Proceedings of the 4th International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/2026.36370

Data-Driven Modeling of Nonlinear Dynamical Systems with Machine Learning

Mingyang Wang

Wilbraham and Monson Academy, Wilbraham, USA
mwang@wma.us

Abstract. Data-driven modelling of nonlinear dynamical systems based on machine learning is reviewed in this paper. Nonlinear dynamics refer to all sorts of complicated and irregular phenomena in mathematics, physics, engineering and life that are not well explained by linear models. Based on the accumulation of observation data from sensors, experiments and high-fidelity simulations, machine learning has recently begun to be applied in system identification, state estimation and the discovery of governing equations. This paper systematically presents the basic ideas of dynamical systems and then introduces several data-driven methods, such as Sparse Identification of Nonlinear Dynamics (SINDy), Neural Ordinary Differential Equations (Neural ODEs) and Koopman operator theory. Based on the above research, a number of methods have been proposed to solve the problems of noise, irregular sampling, etc., in high-dimensional chaos, and their respective advantages are introduced below. At present, the problems of poor data quality, partial observability and lack of interpretability in deep learning models are well-known. In the future, Physics-Informed Machine Learning, uncertainty quantification and robust model methods will be applied in engineering. Data science and applied mathematics will be combined in this paper to offer a comprehensive introduction to the problems and models of complex systems.

Keywords: data-driven modeling, nonlinear dynamical systems, machine learning, sparse identification, neural ordinary differential equations

1. Introduction

Many kinds of complex phenomena in science and other fields can be studied and predicted using nonlinear dynamical systems. There are numerous non-linear phenomena in the chaotic flow of fluids of aerospace engineering and the complex population dynamics of biological ecosystems, as well as the irregular fluctuations of financial markets and prolonged changes in climate science. Over time, the development of the system in the above areas is often highly sensitive to the initial conditions, and thus the resulting fluctuations are complex and cannot be approximated linearly. Traditionally, to learn about these systems, one had to derive them from first principles and gain domain knowledge. With the development of modern sensing technology, many time-series datasets have been obtained through experiments and large-scale numerical simulations, and now a new era has begun. Extraction of governing laws for the future states of the system and characterisation of

DOI: 10.54254/2753-8818/2026.36370

chaotic behaviour from observational data have become research problems at the intersection of data science and applied mathematics [1-3]. To help solve problems or extend the scope of old mathematical models, one can now use machine learning.

Recently, the intersection of machine learning and dynamical systems has achieved good results in nonlinear system identification, time series forecasting and discovery of physical knowledge. Some researchers have developed many ways to obtain the data of complex systems. Sparse Modelling Methods, such as Sparse Identification of Nonlinear Dynamics (SINDy), have been employed to obtain a small number of simple and interpretable governing equations directly from data [1]. At the same time, RNNs and LSTM networks are also kinds of deep learning models that have been used widely to predict the behaviour of high-dimensional chaotic systems. Neural Ordinary Differential Equations (Neural ODEs) can be used to model the continuous-time evolution of a time series by having a neural network parametrize the derivative of the hidden state, and thus they are closely related to discrete deep learning [4, 5]. Methods of operator theory, such as those based on the Koopman operator, can be used to map nonlinear dynamics into infinite-dimensional linear spaces for the purpose of linear analysis [6]. The above methods are often compared with the well-known benchmark systems of the Van der Pol oscillator, the Duffing oscillator, the chaotic Lorenz system and the discrete logistic map. This review examines how all these algorithms use state variables, time steps, initial conditions and noisy observation sequences, and then systematically compares the application of linearization methods, SINDy, deep learning frameworks and Koopman-based methods [2, 3].

The first goal of this paper is to introduce the general ideas of machine learning and data-driven modeling for nonlinear dynamical systems. Some examples of how observational data can help discover physical laws, predict the short-term behaviour of systems and learn about complex models will be presented in this paper. The following sections will be organised as follows: Section 2 introduces the general ideas of nonlinear dynamical systems and system identification. Section 3 will introduce some general data-driven methods and machine learning workflows for this problem. Section 4 is a review of related work and comparisons of these studies. Section 5 lists the deficiencies of the current methods and puts forward some research directions. Finally, Section 6 concludes this paper. Through this systematic review, it has been shown that data science techniques can be applied to connect mathematical models with real-world time-series data and solve many problems in applied mathematics.

2. Basic concepts of nonlinear dynamical systems and system identification

Before carrying out some research on data-driven modeling, one must first learn the foundation of nonlinear dynamical systems. The change in the values of some variables of a system over time is called a dynamic process. The system can be expressed continuously by ordinary differential equations (ODEs), where $x(t)$ is the state vector, t is time, $u(t)$ is the external control input, and f is a nonlinear vector field that determines the evolution of the system. Discrete-time systems can also be expressed in the form of difference equations or iterative maps. Linear systems obey the principle of superposition, and their behaviour can often be analysed using eigenvalues; nonlinear systems exhibit more complex phenomena. There are many equilibria, limit cycles (periodic motion), bifurcations that change the structure of the system based on different parameters, and deterministic chaos [6]. Chaotic systems, such as the Lorenz system, are very sensitive to small changes in their initial conditions; therefore, even though the laws of nature are fixed, it is practically impossible to know the exact state of the system far in advance. Although the world cannot be understood as a

DOI: 10.54254/2753-8818/2026.36370

whole, by calculating the Jacobian matrix at an equilibrium point in a certain area, one can determine the local stability and transient behavior.

To build a mathematical model of a dynamic system based on measured data is called system identification. Most of the system identification in classical control theory is based on linear models, such as ARMAX models, or fixed-form nonlinear models. Modern data-driven modelling is now able to find the unknown function f directly from various kinds of data [7]. The observational data used in the modelling tasks can take the form of continuous-time high-frequency sensor data, irregularly sampled discrete-time measurements, spatially distributed field data, etc. The set of sample state variables has many interaction terms, explicit parameter information and applied control inputs. The quality of the data is not uniform; some have been obtained through very precise deterministic numerical simulation, and others are the result of physical experiments and are therefore subject to environmental noise and measurement errors. Given the structure of this nonlinear system and the properties of the available data, the first step before applying advanced machine learning algorithms for system identification is to gain some knowledge about these characteristics [8].

3. Data-driven methods for nonlinear dynamical system modeling

The basis for building data-driven models of nonlinear dynamical systems is relatively wide, and there are many kinds of regression methods as well as many complex, high-parameter deep learning models. Linear regression and polynomial regression are basic methods to describe the simple changes at the foundation level. However, a large change occurred with the introduction of the Sparse Identification of Nonlinear Dynamics (SINDy) method [3]. SINDy constructs a library of candidate nonlinear functions, such as polynomials and their derivatives, and then uses sparse regression to find the smallest set of terms that can describe the system; sequentially thresholded least squares is one such method. Thus, there are simple, mathematically understandable ODEs that directly reflect the laws of nature. Kernel methods and ensemble learning techniques, such as Random Forests, can be used to model the state transition function non-parametrically, and although they are robust to all kinds of noise, they often lack the interpretability of analytical equations [1].

Given that deep learning can approximate any continuous function, it has been applied to model dynamical systems. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are suitable for processing sequences by nature and can learn all sorts of time-varying relationships and memory effects in many physical systems well. Neural Ordinary Differential Equations (Neural ODEs) have recently changed the field by taking the dynamics of the hidden state to be the solution of an ODE parametrized by a neural network. This continuous-time formulation can handle irregular data sampling naturally, and it is suitable for combining deep learning with traditional numerical integration [4, 5]. Another good way is reservoir computing; it is a fixed, randomly connected recurrent network that projects the input data into a high-dimensional space, and only a simple linear readout layer needs to be trained. Thus, reservoir computing can be used to forecast chaotic systems. Koopman operator methods are used to find observable functions that take the form of neural networks (often autoencoders) to map nonlinear dynamics into an infinite-dimensional linear space, and then conventional linear forecasting and control techniques can be applied to nonlinear problems [6].

The following are the steps in the data processing and validation process. The first stage is to clean the time series, reduce noise, normalise the variables and divide the data into training, validation and test sets. A sliding window and cross-validation need to be carried out on the input data for time series. Typical indicators of the performance of models are the mean absolute error

DOI: 10.54254/2753-8818/2026.36370

(MAE), root mean square error (RMSE) and R-squared (R^2) values. Dynamical systems are also subject to prediction time; that is to say, how long the model will be accurate before errors accumulate to reduce performance, and one should consider the long-term stability of the predicted attractor. Interpretability analysis is increasingly needed to help us understand which state variables and coupling mechanisms affect the general development of the system, so that the learned model can be in line with reality [3].

4. State-of-the-art studies of data-driven nonlinear dynamical system modeling

Recently, scholars at home and abroad have been studying how to apply machine learning for the construction of data-driven models of nonlinear dynamical systems with good results. A problem of the present time is how different methods handle various kinds of systems, such as simple periodic oscillators and complex, chaotic fluid flows. For example, studies of the Lorenz system and the Kuramoto-Sivashinsky equation involve trade-offs between interpretable sparse modelling and high-capacity deep learning. Champion and others [3] have been able to use autoencoders in conjunction with the SINDy framework to jointly discover general coordinate transformations and corresponding sparse governing equations from high-dimensional video data, and it has been shown that parsimonious physical laws can be obtained even when the natural state variables are not known explicitly. Deep learning has been used more recently to improve the accuracy of forecasts in all sorts of cases. DUE (Deep Learning Framework for Unknown Equations) is used to approximate the unknown terms in partial differential equations with neural networks, and good results have been achieved for multivariable coupled systems [2].

Based on the comparison of studies, the sources of data and types of errors are not the same. Based on the above results, SINDy can be used to obtain interpretable governing equations from sufficiently sampled data, but it may be noisy, have unobserved variables, or be very stiff [1, 9]. Caldana and Hesthaven [5] have employed Neural ODEs to address the problem of stiff systems and have found that continuous-time neural networks can effectively model multi-scale changes in time. Goyal and Benner [4] have found that Neural ODEs are suitable for experimental conditions with irregularly sampled or noisy data, and their continuous form can be interpolated and integrated over the missing time intervals. The aim of all the above research is to maintain long-term stability; that is, the reconstructed attractor map should be in agreement with the true system [7, 8].

As shown in the representative studies above, the general workflow of data science for complex systems can be summarised as follows: start with raw time-series data, construct various features (such as delay embeddings or latent space projections), train a model, and finally interpret the physical meaning. Based on the above, both the volume of data and the sampling frequency are high. Deep networks have shown that they can approximate a large amount of fluid dynamics data very well [4], but for small, highly controlled biological data, the inductive bias of sparse identification is needed. Finally, the above research has shown that to realise good results with data-driven modelling, one needs to address problems such as poor data quality, select a suitable mathematical model, set clear performance standards, etc., and so on, in order to better understand complex nonlinear phenomena.

5. Limitations and prospects

Although there have been some achievements in the application of data-driven modelling for nonlinear dynamical systems, many serious problems still need to be addressed. First and foremost, the quality of the data is relatively poor. The observation data obtained from real-world physical or

DOI: 10.54254/2753-8818/2026.36370

engineering systems will always have measurement noise, and there will often be missing values due to sensor failure or transmission errors. Complex nonlinear relations, especially those with high-order derivatives or rapid fluctuations, are very sensitive to the deficiencies mentioned above and often result in the identification of spurious terms by SINDy or overfitting to noise in deep learning models [1, 10]. Additionally, there is a problem of partial observability; that is, in many complicated systems, only some of the state variables can be observed directly. Reconstruction of the entire state space from limited measurements is usually done via Takens' embedding theorem, and it is very complex and uncertain.

Another serious problem is that the system is chaotic in nature. Because they are sensitive to the initial state, even the best mathematical models will gradually accumulate prediction errors over time. Data-driven models will have approximation errors, and if these models are used to make forecasts over a long period, the errors will accumulate exponentially and become unreliable with time [7]. Deep learning models, such as deep RNNs and many Neural ODEs, have achieved good results in the accuracy of short-term prediction, but they are generally not easy to interpret mathematically. These "black-box" models do not show the reasons for the changes in the system and are therefore difficult to confirm with existing knowledge about physics. Therefore, the cross-scenario generalisation ability is often weak; that is to say, a model trained in one dynamic regime may perform poorly in another dynamic regime or under new operating conditions that are not in the training data [2, 8].

Looking ahead, to address the above shortcomings, relevant knowledge from the physical world needs to be incorporated. One of the most promising directions is the development of Physics-Informed Machine Learning (PIML). Known physical constraints, conservation laws, symmetries, etc., can be added to the loss function of a neural network, and then Physics-Informed Neural Networks (PINNs) can reduce the hypothesis space to improve generalization and robustness in low-data regimes [11]. Data from all places can be gathered and combined with high-resolution but infrequent simulation data to improve the accuracy of the model. Explainable AI (XAI) can be developed for the purpose of mathematical modelling to obtain verifiable theories of complex networks. Further studies will be carried out on uncertainty quantification, and a Bayesian framework can be employed to provide confidence intervals for the predictions in safety-critical applications [12]. Finally, active learning and an optimal experimental design will be employed to have the algorithm proactively obtain data, and based on this information, it will dynamically adjust the position and sampling frequency of the sensors to build very reliable models of real-world engineering systems.

6. Conclusion

This paper has presented some data-driven model construction methods for nonlinear dynamical systems and examined the problems of data quality and machine learning design. Sparse Identification of Nonlinear Dynamics can obtain an interpretable governing equation from a suitable candidate library and sufficiently clean observations. Neural Ordinary Differential Equations can be used for continuous-time models and irregular data sampling; other options for sequence prediction are recurrent networks and reservoir computing. Koopman-type methods are used to find observables that can be written as the linear evolution of nonlinear dynamics. All of the above will affect both the performance and the prediction accuracy of the model; this includes the sampling method, noise level, degree of observability, forecast horizon and stability of reconstructed dynamics. None of the above studies have found one method that can be applied to all periodic, stiff, high-dimensional and chaotic systems. Sparse models are usually suitable for equation discovery,

DOI: 10.54254/2753-8818/2026.36370

but if it is necessary to model complex state transitions based on a large amount of data, high-capacity neural networks can also be used. The main defects are still the sensitivity to noisy derivatives, incomplete measurements, long-term accumulation of errors, limited interpretability and weak generalization outside the training region. Future research will introduce physical constraints, uncertainty quantification, explainable models, multi-fidelity data and active experimental design. The same data division and several error and stability indicators should be used, and the computational cost should be shown explicitly in order to compare them more fairly. Generally speaking, when machine learning is used to improve the knowledge of dynamical systems, data-driven modeling is more reliable, and together they can be applied to system identification, prediction and control.

References

[1] North, J. S., Wikle, C. K., & Schliep, E. M. (2023). A review of data-driven discovery for dynamic systems. *International Statistical Review, 91*(3), 464-492. https://doi.org/10.1111/insr.12554

[2] Chen, J., Wu, K., & Xiu, D. (2025). DUE: A deep learning framework and library for modeling unknown equations. *SIAM Review, 67*(4), 873-902. https://doi.org/10.1137/24M1671827

[3] Champion, K., Lusch, B., Kutz, J. N., & Brunton, S. L. (2019). Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences, 116*(45), 22445-22451. https://doi.org/10.1073/pnas.1906995116

[4] Goyal, P., & Benner, P. (2023). Neural ordinary differential equations with irregular and noisy data. *Royal Society Open Science, 10*(7), 221475. https://doi.org/10.1098/rsos.221475

[5] Caldana, M., & Hesthaven, J. S. (2024). Neural ordinary differential equations for model order reduction of stiff systems. *International Journal for Numerical Methods in Engineering, 125*(19), e70060. https://doi.org/10.1002/nme.70060

[6] Klus, S., & Djurdjevac Conrad, N. (2024). Dynamical systems and complex networks: A Koopman operator perspective. *Journal of Physics: Complexity, 5*(4), 041001. https://doi.org/10.1088/2632-072x/ad9e60

[7] Chakraborty, D., Chung, S. W., Arcomano, T., & Maulik, R. (2024). Divide and conquer: Learning chaotic dynamical systems with multistep penalty neural ordinary differential equations. *Computer Methods in Applied Mechanics and Engineering, 432*, 117442. https://doi.org/10.1016/j.cma.2024.117442

[8] Cheng, S., Quilodran-Casas, C., Ouala, S., Farchi, A., Liu, C., Tandeo, P., Fablet, R., Lucor, D., Iooss, B., Brajard, J., Xiao, D., Janjic, T., Ding, W., Guo, Y., Carrassi, A., Bocquet, M., & Arcucci, R. (2023). Machine learning with data assimilation and uncertainty quantification for dynamical systems: A review. *IEEE/CAA Journal of Automatica Sinica, 10*(6), 1361-1387. https://doi.org/10.1109/JAS.2023.123537

[9] Akbarzadeh, M., Halkon, B., & Oberst, S. (2025). Sparse identification of nonlinear dynamics applied to the acoustic levitation of acoustically large objects. *Scientific Reports, 15*(1), 40745. https://doi.org/10.1038/s41598-025-24666-5

[10] Wang, J., Moreira, J., Cao, Y., & Gopaluni, R. B. (2023). Simultaneous digital twin identification and signal-noise decomposition through modified generalized sparse identification of nonlinear dynamics. *Computers & Chemical Engineering, 177*, 108294. https://doi.org/10.1016/j.compchemeng.2023.108294

[11] Yu, J., Lu, L., Meng, X., & Karniadakis, G. E. (2022). Gradient-enhanced physics-informed neural networks for forward and inverse PDE problems. *Computer Methods in Applied Mechanics and Engineering, 393*, 114823. https://doi.org/10.1016/j.cma.2022.114823

[12] Tyralis, H., & Papacharalampous, G. (2024). A review of predictive uncertainty estimation with machine learning. *Artificial Intelligence Review, 57*, Article 94. https://doi.org/10.1007/s10462-023-10698-8