \mathbf{M}^{1/2}\mathbf{Q}$ and $\mathbf{P}' = \mathbf{M}^{1/2}\mathbf{P}$, respectively, and we have $\mathbf{Q}^{\mathrm{T}}\mathbf{M}\mathbf{Q} = \mathbf{I}$. Therefore, we here restrict our attention to the Euclidean case.

Without loss of generality we can always assume an orthonormal basis of the reduced subspace ($\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$), which is convenient when optimizing the subspace (see section 2c). This still leaves the gauge freedom of an arbitrary real orthogonal transformation. A standard representation can be defined by requiring the modes to be uncorrelated, that is, by applying a real orthogonal transformation such that the covariance matrix in the subspace becomes diagonal,

$$\mathbf{Q}^{\mathrm{T}}\mathbf{C}\mathbf{Q} = \mathrm{diag}(\pi_1, \ldots, \pi_J), \tag{8}$$

where π_j gives the variance accounted for by the pattern $\mathbf{q}_j$. When ordering the patterns by decreasing variance the basis is unique for a given subspace up to the sign of the patterns. This corresponds to a PC or EOF analysis in the subspace of the reduced model.

The POPs are the eigenmodes of the system matrix. Generically, the system matrix has distinct eigenvalues and is thus diagonalizable as

$$\mathbf{N}^{-1}\mathbf{A}\mathbf{N} = \mathrm{diag}(\lambda_1, \ldots, \lambda_J), \tag{9}$$

with eigenvalues

$$\lambda_j = \mu_j + i\omega_j \tag{10}$$

and eigenvectors given as the columns of the $J \times J$ matrix $\mathbf{N}$. The POPs $\{\mathbf{p}_j\}_{j=1}^{J}$ in the original state space are given by the columns of the $D \times J$ matrix $\mathbf{P} = \mathbf{Q}\mathbf{N}$, corresponding to an expansion of the anomalies as

$$\hat{\mathbf{y}} = \sum_{j=1}^{J} z_j \mathbf{p}_j. \tag{11}$$

The eigenvalues $\{\lambda_j\}_{j=1}^{J}$, here ordered by nonincreasing real part, the eigenmodes $\{\mathbf{p}_j\}_{j=1}^{J}$ and the expansion coefficients $\{z_j\}_{j=1}^{J}$ are either real or come as complex conjugate pairs. The real modes describe exponentially decaying structures with damping time scale $1/|\mu_j|$. The oscillatory modes describe decaying traveling or standing waves with damping time scale $1/|\mu_j|$ and frequency ω_j or period $T_j = 2\pi/\omega_j$. For $\omega_j > 0$, they undergo the damped POP cycle

$$\mathrm{Re}(\mathbf{p}_j) \to -\mathrm{Im}(\mathbf{p}_j) \to -\mathrm{Re}(\mathbf{p}_j) \to \mathrm{Im}(\mathbf{p}_j) \to \mathrm{Re}(\mathbf{p}_j)$$

at times $t = 0$, $t = T_j/4$, $t = T_j/2$, $t = 3T_j/4$, and $t = T_j$. The POPs are normalized with respect to the (complex) Euclidean norm. There are free-phase factors $e^{i\varphi}$ and $e^{-i\varphi}$ for the complex modes and φ is here chosen such that the norm of the real part of the patterns is maximized. The mode with positive frequency ω_j is listed first for complex conjugate pairs. All of the modes are then uniquely determined up to their sign. The POPs are generally not orthogonal in space unless the system matrix is normal or even symmetric; they are generally also not uncorrelated in time as the stochastic forcing is correlated.

The longest damping time scale of the POP model defines a predictability time scale of the system. It is the time scale at which any deterministic forecast of the POP model relaxes toward the mean, that is, zero forecast. It is also the time scale at which any probabilistic forecast given by the conditional probability density of the POP model converges to the invariant probability density or climatology of the model (Penland 1989).

Alternatively, the system can be modeled by a discrete linear Markov process, that is, a first-order vector autoregressive process or VAR(1) process with time lag τ:

$$\mathbf{z}(t+\tau) = \mathbf{B}_\tau \mathbf{z}(t) + \boldsymbol{\xi}(t). \tag{12}$$

Here, $\mathbf{B}_\tau$ is the finite-time propagator or Green function and $\boldsymbol{\xi}$ is a vector of independent Gaussian white noises with zero mean and covariance matrix $\langle \boldsymbol{\xi}\boldsymbol{\xi} \rangle^{\mathrm{T}} = \boldsymbol{\Omega}_\tau$. The discrete model of Eq. (12) is statistically identical to the continuous model of Eq. (6) sampled at the interval τ provided that the finite-time propagator and the instantaneous system matrix are linked as

$$\mathbf{B}_\tau = \exp(\tau \mathbf{A}) \tag{13}$$

and

$$\mathbf{A} = \frac{1}{\tau} \log \mathbf{B}_\tau. \tag{14}$$

The propagator has the same eigenvectors as the instantaneous system matrix; we have

$$\mathbf{N}^{-1} \mathbf{B}_\tau \mathbf{N} = \boldsymbol{\Lambda}_\tau = \mathrm{diag}[\Lambda_1^{(\tau)}, \ldots, \Lambda_J^{(\tau)}], \tag{15}$$

with

$$\Lambda_j^{(\tau)} = e^{\tau \lambda_j} \tag{16}$$

and

$$\lambda_j = \frac{1}{\tau} \log \Lambda_j^{(\tau)}. \tag{17}$$

The covariance matrix $\boldsymbol{\Gamma}$ of the reduced model of Eq. (12) is the same as above and is linked to the propagator and the noise covariance matrix by the discrete Lyapunov equation

$$\mathbf{B}_\tau \boldsymbol{\Gamma} \mathbf{B}_\tau^{\mathrm{T}} - \boldsymbol{\Gamma} + \boldsymbol{\Omega}_\tau = 0. \tag{18}$$

The maximum and minimum growth supported by the propagator is described by the largest and smallest singular value, respectively, as

$$\sigma_1^{(\tau)} = \max_{\mathbf{z} \neq 0} \frac{|\mathbf{B}_\tau \mathbf{z}|}{|\mathbf{z}|} = \max_{|\mathbf{z}|=1} |\mathbf{B}_\tau \mathbf{z}| \tag{19}$$

and

$$\sigma_J^{(\tau)} = \min_{\mathbf{z} \neq 0} \frac{|\mathbf{B}_\tau \mathbf{z}|}{|\mathbf{z}|} = \min_{|\mathbf{z}|=1} |\mathbf{B}_\tau \mathbf{z}|. \tag{20}$$

For the inference the dataset is projected onto the subspace as $\mathbf{z}_n = \mathbf{Q}^{\mathrm{T}} \mathbf{y}_n$ and a time lag

$$\tau_0 = K \delta t \tag{21}$$

is chosen; the discrete model applied to the dataset is

$$\mathbf{z}_{n+K} = \mathbf{B}_{\tau_0} \mathbf{z}_n + \boldsymbol{\xi}_n, \quad n = 1, \ldots, N-K. \tag{22}$$

In the following, for notational convenience, we use interchangeably the notations $\mathbf{B}_{\tau_0}$ and $\mathbf{B}_K$ for the system matrix and $\boldsymbol{\Omega}_{\tau_0}$ and $\boldsymbol{\Omega}_K$ for the noise covariance matrix. The least squares estimator of the system matrix, minimizing the sum of squared errors

$$F_{\mathrm{dyn}} = \sum_{n=1}^{N-K} |\mathbf{z}_{n+K} - \mathbf{B}_K \mathbf{z}_n|^2, \tag{23}$$

is given by

$$\mathbf{B}_K = \boldsymbol{\Gamma}_K \boldsymbol{\Gamma}_0^{-1} = \mathbf{Q}^{\mathrm{T}} \mathbf{C}_K \mathbf{Q} (\mathbf{Q}^{\mathrm{T}} \mathbf{C}_0 \mathbf{Q})^{-1}, \tag{24}$$

with the covariance matrices

$$\boldsymbol{\Gamma}_0 = \frac{1}{N-K} \sum_{n=1}^{N-K} \mathbf{z}_n \mathbf{z}_n^{\mathrm{T}} = \mathbf{Q}^{\mathrm{T}} \mathbf{C}_0 \mathbf{Q}, \tag{25}$$

$$\boldsymbol{\Gamma}_K = \frac{1}{N-K} \sum_{n=1}^{N-K} \mathbf{z}_{n+K} \mathbf{z}_n^{\mathrm{T}} = \mathbf{Q}^{\mathrm{T}} \mathbf{C}_K \mathbf{Q}, \tag{26}$$

and

$$\mathbf{C}_0 = \frac{1}{N-K} \sum_{n=1}^{N-K} \mathbf{y}_n \mathbf{y}_n^{\mathrm{T}}, \tag{27}$$

$$\mathbf{C}_K = \frac{1}{N-K} \sum_{n=1}^{N-K} \mathbf{y}_{n+K} \mathbf{y}_n^{\mathrm{T}}. \tag{28}$$

For additive noise, as is discussed here, this estimator of the system matrix is also the maximum likelihood estimator. The eigendecomposition

$$\mathbf{B}_{\tau_0} = \mathbf{N} \boldsymbol{\Lambda}_{\tau_0} \mathbf{N}^{-1} \tag{29}$$

is calculated. The instantaneous system matrix and eigenvalues are

$$\mathbf{A} = \frac{1}{\tau_0} \log \mathbf{B}_{\tau_0} \tag{30}$$

and

$$\lambda_j = \frac{1}{\tau_0} \log \Lambda_j^{(\tau_0)}; \tag{31}$$

the propagator for arbitrary time lag τ is

$$\mathbf{B}_\tau = \mathbf{B}_{\tau_0}^{\tau/\tau_0} = \mathbf{N} \boldsymbol{\Lambda}_{\tau_0}^{\tau/\tau_0} \mathbf{N}^{-1}. \tag{32}$$

For stationary or at least weakly stationary datasets, the inference always yields stable system matrices; for all of the eigenvalues we have $|\Lambda_j^{(\tau_0)}| < 1$ and $|\Lambda_j^{(\tau)}| < 1$ for all $\tau_0 > 0$ and $\tau > 0$, or equivalently $\mu_j = \mathrm{Re}(\lambda_j) < 0$ (Penland and Ghil 1993).

A reasonable estimate of the uncertainty in the eigenvalues and associated time scales can be obtained as follows: The standard deviation ρ_{jk} of the estimation error in the matrix element $(B_K)_{jk}$ is given by (Lütkepohl 2005)

$$\rho_{jk}^2 = \frac{1}{N-K} (\Gamma_0^{-1})_{jj} (\Omega_K)_{kk}, \tag{33}$$

where the noise covariance matrix can be estimated as

$$\mathbf{\Omega}_K = \mathbf{\Gamma}_0 - \mathbf{\Gamma}_K \mathbf{\Gamma}_0^{-1} \mathbf{\Gamma}_K^{\mathrm{T}}. \tag{34}$$

Assuming a Gaussian distribution for the parameter errors we generate a large ensemble of errors consistent with the uncertainty, propagate them through to the eigenvalues and time scales and form nonparametric confidence intervals for these. Equation (33) is valid for independent data points. For data with serial correlation $N - K$ is replaced with $(N - K)/\tau_c$, where τ_c is a dimensionless correlation time given as a multiple of the sampling interval δt obtained as the integral of the autocorrelation function estimated from the discrete sample autocorrelation function. The average of the correlation times of the EOF/DMD or OMD modes involved in the model is taken.

If the dynamics truly follow a stochastic linear model then $\mathbf{A}$ and $\mathbf{B}_\tau$ are independent of τ_0 (Penland 1989; Penland and Ghil 1993; Penland and Sardeshmukh 1995). This so-called tau test can thus be used to check for linearity and Markovianity. One would estimate the linear inverse model for various choices of τ_0 and compare the obtained eigenvalues $\{\lambda_j\}_{j=1}^J$ or time scales $\{1/|\mu_j|\}_{j=1}^J$. The tau test passes if the system is linear and Markovian; it generally fails if the system is nonlinear and/or non-Markovian; it also generally fails for nonstationary systems. Moreover, the tau test may spuriously fail due to the Nyquist issue (Penland and Sardeshmukh 1995; Penland 2019), that is, if τ_0 is greater than or close to half of the period of an intrinsic oscillatory mode of variability in the data. Linearity and Markovianity also depend on the number and type of variables and modes chosen for the reduced model.

Given $\mathbf{A}$ the propagator $\mathbf{B}_\tau$ defined by the matrix exponential in Eq. (13) always exists for all τ; in the generic case that $\mathbf{B}_{\tau_0}$ has no real and negative eigenvalues the converse is also true. If $\mathbf{B}_{\tau_0}$ has real and negative eigenvalues no real (matrix) logarithm exists in Eqs. (30) and (31). The discrete model is then Markovian only at the particular time lag τ_0 but there is no corresponding continuous Markovian model and also generally no discrete Markovian model at other time lags τ. In this case $\mathbf{B}_{\tau_0}$ is replaced with $\mathbf{B}_{\tau_0}^* = \mathbf{N}\mathbf{\Lambda}_{\tau_0}^* \mathbf{N}^{-1}$, where $\mathbf{\Lambda}_{\tau_0}^*$ is equal to $\mathbf{\Lambda}_{\tau_0}$ except that the real and negative eigenvalues are replaced with a small positive real number, say, $\varepsilon = 10^{-5}$, effectively eliminating those eigenmodes from the propagator. This situation does not occur for small or moderate τ_0 well below any damping or oscillation time scale present in the reduced model for the chosen number of modes. It may happen at large τ_0, particularly for short datasets, when an eigenmode has virtually decayed to zero but due to sampling uncertainty a real negative eigenvalue with small modulus occurs and it may also happen due to the Nyquist issue (Penland 2019).

b. EOF truncation/DMD

The standard choice for the basis functions spanning the principal subspace for linear inverse modeling are the leading EOFs (Jolliffe 2002). They minimize the state representation error $\sum_{n=1}^N |\mathbf{y}_n - \mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{y}_n|^2$ or equivalently maximize the explained variance

$$\mathrm{explvar0} = \frac{\sum_{n=1}^N |\mathbf{z}_n|^2}{\sum_{n=1}^N |\mathbf{y}_n|^2}, \tag{35}$$

subject to $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$ and are given as the eigenvectors of the covariance matrix, $\mathbf{C}\mathbf{e}_j = \nu_j \mathbf{e}_j$, corresponding to the J largest eigenvalues. The EOFs are arranged as columns of the $D \times J$ matrix $\mathbf{E}$. They are orthogonal in space ($\mathbf{E}^{\mathrm{T}}\mathbf{E} = \mathbf{I}$) and uncorrelated in time [$\mathbf{E}^{\mathrm{T}}\mathbf{C}\mathbf{E} = \mathrm{diag}(\nu_1, \ldots, \nu_J)$]. The eigenvalues are arranged in decreasing order and give the variance accounted for by each mode; we have $\mathrm{explvar0} = \sum_{j=1}^J \nu_j / \sum_{j=1}^D \nu_j$.

We remark that linear inverse modeling with dimension reduction using EOFs is outside of climate science, mainly in the engineering fluid dynamics and dynamical systems communities, more commonly referred to as dynamic mode decomposition (Schmid 2010). As shown by Wynn et al. (2013) the algorithm based on singular value decomposition proposed by Schmid (2010) is exactly the same as linear inverse modeling or POP analysis in an EOF subspace. This type of data-driven analysis occurred much earlier in weather and climate science (Hasselmann 1988; Penland 1989). Also the point of view and the goals of the analysis are slightly broader. Besides the identification of spatial patterns and characteristic time scales the focus is also on prediction; moreover, there is a fully stochastic perspective including probabilistic prediction and uncertainty quantification. Also, in climate science linear inverse modeling is usually applied to anomaly fields whereas DMD is usually applied to the full field without removing the time mean; only very recently a centered DMD was introduced (Hirsh et al. 2020). In the DMD community the focus is on identification of patterns and time scales as well as on the link to the Koopman operator (Rowley et al. 2009; Mezić 2013). Consider an autonomous deterministic dynamical system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}), \tag{36}$$

with state vector $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$ and a scalar-valued observable $g(\mathbf{x})$, that is, a function of the state vector. The Koopman operator $\mathcal{K}$ evolves the observable in time,

$$\mathcal{K}g(\mathbf{x}) = g[\mathcal{F}_\tau(\mathbf{x})], \tag{37}$$

where $\mathcal{F}_\tau$ is the evolution operator associated with $\mathbf{f}$ for a time lag τ. The Koopman operator is a linear operator even for nonlinear systems but is infinite-dimensional even for finite-dimensional systems. It can thus be described by standard spectral theory in terms of eigenvalues and eigenfunctions. The linear operator of the reduced model is now a finite-rank approximation to the Koopman operator in the chosen principal subspace and its eigenvalues and eigenmodes are approximations to the leading Koopman eigenvalues and eigenfunctions. The Koopman operator and the related Perron–Frobenius operator, collectively known as transfer operators, have recently been used for prediction of equatorial Pacific sea surface temperatures (Navarra et al. 2021).

c. OMD

We here discuss the complete linear inverse modeling problem consisting in simultaneous dynamical optimization of the principal subspace and the linear operator. This was already envisaged originally by Hasselmann (1988) but no viable solution presented. The optimal subspace and system matrix are determined by minimizing the objective function

$$F(\mathbf{Q},\mathbf{B}_K)=\sum_{n=1}^{N-K}\left|\mathbf{y}_{n+K}-\mathbf{Q}\mathbf{B}_K\mathbf{Q}^{\mathrm{T}}\mathbf{y}_n\right|^2, \tag{38}$$

subject to the orthonormality constraint $\mathbf{Q}^{\mathrm{T}}\mathbf{Q}=\mathbf{I}$. As detailed in the appendix the objective function can be decomposed as

$$F(\mathbf{Q},\mathbf{B}_K)=F_{\mathrm{dyn}}(\mathbf{Q},\mathbf{B}_K)+F_{\mathrm{pr}}(\mathbf{Q}). \tag{39}$$

Here, the dynamical part

$$\begin{aligned}F_{\mathrm{dyn}}(\mathbf{Q},\mathbf{B}_K)&=\sum_{n=1}^{N-K}\left|\mathbf{Q}^{\mathrm{T}}\mathbf{y}_{n+K}-\mathbf{B}_K\mathbf{Q}^{\mathrm{T}}\mathbf{y}_n\right|^2\\&=\sum_{n=1}^{N-K}\left|\mathbf{z}_{n+K}-\mathbf{B}_K\mathbf{z}_n\right|^2\end{aligned} \tag{40}$$

gives the sum of squared prediction errors within the reduced subspace and depends on both the subspace and the system matrix; the part

$$\begin{aligned}F_{\mathrm{pr}}(\mathbf{Q})&=\sum_{n=1}^{N-K}\left|\mathbf{y}_{n+K}-\mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{y}_{n+K}\right|^2=\sum_{n=1}^{N-K}\left|\mathbf{y}_{n+K}-\mathbf{Q}\mathbf{z}_{n+K}\right|^2\\&=\sum_{n=1}^{N-K}\left|\mathbf{y}_{n+K}\right|^2-\sum_{n=1}^{N-K}\left|\mathbf{z}_{n+K}\right|^2\end{aligned} \tag{41}$$

is the sum of squared projection errors and depends only on the subspace.

For given patterns $\mathbf{Q}$ the system matrix $\mathbf{B}_K=\mathbf{B}_K(\mathbf{Q})$ which minimizes F_{dyn}, and thus also F, is given by Eq. (24). Inserting this result into Eq. (38) yields the objective function as a function of only the patterns

$$F(\mathbf{Q})=\sum_{n=1}^{N-K}\left|\mathbf{y}_{n+K}-\mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{C}_K\mathbf{Q}(\mathbf{Q}^{\mathrm{T}}\mathbf{C}_0\mathbf{Q})^{-1}\mathbf{Q}^{\mathrm{T}}\mathbf{y}_n\right|^2, \tag{42}$$

which is to be minimized with respect to the patterns subject to $\mathbf{Q}^{\mathrm{T}}\mathbf{Q}=\mathbf{I}$. Also F_{dyn} can be written as a function of only $\mathbf{Q}$ by inserting the expression for $\mathbf{B}_K(\mathbf{Q})$ into Eq. (40). The objective function F is only a function of the subspace and not the particular orthonormal basis functions; it is readily verified that $\mathbf{B}_K(\mathbf{QU})=\mathbf{U}^{\mathrm{T}}\mathbf{B}_K(\mathbf{Q})\mathbf{U}$, $F_{\mathrm{dyn}}(\mathbf{QU})=F_{\mathrm{dyn}}(\mathbf{Q})$, $F_{\mathrm{pr}}(\mathbf{QU})=F_{\mathrm{pr}}(\mathbf{Q})$, and $F(\mathbf{QU})=F(\mathbf{Q})$ for any real orthogonal $J\times J$ matrix $\mathbf{U}$ ($\mathbf{U}^{\mathrm{T}}\mathbf{U}=\mathbf{U}\mathbf{U}^{\mathrm{T}}=\mathbf{I}$). Minimizing F is a difficult nonlinear minimization problem, which clearly lies outside the framework of the classical linear pattern identification techniques such as PC/EOF analysis, maximum covariance analysis (MCA) or canonical correlation analysis (CCA) (von Storch and Zwiers 2002; Hannachi 2021). The solution, termed *optimal mode decomposition*, was proposed only relatively recently (Goulart et al. 2012; Wynn et al. 2013) and makes use of advanced optimization techniques on matrix manifolds (Edelman et al. 1998; Absil et al. 2008). The OMD algorithm is presented in the appendix. A MATLAB implementation is publicly available on GitHub: https://github/FKwasniok/OMD/.

Unlike in the case of EOFs/DMD the OMD subspaces for different dimensions of the reduced model J_1 and J_2 are not nested, though in typical applications they are to a very good approximation. One could consider constructing a sequential algorithm in which the patterns are introduced one by one, but this is outside the scope of the present paper.

We introduce the explained variances

$$\mathrm{explvar1}=1-\frac{\sum_{n=1}^{N-K}\left|\mathbf{y}_{n+K}-\mathbf{Q}\mathbf{B}_K\mathbf{Q}^{\mathrm{T}}\mathbf{y}_n\right|^2}{\sum_{n=1}^{N-K}\left|\mathbf{y}_{n+K}\right|^2}=\frac{\sum_{n=1}^{N-K}\left|\mathbf{B}_K\mathbf{z}_n\right|^2}{\sum_{n=1}^{N-K}\left|\mathbf{y}_{n+K}\right|^2} \tag{43}$$

and

$$\mathrm{explvar2}=1-\frac{\sum_{n=1}^{N-K}\left|\mathbf{z}_{n+K}-\mathbf{B}_K\mathbf{z}_n\right|^2}{\sum_{n=1}^{N-K}\left|\mathbf{z}_{n+K}\right|^2}=\frac{\sum_{n=1}^{N-K}\left|\mathbf{B}_K\mathbf{z}_n\right|^2}{\sum_{n=1}^{N-K}\left|\mathbf{z}_{n+K}\right|^2}. \tag{44}$$

The expression explvar1 gives the explained variance as a fraction of the variance in the full state space; the expression explvar2 gives the explained variance relative to the variance in the selected subspace. The OMD maximizes explvar1 as the denominator is a constant and explvar1 is a linear function of F with negative slope; the OMD generally does not maximize explvar2 as the denominator depends on $\mathbf{Q}$. The second equality in Eqs. (43) and (44) is shown in the appendix; it holds in the learning dataset in which the model is estimated and reflects the fact that the OMD model is calibrated in sample, that is, the variance of the model predictions is equal to their (total) explained variance.

Alternatively, one may suggest minimizing F_{dyn} or maximizing explvar2 rather than minimizing F or equivalently maximizing explvar1. PDCs (de la Iglesia and Tabak 2013) are defined by minimizing F_{dyn}. The proposed algorithm is less efficient than the OMD algorithm, particularly at large J, but the OMD algorithm could be readily adapted to minimizing F_{dyn}. Nevertheless, we here prefer to minimize the squared error against a prediction target that is independent of the dimension and the subspace of the reduced model. Moreover, the optimization of F is likely to have a better condition than the optimization of F_{dyn} or explvar2 as the term F_{pr} acts as a kind of regularization term. A similar strategy was pursued in Kwasniok (1996, 2007).

3. System identification: Two pedagogical examples

The OMD method is first demonstrated on two simulation datasets from simple linear dynamical systems representing two prototype situations in which an advantage of the OMD

models over the EOF/DMD models could be expected: (i) a linear operator under nonuniform stochastic forcing and (ii) the truncation of a strongly nonnormal linear operator.

a. Example 1: Nonuniform excitation of the system modes

We consider a system with a D-dimensional state vector $\mathbf{x}$ subject to periodic boundary conditions. One may think of a grid representation of a spatially extended system in one space dimension. The dynamics are governed by a three-dimensional linear subsystem described by Eq. (6) with eigenvalues $\lambda_1 = \mu_1$, $\lambda_2 = \mu_2 + i\omega$, and $\lambda_3 = \mu_2 - i\omega$. We have a single purely damped mode and an oscillatory pair of modes. The variables of the reduced system are linked to the modes $\mathbf{p}_1$, $\mathbf{p}_2$, and $\mathbf{p}_3$ which are arranged as column vectors in the $D \times 3$ matrix $\mathbf{P}$ and are given as

$$P_{k1} = \frac{1}{\sqrt{D}}\left(\sin\frac{2\pi k}{D} + \sin\frac{4\pi k}{D}\right), \tag{45}$$

$$P_{k2} = \frac{1}{\sqrt{D}}\left(\cos\frac{4\pi k}{D} + i\sin\frac{4\pi k}{D}\right), \tag{46}$$

$$P_{k3} = \frac{1}{\sqrt{D}}\left(\cos\frac{4\pi k}{D} - i\sin\frac{4\pi k}{D}\right), \tag{47}$$

for $k = 1, \ldots, D$. The oscillatory modes form a traveling wave with wavenumber 2 and frequency ω; the single mode is a mixture of a wavenumber 1 and a wavenumber 2 pattern. The system modes are normalized ($\mathbf{p}_j^{\mathrm{H}}\mathbf{p}_j = 1$) but not orthogonal as is often the case in climate data, for example, when studying atmospheric teleconnection patterns. For the simulation of the system we switch to a discrete representation with $\delta t = \tau = 1$ and also the system recovery is performed at this time lag, that is, $K = 1$ and $\tau_0 = 1$. The full state vector $\mathbf{x}$ follows the stochastic process

$$\mathbf{x}_{n+1} = \mathbf{P}\mathbf{B}_1\mathbf{P}^{+}\mathbf{x}_n + \mathbf{P}\mathbf{P}^{+}\boldsymbol{\zeta}_n + \eta(\mathbf{I} - \mathbf{P}\mathbf{P}^{+})\boldsymbol{\zeta}_n, \tag{48}$$

with $\mathbf{B}_1 = \mathrm{diag}(e^{\lambda_1}, e^{\lambda_2}, e^{\lambda_3})$ and $\boldsymbol{\zeta}_n$ being a column vector of length D of independent Gaussian white noises with zero mean and unit variance. Equation (48) is obtained by lifting the reduced models of Eq. (12) or (22) to the full state space, that is, multiplying from the left with $\mathbf{P}$ but applying the stochastic forcing in the full space. The propagator matrix $\mathbf{P}\mathbf{B}_1\mathbf{P}^{+}$ has rank three and is (moderately) nonnormal as can be seen from the nonorthogonality of the eigenmodes. The noise in the dynamical subspace is $\boldsymbol{\xi}_n = \mathbf{P}^{+}\boldsymbol{\zeta}_n$; it is correlated between the modes with covariance matrix $\langle \boldsymbol{\xi}_n \boldsymbol{\xi}_n^{\mathrm{T}} \rangle = \boldsymbol{\Omega}_1 = (\mathbf{P}^{\mathrm{H}}\mathbf{P})^{-1}$. The last term in Eq. (48) is the noise outside the dynamical subspace. The parameter $\eta \geq 0$ allows for different configurations of the noise excitation. For $\eta = 0$, only the dynamical modes in the three-dimensional subspace are excited. The covariance matrix of $\mathbf{x}$ is then degenerate and has only rank three; both the EOF/DMD and OMD models are always able to identify the system correctly. The case $\eta = 1$ corresponds to uniform excitation of all components of $\mathbf{x}$; other values of η correspond to nonuniform excitation of the system.

The dimension of the system is here chosen as $D = 50$ and the frequency of the oscillation is set to $\omega = 0.5$, corresponding to a period of 4π. The damping parameters μ_1 and μ_2 are each varied on the interval $[-1, -0.01]$. For each pair of values (μ_1, μ_2) a dataset of length $N = 5000$ is generated by iterating Eq. (48); then EOF/DMD and OMD models are derived from the dataset. Our primary interest is here in system identification rather than the explained variances or predictive skill. We look at the estimates for μ_1, μ_2, and ω given by the least damped real and complex eigenvalues, if any, of the system matrix of the extracted reduced models. The identification of the modes is assessed using the squared projections

$$\alpha = (\mathbf{p}_1^{\mathrm{T}}\mathbf{p}_r)^2 \tag{49}$$

and

$$\beta = \left|\mathbf{p}_2^{\mathrm{H}}\mathbf{p}_{\mathrm{osc}}\right|^2, \tag{50}$$

where $\mathbf{p}_r$ and $\mathbf{p}_{\mathrm{osc}}$ are the normalized least damped real and oscillatory modes, respectively, of the reduced model.

In a noisy system the reduced model with $J = 3$ cannot be expected to exactly identify the three dynamical modes, particularly with EOFs/DMD. It therefore turns out to be beneficial to choose the dimension of the reduced model slightly higher than the actual dimension of the system. Best results are obtained with $J = 6$ for EOF/DMD models and $J = 4$ for OMD models.

Figures 1 and 2 show the results for $\eta = 1$, that is, uniform excitation of the system. The OMD achieves excellent reconstruction, apart from sampling errors, of all the system properties over the whole range of values for μ_1 and μ_2. The reconstruction is virtually perfect if the respective mode is weakly damped; for strongly dissipative modes slight errors occur. For weakly damped, that is, highly predictable modes the EOF/DMD model performs as well as the OMD model. For stronger damping, that is, low predictability the EOF/DMD model underestimates the predictability of the modes as the EOF space becomes entrained with the unpredictable noise in the system. There is then also considerable error in the frequency of the oscillation and the identification of the structure of the modes. This finding is in line with the fluctuation–dissipation relation which implies that under uniform stochastic excitation the most predictable modes are those with the largest variance.

We now examine a case with a higher noise level in the unpredictable modes of the system. Figures 3 and 4 display the results for $\eta = \sqrt{2}$. The OMD still reconstructs all aspects of the system very accurately for all values of μ_1 and μ_2 with the same minor limitations as described above. There is only a slight drop in performance compared to the case $\eta = 1$. The EOF/DMD model performs much worse in this setting. It completely misses the oscillatory modes if the persistence time scale is below about 3 time units as they are overwhelmed by the noise; above this threshold the oscillator is identified with some error in the frequency provided the real mode is strongly damped. If the real mode is also rather persistent the EOF/DMD model identifies the damping parameter of the oscillator correctly but has a large error in the frequency and some error in the pattern structure. The real mode is identified as long as

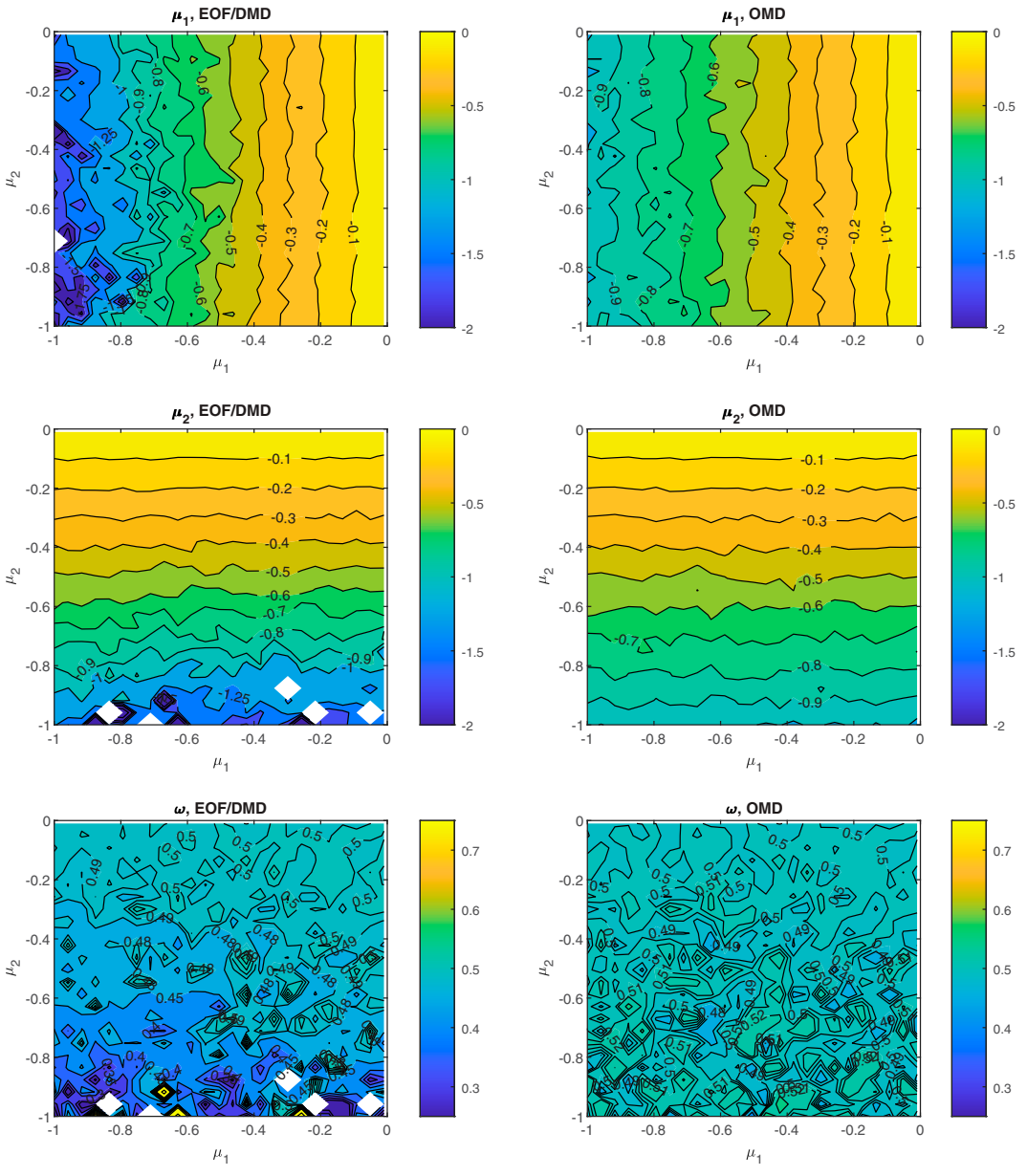


FIG. 1. Example 1: (from top to bottom) Estimated damping parameters μ_1 and μ_2 as well as frequency ω as a function of the true damping parameters μ_1 and μ_2. (left) EOF/DMD models. (right) OMD models. White patches indicate that there is no real or oscillatory mode in the reduced model. The stochastic forcing is uniform ($\eta = 1$).

its time scale is larger than about 2.5 time units and the oscillator is strongly damped but the predictability is overestimated and there is some error in the pattern. If both the real mode and the oscillator are rather persistent there is a large range of values of μ_1 and μ_2 for which the real mode is completely missed by the EOF/DMD model. The interference between the modes hampering the inference is due to their lack of orthogonality in $\mathbf{x}$ space. If the patterns are orthogonal, that is, the system is normal or even Hermitian the identification is easier. The EOF/DMD model then correctly reconstructs both modes if they are above a certain persistence threshold but still misses them completely if they are below the threshold (not shown). The example shown here is a typical case. The details of what the EOF/DMD model can and cannot recover depend on the degree and structure of the overlap between the patterns, that is, the nonnormality of the system as well as on the covariance structure of the noise.

b. *Example 2: Strongly nonnormal system matrix*

For nonnormal system matrices the eigenvalues and eigenmodes describe only the asymptotic dynamics; the short- and medium-term behavior may be characterized by substantial nonmodal growth (Trefethen and Embree 2005; Schmid 2007). The construction of reduced models is then particularly

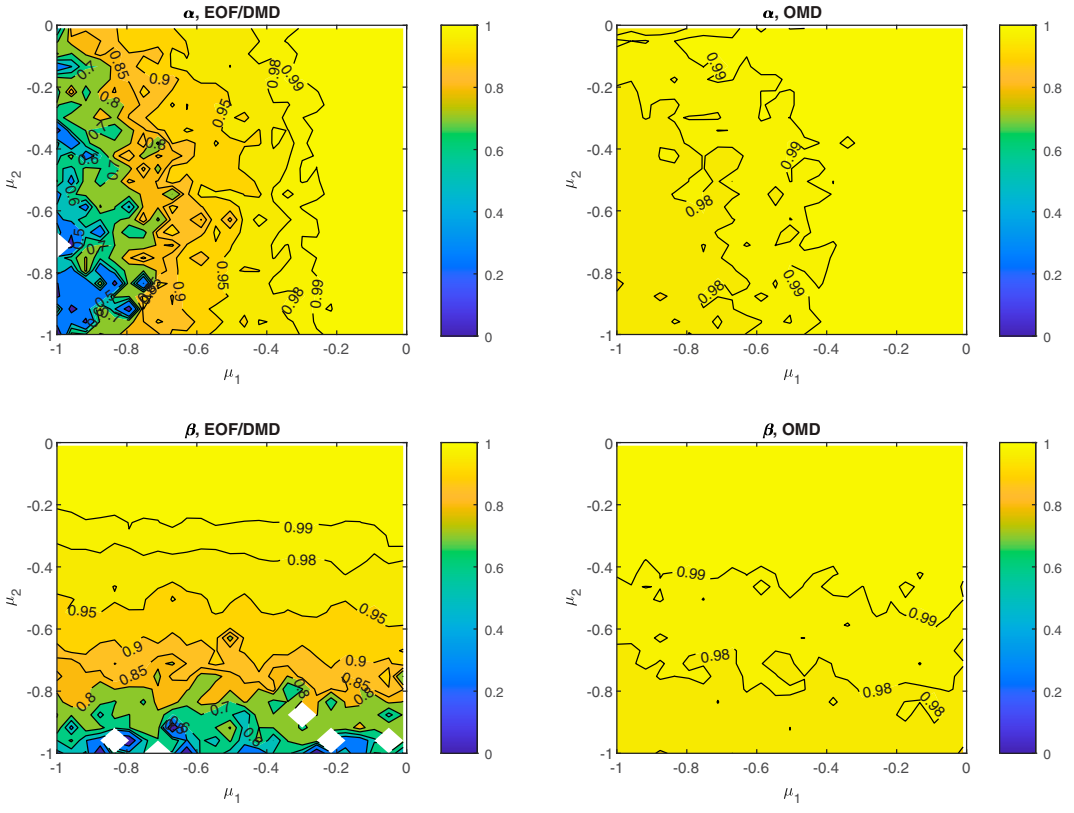


FIG. 2. Example 1: Squared projections (top) α and (bottom) β given in Eqs. (49) and (50) as a function of the true damping parameters μ_1 and μ_2. (left) EOF/DMD models. (right) OMD models. White patches indicate that there is no real or oscillatory mode in the reduced model. The stochastic forcing is uniform ($\eta = 1$).

challenging as the reduced basis would need to contain both the excitation and the response patterns of the system. The issue is here illustrated on a very simple example.

We introduce a scalar measure of nonnormality of a linear operator. The departure from normality of an $L \times L$ matrix $\mathbf{X}$ is defined as (Henrici 1962)

$$\mathrm{dep}(\mathbf{X}) = \left(\sum_{j,k=1}^{L} X_{jk}^2 - \sum_{j=1}^{L} |\lambda_j|^2 \right)^{1/2} = \left(\sum_{j=1}^{L} \sigma_j^2 - \sum_{j=1}^{L} |\lambda_j|^2 \right)^{1/2}, \tag{51}$$

where $\{\sigma_j\}_{j=1}^{L}$ are the singular values and $\{\lambda_j\}_{j=1}^{L}$ are the eigenvalues of $\mathbf{X}$. We have $\mathrm{dep}(\mathbf{X}) \geq 0$ and $\mathrm{dep}(\mathbf{X}) = 0$ if and only if $\mathbf{X}$ is normal. The measure dep summarizes the total nonnormality across all state space dimensions rather than just the maximum growth described by the leading singular value. It already provides useful information when applied to infinitesimal system matrices although the finite-time propagator is always more informative.

We consider the three-dimensional continuous system matrix

$$\overline{\mathbf{A}} = \begin{pmatrix} \mu_1 & c & 0 \\ 0 & \mu_2 & 0 \\ 0 & 0 & \mu_3 \end{pmatrix}, \tag{52}$$

with the real parameters $\mu_1 < 0$, $\mu_2 < 0$, $\mu_3 < 0$, and $c \geq 0$ under uniform and uncorrelated stochastic forcing with unit variance for all components. The overbar denotes the system matrix of the complete system as opposed to the reduced system. The matrix $\overline{\mathbf{A}}$ has eigenvalues μ_1, μ_1, and μ_1, for simplicity here assumed to be all distinct, with corresponding eigenvectors $(1, 0, 0)^{\mathrm{T}}$, $[1, (\mu_2 - \mu_1)/c, 0]^{\mathrm{T}}$, and $(0, 0, 1)^{\mathrm{T}}$. For $c \to 0$, $\overline{\mathbf{A}}$ is normal, the eigenvectors are orthogonal and point in the coordinate directions. For increasing c, the nonnormality increases and the first two eigenvectors become more and more aligned. We have $\mathrm{dep}(\overline{\mathbf{A}}) = c$. The finite-time propagator is readily worked out to be given by

$$\overline{\mathbf{B}}_\tau = e^{\tau \overline{\mathbf{A}}} = \begin{bmatrix} e^{\mu_1 \tau} & \dfrac{c}{\mu_2 - \mu_1}(e^{\mu_2 \tau} - e^{\mu_1 \tau}) & 0 \\ 0 & e^{\mu_2 \tau} & 0 \\ 0 & 0 & e^{\mu_3 \tau} \end{bmatrix}. \tag{53}$$

It has the eigenvalues $e^{\mu_1 \tau}$, $e^{\mu_2 \tau}$, and $e^{\mu_3 \tau}$ with the same eigenvectors as $\overline{\mathbf{A}}$ and

$$\mathrm{dep}(\overline{\mathbf{B}}_\tau) = \frac{c}{\mu_2 - \mu_1}(e^{\mu_2 \tau} - e^{\mu_1 \tau}). \tag{54}$$

The maximum of $\mathrm{dep}(\overline{\mathbf{B}}_\tau)$ is attained for $\tau_m = (\log|\mu_1| - \log|\mu_2|)/(\mu_2 - \mu_1)$.

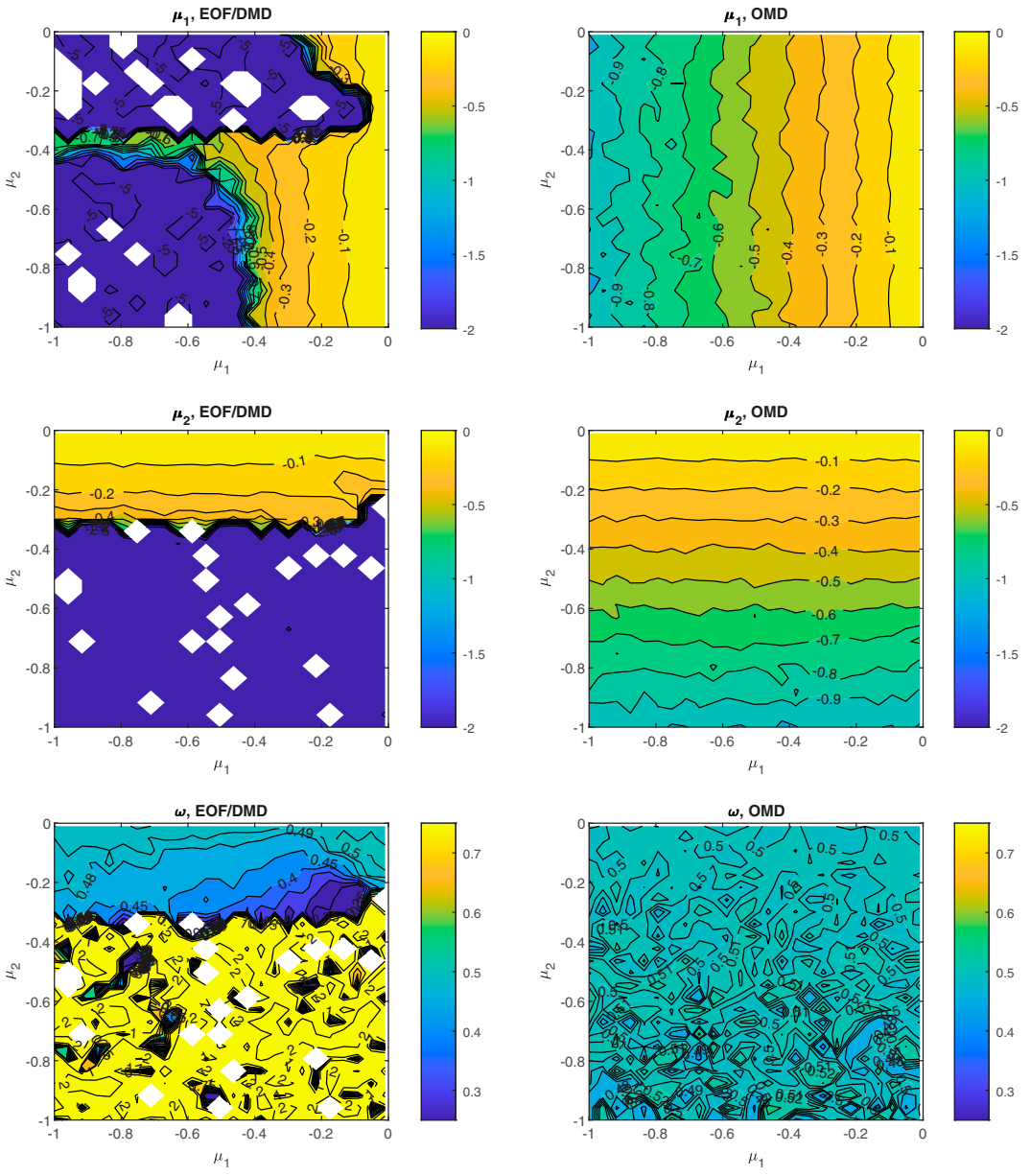


FIG. 3. Example 1: As in Fig. 1, but with nonuniform stochastic forcing ($\eta = \sqrt{2}$).

Here, the parameter setting $\mu_1 = -0.5$, $\mu_2 = -0.25$, $\mu_3 = -0.2$ and $c = 5$ is chosen. The maximum of $\mathrm{dep}(\overline{\mathbf{B}}_\tau)$ is attained for $\tau_m = 4\log 2 \approx 2.77$ with $\mathrm{dep}(\overline{\mathbf{B}}_{\tau_m}) = c$ as $\mu_1 = 2\mu_2$; the general expression for $\mathrm{dep}(\overline{\mathbf{B}}_{\tau_m})$ is more complicated. The exact covariance matrix of the system given by Eq. (7) is

$$\mathbf{C} = \begin{pmatrix} 134.77 & 12.92 & 0 \\ 12.92 & 2.54 & 0 \\ 0 & 0 & 3.03 \end{pmatrix}. \tag{55}$$

The least damped eigenmode points in the x_3 direction but most of the variance is generated by the nonnormality in the x_1–x_2 plane.

A dataset of length $N = 5000$ with sampling interval $\delta t = 1$ is generated by iterating $\overline{\mathbf{B}}_1$ according to

$$\mathbf{x}_{n+1} = \overline{\mathbf{B}}_1 \mathbf{x}_n + \zeta_n, \tag{56}$$

with $\langle \zeta_n \zeta_n^{\mathrm{T}} \rangle = \mathbf{I}$. Then reduced EOF/DMD and OMD models of dimension $J = 2$ are estimated with $K = 1$ and $\tau_0 = 1$. The reduced system matrices $\mathbf{B}_1$ are lifted back to the full state space as $\mathbf{E}\mathbf{B}_1\mathbf{E}^{\mathrm{T}}$ and $\mathbf{Q}\mathbf{B}_1\mathbf{Q}^{\mathrm{T}}$ for the EOF/DMD and OMD case, respectively, to allow for a direct comparison with $\overline{\mathbf{B}}_1$. Table 1 summarizes the characteristics of the EOF/DMD and OMD subspaces and system matrices. The EOF/DMD subspace is essentially the x_1–x_3 plane and the reduced system matrix is almost symmetric. The EOF/DMD model covers the

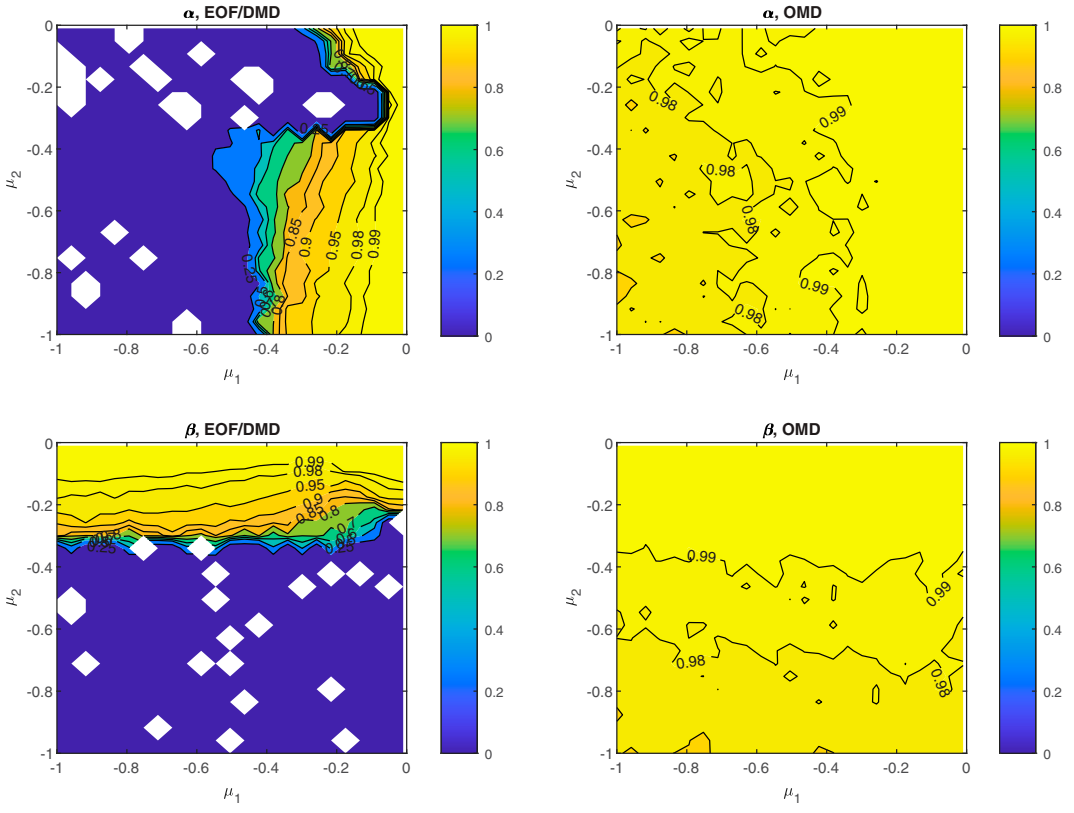


FIG. 4. Example 1: As in Fig. 2, but with nonuniform stochastic forcing ($\eta = \sqrt{2}$).

most persistent modes as evidenced by the eigenvalues of the linear operator but completely misses the nonnormality and the associated nonmodal growth accounting for most of the variance. The OMD subspace is basically the x_1–x_2 plane. It is slightly behind the EOF/DMD subspace in explained variance explvar0 but clearly ahead in the predictive explained variances explvar1 and explvar2. The reduced system matrix almost perfectly captures the nonnormality of the system as evidenced by the singular values and the departure from normality. Figure 5 shows the departure from normality and the largest singular value of the finite-time propagators of the full system and the reduced models over the full range of lead times. The OMD model very faithfully captures the transient growth supported by the full system whereas the EOF/DMD

TABLE 1. Example 2: Basis functions and explained variances of the EOF/DMD and OMD subspaces as well as the system matrices of the full and reduced models together with their eigenvalues, singular values, and departure from normality. Deviations of the basis functions from normalization are due to rounding.

	True	EOF/DMD	OMD
$\mathbf{Q}$	$\begin{pmatrix} 0.995 & 0.006 & -0.096 \\ 0.096 & -0.019 & 0.995 \\ -0.004 & 1.000 & 0.020 \end{pmatrix}$	$\begin{pmatrix} 0.995 & 0.006 \\ 0.096 & -0.019 \\ -0.004 & 1.000 \end{pmatrix}$	$\begin{pmatrix} 0.995 & -0.096 \\ 0.096 & 0.995 \\ -0.005 & -0.001 \end{pmatrix}$
explvar0	100.0%	99.1%	97.9%
explvar1	97.9%	86.5%	96.4%
explvar2	97.9%	87.4%	98.5%
$\overline{\mathbf{B}}_1$	$\begin{pmatrix} 0.607 & 3.445 & 0 \\ 0 & 0.779 & 0 \\ 0 & 0 & 0.819 \end{pmatrix}$	$\begin{pmatrix} 0.928 & 0.091 & -0.068 \\ 0.090 & 0.009 & -0.022 \\ 0.001 & -0.016 & 0.809 \end{pmatrix}$	$\begin{pmatrix} 0.606 & 3.446 & -0.009 \\ 0.001 & 0.767 & -0.001 \\ -0.003 & -0.016 & 0.000 \end{pmatrix}$
$\Lambda_1^{(1)}, \Lambda_2^{(1)}, \Lambda_3^{(1)}$	0.819, 0.779, 0.607	0.937, 0.809, 0.000	0.785, 0.588, 0.000
$\sigma_1^{(1)}, \sigma_2^{(1)}, \sigma_3^{(1)}$	3.582, 0.819, 0.132	0.947, 0.801, 0.000	3.579, 0.129, 0.000
$\mathrm{dep}(\overline{\mathbf{B}}_1)$	3.446	0.077	3.444

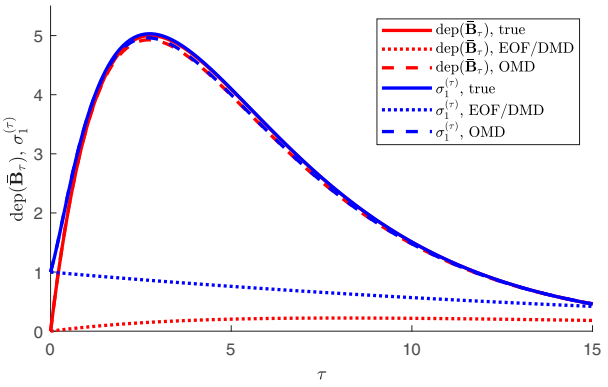


FIG. 5. Example 2: Departure from normality and largest singular value of the finite-time propagator of the full system and the reduced EOF/DMD and OMD models as a function of lead time.

model completely misses it and just displays the decay described by the eigenvalues. Here, due to the low dimension of the system, the leading singular value already completely characterizes the nonnormality and carries virtually the same information as the departure from normality. This can also be seen from the information in Table 1.

We remark that balanced truncation (Moore 1981; Glover 1984; Farrell and Ioannou 2001a) for this example yields virtually the same results as OMD. However, the applicability of balanced truncation is rather restricted. It assumes that the true dynamics of the full system are purely linear driven by stochastic forcing and needs both the system matrix and the noise covariance matrix as input. By contrast, the OMD is fully data driven, can be applied to nonlinear systems, and picks only modes which are active in the data under the nonlinear dynamics.

4. Large-scale atmospheric dynamics

a. The atmospheric model and the dataset

We here use a spectral quasigeostrophic (QG) three-level atmospheric model on the Northern Hemisphere, triangularly truncated at wavenumber 21 (Kwasniok 2007, 2019). The model levels are located at the 250, 500, and 750 hPa pressure surfaces. The governing equations are

$$\frac{\partial q_i}{\partial t} + J(\Psi_i, q_i) = D_i + S_i, \quad i = 1,2,3. \tag{57}$$

Here, Ψ_I and q_i are the streamfunction and the potential vorticity at level i, respectively, and J denotes the Jacobian operator on the sphere. The dissipative terms D_i comprise Newtonian temperature relaxation at all levels, Ekman damping at the lowest level, and hyperviscosity on the time-dependent part of the potential vorticity at all levels. The time-independent but spatially varying forcing terms S_i are diabatic sources of potential vorticity.

The model parameters and forcing are tuned in such a way that the model in a long-term integration exhibits a remarkably realistic mean state and variance pattern of streamfunction and potential vorticity. The model is integrated forward in time using the third-order Adams–Bashforth scheme with a constant time step of 1 h. The details of the model configuration, parameter setting, parameter tuning procedure, and performance versus reanalysis data can be found in Kwasniok (2007) and Kwasniok (2019). The model configuration used here is exactly the same as described in Kwasniok (2019). The model is similar to the one proposed by Marshall and Molteni (1993).

A posttransient long-term integration with the QG model is performed and 25 000 days' worth of data are archived at intervals of 12 h, resulting in a dataset of length $N = 50\,000$. The time mean state is removed from the dataset. The reduced models are constructed from the 500 hPa streamfunction anomaly field which is given at every time instant by $D = 231$ spectral coefficients. The Euclidean scalar product and associated norm in spectral space is used throughout for the projection onto the reduced subspaces and the calculation of EOFs; in physical space this corresponds to the norm streamfunction metric. Linear inverse models based on EOFs/DMD and OMD are considered for dimensions ranging from $J = 1$ to $J = 25$ and for time lags ranging from $\tau_0 = 1$ day to $\tau_0 = 8$ days.

b. Characterization of the OMD modes and subspace

Figure 6 shows the cumulative explained variances explvar0, explvar1 and explvar2 as a function of dimension for various time lags. The explained state variance is maximized by the EOFs/DMD; the OMD always lies below, with the explained variance generally decreasing with increasing time lag τ_0. For example, at the dimension $J = 10$, the EOFs explain 48.9% of the variance of the streamfunction anomaly field whereas for the OMD modes at the time lags $\tau_0 = 1$ day, $\tau_0 = 2$ days, $\tau_0 = 4$ days, and $\tau_0 = 6$ days it is 48.1%, 45.3%, 40.2%, and 36.3%, respectively.

The quantity explvar1 is maximized by the OMD. The improvement in predictive explained variance on the EOFs/DMD is substantial taking into account that the captured variance is lower with the OMD. This is better highlighted by explvar2, the predictive explained variance in the respective subspace, although this quantity is not directly optimized. For example, with $J = 5$ and $\tau_0 = 4$ days explvar2 increases from 20.4% for DMD to 38.1% for OMD. Interestingly, explvar2 is large for rather low dimensions of the OMD models, actually largest at $J = 1$, indicating a dynamically closed system, and then slowly drops with increasing dimension. We remark that for $J = 1$ the OMD finds a mode with an autocorrelation of 0.80 at $\tau_0 = 2$ days and 0.65 at $\tau_0 = 4$ days whereas the autocorrelation of the first EOF at these time lags is only 0.69 and 0.30, respectively. When including more modes in the model, in particular small-scale ones, it becomes more difficult to get a high predictive explained variance for all of them.

Figure 7 gives a comparison of the EOF/DMD and OMD subspaces by displaying for each EOF $\mathbf{e}_k$ the squared projection,

$$\gamma_k = \sum_{j=1}^{J} (\mathbf{q}_j^{\mathrm{T}} \mathbf{e}_k)^2, \tag{58}$$

for various dimensions J and time lags τ_0. We have $0 \le \gamma_k \le 1$ and $\sum_{k=1}^{D} \gamma_k = J$. The EOF $\mathbf{e}_k$ lies in the subspace spanned by

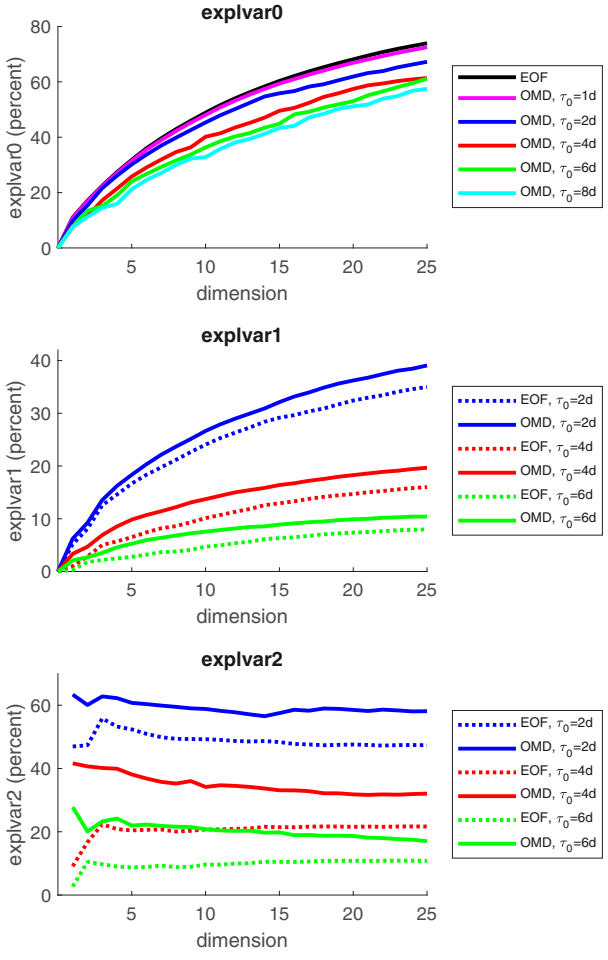


FIG. 6. (top) Explained variance of the 500 hPa streamfunction anomaly field, (middle) predictive explained variance over the time lag τ_0, and (bottom) predictive explained variance over the time lag τ_0 in the respective subspace of the reduced model.

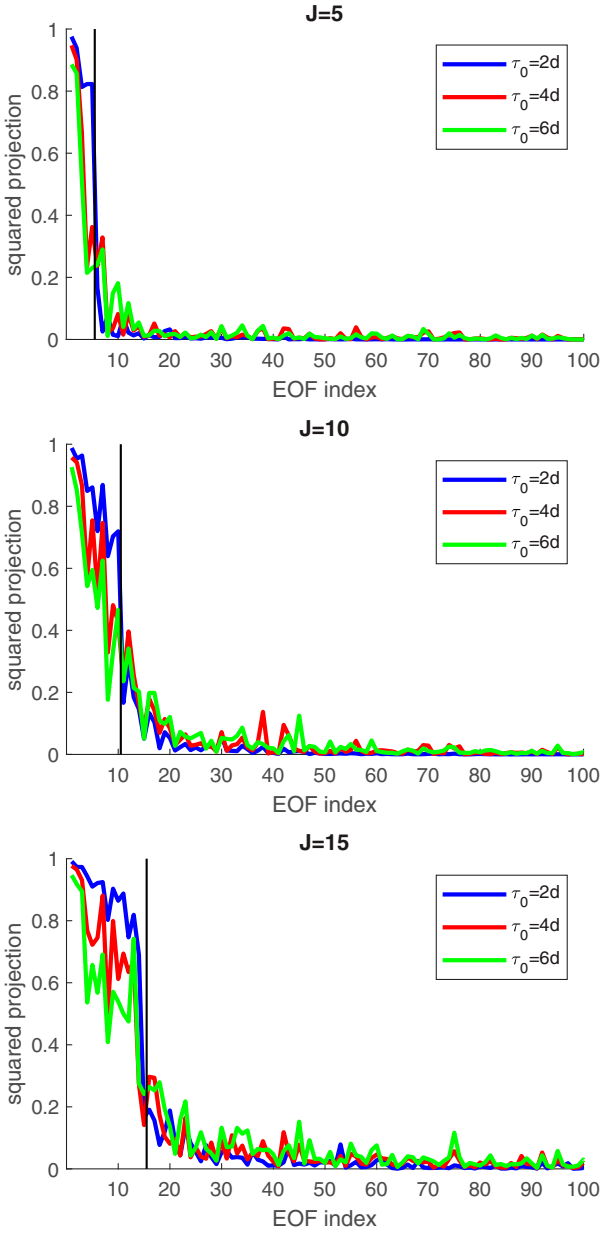


FIG. 7. Squared projection γ_k of individual EOFs $\mathbf{e}_k$ onto the OMD subspace for (top) $J = 5$, (middle) $J = 10$, and (bottom) $J = 15$. The vertical lines indicate the dimensions of the models.

the modes $\{\mathbf{q}_j\}_{j=1}^{J}$ if $\gamma_k = 1$ and is orthogonal to it if $\gamma_k = 0$. For the DMD, we have $\mathbf{q}_j = \mathbf{e}_j$ and thus $\gamma_k = 1$ for $k \leq J$ and $\gamma_k = 0$ for $k > J$.

The EOF/DMD and OMD subspaces are markedly different. Significant contributions from EOF modes $\mathbf{e}_k$ with $k > J$ in the OMD subspaces are observed which increase and involve higher and higher EOFs with increasing time lag τ_0 and dimension J, in line with the results in Fig. 6. These contributions are correlated between different time lags and dimensions.

Figure 8 displays the eigenvalue spectra of the system matrix of the EOF/DMD and OMD models at $\tau_0 = 4$ days for $J = 5$ and $J = 10$. The corresponding damping time scales and oscillation periods are listed in Table 2. We will refer back to these models later when discussing prediction skill in detail. The OMD modes systematically have longer damping time scales, that is, longer predictability time scales than the DMD modes. Both with DMD and OMD, there are real modes and complex conjugate pairs of modes; some oscillatory modes have very long periods such that they are effectively equivalent to two real damped modes, some have periods short enough such that a POP cycle can really be observed.

The least damped eigenmodes of the EOF/DMD and OMD models for $J = 10$ and $\tau_0 = 4$ days are displayed in Fig. 9. They both resemble the Arctic Oscillation; in the pattern from the OMD model the center over the Pacific is shifted to the Asian continent and the dipole structure between the midlatitudes and the subtropics is more pronounced. The pattern correlation between the two patterns is 0.89; the difference is highly significant given the large dataset. The mode from the OMD is more persistent with a damping time scale of 8.6 days versus

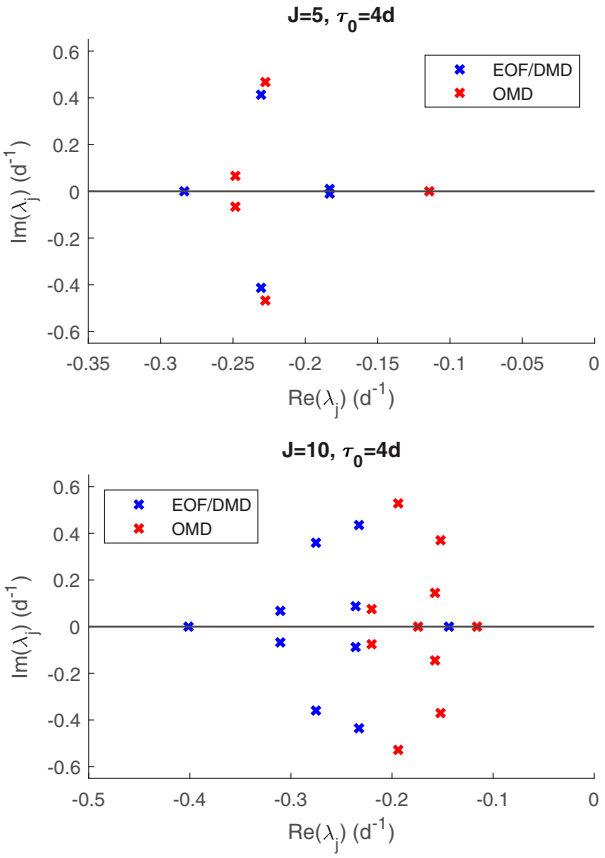


FIG. 8. Eigenvalue spectrum of the system matrix of the EOF/DMD and OMD models for $\tau_0 = 4$ days and (top) $J = 5$ and (bottom) $J = 10$.

7.0 days for the DMD. The positive and negative phases of the Arctic Oscillation also occur as two out of four cluster centers in a hidden Markov model regime analysis of the same QG model (Allen et al. 2020). They are the most persistent regimes with mean residence times of 9.6 and 10.7 days and frequencies of occurrence of 26% and 25% for the positive and negative phase, respectively.

The second eigenmode of the OMD model for $J = 10$ and $\tau_0 = 4$ days is an oscillatory pair and is shown in Fig. 10. It is a low-variance mode and does not bear resemblance with any of the well-known teleconnection patterns. The real part exhibits a wave train structure with three centers extending from the North Atlantic over the pole to the Pacific; the imaginary part has a wave train with four centers extending from Eurasia over the pole to North America.

A more comprehensive overview of the time scales of the eigenmodes of the reduced models as a function of the dimension and the time lag is given in Fig. 11. The OMD consistently finds more predictability than the DMD in virtually all the modes at all dimensions and time lags. At a time lag of $\tau_0 = 1$ day damping time scales larger than 10 days and even beyond 12 days are observed. However, the correlations in the system at larger time lags decay faster than the initial slow decay and thus the time scales in the reduced models

TABLE 2. Damping time scales and oscillation periods of the eigenmodes of the EOF/DMD and OMD models at $\tau_0 = 4$ days for $J = 5$ and $J = 10$.

Model	Mode	Damping time scale (days)	Oscillation period (days)
EOF/DMD	1, 2	5.5	636.2
$J = 5$	3, 4	4.3	15.2
$\tau_0 = 4$ days	5	3.5	
OMD	1	8.7	
$J = 5$	2, 3	4.4	13.4
$\tau_0 = 4$ days	4, 5	4.0	95.4
EOF/DMD	1	7.0	
$J = 10$	2, 3	4.3	14.4
$\tau_0 = 4$ days	4, 5	4.2	71.8
	6, 7	3.6	17.5
	8, 9	3.2	92.9
	10	2.5	
OMD	1	8.6	
$J = 10$	2, 3	6.6	17.0
$\tau_0 = 4$ days	4, 5	6.4	43.4
	6	5.7	
	7, 8	5.2	11.9
	9, 10	4.5	83.4

gradually decrease with increasing time lag. Models fitted at small values of τ_0 overestimate the predictability of the system at larger lead times whereas models fitted at intermediate and large values of τ_0 underestimate the predictability at short lead times. Both for the DMD and the OMD the dynamics of the large-scale modes are markedly non-Markovian as evidenced by the strong dependence of the eigenvalue spectrum on the time lag. This still holds true when fixing the OMD subspace obtained for some time lag τ_0 and refitting the system matrix for other time lags (not shown). An uncertainty analysis as described in the methodology section was performed to test the significance of the findings (not shown). A correlation time of $\tau_c = 10$ corresponding to 5 days is chosen to reflect the serial correlation in the leading large-scale flow modes of the QG model. As evidenced by nonoverlapping 90% confidence intervals the differences in the time scales between DMD and OMD are significant at all values of τ_0 as is the failure of the tau test for both DMD and OMD for time lags $\tau_0 \leq 4$ days; the tau test passes for $\tau_0 > 4$ days within the uncertainty. The QG model probably has no oscillations with periods below 10 days; therefore, the Nyquist issue (Penland 2019) should not occur here and the failure of the tau test at short time lags is a genuine sign of non-Markovianity and/or nonlinearity, probably more non-Markovianity as nonlinear effects are small in the large-scale QG dynamics at these short time scales. Also the dependence of the OMD modes on the time lag τ_0 as shown in Figs. 6 and 7 is an indicator of non-Markovianity. Each of the Markov models obtained at a particular time lag is a different approximation to the non-Markovian dynamics. Memory effects in large-scale atmospheric or oceanic dynamics have also been found by other studies (Kravtsov et al. 2005; Franzke et al. 2009).

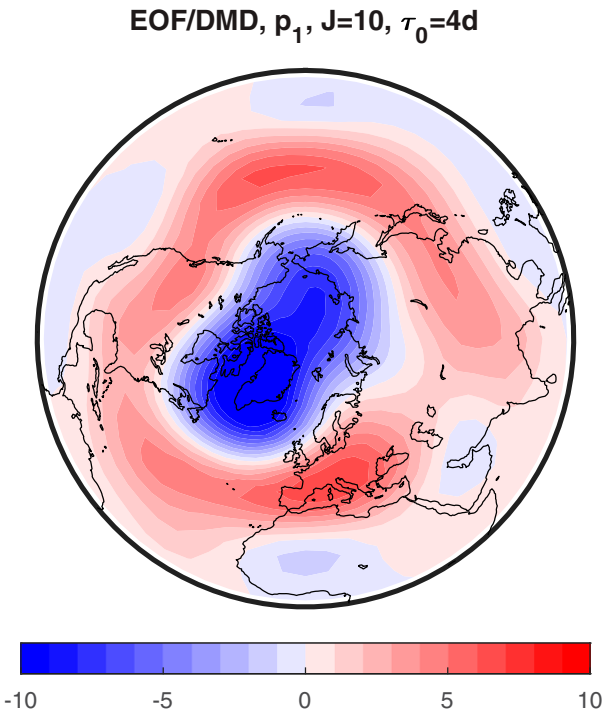
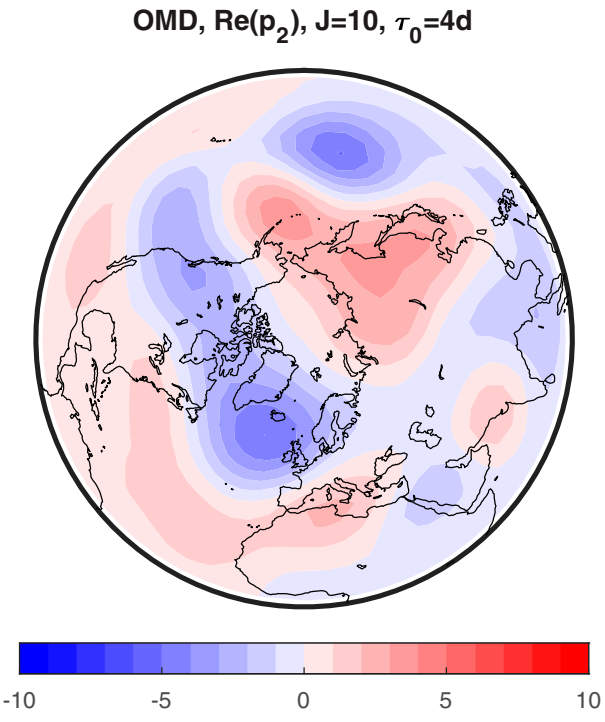
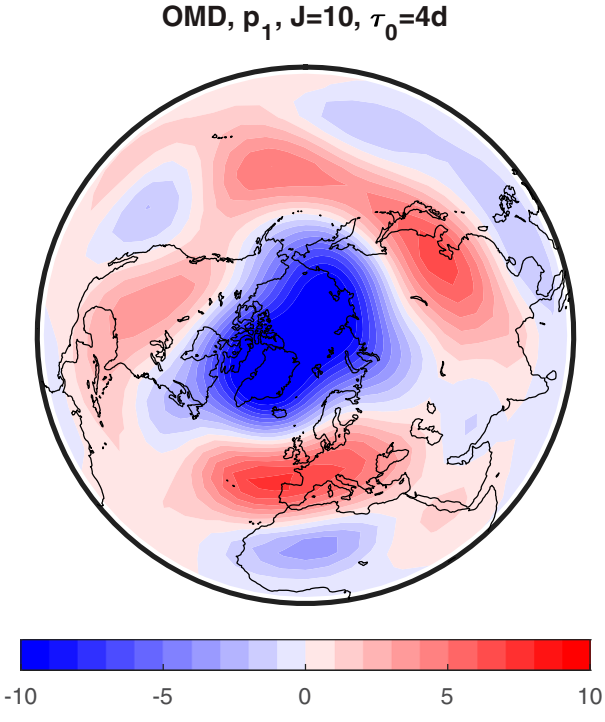


FIG. 9. Streamfunction anomaly patterns of the least damped eigenmode of (top) the EOF/DMD model and (bottom) the OMD model for $J = 10$ and $\tau_0 = 4$ days. The corresponding damping time scales are 7.0 and 8.6 days, respectively. The pattern correlation is 0.89. The pattern amplitudes are given by the standard deviation of the expansion coefficients in the dataset. Units are 10^6 m^2 s^{-1}.





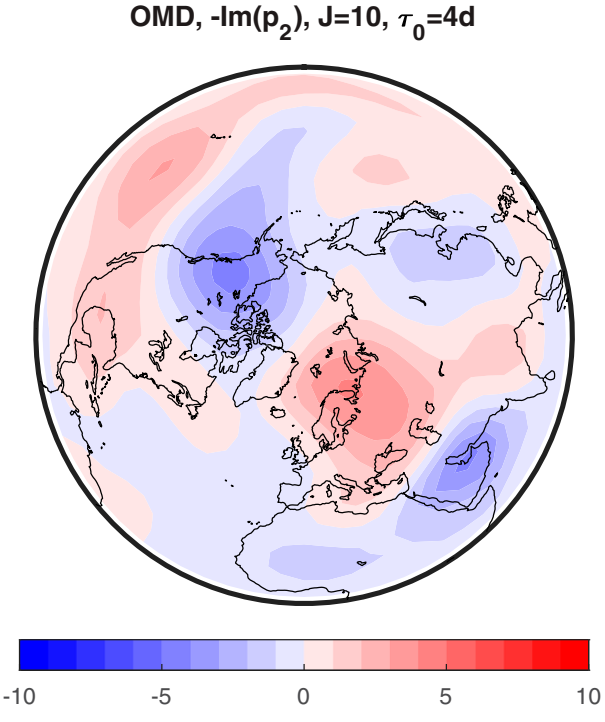


FIG. 10. Streamfunction anomaly patterns of (top) the real and (bottom) the negative imaginary part of the second eigenmode of the OMD model with $J = 10$ and $\tau_0 = 4$ days. The damping time scale is 6.6 days and the oscillation period is 17.0 days. The pattern amplitudes are given by the standard deviation of the expansion coefficients in the dataset. Units are 10^6 m^2 s^{-1}.

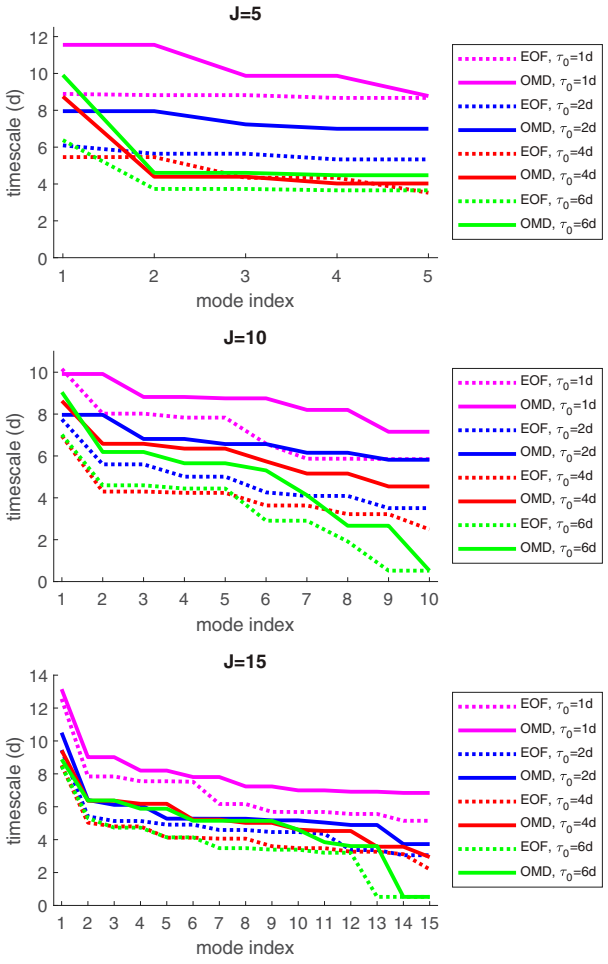


FIG. 11. Damping time scales of the eigenmodes of the EOF/DMD and OMD models for (top) $J = 5$, (middle) $J = 10$, and (bottom) $J = 15$.

c. *Predictive skill*

We now look at the prediction skill when iterating the same model over a range of lead times. For this purpose, a medium value of $\tau_0 = 4$ days for the time lag is fixed. All the reduced models calculated at that time lag are then transformed to the equivalent model with a time step of 12 h as discussed in section 2a and then iterated with a time step of 12 h over a range of lead times between 12 h and 12.5 days, that is, over 25 iterations. Figure 12 displays the anomaly correlation and the relative root mean squared error as a function of lead time for the model dimensions $J = 5$, $J = 10$, and $J = 15$. The forecast skills are calculated out of sample, dividing the dataset into two halves of 25 000 data points each and using the first half as learning dataset and the second half as verification dataset. Any of the OMD models outperforms any of the EOF/DMD models at all lead times. The improvement is largest for intermediate lead times around 4 days, the time lag for which the models are fitted. Among the OMD models the model with $J = 5$ performs best; this is in line with the explained variance explvar2 shown in Fig. 6. We remark that uncertainties in the estimation of the prediction skills are very small in the present data-rich setting and all of the findings are highly significant.

To put the prediction targets of the EOF/DMD and OMD models more on an equal footing Fig. 12 also shows the predictive skill only for the mode with largest variance in the respective subspace, that is, the mode $\mathbf{e}_1$ and $\mathbf{q}_1$, respectively. These two modes are still not exactly the same but they are very close as the mode $\mathbf{e}_1$ is almost fully contained in any OMD subspace with dimension $J \geq 5$ as shown in Fig. 7. Also the dependence of $\mathbf{q}_1$ on J is very weak once $J \geq 5$. The OMD models still clearly improve on the EOF/DMD models. Much of the skill of the OMD model is already present at $J = 5$; the EOF/DMD model catches up with increasing dimension. The OMD models with $J = 10$ and $J = 15$ outperform any of the EOF/DMD models at all lead times; the OMD model with $J = 5$ outperforms the EOF/DMD model with $J = 5$ at all lead times and outperforms any of the EOF/DMD models at lead times larger than 2 days.

d. *Role of nonnormality*

We now look at the degree of nonnormality present in the propagator matrices of the reduced models. Figure 13 shows the departure from normality for the EOF/DMD and OMD models as a function of dimension for various time lags. The nonnormality of the EOF/DMD models increases slowly with dimension for all time lags. The departure from normality of the OMD models is considerably stronger than that of the EOF/DMD models for virtually all dimensions and time lags. At $\tau_0 = 2$ days, this occurs only for dimensions $J \geq 15$ and thus does not seem to involve the most prominent large-scale modes of the QG model. At larger time lags, there is strong nonnormality already at small dimensions of the OMD model, starting at $J = 2$ for $\tau_0 = 4$ days and at $J = 4$ for $\tau_0 = 6$ days and $\tau_0 = 8$ days.

To investigate to what extent and how the nonnormality of the reduced models actually materializes in the nonlinear QG model, Fig. 14 displays boxplots of the growth/decay factors of various models in the dataset. The theoretical range of growth/decay factors as given by the smallest and largest singular values as well as the decay factor associated with the least damped eigenmode are also indicated. The distributions of growth factors of the OMD models are clearly positively skewed with long tails of large values. This is particularly prominent at $\tau_0 = 4$ days and $\tau_0 = 6$ days, less pronounced at $\tau_0 = 8$ days and not really developed at $\tau_0 = 2$ days. For some OMD models the observed growth/decay factor exceeds the decay factor of the least damped mode about 25% of the time, for $\tau_0 = 4$ days and $J = 5$ even more than 35% of the time. For $J = 5$, the full theoretical range of growth factors is actually realized in the dataset and the largest observed growth factors occur at $J = 5$. For higher dimensions the reduced model supports even larger growth factors but they are not realized in the data as the corresponding optimal excitation modes do not lie on the attractor of the nonlinear QG model. For the EOF/DMD models, the positive skew in the growth factor distributions is weaker, if any, both the upper bounds

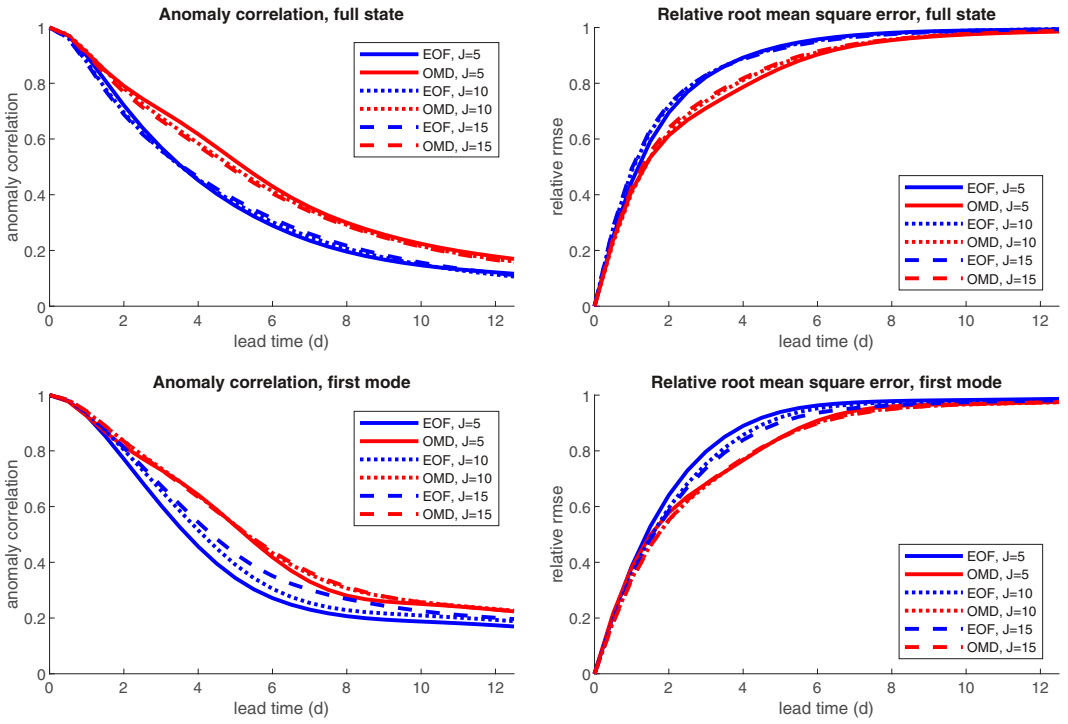


FIG. 12. Prediction skill of the EOF/DMD and OMD models as a function of lead time for various dimensions: (top left) anomaly correlation and (top right) relative root mean squared error for the whole state; (bottom left) anomaly correlation and (bottom right) relative root mean squared error for the mode with maximum variance.

and the actually realized growth factors are much smaller, and exceedances of the growth factor above the decay factor of the least damped mode occur much less often. We conclude that a considerable part of the improvement in predictive explained variance of the OMD models on the EOF/DMD models at the intermediate time lags $\tau_0 = 4$ days and $\tau_0 = 6$ days, and to a lesser extent at $\tau_0 = 8$ days, is due to better capturing the nonnormality of the linear operator and the associated nonmodal growth. This comes on top of the OMD modes being more persistent than the DMD modes at all time lags.

We now investigate the nonnormal growth of the reduced models used above in the prediction experiments, that is, we fix $\tau_0 = 4$ days. Figure 15 displays as a function of τ the maximum growth factor as given by the largest singular value of the propagator matrix $\mathbf{B}_{\tau_0}^{\tau/\tau_0}$ and the maximum growth factor actually realized in the dataset, that is, $\max_{n=1,\ldots,N-K}|\mathbf{B}_{\tau_0}^{\tau/\tau_0}\mathbf{z}_n|/|\mathbf{z}_n|$. For $J = 5$ we see substantial nonmodal growth in the linear operator which reaches its maximum at $\tau = 2.4$ days. Almost the full growth is actually realized in the data; the largest observed growth factor is 2.07 and occurs at $\tau = 2.3$ days. The behavior for $J = 10$ is similar but the full range of growth factors is not realized in the data and the maxima of the growth factors occur at larger lead times. The picture is largely in line with the results presented in Fig. 14. However, when using the model fitted at $\tau_0 = 4$ days for the whole range of lead times nonnormal growth occurs already at short lead times whereas the model fitted at $\tau_0 = 2$ days chooses more persistent modes and has only very little nonnormality (see Fig. 14).

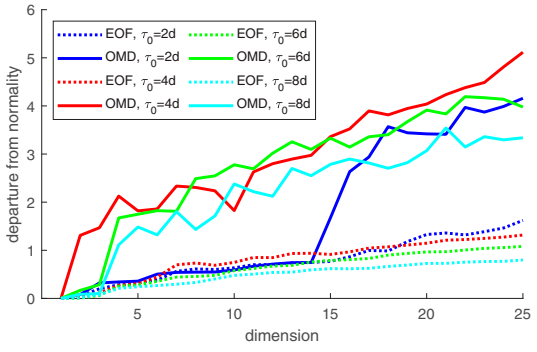


FIG. 13. Departure from normality of the propagator matrix of the EOF/DMD and OMD models as a function of dimension for various time lags.

Figure 16 displays the initial streamfunction anomaly state projected onto OMD space and the pattern resulting from time evolution under the OMD model propagator at a lead time of 2.3 days corresponding to the data point with the largest realized growth factor, 2.07, for $J = 5$ and $\tau_0 = 4$ days (cf. Fig. 15). The initial state has rather small-scale structure with few pronounced features. It does not resemble any of the well-known teleconnection patterns and it is also a low-variance pattern, accounting for only 2.2% of the streamfunction variance. The evolved pattern is a high-variance pattern combining elements of the negative phases of the Pacific–North America (PNA) patterns, the Arctic Oscillation (AO),

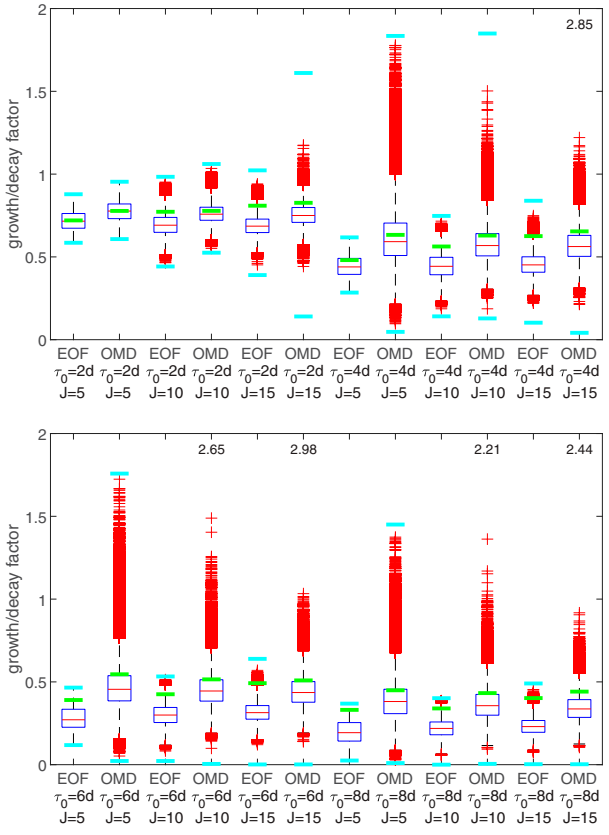


FIG. 14. Box-and-whisker plots of the growth/decay factors of the reduced models in the dataset of the QG model for (top) $\tau_0 = 2$ days and $\tau_0 = 4$ days as well as (bottom) $\tau_0 = 6$ days and $\tau_0 = 8$ days. The red line indicates the median. The bottom and top of the box give the 25th and 75th percentile, respectively, with the distance between the bottom and the top being the interquartile range. The whisker extends to the furthest observations not more than 1.5 times the interquartile range away from the bottom or top of the box. Observations beyond the whisker length are marked individually as outliers. The green line indicates the decay factor of the least damped eigenmode of the system matrix; the cyan lines give the largest and smallest singular values. If the largest singular value is beyond the plot range the value is given as a number.

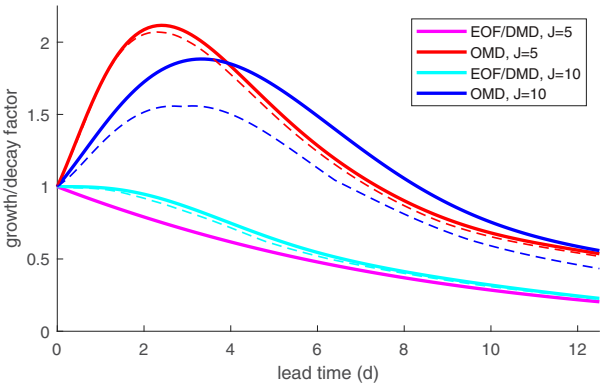


FIG. 15. Maximum growth/decay factor of the propagator matrix of EOF/DMD and OMD models of different dimensions as given by the largest singular value. The corresponding thin dashed curves of the same color give the maximum growth/decay factor actually realized in the dataset of the QG model. The reduced models are fitted at the time lag $\tau_0 = 4$ days.

and the North Atlantic Oscillation (NAO); it explains 9.1% of the variance. The pattern correlation between the evolved pattern and the actual state of the nonlinear QG model at a lead time of 4 days projected onto OMD space is 0.761, compared to an average anomaly correlation of 0.617 at that lead time. This confirms that the ability of the OMD models of capturing the nonnormal growth significantly contributes to the improvement in prediction skill on the EOF/DMD models.

e. *Non-Markovian modeling*

In view of the non-Markovian dynamics of the large-scale flow modes we briefly look at non-Markovian modeling using vector autoregressive (VAR) processes of higher order. A VAR model of order M, or VAR(M) model, is given as

$$\mathbf{z}_{n+1} = \sum_{m=0}^{M-1} \mathbf{F}_m \mathbf{z}_{n-m} + \boldsymbol{\xi}_n, \tag{59}$$

with $J \times J$ coefficient matrices $\mathbf{F}_m$ and again ξ_n being column vectors of length J of white zero mean Gaussian noises. Introducing the time-delay vectors $\mathbf{v}_n = (\mathbf{z}_n^{\mathrm{T}}, \mathbf{z}_{n-1}^{\mathrm{T}}, \ldots, \mathbf{z}_{n-M+1}^{\mathrm{T}})^{\mathrm{T}}$ for $n = M, \ldots, N-1$ the least squares and maximum likelihood estimator for the coefficient matrices is

$$(\mathbf{F}_0 \quad \mathbf{F}_1 \quad \cdots \quad \mathbf{F}_{M-1}) = \mathbf{H}_1 \mathbf{H}_0^{-1}, \tag{60}$$

with the covariance matrices

$$\mathbf{H}_0 = \frac{1}{N-M} \sum_{n=M}^{N-1} \mathbf{v}_n \mathbf{v}_n^{\mathrm{T}}, \tag{61}$$

$$\mathbf{H}_1 = \frac{1}{N-M} \sum_{n=M}^{N-1} \mathbf{z}_{n+1} \mathbf{v}_n^{\mathrm{T}}. \tag{62}$$

The VAR(M) model for $\mathbf{z}_n$ can be equivalently written as a VAR(1) model for $\mathbf{v}_n$. We here restrict our attention to models with prediction time lag and time delays equal to the sampling interval of the time series. Again, the dataset is divided into a learning dataset and a verification dataset of equal length in order to evaluate an out-of-sample prediction skill.

Figure 17 shows the anomaly correlation and the relative root mean squared error for various VAR models based on EOFs/DMD and OMD for $J = 5$. For comparison, the prediction skill of the Markovian models considered above is also indicated again. The higher-order VAR models outperform the Markovian models at all lead times for both EOFs/DMD and OMD; the improvement is particularly pertinent at short

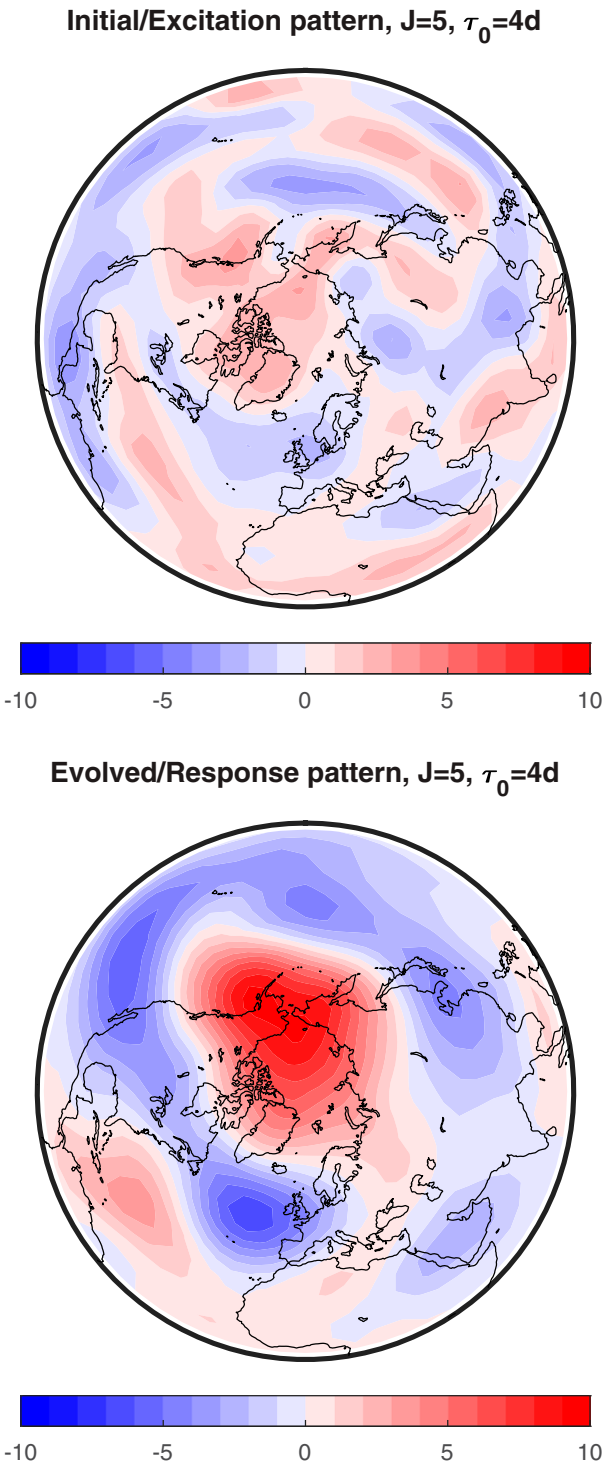


FIG. 16. (top) Initial/excitation and (bottom) evolved/response streamfunction anomaly pattern of the data point with the largest growth factor for the OMD model with $J = 5$ and $\tau_0 = 4$ days. The pattern amplitudes are given by the standard deviation of the expansion coefficients in the dataset of the QG model. Units are $10^6\ \mathrm{m}^2\ \mathrm{s}^{-1}$.

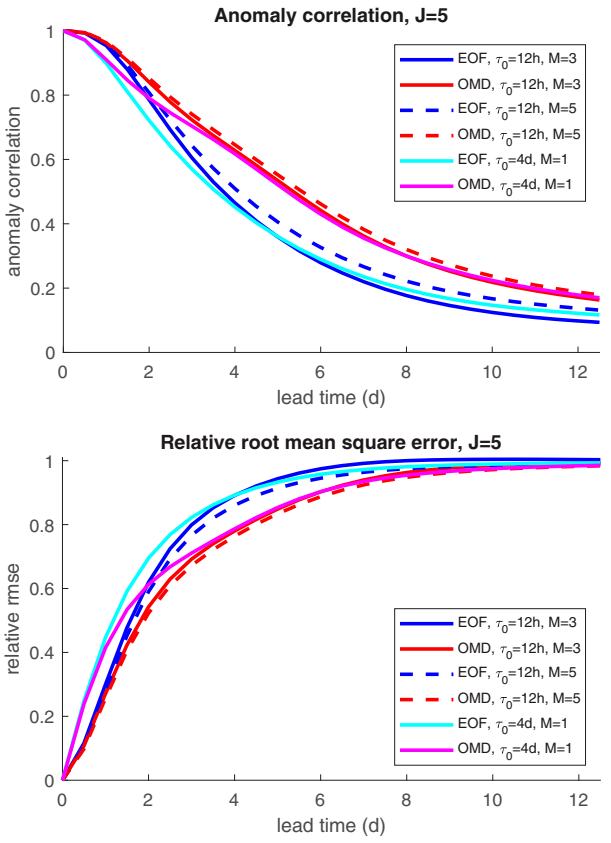


FIG. 17. Prediction skill of non-Markovian EOF/DMD and OMD models for $J = 5$: (top) Anomaly correlation and (bottom) relative root mean squared error.

lead times. Expectedly, the gain from the non-Markovian dynamics is slightly smaller for the OMD modes as these are optimized for Markovian dynamics. The prediction skill does not improve further beyond the order $M = 5$; for the OMD modes almost the full prediction skill is already there at the order $M = 3$. The non-Markovian models based on OMD modes have a higher predictive skill than those based on EOF/DMD modes at all lead times and the advantage is most prominent at intermediate lead times, similarly to the result above for the Markovian models. Again, the uncertainties in the prediction skills are very small given the large amount of data available here and all of the above statements are highly significant.

The question arises how much additional skill could be gained from an extension of the OMD to non-Markovian dynamics, that is, by optimizing the modes under the VAR setting. This would require a new algorithm; note that it is not possible to just apply the existing OMD algorithm to the VAR(1) description in time-delay space as the projection patterns need to be constrained to be the same for all of the delay times. Moreover, it may be worth considering VAR models with different prediction time lags and time delays. These issues will be addressed in a separate study.

5. Conclusions

The complete linear inverse modeling problem involving simultaneous optimization of the principal subspace and the system matrix using optimal mode decomposition (OMD) was investigated. In two simple examples with known linear system dynamics it is found that the OMD approach considerably improves on the conventional EOF/DMD approach in terms of system identification if modal predictability, that is, persistence is rather weak, if the excitation of the eigenmodes is nonuniform, and if there is pronounced nonnormality in the linear operator.

The OMD technique was then explored in the context of an intermediate-complexity atmospheric model with realistic mean state and variability. It finds more predictability than the DMD. At all time lags, this is due to the OMD modes being more persistent than the EOF/DMD modes; at intermediate lead times between 3 and 6 days the advantage of the OMD is largest and stems also from better modeling the nonnormality of the linear dynamics which is a major source of predictability here.

The dynamics of the large-scale modes are found to be considerably non-Markovian. This result calls for an extension of OMD to non-Markovian systems in order to find optimal patterns to model the large-scale dynamics with a vector autoregressive (VAR) model of higher order. However, when using the OMD modes optimized for Markovian dynamics in a non-Markovian model they already outperform the EOF/DMD modes also in this setting.

The OMD appears to be a very attractive candidate for a Markovian or non-Markovian ENSO prediction model as the ENSO phenomenon is approximately linear over quite long time scales involving strongly nonnormal linear operators. Another prediction task might be the Madden–Julian oscillation (MJO).

The OMD modes could also be used as a basis for nonlinear stochastic modeling of large-scale atmospheric, oceanic, or climate processes although they are not strictly optimized for this purpose. In a nonlinear reduced model of large-scale atmospheric dynamics given by projection of the quasigeostrophic equations of motion it was found that most of the improvement obtained from optimizing the basis functions actually stems from a better representation of the linear operator (Kwasniok 2007).

Other applications of the OMD might be reduced-rank data assimilation (Farrell and Ioannou 2001b; Mitchell and Gottwald 2012) and the generation of initial perturbations for ensemble prediction (Demaeyer et al. 2022).

Acknowledgments. This research was supported by the Natural Environment Research Council (NERC) Grant NE/N008693/1 and in part by the National Science Foundation (NSF) under Grant NSF PHY-1748958. The author would like to thank Cécile Penland and two anonymous reviewers for their useful comments which helped improve the quality of the manuscript.

APPENDIX

The OMD Algorithm

Introducing the residuals

$$\mathbf{r}_n = \mathbf{y}_{n+K} - \mathbf{Q}\mathbf{B}_K\mathbf{Q}^{\mathrm{T}}\mathbf{y}_n, \quad n = 1, \ldots, N-K, \tag{A1}$$

the objective function of Eq. (38) is

$$\begin{aligned}
F &= \sum_{n=1}^{N-K} |\mathbf{r}_n|^2 \\
&= \sum_{n=1}^{N-K} |\mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{r}_n + (\mathbf{I} - \mathbf{Q}\mathbf{Q}^{\mathrm{T}})\mathbf{r}_n|^2 \\
&= \sum_{n=1}^{N-K} |\mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{r}_n|^2 + \sum_{n=1}^{N-K} |(\mathbf{I} - \mathbf{Q}\mathbf{Q}^{\mathrm{T}})\mathbf{r}_n|^2 \\
&= \sum_{n=1}^{N-K} |\mathbf{Q}^{\mathrm{T}}\mathbf{r}_n|^2 + \sum_{n=1}^{N-K} |\mathbf{y}_{n+K} - \mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{y}_{n+K}|^2 \\
&= \sum_{n=1}^{N-K} |\mathbf{z}_{n+K} - \mathbf{B}_K\mathbf{z}_n|^2 + \sum_{n=1}^{N-K} |\mathbf{y}_{n+K} - \mathbf{Q}\mathbf{z}_{n+K}|^2 \\
&= \sum_{n=1}^{N-K} |\mathbf{z}_{n+K} - \mathbf{B}_K\mathbf{z}_n|^2 + \sum_{n=1}^{N-K} (\mathbf{y}_{n+K} - \mathbf{Q}\mathbf{z}_{n+K})^{\mathrm{T}}(\mathbf{y}_{n+K} - \mathbf{Q}\mathbf{z}_{n+K}) \\
&= \underbrace{\sum_{n=1}^{N-K} |\mathbf{z}_{n+K} - \mathbf{B}_K\mathbf{z}_n|^2}_{=F_{\mathrm{dyn}}} + \underbrace{\sum_{n=1}^{N-K} |\mathbf{y}_{n+K}|^2 - \sum_{n=1}^{N-K} |\mathbf{z}_{n+K}|^2}_{=F_{\mathrm{pr}}}.
\end{aligned} \tag{A2}$$

The anomaly dataset $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ is arranged in the data matrix

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 & \cdots & \mathbf{y}_{N-K} \end{pmatrix} \in \mathbb{R}^{D\times(N-K)} \tag{A3}$$

and the lagged data matrix

$$\tilde{\mathbf{Y}} = \begin{pmatrix} \mathbf{y}_{K+1} & \cdots & \mathbf{y}_N \end{pmatrix} \in \mathbb{R}^{D\times(N-K)}, \tag{A4}$$

with the data point vectors as columns. The corresponding projected data matrices are

$$\mathbf{Z} = \mathbf{Q}^{\mathrm{T}}\mathbf{Y} = \begin{pmatrix} \mathbf{z}_1 & \cdots & \mathbf{z}_{N-K} \end{pmatrix} \in \mathbb{R}^{J\times(N-K)} \tag{A5}$$

and

$$\tilde{\mathbf{Z}} = \mathbf{Q}^{\mathrm{T}}\tilde{\mathbf{Y}} = \begin{pmatrix} \mathbf{z}_{K+1} & \cdots & \mathbf{z}_N \end{pmatrix} \in \mathbb{R}^{J\times(N-K)}. \tag{A6}$$

In this notation the system matrix is

$$\mathbf{B}_K = \tilde{\mathbf{Z}}\mathbf{Z}^{\mathrm{T}}(\mathbf{Z}\mathbf{Z}^{\mathrm{T}})^{-1}, \tag{A7}$$

with the covariance matrices in the reduced subspace being given by

$$\mathbf{\Gamma}_0 = \frac{1}{N-K}\mathbf{Z}\mathbf{Z}^{\mathrm{T}} \tag{A8}$$

and

$$\mathbf{\Gamma}_K = \frac{1}{N-K}\tilde{\mathbf{Z}}\mathbf{Z}^{\mathrm{T}}. \tag{A9}$$

For any matrix $\mathbf{X} \in \mathbb{R}^{L_1\times L_2}$ the Frobenius norm is defined by

$$\|\mathbf{X}\| = \sqrt{\sum_{j=1}^{L_1}\sum_{k=1}^{L_2} X_{jk}^2}. \tag{A10}$$

It can be written as $\|\mathbf{X}\|^2 = \mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) = \mathrm{tr}(\mathbf{X}\mathbf{X}^{\mathrm{T}})$ where tr denotes the trace of a square matrix. Moreover, we have $\mathrm{tr}(\mathbf{X}_1\mathbf{X}_2) = \mathrm{tr}(\mathbf{X}_2\mathbf{X}_1)$ for any $\mathbf{X}_1 \in \mathbb{R}^{L_1\times L_2}$ and $\mathbf{X}_2 \in \mathbb{R}^{L_2\times L_1}$.

We now look at algorithms for optimizing the OMD objective function. The discussion largely follows Goulart et al. (2012) and Wynn et al. (2013) with some modifications and simplifications increasing the speed and efficiency of the algorithms.

a. The alternating update algorithm

The matrix $\mathbf{Z}$ has a compact singular value decomposition

$$\mathbf{Z} = \mathbf{U}_\mathbf{Z}\mathbf{S}_\mathbf{Z}\mathbf{V}_\mathbf{Z}^\mathrm{T}, \tag{A11}$$

with $\mathbf{U}_\mathbf{Z} \in \mathbb{R}^{J\times J}$, $\mathbf{S}_\mathbf{Z} \in \mathbb{R}^{J\times J}$, and $\mathbf{V}_\mathbf{Z} \in \mathbb{R}^{(N-K)\times J}$. We introduce the auxiliary matrix

$$\begin{aligned}\mathbf{W} &= \mathbf{Y}^\mathrm{T}\mathbf{Q}(\mathbf{Q}^\mathrm{T}\mathbf{Y}\mathbf{Y}^\mathrm{T}\mathbf{Q})^{-1}\mathbf{Q}^\mathrm{T}\mathbf{Y} = \mathbf{Z}^\mathrm{T}(\mathbf{Z}\mathbf{Z}^\mathrm{T})^{-1}\mathbf{Z} \\ &= \mathbf{V}_\mathbf{Z}\mathbf{V}_\mathbf{Z}^\mathrm{T} \in \mathbb{R}^{(N-K)\times(N-K)}\end{aligned} \tag{A12}$$

and observe that $\mathbf{W}$ is symmetric ($\mathbf{W} = \mathbf{W}^\mathrm{T}$) and idempotent ($\mathbf{W}^2 = \mathbf{W}$); it is actually a projector matrix, projecting onto the right singular vectors of $\mathbf{Z}$ corresponding to the nonzero singular values. The identity

$$\mathbf{Q}^\mathrm{T}\tilde{\mathbf{Y}}\mathbf{W} = \mathbf{B}_K\mathbf{Z} \tag{A13}$$

holds. We now get

$$\begin{aligned}F &= \|\tilde{\mathbf{Y}} - \mathbf{Q}\mathbf{B}_K\mathbf{Q}^\mathrm{T}\mathbf{Y}\|^2 \\ &= \|\tilde{\mathbf{Y}} - \mathbf{Q}\mathbf{Q}^\mathrm{T}\tilde{\mathbf{Y}}\mathbf{Y}^\mathrm{T}\mathbf{Q}(\mathbf{Q}^\mathrm{T}\mathbf{Y}\mathbf{Y}^\mathrm{T}\mathbf{Q})^{-1}\mathbf{Q}^\mathrm{T}\mathbf{Y}\|^2 \\ &= \|\tilde{\mathbf{Y}} - \mathbf{Q}\mathbf{Q}^\mathrm{T}\tilde{\mathbf{Y}}\mathbf{W}\|^2 \\ &= \mathrm{tr}[(\tilde{\mathbf{Y}} - \mathbf{Q}\mathbf{Q}^\mathrm{T}\tilde{\mathbf{Y}}\mathbf{W})(\tilde{\mathbf{Y}} - \mathbf{Q}\mathbf{Q}^\mathrm{T}\tilde{\mathbf{Y}}\mathbf{W})^\mathrm{T}] \\ &= \mathrm{tr}[(\tilde{\mathbf{Y}} - \mathbf{Q}\mathbf{Q}^\mathrm{T}\tilde{\mathbf{Y}}\mathbf{W})(\tilde{\mathbf{Y}}^\mathrm{T} - \mathbf{W}\tilde{\mathbf{Y}}^\mathrm{T}\mathbf{Q}\mathbf{Q}^\mathrm{T})] \\ &= \mathrm{tr}(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^\mathrm{T}) - \underbrace{\mathrm{tr}[\mathbf{Q}(\mathbf{Q}^\mathrm{T}\tilde{\mathbf{Y}}\mathbf{W}\tilde{\mathbf{Y}}^\mathrm{T})]}_{=\mathrm{tr}(\mathbf{Q}^\mathrm{T}\tilde{\mathbf{Y}}\mathbf{W}\tilde{\mathbf{Y}}^\mathrm{T}\mathbf{Q})} - \underbrace{\mathrm{tr}[(\tilde{\mathbf{Y}}\mathbf{W}\tilde{\mathbf{Y}}^\mathrm{T}\mathbf{Q})\mathbf{Q}^\mathrm{T}]}_{=\mathrm{tr}(\mathbf{Q}^\mathrm{T}\tilde{\mathbf{Y}}\mathbf{W}\tilde{\mathbf{Y}}^\mathrm{T}\mathbf{Q})} \\ &\quad + \underbrace{\mathrm{tr}[\mathbf{Q}(\mathbf{Q}^\mathrm{T}\tilde{\mathbf{Y}}\mathbf{W}\tilde{\mathbf{Y}}^\mathrm{T}\mathbf{Q}\mathbf{Q}^\mathrm{T})]}_{=\mathrm{tr}(\mathbf{Q}^\mathrm{T}\tilde{\mathbf{Y}}\mathbf{W}\tilde{\mathbf{Y}}^\mathrm{T}\mathbf{Q})} \\ &= \mathrm{tr}(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^\mathrm{T}) - \mathrm{tr}(\mathbf{Q}^\mathrm{T}\tilde{\mathbf{Y}}\mathbf{W}\tilde{\mathbf{Y}}^\mathrm{T}\mathbf{Q}) \\ &= \|\tilde{\mathbf{Y}}\|^2 - \|\mathbf{Q}^\mathrm{T}\tilde{\mathbf{Y}}\mathbf{W}\|^2 \\ &= \|\tilde{\mathbf{Y}}\|^2 - \|\mathbf{B}_K\mathbf{Z}\|^2 \\ &= \sum_{n=1}^{N-K}|\mathbf{y}_{n+K}|^2 - \sum_{n=1}^{N-K}|\mathbf{B}_K\mathbf{z}_n|^2,\end{aligned} \tag{A14}$$

which verifies the second equality in Eqs. (43) and (44). We also have the identity

$$\begin{aligned}\|\mathbf{B}_K\mathbf{Z}\|^2 &= \mathrm{tr}(\mathbf{B}_K\mathbf{Z}\mathbf{Z}^\mathrm{T}\mathbf{B}_K^\mathrm{T}) = (N-K)\mathrm{tr}(\mathbf{\Gamma}_K\mathbf{\Gamma}_0^{-1}\mathbf{\Gamma}_0\mathbf{B}_K^\mathrm{T}) \\ &= (N-K)\mathrm{tr}(\mathbf{\Gamma}_K\mathbf{B}_K^\mathrm{T}).\end{aligned} \tag{A15}$$

Minimizing F is equivalent to maximizing the total explained variance

$$G = \|\mathbf{B}_K\mathbf{Z}\|^2 = \|\mathbf{Q}^\mathrm{T}\tilde{\mathbf{Y}}\mathbf{W}\|^2 = \|\mathbf{Q}^\mathrm{T}\tilde{\mathbf{Y}}\mathbf{V}_\mathbf{Z}\|^2. \tag{A16}$$

An iterative subspace projection method is used which updates $\mathbf{Q}$ and $\mathbf{V}_\mathbf{Z}$ in turn. Each iteration consists of two steps. In the first step, $\|\mathbf{Q}^\mathrm{T}\tilde{\mathbf{Y}}\mathbf{V}_\mathbf{Z}\|^2$ is maximized with respect to $\mathbf{Q}$, holding $\mathbf{V}_\mathbf{Z}$ fixed. The solution is given by taking the columns of $\mathbf{Q}$ as the left singular vectors of $\tilde{\mathbf{Y}}\mathbf{V}_\mathbf{Z}$. In the second step, $\mathbf{V}_\mathbf{Z}$ is updated from the singular value decomposition of $\mathbf{Z}$. The technique is a heuristic scheme rather than a systematic optimization algorithm as there is no guarantee of an increase of G in an iteration; retrograde steps are possible and no convergence can be established. However, for reasons outlined by Goulart et al. (2012), the method performs very well on typical practical problems and provides a very good approximation to the OMD modes. Moreover, it is fast as no gradients are formed and no line search is performed. The algorithm is run for a fixed number of iterations N_au and the pattern matrix $\mathbf{Q}$ with the highest value of G encountered is kept as the approximate solution.

b. The gradient ascent algorithm

The second optimization technique is a gradient ascent algorithm taking into account the underlying special structure of the problem here. The objective function G is maximized over the Grassmann manifold of J-dimensional subspaces of $\mathbb{R}^D$. Each point on the manifold is represented by infinitely many matrices $\mathbf{Q}$ satisfying $\mathbf{Q}^\mathrm{T}\mathbf{Q} = \mathbf{I}$ as there is the gauge freedom of an arbitrary real orthogonal transformation. Searches are performed along geodesics of the Grassmann manifold, that is, paths of shortest distance between two given subspaces. The gradient of G on the Grassmann manifold is given by (Edelman et al. 1998)

$$\nabla G = (\mathbf{I} - \mathbf{Q}\mathbf{Q}^\mathrm{T})\frac{\partial G}{\partial \mathbf{Q}}, \tag{A17}$$

with the matrix of partial derivatives (Wynn et al. 2013)

$$\frac{1}{2}\frac{\partial G}{\partial \mathbf{Q}} = \tilde{\mathbf{Y}}\mathbf{Z}^\mathrm{T}\mathbf{B}_K^\mathrm{T} + \mathbf{Y}\tilde{\mathbf{Z}}^\mathrm{T}\mathbf{B}_K - \mathbf{Y}\mathbf{Z}^\mathrm{T}\mathbf{B}_K^\mathrm{T}\mathbf{B}_K \tag{A18}$$

$$= (N-K)(\mathbf{C}_K\mathbf{Q}\mathbf{B}_K^\mathrm{T} + \mathbf{C}_K^\mathrm{T}\mathbf{Q}\mathbf{B}_K - \mathbf{C}_0\mathbf{Q}\mathbf{B}_K^\mathrm{T}\mathbf{B}_K). \tag{A19}$$

The geodesic curve passing through $\mathbf{Q}$ in the direction ∇G can be parameterized as (Edelman et al. 1998)

$$\overline{\mathbf{Q}}(\theta) = \mathbf{Q}\mathbf{V}\cos(\theta\mathbf{S})\mathbf{V}^\mathrm{T} + \mathbf{U}\sin(\theta\mathbf{S})\mathbf{V}^\mathrm{T}, \tag{A20}$$

where

$$\nabla G = \mathbf{U}\mathbf{S}\mathbf{V}^\mathrm{T} \tag{A21}$$

is a compact singular value decomposition of ∇G with $\mathbf{U} \in \mathbb{R}^{D\times J}$, $\mathbf{S} = \mathrm{diag}(s_1,\ldots,s_J) \in \mathbb{R}^{J\times J}$ and $\mathbf{V} \in \mathbb{R}^{J\times J}$. Here, $\cos(\theta\mathbf{S})$ is the diagonal matrix with the cosines of the diagonal elements of $\theta\mathbf{S}$ on the diagonal and analogously for $\sin(\theta\mathbf{S})$. Having arrived at an estimate $\mathbf{Q}^{(k)}$ after k iterations a simple inaccurate line search using backtracking of the step size is performed along the geodesic curve passing

through $\mathbf{Q}^{(k)}$ in the direction $\nabla G|_{\mathbf{Q}=\mathbf{Q}^{(k)}}$. We consider the sequence

$$\theta_i = \frac{\pi}{2^i \max_j s_j}, \tag{A22}$$

where i is a nonnegative integer. Let i^* be the smallest value such that $G[\overline{\mathbf{Q}}(\theta_{i^*})] > G[\mathbf{Q}^{(k)}]$. In the $(k+1)$th iteration the estimate of $\mathbf{Q}$ is updated from $\mathbf{Q}^{(k)}$ to $\mathbf{Q}^{(k+1)} = \overline{\mathbf{Q}}[\theta^{(k+1)}]$, where $\theta^{(k+1)}$ maximizes $G[\overline{\mathbf{Q}}(\theta_i)]$ over the set $\{\theta_0, \theta_1, \ldots, \theta_{i^*+1}\}$. The algorithm converges to a (local) maximum of G. The iteration is terminated as soon as

$$G[\mathbf{Q}^{(k+1)}] - G[\mathbf{Q}^{(k)}] < \varepsilon \|\tilde{\mathbf{Y}}\|^2, \tag{A23}$$

with some $\varepsilon > 0$. The additional constraint of Eq. (8) is not enforced in the algorithm; it is applied afterward.

The gradient ascent algorithm can be run in two versions which produce identical results but differ in computation time. To evaluate the objective function the first variant projects the dataset onto the current patterns according to Eqs. (A5) and (A6), obtains $\mathbf{\Gamma}_0$, $\mathbf{\Gamma}_K$, and $\mathbf{B}_K$ from Eqs. (A8), (A9), and (A7), and calculates G from Eq. (A15); the gradient of G is evaluated from Eq. (A18). This variant avoids ever computing the covariance matrices $\mathbf{C}_0$ and $\mathbf{C}_K$ in full state space. The second variant once computes $\mathbf{C}_0 = \mathbf{Y}\mathbf{Y}^{\mathrm{T}}/(N-K)$ and $\mathbf{C}_K = \tilde{\mathbf{Y}}\mathbf{Y}^{\mathrm{T}}/(N-K)$, and then uses Eqs. (25), (26), (24), (A15), and (A19); it avoids ever projecting the dataset onto the reduced subspace. The choice of variant depends on the relative sizes of J, D, and N as well as on for how many different values of J we need to compute the OMD.

The simple gradient ascent could be refined using a conjugate gradient algorithm (Edelman et al. 1998; Wynn et al. 2013); however, the improvement is found not to be really significant and the complication is therefore avoided here.

c. The hybrid algorithm

We here combine the alternating update and gradient ascent algorithms to a hybrid algorithm which provides the best practical performance. The alternating update algorithm is run with $N_{\mathrm{au}} = 5$, for example; the exact choice of this parameter is not important. The pattern matrix $\mathbf{Q}$ with the largest value of G encountered is kept and subsequently used as starting point for the gradient ascent algorithm which is run as described above until the stopping criterion of a relative increase in G smaller than $\varepsilon = 10^{-5}$ is satisfied.

The OMD objective function is not convex which opens the possibility of secondary maxima. However, the optimization problem appears to be rather well-behaved in practice. A canonical choice for the first guess to start the alternating update algorithm is given by the EOFs/DMD ($\mathbf{Q} = \mathbf{E}$). All of the results for the QG model are obtained with this strategy. For genuinely noisy datasets such as the two pedagogical examples it rarely happens that the optimization starting with the EOFs gets stuck in a local maximum. This can be resolved by also running one or two optimizations with random initial pattern matrix and keeping the solution with the largest value of G.

For very large D an approximate OMD can be obtained at reduced computation time by applying a static dimension reduction prior to the OMD algorithm. We choose an intermediate dimension D^* with $J < D^* < D$, project the dataset onto the leading D^* EOFs, apply the OMD to the projected dataset, and lift the result back to the original state space. If the dimension of state space is larger than the length of the dataset ($D > N$) one would apply without any approximation the prior dimension reduction with $D^* = N$ to reduce computation time.

Obviously, the OMD is computationally much more costly than the DMD. But note that it is still not an onerous task. For example, the computation time for the OMD of the QG model dataset with $N = 25\,000$, $D = 231$, and $J = 25$ is about 1–2 s on a standard PC; for $J = 10$ or $J = 15$ it is well below 1 s.

A MATLAB implementation of the hybrid algorithm is publicly available on GitHub: https://github/FKwasniok/OMD/.

REFERENCES

Absil, P. A., R. Mahony, and R. Sepulchre, 2008: *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 240 pp.

Allen, S., C. A. T. Ferro, and F. Kwasniok, 2020: Recalibrating wind-speed forecasts using regime-dependent ensemble model output statistics. *Quart. J. Roy. Meteor. Soc.*, **146**, 2576–2596, https://doi.org/10.1002/qj.3806.

Blumenthal, B., 1991: Predictability of a coupled ocean–atmosphere model. *J. Climate*, **4**, 766–784, https://doi.org/10.1175/1520-0442(1991)004<0766:POACOM>2.0.CO;2.

de la Iglesia, M. D., and E. G. Tabak, 2013: Principal dynamical components. *Commun. Pure Appl. Math.*, **66**, 48–82, https://doi.org/10.1002/cpa.21411.

DelSole, T., 1996: Can quasigeostrophic turbulence be modeled stochastically? *J. Atmos. Sci.*, **53**, 1617–1633, https://doi.org/10.1175/1520-0469(1996)053<1617:CQTBMS>2.0.CO;2.

——, 2007: Optimal perturbations in quasigeostrophic turbulence. *J. Atmos. Sci.*, **64**, 1350–1364, https://doi.org/10.1175/JAS3875.1.

Demaeyer, J., S. G. Penny, and S. Vannitsem, 2022: Identifying efficient ensemble perturbations for initializing subseasonal-to-seasonal prediction. *J. Adv. Model. Earth Syst.*, **14**, e2021MS002828, https://doi.org/10.1029/2021MS002828.

Edelman, A., T. A. Arias, and S. T. Smith, 1998: The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, **20**, 303–353, https://doi.org/10.1137/S0895479895290954.

Farrell, B. F., 1989: Optimal excitation of baroclinic waves. *J. Atmos. Sci.*, **46**, 1193–1206, https://doi.org/10.1175/1520-0469(1989)046<1193:OEOBW>2.0.CO;2.

——, and P. J. Ioannou, 1993: Stochastic dynamics of baroclinic waves. *J. Atmos. Sci.*, **50**, 4044–4057, https://doi.org/10.1175/1520-0469(1993)050<4044:SDOBW>2.0.CO;2.

——, and ——, 1996: Generalized stability theory. Part I: Autonomous operators. *J. Atmos. Sci.*, **53**, 2025–2040, https://doi.org/10.1175/1520-0469(1996)053<2025:GSTPIA>2.0.CO;2.

——, and ——, 2001a: Accurate low-dimensional approximation of the linear dynamics of fluid flow. *J. Atmos. Sci.*, **58**, 2771–2789, https://doi.org/10.1175/1520-0469(2001)058<2771:ALDAOT>2.0.CO;2.

——, and ——, 2001b: State estimation using a reduced-order Kalman filter. *J. Atmos. Sci.*, **58**, 3666–3680, https://doi.org/10.1175/1520-0469(2001)058<3666:SEUARO>2.0.CO;2.

Franzke, C., I. Horenko, A. J. Majda, and R. Klein, 2009: Systematic metastable atmospheric regime identification in an AGCM. *J. Atmos. Sci.*, **66**, 1997–2012, https://doi.org/10.1175/2009JAS2939.1.

Gardiner, C., 2010: *Stochastic Methods: A Handbook for the Natural and Social Sciences*. 4th ed. Springer, 468 pp.

Gavrilov, A., D. Mukhin, E. Loskutov, E. Volodin, A. Feigin, and J. Kurths, 2016: Method for reconstructing nonlinear modes with adaptive structure from multidimensional data. *Chaos*, **26**, 123101, https://doi.org/10.1063/1.4968852.

——, A. Seleznev, D. Mukhin, E. Loskutov, A. Feigin, and J. Kurths, 2019: Linear dynamical modes as new variables for data-driven ENSO forecast. *Climate Dyn.*, **52**, 2199–2216, https://doi.org/10.1007/s00382-018-4255-7.

Glover, K., 1984: An optimal Hankel-norm approximation of linear multivariable systems and their L^{∞}-error bounds. *Int. J. Control*, **39**, 1115–1193, https://doi.org/10.1080/00207178408933239.

Goulart, P. J., A. Wynn, and D. Pearson, 2012: Optimal mode decomposition for high dimensional systems. *51st IEEE Conf. on Decision and Control*, Maui, Hawaii, IEEE, 4965–4970, https://doi.org/10.1109/CDC.2012.6426995.

Hannachi, A., 2021: *Patterns Identification and Data Mining in Weather and Climate*. 1st ed. Springer, 624 pp.

Hasselmann, K., 1988: PIPs and POPs: The reduction of complex dynamical systems using principal interaction and oscillation patterns. *J. Geophys. Res.*, **93**, 11 015–11 020, https://doi.org/10.1029/JD093iD09p11015.

Henrici, P., 1962: Bounds for iterates, inverses, spectral variation and fields of values of non-normal matrices. *Numer. Math.*, **4**, 24–40, https://doi.org/10.1007/BF01386294.

Hirsh, S. M., K. D. Harris, J. N. Kutz, and B. W. Brunton, 2020: Centering data improves the dynamic mode decomposition. *SIAM J. Appl. Dyn. Syst.*, **19**, 1920–1955, https://doi.org/10.1137/19M1289881.

Jolliffe, I. T., 2002: *Principal Component Analysis*. 2nd ed. Springer, 502 pp.

Kravtsov, S., D. Kondrashov, and M. Ghil, 2005: Multilevel regression modeling of nonlinear processes: Derivation and applications to climatic variability. *J. Climate*, **18**, 4404–4424, https://doi.org/10.1175/JCLI3544.1.

Kubo, R., 1966: The fluctuation-dissipation theorem. *Rep. Prog. Phys.*, **29**, 255–284, https://doi.org/10.1088/0034-4885/29/1/306.

Kwasniok, F., 1996: The reduction of complex dynamical systems using principal interaction patterns. *Physica D*, **92**, 28–60, https://doi.org/10.1016/0167-2789(95)00280-4.

——, 1997: Optimal Galerkin approximations of partial differential equations using principal interaction patterns. *Phys. Rev. E*, **55**, 5365–5375, https://doi.org/10.1103/PhysRevE.55.5365.

——, 2004: Empirical low-order models of barotropic flow. *J. Atmos. Sci.*, **61**, 235–245, https://doi.org/10.1175/1520-0469(2004)061<0235:ELMOBF>2.0.CO;2.

——, 2007: Reduced atmospheric models using dynamically motivated basis functions. *J. Atmos. Sci.*, **64**, 3452–3474, https://doi.org/10.1175/JAS4022.1.

——, 2018: Detecting, anticipating, and predicting critical transitions in spatially extended systems. *Chaos*, **28**, 033614, https://doi.org/10.1063/1.5022189.

——, 2019: Fluctuations of finite-time Lyapunov exponents in an intermediate-complexity atmospheric model: A multivariate and large-deviation perspective. *Nonlinear Processes Geophys.*, **26**, 195–209, https://doi.org/10.5194/npg-26-195-2019.

Lütkepohl, H., 2005: *New Introduction to Multiple Time Series Analysis*. Springer, 556 pp.

Marshall, J., and F. Molteni, 1993: Toward a dynamical understanding of planetary-scale flow regimes. *J. Atmos. Sci.*, **50**, 1792–1818, https://doi.org/10.1175/1520-0469(1993)050<1792:TADUOP>2.0.CO;2.

Mezić, I., 2013: Analysis of fluid flows via spectral properties of the Koopman operator. *Annu. Rev. Fluid Mech.*, **45**, 357–378, https://doi.org/10.1146/annurev-fluid-011212-140652.

Mitchell, L., and G. A. Gottwald, 2012: Data assimilation in slow–fast systems using homogenized climate models. *J. Atmos. Sci.*, **69**, 1359–1377, https://doi.org/10.1175/JAS-D-11-0145.1.

Moore, B. C., 1981: Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Trans. Automat. Control*, **AC-26**, 17–31, https://doi.org/10.1109/TAC.1981.1102568.

Navarra, A., J. Tribbia, and S. Klus, 2021: Estimation of Koopman transfer operators for the equatorial Pacific SST. *J. Atmos. Sci.*, **78**, 1227–1244, https://doi.org/10.1175/JAS-D-20-0136.1.

Penland, C., 1989: Random forcing and forecasting using principal oscillation pattern analysis. *Mon. Wea. Rev.*, **117**, 2165–2185, https://doi.org/10.1175/1520-0493(1989)117<2165:RFAFUP>2.0.CO;2.

——, 2019: The Nyquist issue in linear inverse modeling. *Mon. Wea. Rev.*, **147**, 1341–1349, https://doi.org/10.1175/MWR-D-18-0104.1.

——, and M. Ghil, 1993: Forecasting Northern Hemisphere 700-mb geopotential height anomalies using empirical normal modes. *Mon. Wea. Rev.*, **121**, 2355–2372, https://doi.org/10.1175/1520-0493(1993)121<2355:FNHMGH>2.0.CO;2.

——, and T. Magorian, 1993: Prediction of Niño 3 sea surface temperatures using linear inverse modeling. *J. Climate*, **6**, 1067–1076, https://doi.org/10.1175/1520-0442(1993)006<1067:PONSST>2.0.CO;2.

——, and P. D. Sardeshmukh, 1995: The optimal growth of tropical sea surface temperature anomalies. *J. Climate*, **8**, 1999–2024, https://doi.org/10.1175/1520-0442(1995)008<1999:TOGOTS>2.0.CO;2.

Rowley, C. W., I. Mezić, S. Bagheri, P. Schlatter, and D. S. Henningson, 2009: Spectral analysis of nonlinear flows. *J. Fluid Mech.*, **641**, 85–113, https://doi.org/10.1017/S0022112009992059.

Sardeshmukh, P. D., and P. Sura, 2009: Reconciling non-Gaussian climate statistics with linear dynamics. *J. Climate*, **22**, 1193–1207, https://doi.org/10.1175/2008JCLI2358.1.

Schmid, P. J., 2007: Nonmodal stability theory. *Annu. Rev. Fluid Mech.*, **39**, 129–162, https://doi.org/10.1146/annurev.fluid.38.050304.092139.

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

Sura, P., M. Newman, C. Penland, and P. Sardeshmukh, 2005: Multiplicative noise and non-Gaussianity: A paradigm for atmospheric regimes. *J. Atmos. Sci.*, **62**, 1391–1409, https://doi.org/10.1175/JAS3408.1.

Trefethen, L. N., and M. Embree, 2005: *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. 1st ed. Princeton University Press, 624 pp.

von Storch, H., and F. W. Zwiers, 2002: *Statistical Analysis in Climate Research*. 1st ed. Cambridge University Press, 995 pp.

——, G. Bürger, R. Schnur, and J.-S. von Storch, 1995: Principal oscillation patterns: A review. *J. Climate*, **8**, 377–400, https://doi.org/10.1175/1520-0442(1995)008<0377:POPAR>2.0.CO;2.

Winkler, C. R., M. Newman, and P. D. Sardeshmukh, 2001: A linear model of wintertime low-frequency variability. Part I: Formulation and forecast skill. *J. Climate*, **14**, 4474–4494, https://doi.org/10.1175/1520-0442(2001)014<4474:ALMOWL>2.0.CO;2.

Wynn, A., D. S. Pearson, B. Ganapathisubramani, and P. J. Goulart, 2013: Optimal mode decomposition for unsteady flows. *J. Fluid Mech.*, **733**, 473–503, https://doi.org/10.1017/jfm.2013.426.

Xu, J.-S., and H. von Storch, 1990: Predicting the state of the Southern Oscillation using principal oscillation pattern analysis. *J. Climate*, **3**, 1316–1329, https://doi.org/10.1175/1520-0442(1990)003<1316:PTSOTS>2.0.CO;2.