RESEARCH ARTICLE | JULY 18 2023

Comparison and evaluation of dimensionality reduction techniques for the numerical simulations of unsteady cavitation

Guiyong Zhang (张桂勇) ; Zihao Wang (王子豪) ; Huakun Huang (黄华坤); Hang Li (李航); Tiezhi Sun (孙铁志)

Physics of Fluids 35, 073322 (2023)
https://doi.org/10.1063/5.0161471

Comparison and evaluation of dimensionality reduction techniques for the numerical simulations of unsteady cavitation

Cite as: Phys. Fluids **35**, 073322 (2023); doi: 10.1063/5.0161471
Submitted: 10 June 2023 · Accepted: 4 July 2023 ·
Published Online: 18 July 2023

View Online Export Citation CrossMark

Guiyong Zhang (张桂勇),[1,2] Zihao Wang (王子豪),[1] Huakun Huang (黄华坤),[3] Hang Li (李航),[4] and Tiezhi Sun (孙铁志)[1,a)]

AFFILIATIONS

[1] State Key Laboratory of Structural Analysis, Optimization and CAE Software for Industrial Equipment, School of Naval Architecture Engineering, Dalian University of Technology, Dalian 116024, China

[2] Collaborative Innovation Center for Advanced Ship and Deep-Sea Exploration, Shanghai 200240, China

[3] Department of Aeronautical and Aviation Engineering, the Hong Kong Polytechnic University, Kowloon, Hong Kong, China

[4] China Ship Development and Design Center, Wuhan 430064, China

[a)] Author to whom correspondence should be addressed: suntiezhi@dlut.edu.cn

ABSTRACT

In the field of fluid mechanics, dimensionality reduction (DR) is widely used for feature extraction and information simplification of high-dimensional spatiotemporal data. It is well known that nonlinear DR techniques outperform linear methods, and this conclusion may have reached a consensus in the field of fluid mechanics. However, this conclusion is derived from an incomplete evaluation of the DR techniques. In this paper, we propose a more comprehensive evaluation system for DR methods and compare and evaluate the performance differences of three DR methods: principal component analysis (PCA), isometric mapping (isomap), and independent component analysis (ICA), when applied to cavitation flow fields. The numerical results of the cavitation flow are obtained by solving the compressible homogeneous mixture model. First, three different error metrics are used to comprehensively evaluate reconstruction errors. Isomap significantly improves the preservation of nonlinear information and retains the most information with the fewest modes. Second, Pearson correlation can be used to measure the overall structural characteristics of the data, while dynamic time warping cannot. PCA performs the best in preserving the overall data characteristics. In addition, based on the uniform sampling-based K-means clustering proposed in this paper, it becomes possible to evaluate the local structural characteristics of the data using clustering similarity. PCA still demonstrates better capability in preserving local data structures. Finally, flow patterns are used to evaluate the recognition performance of flow features. PCA focuses more on identifying the major information in the flow field, while isomap emphasizes identifying more nonlinear information. ICA can mathematically obtain more meaningful independent patterns. In conclusion, each DR algorithm has its own strengths and limitations. Improving evaluation methods to help select the most suitable DR algorithm is more meaningful.

I. INTRODUCTION

With the rapid development in the field of fluid mechanics, dealing with large-scale high-dimensional spatiotemporal data from experiments and computational fluid dynamics (CFD) has become a reality that fluid mechanics researchers must face.[1–3] Dimensionality reduction (DR) is an essential tool in machine learning, aiming to transform high-dimensional data into lower-dimensional representations.[4–6] In the field of fluid mechanics, DR is widely utilized for feature extraction and information simplification.[7] The DR process can be classified into two major categories: linear methods and nonlinear methods.

Currently, linear DR methods are popular in CFD. Linear DR methods are employed to extract important features or patterns from flow fields and provide a low-dimensional representation of the data. This has emerged as a subfield known as data-driven modal analysis.[8] Modal analysis has achieved extensive success in the field of fluid mechanics and has been applied to flow visualization,[9] turbulence analysis,[10] and the reconstruction of sparse flow data.[11] Modal decomposition methods have played a crucial role in understanding high-dimensional, nonlinear, and complex fluid phenomena.[12] However, they fail to capture the highly nonlinear characteristics of fluid dynamics. Additionally, commonly used modal decomposition methods are invasive, requiring modifications to the data during the DR process, such as feature selection, feature extraction, or feature transformation. This inevitably leads to information loss, and as the data volume increases, the proportion of information loss also increases.

Nonlinear DR methods have a relatively short history but have gained popularity in recent years. Nonlinear methods can capture the nonlinear relationships between data points.[13] It is worth noting that deep learning has been applied in this field. A series of powerful nonlinear DR tools based on autoencoders (AEs) have been developed, aiming to learn the nonlinear low-dimensional representations of high-dimensional data.[14,15] However, although deep learning has shown promising results in standard benchmark problems, it comes with high computational costs and limited interpretability for complex fluid problems.[9] In addition to deep learning, manifold learning methods are also applicable for nonlinear dimensionality reduction. Manifold learning is a noninvasive DR technique designed to reveal the low-dimensional structure hidden in high-dimensional data while preserving the structure's similarity. The inherent preservation of structural similarity in manifold learning provides higher interpretability compared to AEs. Among manifold learning techniques, isomap can be considered the most popular, which uses neighborhood or graph-based techniques to find a low-dimensional representation of high-dimensional data.[16] Additionally, techniques such as kernel principal component analysis (KPCA),[17] local linear embedding (LLE),[18] and Laplacian eigenmaps (LEM)[19] are also gaining popularity in the fluid dynamics community. It is important to note that, unlike linear DR and AEs, most manifold learning algorithms for DR are noninvertible, meaning they cannot be mapped back to the original space. This issue will be discussed later.

A comprehensive comparison and evaluation of various DR methods are of great significance for their application in fluid mechanics. Recently, Csala *et al.*[20] provided a comprehensive analysis of these methods, focusing on their qualitative (mode patterns) and quantitative (reconstruction error) discussions. Mode patterns focus on the interpretability of the reduced flow field features, while reconstruction error focuses on the loss of information in the reduced flow field. These two criteria are considered as a consensus in the field of fluid mechanics. However, when we review the original intention of DR, a good method should be able to preserve the structural characteristics of the original data while reducing its dimensionality and retain as much important information as possible. The preservation of the original data structure is often overlooked. The comparison of reconstruction errors alone is not accurate for evaluating DR methods.

In the field of fluid dynamics, complex nonlinear DR algorithms are frequently applied to standard benchmark problems, such as flow around a cylinder.[21] However, understanding linear and nonlinear DR methods is more relevant for complex flow field problems. Cavitation is a complex flow phenomenon involving interactions between phase change and vortical structures.[22] Typical cavitation refers to the phenomenon where, in a liquid, local pressure drops below the liquid's saturation vapor pressure due to a sudden increase in liquid flow velocity or a sudden decrease in pressure, leading to the formation of bubbles inside the liquid.[23,24] These bubbles grow and rupture in high-speed liquid flow, causing liquid vibration and damage. Cavitation phenomena are widely present in fields such as ships, liquid pumps, and hydropower generation, seriously affecting the normal operation and service life of equipment.[25,26] Therefore, cavitation research has always been of great concern. Using linear DR methods to simplify cavitation flow fields and obtain flow field modes is popular and helps to gain a deeper understanding of cavitation phenomena. However, the application of nonlinear DR algorithms to cavitation flow fields has been rarely studied.

To enhance the understanding of the application of DR in fluid dynamics and overcome its challenges, this paper extensively compares and evaluates the performance differences of three DR methods applied to the unsteady cavity flow around a 3D hydrofoil. The rest of this paper is organized as follows: Sec. II introduces the methods used in this study and outlines the detailed information of the dataset. Section III analyzes the different DR methods from three aspects: reconstruction error of the flow field, preservation of data structure, and features of flow field modes. Finally, Section IV summarizes the work and presents the main conclusions of this paper.

II. METHOD AND DATA SET

A. Dimensionality reduction method

In this section, we introduce three DR algorithms, including two linear methods and one nonlinear method, which are discussed in this article. These algorithms are briefly outlined in this paper, and more detailed descriptions can be found in the references.[27–29] For cavitation flow phenomena, the velocity of the flow field reflects most of the flow characteristics of cavitation, so it is used as the object of DR. The goal of all DR methods is to reduce the temporal dimension of the data, which captures the underlying coherent structures and flow features of the fluid, thereby facilitating physical understanding and interpretation. The Scikit-learn Python library[31] is used for implementing DR methods.

1. Principal component analysis (PCA)

PCA is an orthogonal linear transformation method that transforms high-dimensional raw data into a set of low-dimensional orthogonal variables, known as principal components or PCs.[30] To perform PCA, first, the raw data matrix $X \in R^{N\times M}$ is obtained by extracting data from the flow field, which includes N snapshots of M positions in the flow field. The objective of PCA is to find a set of orthonormal eigenvectors that capture the maximum variance of the data. This is achieved by computing the covariance matrix C. The covariance matrix C is as follows:

$$C = \frac{1}{n-1} X^T X. \tag{1}$$

The eigenvectors and eigenvalues of C are then calculated using singular value decomposition (SVD) to obtain the principal components and their corresponding temporal coefficients

$$C = U\Sigma V^T. \tag{2}$$

From the definition of singular value decomposition, each column of U corresponds to a PCA dimension, and each row of V contains the corresponding temporal coefficients. The size of the matrix Σ is $M \times N$, and the non-zero elements on the diagonal correspond to the singular values of the discrete matrix Y. The importance of each mode is given by its associated singular value.

2. Independent component analysis (ICA)

ICA is a method that separates a multivariate signal into independent, non-Gaussian components. PCA determines principal components by finding directions of maximal variance through sorting eigenvectors according to their corresponding eigenvalues in descending order. However, this method has limitations in dealing with non-Gaussian distributed data since variance alone may not capture the primary directions. As a complement to PCA, independent component analysis (ICA) extracts statistically independent components from non-Gaussian data, providing additional practicality. Therefore, PCA and ICA are complementary linear DR methods. ICA is actually an optimization problem that obtains independent components through optimization criteria and algorithms. ICA not only eliminates correlations among variables but also extracts statistically independent components.[31] This method has been widely used in fields such as speech signals, image signals, and electroencephalographic signals, but research on its application in fluid mechanics is lacking.

ICA aims to separate source signals from observed data $X_{\mathrm{n}\times\mathrm{m}}$. Given m source signals $S_{\mathrm{nxm}} = [s_1(t), s_2(t), \ldots, s_m(t)]$ and a mixing matrix $A_{\mathrm{n}\times\mathrm{n}}$ that combines them

$$X_{\mathrm{n}\times\mathrm{m}} = A_{n\times\mathrm{n}} S_{n\times w}. \tag{3}$$

The goal of ICA processing is to find a linear transformation matrix W that separates the observed signals X(t) as much as possible, without knowing the source signals S(t) or the mixing matrix A.

Each separated signal can be expressed as

$$y_j(t) = \sum_{i=1}^{m} w_{ji} x_i(t), j = 1, 2, \ldots, n, \tag{4}$$

where w_{ji} are separation coefficients. The corresponding blind separation model is

$$y(t) = WX(t) = WAS(t), \tag{5}$$

where $\boldsymbol{y}(t) = [y_1(t), y_2(t), \ldots, y_n(t)]^T$ is a matrix composed of n separated signals, called the mixing matrix.

Independence is the criterion for separating source signals. Fast independent component analysis (FastICA) is a commonly used fast optimization iterative algorithm for ICA.[32] It uses the maximization of non-Gaussianity as the objective function to measure independence. The negative entropy $J(x)$ is defined as

$$J(x) = H(x_{\mathrm{gamax}}) - H(x). \tag{6}$$

The calculation of negative entropy is complex, so an approximation $J(y)$ is used to find it

$$J(y) \propto \left(E\{G(y)\} - E\{G(v)\}\right)^2, \tag{7}$$

where $H(x)$ represents the entropy of the pre-whitened data x, while $H(x_{\mathrm{gamax}})$ represents the entropy of the maximally non-Gaussian projection of x. $G(y)$ represents an approximation of negative entropy for a variable y, while $G(v)$ represents an approximation of negative entropy for a variable v that is orthogonal to y.

Specifically, the FastICA algorithm maximizes the non-Gaussianity of the rotated components of pre-whitened data through fixed-point iterations. The iterations are as follows:

$$(1+\alpha)W = E\left(xg\left(W^{\mathrm{T}}x\right)\right) + \alpha W, \tag{8}$$

where x is the pre-whitened data, g is any non-quadratic function, E is the expectation, and α is a small positive constant.

3. Isometric mapping (isomap)

Isomap is a nonlinear DR algorithm that seeks to preserve the intrinsic geometric structure of high-dimensional data in a low-dimensional space.[33] The basic idea behind isomap is to construct a weighted graph that approximates the geodesic distances between data points in the high-dimensional space and then embed this graph in a lower-dimensional space while preserving the pairwise distances as much as possible.[34] The isomap algorithm can be summarized as follows:

(1) Constructing the neighborhood graph G: given a set of sample points $x(i = 1, 2, \ldots, N)$, we can construct a neighborhood graph G based on the following criteria:
Euclidean distance threshold ε: if the Euclidean distance $d(i, j)$ between two sample points i and j is less than or equal to a specified threshold ε, we consider them adjacent and connect nodes i and j with an edge. The weight of the edge is set to the Euclidean distance $d(i, j)$.
k-nearest neighbors: if sample points i and j are each other's k-nearest neighbors, we consider them adjacent and connect nodes i and j with an edge. The weight of the edge is set to the Euclidean distance $d(i, j)$.
By applying these criteria, we can obtain an adjacency graph G that represents the relationships between all the sample points.

(2) Estimating the geodesic distance matrix: in isomap, since it is challenging to accurately obtain the geodesic distance on the manifold structure, we approximate it by using the shortest path on the neighborhood graph G. The geodesic distance matrix is initialized as $d_a(i, j)$. For each sample point $I = 1, 2, \ldots, N$ (where N is the number of sample points), we calculate the updated geodesic distance matrix $D_\sigma = \{d_c(i, j)\}$ using the following formula:

$$d_a(i, j) = \min\{d(i, j), d_c(i, I) + d(I, j)\}. \tag{9}$$

Here, $d_c(i, I)$ represents the sum of distances between sample points i and I, and $d(I, j)$ represents the distance between sample points I and j. If sample points i and j are connected by an edge in the neighborhood graph G, then $d_a(i, j)$ is set to $d(i, j)$. Otherwise, it is set to infinity. By iteratively applying this formula for all pairs of sample points, we can obtain the updated geodesic distance matrix D_σ, which approximates the geodesic distances on the manifold structure.

(3) Obtain low-dimensional embedding: apply the multi-dimensional scaling (MDS) technique[35] to the geodesic distance matrix D_c, set $\tau(D_0) = -\frac{HSH}{2}$, where H is the centering matrix $H_v = \{\delta - 1/N\}$, S is the square matrix of distances $S = \{D_i^2\}$. Use matrix analysis to obtain all eigenvalues and eigenvectors of matrix $\tau(D_{\mathrm{o}})$, sort them in descending order of eigenvalues, select the first d eigenvalues $\lambda, \lambda_2, \ldots, \lambda_{\mathrm{d}}$ and eigenvectors and combine them into a matrix $U = (u_1, u_2, \ldots, u_{\mathrm{d}})$. Then, the final d-dimensional embedding result is

$$Y = \mathrm{daig}\left(\lambda_1^{\frac{1}{2}}, \lambda_2^{\frac{1}{2}}, \ldots, \lambda_d^{\frac{1}{2}}\right) U^x. \tag{10}$$

B. Evaluation method

In the field of fluid mechanics, the conventional evaluation methods for various DR methods have only focused on the interpretability of the flow field features and the reconstruction error. This approach is inaccurate. Therefore, this article proposes to supplement the evaluation method by considering the preservation of the structural characteristics of the original data. This aspect is crucial in assessing the quality of DR methods for fluid mechanics applications, especially when analyzing and utilizing flow fields.

1. Data reconstruction

Maximizing the preservation of original data information is crucial in data DR. This often requires the use of data reconstruction methods to assess the loss of data by reconstructing the reduced data back into the original high-dimensional space. For linear DR methods, the reconstructed high-dimensional data can be obtained by multiplying the low-dimensional data with the linear transformation matrix obtained during the dimensionality reduction process. However, for nonlinear DR methods, the data reconstruction process is more intricate due to their nonlinear characteristics. In this case, the reconstruction method is based on the nearest neighbor reconstruction method. Specifically, for each point in the low-dimensional data, the k nearest neighbors in the high-dimensional space are found, and then the data are reconstructed using these nearest neighbors. This approach is widely used.[20,36–38] Specifically, for each point $\mathbf{y}_i \in \mathbb{R}^r$ in the low-dimensional space, we find its k nearest neighbors $\mathbf{y}_{ij}$ in the high-dimensional space. Then, the reconstruction weights w_{ij} can be obtained by minimizing the following objective function:

$$\min_{w_{ij}} \left\| \mathbf{y}_i - \sum_{j=1}^{k} w_{ij}\mathbf{y}_{ij} \right\|_2^2, \tag{11}$$

$$\sum_{j=1}^{k} w_{ij} = 1. \tag{12}$$

The above optimization problem can be solved, and the regularization parameter λ can be used to handle singularity issues. Once the weights w_{ij} are obtained, the original data point $\mathbf{x}_i \in \mathbb{R}^n$ can be approximated as

$$\mathbf{x}_i \approx \sum_{j=1}^{k} w_{ij}\mathbf{x}_j. \tag{13}$$

Finally, the reconstructed data matrix $\mathbf{X}_r$ can be obtained as

$$\mathbf{X}_r = \mathbf{W}\tilde{\mathbf{X}}, \tag{14}$$

where $\mathbf{W}$ is the weight matrix with elements w_{ij} and $\tilde{\mathbf{X}}$ is the matrix of k nearest neighbors in the high-dimensional space.

2. Reconstruction error

Three DR algorithms, whether linear or nonlinear, aim to construct a low-dimensional embedding Y to capture the most important flow features and map back to the original high-dimensional space while minimizing reconstruction error. To evaluate the reconstruction error of DR algorithms, this study uses three different error criteria for different perspectives of analysis. The first is the relative reconstruction error based on the Frobenius norm ε_r:

$$\varepsilon_r(\mathbf{X}, \mathbf{X}_r) = \frac{\|\mathbf{X} - \mathbf{X}_r\|_F}{\|\ \mathbf{X}\|_F}, \tag{15}$$

where $\mathbf{X}_r$ is the reconstructed $\mathbf{X}$ from the r-dimensional latent space. This error measure is widely used in the field of fluid mechanics.[20] This error integrates time and spatial errors into one value, which facilitates the evaluation of algorithm hyperparameters and the visualization of parameter effects. The second is the standard reconstruction error ε_s

$$\varepsilon_s(\mathrm{X},\ \mathbf{X}_r) = \frac{|\mathrm{X} - \mathbf{X}_r|}{|\mathrm{X}|}. \tag{16}$$

This error can visualize the differences in the spatial values of the flow field at each time step. Finally, the root mean square error (RMSE) integrates the time errors of each spatial point into one value, enabling visualization and analysis of differences

$$\mathrm{RMSE}(\mathbf{X}, \mathbf{X}_r) = \sqrt{(\mathbf{X} - \mathbf{X}_r)^2/\mathrm{N}}. \tag{17}$$

3. Pearson correlation

Pearson correlation is a measure of linear correlation between two sequences.[39] Pearson correlation only considers the trend similarity between two sequences. The first step in constructing Pearson correlation is to compute the covariance between the two sequences, which is then divided by the square root of the product of their sequences. The Pearson correlation ρ_{XY} is defined as follows:

$$\rho_{XY} = \frac{Cov(X, \mathrm{Y})}{\sqrt{D(X)}\sqrt{D(Y)}} = \frac{E((X - EX)(Y - EY))}{\sqrt{D(X)}\sqrt{D(Y)}}, \tag{18}$$

where ρ_{XY} is a dimensionless variable that ranges between -1 and $+1$. A positive value indicates a positive correlation, while a negative value indicates a negative correlation, and its value is an indicator of the strength and direction of the correlation. $Cov(X, Y)$ represents the covariance between X and Y. Covariance measures how the variables vary together. $D(X)$ represents the variance of X. Variance measures the spread or dispersion of the variable X. $D(Y)$ represents the variance of Y. Variance measures the spread or dispersion of the variable Y. $E((X - EX)(Y - EY))$ represents the expectation or expected value of the product of the standardized deviations of X and Y. The standardized deviation of a variable is obtained by subtracting its mean ($E(X)$ or $E(Y)$) and dividing by the square root of its variance.

4. Dynamic time warping

Dynamic time warping (DTW) is a method used to compare the similarity between two sequences.[40] By aligning two sequences to find the best match between them, DTW calculates their similarity, with smaller values indicating greater similarity. DTW considers mainly the numerical similarity between two sequences. Suppose we have two single-dimensional sequences

$x(i), i = 1, 2, \ldots, m$ and $y(j), j = 1, 2, \ldots, n$, and construct the distance matrix $\boldsymbol{D}$. $\boldsymbol{D}$ is defined as

$$\boldsymbol{D} = \begin{bmatrix} d(m,1) & d(m,2) & \cdots & d(m,n) \\ \vdots & \vdots & \cdots & \vdots \\ d(2,1) & d(2,2) & \cdots & d(2,n) \\ d(1,1) & d(1,2) & \cdots & d(1,n) \end{bmatrix}, \tag{19}$$

where $d(i,j) = (x(i) - y(j))^2$ is the Euclidean distance between every two data points. The DTW distance is the sum of the accumulated distances along the optimal path $r(i,j)$. The optimal path must satisfy three constraints: boundary, continuity, and monotonicity, which means the path must start at $d(1,1)$ and end at $d(m,n)$, and cannot cross or backtrack on any matches. Mathematically, the DTW distance can be expressed as

$$r(i,j) = d(i,j) + \min\{r(i-1, j-1) \tag{20}$$

$$r(i-1, j), r(i, j-1)\} \tag{21}$$

$$DTW = \min r(m,n). \tag{22}$$

5. K-means clustering based on uniform sampling

This article proposes a method for identifying the micro-structural features of data using clustering. Clustering can divide data into different categories based on their characteristics, and different clusters represent the micro-structural features of the data. If high-dimensional data and low-dimensional data have similar micro-structural features, their clustering results will also be similar.

The K-means algorithm is a classic clustering algorithm; however, it is very sensitive to the selection of initial centroids, and different initial centroids may lead to different clustering results.[41] The default method for selecting initial centroids in the K-means algorithm is random selection. The K-means algorithm is a classic clustering algorithm, but it is highly sensitive to the initial selection of centroids. Different initial centroids can lead to different clustering results. The default method for selecting initial centroids in the K-means algorithm is random selection. Modifications to the initial centroids can result in a series of improved versions of the K-means algorithm designed for specific purposes. In recent work, Wang *et al.*[42] proposed an enhanced version of the K-means algorithm called time series K-means (TK-means), which improves the definition of initial centroids by considering the temporal continuity of clustering results across different snapshots under the first-order Markov assumption.

In this article, we propose to use uniform sampling to select initial centroids. This ensures that the initial conditions for clustering high-dimensional and low-dimensional data are the same, resulting in comparable evaluation results. This method can be called the uniform sampling-based K-means algorithm (UsK-means). The specific procedure of the UsK-means algorithm is explained as follows:

a) Use uniform sampling to obtain the initial centroids of the data.
b) For each data point, calculate its Euclidean distance to each centroid and assign it to the closest cluster.
c) For each cluster, re-calculate its centroid by taking the mean coordinates of all data points in the cluster.
d) Iterate the above steps until the stopping criteria are met, such as reaching the specified maximum number of iterations or the objective function is below a predetermined threshold. The objective function is the sum of squared Euclidean distances between data points and their nearest centroid (*SSE*)

$$SSE(\mathrm{p}, \mathrm{c}) = \sum\nolimits_{i=1}^{n} \|p_i - c_j\|^2, \tag{23}$$

where p_i is the location of a data point, c_j is the nearest center of mass of the data point, and $\|p_i - c_j\|$ is the Euclidean distance.

6. Clustering similarity measures

Cluster similarity metrics are used to measure the similarity or distance between high-dimensional and low-dimensional clustering results. In this paper, we adopt three different cluster similarity metrics: Fowlkes–Mallows index (FMI),[43] normalized mutual information (NMI),[44] and adjusted Rand index (ARI).[45]

FMI measures the degree of matching between clustering results and true class labels, where a larger value indicates a better match. It is calculated by converting clustering results and true labels into two sets, and computing the ratio of the intersection size to the union size of these two sets. The formula for FMI is as follows:

$$\mathrm{FMI} = \frac{T_P}{\sqrt{(T_P + F_P)(T_P + T_N)}}, \tag{24}$$

where T_P is the true positive, implying that the actual positive samples are correctly predicted as a positive samples. T_N is the true negative, meaning that the actual negative samples are correctly predicted as a negative example. F_P is false positive, meaning that the actual negative samples are mis-predicted as a positive sample. F_N is false negative, meaning that the actual positive samples are mis-predicted as a negative sample.

NMI takes into account both the matching degree between clustering results and true labels and their information entropy. It is calculated by converting clustering results and true labels into two probability distributions, computing their mutual information, and then dividing the result by the sum of the entropies of clustering results and true labels. The formula for NMI is as follows:

$$\mathrm{NMI} = \frac{\mathrm{I}(\mathrm{X}, \mathrm{Y})}{\sqrt{H(X)\mathrm{H}(\mathrm{Y})}}, \tag{25}$$

where X, Y is the result of two clusters. $\mathrm{I}(\mathrm{X}, \mathrm{Y})$ denotes the mutual information between X and Y, and $H(X)$ and $H(Y)$ denote the entropy of X and Y, respectively. The value of NMI is in the range of [0, 1], and the larger the value, the more information is shared with the real results, i.e., the better the clustering effect.

ARI not only considers the matching degree between clustering results and true labels but also takes into account the matching degree between random clustering results and true labels. It is calculated by converting clustering results and true labels into two sets, computing their Rand index, subtracting the expected value of the random Rand index, and then dividing the result by the difference between the maximum and expected values of the random Rand index. The formula for ARI is as follows:

$$RI = \frac{TP + TN}{TP + FP + TN + FN}, \tag{26}$$

$$ARI = \frac{RI - E[RI]]}{\max(RI) - E[RI]]}, \tag{27}$$

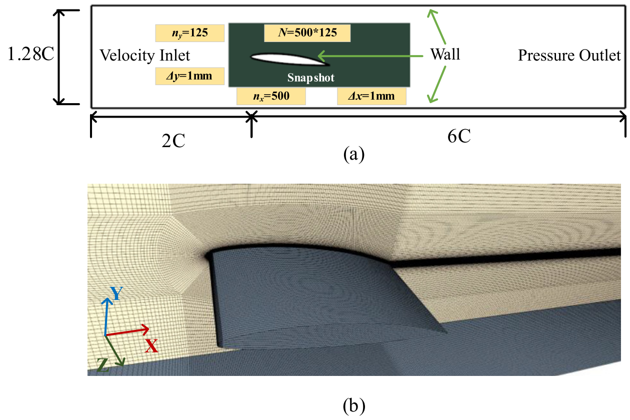
where $E[RI]$ is the expected value of the Rand index under the null hypothesis of random cluster assignments.

C. Cavitation flow data sets

The study focuses on the cavitation flow around an NACA66 hydrofoil and generates the dataset required for the DR algorithm. The NACA66 hydrofoil has been extensively studied experimentally by Leroux *et al.*[46] and is considered a classic benchmark example for cavitation flow research.[24,25]

1. Numerical setup

In this study, a three-dimensional unsteady cloud cavitation flow around the hydrofoil is simulated using the compressible Navier–Stokes equations based on a homogeneous multiphase flow model. The momentum equation is solved using a second-order upwind scheme with second-order temporal discretization. Turbulence is modeled using the Reynolds-averaged Navier–Stokes shear stress transport (RANS SST) k-ω model,[47] while the Schnerr–Sauer model is used to describe the cavitation process.[48] The volume of fluid (VOF) method is used to capture the interface between different phases.[49] Numerical simulations are conducted using the Star-CCM+ flow solver. The sketch of the computational domain and boundary conditions are shown in Fig. 1(a). The chord length of the NACA66 hydrofoil is $C = 0.15$ m, and the angle of attack is 6°. To be consistent with the experimental setup,[46] the entire computational domain has a length of 8C and a height of 1.28C. To reduce the required computational resources, the span of the computational domain is set to 0.3C, which is a widely used technique for simulating cavitation flows.[50] The leading edge of the hydrofoil is located 2C away from the inlet of the computational domain, and the trailing edge is 5C away from the outlet. The inlet and outlet boundary conditions are set to velocity inlet and pressure outlet, respectively. The upper and lower walls of the computational domain and the surface of the hydrofoil are set to non-slip walls, while all other boundaries of the computational domain are set to symmetric planes. The initial temperature is set to 298 K, the inlet velocity is set to 5.33 m/s, and the cavitation number is set to 1.25. The details of the three-dimensional mesh around the hydrofoil are shown in Fig. 1(b). A structured mesh is generated using ICEM. The mesh independence study of the model can be found in the paper by Wang *et al.*,[21] and the final mesh consists of approximately 4.54×10^6 mesh elements, selected to balance computational accuracy and efficiency. The time step is $\Delta t = 2.5 \times 10^{-5}$, which ensures that the maximum Courant number in the flow field remains below 0.8, which ensures the stability of the numerical simulations.[51] The maximum number of iterations per time step is limited to 20. In this study, we focus on the flow around the hydrofoil, so the sampling area of the flow field snapshots is concentrated around the hydrofoil attachment. Its size is $n_x \times n_y = 500 \times 125$, with a sampling point spacing of 1 mm.



(a)

(b)

FIG. 1. (a) Sketch of the computational domain and boundary conditions. (b) Three-dimensional mesh around the hydrofoil.

2. Numerical method validation studies

To validate the numerical method, Fig. 2 shows a comparison between the simulated and experimental shapes of the cavity represented by the iso-surface of water volume fraction equal to 0.8. The selection of water volume fraction is typically based on the balance between the clarity of visualizing and the accuracy of cavity morphology. Higher water volume fractions are generally advantageous for achieving a clearer visualization of the details of small-scale cavities. Therefore, an iso-surface with a water volume fraction of 0.8 is used to represent the morphology of the cavities. The results indicate that the shape of the cavity exhibits a regular periodic variation. The numerical results capture well the characteristics of each stage of the unsteady cavitation in the experimental visualization, including the growth and breakage of the sheet cavity, as well as the subsequent generation and downstream movement of the cloud cavity. The numerical simulation results are in good agreement with the experimental results.

Pressure fluctuations display the unsteady flow characteristics of the flow field. To validate the unsteady evolution feature of the numerical method, Fig. 3(a) compares the simulated pressure fluctuations with the experimental results of Leroux *et al.*[46] The pressure monitoring point is located in the middle of the suction surface of the hydrofoil ($X/C = 0.7$), consistent with the experimental conditions. The agreement between numerical and experimental data indicates that the numerical simulation accurately captures the unsteady evolution process of the cavitation flow field around the hydrofoil. Figure 3(b) shows the power spectral density (PSD) of the pressure fluctuation process obtained after a fast FFT transformation as a function of frequency. The numerical simulation result (3.82 Hz) is in good agreement with the experimental result (3.625 Hz) within the error range.

III. RESULTS AND DISCUSSION

This study compares and evaluates the performance differences of three DR methods (PCA, ICA, isomap) applied to unsteady cavity flows around a three-dimensional hydrofoil. First, the reconstructed flow field and reconstruction error are compared to evaluate the ability of the three methods to preserve flow field information. In addition, the correlation and clustering similarity before and after DR are compared to evaluate the ability of the three methods to preserve data structure features. Finally, the obtained temporal modes after DR are

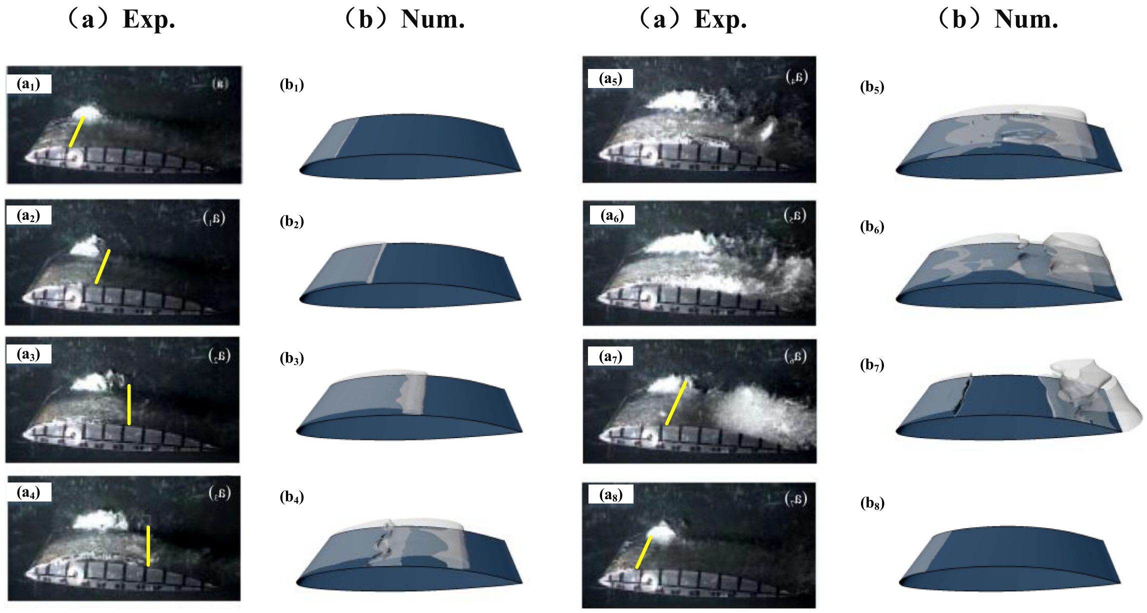


FIG. 2. Comparison of cavity shapes between (a) experimental results[46] and (b) numerical predictions.

compared to evaluate the ability of the three methods to extract underlying coherent structures and flow characteristics.

A. Comparison of reconstruction error

The dimension of the low-dimensional embedding space is a common parameter in DR algorithms, and this study uses a default setting of ten dimensions to enable a fair comparison of different methods. Linear DR methods are simple and do not require hyperparameter selection. However, for the nonlinear isomap algorithm, the only free parameter is k, which represents the number of nearest neighbors between data points. If k is too small, the data may be divided into many disconnected regions, and the manifold structure may not be fully reflected, failing to reveal the true dimensionality of the data and leading to inaccurate DR results. If k is too large, unrelated data points may be included in the neighborhood, and the influence of noise data may increase, thereby affecting the accuracy of DR results. In this study, the value of k is determined by the relative reconstruction error, i.e., the "inflection point" where the relative reconstruction error and k-value curve show a significant decrease.

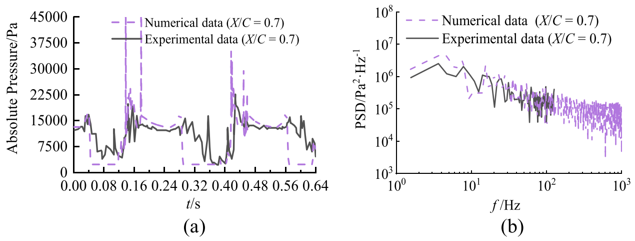


FIG. 3. Comparisons of (a) experimental results[46] with predicted pressure fluctuations at $X/C = 0.7$ and (b) power spectral density (PSD) at $X/C = 0.7$.

Figure 4 shows the relative reconstruction error curve as a function of the number of nearest neighbors (k). The curve has a turning point at $k = 20$, where the relative reconstruction error is minimized, making $k = 20$ the default parameter for subsequent calculations in the isomap algorithm. Furthermore, the behavior of the relative reconstruction error changes on either side of $k = 20$, where it rapidly decreases for k values below 20 and gradually increases for k-values above 20. This indicates that the influence of k-values is more significant when they are too small because the correct manifold structure has not been captured. Moreover, the value of $k = 20$ is twice the dimensionality of the selected low-dimensional space (10), indicating

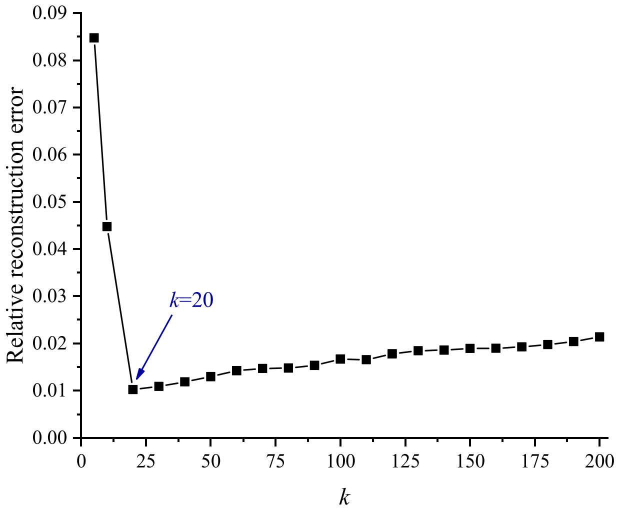


FIG. 4. The relative reconstruction error curve as the k-value changes.

that the number of nearest neighbors used for embedding must be significantly higher than the dimensionality of the low-dimensional space (at least two times). This conclusion has also been found in Csala's research.[20]

Figure 5(a) shows the relative reconstruction errors as the number of modes increases for different DR methods. It is important to note that in the figures of this paper, "Mode" represents the first N modes, while "Mode" refers to the Nth mode. The curves for PCA and ICA show that the relative reconstruction error decreases logarithmically with an increasing number of modes and does not converge. Isomap's reconstruction error is significantly lower than that of PCA and ICA. The curves for isomap show a rapid decrease in the relative reconstruction error for the first three modes, followed by convergence. Therefore, isomap is a better algorithm for DR purposes, as it preserves the most information with the least number of modes. It is worth noting that the values and trends of the relative reconstruction error for PCA and ICA are almost the same. PCA and ICA are both linear DR methods with different objectives in theory. PCA aims to find the DR directions with the maximum variance, while ICA aims find the DR directions with the data into several independent signals. Figure 5(b) depicts the DR directions associated with the first ten modes in PCA and ICA. In the context of linear DR, the outcome of the reduction process is obtained by utilizing a linear transformation matrix to convert the high-dimensional original data into a lower-dimensional space. The eigenvectors of the linear transformation matrix define the new coordinate system for the reduced data within the lower-dimensional space. The orientations of the coordinate axes symbolize the directions of the respective eigenvectors for each mode. The first five modes have very similar DR directions, while the last five modes have differences. However, since the first five modes contain most of the data information, the reconstructed PCA and ICA have similar relative reconstruction errors.

Figure 6 displays velocity and the standard reconstruction error for different DR methods at six typical moments. The cavity region and the wake region show a clear low velocity characteristic. The errors of PCA and ICA mainly distribute in the cavity region and wake region, especially with large errors on the surface of the cavity. This is because these regions contain more nonlinear information, and the surface of the cavity is the interface between water vapor and water, where the nonlinear features are stronger. Isomap error is only present in the surface region of the cavity. Isomap, as a nonlinear DR algorithm, greatly improves the loss of nonlinear information, especially the high preservation of nonlinear information in the wake region.

As shown in Fig. 5, the errors of different DR methods vary with the number of modes. Figure 7 visualizes the spatial distribution of errors at different numbers of modes using RMSE. The errors of PCA and ICA gradually decrease in the wake and cloud cavity regions as the number of modes increases. The errors in the sheet cavity region near the hydrofoil leading edge persist. Isomap's error remains unchanged significantly after the number of modes exceeds three, consistent with the conclusion of Fig. 5. Isomap's error is significantly smaller than that of PCA and ICA in all regions.

B. Comparison of data structure

The relationship between the distances of data points in high-dimensional space and those in low-dimensional space reflects the overall characteristics of the structures of high-dimensional and low-dimensional data. In this study, we first calculated the Euclidean distance between each point and all other points in its dimensional space, and then took the average value to represent the distance feature of each point. Finally, the distance feature of each point was flattened and arranged in a distance sequence. The relationship between high-dimensional and low-dimensional distance sequences includes similarities in trend and numerical value, which are measured by Pearson correlation and DTW, respectively.

Figure 8(a) shows the Pearson correlations between the distance sequences of high-dimensional and low-dimensional data for different numbers of modes. A higher Pearson correlation indicates closer trend features. As shown in the figure, the Pearson correlations of both PCA and ICA algorithms increase with the number of modes, while the Pearson correlation of isomap converges after the number of modes reaches 3. Comparing the absolute values of Pearson correlations, it can be seen that PCA performs the best in preserving the overall data features, followed by isomap and then ICA. This indicates that the

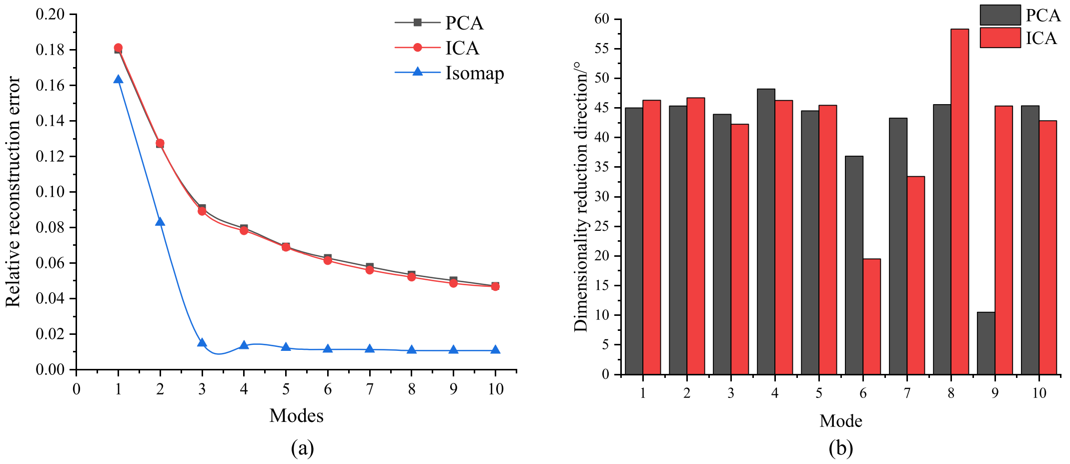


FIG. 5. (a) Relative reconstruction error with a different number of modes. (b) First ten modes of PCA and ICA and their corresponding DR directions.

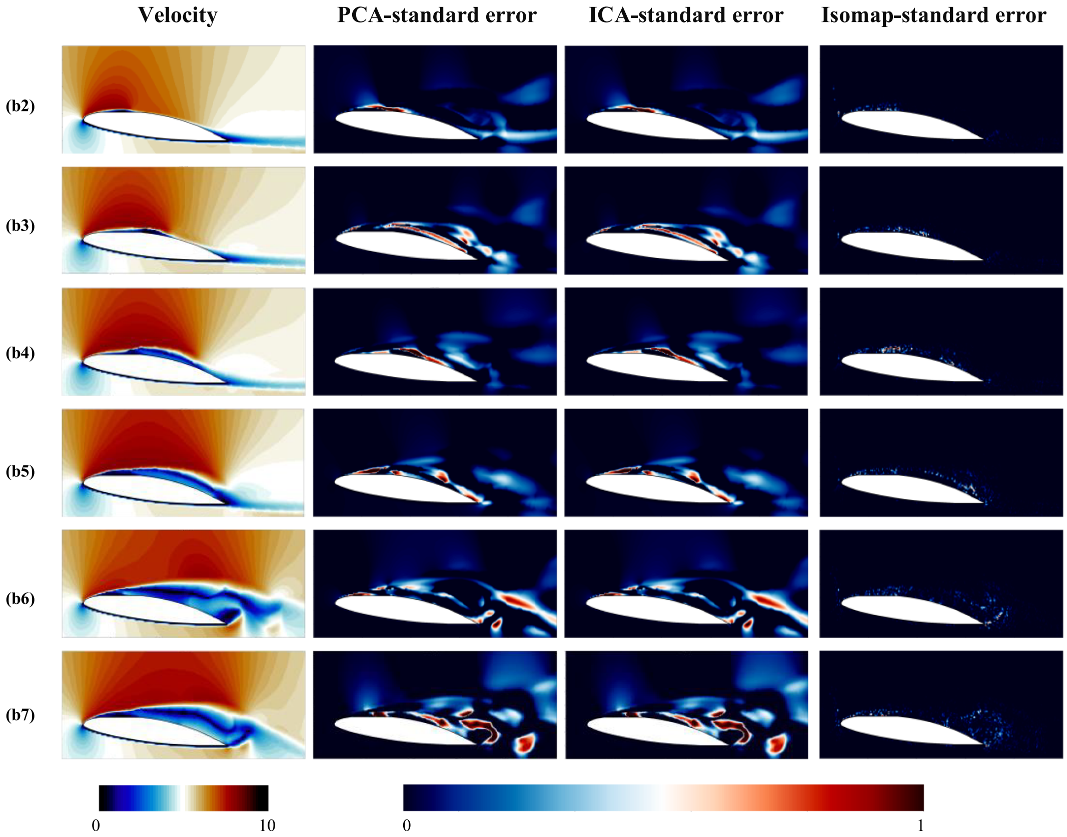


FIG. 6. Velocity and standard reconstruction error for different DR methods at six typical moments.

PCA algorithm shows the best performance in preserving the overall data structure.

Figure 8(b) shows the DTW between the high-dimensional and low-dimensional distance sequences for different numbers of modes. Interestingly, the DTW of PCA and ICA algorithms decreases as the number of modes increases, indicating that the two distance sequences become more similar. However, isomap shows the opposite performance. In addition, the distribution range of the DTW values of different DR algorithms is also noteworthy. The DTW of PCA gradually decreases within a large range (10–500), while that of ICA gradually decreases within a small range (2629–2631) and that of isomap gradually increases within a large range (600–2600). This phenomenon is worth further investigation.

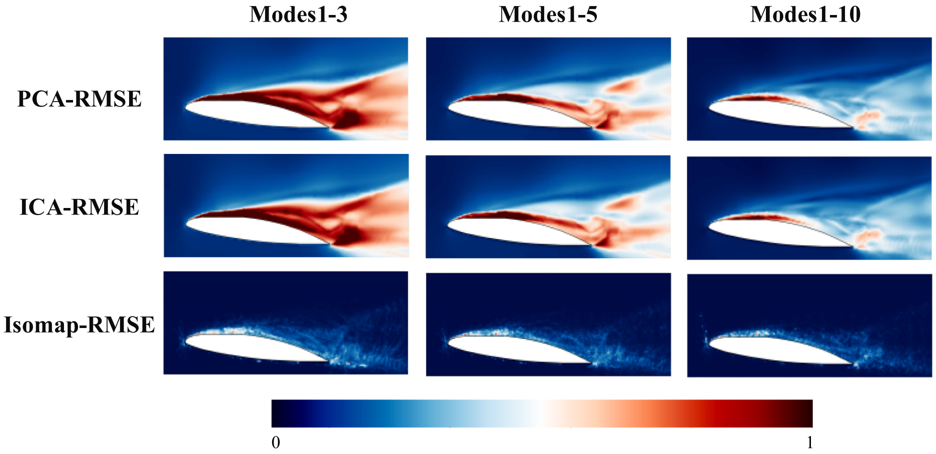


FIG. 7. Spatial distribution of RMSE at different numbers of modes for different DR methods.

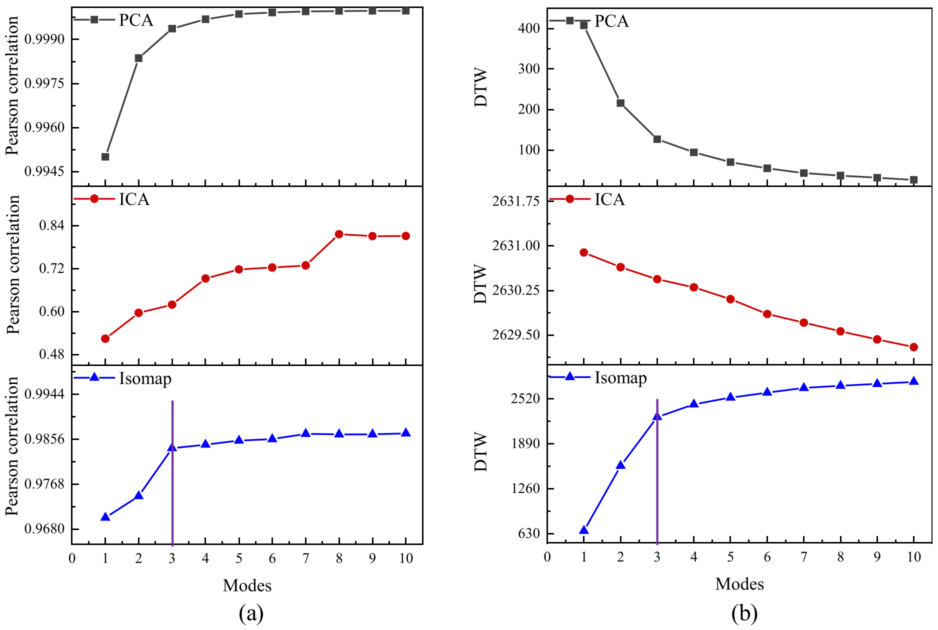


FIG. 8. (a) Pearson correlations between the distance sequences of high-dimensional and low-dimensional data for different numbers of modes. (b) DTW between the high-dimensional and low-dimensional distance sequences for different numbers of modes.

When we project high-dimensional data into a low-dimensional space, sometimes there may be changes in the distances between points in the space. The relationship between the distances among the points in the low-dimensional space and those in the original high-dimensional space depends on the DR technique used and the characteristics of the data itself. In some cases, the distances among the points in the low-dimensional space may reflect those in the original high-dimensional space, which is often observed in linear DR techniques. PCA is a linear transformation method that can preserve the principal variance direction of the dataset, and the Euclidean distance among the points in the low-dimensional space can typically reflect that in the original high-dimensional space. Although ICA is also a linear DR technique, it relies on a nonlinear transformation of the matrix, achieved through the mixing and unmixing matrices of independent components. Thus, due to the nonlinear transformation, the positions of the reduced data points in the space may deviate greatly from those of the original data points. In the case of the isomap algorithm, it is a manifold learning technique. Isomap constructs the connections among data points by calculating the geodesic distance between each pair of data points in the high-dimensional space and then maps the high-dimensional data to the low-dimensional space using a method similar to MDS. The geodesic distance in the isomap method is measured along the curve path of the data manifold in the high-dimensional space, rather than the Euclidean distance in a straight line. Therefore, in the low-dimensional space after DR, the neighboring data points measured by the Euclidean distance may be closer, while the distance between distant data points may be stretched, leading to a large difference in the distances between the points in the reduced space and those in the original high-dimensional space. However, this does not necessarily mean that the DR effect of isomap is poor. The evaluation of a DR algorithm needs to be considered comprehensively.

Based on the above discussions, we can use Pearson correlation to measure the overall structure of the data, which is not influenced by the numerical values of the distances among points but considers the global change characteristics. We cannot use DTW to measure the overall structure of the data because we cannot use the numerical values of the distances among points to measure the overall structure of the data at a deep level.

In Fig. 8, we evaluated the differences in the overall structure of data among different DR algorithms. In this section, we propose the UsK-means method to identify the micro-structural features of data through clustering. The uniform sampling method used in this algorithm ensures comparability between clustering results. Three clustering evaluation metrics are used to measure the similarity between clustering results and further evaluate the similarity of micro-structural features of data. Although these metrics have different principles, their goals are the same: the larger the value, the more similar the clustering results. Since the number of clusters is an unknown parameter for clustering algorithms, we use the three clustering similarity metrics to search for the optimal number of clusters. Figure 9 shows the FMI, NMI, and ARI of three DR algorithms under different numbers of clusters. The patterns of change in the three clustering similarity metrics are similar, and they reach a convergence state after 150 clusters. Therefore, we choose 150 as the default number of clusters.

Figure 10 shows the FMI, NMI, and ARI of the three DR algorithms under different numbers of modalities. The three clustering similarity metrics exhibit similar patterns as the number of modalities increases. As the number of modalities increases, the information contained in the reduced space gradually increases and approaches that of the original high-dimensional space, leading to an increase in clustering similarity. PCA still demonstrates better capability of preserving local data structure, followed by isomap, and finally ICA. It is noteworthy that the clustering similarity between PCA and isomap is almost identical for the first three modalities. However, for modalities beyond three, the clustering similarity of isomap converges, while that of PCA continues to increase.

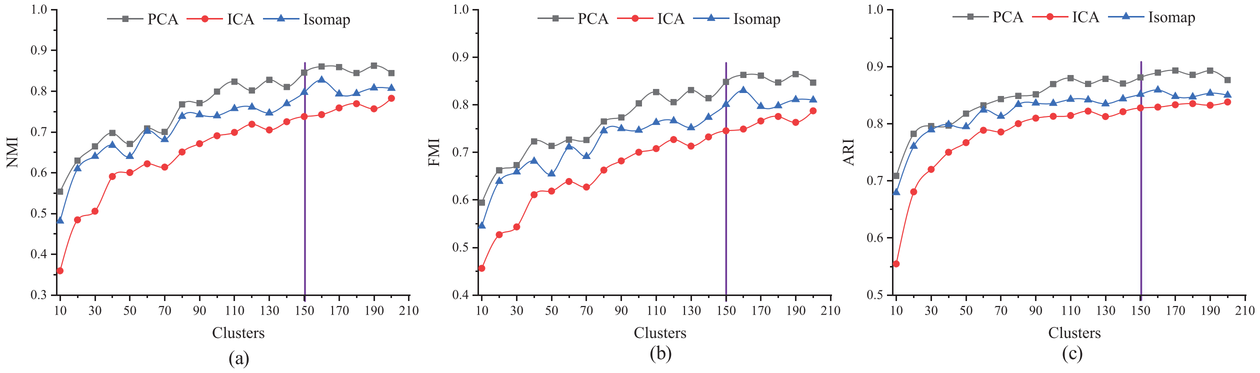


FIG. 9. Clustering similarity of three DR algorithms under different numbers of clusters: (a) NMI, (b) FMI, and (b) ARI.

Figure 11 displays the spatial distribution of clustering results of different DR methods under different modal numbers. The local structural features of data are visualized in the flow field space using different colors to represent different clusters. However, we chose a gradient color bar as the color mode, as it is not feasible to visualize 150 different colors, which would make the visualization less interpretable. This compromise solution can still roughly convey the differences in local structural features of data. The reference clustering results are obtained from the original high-dimensional space. With the increase in modal numbers, the clustering results of PCA and ICA gradually become more reasonable and approach the reference results. Isomap reaches a convergence state in the clustering results after the first three modalities, and there is no significant change in the clustering results with the increase in modal numbers. Obviously, the clustering results of PCA are the closest to the reference results, followed by isomap, and then by ICA. It is worth noting that the clustering results of isomap are smoother, perhaps because this algorithm is based on nearest neighbors and therefore merges features that are close in space.

C. Comparison of flow mode

In fluid mechanics, DR is used to reduce the time dimension of data and obtain a low-dimensional representation that captures the underlying coherent structures and flow features of the fluid, providing a better understanding of the essence of fluid motion. This is one of the main reasons why modal decomposition techniques such as PCA proper orthogonal decomposition (POD) are popular in fluid mechanics.

To understand the meaning of different modes after DR, the contribution of each mode is usually ranked, as it allows us to obtain the percentage of information that each mode represents for the original flow field. Various statistical indicators such as variance explained ratio, chi-square value, and mutual information are commonly used for contribution ranking. Among them, variance ratio is the most commonly used ranking indicator and has been used in combination with PCA. In this paper, it is also applied to the ranking task of ICA and isomap. Figure 12 shows the variance ratio of different modes under different DR methods. PCA and isomap exhibit the same distribution pattern, with mode 1 having the highest variance ratio. The higher the mode number, the less important. It is worth noting that the variance ratio of ICA is the same for each mode. This is because the data may be uniformly distributed across all dimensions. In this case, there is no major direction of change, and all dimensions are equally important in explaining the data variation.

Figure 13 shows the flow modes obtained using different DR methods. Based on the comparison of the velocity field distribution shown in Fig. 6, the distribution characteristics of PCA and isomap modes can be observed. The distribution characteristics of the modes are mainly reflected in the coherent structures of positive and negative

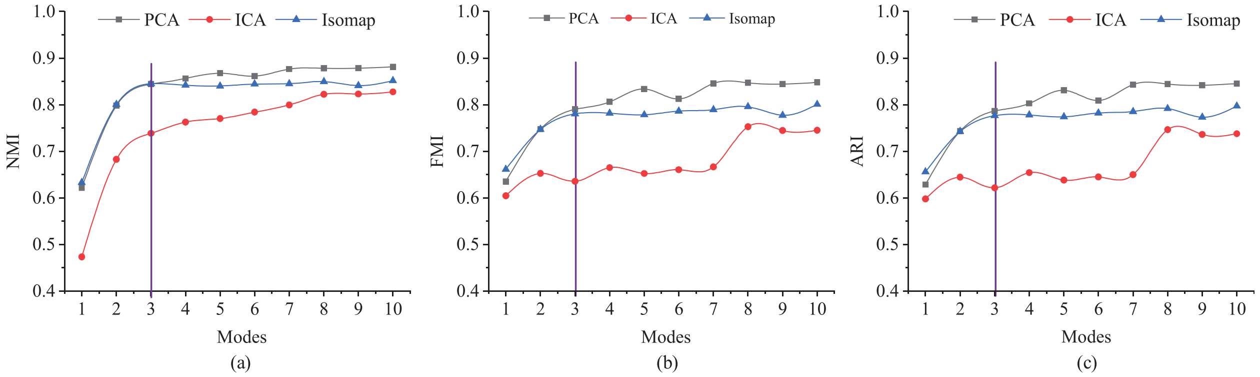


FIG. 10. Clustering similarity of three DR algorithms under different numbers of modes: (a) NMI, (b) FMI, and (c) ARI.

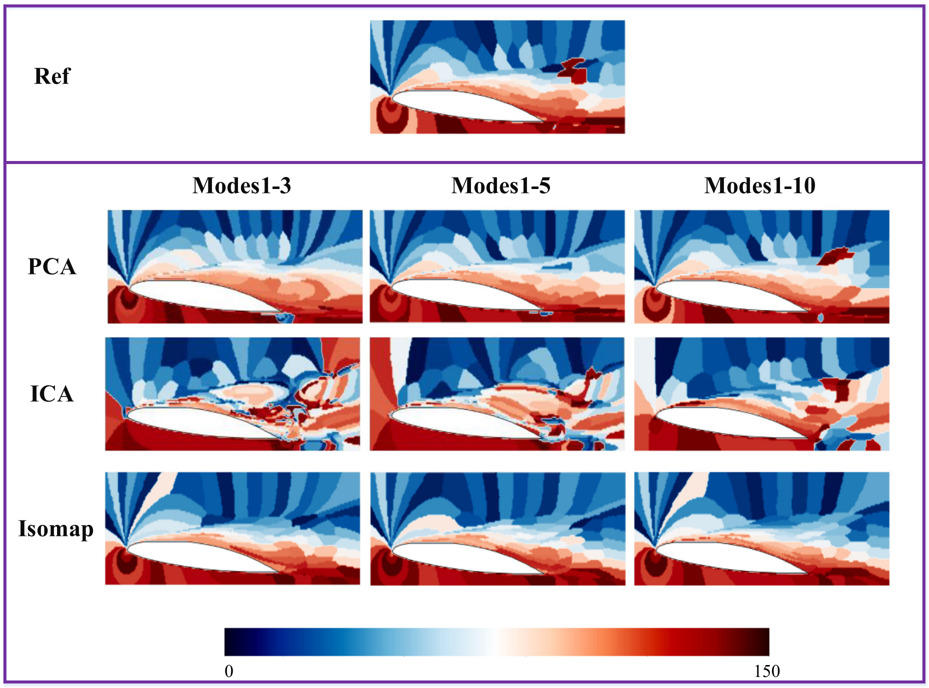


FIG. 11. Spatial distribution of clustering results at different numbers of modes for different DR methods.

values. In mode 1, the coherent structure of negative values mainly represents the high-speed region of the velocity field. The negative value region of PCA is mainly located above the hydrofoil, while that of isomap is mainly located behind the hydrofoil. The coherent structure of positive values represents the low-speed region, which is also the region of cavity. The positive value region of PCA is more dispersed, while that of isomap is more concentrated. In mode 2, the coherent structure distribution of PCA and isomap is similar, and the coherent structure of positive values represents the region of cavity. In mode 3, there are differences in the coherent structure distribution of PCA and isomap. PCA focuses more on the sheet cavity region above the hydrofoil, while isomap focuses more on the cloud cavity region behind the hydrofoil. In modes 4–10, there is a stable coherent structure at the leading edge of the hydrofoil, followed by the alternating distribution of positive and negative coherent structures of different scales in other regions of the hydrofoil. As different modes contain temporal process information, this phenomenon is considered as a stable coherent structure and subsequent decomposition process. The coherent structure decomposition of PCA is more concentrated in the sheet cavity region, while that of isomap is more concentrated in the cloud cavity region. Overall, PCA focuses more on the features of sheet cavity, while isomap focuses more on the features of cloud cavity. Sheet cavity occupies most stages in the entire cavity evolution process and contains more energy and time information, which is the goal of PCA, i.e., to retain most of the information in the flow field. Cloud cavity represents the most complex flow process in the entire cavity evolution process and contains more nonlinear information. This is the reason why isomap can better retain this nonlinear information.

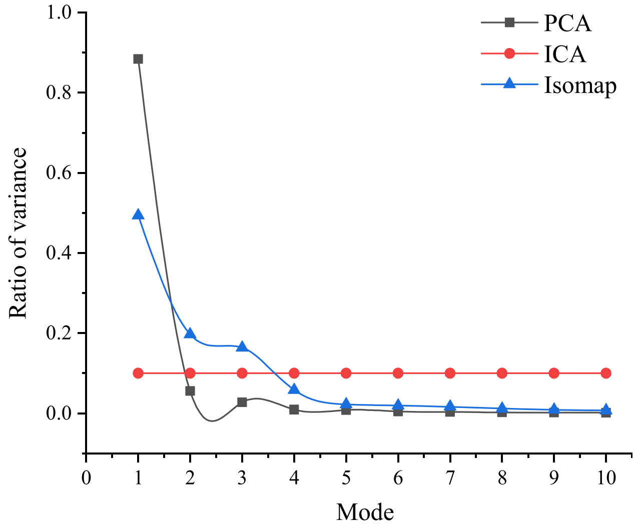


FIG. 12. Variance ratio at different modes for different DR methods.

It is worth noting that Fig. 12 proves the equal importance of ICA in obtaining different flow modes. This means that we cannot rank them directly. Therefore, in comparing with the other two algorithms, this paper manually sorted the modes based on the mode distribution of the other two methods. This sorting method is only for better comparison of the three algorithms and does not have any reference value in practical applications. The mode distribution of ICA is closer to PCA, and the conclusion is similar to that of PCA because they are both linear DR methods. In ICA, modes are unordered.

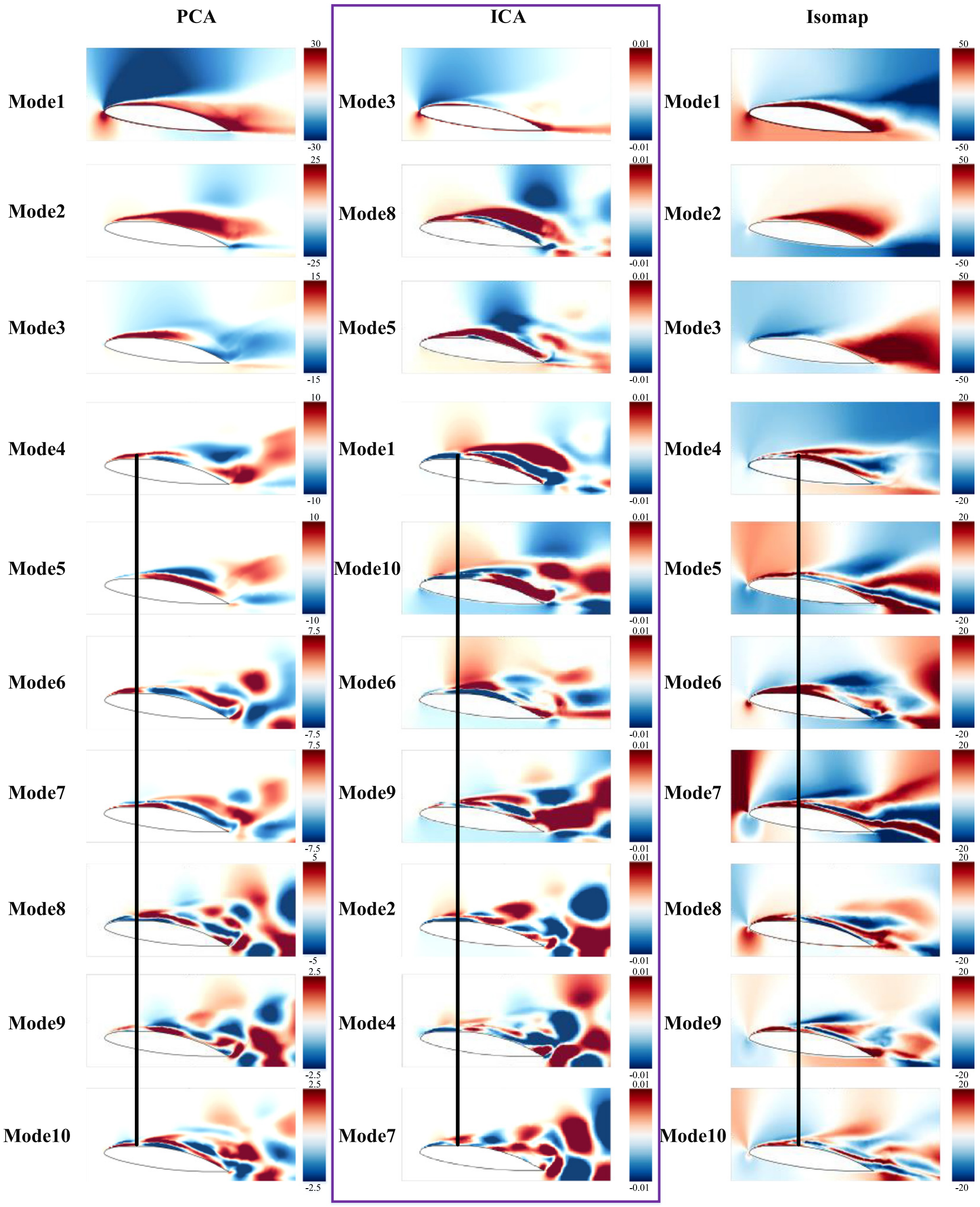


FIG. 13. Flow mode at different modes for different DR methods.

This makes it impossible for us to systematically analyze the similarity between different modes through changes in mode numbering, making the physical interpretability of ICA modes lower than that of PCA. However, an important advantage of ICA is that it can find independent components, which may be more useful than variance in some fields. Mathematically, independence is a more stringent constraint, so the modes obtained through ICA may be more meaningful in certain situations.

IV. CONCLUSION

In this paper, we improve the evaluation system for DR methods and comprehensively compare and evaluate the performance differences of three DR methods (PCA, ICA, isomap) applied to the unstable cavitation flow on a 3D hydrofoil. The main conclusions are as follows:

(1) Three different error criteria are used to evaluate the reconstruction errors of the three DR methods. The relative reconstruction error, based on the Frobenius norm criterion, suggests that PCA and ICA exhibit similar levels of relative reconstruction errors. Isomap is considered a superior algorithm to both PCA and ICA, as it preserves the maximum amount of information with the fewest number of modes. The standard reconstruction error shows that isomap significantly improves the loss of nonlinear information, especially in the wake flow region. The RMSE indicates that the error of isomap is significantly smaller than that of PCA and ICA in all flow field regions, and the error remains constant after the number of modes exceeds three.
(2) Pearson correlation and DTW are used to evaluate the degree of preservation of the overall characteristics of the data structure of the three DR methods. In the process of DR, the magnitudes of distances between points in space may undergo changes, which can impact the evaluation outcomes of DTW. Pearson correlation measures the overall structure of the data, considering global patterns of variation. It is regarded as a superior indicator for quantifying the overall characteristics of the data. PCA performs the best in preserving the overall data characteristics, followed by isomap and then ICA.
(3) Cluster similarity is used to evaluate the degree of preservation of the local characteristics of the data structure of the three DR methods. The UsK-means method proposed in this paper guarantees the comparability between the clustering results of high-dimensional data and low-dimensional data. PCA still shows better ability to preserve local data structure, followed by isomap, and then ICA. For the first three modes, the cluster similarity between PCA and isomap is almost the same. For more than three modes, the cluster similarity of isomap converges, while that of PCA continues to increase. Isomap, based on the nearest neighbors, merges features that are close to each other in space.
(4) Flow modes are used to evaluate the differences in the recognition of flow characteristics of the three DR methods. PCA focuses more on identifying the features of sheet cavities because they occupy most of the information in the flow field. Isomap focuses more on identifying the features of cloud cavities because the behavior of cloud cavities contains more nonlinear information. The modes in ICA cannot be sorted or ranked, as all modes are equally important in explaining the variations in the data. This results in a lower level of physical interpretability for the modes obtained from ICA compared to the modes in PCA. However, the independent modes obtained through ICA hold greater mathematical significance.

ACKNOWLEDGMENTS

This work was partially funded by the National Natural Science Foundation of China (Grant Nos. 52061135107 and 52071062), the Fundamental Research Fund for the Central Universities (Grant No. DUT20TD108), the Liao Ning Revitalization Talents Program (Grant No. XLYC1908027), the Dalian Innovation Research Team in Key Areas (Grant No. 2020RT03), the Dalian High-level Talent Innovation Support Program (Youth Technology Star) (Grant No. 2020RQ003), and the computation support of the Supercomputing Center of Dalian University of Technology.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Guiyong Zhang: conceptualization (equal); data curation (equal); funding acquisition (lead); investigation (equal); methodology (equal); project administration (lead); resources (equal); software (equal); validation (lead); visualization (equal); writing—original draft (lead); and writing—review and editing (lead). **Zihao Wang:** conceptualization (lead); data curation (equal); funding acquisition (lead); investigation (equal); methodology (equal); project administration (lead); resources (equal); software (equal); validation (lead); visualization (equal); writing—original draft (lead); and writing—review and editing (equal). **Huakun Huang:** formal analysis (equal); investigation (equal); methodology (equal); resources (equal); software (equal); validation (equal); and writing—review and editing (equal). **Hang Li:** conceptualization (equal); formal analysis (equal); methodology (equal); software (equal); and visualization (equal). **Tiezhi Sun:** data curation (equal); formal analysis (equal); funding acquisition (equal); project administration (lead); supervision (lead); visualization (equal); and writing—review and editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] A. Pollard, L. Castillo, L. Danaila, and M. Glauser, "Challenges for large eddy simulation of engineering flows," in *Whither Turbulence and Big Data in the 21st Century?* (Springer, Cham, Switzerland, 2017).
[2] S. L. Brunton, B. R. Noack, and P. Koumoutsakos, "Machine learning for fluid mechanics," Annu. Rev. Fluid Mech. **52**, 477–508 (2020).
[3] A. Man, M. Jadidi, A. Keshmiri, H. Yin, and Y. Mahmoudi, "A divide-and-conquer machine learning approach for modeling turbulent flows," Phys. Fluids **35**(5), 055110 (2023).
[4] B. Haddadi, C. Jordan, and M. Harasek, "Cost efficient CFD simulations: Proper selection of domain partitioning strategies," Comput. Phys. Commun. **219**, 121–134 (2017).
[5] C. Liu, R. Jiang, D. Wei, C. Yang, Y. Li, F. Wang, and X. Yuan, "Deep learning approaches in flow visualization," Adv. Aerodyn. **4**, 17 (2022).
[6] B. R. Noack, K. Afanasiev, M. Morzyński, G. Tadmor, and F. Thiele, "A hierarchy of low-dimensional models for the transient and post-transient cylinder wake," J. Fluid Mech. **497**, 335–363 (2003).
[7] B. R. Noack, P. Papas, and P. A. Monkewitz, "The need for a pressure-term representation in empirical Galerkin models of incompressible shear flows," J. Fluid Mech. **523**, 339–365 (2005).
[8] K. Taira, S. L. Brunton, S. Dawson, C. W. Rowley, T. Colonius, B. J. McKeon, O. T. Schmidt, S. Gordeyev, V. Theofilis, and L. S. Ukeiley, "Modal analysis of fluid flows: An overview," AIAA J. **55**(12), 4013–4041 (2017).

[9] Z. Wang, G. Zhang, T. Sun, C. Shi, and B. Zhou, "Data-driven methods for low-dimensional representation and state identification for the spatiotemporal structure of cavitation flow fields," Phys. Fluids **35**, 033318 (2023).

[10] M. Karami, H. Hangan, L. Carassale, and H. Peerhossaini, "Coherent structures in tornado-like vortices," Phys. Fluids **31**(8), 085118 (2019).

[11] X. Xing, M. H. Dao, B. Zhang, J. Lou, W. S. Tan, Y. Cui, and B. C. Khoo, "Fusing sensor data with CFD results using gappy POD," Ocean. Eng. **246**, 110549 (2022).

[12] K. Taira, M. S. Hemati, S. L. Brunton, Y. Sun, K. Duraisamy, S. Bagheri, S. T. M. Dawson, and C.-A. Yeh, "Modal analysis of fluid flows: Applications and outlook," AIAA J. **58**, 998–1022 (2020).

[13] M. A. Mendez, J. Dominique, M. Fiore, F. Pino, P. Sperotto, and J. Berghe, "Challenges and opportunities for machine learning in fluid mechanics," arXiv:2202.12577 (2022).

[14] N. Omata and S. Shirayama, "A novel method of low-dimensional representation for temporal behavior of flow fields using deep autoencoder," AIP Adv. **9**, 015006 (2019).

[15] H. Eivazi, H. Veisi, M. H. Naderi, and V. Esfahanian, "Deep neural networks for nonlinear model order reduction of unsteady flows," Phys. Fluids **32**(10), 105104 (2020).

[16] F. Tauro, S. Grimaldi, and M. Porfiri, "Unraveling flow patterns through nonlinear manifold learning," PLoS One **9**(3), e91131 (2014).

[17] A. Ehlert, C. N. Nayeri, M. Morzynski, and B. R. Noack, "Locally linear embedding for transient cylinder wakes," arXiv:1906.07822 (2019).

[18] M. Mendez, "Linear and nonlinear dimensionality reduction from fluid mechanics to machine learning," Meas. Sci. Technol. **34**(34), 042001 (2023).

[19] L. Pyta and D. Abel, "Nonlinear model reduction of the Navier–Stokes-equations," in *Proceedings of the American Control Conference (ACC)* (IEEE, 2016), pp. 5249–5254.

[20] H. Csala, S. T. M. Dawson, and A. Arzani, "Comparing different nonlinear dimensionality reduction techniques for data-driven unsteady fluid flow modeling," Phys. Fluids **34**, 117119 (2022).

[21] Z. Wang, X. Zhang, Y. Wang, and J. Liu, "Comparative study between turbulence models in unsteady cavitating flow with special emphasis on shock wave propagation," Ocean. Eng. **240**, 109988 (2021).

[22] C. E. Brennen, *Fundamentals of Multiphase Flow* (Cambridge University Press, Cambridge, 2005).

[23] J. P. Franc and J. M. Michel, *Fundamentals of Cavitation* (Springer, Berlin, 2006), p. 76.

[24] T. Sun, Z. Wang, L. Zou, and H. Wang, "Numerical investigation of positive effects of ventilated cavitation around a NACA66 hydrofoil," Ocean. Eng. **197**, 106831 (2020).

[25] T. Sun, Z. Wang, L. Zou, Z. Sun, and Z. Zong, "Numerical Investigation of the Natural and Ventilated Cavitation Dynamics Around NACA66 Hydrofoil," in *The 29th International Ocean and Polar Engineering Conference* (OnePetro, 2019).

[26] J. A. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction* (Springer, 2007).

[27] R. Vidal, Y. Ma, and S. S. Sastry, *Generalized Principal Components Analysis* (Springer, 2016).

[28] L. Van Der Maaten, E. Postma, and J. van den Herik, "Dimensionality reduction: A comparative review," J. Mach. Learn. Res. **10**(66–71), 13 (2009).

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," J. Mach. Learn. Res. **12**, 2825–2830 (2011).

[30] K. Pearson, "On lines and planes of closest fit to systems of points in space," London Edinburgh Dublin Philos. Mag. J. Sci. **2**(11), 559–572 (1901).

[31] J. Karhunen, E. Oja, L. Wang, R. Vigario, and J. Joutsensalo, "A class of neural networks for independent component analysis," IEEE Trans. Neural Networks **8**, 486 (2000).

[32] E. Bingham and A. Hyvärinen, "A fast fixed-point algorithm for independent component analysis of complex valued signals," Int. J. Neur. Syst. **10**, 1–8 (2000).

[33] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," Science **290**(5500), 2319–2323 (2000).

[34] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms* (MIT Press, 2022).

[35] W. S. Torgerson, "Multidimensional scaling. I. Theory and method," Psychometrika **17**(4), 401–419 (1952).

[36] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley, "Locally linear embedding and its variants: Tutorial and survey," arXiv:2011.10925 (2020).

[37] T. Franz, R. Zimmermann, S. Görtz, and N. Karcher, "Interpolation-based reduced-order modelling for steady transonic flows via manifold learning," Int. J. Comput. Fluid Dyn. **28**(3–4), 106–121 (2014).

[38] L. K. Saul and S. T. Roweis, "Think globally, fit locally: Unsupervised learning of low dimensional manifolds," J. Mach. Learn. Res. **4**, 119–155 (2003).

[39] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing* (Springer, 2009), pp. 1–4.

[40] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. Acoust., Speech, Signal Process. **26**(1), 43–49 (1978).

[41] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (University of California Press, Berkeley, 1967), Vol. 1, pp. 281–297.

[42] Z. Wang, X. Xing, T. Sun, and G. Zhang, "Segmentation of unsteady cavitation flow fields based on multivariate spatiotemporal hierarchical clustering," Phys. Fluids **35**(5), 053317 (2023).

[43] E. B. Fowlkes and C. L. Mallows, "A method for comparing two hierarchical clusterings," J. Am. Stat. Assoc. **78**(383), 553–569 (1983).

[44] M. Meilă and J. Shi, "A random walks view of spectral segmentation," in *International Workshop on Artificial Intelligence and Statistics* (PMLR, 2001), pp. 203–208.

[45] L. Hubert and P. Arabie, "Comparing partitions," J. Classif. **2**, 193–218 (1985).

[46] J. B. Leroux, J. A. Astolfi, and J. Y. Billard, "An experimental study of unsteady partial cavitation," J. Fluid Eng. **126**, 94–101 (2004).

[47] F. Menter, M. Kuntz, and R. Langtry, "Ten years of industrial experience with the SST turbulence model," Turbul. Heat Mass Transfer **4**(1), 625–632 (2003).

[48] G. H. Schnerr and J. Sauer, "Physical and numerical modeling of unsteady cavitation dynamics," in Fourth International Conference on Multiphase Flow, 2001.

[49] M. Passandideh-Fard and E. Roohi, "Transient simulations of cavitating flows using a modified volume-of-fluid (VOF) technique," Int. J. Comput. Fluid Dyn. **22**(1–2), 97–114 (2008).

[50] P. Sagaut, *Large Eddy Simulation for Incompressible Flows* (Springer, 2002).

[51] O. Coutier-Delgosha, R. Fortes-Patella, and J. L. Reboud, "Evaluation of the turbulence model influence on the numerical simulations of unsteady cavitation," J. Fluid Eng. **125**, 38–45 (2003).