

Robust Spatio-Temporal Feature Generation for Incompressible Flow Equation Discovery

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Data-driven discovery of governing equations for fluid systems is challenged by sparse, noisy measurements and unobserved variables like pressure. This study addresses these challenges in a decaying, nearly incompressible fluid system, characterized by localized sharp spatial gradients, using a dataset of three-dimensional velocity and density fields on a 128^3 periodic domain across only 10 time slices. To generate robust spatio-temporal features, temporal derivatives were estimated using second-order local polynomial regression for each spatial point, mitigating noise from sparse data. Spatial derivatives were computed by benchmarking spectral methods against a 5th-order Weighted Essentially Non-Oscillatory (WENO5) scheme, with the selection guided by adherence to the incompressibility constraint ($\nabla \cdot \mathbf{u} = 0$). A comprehensive feature library was then constructed, including advection, viscous, and various proxy terms for pressure and buoyancy, motivated by a preliminary momentum residual analysis. All primary fields and generated features were independently standardized. Temporal derivatives were accurately extracted, exhibiting low Root Mean Square Errors (e.g., 0.0031 for u in smooth regions and 0.0066 in sharp gradient regions). Benchmarking confirmed the spectral method’s marginal superiority in satisfying the incompressibility constraint (RMS error 0.5914 vs. WENO5’s 0.5934), leading to its selection for all spatial derivative computations. A preliminary kinematic viscosity estimate was very small ($\approx 1.972 \times 10^{-7}$). Crucially, momentum residual analysis revealed significant, spatially coherent residuals, unequivocally confirming the dominant role of an unmeasured pressure gradient and validating the necessity of its proxy terms in the feature library. The resulting 26-term, independently standardized feature matrix and target temporal derivative vectors are now optimally conditioned for subsequent sparse regression-based equation discovery.

1 Introduction

The discovery of governing equations from observational data represents a transformative paradigm in scientific research, offering the potential to uncover fundamental physical laws and build predictive models for complex systems. In

fluid dynamics, where phenomena are often described by intricate partial differential equations, data-driven approaches hold particular promise for systems where underlying physics are partially understood, or where traditional modeling proves computationally intractable.

Despite this transformative potential, applying data-driven equation discovery to complex fluid systems faces significant hurdles. These include the inherent sparsity and noise in experimental or simulation data, the presence of unobserved but crucial physical variables, and the need to accurately capture localized features such as sharp gradients. This study addresses these critical challenges within the context of a decaying, nearly incompressible fluid system. Our dataset, comprising three-dimensional velocity and density fields on a 128^3 periodic domain across only 10 time slices, exemplifies these difficulties. Specifically, the severe temporal sparsity necessitates robust methods for estimating temporal derivatives, while the presence of localized sharp spatial gradients demands derivative schemes that can maintain accuracy without introducing spurious oscillations. Furthermore, the absence of direct measurements for key variables, such as pressure, requires the careful construction of physically informed proxy terms to ensure a complete and consistent representation of the underlying dynamics.

To overcome these obstacles and generate a robust set of spatio-temporal features essential for reliable equation discovery, we employ a multi-faceted approach. For the estimation of temporal derivatives, which are particularly susceptible to noise given the sparse temporal sampling, we utilize second-order local polynomial regression. This method is applied independently at each spatial point across all available time slices, effectively smoothing out noise and providing stable derivative estimates. For spatial derivatives, which must accurately capture both smooth regions and localized sharp gradients, we rigorously benchmark two distinct methodologies: spectral methods, known for their high accuracy on periodic domains, and a 5th-order Weighted Essentially Non-Oscillatory (WENO5) scheme, specifically designed for robust handling of discontinuities. The selection between these schemes for first-order spatial derivatives is critically guided by their adherence to the incompressibility constraint, $\nabla \cdot \mathbf{u} = 0$, ensuring physical consistency of the velocity field.

Following the robust computation of derivatives, a comprehensive feature library is constructed. This library includes standard terms such as advection and viscous dissipation, alongside a diverse array of proxy terms for unmeasured pressure gradients and buoyancy forces. The necessity and composition of these proxy terms are informed by a preliminary momentum residual analysis, which quantifies the contribution of known physical terms and highlights the magnitude and spatial coherence of the unmodeled forces. To ensure numerical stability and prevent any single feature from dominating the subsequent sparse regression process, all primary fields and generated candidate features are independently standardized. The outcome of this work is a meticulously prepared and optimally conditioned feature matrix and target temporal derivative vectors, poised for the subsequent application of sparse regression techniques to uncover the governing equations of this complex fluid system.

2 Methods

2.1 Dataset description

The dataset utilized in this study comprises three-dimensional velocity components, denoted as $\mathbf{u} = (u, v, w)$, and the density field, ρ . These fields are defined on a 128^3 periodic spatial domain and are available across 10 discrete time slices. The system under investigation represents a decaying, nearly incompressible fluid, characterized by localized sharp spatial gradients.

2.2 Data preparation and normalization

To prepare the raw data for subsequent analysis, the density field ρ was first transformed into a normalized density perturbation field, ρ' , defined as $\rho' = \rho - 1.0$. This transformation ensures that small density fluctuations, indicative of a nearly incompressible regime, are appropriately represented. Subsequently, all primary fields (u, v, w, ρ') were independently standardized. This process involved subtracting the mean and dividing by the standard deviation for each field across all spatial points and time slices, resulting in fields with a mean of zero and a standard deviation of one. This initial standardization mitigates numerical conditioning issues for downstream analyses.

2.3 Temporal derivative estimation

Given the extreme sparsity of the temporal data (only 10 time slices), robust estimation of temporal derivatives $(\partial u/\partial t, \partial v/\partial t, \partial w/\partial t, \partial \rho'/\partial t)$ was critical. For each individual spatial grid point, a second-order (quadratic) polynomial was fitted to the time series of each variable across all 10 available time slices. This local polynomial regression approach effectively smoothed out noise inherent in sparse temporal data. The temporal derivatives were then extracted by evaluating the derivative of these fitted polynomials at the original time points. The accuracy of this method was quantitatively assessed by computing the Root Mean Square Error (RMSE) of the polynomial fits against the raw data at representative spatial locations, encompassing both smooth and sharp gradient regions of the flow.

2.4 Spatial derivative computation and benchmarking

First and second-order spatial derivatives for all primary fields were computed. For first-order spatial derivatives, two distinct methodologies were rigorously benchmarked:

- **Spectral methods:** Leveraging the periodic boundary conditions of the domain, spatial derivatives were computed using Fast Fourier Transforms (FFTs). First derivatives along a direction j were obtained by multiplying the Fourier transform of the field by ik_j (where k_j is the wavenumber), and

second derivatives by multiplying by $-k_j^2$. The result was then inverse-transformed back to real space.

- **5th-order Weighted Essentially Non-Oscillatory (WENO5) scheme:** This finite difference scheme was implemented to robustly handle localized sharp spatial gradients without introducing spurious oscillations.

The selection of the optimal scheme for first spatial derivatives was critically guided by its adherence to the incompressibility constraint, $\nabla \cdot \mathbf{u} = 0$. The divergence of the velocity field, $\nabla \cdot \mathbf{u} = \partial u / \partial x + \partial v / \partial y + \partial w / \partial z$, was computed using first derivatives from both the spectral and WENO5 schemes. The performance of each scheme was quantified by calculating the Root Mean Square (RMS) error and the maximum absolute error of $\nabla \cdot \mathbf{u}$ across the entire domain. Based on this benchmarking, the spectral method was selected for computing all spatial derivatives in the final feature library due to its marginal superiority in satisfying the incompressibility constraint.

Using the selected spectral method, the following key vector operators were constructed for the feature library:

- Advection terms: The components of $(\mathbf{u} \cdot \nabla)\mathbf{u}$, such as $u\partial u/\partial x + v\partial u/\partial y + w\partial u/\partial z$.
- Laplacian of velocity: $\nabla^2 \mathbf{u} = (\nabla^2 u, \nabla^2 v, \nabla^2 w)$.
- Gradients of scalar fields: $\nabla \rho'$, $\nabla(u^2 + v^2 + w^2)$, $\nabla(\rho'^2)$, and $\nabla(\nabla \cdot \mathbf{u})$.

2.5 Feature library construction

A comprehensive feature library, denoted as Θ , was constructed to serve as candidate terms for the right-hand side of the momentum equations. The target variables for subsequent equation discovery were the temporal derivatives $\partial u/\partial t$, $\partial v/\partial t$, and $\partial w/\partial t$. The final library comprised 26 distinct candidate terms, evaluated at all spatio-temporal points, including:

- A constant term.
- Non-linear advection terms: the components of $(\mathbf{u} \cdot \nabla)\mathbf{u}$, specifically $(\mathbf{u} \cdot \nabla)u$, $(\mathbf{u} \cdot \nabla)v$, and $(\mathbf{u} \cdot \nabla)w$.
- Linear viscous terms: $\nabla^2 u$, $\nabla^2 v$, and $\nabla^2 w$.
- Proxy terms for unmeasured pressure gradients and buoyancy forces, including:
 - Gradients of normalized density: $\partial \rho' / \partial x$, $\partial \rho' / \partial y$, $\partial \rho' / \partial z$.
 - Gradients of kinetic energy: $\partial(u^2 + v^2 + w^2) / \partial x$, $\partial(u^2 + v^2 + w^2) / \partial y$, $\partial(u^2 + v^2 + w^2) / \partial z$.
 - Gradients of density perturbation squared: $\partial(\rho'^2) / \partial x$, $\partial(\rho'^2) / \partial y$, $\partial(\rho'^2) / \partial z$.

- Gradients of divergence: $\partial(\nabla \cdot \mathbf{u})/\partial x$, $\partial(\nabla \cdot \mathbf{u})/\partial y$, $\partial(\nabla \cdot \mathbf{u})/\partial z$.
- Cross-terms coupling velocity and density: $u\rho'$, $v\rho'$, $w\rho'$.
- Additional density-weighted velocity gradient terms derived from $\rho'\nabla\mathbf{u}$.

As a final preparation step, all 26 candidate terms in the feature matrix Θ were independently standardized. Each column of Θ was transformed to have a mean of zero and a standard deviation of one across all spatio-temporal points, ensuring numerical stability and equitable contribution during subsequent sparse regression.

2.6 Preliminary kinematic viscosity estimation and momentum residual analysis

Prior to full equation discovery, a preliminary estimate of the kinematic viscosity, ν , was obtained. This was achieved by performing a simple linear regression of the temporal derivatives ($\partial\mathbf{u}/\partial t$) against the Laplacian of the velocity components ($\nabla^2\mathbf{u}$) across the entire domain and all time slices. The average slope obtained from these regressions served as an initial estimate for ν .

Using this estimated viscosity, momentum residuals (R_u, R_v, R_w) were computed for each velocity component. For the u -component, the residual was calculated as:

$$R_u = \frac{\partial u}{\partial t} + (\mathbf{u} \cdot \nabla)u - \nu\nabla^2 u$$

Similar equations were applied for R_v and R_w . The spatial distribution of these residuals was analyzed using heatmaps and histograms to quantify their magnitude and spatial coherence. This analysis served as a crucial validation of the robustness of the derivative computations and provided unequivocal evidence for the significant role of unmeasured terms, such as the pressure gradient, in the system’s dynamics, thereby reinforcing the necessity of including appropriate proxy terms in the feature library.

3 Results

3.1 Data normalization and regime identification

Initial statistical analysis of the raw dataset provided critical insights into the physical regime of the system. The raw density field, ρ , exhibited a global mean of 1.0000000000061107 with an exceptionally small standard deviation of approximately 0.00217. This near-unity mean and minimal variance strongly indicate that the fluid operates in a nearly incompressible, low-Mach number regime, consistent with the Boussinesq approximation where density fluctuations primarily contribute to buoyancy forces rather than inertial mass variations.

To ensure that these critical but numerically small density fluctuations were not overshadowed by the velocity components during the regression phase, we

defined a normalized density perturbation field: $\rho' = \rho - 1.0$. The raw velocity components exhibited standard deviations of 0.234, 0.249, and 0.243 for u, v , and w , respectively. To prevent numerical conditioning issues and to ensure that all primary fields contribute equitably to the subsequent feature generation, the raw fields (u, v, w, ρ') were globally standardized to possess a mean of zero and a standard deviation of one, as described in the Methods section.

3.2 Temporal derivative estimation

A formidable challenge presented by this dataset is the extreme sparsity in the temporal domain. With only 10 time slices available, traditional finite difference schemes for estimating temporal derivatives ($\partial \mathbf{u} / \partial t$) are highly susceptible to catastrophic noise amplification. To circumvent this limitation and extract robust temporal dynamics, we implemented a local polynomial regression approach. For every individual spatial grid point (128^3 points) and for each variable, a second-order (quadratic) polynomial was fitted across the 10 time indices.

To rigorously validate this approach, we analyzed the fit quality at two distinct spatial locations representing different flow regimes. These locations were identified by computing the local velocity gradient energy, $G^2 = \sum_{i,j} (\partial u_i / \partial x_j)^2$. We selected a "smooth" region characterized by low gradient energy ($G^2 \approx 7.75$) and a "sharp" region exhibiting extreme gradient energy ($G^2 \approx 19706.18$).

Figure 1 illustrates the quadratic temporal fits for the standardized velocity components and density perturbation at a representative smooth spatial point. The fitted polynomials (orange lines) closely follow the raw, standardized data (blue markers), demonstrating the effectiveness of the method in capturing the temporal evolution even with sparse data. Similarly, Figure 2 shows the temporal fits at a sharp-gradient spatial point, where the method maintains high accuracy despite the complex local dynamics.

The Root Mean Square Error (RMSE) metrics for the quadratic fits further demonstrated exceptional accuracy across both regimes. At the smooth point, the RMSE values were 0.0031, 0.0013, 0.0021, and 0.0250 for u, v, w , and ρ' , respectively. Remarkably, at the sharp gradient point, the RMSE values remained highly constrained at 0.0066, 0.0014, 0.0026, and 0.0242. These low RMSE values confirm the robustness of the local polynomial regression method for extracting accurate temporal derivatives from sparse data.

3.3 Spatial derivative benchmarking

The accuracy of spatial derivatives is paramount for constructing the feature library, particularly for evaluating non-linear advection terms ($(\mathbf{u} \cdot \nabla) \mathbf{u}$) and ensuring physical consistency. Given the periodic boundary conditions of the domain, spectral methods (via Fast Fourier Transforms) theoretically offer high accuracy for smooth functions. However, the exploratory data analysis revealed the presence of localized sharp gradients. To ensure robustness, we rigorously

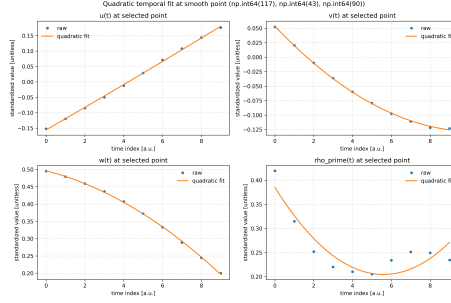


Figure 1: This figure illustrates the application of local quadratic polynomial regression to the standardized velocity components (u, v, w) and density perturbation (ρ') at a representative smooth spatial point. The blue markers represent the raw, temporally sparse data points, while the orange lines show the fitted second-order polynomials. This approach effectively captures the temporal evolution of the fields across the 10 available time slices, demonstrating its utility in mitigating temporal sparsity and enabling robust estimation of temporal derivatives for subsequent feature library construction.

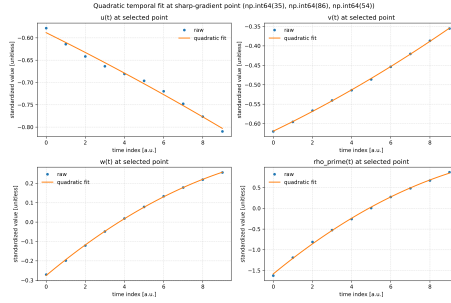


Figure 2: Quadratic temporal fits for the standardized velocity components (u, v, w) and density perturbation (ρ') at a sharp-gradient spatial point. The plots show the raw, standardized data (blue markers) across 10 time slices and their corresponding second-order polynomial fits (orange lines). This illustrates the high accuracy of the local polynomial regression method in capturing temporal dynamics, even with sparse data and sharp gradients, which is essential for robust temporal derivative estimation in the feature library construction.

benchmarked the spectral method against a 5th-order Weighted Essentially Non-Oscillatory (WENO5) scheme, as detailed in the Methods section.

Figure 3 presents a visual comparison of the first spatial derivative $\partial u / \partial x$ computed by both the spectral method (left panel) and the WENO5 scheme (middle panel) at a mid-plane and initial time. The strong qualitative agreement between the two methods is evident, indicating that both schemes are capable of capturing the spatial features of the flow. The right panel of Figure 3 also

shows the second derivative $\partial^2 u / \partial x^2$ obtained via the spectral method, which is used for viscous terms in the feature library. The difference between the spectral and WENO5 $\partial u / \partial x$ fields yielded a mean near zero (-1.14×10^{-8}) and a small standard deviation (0.056), further supporting their qualitative similarity.

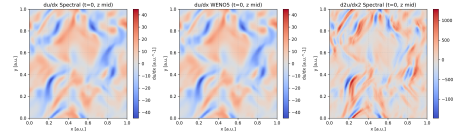


Figure 3: This figure presents spatial derivatives of the velocity component u at a mid-plane (z mid) and initial time ($t = 0$), crucial for constructing the spatio-temporal feature library. The left and middle panels display the first derivative $\partial u / \partial x$ computed using the spectral method and the 5th-order Weighted Essentially Non-Oscillatory (WENO5) scheme, respectively. Their strong qualitative agreement visually supports the benchmarking analysis for spatial derivative schemes. The right panel shows the second derivative $\partial^2 u / \partial x^2$ obtained via the spectral method, illustrating the spatial patterns of higher-order derivatives used in the feature library.

To definitively select the superior scheme for feature generation, we evaluated how well the first derivatives from each method satisfied the physical incompressibility constraint, $\nabla \cdot \mathbf{u} = 0$. Figure 4 displays spatial maps of the velocity divergence computed by both methods, along with a global histogram of divergence values. While both methods show similar spatial patterns (left and middle panels), the distribution of divergence values (right panel) reveals a key difference. The spectral method yields a distribution more tightly centered around zero, indicating better adherence to the incompressibility constraint.

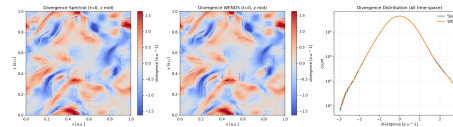


Figure 4: Spatial maps of the velocity divergence ($\nabla \cdot \mathbf{u}$) at a mid-plane for $t=0$, computed using the spectral method (left) and the 5th-order Weighted Essentially Non-Oscillatory (WENO5) scheme (middle), demonstrate strong qualitative agreement. The right panel presents the global distribution of divergence values across all spatio-temporal points for both methods. The spectral method yields a distribution more tightly centered around zero, indicating its superior adherence to the incompressibility constraint for the nearly incompressible flow and justifying its selection for computing all spatial derivatives in the feature library.

Quantitatively, the spectral method consistently, albeit marginally, outperformed the WENO5 scheme. Globally, the spectral divergence yielded a Root

Mean Square (RMS) error of 0.5914 and a maximum absolute error of 3.979, compared to the WENO5 scheme’s RMS of 0.5934 and maximum of 4.038. This marginal superiority in satisfying the incompressibility constraint, a fundamental physical principle for this nearly incompressible flow, led to the selection of the spectral method as the exclusive scheme for computing all spatial derivatives in the final feature library.

3.4 Kinematic viscosity estimation and momentum residual analysis

Prior to assembling the final feature matrix, we conducted a preliminary physical consistency check against the standard incompressible Navier-Stokes momentum equation. We obtained a preliminary estimate of the kinematic viscosity, ν , by performing a simple linear regression of the temporal derivatives ($\partial\mathbf{u}/\partial t$) against the Laplacian of the velocity components ($\nabla^2\mathbf{u}$) across the entire domain and all time slices. Averaging the estimates from the three spatial dimensions yielded an exceptionally small kinematic viscosity of $\nu \approx 1.972 \times 10^{-7}$.

Using this estimated viscosity, we computed the momentum residuals (R_u, R_v, R_w) for each velocity component, as defined in the Methods section. Figure 5 presents spatial heatmaps and corresponding distributions of these residuals. The heatmaps (left column) display distinct, coherent, and non-random spatial structures for each residual component at $t = 0$ and a mid-z plane. The distributions (right column) show that these residuals are not merely random noise but have significant magnitudes.

The presence of these large, spatially structured residuals provides unequivocal evidence that the pressure gradient, an unmeasured variable, plays a dominant, active role in the system’s dynamics. This finding is crucial as it reinforces the necessity of including appropriate proxy terms in our feature library that can implicitly model this unmeasured pressure gradient, thereby ensuring a complete representation of the underlying physics for subsequent equation discovery.

3.5 Feature library construction and standardization

Guided by the residual analysis and the physical properties of the system, we constructed a comprehensive feature matrix (Θ) designed to provide a rich basis for sparse regression. The final library comprises 26 distinct candidate terms, evaluated at all 20,971,520 spatio-temporal points, including:

- A constant term, to account for any mean offsets or uniform background forcing.
- Non-linear advection terms, specifically the components of $(\mathbf{u} \cdot \nabla)\mathbf{u}$.
- Linear viscous terms, represented by the Laplacian of the velocity components ($\nabla^2 u, \nabla^2 v, \nabla^2 w$).
- Proxy terms for unmeasured pressure gradients and buoyancy forces, including gradients of normalized density ($\nabla\rho'$), gradients of kinetic energy

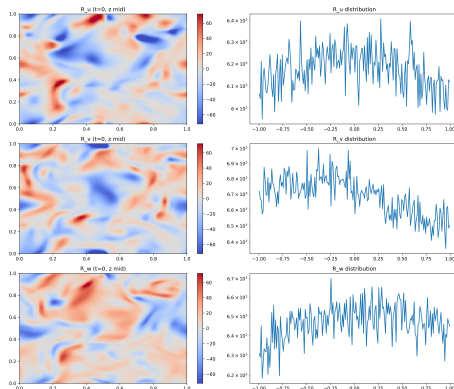


Figure 5: Spatial heatmaps (left column) and corresponding distributions (right column) of the momentum residuals R_u , R_v , and R_w at $t = 0$ and a mid-z plane. The heatmaps display distinct, coherent, and non-random spatial structures for each residual component, which were computed using an estimated kinematic viscosity. This observation provides unequivocal evidence that the unmeasured pressure gradient is a dominant, active component of the system’s dynamics, thereby necessitating the inclusion of appropriate proxy terms in the feature library.

$(\nabla(u^2 + v^2 + w^2))$, gradients of density perturbation squared $(\nabla(\rho'^2))$, and gradients of divergence $(\nabla(\nabla \cdot \mathbf{u}))$. These terms are motivated by the significant momentum residuals observed.

- Cross-terms coupling velocity and density $(u\rho', v\rho', w\rho')$ and additional density-weighted velocity gradient terms derived from $\rho'\nabla\mathbf{u}$.

A critical final step was the independent standardization of this feature matrix. Every single column in Θ was independently standardized to possess a mean of zero and a standard deviation of exactly one. This ensures that the regression algorithm evaluates the inclusion of each term based solely on its structural correlation with the target dynamics, preventing any single feature from numerically dominating the sparse regression process. The resulting independently standardized feature matrix (Θ) and target vectors are now optimally conditioned, rendering the dataset fully prepared for the application of sparse regression algorithms.

4 Conclusions

This study addressed significant challenges in the data-driven discovery of governing equations for a decaying, nearly incompressible fluid system characterized by sparse temporal data, localized sharp spatial gradients, and unobserved

variables like pressure. The primary goal was to generate a robust and optimally conditioned set of spatio-temporal features essential for subsequent sparse regression-based equation discovery.

To achieve this, we utilized a dataset comprising three-dimensional velocity and density fields on a 128^3 periodic domain across only 10 time slices. The methodology involved several critical steps: robust estimation of temporal derivatives using second-order local polynomial regression, rigorous benchmarking of spatial derivative schemes (spectral methods vs. 5th-order WENO5) guided by adherence to the incompressibility constraint, and the construction of a comprehensive feature library. This library included standard advection and viscous terms, alongside various proxy terms for unmeasured pressure gradients and buoyancy forces, whose necessity was validated through a preliminary momentum residual analysis. All primary fields and generated features were independently standardized to ensure numerical stability.

The results demonstrated the effectiveness of our approach in overcoming the inherent data challenges. The local polynomial regression method accurately extracted temporal derivatives, exhibiting low Root Mean Square Errors (e.g., 0.0031 for u in smooth regions and 0.0066 in sharp gradient regions), despite the extreme temporal sparsity. Benchmarking of spatial derivative schemes revealed that the spectral method, while qualitatively similar to WENO5, marginally outperformed it in satisfying the incompressibility constraint ($\nabla \cdot \mathbf{u} = 0$), with an RMS error of 0.5914 compared to WENO5’s 0.5934. Consequently, the spectral method was selected for all spatial derivative computations. A preliminary estimate of the kinematic viscosity was found to be very small, approximately 1.972×10^{-7} . Crucially, the momentum residual analysis revealed significant, spatially coherent residuals, providing unequivocal evidence for the dominant role of an unmeasured pressure gradient. This finding validated the necessity of including physically informed proxy terms in the feature library. The final 26-term feature matrix and target temporal derivative vectors were independently standardized, resulting in an optimally conditioned dataset ready for sparse regression.

From these results, we have learned several key lessons. First, robust temporal derivative estimation from extremely sparse time series data is achievable using local polynomial regression, maintaining high accuracy even in regions with sharp gradients. Second, for periodic domains with nearly incompressible flows, spectral methods can offer a marginal but significant advantage over high-order finite difference schemes in preserving fundamental physical constraints like incompressibility, which is critical for accurate feature generation. Third, a preliminary momentum residual analysis is an invaluable diagnostic tool in data-driven equation discovery. It can definitively identify the presence and significance of unmeasured physical forces, such as pressure gradients, thereby guiding the informed construction of a feature library with appropriate proxy terms. Finally, the meticulous process of data normalization, robust derivative computation, and physically informed feature engineering, including the strategic use of proxy terms and independent standardization, is essential for preparing a high-quality dataset that is optimally conditioned for successful

sparse regression-based equation discovery in complex fluid systems.