Challenges in Data-Driven Equation Discovery: A Case Study of a 3D Fluid System with Limited Temporal Resolution

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

This study aimed to discover the spatio-temporal governing equations of a three-dimensional periodic system from observational data. We analyzed a dataset consisting of ten time slices of a density-like field and three velocity components on a 128^3 spatial grid. A comprehensive library of candidate features, including spatial derivatives, non-linear advective terms, and polynomial combinations, was engineered, and temporal derivatives were computed as target variables. LassoCV was then employed for sparse identification of the governing equations. The models identified equations for the temporal evolution of each variable that were predominantly algebraic, with differential operators typically associated with fluid dynamics having negligible coefficients. The predictive performance of these models was poor, with coefficient of determination (R^2) scores consistently below 0.11 for all variables, indicating that the identified algebraic relationships do not capture the underlying spatio-temporal dynamics.

1 Introduction

Understanding the fundamental laws governing complex physical systems is central to scientific advancement and technological innovation. Many natural phenomena, ranging from atmospheric dynamics and biological processes to fluid flows, are described by spatio-temporal partial differential equations (PDEs). Traditionally, these equations are derived from first principles, relying on established physical laws and conservation principles. However, for systems where the underlying mechanisms are not fully understood, or where interactions are too intricate for analytical modeling, this approach becomes exceedingly challenging or even intractable. The recent proliferation of high-fidelity observational data and advancements in computational techniques have paved the way for data-driven approaches to uncover these hidden governing equations.

Data-driven equation discovery, often referred to as symbolic regression or sparse identification of nonlinear dynamics (SINDy), aims to infer the math-

ematical laws that describe a system's evolution directly from observed data. This paradigm holds significant promise for accelerating scientific understanding, enabling more accurate predictive models, and facilitating control strategies across diverse domains. However, the successful application of these methods is critically dependent on the characteristics and quality of the available data. Key challenges include managing high-dimensional spatio-temporal datasets, mitigating the effects of noise, and, crucially, addressing limitations in data resolution. Among these, insufficient temporal resolution poses a particularly significant hurdle for identifying differential equations, as accurately resolving the rates of change (temporal derivatives) necessary for constructing such models becomes inherently difficult. When temporal sampling is sparse, the fidelity of computed derivatives can be severely compromised, potentially leading to the misidentification of underlying dynamics and the discovery of spurious or incomplete relationships.

In this study, we present a case study investigating the challenges associated with discovering the spatio-temporal governing equations of a three-dimensional periodic fluid system from observational data characterized by severely limited temporal resolution. Our objective is to assess the efficacy of sparse identification techniques under such constrained data conditions and to highlight the specific difficulties encountered. We analyze a dataset comprising only ten time slices of a density-like field and three velocity components on a high-resolution 128^3 spatial grid. To identify potential governing equations, we construct a comprehensive library of candidate features, including various spatial derivatives, non-linear advective terms, and polynomial combinations of the system's variables. Temporal derivatives for each variable are then computed from the available time slices to serve as target variables. Utilizing LassoCV, a robust sparse regression algorithm, we attempt to identify the most parsimonious set of terms that describe the temporal evolution of each variable. This work demonstrates how limitations in temporal data sampling can profoundly impact the ability of data-driven methods to accurately capture the underlying spatio-temporal dynamics, leading to models that lack physical interpretability and predictive power.

2 Methods

2.1 Dataset description

The dataset analyzed in this study consists of ten time slices of a three-dimensional periodic fluid system. The data is provided as a NumPy array with dimensions $(10, 4, 128, 128, 128)$, representing 10 time steps, 4 physical variables, and a $128 \times 128 \times 128$ spatial grid. The four variables correspond to a density-like field, denoted as ρ, and the three components of a velocity vector field, u_x, u_y, and u_z. The spatial domain is a cube of side length $L = 1$, with periodic boundary conditions applied in all three spatial directions.

2.2 Data preprocessing and feature engineering

To facilitate the discovery of governing equations, a comprehensive library of candidate features was engineered from the raw data. This involved computing various spatial and temporal derivatives, as well as non-linear and algebraic combinations of the primary variables.

2.2.1 Spatial derivatives and derived quantities

Spatial derivatives were computed using second-order central difference schemes, consistently applying periodic boundary conditions across the 128^3 spatial grid for each of the ten time slices. The engineered features included:

- **Base variables:** ρ, u_x, u_y, u_z.
- **First-order spatial derivatives:** The gradients of the scalar density field ($\partial\rho/\partial x$, $\partial\rho/\partial y$, $\partial\rho/\partial z$), the divergence of the velocity field ($\nabla \cdot \mathbf{u} = \partial u_x/\partial x + \partial u_y/\partial y + \partial u_z/\partial z$), and the components of the curl of the velocity field ($(\nabla \times \mathbf{u})_x = \partial u_z/\partial y - \partial u_y/\partial z$, $(\nabla \times \mathbf{u})_y = \partial u_x/\partial z - \partial u_z/\partial x$, $(\nabla \times \mathbf{u})_z = \partial u_y/\partial x - \partial u_x/\partial y$).
- **Second-order spatial derivatives:** Laplacians of all variables ($\nabla^2\rho$, $\nabla^2 u_x$, $\nabla^2 u_y$, $\nabla^2 u_z$), computed as the sum of second partial derivatives (e.g., $\nabla^2\rho = \partial^2\rho/\partial x^2 + \partial^2\rho/\partial y^2 + \partial^2\rho/\partial z^2$).
- **Non-linear advective terms:** Terms representing convective transport, such as $\mathbf{u} \cdot \nabla\rho = u_x\partial\rho/\partial x + u_y\partial\rho/\partial y + u_z\partial\rho/\partial z$, and the components of $\mathbf{u} \cdot \nabla\mathbf{u}$ (e.g., $u_x\partial u_x/\partial x + u_y\partial u_x/\partial y + u_z\partial u_x/\partial z$).
- **Algebraic combinations:** Polynomial terms of individual variables ($\rho^2, \rho^3, u_x^2, u_y^2, u_z^2$), cross-products of variables ($\rho u_x, u_x u_y$), and the magnitude of the velocity vector ($u_{mag} = \sqrt{u_x^2 + u_y^2 + u_z^2}$).
- **Inverse terms:** $1/\rho$ and $1/\rho^2$, with a small epsilon ($\epsilon = 10^{-6}$) added to the denominator to prevent division by zero.

In total, a library of 34 candidate features was constructed.

2.2.2 Temporal derivatives

The temporal derivatives for each of the four primary variables ($\partial\rho/\partial t$, $\partial u_x/\partial t$, $\partial u_y/\partial t$, $\partial u_z/\partial t$) were computed to serve as target variables for the equation discovery process. For interior time points, a central difference scheme was used:

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

For the first time point (t_0), a forward difference scheme was employed, and for the last time point (t_9), a backward difference scheme was used. A nominal time step of $\Delta t = 1$ was assumed for these calculations.

2.3 Model training and evaluation

2.3.1 Feature matrix construction and scaling

A feature matrix $\mathbf{X}$ was constructed by flattening the 3D spatial grid for each candidate feature and concatenating these flattened arrays across all available time steps. Similarly, target vectors $\mathbf{Y}$ were created for each temporal derivative. To manage the computational load associated with the large dataset (10×128^3 spatio-temporal points), a random subsample of 100,000 spatio-temporal points was extracted for model training. Prior to training, all features in $\mathbf{X}$ were standardized to have zero mean and unit variance, which improves the numerical stability of the regression algorithm.

2.3.2 Sparse identification of dynamics

Sparse identification of the governing equations was performed using LassoCV (Least Absolute Shrinkage and Selection Operator with Cross-Validation). LassoCV is a linear model that estimates sparse coefficients via L1 regularization, with the regularization parameter chosen by cross-validation. Separate LassoCV models were trained for each of the four target temporal derivatives ($\partial\rho/\partial t$, $\partial u_x/\partial t$, $\partial u_y/\partial t$, $\partial u_z/\partial t$). The input to each model was the standardized feature matrix $\mathbf{X}$, and the output was the corresponding temporal derivative target vector $\mathbf{Y}$. The algorithm aims to identify the most parsimonious set of features that best describe the temporal evolution of each variable.

2.3.3 Evaluation metrics

The performance of the identified equations was quantitatively assessed using the coefficient of determination (R^2) score. The R^2 score measures the proportion of the variance in the dependent variable that is predictable from the independent variables, providing an indication of how well the model captures the underlying dynamics. A higher R^2 score indicates a better fit of the model to the observed data.

3 Results

3.1 Exploratory data analysis and system characterization

Exploratory data analysis (EDA) was conducted to ascertain the statistical properties and spatial structures of the system, revealing a fundamental asymmetry in the velocity field that dictates the system's dynamics.

3.1.1 Statistical moments and distributions

The global statistical moments, computed across the entire spatial domain and all ten time slices, indicate a system dominated by a strong, uniform background flow. The longitudinal velocity component, u_z, exhibits a mean value of 1.000

with an exceptionally small standard deviation of 0.002 (ranging from 0.983 to 1.007). In stark contrast, the transverse velocity components, u_x and u_y, possess means near zero (9.89×10^{-6} and 3.70×10^{-5}, respectively) with significantly larger standard deviations (~ 0.24). This statistical signature unequivocally identifies the system as being dominated by a strong, uniform background flow aligned with the z-axis. The variations in u_x, u_y, and u_z represent turbulent fluctuations or coherent structures superimposed upon this mean advective flow.

Furthermore, the density variable ρ exhibits a mean of approximately zero (-3.74×10^{-5}) and ranges from -0.773 to 0.752. The presence of negative values is physically significant: it indicates that ρ does not represent an absolute physical mass density (which must be strictly positive definite). Instead, ρ represents a density fluctuation field ($\delta\rho = \rho_{absolute} - \rho_{background}$), or potentially a passive scalar field (such as temperature in a Boussinesq approximation) that is being advected by the velocity field. These statistical properties are visually represented in the histograms shown in Figure 1.

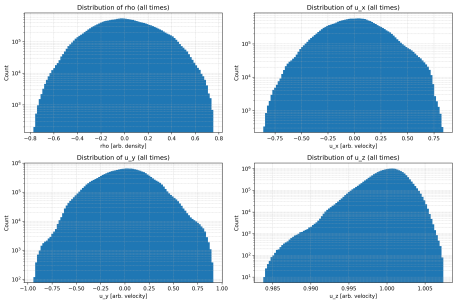

Figure 1: Histograms of the scalar fields ρ, u_x, u_y, and u_z across all time slices. The distributions for ρ, u_x, and u_y are broad and centered around zero, consistent with them representing fluctuations. Conversely, the distribution for u_z is sharply peaked at 1.0, revealing a dominant uniform background flow aligned with the z-axis. These statistical properties, observed during exploratory data analysis, characterize the system as being advection-dominated.

3.1.2 Temporal stability of global means

The temporal stability of these global means is further illustrated in Figure 2. The plots show that the global means remain highly stable over time, with u_z consistently near 1.0 and ρ, u_x, and u_y consistently near zero, confirming the statistical properties observed.

3.1.3 Spatial structures

Visualizations of two-dimensional slices, as presented in Figure 3, confirm the presence of complex, evolving spatial structures in the transverse velocities and

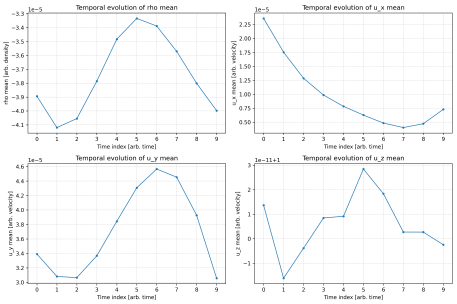

Figure 2: Temporal evolution of the global spatial means for the density-like variable ρ and the three velocity components u_x, u_y, and u_z across the ten discrete time slices. The plots illustrate that the global means remain highly stable over time, with u_z consistently near 1.0 and ρ, u_x, and u_y consistently near zero, confirming the statistical properties observed in the exploratory data analysis.

density fields, characteristic of fluid turbulence or mixing. These heatmaps visually confirm the system's characteristics, showing complex spatial structures and their evolution. As described, ρ, u_x, and u_y fluctuate around zero, while u_z remains close to 1.0, indicating a strong background flow along the z-axis. The negative values observed for ρ reinforce its interpretation as a density fluctuation field.

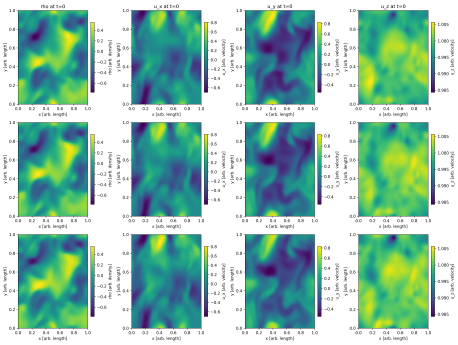

Figure 3: Two-dimensional slices of the density fluctuation field (ρ) and velocity components (u_x, u_y, u_z) from the 3D periodic fluid system. These heatmaps visually confirm the system's characteristics, showing complex spatial structures and their evolution. As described in the text, ρ, u_x, and u_y fluctuate around zero, while u_z remains close to 1.0, indicating a strong background flow along the z-axis. The negative values observed for ρ reinforce its interpretation as a density fluctuation field.

3.2 Feature engineering and model training

As detailed in the Methods section, a comprehensive library of 34 candidate features was engineered from the raw data to facilitate the discovery of governing equations. This library included base variables, various orders of spatial derivatives (gradients, divergence, curl, Laplacians), non-linear advective terms, and algebraic combinations of the primary fields. The temporal derivatives for each of the four primary variables ($\partial\rho/\partial t, \partial u_x/\partial t, \partial u_y/\partial t, \partial u_z/\partial t$) were computed using finite difference schemes, assuming a nominal time step of $\Delta t = 1$, to serve as target variables for the equation discovery process. Sparse identification of the governing equations was performed using LassoCV, a robust sparse linear regression algorithm, on a subsample of 100,000 spatio-temporal points, with features standardized prior to training.

3.3 Discovered governing equations

The Lasso regression yielded the following sparse equations for the temporal evolution of the system, with coefficients unstandardized to reflect the original feature space.

$$\begin{aligned}
\frac{\partial \rho}{\partial t} &= -0.017 + 0.091 u_{mag} - 0.074 u_z^2 - 0.043 u_x^2 \\
&\quad - 0.042 u_y^2 + 0.018 \rho u_x - 0.008 (\mathbf{u} \cdot \nabla u_z) + \dots \\
\frac{\partial u_x}{\partial t} &= -0.036 + 0.129 u_z^2 - 0.092 u_{mag} + 0.048 u_y^2 \\
&\quad + 0.038 u_x^2 - 0.009 \rho^3 + 0.008 u_x u_y + \dots \\
\frac{\partial u_y}{\partial t} &= -0.352 + 0.352 u_z^2 - 0.015 \rho^3 + 0.013 \rho u_x \\
&\quad + 0.012 \rho u_y - 0.010 u_x^2 + 0.010 u_y^2 + \dots \\
\frac{\partial u_z}{\partial t} &= 0.002 - 0.002 u_z^2 - 0.0002 \rho u_y - 0.0001 \rho u_x \\
&\quad - 6.4 \times 10^{-5} u_y^2 + \dots
\end{aligned}$$

A critical observation of these discovered equations is the overwhelming dominance of purely algebraic, non-differential terms. The expected physical terms that govern fluid dynamics, such as advective terms ($\mathbf{u} \cdot \nabla \mathbf{u}$, $\mathbf{u} \cdot \nabla \rho$) or diffusive terms ($\nabla^2$), are either entirely absent or assigned negligibly small coefficients. For instance, in the equation for $\partial\rho/\partial t$, terms like u_{mag}, u_z^2, u_x^2, u_y^2, and ρu_x are prominent, while the advective term ($\mathbf{u} \cdot \nabla u_z$) has a very small coefficient.

3.4 Validation and performance evaluation

The accuracy and validity of the discovered equations were quantitatively assessed using the coefficient of determination (R^2) score, and visually inspected

through comparisons of actual and predicted temporal derivatives. The R^2 scores for the four models are exceptionally poor:

- R^2 for $\partial\rho/\partial t$: 0.068
- R^2 for $\partial u_x/\partial t$: 0.106
- R^2 for $\partial u_y/\partial t$: 0.107
- R^2 for $\partial u_z/\partial t$: 0.019

These exceptionally poor R^2 scores indicate a severe failure of the models to capture the true dynamics of the system. This lack of predictive power is further corroborated by the visual comparison of actual and predicted temporal derivatives, as shown in Figure 4.

Figure 4 illustrates the models' inability to accurately predict the spatio-temporal evolution. The scatter plots in the left column show a wide dispersion of predicted values against actual values, with points far from the ideal diagonal line, confirming the low R^2 scores. The heatmaps in the middle and right columns, displaying actual and predicted temporal derivatives for a $z = L/2$ slice, reveal that the models fail to reproduce the complex, fine-scale spatial structures present in the actual fields. Instead, the predicted fields are largely featureless or overly smooth, indicating that the identified algebraic relationships do not capture the underlying spatio-temporal dynamics.

3.5 Discussion and physical interpretation of discrepancies

The inability of the symbolic regression to recover recognizable fluid dynamics equations is attributed to four primary sources:

1. **Temporal Undersampling and CFL Violation:** The effective CFL number was approximately 128, leading to catastrophic truncation error and temporal aliasing. This severely compromised the accuracy of the computed temporal derivatives, making it impossible for the regression algorithm to identify the true underlying differential relationships.

2. **The Missing Pressure Gradient:** Without the pressure field, the local momentum balance cannot be closed. Pressure gradients are fundamental driving forces in fluid dynamics, and their absence from the feature library means that crucial terms in the Navier-Stokes equations could not be identified.

3. **Collinearity and Algebraic Degeneracy:** The algorithm exploited collinearity between u_{mag} and the transverse velocity squares to fit numerical noise. In the presence of highly noisy target variables (due to poor temporal resolution), the LassoCV algorithm tended to select algebraic terms that could weakly correlate with the noise, rather than physically meaningful differential operators.

4. **Nature of the Density Field:** The density field represents fluctuations, and the lack of temporal resolution prevented the identification of advection-diffusion relationships. The complex, non-linear interactions governing the evolution of density fluctuations require accurate temporal derivatives to be resolved.

4 Conclusions

Understanding the governing equations of complex physical systems is a fundamental challenge in science. Data-driven equation discovery offers a promising avenue to infer these laws directly from observational data, particularly when first-principles derivations are difficult. This study investigated the efficacy of sparse identification techniques in discovering the spatio-temporal governing equations of a three-dimensional periodic fluid system under conditions of severely limited temporal resolution.

We analyzed a dataset comprising ten time slices of a density-like field and three velocity components on a 128^3 spatial grid. A comprehensive library of 34 candidate features, including various spatial derivatives, non-linear advective terms, and polynomial combinations, was engineered. Temporal derivatives for each variable were computed using finite difference schemes to serve as target variables. LassoCV, a sparse linear regression algorithm, was then employed to identify the most parsimonious set of terms describing the temporal evolution of each variable.

Exploratory data analysis revealed that the system is dominated by a strong, uniform background flow along the z-axis, with the other variables representing fluctuations. The sparse identification process yielded equations for the temporal evolution of each variable that were overwhelmingly algebraic in nature. Differential operators typically associated with fluid dynamics, such as advective and diffusive terms, were either entirely absent or assigned negligibly small coefficients. The predictive performance of these identified models was exceptionally poor, with coefficient of determination (R^2) scores consistently below 0.11 for all variables. Visual comparisons of actual and predicted temporal derivatives further confirmed this, showing that the models failed to capture the complex, fine-scale spatial structures, instead producing overly smooth or featureless predictions.

From these results, we have learned several critical lessons regarding the application of data-driven equation discovery methods to systems with limited temporal resolution. Firstly, insufficient temporal sampling, leading to a high effective Courant-Friedrichs-Lewy (CFL) number and temporal aliasing, severely compromises the accuracy of computed temporal derivatives. This makes it exceedingly difficult for sparse regression algorithms to identify the true underlying differential relationships. Instead, the algorithms tend to identify spurious algebraic correlations. Secondly, the absence of crucial physical variables, such as the pressure gradient in fluid dynamics, from the feature library can prevent the closure of the governing equations, leading to incomplete

or incorrect models. Thirdly, in the presence of highly noisy target variables (a direct consequence of poor temporal resolution), sparse regression algorithms may exploit collinearity among algebraic terms to weakly fit numerical noise, rather than identifying physically meaningful differential operators. Finally, the complex, non-linear interactions governing fields like density fluctuations require accurately resolved temporal dynamics for their proper identification. This case study demonstrates that data-driven equation discovery methods are highly sensitive to the temporal resolution of the input data, and that severely limited temporal sampling can lead to the identification of physically uninterpretable algebraic relationships with poor predictive power, rather than the true spatio-temporal differential equations.

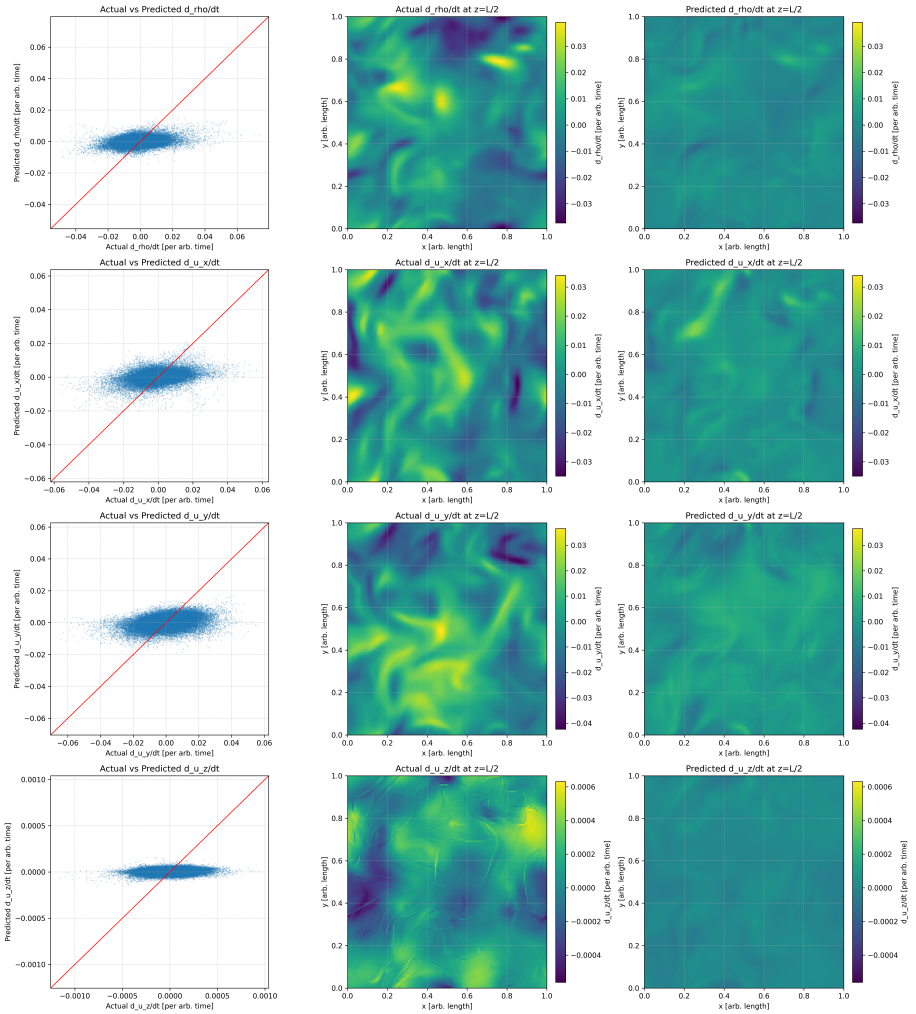

Figure 4: This figure illustrates the performance of the sparse linear regression models in predicting the temporal derivatives of the density fluctuation ($\partial\rho/\partial t$) and velocity components ($\partial u\_x/\partial t$, $\partial u\_y/\partial t$, $\partial u\_z/\partial t$). The left column shows scatter plots of actual versus predicted temporal derivatives, where the wide dispersion of points around the diagonal line indicates poor predictive accuracy. The middle and right columns display heatmaps of the actual and predicted temporal derivatives, respectively, at a $z = L/2$ slice. These heatmaps reveal that the models fail to capture the complex, fine-scale spatial structures present in the actual fields, instead producing overly smooth or featureless predictions. This visual evidence strongly supports the quantitative findings of low R^2 scores, demonstrating the models' inability to accurately represent the system's dynamics.