Data Fusion using gappy-POD

Charles E. Tinney* Yingjun Zhao-Dubuc†
The University of Texas at Austin, Austin, TX 78713, USA
Oliver Schmidt‡
University of California at San Diego, La Jolla, CA 92093, USA
Christophe Bogey§
University of Lyon, Lyon, 69134 Ecully Cedex, France
Yang Zhang¶ Louis N. Cattafesta‖
Illinois Institute of Technology, Chicago, IL 60616, USA

A framework for fusing numerical and experimental databases using heterogeneous forms of the gappy-POD is presented in order to reconstruct real observations from reduced sensor sets. The problem is demonstrated in the context of the jet noise problem where the observation space comprises wavefronts corresponding to the unsteady pressure field within the hydrodynamic periphery of a Mach 3 jet flow. The data fusion approach leverages a numerical model of the unsteady density field (generated by way of large eddy simulation) for training, followed by entries from experiments (high speed schlieren) for real-time reconstruction. The compatibility of these databases is evaluated to ensure that observations are generated using the smallest number of synthesized basis functions. The input to the reconstruction is a sensor set from experimentally captured schlieren images of the same jet flow. A further simplification utilizes heterogeneous forms of the gappy-POD so that only a reduced sensor set is required for the reconstruction. This comprises the XX-topos and YX forms where the linear algebraic system of equations are formed using the eigenvectors and expansion coefficients, respectively while the gappy sensor set is determined using a random number generator. An error analysis reinforces the general requirement that $2n+1$ sensors are needed to properly resolve the nth POD mode. Sample reconstructions of the density field are performed in real time using only 10% of available POD modes and 20% of the sensors and are shown to generate qualitatively satisfactory images of wavefronts evolving across space and time.

I. Nomenclature

a, b, c, d, e	=	POD expansion coefficients
a_∞	=	sound speed of ambient gas
a_j	=	sound speed of jet gas
$\delta_x, \delta t$	=	space and time increments
D_j	=	jet diameter
E	=	sum of eigenvalues
ϵ	=	mean square error
I	=	identity matrix
λ, ω	=	eigenvalues
Λ_1	=	normalized eigenvalue

*Senior Research Scientist; Applied Research Laboratories, ✉ charles.tinney@arlut.utexas.edu, Associate Fellow AIAA
†Research Associate; Applied Research Laboratories, ✉ yingjun.zhao@arlut.utexas.edu, Member AIAA.
‡Associate Professor; Mechanical and Aerospace Engineering, ✉ oschmidt@ucsd.edu, Senior Member AIAA.
§CNRS Research Director; Fluid Mechanics and Acoustics Laboratory, ✉ christophe.bogey@ec-lyon.fr, Senior Member AIAA.
¶Research Assistant Professor; Mechanical Materials & Aerospace Engineering, Member AIAA.
‖John G. and Jane E. Olin Endowed Department Chair, Mechanical Materials & Aerospace Engineering, Fellow AIAA.

M_j = Mach number
M_a = acoustic Mach number
M_c = convective Mach number
n, q = POD mode numbers
p = numerical database
$p_\dagger$ = reference pressure
$\underline{P}$ = experimental database
$\tilde{p}$ = low-order reconstruction of p
$\hat{p}$ = gappy-POD reconstruction of $\hat{p}$
$\mathbf{P}$ = gappy-POD output matrix
$\mathbf{\Pi}$ = Eigenvalue cumulative sum
$\mathbf{\Phi}, D$ = gappy-POD input matrix
Π = normalized cumulative sum of eigenvalues
$\phi,\ \psi$ = eigenvectors
R = POD covariance matrix
R_g = gas constant
χ = data split ratio
$\rho_\dagger$ = reference density
U_j = jet exit velocity
x, y = independent variables, often space and time
x', y' = separation in x and y
$\tilde{\mathbf{x}},\ \tilde{\mathbf{z}}$ = gappy sensor sets
$|\ .\ |$ = cardinality of a vector
x_1, x_2, x_3 = Euclidean space vectors
ξ = arbitrary constant

II. Introduction

Solutions to real world problems in engineering and physics are now generating overwhelming amounts of information that are often too large and cumbersome to sift through. As such, there continues to be a growing demand for data reduction methods capable of reducing the scale and complexity of these databases all the while retaining the essential physics needed to solve problems of practical importance. Robust and reliable computer models are also now being developed with the kind of sophistication that no longer requires rigorous experimental validation. Depending on the complexity of the problem, the overhead needed to generate these numerical solutions can be overwhelming. One such way of reducing this expense is by way of data fusion whereby multiple sources of data are leveraged to generate an outcome that is more superior than its individual parts. A growing number of demonstrations of data fusion applied to problems in fluid mechanics can be found throughout the literature and involve well known techniques such as proper orthogonal decomposition (POD) and stochastic estimation,[1,2] dynamic mode decomposition (DMD),[3] artificial and deep neural networks,[4,5] Bayesian statistics,[6] gappy-POD,[7,8] as well as an assortment of other methods and combinations thereof.[9–11]

The motivation of this study is to provide a mathematical framework for fusing numerically generated data with sparse sensor sets from field measurements for real time reconstruction of spatially and temporally resolved field data. The approach is not limited to problems in fluid mechanics and could easily be extended to other disciplines such as structural health monitoring. The approach builds upon the gappy-POD formulation of Everson and Sirovich[12] which has seen widespread use in repairing missing or damaged data[13–16] or for optimizing the location of reduced sensors sets.[17–20] Our eventual goal is to guide the refinement of less sophisticated numerical models that still encompass sufficient fidelity for distilling the essential physics. These numerical models can be used for offline training and then combined with in situ measurements from strategically placed sensors for real time reconstructions; information that was truncated from the model is returned by way of the sparse sensor sets. For now, we will focus on the process of fusing numerical and experimental databases using heterogeneous forms of the gappy-POD. An outline of this paper is as follows. The conventional POD is first described followed by an outline of the mathematical framework for the *XX-topos* and *YX* forms of the gappy-POD for data fusion using the naming convention of Zhao and Tinney.[21] These are the only known heterogeneous forms of the gappy-POD that support data fusion where the input to the reconstructed database comprises a reduced sensor set, relative to the training data. The technique is then demonstrated using signatures corresponding to propagating wavefronts in the near-field of a high Mach number supersonic jet.

III. Data fusion using gappy-POD with conventional kernels

Formulations for Lumley's POD[22–24] are found throughout the open literature[25–29] with applications to problems involving both numerically and experimentally generated data. We begin by providing a short review of the conventional POD framework in order to establish a nomenclature for subsequent discussion. The starting point is a kernel which is formed in a space with independent coordinates (x and y) where $p \in \mathbb{C}^{x,y}$. In this case, x' denotes a separation along x while the p is confined by a Hilbert space so that,

$$R(x,x') = \sum_{y} p(x,y)p^*(x',y) \tag{1}$$

A shorter notation that uses inner product operations $\langle\rangle$ is written as $R_{xx'} = \langle p(x,y), p(x',y)\rangle_{xx'}$ and will be used interchangeably throughout this discussion. Eq. (1) is a self-adjoint covariance matrix $\det([R_{xx'}]^{-1}) = 0$ that can be reconstructed from a denumerable set of solutions obtained from the following generalized eigenvalue problem

$$R(x,x')\phi(x',n) = \lambda(n)\phi(x,n) \tag{2}$$

In the context of machine learning, the formation of $R(x,x')$ is the process of training, and is accomplished here using an entirely numerically generated dataset defined by $p(x,y)$. Thus, a solution space comprised of eigenvalues $\lambda(n)$ and eigenvectors $\phi(x,n)$ are solutions based on numerically / synthesized observations. These synthesized eigenvectors are used to expand p as follows,

$$a(y,n) = \sum_{x} p(x,y)\phi(x,n) \tag{3}$$

$$\hat{p}^{(m)}(x,y) = \sum_{n}^{m} a(y,n)\phi(x,n) \tag{4}$$

where the reduced-order model $\hat{p}$ is formed using any combination of solutions $n \in m$ while the number of available solutions is confined by the rank of the kernel, $m = \text{rank}(R_{xx'})$.

Alternatively, the data can be reconstructed by fusing the numerically generated POD basis set with field measurements of the same shared observations; these field measurements are statistically independent of the numerically generated database and are labeled $\underline{P}(x,y)$. A further simplification can be performed using a gappy sensor set where $\tilde{\mathbf{x}}$ is a subset of sensors in x so that $\tilde{\mathbf{x}} \subseteq x$; missing (or damaged) sensors $\tilde{\mathbf{z}}$ are unique so that $\tilde{\mathbf{x}} \neq \tilde{\mathbf{z}}$, $\tilde{\mathbf{z}} \subseteq x$, and thus $|\tilde{\mathbf{x}}| + |\tilde{\mathbf{z}}| = |x|$. We use $|x|$ to define the cardinality of x (the number of elements). In order to boost the accuracy of the reconstructed variable on account of the gappy sensor set we will resort to heterogeneous forms of the gappy-POD which leverage a linear algebraic system of equations to minimize the error between the gappy sensor set $p(\tilde{\mathbf{x}},y)$ and the reconstructed variable $\hat{p}^{(m)}(x,y)$.

Following Zhao and Tinney,[21] three heterogeneous forms of the gappy-POD are available to choose from and are referred to as *XX-topos*, *XX-chronos* and *YX*; formulations for these have been extended to the spectral domain[16,20] and for occasions when higher-order effects are warranted.[30] For now, we will confine our demonstration to linear terms only and for the conventional form of the POD. Where data fusion is concerned, and for systems where the database to be constructed comprises the reduced sensor set, the only tenable formulations are the *XX-topos* and *YX* forms and are written as follows.

1. *XX-topos* form for data fusion. The framework for assimilating the *XX-topos* form of the gappy-POD to fuse data begins with the same linear algebraic system of equations that form the foundation for stochastic estimation,[31]

$$\langle\phi(\tilde{\mathbf{x}},n)\phi(\tilde{\mathbf{x}},q)\rangle_{nq}\ b(y,n) = \langle\underline{P}(\tilde{\mathbf{x}},y)\phi(\tilde{\mathbf{x}},n)\rangle_{yn} \tag{5}$$

where the correlating terms on both sides of Eq. (5) are the spatial (*topos*) eigenvectors $\phi(x,n)$ from the solution to Eq. (2) using the numerical database. A process diagram for data fusion using the *XX-topos* form of the gappy-POD is shown in Fig. 1 (adapted from Zhao and Tinney[21]) and assumes $|x| = |y|$ for simplicity. The caveat to fusing databases is in generating the term on the right hand side of Eq. (5) where we have inserted $\underline{P}(\tilde{\mathbf{x}},y)$ from the experiments in place of $p(\tilde{\mathbf{x}},y)$ from the numerical database (as is traditionally done). In matrix form, Eq. (5) can be written as $[\mathbf{\Phi}(n,q)]\ [\mathbf{b}(y,n)] = [\underline{\mathbf{P}}(y,n)]$ and is rearranged and solved for $[\mathbf{b}] = [\mathbf{\Phi}]^{-1}\left[\underline{\mathbf{P}}\right]$ where $[\mathbf{b}]$ is the transfer function that fuses numerical and experimental databases. For ill-conditioned matrices or badly scaled data, a suitable replacement is offered,[32]

$$[\Phi'(n,q)] = (1+\xi)I(n,q) - \langle\phi(\tilde{\mathbf{z}},n)\phi(\tilde{\mathbf{z}},q)\rangle_{nq} \tag{6}$$

where the identity matrix $I(n,q) = \langle \phi(x,n)\phi(x,q) \rangle_{nq}$ is a natural artifact of the orthonormal basis while an arbitrary constant ξ is inserted to regularize the matrix. Values for ξ are unique to the problem, though a good starting point may be something in the range $10^{-10} \leq \xi \leq 10^{-12}$. The newly formed coefficients are used to approximate the data,

$$\hat{\underline{P}}_1^{(m)}(x,y) = \sum_n^m b(y,n)\phi(x,n) \tag{7}$$

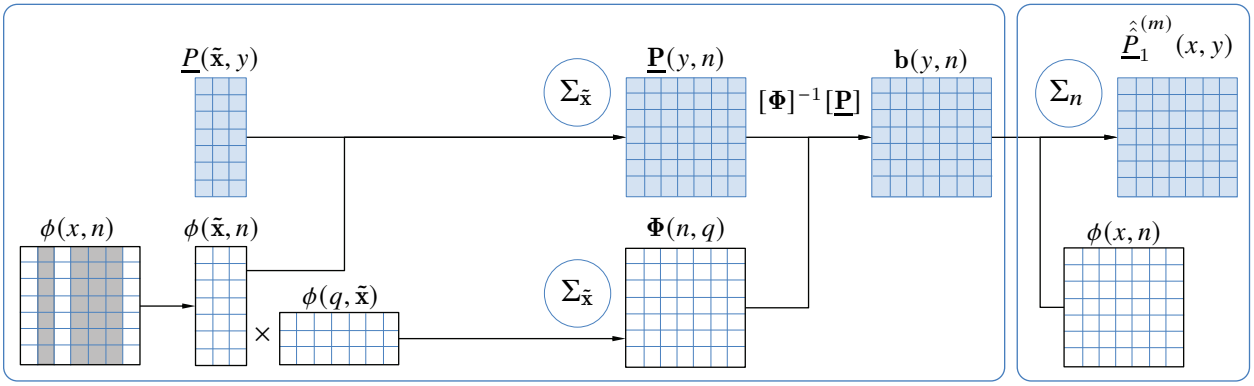


Figure 1: **Process diagram for the heterogeneous *XX-topos* form of the gappy-POD for data fusion.**

2. *YX* form for data fusion. For this alternative form, we first decompose the y-direction using a kernel constructed from the opposing coordinate $R(y,y') = \langle p(x,y), p(x,y') \rangle_{yy'}$ where y' denotes a separation along y; the properties of $R(y,y')$ are the same as $R(x,x')$ and are used to generate a new set of eigenvectors $\psi(y,n)$ and eigenvalues $\omega(n)$ from the following eigenvalue problem solution,

$$R(y,y')\psi(y',n) = \omega(p)\psi(y,n) \tag{8}$$

Two new sets of coefficients are needed from the training database

$$c(x,n) = \sum_y p(x,y)\psi(y,n) \tag{9}$$

$$d(\tilde{\mathbf{x}},n) = \sum_y p(\tilde{\mathbf{x}},y)\psi(y,n) \tag{10}$$

where Eq. (9) employs the full sensor set while Eq. (10) is formed from a subset of sensors that matches the gappy sensor set of the reconstructing database. The linear algebraic system of equations now becomes

$$\langle d(\tilde{\mathbf{x}},n)d(\tilde{\mathbf{x}},q) \rangle_{nq}\ b(y,n) = \langle \underline{P}(\tilde{\mathbf{x}},y)d(\tilde{\mathbf{x}},n) \rangle_{yn} \tag{11}$$

where the gappy sensor set from the reconstructing database is inserted to form the right hand side term. This is written in matrix form as $[\mathbf{D}(n,q)]\ [\mathbf{b}(y,n)] = [\underline{\mathbf{P}}(y,n)]$. An alternative expression for the left hand side term is provided as $[\mathbf{D}'(n,q)] = (1+\xi)\omega(n,q) - \langle e(\tilde{\mathbf{z}},n)e(\tilde{\mathbf{z}},q) \rangle_{nq}$ and is substituted for occasions when the matrix is ill-conditioned. In this case, the eigenvalue matrix is replaced by $\omega(n,q) = \langle c(x,n)c(x,q) \rangle_{nq}$ while a new set of POD expansion coefficients are constructed from the opposing gappy sensor set $e(\tilde{\mathbf{z}},n) = \langle p(\tilde{\mathbf{z}},y)\psi(y,n) \rangle_{\tilde{z}n}$. At last, a low-order reconstruction of the data is accomplished using

$$\hat{\underline{P}}_2^{(m)}(x,y) = \sum_n^m b(y,n)c(x,n) \tag{12}$$

A process diagram for the *YX* form is provided in Fig. 2. If $|x| = |y|$ and $R(x,x')$ is constructed from the same data as $R(y,y')$, then it can be shown how $c(x,n) = \phi(x,n)$ and $a(y,n) = \Psi(y,n)$ for all n.

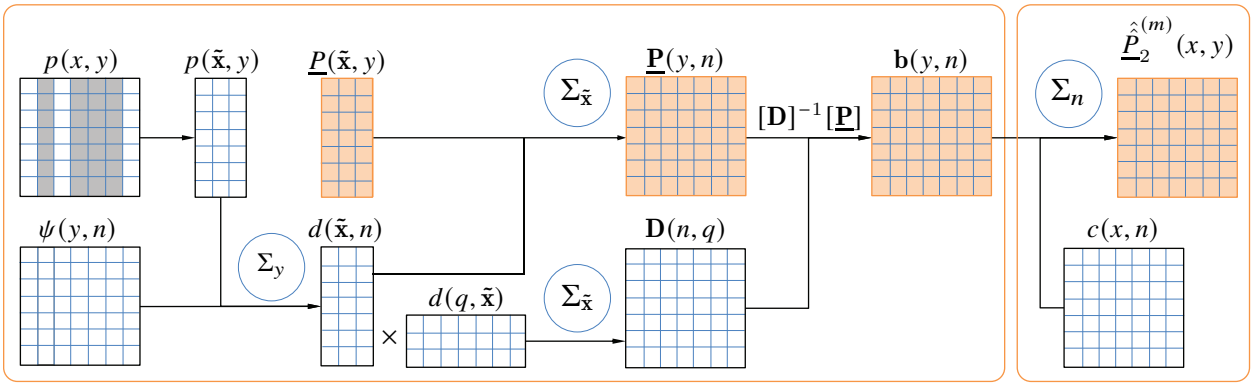


Figure 2: **Process diagram for the heterogeneous *YX* form of the gappy-POD for data fusion.**

IV. Numerical database: large eddy simulation

To demonstrate the utility of fusing data using reduced sensors sets, two separate databases are required and of the same physical system. The first of these is a numerically generated large-eddy simulation (LES) of a Mach 3 jet flow which will be used for training purposes.

This jet LES was described in Pineau and Bogey.[33] For the most part, the nozzle-exit parameters of the jet were chosen to match the physical jet used to generate the experimental database (to be described after). That is, the Mach 3 jet is perfectly expanded and exits from a straight-pipe nozzle with a diameter of $D_j = 2r_0 = 0.7$ mm at a centerline velocity of $U_j = 615$ m/s. The LES was carried out by solving the filtered compressible Navier–Stokes equations in cylindrical coordinates using low-dispersion and low-dissipation finite-difference schemes.[34] The computational domain extends to $x_1 = 35D_j$ in the axial direction and $x_2 = 7.5D_j$ in the normal direction, for a total of 325 million points.[33] More details on the LES methods and the grid design can be found in previous papers.[33,35,36] To match the experimental database (schlieren images of wavefronts in the hydrodynamic periphery of the jet flow), the derivatives of the numerically-generated density field $\partial\rho/\partial x_1$ and $\partial\rho/\partial x_2$ are computed and then integrated along lines in x_3 and perpendicular to the flow. The 3-D density field, not stored during the LES, is obtained by summing the contributions from the first five azimuthal modes (0 to 4) obtained every $\delta t^* = \delta t \cdot U_j/D_j = 0.0974$, as was done in Pineau and Bogey[37] to capture the steepened acoustic waves radiating in the near fields of temporally-developing round jets at Mach numbers of 2 and 3. Lines of sight integrations were performed between $x_3 = -L_{x3}$ and $x_3 = L_{x3}$ with $L_{x3} = 5.5D_j$ in increments of $\delta x_3 = 0.025D_j$. The integrated result was checked so as not to depend significantly on the two integration parameters L_{x3} and δx_3.

A sample snapshot of a numerical schlieren of both the flow and immediate sound field is shown in Fig. 3a. The illustration reveals density gradients corresponding to pressure waves propagating from the jet shear layer and at inclined angles relative to the jet axis. These wavefronts are the result of large turbulent structures convecting along the jet axis at supersonic speeds that become acoustically matched with the ambient gas to generate high intensity pressure waves. A line array of virtual sensors are placed parallel to the jet axis at $x_2/D_j = 4.0$ from which the numerical schlieren field is captured and stored for subsequent training. The space time evolution of density waves passing through this sensor array is shown in Fig. 3b and reveals wavefronts that remain intact over distances much greater than the potential core length. Like the overall sound pressure level, an equivalent overall sound density level (OASDL, [dB], re: 20μPa) is computed by assuming ideal gas effects $\rho_\dagger = p_\dagger/R_g/T$ and a commonly adopted reference pressure $p_\dagger = 20\mu$Pa. The findings are illustrated in Fig. 3c and demonstrate how density gradients in the computational domain for the sideline observer at $x_2/D_j = 4.0$ are most intense around $x_1/D_j = 10$. A window identifying the location in the flow where experimental schlieren measurements are performed is also provided.

V. Experimental database: digital schlieren

The experimental database used for reconstruction is the one described by Valdez and Tinney[38] and comprises spatially and temporally resolved schlieren images of Mach waves captured in the immediate vicinity of a laboratory scale Mach 3 jet flow. The nozzle contour is axisymmetric and produces a shock

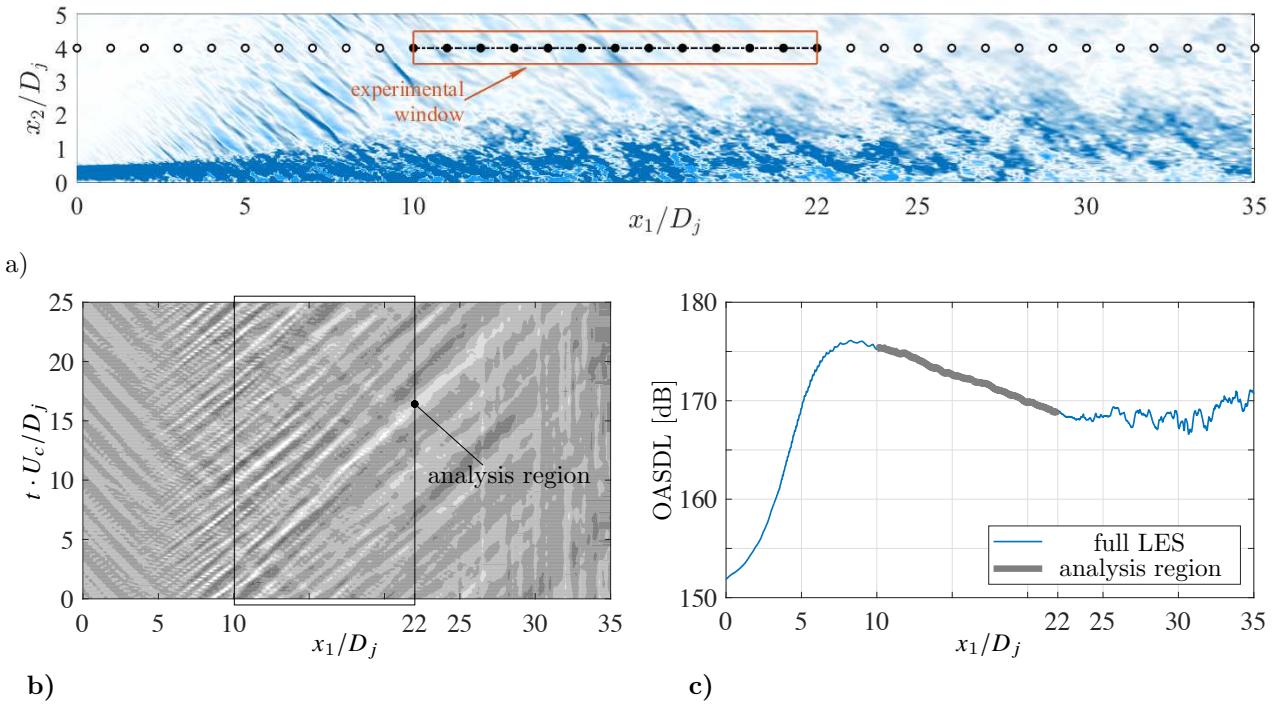
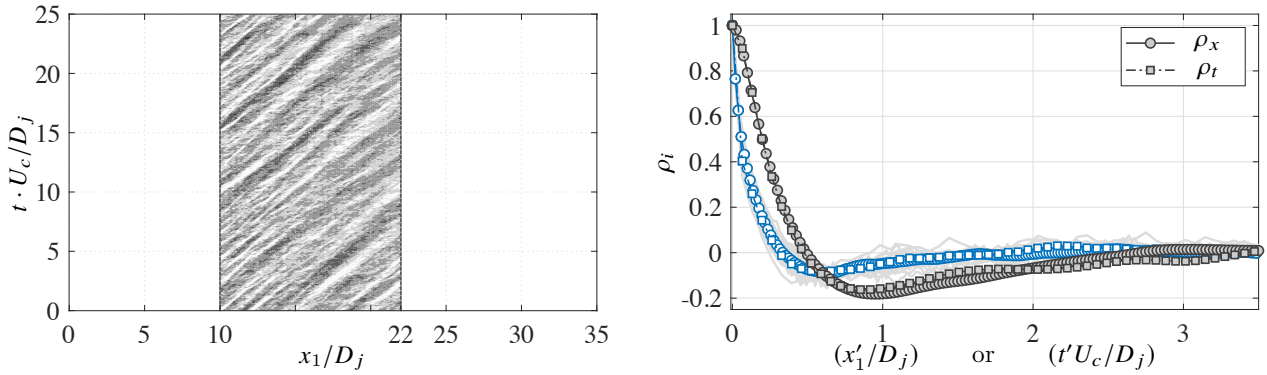


Figure 3: **a) Numerical schlieren of the large eddy simulation of a Mach 3 jet flow with its immediate sound field and using the first five azimuthal modes of the density field ($\partial\rho/\partial x_2$). Points in space identify where experimental schlieren is captured. b) Numerical schlieren of Mach waves convecting through space and time at $x_2/D_j = 4.0$. c) Corresponding overall sound density level [dB, re: 20μPa].**

free jet of air at a nozzle pressure ratio (NPR; ratio of plenum pressure P_o to ambient pressure P_∞) of 36.7 corresponding to a gas dynamic Mach number of 3.0 ($M_j = U_j/a_\infty$). The jet temperature is that of the surrounding air (total temperature, T_o = 302 K) while the convective acoustic Mach number is estimated to be around M_c = 1.43 ($M_a = U_j/a_\infty$, $M_c = \beta M_a$ and β = 0.8) using a sound speed of a_j = 208 m/s. The analysis is confined to a total of 8,192 digital images acquired uninterrupted at a rate of $f_s = 300,000$ frames per second (fps) using a i-SPEED 726 monochromatic camera and a field-of-view of 66×620 pixels; this sample rate equates to a non-dimensional time increment of $\delta t^* = 0.0819$. The image magnification is $\Delta_p = 5{\times}10^{-4}$ meters per pixel so that the resulting vertical and horizontal boundaries of the schlieren window are $[x_{1,start}, x_{1,end}]/D_j = [10.0, 22.0]$ and $[x_{2,start}, x_{2,end}]/D_j = [3.5, 4.5]$, respectively. The location of this window, relative to the jet flow, is shown in Fig. 3a. In this region of the sound field, imaged pressure waves are intense and are dominated by Mach waves. Given the sampling speed of the camera and its magnification, the dominant pressure waves traverse $U_c/(f_s\Delta_p) = 3.33$ pixels per frame based on a convective velocity of 500 m/s. The final grid for both the numerical and experimental databases are matched to have equal increments of $\delta x_1^* = \delta x_1/D = 0.025$ and $\delta t^* = 0.0819$. This equates to a total of 1400 sensors between $x_1/D_j = 0$ and 35 (480 sensors between $10 \le x_1/D_j \le 22$) and was checked to ensure that spurious noise was not introduced in the interpolation. At last, the gray scale contrast from the experimental schlieren (ξ_e) is scaled to match the numerical schlieren ($\nabla\rho_n$) so that $\nabla\rho_e(x) = \xi_e(x)\sigma_n(x)/\sigma_e(x)$ where $\sigma_e(x)$ and $\sigma_n(x)$ are the standard deviations of the ξ_e and $\nabla\rho_n$, respectively.

A sample of the experimental schlieren data is shown in Fig. 4a to provide a visual of the spatial and temporal evolution of young Mach waves generated by this Mach 3 jet and are complementary to the microphone measurements of Tinney and Jordan.[28] The features are also strikingly similar to the numerical schlieren snapshots in Fig. 3. Auto- and cross-correlations of these density waves are shown in Fig. 4b and are calculated using separations in space ρ_x and time ρ_t, and for both numerical (black symbols) and experimental (blue) databases. Spatial separations are made non-dimensional using jet diameter while time separations are made dimensionless by tU_c/D. Additional lines represent a multitude of different starting points in space and time, while symbols reflect their respective averages. For a given dataset, both correlation

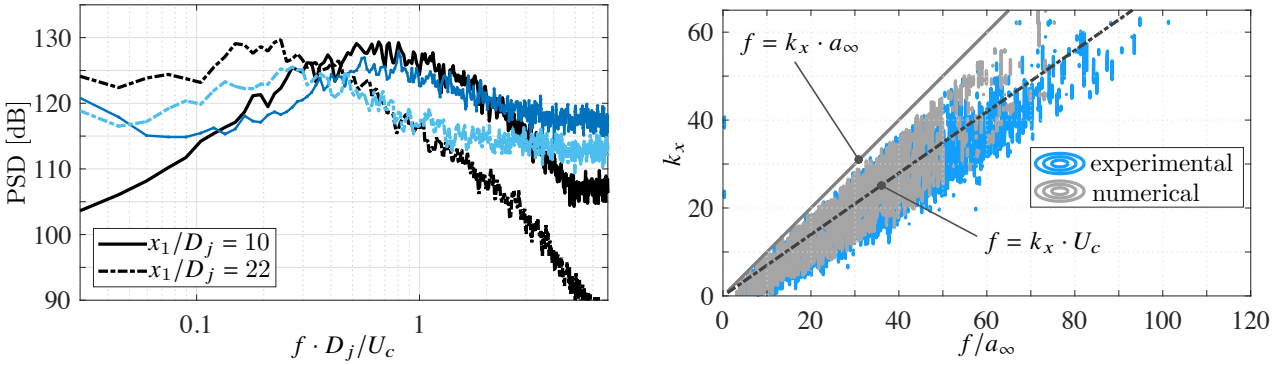
functions follow nearly identical shapes thereby demonstrating the invariance of these wavefronts to both space and time.



a) **b)**

Figure 4: **a) Experimental schlieren of Mach waves convecting through space and time at $x_2/D_j = 4.0$. b) Space-delay (ρ_x) and time-delay (ρ_t) correlation coefficients at various points between $10 \leq x_1/D_j \leq 22$ computed using experimental (closed symbols) and numerical (open symbols) databases.**

Ensemble averaged power spectral densities (PSD) of the numerical and experimental schlieren data are computed for observers located at $x_1/D_j = 10$ and 22 and shown in Fig. 5a. Dark and light color contrasts correspond to experimental and numerical schlieren databases, respectively. These spectra are computed using 29 overlapping bins with a spectral resolution of $\delta f = 293$ Hz which are then converted to dB using the same reference density as before. Peak frequencies are nearly identical for the numerical and experimental databases and are seen to shift to lower values at the lower downstream location. This reinforces the notion that the prominent sources of noise are from large scale turbulence that loses speed by dispersing its momentum to smaller scales after the collapse of the potential core, and that both databases are preserving these kinds of features. The wavenumber frequency spectra ($k_x - f$) is then computed and shown in Fig. 5b. Spatial transforms are confined to $10 \leq x_1/D_j \leq 22$ where the databases overlap. Two lines are drawn and correspond to wavenumbers that match with the sound speed of the surrounding gas ($f = k_x \cdot a_\infty$), as well as wavenumbers that match with the jet convection velocity ($f = k_x \cdot U_c$). Both the numerical and experimental databases demonstrate stark agreement to the latter of the two thereby reinforcing the location and speed of the acoustic source mechanisms.



a) **b)**

Figure 5: **a) Power spectral densities [dB] of the numerical and experimental schlieren at $x_2/D_j = 4.0$ and two axial stations in the sound field. b) Wavenumber frequency spectra computed for $10 \leq x_1/D_j \leq 22$.**

VI. Results

The results of the data fusion process are now discussed and begin with solutions from the *XX-topos* form followed by the *YX* form. We will adopt p in place of ρ to described the data fields being fused. The analysis focuses on circumstances where both numerical and experimental databases are confined to the same region in space ($10 \leq x_1/D_j \leq 22$). This simplifies programming although it is not a requirement of the technique. Both databases are identical in size and comprise $|x| = 481$ sensors with $|y| = 8192$ continuous samples in time. Likewise, the notation used in Section III will be adopted here where the numerical schlieren database is identified by $p(x, y)$ and the experimental database by $\underline{P}(x, y)$ while y is synonymous with time t. In order to gauge their compatibility, the kernel is constructed using data splits ranging from one that is entirely numerical ($\chi = 1.00$), to one that is entirely experimental ($\chi = 0.00$) where $\text{rank}(R_{xx'}) = |x|$. Blending data like this is not new and has been used to generate global POD modes for flow control purposes.[39]

Starting with Fig. 6a, the convergence of eigenvalues are shown for different data splits and superposed one another. Eigenvalues in the outer illustration are normalized by their respective cumulative sum and expressed as a percentage, while the figure insert displays convergence relative to the first and most energetic eigenvalue (since $\lambda_1 \geq \lambda_2 \geq ...\lambda_n$). A smooth cascade of energy is manifest which is nearly indistinguishable for all values of χ and for the first 5% of available POD modes. Since the relative weight of the first few POD modes are nearly identical between databases, they are expected to contribute equally to the reconstructed signal and absent of the data split. The slower decay with the experimental database ($\chi = 0.00$) for higher POD mode numbers may be the consequence of a poor signal-to-noise ratio as was demonstrated by the slower spectral decay in the higher frequencies shown in Fig. 5a and the smaller integral scales in Fig. 4b. Alongside the eigenvalues, the first four spatial eigenvectors are compared in Fig. 6b for the same data splits used in Fig. 6a. It is known that jet flows possess a dominant axial instability wave that undergoes growth, saturation and decay. This envelop occurs over several nozzle diameters and extends past the collapse of the potential core. The first two mode shapes in Fig. 6b appear to have reached their growth limit early on and are assumed to be part of this axial instability wave. The next two POD mode shape pairs ($n = 3$ and $n = 4$) have similar wavelengths and shapes while extending further downstream. Phase portraits of the first four eigenvector pairs are shown in Fig. 7 using solutions for $\chi = 1.00$ (Fig. 7a,b) and $\chi = 0.00$ (Fig. 7c,d). The findings demonstrate that these eigenvector pairs are spatially compatible and that accurate reconstructions of real observations should be possible using only a small fraction of synthesized basis functions.

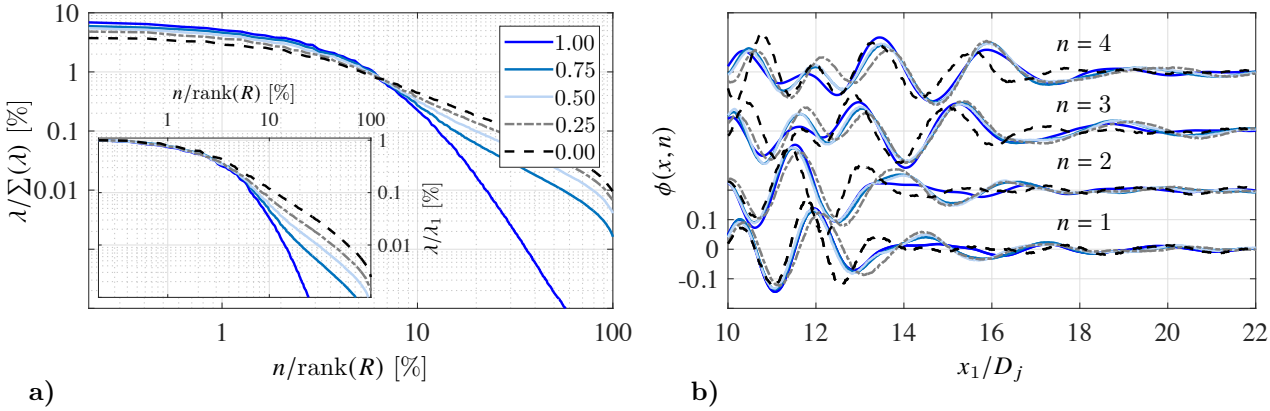


Figure 6: **a) Eigenvalue convergence and b) first four eigenvectors. Legend identifies the training split between one that is entirely numerical ($\chi = 1.00$) and one that is entirely experimental ($\chi = 0.00$).**

An evaluation of the dominant wavenumbers that make up the POD eigenvectors provides insight concerning the scales of the source field responsible for generating these wavefronts. This is accomplished using an equivalent wavenumber and is defined in one of two ways. The first is based on a definition proposed by Moser[40] and is denoted by k_{eq}. Alternatively, one can count the number of zero crossings (N_z: the number of instances that the waveform flips sign and intercepts the x axis at $y = 0$) and is a primitive form of pitch detection. In this case, $k_{eq} = N_z/2$. The findings are shown in Fig. 7a using both definitions (solid lines for N_z and dashed lines for k_{eq}) applied to both databases. Within the first 5% of the available

POD modes, equivalent wavelengths are on the order of several nozzle diameters. Above this, wavenumbers increase monotonically to values that are fractions of the jet diameter. The significance of this is that if one is interested in distilling frequencies of relevance to the sound field, a good starting point would be ones that match with the convective speed of the jet, i.e. $f_{eq} = k_{eq} \cdot U_c$ following the wavenumber frequency spectra in Fig. 5b. Overall, both databases yield complementary results using both definitions.

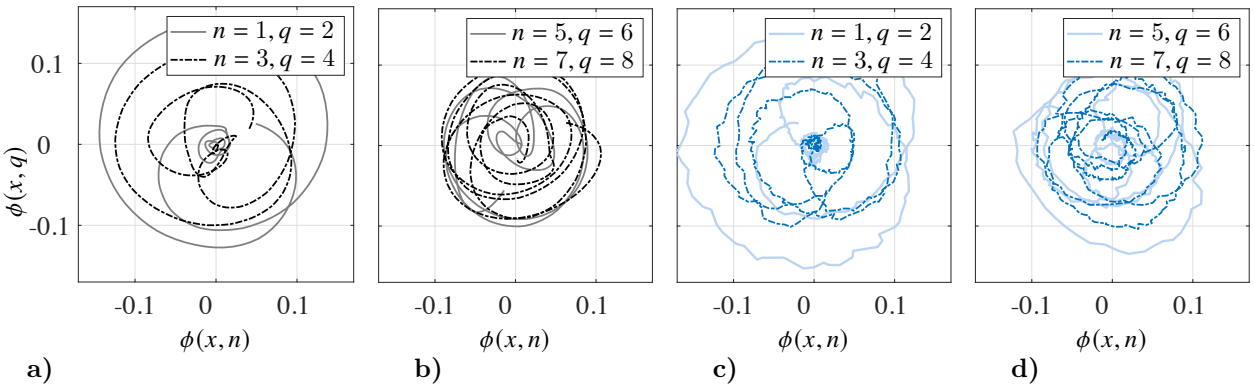


Figure 7: **a-b)Phase portraits of eigenvector pairs generated from entirely numerical ($\chi = 1.00$) or c-d) experimental ($\chi = 0.00$) databases.**

A. Data fusion using the *XX-topos* form

Findings from the data fusion process are now discussed and begin with results from the *XX-topos* form of the technique. While there are numerous approaches for optimizing the location of the gappy sensor set, they often encompass a trade-off between speed, accuracy and sensor characteristics. In previous work of Zhao and Tinney[21] it was demonstrated how the optimal sensor set is one that provides broad coverage of the data field when the data is stochastic. As such, the gappy sensor set is determined here using a random sequence of integers and is shown in Fig. 8b. Relative to standard greedy methods, the use of a random number generator is much less costly, and for nearly the same accuracy. Thus, there is a diminishing return in accuracy when running a traditional greedy algorithm.

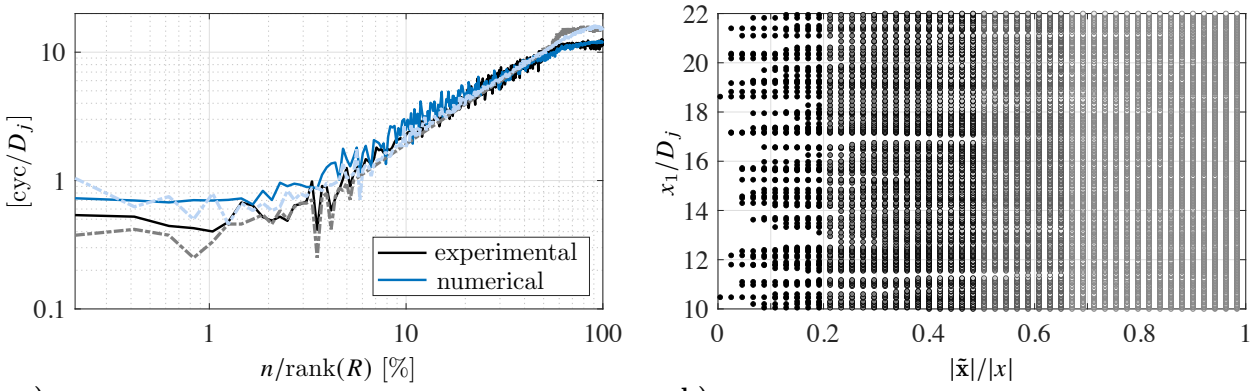


Figure 8: **a) Equivalent wavenumbers of the POD modes and number of zero crossings per jet diameter. b) Location of the gappy sensor set using a random number generator.**

Mean square errors between the unfiltered database and its low-order counterpart are plotted in Fig. 9 for a range of sensor sets and POD mode numbers. Since the total resolved signal is computed as

$$E = \sum_x \sum_y p(x,y)^2 = \sum_n \lambda(n) \tag{13}$$

and its cumulative sum by

$$\Pi^{(m)} = \sum_{n}^{m} \lambda(n)/E \tag{14}$$

then the mean square error in the reconstructed signal, for a given combination of POD modes, is simply $\epsilon^{(m)} = 1 - \Pi^{(m)}$. As such, errors are computed using,

$$\hat{\hat{\epsilon}}(m) = \frac{\sum_x \sum_y \hat{\underline{P}}^{(m)}(x,y)^2}{\sum_n^m \lambda(n)} \tag{15}$$

and then presented as $1/\hat{\hat{\epsilon}}(m)$ so that values range between 0 (for 100% error) and 1 (for 0% reconstructed error). Fig. 9a shows the mean square errors using the numerical database as the reconstructed signal. The findings are plotted as a ratio of the cardinality of the gappy sensor set relative to the number of POD modes used in the reconstruction. This is the benchmark to compare to since the same databases are used for training and reconstruction. Overall, for a given POD mode number, errors are shown to vanish quite rapidly when $|\tilde{\mathbf{x}}| \geq 2n$ and with increasing mode number. Replacing the reconstructed database with the experimental schlieren measurements ($p(x,t) \rightarrow \underline{P}(x,t)$) generates nearly identical mean square errors, as shown in Fig. 9b. Once again, the findings reinforce the compatibility of these two databases and the robustness of the gappy-POD.

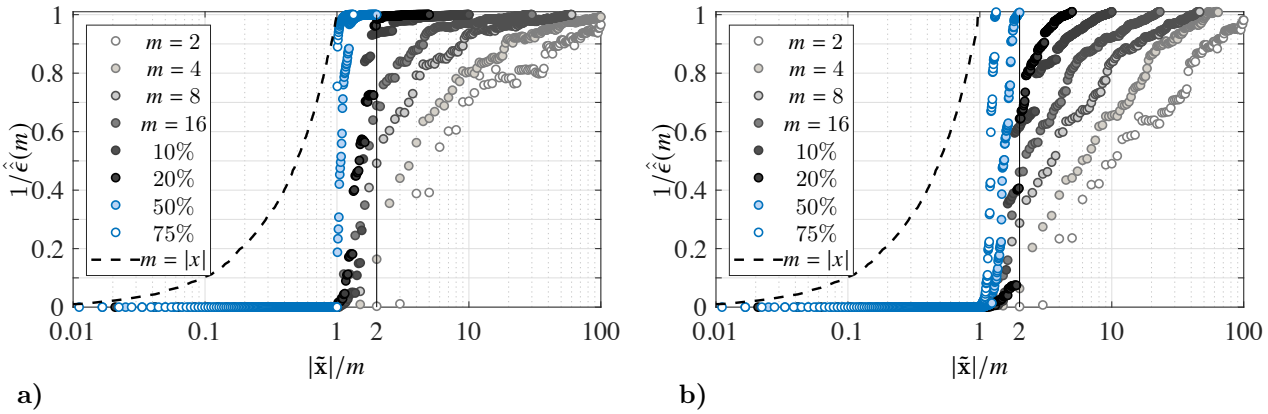


Figure 9: **a) Mean square errors in the *XX-topos* form of the gappy POD using the a) numerical $p(x,t)$ and b) experimental $\underline{P}(x,t)$ databases for reconstruction.**

Sample reconstructions of the experimental waveforms are shown in Fig. 10 to demonstrate qualitatively how compatible the numerical and experimental databases are. The training process is confined to the numerical database only ($\chi = 1.00$). A sample signal from the raw experimental database is shown in Fig. 10a alongside a gappy-POD reconstruction using the first $m = 24$ POD modes in Fig. 10b. This equates to 5% of the available spatial modes and accounts for approximately 30% of the resolved signal. These snapshots correspond to the same points in time. A third low-order reconstruction is shown in Fig. 10c using the *XX-topos* form of the gappy-POD with the first $m = 24$ numerical POD modes and a gappy sensor set comprised of $\tilde{\mathbf{x}} = 48$ sensors. This equates to a 90% reduction in the number of available sensors. Closer inspection of the axial waveforms at arbitrary instants in space $p(t)$ and time $p(x)$ is provided in Fig. 10d and reveals the differences between the original input waveforms and the gappy-POD reconstructions. Both low-order reconstructions, one using experimental POD modes and the other using numerically generated POD modes with a gappy sensor set, are shown to capture the large-scale features of the raw signal. The resemblance between reconstructions that employ experimental eigenvectors versus ones that are generated using the numerically generated observations and the gappy sensor set is remarkable.

B. Data fusion using the *XY* form

The last of these results focuses on data fusion using the *YX* form of the gappy POD. Similar to before, solutions to Eq. (8) are obtained for data splits over the range $0.00 \leq \chi \leq 1.00$ and for a kernel where

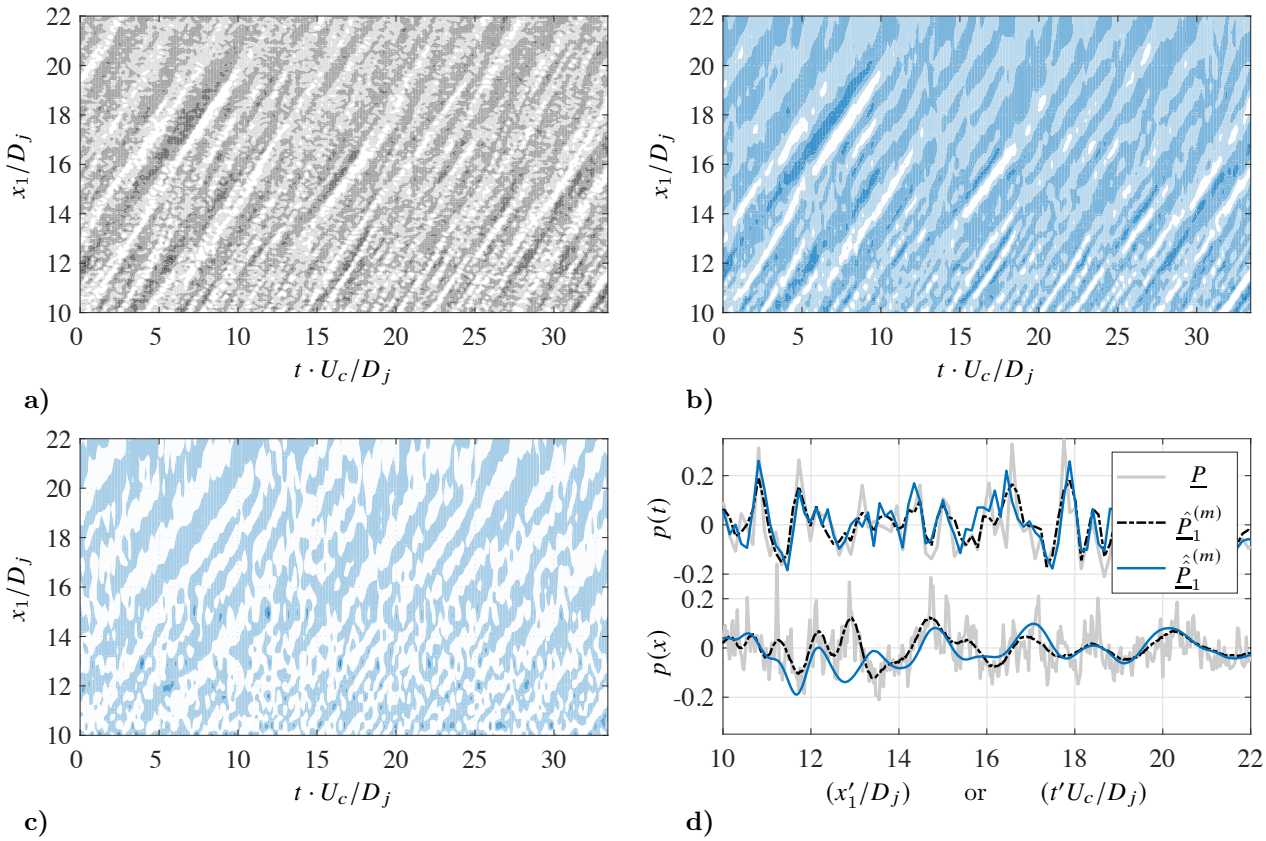


Figure 10: **Comparison of experimental database at $x_2/D_j = 4$. a) Unfiltered waveforms $\underline{P}(x,t)$, b) low-order reconstruction with $m = 24$ POD modes, c) gappy-POD reconstruction using the *XX-topos* form with $m = 24$ POD modes and $\tilde{\mathbf{x}} = 48$ sensors. d) Same reconstructed waveforms at arbitrary slices in space and time.**

$\text{rank}(R_{yy'}) = |x|$. To eliminate the formation of neglected modes, the training database is truncated so that $|y| = |x|$. As such, fewer samples are included in the inner product operation when forming $R_{yy'}$, relative to $R_{xx'}$. Therefore, solutions to Eq. (8) are not expected to match those from Eq. (1) (i.e. $\lambda(n) \neq \omega(n)$), even though $\text{rank}(R_{yy'}) = \text{rank}(R_{xx'})$. Eigenvalue convergences are shown in Fig. 11a and are complementary to the findings in Fig. 6 where data splits are concerned. In this case, the percent contribution of subsequent POD modes relative to the first is unaffected by the data split over the first 15% of available modes. The smooth decay is emblematic of homogeneous turbulence in the source field and is complementary to the eigenvalues in Fig. 6.

Mean square errors for the *YX* form of the gappy POD using different combinations of POD modes and gappy sensors sets are shown in Fig. 11b. Reconstructions are performed using the experimental database only and are shown to generate similar errors as the ones using the *XX-topos* form and illustrated in Fig. 9b. Once again, these errors reduce when $|\tilde{\mathbf{x}}| > 2n$. When $m/\text{rank}(R) \gtrsim 0.5$, the number of sensors in the gappy sensor set required to generate the same relative error reduces to $|\tilde{\mathbf{x}}| \sim n$. It is postulated that this is related to either a nyquist effect with the POD modes, or the lack of physical relevance in the higher POD mode numbers.

Low-dimensional reconstructions of the experimental waveforms using the *YX* form of the gappy-POD are shown in Fig. 12 to complement the findings in Fig. 10c,d using the *XX-topos* form and the same POD mode and gappy sensor set combinations. The original input waveforms are shown in Fig. 10a as a comparison. Slices of the original and reconstructed waveforms across space and time are shown independently in Fig. 12b. Qualitatively, the spatial waveform reconstruction ($p(x)$ in Fig. 12b) is not as crisp with the *YX* form as it is when reconstructed using the *XX-topos* form in Fig. 10b. Albeit, differences between the *YX* and *XX-topos* forms are nearly indistinguishable for the reconstructions across time $p(t)$. The discrepancies between spatial

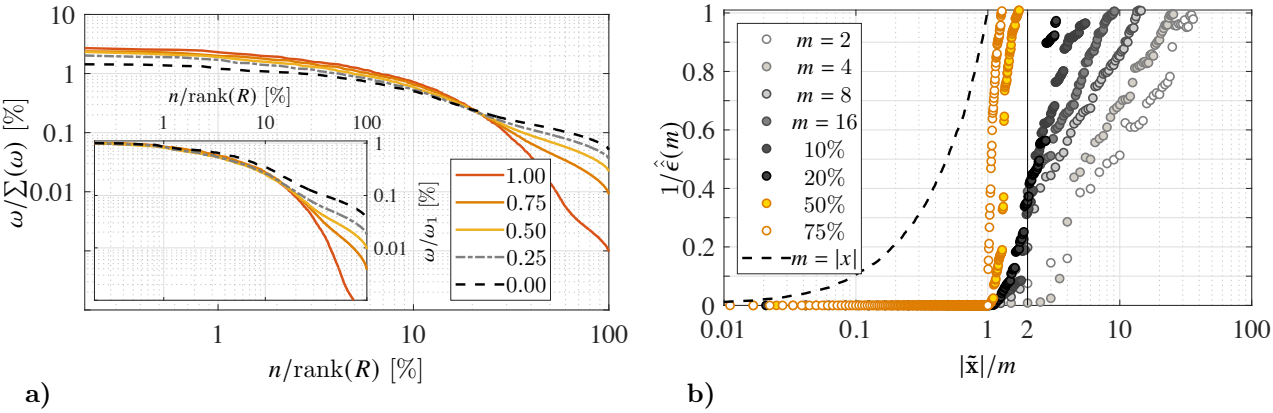


Figure 11: **a) Eigenvalue convergence for $R(y, y')$ for training splits between $\chi = 1.00$ and 0.00. b) Convergence error for YX form of the gappy POD using the experimental database for reconstruction.**

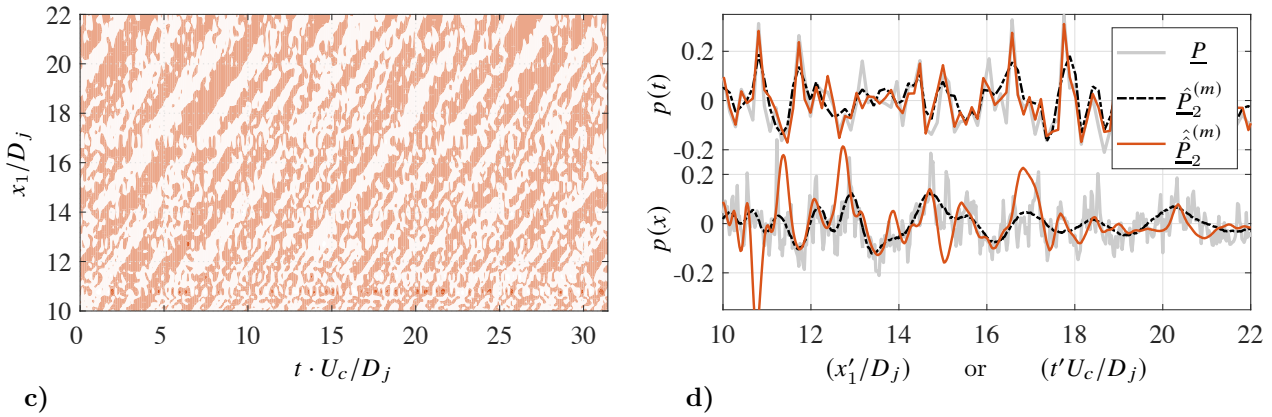


Figure 12: **a) Gappy-POD reconstruction using the YX form with $m = 24$ POD modes and $\tilde{\mathbf{x}} = 48$ sensors. b) Comparison of experimental database at $x_2/D_j = 4$ to the YX form of the same reconstructed waveforms at arbitrary slices in space and time.**

and temporal accuracies may have more to do with differences in the sample set sizes used to construct $R_{xx'}$ and $R_{yy'}$, and less to do with differences between the YX and XX-*topos* forms of the gappy-POD where data fusion is concerned.

There are different approaches to training which do not necessarily require compatible databases. Since the POD basis is orthonormal, ($\phi(x, n)\phi(x, q) = \delta_{nq}$ where δ_{nq} is the kronecker delta; $\delta_{nq} = 1$ for $n = q$ and 0 for $n \neq q$) then the reconstructed signal will always be fully recovered, absent of the training data used to generate the POD basis. Albeit, the more compatible the database, the fewer the number of modes are required for reconstruction. For example, Willis et al.[5] adopted a transfer learning approach built off of the pre-trained SqueezeNet convolution neural network and with 3.2 million images from the ImageNet database. While effective, the training process was slow, (taking several minutes, if not hours, to complete), unlike the fractions of seconds that it takes to compile the numerical database used here. Thus, the more compatible the database, the less costly the training, and the more accurate the reconstruction using a gappy sensor set.

Acknowledgments

Partial support for the first two authors was graciously provided through an internal research and development award provided by the Signal and Information Sciences Laboratory of the Applied Research Labo-

ratories, The University of Texas at Austin. Special thanks to John Valdez for generating the high-speed schlieren images used in this study.

References

[1]Tinney, C. E., Ukeiley, L. S., and Glauser, M. N., "Low-dimensional Characteristics of a Transonic Jet. Part 2. Estimate and far-field prediction," *Journal of Fluid Mechanics,* Vol. 615, 2008, pp. 53-92, https://doi.org/10.1017/S0022112008003601

[2]Yang, Z., Cattafesta, L. N., and Ukeiley, L., "Spectral analysis modal methods (SAMMs) using non-time-resolved PIV." *Experiments in Fluids* Vol. 61, No. 226, 2020, https://doi.org/10.1007/s00348-020-03057-8

[3]Berry, M. G., Stack, C. M., Magstadt, A. S., Ali, M. Y., Gaitonde, D. V., and Glauser, M. N., "Low-dimensional and data fusion techniques applied to a supersonic multistream single expansion ramp nozzle," *Physical Review Fluids 2,* Vol. 100504, 2017, pp. 1-26, https://doi.org/10.1103/PhysRevFluids.2.100504

[4]Tenney, A. S., Glauser, M. N., Ruscher, C., and Berger, Z., "Applications of Artificial Neural Networks to Stochastic Estimation and Jet Noise Modeling," *AIAA Journal,* Vol. 58, No. 2, 2020, pp. 647-658, https://doi.org/10.2514/1.J058638

[5]William, W. A., Valdez, J. A., Tinney, C. E., and Hamilton, M. F., "A machine learning approach to identifying waveform coalescence in supersonic jet noise," *AIAA Aviation 2023 Forum*, AIAA Paper 2023-2514, 2023, pp. 1-18, https://doi.org/10.2514/6.2023-3353

[6]Renganathan, S. A., Harada, K., and Mavris, D. N., "Aerodynamic data fusion toward the digital twin paradigm," *AIAA Journal,* Vol. 58, No. 9, 2020, pp. 3902-3918, https://doi.org/10.2514/1.J059203

[7]Jiang, C., Soh, Y. C., and Li, H., "Sensor and CFD data fusion for airflow field estimation," *Applied Thermal Engineering,* Vol. 92, 2016, pp. 149-161, https://doi.org/10.1016/j.applthermaleng.2015.09.078

[8]Ruscher, C., Dannenhoffer III, J., and Glauser, M., "Repairing occluded data for a Mach 0.6 jet via data fusion," *AIAA Journal,* Vol. 55, No. 1, 2017, pp. 255-264, https://doi.org/10.2514/1.J054785

[9]Zimmermann, R., Vendl, A., and Görtz, S., "Reduced-order modeling of steady flows subject to aerodynamic constraints," *AIAA Journal,* Vol/ 52, No. 2, 2014, pp. 255-266, https://doi.org/10.2514/1.J052208

[10]Bertram, A., Bekemeyer, P., and Held, M., "Bayesian gappy proper orthogonal decomposition for aerodynamic data fusion," *AIAA Journal,* Vol. 61, No. 9, 2023, pp. 4032-4044, https://doi.org/10.2514/1.J062356

[11]Anhichem, M., Timme, S., Castagna, J., Peace, A. J., and Maina, M., "Data Fusion of Wing Pressure Distributions Using Scalable Gaussian Processes," *AIAA Journal,* Vol. 62, No. 5, 2024, pp. 1946-1961, https://doi.org/10.2514/1.J063317

[12]Everson, R. and Sirovich, L., "The Karhunen-Loeve procedure for gappy data," *Journal of the Optical Society of America,* Vol. 12, 1995, pp. 1657-1664, https://doi.org/10.1364/JOSAA.12.001657

[13]Venturi, D., and Karniadakis, G., "Gappy data and reconstruction procedures for flow past a cylinder," *Journal of Fluid Mechanics*, Vol. 519, pp. 315-336, 2004

[14]Murray, N.E., and Ukeiley, L.S., "An application of Gappy POD," *Experiments in Fluids,* Vol. 42, No. 1, 2007, pp. 79-91.

[15]Raben, S. G., Charonko, J. ., and Vlachos, P. P., "Adaptive gappy proper orthogonal decomposition for particle image velocimetry data reconstruction," *Measurement Science and Technology*, Vol. 23, No. 025303, pp. 1-16, 2012

[16]Nekkanti, A. and Schmidt, O., "Gappy spectral proper orthogonal decomposition," *Journal Of Computational Physics,* Vol. 478, 2023, pp. 111950, https://doi.org/10.1016/j.jcp.2023.111950

[17]Bui-Thanh, T,. Damodaran, M., and Willcox, K., "Aerodynamic data reconstruction and inverse design using proper orthogonal decomposition," *AIAA Journal,* Vol. 42, No. 8, 2004, pp. 1505-1516.

[18]Willcox, K., "Unsteady flow sensing and estimation via the gappy proper orthogonal decomposition," *Computer and Fluids,* Vol. 35, 2006, pp. 208-226.

[19]Yildirim, B., Chryssostomidis, C. and Karniadakis, G., "Efficient sensor placement for ocean measurements using low-dimensional concepts," *Ocean Modelling,* Vol. 27, 2009, pp. 160-173.

[20]Tinney, C. E., Zhao-Dubuc, Y., and Valdez, J., "Frequency Domain gappy-POD for Rotor Acoustic Measurement Optimization," *Vertical Flight Society Forum 80,* May, 2024, pp. 1-12, https://doi.org/10.4050/F-0080-2024-1121

[21]Zhao,-Dubuc, Y., and Tinney, C. E., "Homogeneous and heterogeneous forms of the gappy-POD applied to the ship airwake," *AIAA Aviation 2024 Forum*, AIAA Paper 2024-4189, 2024, pp. 1-21, https://doi.org/10.2514/6.2024-4189

[22]Lumley, J. L., Stochastic Tools in Turbulence, *Academic Press, Inc.* New York, 1970.

[23]Lumley, J. L., "The Structure of Inhomogenous Turbulent Flows," In *Proceedings of the International Colloquium on the Fine Scale Structure of the Atmosphere and its Influence on Radio Wave Propagation*, (ed. Yaglom A. M., and Tatarski V. I.) Doklady Akademii Nauk SSSR, Moscow, 1967, pp. 166-178.

[24]Sirovich, L., "Turbulence and the Dynamics of Coherent Structures," *Journal of Applied Mathematics,* Vol. 45, 1987, pp. 561-590.

[25]Aubry, N., "On the hidden beauty of the proper orthogonal decomposition," *Theoretical And Computational Fluid Dynamics,* Vol. 2, 1991, pp. 339-352, https://doi.org/10.1007/BF00271473

[26]Glauser, M.N., and George, W.K., "Orthogonal Decomposition of the Axisymmetric Jet Mixing Layer Including Azimuthal Dependence," *Advances in Turbulence,* (ed. Comte-Bellot, G., and Mathieu, J.) 1987, pp. 357-366, https://doi.org/10.1007/978-3-642-83045-7-40

[27]Bonnet, J.P., Delville, J., Glauser, M.N., Antonia, R.A., Bisset, D.K., Cole, D.R., Fiedler, J.H., Garem, J.H., Hilberg, D., Jeong, J., Kevlahan, N.K.R., Ukeiley, L., and Vincendeau, E., "Collaborative testing of eddy structure identification methods in free turbulent shear flows." *Experiments in Fluids* Vol. 25, 1998, pp. 197-225, https://doi.org/10.1007/s003480050224

[28]Tinney, C.E., and Jordan, P., "The near pressure field of co-axial subsonic jets," *Journal of Fluid Mechanics,* Vol. 611, 2008, pp. 175-204, https://doi.org/10.1017/S0022112008001833

[29]Towne, A., Schmidt, O. T., and Colonius, T., "Spectral proper orthogonal decomposition and its relationship to dynamic mode decomposition and resolvent analysis," *Journal of Fluid Mechanics*, Vol. 847, 2018, pp. 821-867, https://doi.org/10.1017/jfm.2018.283

[30]Tinney, C. E., Zhao-Dubuc, Y., and Valdez, J., "Quadratic gappy-POD applied to the acoustic near-field of high Mach number jets," *AIAA/CEAS Aeroacoustics conference,* 2024, pp. 1-16, https://doi.org/10.2514/6.2024-3254

[31]Adrian, R.J., "Conditional eddies in isotropic turbulence," *Physics of Fluids,* Vol. 22, No. 11, 1979, pp. 2065-2070.

[32]Mokhasi, P., "Strategies for the real-time prediction of velocity fields," *PhD Thesis*, Illinois Institute of Technology, Chicago, IL, 2005

[33]Pineau, P., and Bogey, C., "Numerical investigation of wave steepening and shock coalescence near a cold Mach 3 jet," *Journal of the Acoustical Society of America*, Vol. 149, No. 1, pp. 357-370, 2021, https://doi.org/10.1121/10.0003343

[34]Bogey, C., and Bailly, C., "A family of low dispersive and low dissipative explicit schemes for flow and noise computations," *Journal of Computational Physics*, Vol. 194, No. 1, pp. 194-214, 2004, https://doi.org/10.1016/j.jcp.2003.09.003

[35]Bogey, C., "Grid sensitivity of flow field and noise of high-Reynolds-number jets computed by large-eddy simulation," *International Journal of Aeroacoustics*, Vol. 17, No. 4-5, pp. 399-424, 2018, https://doi.org/10.1177/1475472X18778287

[36]Pineau, P., and Bogey, C., "Links between steepened Mach waves and coherent structures for a supersonic jet," *AIAA Journal*, Vol. 59, No. 5, pp. 1673-1681, 2021, https://doi.org/10.2514/1.J059859

[37]Pineau, P., and Bogey, C., "Steepened Mach waves near supersonic jets: study of azimuthal structure and generation process using conditional averages," *Journal of Fluid Mechanics*, Vol. 880, 2019, pp. 594-619, https://doi.org/10.1017/jfm.2019.729

[38]Valdez, J. A., and Tinney, C. E., "Measurements of a Mach 3 jet using high-speed optical flow techniques," *AIAA Aeroacoustics,* AIAA Paper 2018-3148, 2018, https://doi.org/10.2514/6.2018-3148

[39]Glauser, M. N., Young, M. J., Higuchi, H., Tinney, C. E., and Carlson, H., "POD Based Experimental Flow Control on a NACA-4412 Airfoil (Invited)," *AIAA Aerosciences Meeting*, AIAA Paper 2004-0575, 2004, pp. 1-9.

[40]Moser, R. D., "Kolmogorov inertial range spectra for inhomogeneous turbulence," *Phys. Fluids,* Vol. 6, No. 2, 1994, pp. 794-801, https://doi.org/10.1063/1.868317