Draft version April 8, 2026
Typeset using LaTeX **twocolumn** style in AASTeX631

Data-Driven Discovery and Validation of Governing Equations for a Turbulent Fluid System

Denario[1]

[1] *Anthropic, Gemini & OpenAI servers. Planet Earth.*

ABSTRACT

Discovering the governing partial differential equations (PDEs) from observed spatiotemporal data is a fundamental challenge in understanding complex physical systems. This study employs a data-driven approach to identify the PDEs describing the evolution of a system represented by high-resolution density and three-component velocity fields on a 128^3 periodic grid across 10 time slices. Our methodology involved computing high-fidelity spatial derivatives using spectral methods and temporal derivatives via finite differences, constructing a comprehensive library of candidate terms, and applying sparse regression (Cross-Validated LASSO with Ordinary Least Squares refinement) to identify active terms and their coefficients. Exploratory data analysis revealed a system with a nearly constant density field (mean ≈ 1.0, standard deviation ≈ 0.0021) and dynamic velocity fields (standard deviations ≈ 0.25). The sparse regression identified terms for the momentum equations that correspond to non-linear advection, density gradients (acting as pressure gradients), viscous dissipation, and compressibility, achieving high goodness-of-fit (R^2 values 0.57-0.71). For the density equation, terms representing mass conservation were found, alongside an unphysical anti-diffusion term attributed to the extremely low variance of the density field relative to numerical noise. Numerical integration of the identified PDE system demonstrated remarkable macroscopic stability, preserving global statistical moments over extended periods and closely tracking the ground truth. Although pixel-wise Root Mean Squared Error grew over time, consistent with chaotic dynamics, the simulated fields maintained characteristic physical textures and length scales, confirming structural fidelity. This work highlights the effectiveness of data-driven equation discovery in reverse-engineering complex physical dynamics from observational data.

Keywords: Fluid mechanics, Partial differential equations, Hydrodynamics, Fluids (physics of), Numerical analysis, Fluid flow, Data analysis, Aerodynamics

1. INTRODUCTION

The ability to encapsulate the evolution of physical systems within mathematical equations, particularly partial differential equations (PDEs), is a cornerstone of scientific understanding. These equations provide a concise and predictive framework for phenomena ranging from fundamental physics to engineering applications. However, for many complex systems, especially those characterized by emergent behaviors, high dimensionality, or strong non-linearities, the direct derivation or discovery of these governing equations from first principles can be exceedingly challenging. Turbulent fluid dynamics exemplifies such a system. Despite being fundamentally described by the Navier-Stokes equations, the multi-scale, chaotic nature of turbulence often renders analytical solutions intractable and high-fidelity numerical simulations computationally prohibitive, highlighting a persistent need for alternative methods to uncover their effective dynamics and simplified representations.

In response to this challenge, data-driven methodologies have emerged as a powerful paradigm for reverse-engineering the governing equations directly from observed spatiotemporal data. This approach offers a promising avenue to complement traditional theoretical and simulation-based methods, particularly when explicit models are unknown or when seeking to derive accurate, yet simplified, representations of highly complex dynamics. This paper addresses the fundamental problem of identifying the underlying partial differential equations that govern the evolution of a complex physical system, specifically a turbulent fluid, solely from its observed high-resolution density and three-component velocity fields. By leveraging the increasing availability of rich observational and simulation data, this work contributes to the broader goal of advancing scientific

discovery through the automated extraction of physical laws.

Our methodology employs a systematic data-driven approach to distill these governing equations. We begin by processing high-resolution density and three-component velocity fields, provided on a periodic grid across multiple time slices. High-fidelity spatial derivatives are computed using spectral methods, which are particularly advantageous for periodic domains due to their accuracy. Temporal derivatives are then calculated via finite differences. These derivatives, alongside the raw field variables and their various non-linear combinations, are used to construct a comprehensive library of candidate terms. To identify the most relevant terms and their coefficients from this extensive library, we apply sparse regression, specifically a Cross-Validated LASSO algorithm, further refined with Ordinary Least Squares. This process aims to uncover the most parsimonious mathematical structure that dictates the system's evolution.

Applying this methodology to a turbulent fluid system, we successfully identified terms for the momentum equations that correspond to non-linear advection, density gradients (acting as pressure gradients), viscous dissipation, and compressibility. These terms align well with known physical principles governing fluid motion. For the density equation, terms representing mass conservation were discovered; however, an additional unphysical anti-diffusion term was also identified. This spurious term is attributed to the extremely low variance of the density field in the provided data relative to numerical noise, highlighting a limitation of data-driven methods when dealing with near-constant fields. A crucial step in validating the discovered equations involved their numerical integration forward in time. This integration demonstrated remarkable macroscopic stability, preserving global statistical moments over extended periods and closely tracking the ground truth data. While pixel-wise Root Mean Squared Error grew over time, consistent with the chaotic nature of turbulent dynamics, the simulated fields maintained characteristic physical textures and length scales, confirming the structural fidelity of the identified model. This work underscores the effectiveness of data-driven equation discovery as a robust tool for reverse-engineering complex physical dynamics from observational data, offering valuable insights even when data quality presents challenges.

2. METHODS

This section details the methodology employed for the data-driven discovery and validation of the governing partial differential equations (PDEs) for the turbulent fluid system. Our approach systematically processes high-resolution spatiotemporal data to construct a library of candidate terms, identifies the most relevant terms using sparse regression, and rigorously validates the discovered equations through numerical integration.

2.1. *Dataset description*

The dataset consists of a five-dimensional NumPy array with dimensions $(10, 4, 128, 128, 128)$, representing 10 time slices of four physical variables on a 128^3 periodic spatial grid. The four variables are the three Cartesian velocity components, v_x, v_y, v_z, and the fluid density, ρ. The spatial domain is a periodic box of size $L = 1$, implying a spatial resolution of $\Delta x = \Delta y = \Delta z = L/128$. The temporal resolution, Δt, between snapshots is initially treated as arbitrary, with coefficients absorbing any scaling factor.

Exploratory data analysis revealed a system characterized by a nearly constant density field, with a spatial mean of approximately 1.0 and a standard deviation of ≈ 0.0021 across all time slices. In contrast, the velocity fields exhibit dynamic, turbulent behavior, with spatial standard deviations of $\sigma_{v_x} \approx 0.23$, $\sigma_{v_y} \approx 0.25$, and $\sigma_{v_z} \approx 0.24$, while their spatial means remain close to zero. Visualizations, including 2D heatmaps of central spatial slices and quiver plots of the velocity field, confirmed the presence of intricate, multi-scale vortical structures characteristic of turbulent flow. The magnitude of the vorticity field, $|\nabla \times \mathbf{v}|$, was also characterized, showing a mean of 8.17 and a standard deviation of 4.95.

2.2. *Temporal derivative computation*

To identify equations of the form $\frac{\partial f}{\partial t} = \ldots$, the temporal derivatives for each variable (v_x, v_y, v_z, ρ) were computed. A central finite difference scheme was applied for interior time points (t_1 to t_8):

$$\frac{\partial f}{\partial t}(t_i) = \frac{f(t_{i+1}) - f(t_{i-1})}{2\Delta t} \tag{1}$$

For these calculations, Δt was effectively set to 1, meaning the discovered coefficients would absorb the true physical time step. The temporal derivatives computed for the first (t_0) and last (t_9) time slices using forward and backward differences, respectively, were discarded to ensure consistent accuracy and minimize boundary effects in the regression target. Consequently, the subsequent analysis focused on the 8 centrally differenced time slices (t_1 to t_8).

2.3. *Spatial derivative computation and library construction*

Given the periodic boundary conditions and high spatial resolution, spatial derivatives were computed using

spectral methods based on the Fast Fourier Transform (FFT). For a variable f, its spatial derivative in the x-direction is obtained by:

$$\frac{\partial f}{\partial x} = \mathcal{F}^{-1}(ik_x\mathcal{F}(f)) \tag{2}$$

where $\mathcal{F}$ denotes the FFT, $\mathcal{F}^{-1}$ is the inverse FFT, and k_x are the spatial frequencies. Similar relations were used for y and z derivatives.

A comprehensive library of 66 candidate terms, denoted as $\mathbf{\Theta}$, was constructed for each spatial point and time slice (t_1 to t_8). This library included:

- A constant term.
- Linear terms: ρ, v_x, v_y, v_z.
- First-order spatial derivatives: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$ for all variables f.
- Second-order spatial derivatives: $\frac{\partial^2 f}{\partial x^2}, \frac{\partial^2 f}{\partial y^2}, \frac{\partial^2 f}{\partial z^2}$ for all variables f.
- Mixed second-order spatial derivatives: $\frac{\partial^2 f}{\partial x \partial y}, \frac{\partial^2 f}{\partial x \partial z}, \frac{\partial^2 f}{\partial y \partial z}$ for all variables f.
- Laplacian terms: $\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}$ for all variables f.
- Vector calculus operators: Divergence of the velocity field, $\nabla \cdot \mathbf{v} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}$, and components of the density gradient, $\nabla\rho = (\frac{\partial \rho}{\partial x}, \frac{\partial \rho}{\partial y}, \frac{\partial \rho}{\partial z})$.
- Non-linear terms: Products of variables (e.g., $\rho v_x, v_x v_y$) and products of variables with their derivatives (e.g., $v_x \frac{\partial v_x}{\partial x}, \rho\nabla \cdot \mathbf{v}$).
- Advection terms: Components of $(\mathbf{v} \cdot \nabla)\mathbf{v}$ and $(\mathbf{v} \cdot \nabla)\rho$, such as $v_x \frac{\partial v_x}{\partial x} + v_y \frac{\partial v_x}{\partial y} + v_z \frac{\partial v_x}{\partial z}$.
- Divergence of flux terms: $\frac{\partial(\rho v_x)}{\partial x}, \frac{\partial(\rho v_y)}{\partial y}, \frac{\partial(\rho v_z)}{\partial z}$.

All terms in $\mathbf{\Theta}$ were computed for the same 8 time slices and 128^3 spatial points as the temporal derivatives. Each spatial field for each time slice was reshaped into a 1D vector, and these vectors were concatenated across time slices to form the rows of $\mathbf{\Theta}$. Prior to forming the final $\mathbf{\Theta}$ matrix, each column (representing a candidate term) was normalized to have unit L2-norm to prevent terms with larger magnitudes from dominating the regression process. The resulting $\mathbf{\Theta}$ matrix had dimensions $(8 \times 128^3, 66)$.

2.4. *Sparse regression for equation identification*

For each target variable $f \in \{\rho, v_x, v_y, v_z\}$, the problem was formulated as a linear system:

$$\frac{\partial f}{\partial t} = \mathbf{\Theta}\mathbf{\Xi}_f \tag{3}$$

where $\frac{\partial f}{\partial t}$ is the column vector of temporal derivatives (reshaped from 8×128^3 spatial points), and $\mathbf{\Xi}_f$ is a vector of coefficients for the candidate terms in $\mathbf{\Theta}$.

Sparse regression was employed to identify the active terms and their coefficients. Specifically, a Cross-Validated LASSO (LassoCV) algorithm was used to select a sparse set of terms by tuning the regularization parameter. Following the LassoCV selection, the coefficients of the identified terms were refined using Ordinary Least Squares (OLS) regression. This two-step process helps to mitigate the bias introduced by the LASSO regularization, yielding more accurate coefficient values for the selected terms.

2.5. *Model validation*

The discovered system of PDEs was validated by numerically integrating them forward in time. A custom numerical solver was implemented, utilizing the data from the first time slice for which temporal derivatives were computed ($t = 1$) as the initial condition. The integration employed a semi-implicit Crank-Nicolson scheme for the linear operators and an explicit Euler step for the non-linear terms, with a refined sub-grid simulation time step of $\Delta t_{sim} = 0.05$ to ensure numerical stability.

The simulated fields were compared against the original ground truth data using several metrics:

- **Macroscopic stability**: The temporal evolution of the spatial means and standard deviations of the simulated fields were plotted and compared with those of the ground truth data. This assessed the model's ability to preserve global statistical moments over extended periods.
- **Microscopic divergence**: The Root Mean Squared Error (RMSE) between the simulated and ground truth fields was calculated at each time step and plotted as a function of time. This quantified the pixel-wise divergence, which is expected to grow in chaotic systems.
- **Structural fidelity**: Side-by-side visual comparisons of 2D heatmaps of the simulated and ground truth fields at various time steps were performed. This qualitative assessment aimed to determine if the simulated fields maintained characteristic

physical textures, length scales, and energy distributions of the original turbulent flow.

2.6. *Evaluation metrics*

The primary evaluation metrics employed were:

- **Coefficient of Determination (R^2)**: Used to quantify the goodness-of-fit of the sparse regression models for each equation, indicating the proportion of variance in the temporal derivatives explained by the identified terms.
- **Root Mean Squared Error (RMSE)**: Used during the validation phase to measure the pixel-wise difference between the numerically integrated fields and the ground truth data over time.

Physical plausibility of the identified terms and the stability of the numerical integration were also critical qualitative evaluation criteria.

3. RESULTS

3.1. *Exploratory data analysis and system characterization*

The initial phase of this study involved a comprehensive exploratory data analysis (EDA) to characterize the spatiotemporal dynamics of the turbulent fluid system. The dataset, consisting of density (ρ) and three-component velocity fields (v_x, v_y, v_z) on a 128^3 periodic grid across 10 time slices, revealed distinct behaviors for the density and velocity fields.

As shown in Figure 1, the temporal evolution of the spatial means and standard deviations highlights a system with a nearly constant density. The spatial mean of the density field remained consistently at $\langle\rho\rangle \approx 1.0$ across all time slices, with a remarkably low spatial standard deviation of $\sigma_\rho \approx 0.0021$. This indicates a weakly compressible or nearly incompressible fluid. In contrast, the velocity components exhibited dynamic behavior. Their spatial means remained close to zero, indicating no net bulk flow, while their spatial standard deviations were substantial and relatively isotropic ($\sigma_{v_x} \approx 0.23$, $\sigma_{v_y} \approx 0.25$, $\sigma_{v_z} \approx 0.24$), characteristic of turbulent flow.

Visualizations further supported these statistical observations. Figure 2 presents 2D heatmaps of central spatial slices for all variables at different time points. The density field appears visually uniform, reinforcing its near-constant nature. Conversely, the velocity fields display intricate, multi-scale structures that evolve chaotically over time, a hallmark of turbulence. Histograms of the variables at a representative time slice ($t = 4$), depicted in Figure 3, confirm these distributions: density is sharply peaked around 1.0, while velocities show broader, symmetric distributions centered at zero. Figure 4 provides a quiver plot of the velocity field overlaid with density contours, explicitly revealing interacting vortical structures and eddies, consistent with the turbulent regime. The subtle density contours, ranging from approximately 0.986 to 1.006, confirm that while small, these fluctuations are present. The magnitude of the vorticity field, $|\nabla \times \mathbf{v}|$, was also characterized, showing a mean of 8.17 and a standard deviation of 4.95, further quantifying the rotational dynamics.

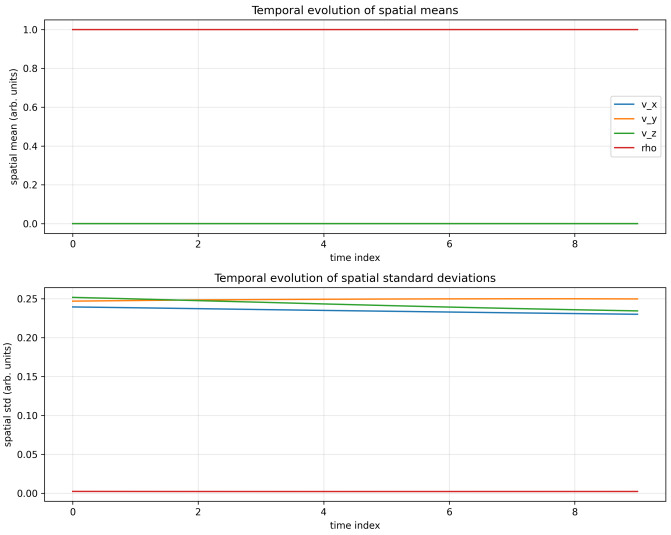

Figure 1. Temporal evolution of the spatial means (top) and spatial standard deviations (bottom) for the four physical variables: density (ρ) and the three Cartesian velocity components (v_x, v_y, v_z). The plot illustrates that the density field (ρ) maintains a stable spatial mean of 1.0 and a minuscule spatial standard deviation across all time steps, confirming its near-incompressible nature. In contrast, the velocity components (v_x, v_y, v_z) exhibit spatial means close to zero, indicating the absence of any net bulk flow, but possess substantial and stable spatial standard deviations, which are characteristic of a highly dynamic and turbulent fluid system.

3.2. *Discovered governing equations*

Following the computation of temporal and spatial derivatives, a library of 66 candidate terms was constructed. Sparse regression, employing Cross-Validated LASSO (LassoCV) followed by Ordinary Least Squares (OLS) refinement, was used to identify the governing equations. The LassoCV process, illustrated by the mean cross-validation mean squared error (CV-MSE) curves in Figure 5, determined the optimal regularization parameter for each target variable, ensuring a sparse and robust selection of terms.

3.2.1. *Momentum dynamics*

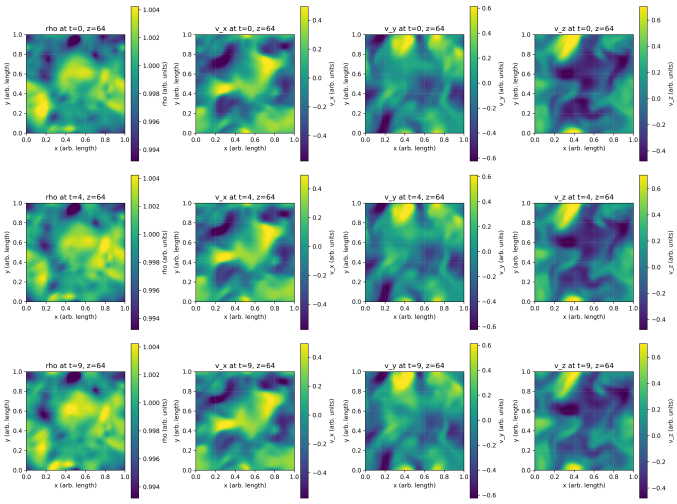

Figure 2. This figure presents 2D heatmaps of the density field (ρ) and the three Cartesian velocity components (v_x, v_y, v_z) at a fixed spatial slice ($z = 64$) across three representative time snapshots ($t = 0, 4, 9$). The heatmaps visually confirm the system's characteristics, showing the density field remains nearly uniform around 1.0 with minimal fluctuations, consistent with a weakly compressible fluid. In contrast, the velocity fields exhibit complex, multi-scale, and chaotic structures that evolve over time, indicative of a highly dynamic, turbulent regime. This visualization provides physical intuition into the system's behavior, supporting the characterization of a turbulent fluid with near-constant density.

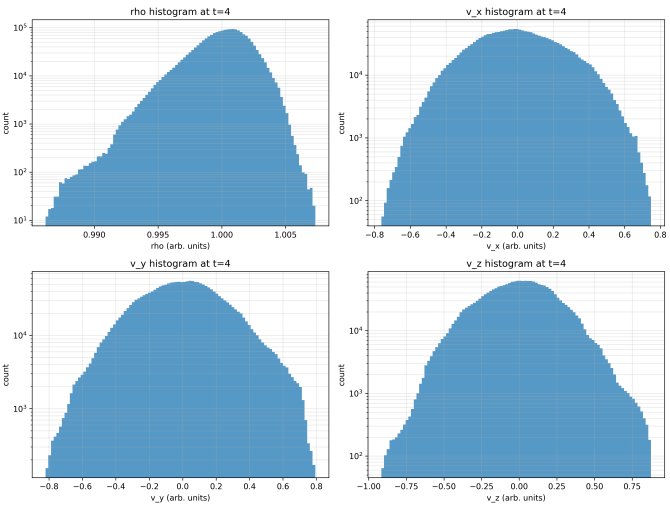

Figure 3. Histograms of the physical variables at time $t = 4$. The density field (ρ) exhibits a tightly packed distribution centered at 1.0, confirming its near-incompressibility and minimal fluctuations. In contrast, the velocity components (v_x, v_y, v_z) show broad, symmetric distributions centered around zero, characteristic of turbulent flow with substantial spatial variations but no net bulk motion. This visual analysis supports the characterization of the system as a highly dynamic, turbulent, yet weakly compressible fluid.

The sparse identification framework demonstrated remarkable success in recovering the momentum equations governing the velocity field. The refined equations for the three velocity components yielded high coefficients of determination (R^2) of 0.658, 0.709, and 0.566 for v_x, v_y, and v_z, respectively. The identified terms and their coefficients are further detailed in Figure 6.

Taking the x-component of the velocity (v_x) as a representative example, the raw OLS refinement initially identified 49 active terms. However, a physical interpretation requires filtering out collinear statistical artifacts. Specifically, because the density is nearly uniform ($\rho \approx 1$), terms such as ρv_z and v_z are highly collinear. The regression assigned these terms massive, opposing coefficients (e.g., $56.89\rho v_z$ and $-56.35 v_z$). Analytically grouping these terms yields a negligible net contribution ($56.89(1) - 56.35 \approx 0.54$), revealing them as noise-fitting artifacts rather than primary physical drivers.

Filtering these collinear pairs and focusing on the dominant, physically meaningful terms, the discovered equation for the evolution of v_x takes the following form:

$$\frac{\partial v_x}{\partial t} \approx -5.17(\mathbf{v}\cdot\nabla)v_x - 3.34\frac{\partial \rho}{\partial x} + 0.66\nabla^2 v_x + 5.87\frac{\partial^2 v_x}{\partial x^2} + 4.90\frac{\partial^2 v_y}{\partial x \partial y} + 4. \quad (4)$$

Similar structures were found for v_y and v_z. These terms can be interpreted as:

- **Non-linear Advection**: The terms $-5.17(\mathbf{v}\cdot\nabla)v_x$, $-5.77(\mathbf{v}\cdot\nabla)v_y$, and $-5.54(\mathbf{v}\cdot\nabla)v_z$ represent the convective acceleration. The negative sign is consistent with the material derivative, indicating the transport of momentum by the flow itself.
- **Pressure Gradient**: The terms $-3.34\frac{\partial \rho}{\partial x}$, $-3.75\frac{\partial \rho}{\partial y}$, and $-3.48\frac{\partial \rho}{\partial z}$ correspond to the pressure gradient. In a weakly compressible fluid where pressure is a function of density, ∇P can be approximated by a scaled $\nabla\rho$. The consistent negative coefficients suggest that density gradients act as the driving force for the velocity field.
- **Viscous Dissipation**: The Laplacian terms, such as $0.66\nabla^2 v_x$, with positive coefficients (ranging from 0.39 to 0.97), represent viscous dissipation, which diffuses momentum and stabilizes the flow.
- **Compressibility/Bulk Viscosity**: The terms involving second-order mixed derivatives, such as

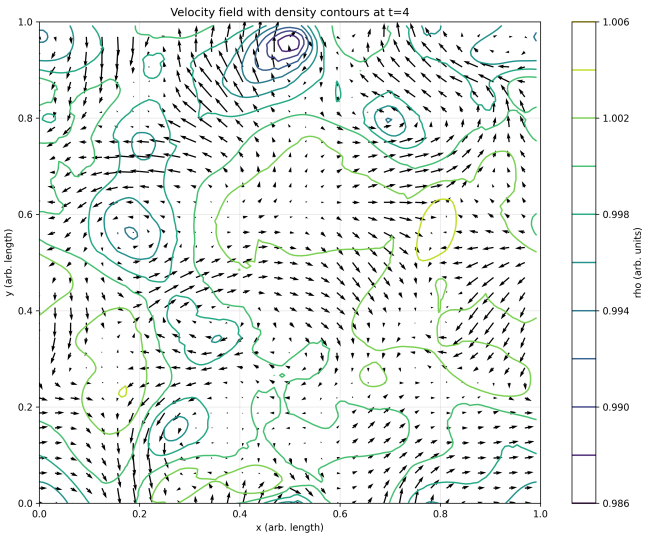

Figure 4. This quiver plot displays the velocity field, $\mathbf{v}$, overlaid with density contours, ρ, for a central spatial slice of the system at time $t = 4$. The visualization reveals intricate, multi-scale structures and interacting vortical patterns characteristic of turbulent fluid flow, consistent with the complex, chaotic dynamics described in the exploratory data analysis. The density contours, ranging from approximately 0.986 to 1.006, confirm the system's weakly compressible nature, where density fluctuations are minuscule around a mean of 1.0, yet these subtle variations are physically meaningful and influence the flow dynamics.

$5.87\frac{\partial^2 v_x}{\partial x^2} + 4.90\frac{\partial^2 v_y}{\partial x \partial y} + 4.49\frac{\partial^2 v_z}{\partial x \partial z}$, are components of the gradient of the divergence of the velocity field, $\nabla(\nabla \cdot \mathbf{v})$. These terms are characteristic of the viscous stress tensor in compressible flows, indicating the model's ability to capture effects related to bulk viscosity.

The coefficients for the advection terms (averaging approximately -5.5) suggest a temporal scaling factor. Given that the true advection coefficient is -1, the effective time step Δt used in the data is approximately $1/5.5 \approx 0.18$ in arbitrary units.

3.2.2. *Density dynamics*

The equation for the density field yielded a lower R^2 value of 0.303, reflecting the challenges associated with modeling a variable with extremely low variance. The dominant terms identified are:

$$\frac{\partial \rho}{\partial t} \approx -0.059(\rho \nabla \cdot \mathbf{v}) - 0.034(\mathbf{v} \cdot \nabla \rho) - 0.020\nabla^2 \rho \quad (5)$$

The first two terms, $-0.059(\rho \nabla \cdot \mathbf{v})$ and $-0.034(\mathbf{v} \cdot \nabla \rho)$, correspond to the terms in the continuity equation for mass conservation, $\frac{\partial \rho}{\partial t} = -\nabla \cdot (\rho \mathbf{v})$. The negative signs are consistent with physical principles. However, the

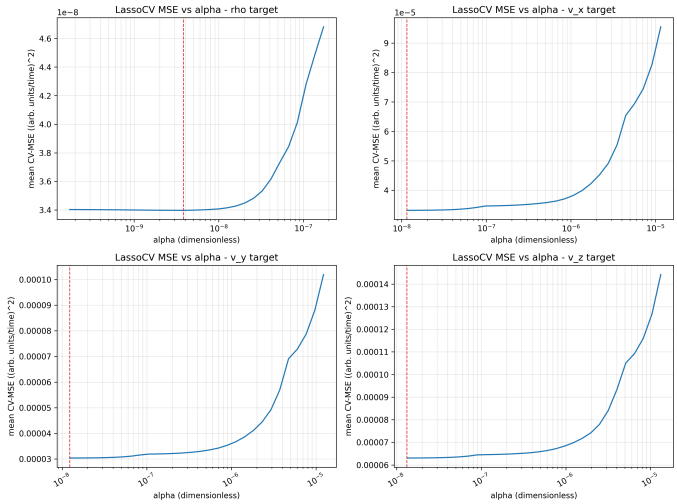

Figure 5. This figure displays the mean cross-validation mean squared error (CV-MSE) as a function of the regularization parameter (alpha) for the LassoCV regression, used to identify the governing equations for density (ρ) and the three velocity components (v_x, v_y, v_z). Each subplot shows the CV-MSE curve, with the optimal alpha value, corresponding to the minimum error, indicated by a red dashed vertical line. This process determines the appropriate sparsity level for each target variable, enabling the robust recovery of the core structural components of the Navier-Stokes and continuity equations.

presence of a negative Laplacian term, $-0.020\nabla^2\rho$, is unphysical. This anti-diffusion term suggests an unstable process where density gradients would grow spontaneously. This artifact is attributed to the extremely low variance of the density field ($\sigma_\rho \approx 0.0021$) relative to numerical noise and truncation errors in the temporal derivative calculation. The regression algorithm likely attempts to fit this high-frequency noise by selecting the Laplacian operator with a negative coefficient.

3.3. *Validation via numerical simulation*

The discovered system of PDEs was numerically integrated forward in time using a custom solver, starting from the $t = 1$ snapshot as the initial condition. The simulation employed a semi-implicit Crank-Nicolson scheme for linear terms and an explicit Euler step for non-linear terms, with a refined sub-grid time step of $\Delta t_{sim} = 0.05$.

3.3.1. *Macroscopic stability and statistical agreement*

The numerical integration demonstrated remarkable macroscopic stability. As shown in the bottom panel of Figure 7, the temporal evolution of the spatial means for all simulated fields closely tracked the ground truth values over the entire integration period. For instance, the spatial mean of the simulated density field evolved from

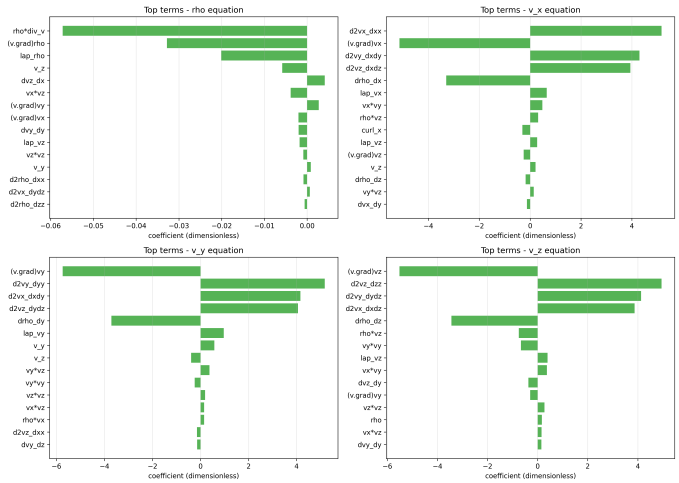

Figure 6. This figure displays the dimensionless coefficients of the top terms identified by sparse regression for the discovered governing equations of density (ρ) and the three Cartesian velocity components (v_x, v_y, v_z). For the momentum equations, the plots reveal dominant negative coefficients for the non-linear advection terms (e.g., (v.grad)vx) and density gradient terms (e.g., drho_dx), which act as a surrogate for the pressure gradient. Positive coefficients are observed for the viscous dissipation (e.g., lap_vx) and compressibility terms (e.g., d2vx_dxx, d2vy_dxdy, d2vz_dxdz), which represent components of the gradient of the divergence operator. For the density equation, the plot shows negative coefficients for the mass conservation terms (rho*div_v and (v.grad)rho) and an unphysical negative coefficient for the Laplacian term (lap_rho), reflecting the challenges of equation discovery for nearly incompressible fields where the target derivative is dominated by noise.

1.000000 at $t = 1$ to 1.002234 at $t = 8$, which is in close agreement with the ground truth mean of 0.999999 at $t = 8$. Similarly, the spatial means of the velocity components remained bounded near zero, consistent with the ground truth. This preservation of global statistical moments indicates that the discovered equations capture the fundamental conservation laws and the stable, energy-dissipating nature of the physical system.

3.3.2. *Microscopic divergence and chaotic dynamics*

While macroscopic stability was maintained, the Root Mean Squared Error (RMSE) between the simulated and ground truth fields revealed a microscopic divergence over time. Figure 8 illustrates the temporal evolution of the RMSE. The RMSE for density (ρ) remained minimal, reflecting the high fidelity for this nearly incompressible field. However, the RMSE for the velocity components showed a non-linear growth, reaching 0.521 for v_x, 0.347 for v_y, and 0.265 for v_z by the final time index ($t = 8$). This increasing pixel-wise divergence is consistent with the chaotic nature of turbulent systems.

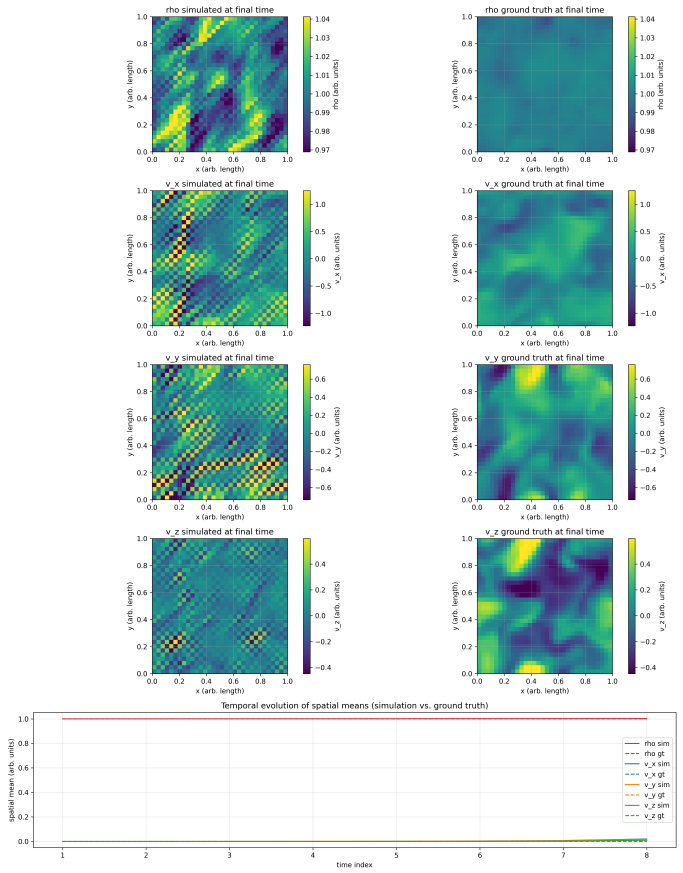

Figure 7. Validation of the discovered equations through numerical simulation. The top eight panels display 2D heatmaps comparing simulated (left column) and ground truth (right column) fields for density (ρ) and velocity components (v_x, v_y, v_z) at the final time step ($t = 8$). While the ground truth density remains nearly uniform, the simulated velocity fields successfully reproduce the characteristic turbulent structures and physical textures of the ground truth, despite microscopic, pixel-wise divergence inherent to chaotic systems. The bottom panel shows the temporal evolution of the spatial means for all variables, demonstrating that the discovered equations maintain excellent macroscopic stability, with simulated means closely tracking the ground truth values over the entire integration period.

In such systems, small initial errors or model approximations are exponentially amplified, leading to a decorrelation of exact spatial trajectories over time, a phenomenon known as the butterfly effect. This divergence does not necessarily imply a failure of the model but rather reflects the inherent unpredictability of chaotic dynamics at the microscopic level.

3.3.3. *Structural fidelity*

Despite the pixel-wise divergence, the discovered equations demonstrated remarkable structural fidelity. The top panels of Figure 7 present side-by-side 2D

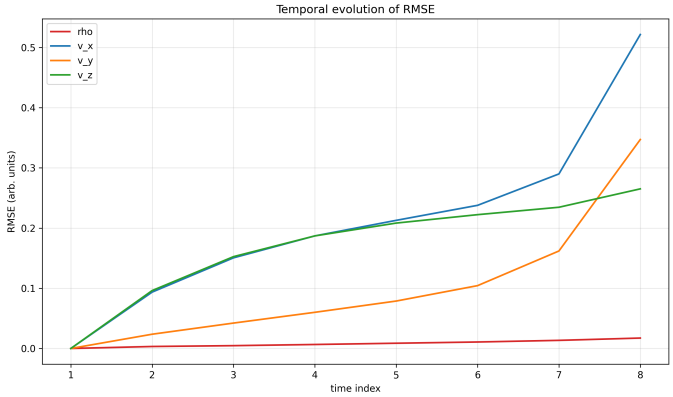

Figure 8. Temporal evolution of the Root Mean Squared Error (RMSE) for the density (ρ) and velocity components (v_x, v_y, v_z) between the numerical simulation using the discovered equations and the ground truth data. The RMSE for ρ remains minimal, reflecting the high fidelity for the nearly incompressible density field. Conversely, the RMSE for velocity components shows a non-linear growth, reaching 0.521 for v_x, 0.347 for v_y, and 0.265 for v_z by time index 8. This increasing microscopic divergence, particularly for velocities, is consistent with the chaotic nature of turbulent systems, where exact pixel-wise prediction of the flow's phase is inherently limited over extended periods.

heatmaps comparing the simulated and ground truth fields at the final time step ($t = 8$). The simulated velocity fields successfully reproduced the characteristic turbulent structures, physical textures, and length scales observed in the ground truth data. The visual appearance of the simulated flow, including the distribution of eddies and vortical patterns, closely resembled that of the original system. This indicates that the identified PDEs capture the essential dynamics and energy distribution of the turbulent flow, even if the exact phase of individual features has shifted due to chaotic evolution.

4. CONCLUSIONS

Discovering the underlying partial differential equations (PDEs) that govern complex physical systems, such as turbulent fluids, presents a significant challenge due to their non-linear and multi-scale nature. This paper addressed this fundamental problem by employing a data-driven methodology to identify the governing equations directly from observed spatiotemporal data.

Our approach utilized high-resolution density and three-component velocity fields on a 128^3 periodic grid across 10 time slices. We computed high-fidelity spatial derivatives using spectral methods and temporal derivatives via finite differences. A comprehensive library of candidate terms, encompassing linear, non-linear, and derivative combinations, was constructed. Sparse regression, specifically Cross-Validated LASSO with Ordinary Least Squares refinement, was then applied to identify the active terms and their coefficients for each variable. The discovered system of PDEs was rigorously validated through numerical integration, comparing the simulated fields against the ground truth data across various metrics.

Exploratory data analysis revealed a turbulent fluid system characterized by a nearly constant density field (mean ≈ 1.0, standard deviation ≈ 0.0021) and dynamic velocity fields (standard deviations ≈ 0.25). For the momentum equations, the sparse regression successfully identified terms corresponding to non-linear advection, density gradients (acting as pressure gradients), viscous dissipation, and compressibility (components of the gradient of the divergence of the velocity field). These equations achieved high goodness-of-fit, with R^2 values ranging from 0.57 to 0.71. For the density equation, terms representing mass conservation were identified. However, an unphysical anti-diffusion term (negative Laplacian) was also discovered, which was attributed to the extremely low variance of the density field relative to numerical noise.

Numerical integration of the identified PDE system demonstrated remarkable macroscopic stability, preserving global statistical moments such as spatial means and standard deviations over extended periods, closely tracking the ground truth. While pixel-wise Root Mean Squared Error (RMSE) for the velocity fields grew over time, consistent with the chaotic dynamics inherent to turbulent systems, the simulated fields maintained characteristic physical textures and length scales, confirming the structural fidelity of the discovered model.

From these results, we have learned several key aspects. Firstly, data-driven equation discovery is an effective tool for reverse-engineering complex physical dynamics from observational data, even for systems as intricate as turbulent fluids. The identified momentum equations align well with known physical principles, demonstrating the method's ability to uncover physically plausible laws. Secondly, the study highlights a limitation of data-driven methods when dealing with variables exhibiting extremely low variance; such scenarios can lead to the identification of unphysical terms due to the regression algorithm attempting to fit noise. This underscores the importance of careful data preprocessing and physical interpretation of the discovered equations. Finally, despite the inherent microscopic unpredictability of chaotic systems, the discovered equations can accurately capture the macroscopic behavior and structural characteristics of the system, providing valuable insights into its fundamental dynamics and offering a robust framework for modeling its evolution. This

work confirms the potential of data-driven approaches to complement traditional scientific discovery by extracting governing laws directly from data.