Theor. Comput. Fluid Dyn.
DOI 10.1007/s00162-010-0184-8

ORIGINAL ARTICLE

Zhanhua Ma · Sunil Ahuja · Clarence W. Rowley

Reduced-order models for control of fluids using the eigensystem realization algorithm

Received: 23 June 2009 / Accepted: 19 November 2009

Abstract As sensors and flow control actuators become smaller, cheaper, and more pervasive, the use of feedback control to manipulate the details of fluid flows becomes increasingly attractive. One of the challenges is to develop mathematical models that describe the fluid physics relevant to the task at hand, while neglecting irrelevant details of the flow in order to remain computationally tractable. A number of techniques are presently used to develop such reduced-order models, such as proper orthogonal decomposition (POD), and approximate snapshot-based balanced truncation, also known as balanced POD. Each method has its strengths and weaknesses: for instance, POD models can behave unpredictably and perform poorly, but they can be computed directly from experimental data; approximate balanced truncation often produces vastly superior models to POD, but requires data from adjoint simulations, and thus cannot be applied to experimental data. In this article, we show that using the Eigensystem Realization Algorithm (ERA) (Juang and Pappa, J Guid Control Dyn 8(5):620–627, 1985) one can theoretically obtain exactly the same reduced-order models as by balanced POD. Moreover, the models can be obtained directly from experimental data, without the use of adjoint information. The algorithm can also substantially improve computational efficiency when forming reduced-order models from simulation data. If adjoint information is available, then balanced POD has some advantages over ERA: for instance, it produces modes that are useful for multiple purposes, and the method has been generalized to unstable systems. We also present a modified ERA procedure that produces modes without adjoint information, but for this procedure, the resulting models are not balanced, and do not perform as well in examples. We present a detailed comparison of the methods, and illustrate them on an example of the flow past an inclined flat plate at a low Reynolds number.

Keywords Flow control · Model reduction · Eigensystem Realization Algorithm · Balanced truncation

1 Introduction

In the last decade, substantial developments have been made in the area of model-based feedback flow control of fluids: for instance, see the recent reviews by [7,8,17]. In many applications, the focus is on how to apply

Communicated by T. Colonius

Z. Ma (✉), S. Ahuja, C. W. Rowley
Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544, USA
E-mail: zma@princeton.edu

S. Ahuja
E-mail: sahuja@princeton.edu

C. W. Rowley
E-mail: cwrowley@princeton.edu

Published online: 03 February 2010

actuation to maintain the flow around an equilibrium state of interest, for instance to delay transition to turbulence, or control separation on a bluff body. Linear control theory provides efficient tools for the analysis and design of feedback controllers. However, a significant challenge is that models for flow control problems are often very high dimensional, e.g., on the order of $\mathcal{O}(10^{5\sim 9})$, so large that it becomes computationally infeasible to apply linear control techniques. To address this issue, model reduction, by which a low-order approximate model is obtained, is therefore widely employed.

Several techniques are available for model reduction, many of which involve projection onto a set of modes. These may be global eigenmodes of a linearized operator [3], modes determined by proper orthogonal decomposition (POD) of a set of data [13], and various variants of POD, such as including shift modes [21]. An approach that is used widely for model reduction of linear systems is balanced truncation [20], and while this method is computationally intractable for systems with very large state spaces (dimension $\gtrsim 10^5$), recently an algorithm for computing approximate balanced truncation from snapshots of linearized and adjoint simulations has been developed [23] and successfully applied to a variety of high-dimensional flow control problems [1,4,14]. In this method, sometimes called balanced POD, one obtains two sets of modes (primal and adjoint) that are bi-orthogonal, and uses those for projection of the governing equations, just as in standard POD. Compared to most other methods, including POD, balanced truncation has key advantages, such as *a priori* error bounds, and guaranteed stability of the reduced-order model (if the original high-order system is stable). Balanced POD is an approximation of exact balanced truncation that is computationally tractable when the number of states is very large (for instance, up to 10^7), and typically produces models that are far more accurate than standard POD models. For instance, for a linearized channel flow investigated in [14], even though the first 5 POD modes capture over 99.7% of the energy in a dataset exhibiting large transient growth, a low-dimensional model obtained by projection onto these modes completely misses the transient growth. By contrast, a 3-mode balanced POD model captures the transient growth nearly perfectly; to do as well with a standard POD model, 17 modes were required.

The main steps of balanced POD include (a) taking snapshots from impulse responses of the linearized and adjoint systems, (b) computing a singular value decomposition (SVD) of a matrix formed from inner products of these snapshots, (c) constructing primal modes and adjoint modes from the resulting singular vectors, and (d) projecting the high-dimensional dynamics onto these modes.

While effective in many examples, balanced POD also faces challenges, especially for use with experimental data. The main restriction is that balanced POD requires snapshots of impulse-response data from an adjoint system, and adjoint information is not available for experiments.

To address this issue, here we describe an algorithm widely used for system identification and model reduction, the eigensystem realization algorithm (ERA) [15]. This algorithm has been used for problems in fluid mechanics, primarily as a system-identification technique for flow control [5,6], but also for model reduction [11,24]. Our main result, presented in Sect. 2, is that, for linear systems, ERA theoretically produces *exactly the same* reduced-order models as balanced POD, with no need of an adjoint system, and at an order of magnitude lower computational cost. This result implies that one can realize approximate balanced truncation even in experiments, and can also improve computational efficiency in simulations. We note that ERA and snapshot-based approximate balanced truncation have been applied together in a model reduction procedure in [10]. However, the theoretical equivalence between these two algorithms was not explored in that work.

We present a comparison between balanced POD and ERA, and show that if adjoint information is available, balanced POD also has its own advantages. In particular, balanced POD provides sets of bi-orthogonal primal/adjoint modes for the linear system, and can be directly generalized to unstable systems. In Sect. 3, we discuss a modified ERA algorithm that, in the absence of adjoint simulations, uses "pseudo-adjoint modes" to compute reduced-order models; however, this method does not produce balanced models, and performs worse than balanced POD in examples. In Sect. 4, we illustrate these methods using a numerical example of the two-dimensional flow past an inclined plate, at a low Reynolds number.

2 The eigensystem realization algorithm as snapshot-based approximate balanced truncation

In this section, we summarize the steps involved in approximate balanced truncation (balanced POD), and the Eigensystem Realization Algorithm, and show that they are equivalent.

Balanced truncation involves first constructing a coordinate transformation that "balances" a linear input–output system, in the sense that certain measures of controllability and observability (the Gramian matrices) become diagonal and identical [20]. A reduced-order model is then obtained by truncating the least controllable and observable states, which correspond to the smallest diagonal entries in the transformed system.

Unfortunately, the exact balanced truncation algorithm is not tractable for the large state dimensions encountered in fluid mechanics. However, an approximate, snapshot-based balanced truncation algorithm, referred to as Balanced Proper Orthogonal Decomposition (balanced POD) was proposed in [23], and has been used successfully in several examples [1,4,14].

The second technique, the ERA, has been used both for system identification and for model reduction, and it is well known that the models produced by ERA are approximately balanced [12,16]. Here we show further that, theoretically, ERA produces exactly the same reduced-order models as balanced POD. This equivalence indicates that ERA can be regarded as an approximate balanced truncation method, in the sense that, before truncation, it implicitly realizes a coordinate transformation under which a pair of approximate controllability and observability Gramians are exactly balanced. This feature distinguishes ERA from other model reduction methods that first realize truncations and then balance the reduced-order models. Note that in ERA the Gramians, and the balancing transformation itself, are never explicitly calculated, as we will also show in the following discussions.

For both techniques, we will consider a high-dimensional, stable, discrete-time linear system, described by

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) \\ y(k) &= Cx(k), \end{aligned} \tag{1}$$

where $k \in \mathbb{Z}$ is the time step index, $u(k) \in \mathbb{R}^p$ denotes a vector of inputs (for instance, actuators or disturbances), $y(k) \in \mathbb{R}^q$ a vector of outputs (for instance, sensor measurements, or simply quantities that one wishes to model), and $x(k) \in \mathbb{R}^n$ denotes the state variable (for instance, flow variables at all gridpoints of a simulation). These equations may arise, for instance, by discretizing the Navier-Stokes equations in time and space, and linearizing about a steady solution, as will be demonstrated in the example in Sect. 4. The goal is to obtain an approximate model that captures the same relationship between inputs u and outputs y, but with a much smaller state dimension:

$$\begin{aligned} x_r(k+1) &= A_r x_r(k) + B_r u(k) \\ y(k) &= C_r x_r(k) \end{aligned} \tag{2}$$

where the reduced state variable $x_r(k) \in \mathbb{R}^r$, $r \ll n$. We consider the discrete-time setting, because we are primarily interested in discrete-time data from simulations or experiments.

2.1 Snapshot-based approximate balanced truncation (balanced POD)

Here, we give only a brief overview of the balanced POD algorithm, and for details of the method, we refer the reader to [23]. The algorithm involves three main steps:

Step 1: *Collect snapshots.* Run impulse-response simulations of the primal system (1) and collect $m_c + 1$ snapshots of *states* $x(k)$ in $m_c P + 1$ steps:

$$X = [B \quad A^P B \quad A^{2P} B \quad \cdots \quad A^{m_c P} B], \tag{3}$$

where P is the sampling period. In addition, run impulse-response simulations for the adjoint system

$$z(k+1) = A^* z(k) + C^* v(k) \tag{4}$$

where the asterisk * stands for adjoint of a matrix, and collect $m_o + 1$ snapshots of states $z(k)$ in $m_o P + 1$ steps:

$$Y = [C^* \quad (A^*)^P C^* \quad (A^*)^{2P} C^* \cdots \quad (A^*)^{m_o P} C^*]. \tag{5}$$

Calculate the generalized Hankel matrix,

$$H = Y^* X. \tag{6}$$

Step 2: *Compute modes.* Compute the singular value decomposition of H:

$$H = U\Sigma V^* = [U_1 \; U_2] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^* \\ V_2^* \end{bmatrix} = U_1 \Sigma_1 V_1^* \tag{7}$$

where the diagonal matrix $\Sigma_1 \in \mathbb{R}^{n_1 \times n_1}$ is invertible and includes all non-zero singular values of $H, n_1 = \text{rank}(H)$, and $U_1^* U_1 = V_1^* V_1 = I_{n_1 \times n_1}$. Choose $r \leq n_1$. Let U_r and V_r denote the sub-matrices of U_1 and V_1 that include their first r columns, and Σ_r the first $r \times r$ diagonal block of Σ_1. Calculate

$$\Phi_r = X V_r \Sigma_r^{-\frac{1}{2}}; \quad \Psi_r = Y U_r \Sigma_r^{-\frac{1}{2}}. \tag{8}$$

where the columns of Φ_r and Ψ_r are, respectively, the first r primal and adjoint modes of system (1). The two sets of modes are bi-orthogonal: $\Psi_r^* \Phi_r = I_{r \times r}$.

Step 3: *Project dynamics.* The system matrices in the reduced-order model (2) are

$$A_r = \Psi_r^* A \Phi_r; \quad B_r = \Psi_r^* B; \quad C_r = C\Phi_r. \tag{9}$$

Note that the $n \times n$ controllability/observability Gramians are approximated by the matrices XX^* and YY^*. The reduced-order model (2) is obtained by considering a subspace $x = \Phi_r x_r$, and projecting the dynamics (1) onto this subspace using the adjoint modes given by Ψ_r. It was shown in [23] that Φ_r and Ψ_r, respectively, form the first r columns of the balancing transformation/inverse transformation that *exactly* balance the approximate controllability/observability Gramians XX^* and YY^*; see more discussion in Sect. 3.

2.2 The eigensystem realization algorithm

The ERA was proposed in [15] as a system identification and model reduction technique for linear systems. The algorithm follows three main steps [15,16]:

Step 1: Run impulse-response simulations/experiments of the system (1) for $(m_c + m_o)P + 2$ steps, where m_c and m_o, respectively, reflect how much effect is taken for considering controllability and observability, and P again is the sampling period. Collect the snapshots of the *outputs* y in the following pattern:

$$\Big(CB, \quad CAB, \quad CA^P B, \quad CA^{P+1} B, \quad \ldots \\ CA^{m_c P} B, \quad CA^{m_c P+1} B, \quad \ldots \quad CA^{(m_c+m_o)P} B, \quad CA^{(m_{c+m_o})P+1} B\Big). \tag{10}$$

The terms $CA^k B$ are commonly called *Markov parameters*. Construct a generalized Hankel matrix $H \in \mathbb{R}^{q(m_o+1) \times p(m_c+1)}$

$$H = \begin{bmatrix} CB & CA^P B & \cdots & CA^{m_c P} B \\ CA^P B & CA^{2P} B & \cdots & CA^{(m_c+1)P} B \\ \vdots & \vdots & \ddots & \vdots \\ CA^{m_o P} B & CA^{(m_o+1)P} B & \cdots & CA^{(m_c+m_o)P} B \end{bmatrix}. \tag{11}$$

Step 2: Compute SVD of H, exactly as in (7), to obtain U_1, V_1, Σ_1. Let $r \leq \text{rank}(H)$. Let U_r and V_r denote the sub-matrices of U_1 and V_1 that include their first r columns, and Σ_r the first $r \times r$ diagonal block of Σ_1.

Step 3: The reduced A_r, B_r and C_r in (2) are then defined as

$$\begin{aligned} A_r &= \Sigma_r^{-\frac{1}{2}} U_r^* H' V_r \Sigma_r^{-\frac{1}{2}}; \\ B_r &= \text{the first } p \text{ columns of } \Sigma_r^{\frac{1}{2}} V_r^*; \\ C_r &= \text{the first } q \text{ rows of } U_r \Sigma_r^{\frac{1}{2}} \end{aligned} \tag{12}$$

where

$$H' = \begin{bmatrix} CAB & CA^{P+1}B & \cdots & CA^{m_cP+1}B \\ \vdots & \vdots & \ddots & \vdots \\ CA^{m_oP+1}B & CA^{(m_o+1)P+1}B & \cdots & CA^{(m_c+m_o)P+1}B \end{bmatrix}, \tag{13}$$

which can again be constructed directly from the collected snapshots (10).

2.3 Theoretical equivalence between ERA and balanced POD

The main result of this article is summarized in the following proposition.

Proposition *The reduced system matrices A_r, B_r and C_r generated in balanced POD and ERA, respectively, by (9) and (12), are theoretically identical.*

Proof The first observation is that, with X and Y given by (3) and (5), the generalized Hankel matrices obtained in balanced POD and ERA, respectively, by (6) and (11), are theoretically identical. The theoretical equivalence between the two algorithms then follows: First, H' given in (13) satisfies $H' = Y^*AX$, which, together with the relation (8), implies the matrices A_r obtained in the two algorithms are identical. To show the equivalence of B_r, first note that when left-multiplied by $\Sigma_1^{-\frac{1}{2}}U_1^*$ on both sides, SVD (7) leads to $\Sigma_1^{-\frac{1}{2}}U_1^*H = \Sigma_1^{\frac{1}{2}}V_1^*$. By definition of U_r, V_r, Σ_r, it implies $\Sigma_r^{-\frac{1}{2}}U_r^*H = \Sigma_r^{\frac{1}{2}}V_r^*$. (Note that it does *not* imply $H = U_r\Sigma_rV_r^*$, since $U_rU_r^*$ is not the identity.) Thus, in balanced POD,

$$B_r = \Psi_r^*B = \Sigma_r^{-\frac{1}{2}}U_r^*Y^*B = \Sigma_r^{-\frac{1}{2}}U_r^*\begin{bmatrix} CB \\ CA^PB \\ \vdots \\ CA^{m_oP}B \end{bmatrix},$$

which equals the first p columns of $\Sigma_r^{-\frac{1}{2}}U_r^*H = \Sigma_r^{\frac{1}{2}}V_r^*$, which is the matrix B_r given by ERA. Similarly, the SVD (7) leads to $HV_1\Sigma_1^{-\frac{1}{2}} = U_1\Sigma_1^{\frac{1}{2}}$ and then $HV_r\Sigma_r^{-\frac{1}{2}} = U_r\Sigma_r^{\frac{1}{2}}$. Thus, in balanced POD, $C_r = C\Phi_r = CXV_r\Sigma_r^{-\frac{1}{2}}$, which equals the first q rows of $HV_r\Sigma_r^{-\frac{1}{2}} = U_r\Sigma_r^{\frac{1}{2}}$, the matrix C_r given by ERA. □

In practice, these two algorithms may generate slightly different reduced-order models, because the Hankel matrices calculated in the two algorithms are usually not exactly the same, due to small numerical inaccuracies in adjoint simulations, and/or in matrix multiplications needed to compute the sub-blocks in the Hankel matrices. In the following discussions, we compare these two algorithms in more detail.

2.4 Comparison between ERA and balanced POD

While ERA and balanced POD produce theoretically identical reduced-order models, the techniques differ in several important ways, both conceptually and computationally. Neither ERA nor balanced POD calculate Gramians explicitly, but balanced POD does construct approximate controllability and observability matrices X and Y^*, from which one calculates the generalized Hankel matrix H and balancing transformation. Balanced POD thus incurs additional computational cost, because one needs to construct the adjoint system (4), run adjoint simulations for Y, and then calculate each block of H by matrix multiplication. Thus we see that the advantages of ERA include:

1. **Adjoint-free:** ERA is a feasible balanced truncation method for experiments, since it needs only the output measurements from the response to an impulsive input. Note that ERA has been successfully applied in several flow control experiments [5,6], as a system-identification technique rather than a balanced-truncation method. In practice, input–output sensor responses are often collected by applying a broadband signal to the inputs, and the ARMARKOV method [2,18] can then be used to identify the Markov parameters, or even directly the generalized Hankel matrix, from the input–output data history.

Table 1 Comparison of the computational times required for various steps of the algorithms using balanced POD and ERA

Steps in computing reduced-order models	Approximate time (CPU hours)	
	Balanced POD	ERA
1. Linearized impulse response	2	4
2. Computation of POD modes	2	2
3. Adjoint impulse responses (10 in number)	30	–
4. Computation of the Hankel matrix	7	0.2
5. Singular value decomposition	0.05	0.05
6. Computation of modes	1	–
7. Computation of models	0.02	0.02

The times are given for a 10-mode output projected system. The Hankel matrices are constructed using 200 state-snapshots from each linearized and adjoint simulations for balanced POD, and 400 Markov parameters (outputs) for ERA

2. **Computational efficiency:** For large problems, typically the most computationally expensive component of computing balanced POD is constructing the generalized Hankel matrix H in (6), as this involves computing inner products of all of the (large) primal and adjoint snapshots with each other. ERA is significantly more efficient at constructing the matrix H in (11), since only the first row and last column of block matrices, i.e., the $(m_c + m_o + 1)$ Markov parameters, need to be obtained by matrix multiplication. All the other $m_c \times m_o$ block matrices in H are copies of other blocks, and need not be recomputed. For balanced POD, the matrix H is obtained by computing all the $(m_c + 1) \times (m_o + 1)$ matrix multiplications (inner products) between corresponding blocks in Y^* and X in (6). Thus, for example, if $m_c = m_0 = 200$, the computing time needed for constructing H in ERA will be about only 1% of that in balanced POD. See Table 1 for a detailed comparison on computational efficiency between balanced POD and ERA in the example of the flow past an inclined flat plate.

At the same time, balanced POD also provides its own advantages:

1. **Sets of bi-orthogonal primal/adjoint modes:** Balanced POD provides sets of bi-orthogonal primal/adjoint modes, the columns of Φ_r and Ψ_r. In comparison, without the adjoint system, ERA cannot provide the primal and adjoint modes. At best, the primal modes may be computed, using the first equation in (8), if the matrix X (3) is stored (in addition to the Markov parameters). But the adjoint modes cannot be computed without solutions of the adjoint system. In this sense, balanced POD incorporates more of the physics of the system (the two sets of bi-orthogonal modes), while ERA is purely based on input–output data of the system. The primal/adjoint modes together can be useful for system analysis and controller/observer design purposes in several ways: for instance, in flow control applications, a large-amplitude region from the most observable mode (the leading adjoint mode) can be a good location for actuator placement. Also, although balanced POD is a linear method, a nonlinear system can be projected onto these sets of modes to obtain a nonlinear low-dimensional model. For instance, the transformation $x = \Phi_r x_r, x_r = \Psi_r^* x$ can be employed to reduce a full-dimensional nonlinear model $\dot{x} = f(x)$ to a low-dimensional system $\dot{x}_r = \Psi_r^* f(\Phi_r x_r)$. Finally, if parameters (such as Reynolds number or Mach number) are present in the original equations, balanced POD can retain these parameters in the reduced-order models. When the values of parameters change, the reduced-order model by balanced POD may still be valid and perform well; see [14] for an application to linearized channel flow.
2. **Unstable systems:** Balanced POD has been extended to neutrally stable [19] and unstable systems [1]. In those cases, one first calculates the right/left eigenvectors corresponding to the neutral/unstable eigenvalues of the state-transition matrix A, using direct/adjoint simulations. Using these eigenvectors, the system is projected onto a stable subspace and then balanced truncation is realized for the stable subsystem. ERA for unstable systems is still an open problem, if adjoint operators are not available. However, we note that, once the stable subsystem is obtained, ERA can still be applied to it and efficiently realize its approximate balanced truncation.

ERA for systems with high-dimensional outputs. The method of output projection proposed in [23] makes it computationally feasible to realize approximate balanced truncation for systems with high-dimensional outputs—for instance, if one wishes to model the entire state x, say the flow field in the entire computational or experimental domain. This method involves projecting the outputs onto a small number of POD modes, determined from snapshots of y from the impulse-response dataset. This method can be directly incorporated into ERA as follows: First, run impulse response simulations of the original system and collect Markov parameters as usual. Then, compute the leading POD modes of the dataset of Markov parameters and stack them as columns of a matrix Θ. Left multiply those Markov parameters by Θ^* to project the outputs onto these POD modes. A generalized Hankel matrix is then constructed using these modified Markov parameters, and the usual steps of ERA follow.

3 A modified ERA method using pseudo-adjoint modes

We have seen that one of the drawbacks of ERA is that it does not provide modes that could be used, for instance, for projection of nonlinear dynamics, or to retain parameters in the models. More precisely, using ERA, one may still obtain primal modes $\Phi_1 = XV_1\Sigma_1^{-1/2}$ as in balanced POD (see (7–8)), as long as the state snapshots are collected and stored in X. But it is not possible to obtain the corresponding adjoint modes Ψ_1 necessary for projection, without performing adjoint simulations to gather snapshots for the matrix Y. This is a severe drawback, as adjoint solutions can be expensive to perform, and are not available for experimental data. One idea, proposed in [22], is to define a set of approximate adjoint modes using the Moore-Penrose pseudo-inverse of Φ_1:

$$\tilde{\Psi}_1 = \Phi_1(\Phi_1^*\Phi_1)^{-1}. \tag{14}$$

We will call the adjoint modes as defined above the *pseudo-adjoint modes* corresponding to the modes Φ_1. The system matrices of a r-dimensional reduced-order model ($r \leq \operatorname{rank}(H)$) generated by this approach then read

$$A_r = \tilde{\Psi}_r^* A \Phi_r; \quad B_r = \tilde{\Psi}_r^* B; \quad C_r = C\Phi_r, \tag{15}$$

where Φ_r and $\tilde{\Psi}_r$ are, respectively, the first $n \times r$ sub-blocks of Φ_1, $\tilde{\Psi}_1$.

While this idea does produce a set of modes that can be used for projection, we show in this section that, unfortunately, the resulting transformation is *not* a balancing transformation, and does not produce models that are an approximation to balanced truncation. In fact, the resulting models are closer to those produced by the the standard POD/Galerkin method: as with standard POD/Galerkin, the method performs well as long as the most controllable and most observable directions coincide. However, when these directions differ (as is the case for many problems of interest, including the example in Sect. 4), the method performs poorly. These systems in which controllable and observable directions do not coincide are precisely the systems for which balanced POD and ERA give the most improvement over the more traditional POD/Galerkin approach.

3.1 Transformed approximate Gramians

First, let us recall in what sense the model-reduction procedures described in Sect. 2 are approximations to balanced truncation. Suppose that we have an approximation of the controllability and observability Gramians, factored as

$$W_c = XX^*, \qquad W_o = YY^*, \tag{16}$$

where X and Y are the matrices of snapshots from (3) and (5). In balanced POD, we define the primal modes as columns of $\Phi_1 = XV_1\Sigma_1^{-\frac{1}{2}}$, and the adjoint modes as columns of $\Psi_1 = YU_1\Sigma_1^{-\frac{1}{2}}$, where U_1, V_1, and Σ_1 are defined in (7). We will assume in this section that the number of columns of X and Y (the number of snapshots, m_c and m_o, respectively) is smaller than the number of rows (the state dimension, n), which is always true for the large fluids systems of interest here.

Then balanced POD is an approximation to balanced truncation in the following sense: as shown in the appendix of [23] (the proof of Proposition 2), we may construct a full (invertible, $n \times n$) transformation

$$T = [\Phi_1 \ \Phi_2] \tag{17}$$

by choosing Φ_2 such that $\Psi_1^*\Phi_2 = 0$. That is, columns of Φ_2 are orthogonal to the adjoint modes, which are columns of Ψ_1. The inverse transformation then has the form

$$T^{-1} = \begin{bmatrix} \Psi_1^* \\ \Psi_2^* \end{bmatrix} \tag{18}$$

where Ψ_1 is the matrix of adjoint modes, and Ψ_2 is defined by (18). Then, Proposition 2 of [23] states that the transformed approximate Gramians (16) have the form

$$T^{-1}W_c(T^{-1})^* = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & M_1 \end{bmatrix}, \qquad T^*W_oT = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & M_2 \end{bmatrix}, \tag{19}$$

and furthermore the product of the approximate Gramians, in the transformed coordinates, is

$$T^{-1}W_cW_oT = \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix}. \tag{20}$$

In this sense, the transformation T balances the approximate Gramians as closely as possible: the Gramians are block diagonal, and the upper-left blocks are equal and diagonal. Furthermore, all of the states in the lower-right block (i.e., involving M_1 and M_2 above) are either unobservable or uncontrollable, as they do not appear in the product of the Gramians.

However, if the pseudo-adjoint modes $\tilde{\Psi}_1$ are used in place of the true adjoint modes Ψ_1, then this result does not hold, as we now show. Note that, in order for the first block of rows of T^{-1} to equal $\tilde{\Psi}_1^*$, we must now define

$$\tilde{T} = [\Phi_1 \; \tilde{\Phi}_2] \tag{21}$$

where $\tilde{\Psi}_1^*\tilde{\Phi}_2 = 0$. Since the range of $\tilde{\Psi}_1$ equals the range of Φ_1, this is then equivalent to choosing $\tilde{\Phi}_2$ such that its columns are orthogonal to the columns of Φ_1 (the *primal* modes), while when the "true" adjoint modes are used, columns of Φ_2 are chosen to be orthogonal to the *adjoint* modes Ψ_1.

Defining $\tilde{\Psi}_2$ by

$$\tilde{T}^{-1} = \begin{bmatrix} \tilde{\Psi}_1^* \\ \tilde{\Psi}_2^* \end{bmatrix}, \tag{22}$$

one can then show that, as long as $\operatorname{rank}(X) \le \operatorname{rank}(Y)$,[1] the transformed Gramians have the form

$$\tilde{T}^{-1}W_c(\tilde{T}^{-1})^* = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \tilde{M}_1 \end{bmatrix}, \qquad \tilde{T}^*W_o\tilde{T} = \begin{bmatrix} \Sigma_1 & M_3 \\ M_3^* & \tilde{M}_2 \end{bmatrix}, \qquad \tilde{T}^{-1}W_cW_o\tilde{T} = \begin{bmatrix} \Sigma_1^2 & \Sigma_1 M_3 \\ \tilde{M}_1 M_3^* & 0 \end{bmatrix}, \tag{23}$$

with

$$M_3 = \Sigma_1 \Psi_1^* \tilde{\Phi}_2, \tag{24}$$

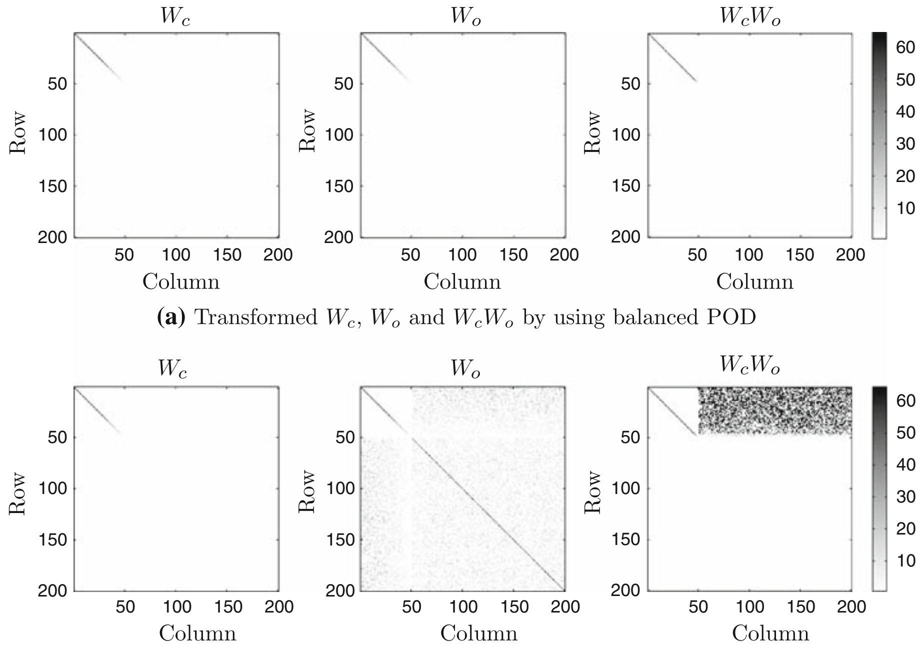
where $\Psi_1 = YU_1\Sigma_1^{-1/2}$ are the true adjoint modes. Note that, when the true adjoint modes are used to define the inverse (18), then $M_3 = 0$, since $\Psi_1^*\Phi_2 = 0$. However, when pseudo-adjoint modes are used, then M_3 is no longer zero, and in fact, can be quite large.

An example is shown in Fig. 1, which shows the magnitude of the elements of the transformed Gramians, where X and Y in (16) are chosen at random. Note that when true adjoint modes are used, the transformed Gramians are equal and diagonal, while when the pseudo-adjoint modes are used, the off-diagonal blocks of the transformed observability Gramian, and the product of the Gramians have significant magnitude.

Thus, when pseudo-adjoint modes are used, the resulting transformation is not, in general, a balancing transformation: even though the upper-left blocks of the transformed Gramians are still equal and diagonal, the transformed observability Gramian is not block diagonal, and so its eigenvalues and eigenvectors do not correspond to those of the transformed controllability Gramian. Note that this is the whole point of balanced truncation: to transform to coordinates in which the most controllable directions (dominant eigenvectors of W_c) correspond to the most observable directions (dominant eigenvectors of W_o). Therefore, while the approximate balanced truncation procedure described in Sect. 2.1 exactly balances the approximate Gramians, transforming by pseudo-adjoint modes does not represent balancing in any meaningful sense.

Note that the matrix M_3 describes the degree to which projection using pseudo-adjoint modes fails to balance the approximate Gramians. This matrix equals zero if the adjoint modes (columns of Ψ_1) are spanned by the primal modes (columns of Φ_1). However, M_3 is the largest when the dominant adjoint modes (columns

[1] If $\operatorname{rank}(X) > \operatorname{rank}(Y)$, then the situation is worse, and the transformed controllability Gramian is not block diagonal, nor does its upper-left block equal Σ_1.



(a) Transformed W_c, W_o and W_cW_o by using balanced POD

(b) Transformed W_c, W_o and W_cW_o by projection using pseudo-adjoint modes

Fig. 1 Transformed Gramian matrices: (**a**) using true adjoint modes (Eqs. 19–20) and (**b**) using pseudo-adjoint modes (Eq. 23). Here, X and Y in (16) are random matrices with $n = 200$ states and $m_c = m_o = 50$ snapshots

of Ψ_1) are nearly orthogonal to the dominant primal modes (columns of Φ_1). Unfortunately, this is the case in many problems of interest, in particular those involving non-normality: the directions spanned by the primal modes often do not coincide with the directions spanned by the adjoint modes.

In the next section, we apply this approach to the flow past a flat plate, and compare it to the methods described in Sect. 2.

4 Example: flow past an inclined flat plate

In this section, we illustrate the application of ERA as an approximate balanced truncation method using a numerical example, by obtaining reduced-order models of a large-dimensional fluid system. We compare the resulting models with those obtained using the balanced POD method of [23], ERA with pseudo-adjoint modes as described in Sect. 3, and the standard POD/Galerkin method [13].

4.1 Model problem and parameters

The model problem that we consider is a two-dimensional uniform flow over a flat plate inclined at an angle $\alpha = 25°$, at a low Reynolds number $Re = 100$. At these conditions, the flow asymptotically reaches a stable steady state, the streamlines of which are plotted in Fig. 2. The numerical method used for all computations is a fast formulation of the immersed boundary method developed by [9], and solves for the vorticity field at each time step. We treat farfield boundary conditions using the multiple-grid scheme described in [9] (Sect. 4) with five nested grids, each with 250×250 points. The finest grid covers the region $[-2, 3] \times [-2.5, 2.5]$, and the largest grid covers the region $[-32, 48] \times [-40, 40]$, where lengths are nondimensionalized by the chord of the flat plate, whose center is located at the origin. The time step used for all simulations is 0.01 (nondimensionalized by chord and freestream velocity). The numerical model is the same as that considered in [1] where balanced POD is applied for feedback controller design to stabilize an *unstable* steady state corresponding to a high angle of attack. However, here we consider the case of a *stable* steady state (with an angle of attack at 25°), for comparison of reduced-order models.

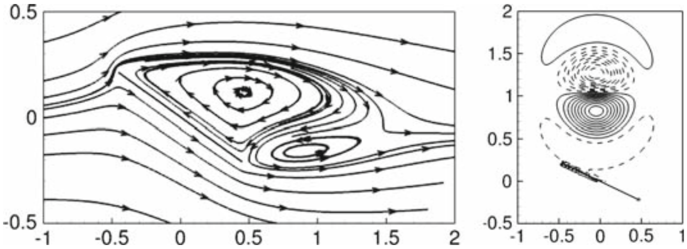


Fig. 2 Streamlines of the stable steady state past a flat plate at $\alpha = 25°$ (*left*), and the contour-lines of the vorticity field obtained from an impulsive input to the actuator (*right*)

4.2 Input and output

The governing equations are first linearized about the stable steady state, resulting in a high-dimensional model in the form of Eq. 1, where the state x consists of the discrete vorticity field at the grid points. See [1] for the details of the linearized (and adjoint) equations and their numerical formulations. The system input u is a disturbance (or actuator) shown in Fig. 2, modeled as a localized body force in the vicinity of the leading edge. We consider the output to be the entire velocity field: this is important for capturing the flow physics, and is often needed to represent cost functions used in optimal control design. Since the output is very high-dimensional, in ERA and balanced POD reduction procedures we use output projection described at the end of Sect. 2, projecting the velocity field onto the leading POD modes of the velocity snapshots obtained from the impulse response simulation.

4.3 Reduced-order models

ERA is applied to the full-dimensional linearized system to construct a reduced-order model. With a sampling period of 50 time steps, 400 adjacent pairs of Markov parameters, as in (10), are collected from an impulse response simulation. Since these parameters are a projection of the velocity fields onto the leading POD modes, for an output projection of order m, the number of inner products required is $4m \times 10^2$ for construction of each H and H' (see Sect. 2.4).

For comparison, balanced POD is also used to compute the same reduced-order models. Adjoint simulations are performed with the POD modes as initial conditions to compute the matrix Y of (5). The matrices X and Y are assembled by stacking 200 snapshots from the linearized and each of the adjoint simulations, and in turn, the generalized Hankel matrix $H = Y^*X$ is computed. For an output projection of order m, the number of inner products required to compute H is $4m \times 10^4$, which is 50 times more than that to compute H and H' in total for ERA.

We also compare reduced-order models using standard POD modes, and ERA with pseudo-adjoint modes, as described in Sect. 3. The first 100 primal modes are used to compute the pseudo-adjoint modes.

For the given case, a comparison between the computational cost using ERA and using balanced POD is shown in Table 1. Results verify that ERA substantially improves computational efficiency in forming reduced-order models.

Next, we compare the reduced-order models. Figure 3 shows the leading two primal modes and true adjoint modes from balanced POD, compared with the leading two pseudo-adjoint modes. The pseudo-adjoint modes look quite different from the true adjoint modes, and the flow structures actually more closely resemble the leading *primal* modes. This result is not surprising, since the pseudo-adjoint modes are always linear combinations of the snapshots from the primal simulations, while the true adjoint modes are linear combinations of snapshots from adjoint simulations. Following the discussion in the last section, the poor approximation of the adjoint modes suggests that the pseudo-adjoint modes may produce poor reduced-order models for this example, as we will verify below.

Figure 4 shows the diagonal values of the controllability and observability Gramians, as well as the empirical Hankel singular values, for reduced-order models obtained from three different methods: ERA, balanced POD, and ERA with pseudo-adjoint modes. The models obtained using ERA are more accurate in the sense that the three sets of curves are almost indistinguishable, for all orders of output projection. However, for balanced POD, the diagonal values of the observability Gramians are accurate only for certain leading modes,

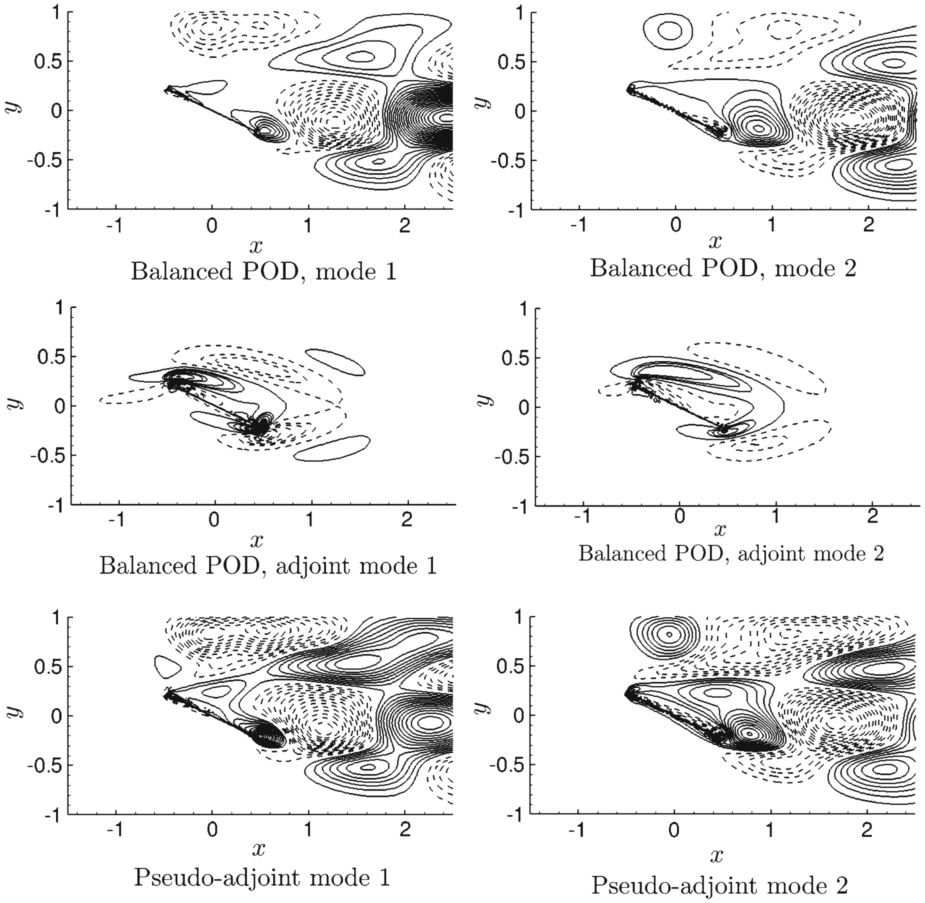


Fig. 3 The first two primal and adjoint modes computed by using balanced POD, and the first two pseudo-adjoint modes computed by using (14) and the first 100 primal modes. Modes are illustrated using contour plots of the vorticity field

the number of which depends on and increases with the order of output projection. This inaccuracy can be attributed to a slight inaccuracy in the adjoint formulation, which in turn results from an approximation in the multi-domain approach used to treat farfield boundary conditions in the immersed boundary method of [9]; see [1] for more details. Thus, ERA is advantageous as it does not need any adjoint simulations and results in more balanced Gramians. On the other hand, ERA with pseudo-adjoint modes generates poorly balanced controllability and observability Gramians, as shown in Fig. 4c. This is because the leading primal modes and adjoint modes are supported very differently in the spatial domain, and thus the pseudo-adjoint modes, based on linear combination of leading primal modes, poorly approximate the true adjoint modes.

4.4 Model performance

We can quantify the performance of the various reduced-order models by computing error norms. One such measure is the 2-norm of the error between the impulse response of the full linearized system, denoted $G(t)$, and that of a reduced-order model with r modes, denoted by $G_r(t)$. We first compute the 2-norm of the error between the full system (with the entire velocity field as output) and the output-projected system of order 20, shown as the horizontal dashed line in Fig. 5. This is the lower error bound for any reduced-order model of the given output-projected system. Results shown in Fig. 5 indicate that the first several low-order models obtained by ERA and balanced POD generate slightly different 2-norms of error, presumably because of the slight inaccuracy in the adjoint, mentioned previously. For most orders, however, they agree; and both error norms converge to the lower bound as the order of the model increases. By running more simulation tests, we observe that with higher-order output projections, ERA and balanced POD error norms converge to each other faster when the order of model increases.

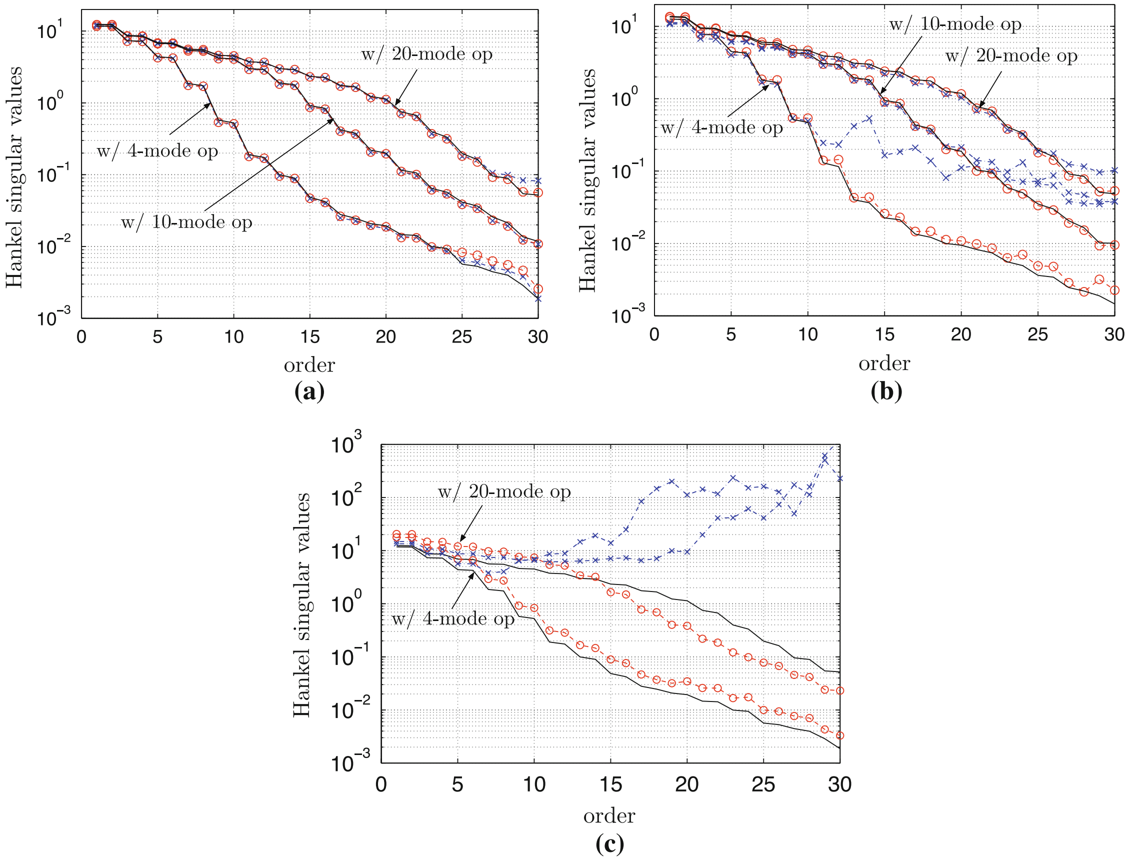


Fig. 4 Comparison of Gramians computed using (**a**) ERA, (**b**) balanced POD, and (**c**) ERA with pseudo-adjoint modes: The empirical Hankel singular values (*solid line*) and the diagonal elements of the controllability (*circle*) and observability (*times*) Gramians with different order of modes (e.g., 4, 10, 20) in output projection

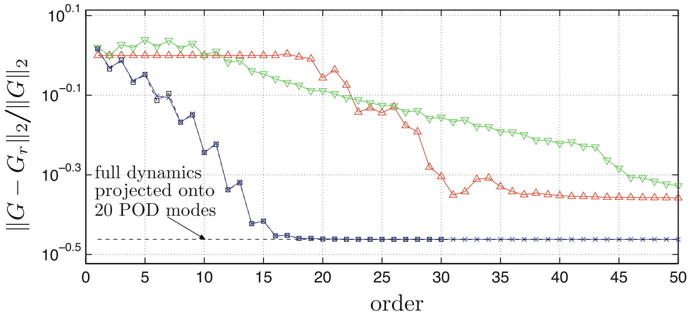


Fig. 5 H_2—norm of the error with increasing order of the reduced-order models: exact output of the output-projected system (*dashed line*); models obtained using balanced POD (*square*), ERA (*times*), ERA with pseudo-adjoint modes (*down-pointing triangle*), and POD (*up-pointing triangle*). A 20-mode output projection is used in ERA, balanced POD, and ERA with pseudo-adjoint modes

Figure 5 also shows the 2-norm error plots for models by ERA with pseudo-adjoint modes, using 20-mode output projection, and for models computed using standard POD. Errors of models by ERA with pseudo-adjoint modes converge to the lower bound much slower than ERA/balanced POD. Errors of models by POD do not start converging until more than nearly 20 modes are used, and they converge to a larger error bound than ERA/balanced POD, again because POD models do not capture the input–output dynamics as well as balanced truncation based models.

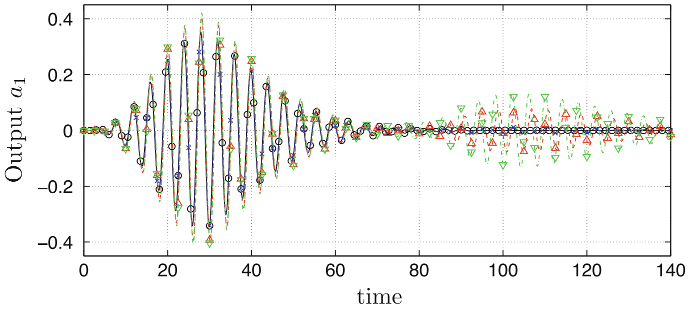


Fig. 6 The first output, output a_1, from the impulse-response simulation: results of full-simulation (*circle*), compared with those of 16-mode reduced order model by ERA (*times*), 30-mode model by ERA with pseudo-adjoint modes (*down-pointing triangle*), and 30-mode model by POD (*up-pointing triangle*). A 20-mode output projection is used in ERA and ERA with pseudo-adjoint modes

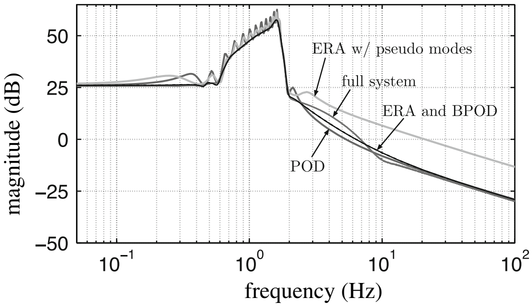


Fig. 7 Singular-value plots: The full system and 30-mode models obtained using balanced POD, ERA, ERA with pseudo-adjoint modes, and POD, all with a 20-mode output projection. ERA and balanced POD models generate almost identical plots

In the time domain, a comparison of the transient response to an impulsive disturbance is shown in Fig. 6, in which the first output of the reduced-order model is plotted, for a 16-mode model determined by ERA, and for 30-mode models by POD and ERA with pseudo-adjoint modes. The 16-mode ERA model already accurately predicts the response for all times. The higher-dimensional, 30-mode models using POD and pseudo-adjoint modes are both stable, and perform reasonably well; however, they overpredict the response, particularly after time $t \approx 80$.

We also compare the frequency response of reduced-order models to that of the full system, or more precisely, the full output-projected system. One way to represent the response of a single-input multiple-output system is by a singular-value plot, a plot of the maximum singular value of the transfer function matrix as a function of frequency. To generate this plot, a very long simulation of 5×10^5 time steps for the full system is performed, with a random input sampled from a uniform distribution in the range $(-0.5, 0.5)$. The output snapshots are projected onto leading POD modes. The magnitude of the transfer function is then computed from the cross spectrum of the input and outputs (using the Matlab command `tfestimate`). Finally, singular-value plots for the full output-projected systems are obtained, with a typical case shown in Fig. 7.

A typical set of singular-value plots of different reduced-order models are presented in Fig. 7. Results shown in the figure indicate that ERA and balanced POD 30-mode models are almost identical, and are close to the corresponding full output-projected system. In comparison, Fig. 7 also shows singular-value plots for 30-mode models by ERA with pseudo-adjoint modes and by POD. Note that for computational feasibility, here the output of the POD model is the first twenty reduced states, i.e., the full-dimensional output of the POD model are projected onto the leading twenty POD modes. The frequency responses of the models by POD and ERA with pseudo-adjoint modes capture the resonant peak, but do not match well for frequencies far away from the resonant peak. These two models both generate spurious peaks in the frequency range of [0.1, 2].

5 Discussion

We report that, theoretically, the ERA and snapshot-based approximate balanced truncation (balanced POD) produce exactly the same reduced-order models. This equivalence implies that ERA balances a pair of approximate Gramians and thus can be regarded as an approximate balanced truncation method. Compared to balanced POD, the main features of ERA are that it does not require data from adjoint systems and therefore can be used with experimental data; furthermore, its construction of the generalized Hankel matrix is computationally an order-of-magnitude cheaper than balanced POD. Numerical results indicate that ERA can be more accurate than balanced POD in practice, since there can be slight inaccuracies in the adjoint operator used with balanced POD. Balanced POD does have its own advantages, however: unlike ERA, it produces sets of bi-orthogonal modes that are useful for other purposes. Nonlinear models may be obtained by projection onto these modes; and parameters such as Reynolds number can be retained in the reduced-order models generated using these modes. Balanced POD has also been generalized for unstable systems. We also examine a modified ERA approach in which one constructs sets of bi-orthogonal modes without using adjoint information, using a matrix pseudo-inverse, as in [22]. Although this approach provides sets of bi-orthogonal modes (primal/pseudo-adjoint modes), in general it cannot be regarded as an approximate balanced truncation method, since it does not balance the approximate Gramians.

We have demonstrated the methods on a model problem consisting of a disturbance interacting with the flow past an inclined flat plate. As expected, balanced POD models perform nearly identically to ERA models. The small differences result because the adjoint simulation required for balanced POD is not a perfect adjoint at the discrete level. Both procedures work significantly better than standard POD models, or ERA models using pseudo-adjoint modes for projection.

Finally, we emphasize that throughout we have considered only stable, linear models. Possible future directions of this study include a generalization to unstable systems, and ultimately to nonlinear systems.

Acknowledgments The authors gratefully acknowledge the support by U. S. Air Force Office of Scientific Research grants FA9550-05-1-0369 and FA9550-07-1-0127. The authors thank Kunihiko Taira for help with the numerical solver.

References

1. Ahuja, S., Rowley, C.W.: Feedback control of unstable steady states of flow past a flat plate using reduced-order estimators. J. Fluid Mech. (accepted) (2009)
2. Akers, J.C., Bernstein, D.S.: ARMARKOV least-squares identification. In: Proceedings of the ACC, pp. 186–190. Albuquerque, NM, USA (1997)
3. Åkervik, E., Hœpffner, J., Ehrenstein, U., Henningson, D.S.: Optimal growth, model reduction and control in a separated boundary-layer flow using global eigenmodes. J. Fluid Mech. **579**, 305–314 (2007)
4. Bagheri, S., Brandt, L., Henningson, D.S.: Input–output analysis, model reduction and control of the flat-plate boundary layer. J. Fluid Mech. **620**, 263–298 (2009)
5. Cabell, R.H., Kegerise, M.A., Cox, D.E., Gibbs, G.P.: Experimental feedback control of flow-induced cavity tones. AIAA J. **44**(8), 1807–1815 (2006)
6. Cattafesta, L.N. III, Garg, S., Choudhari, M., Li, F.: Active control of flow-induced cavity resonance. AIAA Paper 97-1804 (1997)
7. Cattafesta, L.N. III., Song, Q., Williams, D.R., Rowley, C.W., Alvi, F.S.: Active control of flow-induced cavity oscillations. Prog. Aerosp. Sci. **44**, 479–502 (2008)
8. Choi, H., Jeon, W.P., Kim, J.: Control of flow over a bluff body. Ann. Rev. Fluid Mech. **40**, 113–139 (2008)
9. Colonius, T., Taira, K.: A fast immersed boundary method using a nullspace approach and multi-domain far-field boundary conditions. Comput. Methods Appl. Mech. Eng. **197**(25–28), 2131–2146 (2008)
10. Djouadi, S.M., Camphouse, R.C., Myatt, J.H.: Empirical reduced-order modeling for boundary feedback flow control. J. Control Sci. Eng. **2008**, Article ID 154956, 11 pages. doi:10.1155/2008/154956 (2008)
11. Gaitonde, A.L., Jones, D.P.: Reduced order state-space models from the pulse responses of a linearized CFD scheme. Int. J. Numer. Methods Fluids **42**, 581–606 (2003)
12. Gawronski, W.: Balanced Control of Flexible Structures. Springer, London (1996)
13. Holmes, P., Lumley, J.L., Berkooz, G.: Turbulence, Coherent Structures, Dynamical Systems and Symmetry. Cambridge University Press, Cambridge, UK (1996)
14. Ilak, M., Rowley, C.W.: Modeling of transitional channel flow using balanced proper orthogonal decomposition. Phys. Fluids **20**, 034,103 (2008)
15. Juang, J.N., Pappa, R.S.: An eigensystem realization algorithm for modal parameter identification and model reduction. J. Guid. Control. Dyn. **8**(5), 620–627 (1985)
16. Juang, J.N., Phan, M.Q.: Identification and Control of Mechanical Systems. Cambridge University Press, Cambridge (2001)
17. Kim, J., Bewley, T.R.: A linear systems approach to flow control. Ann. Rev. Fluid Mech. **39**, 383–417 (2007)
18. Lim, R.K., Phan, M.Q., Longman, R.W.: State-space system identification with identified Hankel matrix. Mechanical and Aerospace Engineering Tech. Report 3045, Princeton University (1998)
19. Ma, Z., Rowley, C.W.: Low-dimensional linearized models for systems with periodic orbits, with application to the Ginzburg-Landau equation. AIAA Paper 2008-4196, 4th Flow Control Conference (2008)
20. Moore, B.C.: Principal component analysis in linear systems: Controllability, observability, and model reduction. IEEE Trans. Automat. Contr. **26**(1), 17–32 (1981)

21. Noack, B., Afanasiev, K., Morzyński, M., Tadmor, G., Thiele, F.: A hierarchy of low-dimensional models for the transient and post-transient cylinder wake. J. Fluid Mech. **497**, 335–363 (2003)
22. Or, A.C., Speyer, J.L., Carlson, H.A.: Model reduction of input-output dynamical systems by proper orthogonal decomposition. J. Guid. Control Dyn. **31**(2), 322–328 (2008)
23. Rowley, C.W.: Model reduction for fluids using balanced proper orthogonal decomposition. Int. J. Bifurc. Chaos **15**(3), 997–1013 (2005)
24. Silva, W.A., Bartels, R.E.: Development of reduced-order models for aeroelastic analysis and flutter prediction using the CFL3Dv6.0 code. J. Fluids Struct. **19**, 729–745 (2004)