communications earth & environment

Perspective

https://doi.org/10.1038/s43247-024-01760-6

Towards data-driven discovery of governing equations in geosciences

Wenxiang Song [1], Shijie Jiang [2,3], Gustau Camps-Valls [4], Mathew Williams [5], Lu Zhang [1,6], Markus Reichstein [2,3], Harry Vereecken [7], Leilei He [1], Xiaolong Hu [1] & Liangsheng Shi [1]

Governing equations are foundations for modelling, predicting, and understanding the Earth system. The Earth system is undergoing rapid change, and the conventional approaches for establishing governing equations, such as empirical generalisations, are becoming increasingly challenging to deal with the complexity and diversity of the geoscience processes we study today. In this Perspective, we explore data-driven equation discovery, a novel scientific artificial intelligence pathway, for advancing geosciences. Data-driven equation discovery identifies hidden patterns from data and transforms them into interpretable equation representations, automating and accelerating equation discovery processes. It provides a practical approach for geoscientists to model and understand complex geoscience processes based on big Earth data. The final vision is to uncover new clear, describable, and quantifiable equations in various geoscience disciplines. We summarize opportunities and highlight that challenges in this field should be addressed by interdisciplinary collaborations.

Modelling and understanding geoscience processes is crucial to predicting natural phenomena and mitigating the impacts of environmental challenges under global change. Earth system processes are typically represented by governing equations in the form of symbolic models, which describe how the values of unknown variables change in response to variations in one or more known variables[1]. Governing equations inherently entail concepts of time, space, causality, and generality, defining the evolution of geophysical, -chemical, -biological, -mechanical and ecological processes[2] with interpretability and accessibility. Historically, the paradigm for establishing governing equations in geosciences has been rooted in constructive and principled theories (Box 1). The former approach derives equations for phenomena based on first principles, such as conservation laws, symmetries, physical regulations, and phenomenological behaviours[3]. The latter approaches are empirical or semi-empirical generalisations summarised and parameterised to capture the main features. For centuries, the classical paradigm has resulted in ubiquitous canonical governing equations in geosciences across various scales and processes, which are illustrated in Fig. 1a. These equations are fundamental to Earth and climate sciences[4].

Despite the historical success, the natural processes we study today are often more complex and multifaceted rather than simple and single processes, such as modelling coupled dynamic components of the Earth system (e.g., atmosphere, ocean, biosphere, cryosphere, carbon-water-nutrient cycling, ecological dynamics). Limited knowledge makes it hard to define accurate variables, and simplifying assumptions can lead to errors and oversimplifications that don't reflect real-world complexity[5]. This is true even for basic equations, such as empirical equations for the physical properties of gases that are not entirely ideal[6], let alone in subsystems dominated by nonlinearity, stochasticity, multiscale couplings, nonequilibrium behaviour, and spontaneous behaviour[6]. In addition, scientific discoveries that adhere to the classical paradigm rely on the creative and intellectual insight of scientists and require continuous trial-and-error approaches for incremental improvement[7]. Nevertheless, scientists are also limited in processing and analysing hidden patterns that are not immediately apparent in complex datasets. Consequently, progress in establishing and refining the governing equations in these systems has been slow over the past several decades.

With advances in sensor and data storage technologies, diversified data within the Earth system have become more accessible, which offers an alternative chance to understand the Earth system[8,9]. These data are primarily used to develop data-driven predictive models[10,11], often making accurate predictions by identifying complex patterns in large datasets. Nevertheless, they sometimes tend to be associated with increased model and computational complexity and reduced transparency[12]. The fundamental goal of geosciences is to derive a concise, interpretable, and meaningful understanding of complex natural phenomena.

[1]State Key Laboratory of Water Resources Engineering and Management, Wuhan University, Wuhan, China. [2]Department of Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena, Germany. [3]ELLIS Unit Jena, Jena, Germany. [4]Image Processing Laboratory (IPL), Universitat de València, Valencia, Spain. [5]School of GeoSciences, University of Edinburgh, Edinburgh, UK. [6]CSIRO Land and Water, Black Mountain, Canberra, Australia. [7]Institute of Bio- and Geosciences: Agrosphere (IBG-3), Forschungszentrum Jülich, Jülich, Germany. e-mail: liangshs@whu.edu.cn

Box 1 | Conventional and data-driven equation discovery

Conventional equation discovery

The classical paradigm for establishing governing equations in geoscience includes first-principle approaches and (semi-)empirical approaches. They are not used in isolation but are often mixed for equation discovery.

First-principle approaches

First-principle approaches can be philosophically summarised as a cycle that begins with observations and intuition, leading to the formulation of hypotheses. These hypotheses are then subject to validation by experiments, resulting in their either acceptance or rejection. This process requires the iterative refinement of equations and, if necessary, the formulation of new hypotheses. The key step here is proposing hypotheses based on intuition, which involves specifically (1) precisely defining the system and identifying relevant variables, (2) introducing relevant assumptions and simplifications tailored to the specific problem, and (3) choosing the fundamental laws applicable to the system (e.g., thermodynamic, or hydrodynamic principles) as a basis for mathematical deductions. Deriving equations from these principles requires strict adherence to rigorous mathematical methodologies.

Empirical approaches

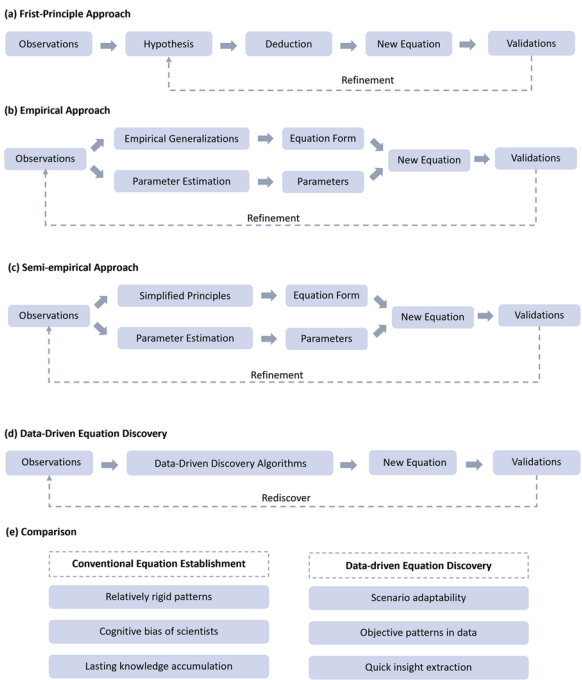
Empirical approaches involve identifying patterns within data through observation and experimentation. Typically, scientists begin by constructing mathematical equations that include several parameters based on the nature of the observed relationship (such as linear or non-linear). Estimating these parameters usually involves a statistical data analysis to find the parameters that best fit the observed patterns or trends. Once the governing equation is defined, validation is likewise required by applying the equation to a data set different from that used to estimate the parameters or by comparing the model's predictions to known outcomes. The process generally involves iteration and refinement, as well as experimenting with different equation forms to find one that adequately fits the new data set and maintains relevance.

Semi-em`pirical approaches

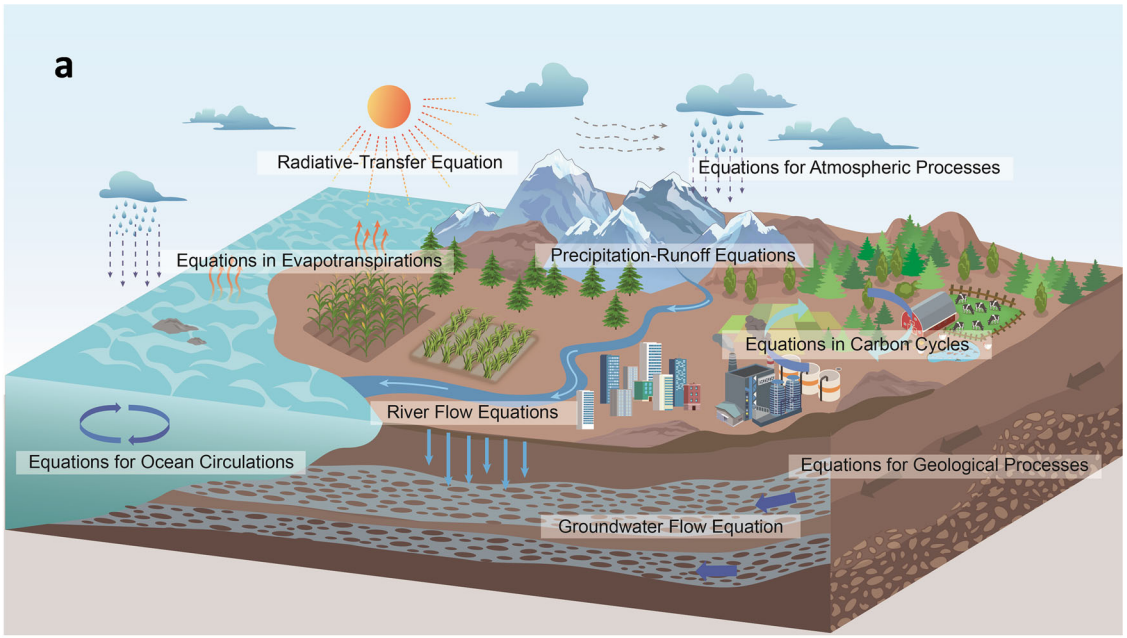
Semi-empirical approaches also referred to as phenomenological approaches, rely on theoretical assumptions about the characteristics of the system rather than being entirely empirical. It combines principles with empirical data to create models that are both scientifically grounded and practically applicable. A typical example of a semi-empirical approach is the Penman-Monteith model for evapotranspiration, which is partially based on the physical principles of conservation energy and transport processes. Of note, however, is the model's empirical treatment of the entire canopy as a single leaf, with stomatal conductance aggregating to a collective value representing the effective canopy conductance.

Data-driven discovery of governing equations

The data-driven discovery of governing equations is the simultaneous identification of the explicit equation structure and the corresponding coefficients from given observations. Given a data set $\{\mathbf{x_i}; \mathbf{y_i}\}_{i=1}^{n}$, where $\mathbf{x_i} \in \mathbf{R^d}$ is the input vector, and $\mathbf{y_i} \in \mathbf{R}$ is a scalar output, the explicit form of symbolic expression $\mathbf{y} = \mathbf{F}(\mathbf{x})$ is the expected result to be discovered. Here, $\mathbf{F}$ denotes a function class consisting of mappings, such as $\mathbf{x}$, $\mathbf{x}^2$, $\sin(\mathbf{x})$, and $\frac{\partial \mathbf{x}}{\partial \mathbf{t}}$. The discovery process can be implemented by different algorithms, as elaborated in the main text. Like the conventional paradigm, this approach requires careful validation and possible iteration based on performance metrics guided by performance metrics.



In this Perspective, we introduce and discuss the data-driven equation discovery and argue that it can integrate the power of data-driven methods and the strengths of governing equations. Figure 2 provides a comparison of these methods. Data-driven equation discovery is defined as automatically distilling the hidden patterns from data and transforming them into an interpretable and concise symbolic representation. As a result, it combines the ability of data-driven models to extract laws that conform to predictive patterns with the simplicity and transparency of equations. Data-driven equation discovery is relevant to practical geoscience applications and potentially essential for pioneering geoscientific discoveries. Specifically, it



b

	Algebraic equations	Ordinary differential equations	Partial differential equations
Hydrology	Bernoulli's law	Darcy's law	Richardson-Richards equation
Ecology	Biomass Allocation Equation	Lotka-Volterra equations	Fisher-KPP equation
Seismology	Exponential decay equation	Damped harmonic oscillator model	Wave equation
Atmosphere	Ideal gas equation of state	Lorenz system	Navier-Stokes Equations

Fig. 1 | Examples of governing equations in geosciences. a Geoscience processes in the Earth system are described by different governing equations derived from the conventional equation discovery paradigm. **b** Representative governing equations in four typical geoscience domains: hydrology, ecology, seismology, and atmospheric science. These equations are of various forms, including algebraic, ordinary, and partial differential equations.

represents an opportunity to move beyond conventional (semi-)empirically parameterised equations (e.g., numerous empirical equations in evapotranspiration modelling[13] and autotrophic respiration modelling[14]), thereby improving the modelling accuracy with transparency. It may also resolve controversies in the forms of various conventional governing equations, such as the ongoing debate over the structure of the advection-diffusion equation[15]. Moreover, the discovery and formulation of equations are naturally driven by data, which occurs spontaneously. This process not only overcomes the difficulties associated with calibrating and estimating equation parameters but also accelerates scientific discovery by improving the efficiency and effectiveness of exploration processes.

We first provide overviews of the conventional and data-driven equation discovery and discuss how the emergence of the new data-driven discovery could bring significant advantages to geoscience. We then underscore the potential challenges and envision advancing geosciences through data-driven discovery. We aim to foster a deep integration of data-driven discovery into the practice of geoscientists, contributing to more accurate, efficient, and comprehensive modelling, understanding, and management of the complex Earth system.

Conventional governing equations in geosciences

Conventionally, in geoscience, the philosophy of deriving governing equations (i.e., mathematical modelling) is based on first principles or (semi-)empirical approaches[3] (see Box 1). Scientists first postulate and conceptualise a formulation based on observations or (theoretical) experiments. This formulation is then subject to validation, refinement, and updating, driven by logical reasoning and scientific or engineering research insights. Many classical governing equations are initially derived by empirical summarization, with subsequent scientific progress revealing their derivability from first principles. Figure 1b provides example equations in different disciplines within geosciences.

However, establishing such equations requires a deep understanding of complex processes by experienced scientists. In cases where a system is not thoroughly understood, the equation-building process can be susceptible to human cognitive biases, particularly in determining which simplifications and assumptions are reasonable or in selecting the most appropriate physical principles. For instance, the formation and dissipation of clouds remain poorly understood, resulting in physics-based cloud parameterization equations that are based on incomplete knowledge and are prone to

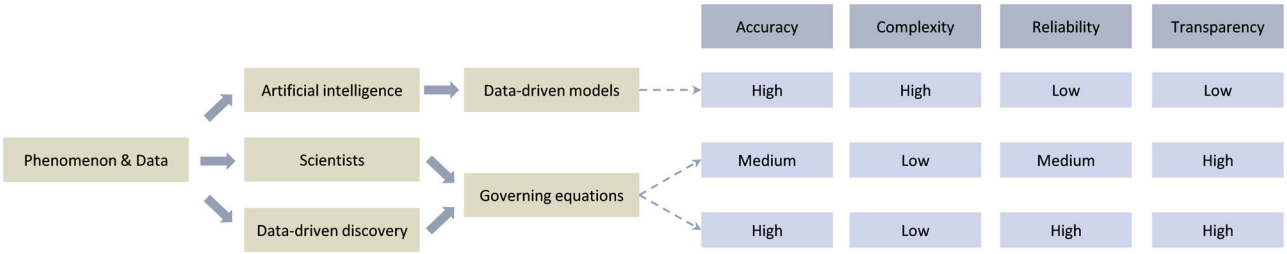


Fig. 2 | Comparison of different approaches for modelling and understanding geoscience processes. Based on phenomena and observed data, scientists summarised and proposed governing equations, the equations are accurate, reliable, transparent to a certain extent and also simple. The pure data-driven models based on artificial intelligence perform higher accuracy while lacking reliability and transparency and are too complex. The new data-driven equation discovery can integrate their merits.

inaccuracies[16], thereby weakening our ability to predict climate dynamics. Geoscientists often simplify fine-scale phenomena through parameterisation, such as first-order degradation of empirically defined carbon pools[17]. In some cases, they may also rely on empirical formulas, guided by their intuition, to capture the salient features of these processes. A typical example is evapotranspiration modelling, where empirical equations describe many physical transformations. For instance, stomatal conductance, a key intermediate variable, is commonly expressed as a product of several environmental factors[18] or through a linear relationship with the rate of photosynthesis[19]. Similarly, aerodynamic and thermal dynamics roughness lengths are estimated using various semi-empirical models[20–22]. These approaches often result in crucial yet intangible parameters that are difficult to determine in practice. Furthermore, human factors can introduce errors into the derived equations or make the equation form questionable. For instance, the derivation of equations that describe unsaturated soil moisture movement based on the Darcy-Buckingham law has long been controversial in hydrology[23]. Furthermore, determining the appropriate form of reaction-diffusion equations is a continuing debate primarily influenced by scale effects[15], which consistently limits our understanding and modelling of complex subsurface flow. Additionally, the conventional paradigm of equation discovery, which is mainly scientist-driven and reliant on intuition, often necessitates an iterative process of trial and error that may lead to a slow pace of scientific progress. In summary, despite its historical achievements, this paradigm may not effectively capture the ever-growing demands for deeper scientific understanding of the increasingly complex Earth system processes we study today.

Towards new data-driven equation discovery

Recently, the big Earth data[9] has been accessible, which is characterized by its considerable volume, diverse sources, and rapid generation (e.g., CMIP-6 data[24]). In addition, with the increasing abundance of computational resources, scientific artificial intelligence approaches have emerged[25]. There is a growing effort on the automatic discovery of governing equations directly from data[26,27]. To the best of our knowledge, the earliest study of data-driven equation discovery can be attributed to Gerwin[28], Langley[29], Falkenhainer and Michalski[30], who proposed heuristic methods to derive the mathematical functions from a large and complex space of possible formulations using informed search. Data-driven equation discovery was starting to become feasible. Subsequently, Koza demonstrated that genetic programming (GP) could discover symbolic governing equations from data[31]. During this time, GP was successfully applied in geoscience. For instance, Babovic and Keijzer[32] discovered the equation describing the additional resistance to flow induced by flexible vegetation from data.

The modern paradigm of data-driven discovery is traced to seminal work by Bongard and Lipson[33] and Schmidt and Lipson[34] through improved GP, who successfully automated the discovery of equations for dynamical systems and conservation laws from data. However, GP typically has inherent limitations, such as computational intensity, susceptibility to overfitting, and difficulties with convergence if not properly balanced[35,36]. These limitations become prohibitive for high-dimensional systems described by PDEs. However, PDEs play a critical role in simulating dynamic systems and phenomena with spatial and temporal variations, including applications in geosciences such as climate modelling and natural disaster prediction[5]. A few years later, Brunton et al.[35] introduced a new data-driven discovery framework known as sparse regression to address this challenge and reignite enthusiasm in the field[28,36]. It led to numerous subsequent works to extend to discover chaotic and complex PDEs from data, such as PDEs with parameter dependencies[37,38], which significantly expand the potential applications within geoscience. In addition, the rapid development of deep learning technologies has begun to address a long-standing high sensitivity to noise and data-hungry[39–41]. In addition, the studies on the identification of coordinates for governing equations, state variables, and implicit representations (i.e., learning an operator that encapsulates the system characteristics)[42,43], may provide another avenue for data-driven equation discovery.

Nowadays, this emerging data-driven discovery paradigm has underpinned wide-ranging applications, including biology[6,44], materials[45], and also geosciences, such as subsurface hydrology[46], ocean modeling[47–49] and climate science[16,50,51]. Despite its enormous potential and widespread attention[42], its opportunities remain underappreciated in the geoscience community, primarily because of the disconnect between the advances and the challenges and needs of geosciences. Therefore, we introduce the data-driven equation discovery in detail and discuss leveraging it to benefit geosciences in the following sections.

How to realise data-driven equation discovery in geosciences

The overview of the data-driven equation discovery workflow in practice is shown in Fig. 3, and an example is given in Box 2. An important part of data-driven equation discovery is to select proper approaches, whose objective is to employ reasonable strategies to reduce search space effectively, as brute force search is considered non-deterministic polynomial-hard (NP-hard)[52], which means that solving it quickly becomes impractical as the size of the problem grows. The equation discovery from data differs from traditional inverse modelling and black-box system identification. The latter aims to estimate the parameters or coefficients from data[53], where the equation structure is usually partly given.

Figure 4 provides an overview of methods for realising data-driven equation discovery in geosciences. A detailed description of approaches is given in Supplementary Note 1. We divide data-driven equation discovery approaches into two primary categories: symbolic regression and sparse selection algorithms, based on whether the algorithms can generate an infinite variety of equation forms. Symbolic regression utilises various search methods to generate infinite combinations of symbolic formulae, mainly including genetic programming[33,34], heuristic symbolic regression[28,54], mixed-integer nonlinear programming approaches[55,56], deep reinforcement learning[39,57], and large-scale pre-trained Transformers[58,59]. They only require data on the variables of interest, including preprocessed data (e.g., derivatives). The ability of symbolic regression makes it well suited to uncovering complex governing equations that describe the underlying symbolic relationships between multiple variables in geoscience. For

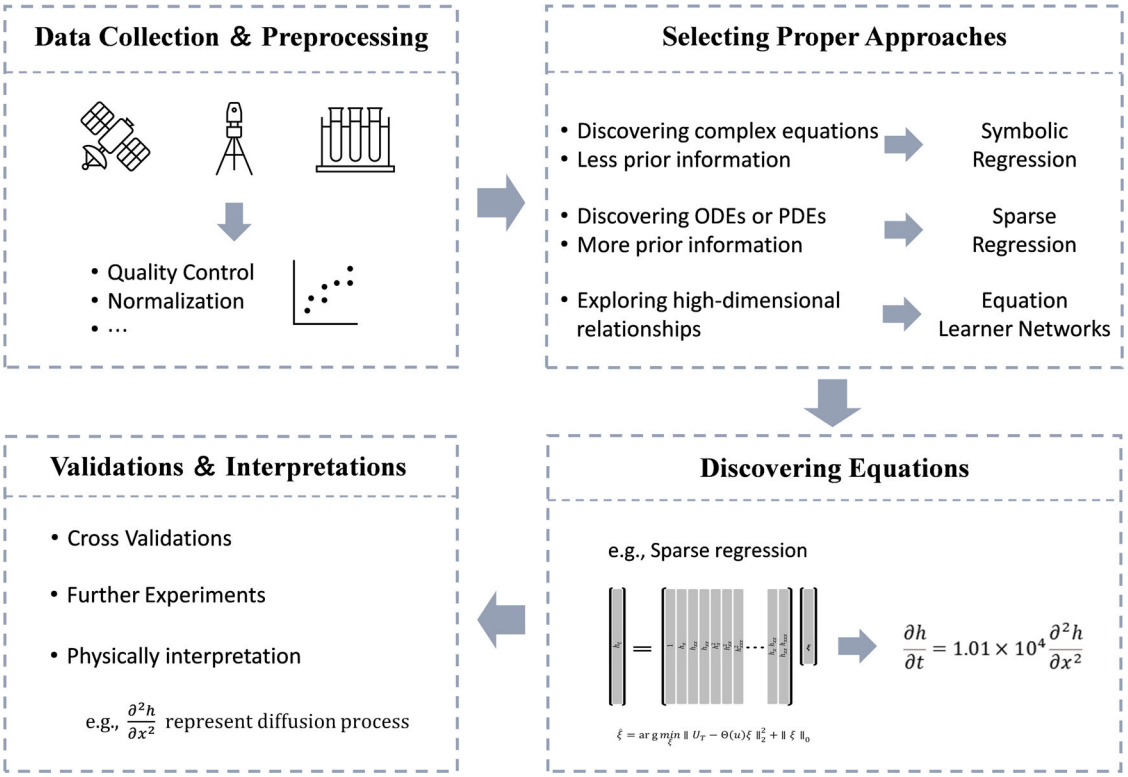


Fig. 3 | The overview of the data-driven equation discovery workflow in practice. The first step is data collection and preprocessing, then selecting proper approaches based on specific tasks. A detailed description of different approaches can be found in Supplementary Note 1. Based on the selected algorithms, the governing equations can be discovered. Finally, the discovered equations should be validated and physically interpreted.

instance, it can be employed to explore intricate governing equation relationships between evapotranspiration flux and various meteorological parameters and vegetation variables using large amounts of data, where the exact physical mechanism is still unclear. In contrast, sparse selection algorithms, including sparse regression[35,36,60] and equation learner networks (EQL)[61,62] aim to select the most appropriate equation from a predefined pool of symbolic combinations. They are efficient for systems with a solid understanding of the underlying functional form. Sparse regression and EQL have their own application scopes. Sparse regression is widely used to discover the underlying PDEs because there are often discernible patterns in the modelling of PDEs. EQL networks can seamlessly interface with high-dimensional data, such as satellite imagery, enabling end-to-end learning processes and exploring these hidden mechanisms behind high-dimensional data.

The benchmarks of accuracy, speed, and tolerance to data noise for symbolic regression methods[63–66] and sparse regression[67] have been performed. It has shown that sparse regression has a low computational cost and few hyperparameters. In contrast, symbolic regression approaches provide an opportunity to discover underlying equations with complex structures, while the main limitation is computational cost. Deep learning-based approaches are more robust to noisy and sparse data such as deep reinforcement learning. They are therefore recommended to deal with geoscientific applications where flawed datasets are common[43]. In terms of required prior information, sparse regression needs more domain expert input and assumptions, such as the general form of the underlying governing equations. Therefore, sparse regression could fail if incorrect or incomplete prior information is introduced, i.e., the candidate library matrix in sparse regression[36]. In contrast, symbolic regression can realise learning from scratch, while we can incorporate some physical information in different ways to discover the underlying equations accurately.

Nowadays, most of the proposed algorithms have open-source code available. For example, PySR and SymbolicRegression.jl[68], implemented in the Python and Julia languages, encapsulate symbolic regression methods. PySindy[69], developed in Python, can be used for sparse regression. These tools can significantly lower the technical barriers to implementing advanced and complex algorithms, paving the way for geoscientists to engage in data-driven equation discovery. It is worth noting that this field is rapidly evolving, so close attention is necessary to obtain algorithms with superior performance, especially considering accuracy, speed, and robustness.

Opportunities for advancing geosciences

Data-driven equation discovery provides promising opportunities for advancing geosciences. Figure 5 summarizes these aspects, and the detailed descriptions are as follows.

Enhancing classical governing equations

The conventional derivation of equations inevitably involves a degree of empiricism, such as selecting and defining variables, conditional assumptions, and simplifications. The new paradigm offers an alternative to these locally empirical methods and promotes improved subsequent derivation, leading to better structure governing equations. For example, it allows the exploration of improved forms of water retention curve equations in subsurface hydrology[70] or the study of moisture sensitivity of soil heterotrophic respiration[71].

Replacing black-box models with explicit expressions

Due to long-standing challenges in parameter calibration and estimation and precision issues, many geoscience equations are being replaced by black-box models, such as those based on machine learning. Through the data-driven discovery paradigm, it is possible to derive equation models that maintain consistent performance and offer greater interpretability and physical relevance. For instance, in hydrology, this approach allows for deriving hydro-pedotransfer functions with precise and explicit forms[72]. Additionally, these explicit governing equations may facilitate a more

Box 2 | Example workflow of data-driven equation discovery

Goal

This example shows the discovery of the governing PDE of one-dimensional groundwater flow without source or sink terms in a saturated, homogeneous, confined aquifer to illustrate the general and practical workflow of discovering governing equations from data that can be widely applied in geoscience. Here, $\frac{\partial \mathbf{h}}{\partial \mathbf{t}} = \mathbf{S_s}^{-1}\mathbf{K}\frac{\partial^2 \mathbf{h}}{\partial \mathbf{x}^2}$ it is assumed to be the true governing PDE, where $\mathbf{S_s} = 10^{-2}\mathbf{m}^{-1}$ is the specific storage, and $\mathbf{K} = 0.01\mathbf{m}\mathbf{d}^{-1}$ is the hydraulic conductivity. The goal is to rediscover this known equation from observational data.

Data collection and preprocess

As the first step, discovering PDEs requires spatial and temporal state variable data (in this case, hydraulic head), typically obtained through laboratory experiments or field monitoring. Generally, data is recommended to be collected as densely as possible in space and time to accurately capture the system behaviour, depending on practical sensor capabilities and cost. The requirement for the length of the time series depends on whether short-term or long-term system behaviour is of interest. On the other hand, the number of samples can be manageable, and usually, an order of magnitude of hundreds to thousands can be sufficient. In this specific example, the data are collected at 100 points per 1**m**, and the data monitoring starts on day 0 and ends on day 100, with records taken at 0.1**d** intervals. The data should be preprocessed, including quality control and normalisation. For example, when dealing with a small number of missing values, it is recommended to utilise spline or DNNs to fit all the collected data to reduce the influence of the data gap. Normalisation can be used to reduce the impact of different scales of potential equation terms.

Methods selection

Then, appropriate data-driven discovery approaches must be selected based on the objective and tailored to each case's specifics. For PDE discovery tasks, sparse regression is considered here. The detailed selection criteria are given in the main text.

Equation discovery

When implementing the sparse regression, the first step is to build a comprehensive candidate library. The purpose of creating such an overcomplete library is to ensure that it includes all possible terms that could accurately represent the dynamics described by the equation. This extensive collection might consist of but is not limited to, various polynomial terms, trigonometric functions, exponential functions, and their derivatives. For example, derivatives up to the third order and their second-order combinations can be considered as candidate terms, i.e., $\Theta = [\mathbf{h}, \frac{\partial \mathbf{h}}{\partial \mathbf{x}}, \frac{\partial^2 \mathbf{h}}{\partial \mathbf{x}^2}, \frac{\partial^3 \mathbf{h}}{\partial \mathbf{x}^3}, \mathbf{h}^2, \mathbf{h}\frac{\partial \mathbf{h}}{\partial \mathbf{x}}, \mathbf{h}\frac{\partial^2 \mathbf{h}}{\partial \mathbf{x}^2}, \mathbf{h}\frac{\partial^3 \mathbf{h}}{\partial \mathbf{x}^3}, (\frac{\partial \mathbf{h}}{\partial \mathbf{x}})^2, \frac{\partial \mathbf{h}}{\partial \mathbf{x}}\frac{\partial^2 \mathbf{h}}{\partial \mathbf{x}^2}, \frac{\partial \mathbf{h}}{\partial \mathbf{x}}\frac{\partial^3 \mathbf{h}}{\partial \mathbf{x}^3}, (\frac{\partial^2 \mathbf{h}}{\partial \mathbf{x}^2})^2, \frac{\partial^2 \mathbf{h}}{\partial \mathbf{x}^2}\frac{\partial^3 \mathbf{h}}{\partial \mathbf{x}^3}, (\frac{\partial^3 \mathbf{h}}{\partial \mathbf{x}^3})^2]$.

Subsequently, the derivatives in the candidate library must be estimated from discrete data. Several methods are available to accomplish this goal. For example, the finite difference method can be applied to approximate these derivatives from discrete data points such that the overdetermined system $\mathbf{U_t} = \Theta(\mathbf{U})\Xi$ can be constructed. Here Ξ is unknown and can be solved by various sparsity-promoting regression techniques, e.g., STRidge[28]. Finally, the discovered PDE, in this case, is $\frac{\partial \mathbf{h}}{\partial \mathbf{t}} = 0.99\frac{\partial^2 \mathbf{h}}{\partial \mathbf{x}^2}$, which is close to the true governing equation.

Validation and interpretation

It is crucial to conduct a rigorous validation of the discovered PDE, such as cross-validation with independent data, to assess its robustness and generalizability. For example, this PDE could be applied to predict water flow in different aquifers. Moreover, it is important to ensure that the equation is not only mathematically sound, physically interpretable, and consistent with established geoscientific principles for its meaningful application.

straightforward assessment of potential underlying biases learned from the data, offering an advantage over the opacity of black-box models. For instance, recently, it showed that data-driven equation discovery can learn new physics for the atmosphere and replace costly modules in cloud parameterizations[16,47].

Improving traditional controversial governing equations

When a system is not entirely understood, the derivation process may be prone to cognitive biases. These biases can introduce errors in the equations or lead to significant controversy in their formulation. The data-driven paradigm can address such controversies. A pertinent example is using fractional-order equations in Earth systems characterised by scale or memory effects[73], whose rationality could be clearer and often sparks debate. Applying the new approach could provide clarity and resolve these ongoing controversies.

Uncovering missing equations

The new paradigm is adept at uncovering previously unrecognised variables or processes, particularly in data-rich scenarios. Integrating interdisciplinary data across the geosciences can reveal complex interactions that may remain elusive when individual disciplines are considered separately. For example, climate scientists can use these newly discovered equations to refine climate models, deepening our understanding and improving climate change predictions. An explicit expression for the concentration-flow relationship (C-Q) may be found in water quality science. Similarly, it may be possible to establish equations linking vegetation structure to radar backscatter in satellite biomass mapping[74,75].

Accelerating scientific discoveries and high-quality data collection

The real-time nature of the data-driven discovery approach bypasses the need for slow, theory-based development from first principles, potentially accelerating the pace of scientific discovery in geoscience. Moreover, the increasing emphasis on data-driven methods in this field may promote advancements in data collection technologies, leading to the acquisition of more diverse and high-quality datasets.

Challenges and potential solutions

Despite the promise, several challenges must be addressed before fully realising the benefits of data-driven discovery for geosciences. From our perspective, these challenges encompass three main aspects: data, geoscience processes, and validations, which are briefly summarised in Table 1. These challenges do not diminish the potential of the new paradigm but rather represent opportunities for collaboration between geoscientists and data scientists to promote artificial intelligence as a truly powerful tool to advance geoscience.

Data perspective

(1) Discovering governing equations from sparse and noisy geoscientific data: While data is becoming increasingly abundant, there are still instances in many geoscientific domains where accurate and extensive datasets still need to be improved. This data sparsity is characterised by temporal and spatial coverage inconsistencies, which will persist as a long-term feature despite the potential for richer datasets in the future[76]. Additionally, they are frequently corrupted by noise, involving

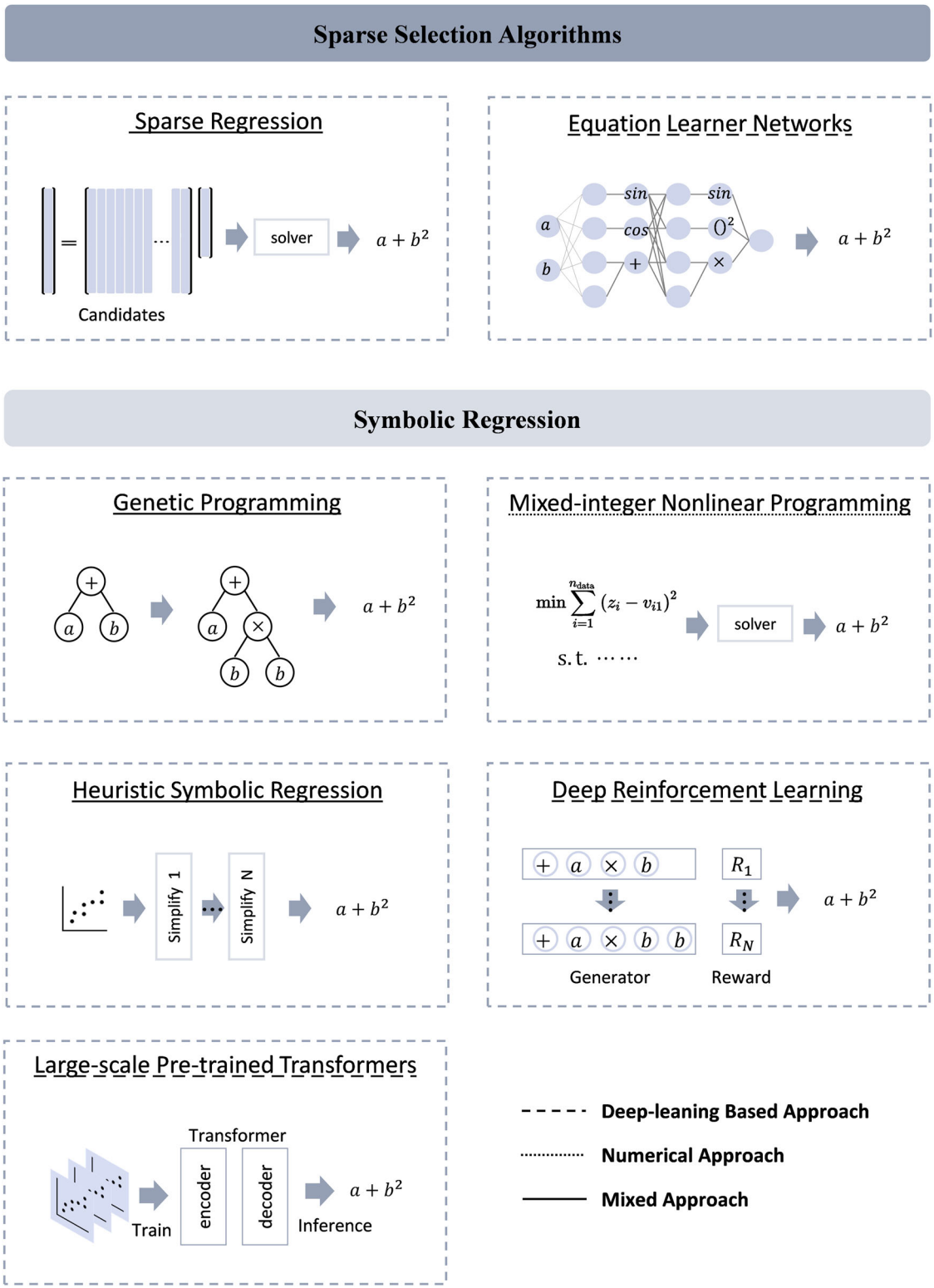


Fig. 4 | Data-driven equation discovery approaches in geosciences. The approaches can be divided into two main categories: symbolic regression and sparse selection algorithms. It is noted that some methods inherently utilise deep learning techniques as part of their core algorithms. At the same time, some can be structured to be compatible with deep learning, allowing for integration that enhances their capabilities. A detailed description of each approach is given in Supplementary Note 1.

diverse noise sources, uncertainties, data missing, and gaps. Data-driven discovery is required to deal with sparse and noisy data[29,40]. Generally, direct observations often involve state variables such as temperature, pressure, and concentration. However, equation discovery tasks typically rely on derivatives of these variables with respect to space and time to capture dynamic changes. Obtaining these derivatives involves numerical approximations that can introduce errors affecting the accuracy of the discovered equations[77]. Conventional finite difference methods[78] would quickly deteriorate when dealing with sparse and noisy data. Fortunately, several approaches

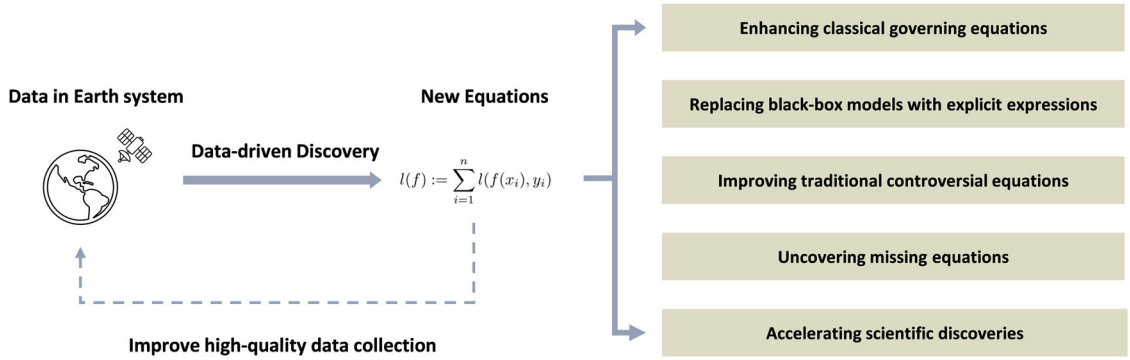


Fig. 5 | The overview of opportunities for using data-driven equation discovery to advance geosciences. Based on various datasets collected in the Earth system, data-driven discovery is expected to discover new equations, thereby enhancing existing equations, model transparency, and finally accelerating scientific discoveries. In turn, it will also facilitate the collection of higher-quality data.

have been developed to discover equations from such data, such as smoothing methods[79,80], weak-form formulas[81], targeted denoising methods[82,83], and deep neural networks[40,55]. The selection depends on the specific types of geoscience processes and should be carefully chosen[84,85]. For instance, the low-rank property of physical system dynamics can be utilised to preprocess large-scale observational datasets[86]. Continued efforts are needed to deal with datasets characterised by extreme noise and significant sparsity.

(2) Distilling underlying equations from high-dimensional big Earth data: Recently, certain geoscience domains have experienced notable shifts in data access, mainly attributed to the proliferation of satellite-based and in-situ sensors[10,76] (e.g., International Soil Moisture Network[87] provides a large amount of in-situ soil moisture measurements). Harnessing these vast and varied data and extracting meaningful insights has proven to be a challenge[76,88]. Data-driven equation discovery has provided opportunities to make sense of this data deluge. For example, it has been demonstrated that EQL networks can be seamlessly integrated with high-dimensional data to explore hidden mechanisms and discover governing equations[89]. When dealing with these complex in-suite data, one limitation is that these data need an effective coordinate system. Data-driven equation discovery approaches may only succeed with proper coordinates. For example, when dealing with irregular measurements (e.g., temperature at different locations) in a complicated geometry, coordinate transformations are inevitable to obtain equations for temperature dynamics. Fortunately, systematic and automated discovery of the latent coordinate representation has been realised, such as deep autoencoder networks[90–94]. It is possible to discover proper coordinates and equations from unorganized measurement data. Another limitation is the need for considerable computational resources when dealing with large datasets[40]. The amount of data that can be ingested and utilised positively correlates with available computing resources. Recently, some emerging efficient computing methods, including parallel processing, distributed computing, and dedicated hardware such as GPUs, have shown promise in solving this challenge[12].

(3) Leveraging imbalance data to find governing equations: It is an obvious feature that some parts of the Earth system have more available data than others; for example, above-ground data are richer than below-ground[76]. Imbalanced data can lead to potential implicit biases in data-driven models, thus affecting the discovered equations. For instance, the accuracy of wet and dry end coefficients is reduced when the soil water flow equation is derived from datasets with fewer observations of extreme wet and dry scenarios[95]. To minimise the impact of data imbalance, data preprocessing is straightforward but effective, such as controlling the data distribution or augmenting the data with deep learning methods to make the data richer and more balanced. However, such simple preprocessing may only be feasible for univariate governing equation discovery. However, using unbalanced multivariate data to find governing equations for multiple processes still needs further research. One possible strategy is to integrate multi-fidelity deep learning[96] and generative deep learning[97] with equation discovery tasks. Multi-fidelity deep learning can integrate data from multiple sources to improve the accuracy of discovered equations. Generative deep learning can create synthetic data that enhances the dataset, enabling more accurate and robust identification of the underlying equations governing the system. In summary, biases hidden in unbalanced datasets should be treated with caution and the equations found need to be carefully validated to ensure reasonable results.

Table 1 | Overview of challenges for data-driven equation discovery in geosciences

Perspective	Challenges
Data	Discovering equations from sparse and noisy data
	Distilling equations from high-dimensional big Earth data
	Dealing with imbalanced datasets
Geoscience	Discovering equations of nonlinear processes with parameter dependencies
	Extracting equations across multiple spatial and temporal scales
	Identifying multivariable equations of multiple connected processes
	Dealing and quantifying with uncertainties
Validation	Physical interpretation and comprehensive validations

Geoscience perspective

Complex geophysical, -chemical, and -biological processes are common in geoscience and play a crucial role in shaping the Earth's surface, climate, and geological features. These processes involve multiple interacting components, can occur on transient to extended time scales, and usually span multiple spatial scales. Data-driven equation discovery provides effective ways to describe these interactions but also faces several challenges, as listed below. These challenges are often interrelated rather than isolated in real-world scenarios, necessitating a holistic consideration.

(1) Equation discovery for nonlinear processes with parameter dependencies: Many geoscience processes exhibit nonlinear behaviour and

can vary significantly in space and time, and the heterogeneity can introduce parametric variability into the underlying governing equations. Several challenges remain to overcome. For instance, sparse regression can discover nonlinear PDEs when nonlinear terms are included in the candidates, but this can be difficult when dealing with a new system[98]. In addition, symbolic regression may easily converge prematurely when searching for complex nonlinear equations, which is inefficient and impractically slow. A worthwhile step could be integrating some geophysical information, such as symmetry and dimensional analysis, as it can speed up the search process. On the other hand, due to the coupled effects of parametric dependencies and equation structure on geoscientific dynamics, it is hard to separate and identify them. Group sparse regression[38,99,100] and the kernel approach[101] have been applied to resolve it. Nevertheless, it is still intractable when dealing with highly nonlinear and complex coefficient fields, common in various geoscientific domains. For example, hydraulic conductivity, one of the parameters in the groundwater flow equation, can vary in magnitude by several orders of magnitude on microscopic spatial scales. Furthermore, key vegetation parameters in global carbon cycle models also vary spatially, mainly as a function of biodiversity[102]. In addition, current methods rely primarily on assumptions such as smoothness and symmetry, which are absent in some systems. Therefore, further approach development is still needed[101].

(2) From data to governing equations for multiscale interactions: Geoscience processes can span multiple spatial and temporal scales. While micro-scale interactions may be well described by governing equations derived from first principles, the macroscopic behaviour may sometimes fail to follow directly. For example, in the groundwater flow, the assumptions of homogeneity and continuity are typically based on the representative elementary volume scale, a virtual volume that may not hold at larger or smaller scales[103]. Despite these modelling difficulties, observational data is much more accessible for some macroscopic interactions. A notable direction is combining the data-driven discovery with microscopically simulated data[46,104,105]. Since microsimulation does not assume any macroscopic governing equations a priori, it can be a valuable approach to verify already derived governing equations[106] and to reveal yet unknown macroscopic governing equations[104]. For instance, macroscale PDE for proppant transport in subsurface geoscience has been successfully discovered[104]. Experimental data can also be explored: an example is the quantitatively accurate equation for weakly turbulent fluid flow, albeit in a complicated and high-dimensional nonequilibrium system, discovered from velocity field measurements[107]. These pioneering works have demonstrated the potential of discovering multiscale interactions governing equations from data.

(3) Identifying equations with multiple connected processes: Geoscience processes often involve numerous interacting factors and require multiple and multivariable governing equations to describe them. Data-driven equation discovery can potentially find multiple interrelated process equations, which may lead to insights into cross-temporal and cross-scale linkages. For instance, the study of terrestrial ecosystem dynamics can significantly benefit from holistically identifying the equations of multiple interacting factors such as vegetation succession and competition, root zone water transport, plant allocation to leaves, stems, and roots, and impacts of fire on vegetation states and atmospheric emissions[108]. An essential prerequisite for identifying equations with multiple interrelated processes is the definition of appropriate variables. In a high-dimensional system, the relevant set of state variables is typically unknown, and identifying them is generally a laborious task that demands considerable scientific effort. Defining compact and complete variables is essential for discovering parsimonious governing equations. The automatic identification of interpretable and physically consistent state variables remains a challenging and intractable problem. Methods such as geometric manifold learning, a machine learning approach for dimensionality reduction by uncovering the intrinsic structure or geometry of high-dimensional data, have been devised to automate the discovery of fundamental variables hidden in time-series data[109] or high-dimensional data (e.g., video data)[110]. These advances in data-driven discovery methods hold promise for addressing the bewildering variety of information that confronted early scientists[109–111].

(4) Extracting equations for geoscience processes with uncertainty and identifiability: In various processes and phenomena within the Earth system, there are inherent, unavoidable factors of uncertainty. This uncertainty can originate from multiple aspects, for example, natural variability can introduce stochastic elements. It is crucial to discover equations to capture the uncertainty and extrapolate better in different uncertain scenarios[112]. Discovering stochastic equations from data in the presence of such uncertainty is a complex task. However, approaches such as variational Bayesian inference make it feasible to learn stochastic governing equations and quantify uncertainties directly from data[113–116]. In addition, adopting techniques such as sensitivity analysis[117] and ensemble methods[118] is promising to address the uncertainties in the equation discovery tasks. Moreover, many geoscience processes often occur as nonequilibrium, such as transient behaviour and critical thresholds or abrupt transitions, making it challenging to identify equations that account for sudden changes in behaviour. The task of inferring non-stationary dynamics from stochastic observations, explored in recent studies[119], is a critical step in this direction.

Validation perspective

Generally, the formulation of governing equations should follow Occam's Razor, balancing parsimony with accuracy[120]. However, assessing the complexity of the underlying equation before its discovery and validation remains challenging. Pareto frontier analysis[34] is recommended to address this, which involves using a series of progressively complex formulas to improve accuracy incrementally. For instance, independent validation for sets of proposed governing equations for the carbon cycle has allowed the determination of their optimal complexity given the information content of the calibration data[121]. Moreover, information criteria have been used to select the best equations that balance model parsimony and predictive power, such as the Akaike information criterion[122], the Bayesian information criterion[123] and the Bayesian machine scientist[113]. However, these information criteria, which are often derived under the assumption that the likelihood function is based on Gaussian errors, may not work well when dealing with non-Gaussian noise, which is common in geoscience, as many complex natural processes are not well-described by simple Gaussian distributions. In addition, automated interpretation of newly revealed governing equations is generally limited and still requires careful validation by geoscientific domain experts to ensure that the equations align with established principles and theories[55]. In practice, it is helpful to incorporate known physical constraints[124] into data-driven discovery approaches or to leverage prior knowledge to guide discovery approaches. Models must obey built-in conservation laws or certain symmetries for the discovered equations to be consistent with established principles. It is worth noting that the selection of constraints should be reasonable, as it might also introduce biases. Furthermore, data-driven equation discovery has been preliminarily shown to understand hidden functional relationships and generalize them from observations to unknown parameter spaces[62]. It initially indicates that it is a powerful tool to help us model complex geoscience processes, but further validation is needed in the future.

Summary and future perspectives

Geoscience communities are confronted with increasingly intricate scientific questions, prompting the exploration of more advanced methods to resolve these challenges better. In this Perspective, our contributions are introducing the data-driven equation discovery to meet the unique needs of modern geosciences. Through the detailed discussions about the potential

opportunities, we advocate that the new data-driven discovery is helpful in modelling and understanding numerous processes within the Earth system, especially those with potentially complicated mechanisms and available observational datasets. It is highly relevant to a wide range of geoscientists in their everyday research routines, aligning with the diverse research needs across the field. We argue that although the discovered equations are not necessarily meant to be causal, they frequently serve the purpose of creating a highly detailed testbed for the study of feedback. This is the first time we have objective measures of (semi)parametric model evaluation, intercomparison and selection learned from data.

We advocate that this emerging field provides opportunities for interdisciplinary collaboration, enabling the cooperative development of more advanced and adapted methods for geoscience, as they cannot be solved by either geoscientists or data scientists alone. Developing these interdisciplinary approaches and using interdisciplinary data in geosciences can reveal scientific insights that would be difficult to discover if individual disciplines were studied in isolation but are easy and feasible for data-driven equation discovery.

Furthermore, it is important to note that, like most data-driven methodologies, the selection of datasets can introduce bias, which can subsequently impact the final equations generated. To mitigate this, techniques such as cross-validation should be employed to minimise the potential for errors. Moreover, the equations obtained must be interpreted in a logical and rational manner to guarantee their scientific validity and coherence. In conclusion, we highlight that data-driven equation discovery should be employed for scientific discovery in a comprehensive and responsible manner.

We believe that data-driven equation discovery is expected to consistently facilitate our comprehension of geoscience processes and even reshape the foundations of geosciences. The discovered insights have the potential to challenge existing geoscientific theories and models. While the initial response from the community may lean towards scepticism or resistance, sustained scientific validation over time could lead these innovative insights to redefine fundamental concepts in geosciences. In the past few years, we have witnessed artificial intelligence's remarkable and rapid success in applied geoscience endeavours. We anticipate that data-driven methods will soon offer similarly significant contributions to our scientific understanding by aiding in the discovery of governing equations, which have seen slow progress in the field over the past decades.

Data availability

The data of the case study shown in Box 2 can be found at https://doi.org/10.5281/zenodo.13735843.

Code availability

The case study shown in Box 2 can be reproduced with Python code provided at https://doi.org/10.5281/zenodo.13735843.

Received: 29 June 2024; Accepted: 2 October 2024;
Published online: 14 October 2024

References

1. Gershenfeld, N. A. *The Nature of Mathematical Modeling*. (Cambridge university press, 1999).
2. Willcox, K. E., Ghattas, O. & Heimbach, P. The imperative of physics-based modeling and inverse theory in computational science. *Nat. Comput. Sci.* **1**, 166–168 (2021).
3. Spencer, H. *First Principles*. vol. 1 (JA Hill, 1904).
4. Scholkopf, B. et al. Toward Causal Representation Learning. *Proc. IEEE* **109**, 612–634 (2021).
5. Bokulich, A. & Oreskes, N. Models in Geosciences. *Springer Handb*. 891–911 https://doi.org/10.1007/978-3-319-30526-4_41 (2017).
6. Maddu, S., Cheeseman, B. L., Müller, C. L. & Sbalzarini, I. F. Learning physically consistent differential equation models from data using group sparsity. *Phys. Rev. E* **103**, 1–13 (2021).
7. Karpatne, A. et al. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Trans. Knowl. Data Eng.* **29**, 2318–2331 (2017).
8. Bzdok, D., Nichols, T. E. & Smith, S. M. Towards algorithmic analytics for large-scale datasets. *Nat. Mach. Intell.* **1**, 296–306 (2019).
9. Vance, T. C., Huang, T. & Butler, K. A. Big data in Earth science: Emerging practice and promise. *Science* **383**, eadh9607 (2024).
10. Bergen, K. J., Johnson, P. A., De Hoop, M. V. & Beroza, G. C. Machine learning for data-driven discovery in solid Earth geoscience. *Science* **363**, eaau0323 (2019).
11. Reichstein, M. et al. Deep learning and process understanding for data-driven Earth system science. *Nature* **566**, 195–204 (2019).
12. Karniadakis, G. E. et al. Physics-informed machine learning. *Nat. Rev. Phys.* **3**, 422–440 (2021).
13. Poyen, E. F. B., Ghosh, A. K. & PalashKundu, P. Review on different evapotranspiration empirical equations. *Int. J. Adv. Eng. Manag. Sci.* **2**, 239382 (2016).
14. Thomas, R. Q. et al. Alternate Trait-Based Leaf Respiration Schemes Evaluated at Ecosystem-Scale Through Carbon Optimization Modeling and Canopy Property Data. *J. Adv. Model. Earth Syst.* **11**, 4629–4644 (2019).
15. Sun, L., Qiu, H., Wu, C., Niu, J. & Hu, B. X. A review of applications of fractional advection–dispersion equations for anomalous solute transport in surface and subsurface water. *Wiley Interdiscip. Rev. Water* **7**, e1448 (2020).
16. Grundner, A., Beucler, T., Gentine, P. & Eyring, V. Data-Driven Equation Discovery of a Cloud Cover Parameterization. *J. Adv. Model. Earth Syst.* **16**, e2023MS003763 (2024).
17. Luo, Y., Keenan, T. F. & Smith, M. Predictability of the terrestrial carbon cycle. *Glob. Change Biol.* **21**, 1737–1751 (2015).
18. Jarvis, P. The interpretation of the variations in leaf water potential and stomatal conductance found in canopies in the field. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **273**, 593–610 (1976).
19. Ball, J. T. An analysis of stomatal conductance. (1988).
20. Su, Z., Schmugge, T., Kustas, W. P. & Massman, W. J. An evaluation of two models for estimation of the roughness height for heat transfer between the land surface and the atmosphere. *J. Appl. Meteorol. Climatol.* **40**, 1933–1951 (2001).
21. Gokmen, M. et al. Integration of soil moisture in SEBS for improving evapotranspiration estimation under water stress conditions. *Remote Sens. Environ.* **121**, 261–274 (2012).
22. Raupach, M. Drag and drag partition on rough surfaces. *Bound.-Layer Meteorol.* **60**, 375–395 (1992).
23. Narasimhan, T. N. Something to think about….Darcy-Buckingham Law. *Groundwater* **99**, 5–6 (1997).
24. Stockhause, M. & Lautenschlager, M. CMIP6 Data Citation of Evolving Data. *Data Sci. J.* **16**, 30 (2017).
25. Wang, H. et al. Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
26. Fortunato, S. et al. Science of science. *Science* **359**, eaao0185 (2018).
27. Waltz, D. & Buchanan, B. G. Automating Science. *Science* **324**, 43–44 (2009).
28. Gerwin, D. Information processing, data inferences, and scientific generalization. *Behav. Sci.* **19**, 314–325 (1974).
29. Langley, P. Data-driven discovery of physical laws. *Cogn. Sci.* **5**, 31–54 (1981).
30. Falkenhainer, B. C. & Michalski, R. S. Integrating quantitative and qualitative discovery: the ABACUS system. *Mach. Learn.* **1**, 367–401 (1986).
31. Koza, J. R. *Genetic Programming: A Paradigm for Genetically Breeding Populations of Computer Programs to Solve Problems*. 34 (Stanford University, Department of Computer Science Stanford, CA, 1990).

32. Babovic, V. & Keijzer, M. Genetic programming as a model induction engine. *J. Hydroinformatics* **2**, 35–60 (2000).
33. Bongard, J. & Lipson, H. Automated reverse engineering of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA.* **104**, 9943–9948 (2007).
34. Schmidt, M. & Lipson, H. Distilling free-form natural laws from experimental data. *Science* **324**, 81–85 (2009).
35. Brunton, S. L., Proctor, J. L., Kutz, J. N. & Bialek, W. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA.* **113**, 3932–3937 (2016).
36. Rudy, S. H., Brunton, S. L., Proctor, J. L. & Kutz, J. N. Data-driven discovery of partial differential equations. *Sci. Adv.* **3**, 1–7 (2017).
37. Schaeffer, H., Tran, G. & Ward, R. *Learning Dynamical Systems and Bifurcation via Group Sparsity.* **1**, 16 (2017).
38. Rudy, S., Alla, A., Brunton, S. L. & Kutz, J. N. Data-driven identification of parametric partial differential equations. *SIAM J. Appl. Dyn. Syst.* **18**, 643–660 (2019).
39. Petersen, B. K. et al. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. *ICLR* (2019).
40. Chen, Z., Liu, Y. & Sun, H. Physics-informed learning of governing equations from scarce data. *Nat. Commun.* **12**, 1–13 (2021).
41. Both, G. J., Choudhury, S., Sens, P. & Kusters, R. DeepMoD: Deep learning for model discovery in noisy data. *J. Comput. Phys.* **428**, 109985 (2021).
42. Camps-Valls, G. et al. Discovering Causal Relations and Equations from Data. *Phys. Rep.* **1044**, 1–68 (2023).
43. Brunton, S. L. & Kutz, J. N. Promising directions of machine learning for partial differential equations. *Nat. Comput. Sci*. https://doi.org/10.1038/s43588-024-00643-2 (2024).
44. Lagergren, J. H., Nardini, J. T., Michael Lavigne, G., Rutter, E. M. & Flores, K. B. Learning partial differential equations for biological transport models from noisy spatio-temporal data. *Proc. R. Soc. Math. Phys. Eng. Sci*. **476**, 20190800 (2020).
45. Brunton, S. L. & Nathan Kutz, J. Methods for data-driven multiscale model discovery for materials. *JPhys Mater*. **2**, 044002 (2019).
46. Zeng, J., Xu, H., Chen, Y. & Zhang, D. Deep learning discovery of macroscopic governing equations for viscous gravity currents from microscopic simulation data. *Comput. Geosci*. https://doi.org/10.1007/s10596-023-10244-z (2023).
47. Ross, A., Li, Z., Perezhogin, P., Fernandez-Granda, C. & Zanna, L. Benchmarking of Machine Learning Ocean Subgrid Parameterizations in an Idealized Model. *J. Adv. Model. Earth Syst.* **15**, e2022MS003258 (2023).
48. Zanna, L. & Bolton, T. Data-Driven Equation Discovery of Ocean Mesoscale Closures. *Geophys. Res. Lett.* **47**, e2020GL088376 (2020).
49. Perezhogin, P., Zhang, C., Adcroft, A., Fernandez-Granda, C. & Zanna, L. Implementation of a data-driven equation-discovery mesoscale parameterization into an ocean model. Preprint at http://arxiv.org/abs/2311.02517 (2023).
50. Xu, H. et al. Interpretable AI-Driven Discovery of Terrain-Precipitation Relationships for Enhanced Climate Insights. *arXiv*. https://doi.org/10.48550/arXiv.2309.15400 (2023).
51. Jakhar, K., Guan, Y., Mojgani, R., Chattopadhyay, A. & Hassanzadeh, P. Learning Closed-Form Equations for Subgrid-Scale Closures From High-Fidelity Data: Promises and Challenges. *J. Adv. Model. Earth Syst.* **16**, e2023MS003874 (2024).
52. Virgolin, M. & Pissis, S. P. Symbolic Regression is NP-hard. *TMLR* **1**, 1–11 (2022).
53. Nakamura, G. & Potthast, R. *Inverse Modeling*. (IOP Publishing, 2015).
54. Udrescu, S. M. & Tegmark, M. AI Feynman: A physics-inspired method for symbolic regression. *Sci. Adv*. **6**, eaay2631 (2020).
55. Cornelio, C. et al. Combining data and theory for derivable scientific discovery with AI-Descartes. *Nat. Commun.* **14**, 1777 (2023).
56. Cozad, A. & Sahinidis, N. V. A global MINLP approach to symbolic regression. *Math. Program.* **170**, 97–119 (2018).
57. Kim, J. T., Kim, S. & Petersen, B. K. An interactive visualization platform for deep symbolic regression. *IJCAI* **2021-Janua**, 5261–5263 (2020).
58. Biggio, L., Bendinelli, T., Neitz, A., Lucchi, A. & Parascandolo, G. Neural Symbolic Regression that Scales. *PMLR* (2021).
59. Valipour, M., You, B., Panju, M. & Ghodsi, A. SymbolicGPT: A Generative Transformer Model for Symbolic Regression. *arXiv* (2021).
60. Egan, K., Li, W. & Carvalho, R. Automatically discovering ordinary differential equations from data with sparse regression. *Commun. Phys.* **7**, 20 (2024).
61. Martius, G. & Lampert, C. H. Extrapolation and learning equations. *arXiv* 1610.02995 (2016).
62. Sahoo, S. S., Lantpert, C. H. & Martius, G. Learning equations for extrapolation and control. *ICML* **10**, 7053–7061 (2018).
63. Orzechowski, P., Cava, W. L. & Moore, J. H. Where are we now? A large benchmark study of recent symbolic regression methods. *GECCO 2018 - Proc. 2018 Genet. Evol. Comput. Conf*. 1183–1190 https://doi.org/10.1145/3205455.3205539 (2018).
64. Žegklitz, J. & Pošík, P. Benchmarking state-of-the-art symbolic regression algorithms. *Genet. Program. Evolvable Mach.* **22**, 5–33 (2021).
65. La Cava, W. et al. Contemporary Symbolic Regression Methods and their Relative Performance. *NeurIPS* (2021).
66. Suseela, S. S., Feng, Y. & Mao, K. A Comparative Study on Machine Learning algorithms for Knowledge Discovery. *ICARCV* 131–136 https://doi.org/10.1109/ICARCV57592.2022.10004302 (2022).
67. Kaptanoglu, A. A., Zhang, L., Nicolaou, Z. G., Fasel, U. & Brunton, S. L. Benchmarking sparse system identification with low-dimensional chaos. *Nonlinear Dyn*. https://doi.org/10.1007/s11071-023-08525-4 (2023).
68. Cranmer, M. Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl. *arXiv* (2023).
69. de Silva, B. et al. PySINDy: A Python package for the sparse identification of nonlinear dynamical systems from data. *J. Open Source Softw.* **5**, 2104 (2020).
70. Vereecken, H. et al. Soil hydrology in the Earth system. *Nat. Rev. Earth Environ.* **3**, 573–587 (2022).
71. Jian, J. et al. Leveraging observed soil heterotrophic respiration fluxes as a novel constraint on global-scale models. *Glob. Change Biol.* **27**, 5392–5403 (2021).
72. Weber, T. K. D. et al. Hydro-pedotransfer functions: a roadmap for future development. *Hydrol. Earth Syst. Sci.* **28**, 3391–3433 (2024).
73. Rahmati, M. et al. Soil is a living archive of the Earth system. *Nat. Rev. Earth Environ.* **4**, 421–423 (2023).
74. Santoro, M., Cartus, O. & Fransson, J. E. S. Integration of allometric equations in the water cloud model towards an improved retrieval of forest stem volume with L-band SAR data in Sweden. *Remote Sens. Environ.* **253**, 112235 (2021).
75. Khabbazan, S. et al. The influence of surface canopy water on the relationship between L-band backscatter and biophysical variables in agricultural monitoring. *Remote Sens. Environ.* **268**, 112789 (2022).
76. Sahnoun, K. & Benabadji, N. Data Cubes for Earth System Research: Challenges Ahead. *arXiv* **2**, 1–4 (2023).
77. Cortiella, A., Park, K. C. & Doostan, A. A Priori Denoising Strategies for Sparse Identification of Nonlinear Dynamical Systems: A Comparative Study. *J. Comput. Inf. Sci. Eng.* **23**, 1–34 (2022).
78. LeVeque, R. J. *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems*. (SIAM, 2007).

79. Fan, J. & Gijbels, I. *Local Polynomial Modelling and Its Applications*. (Routledge, 2018).
80. Schaeffer, H. Learning partial differential equations via data discovery and sparse optimization. *Proc. R. Soc. Math. Phys. Eng. Sci.* **473**, 20160446 (2017).
81. Schaeffer, H. & McCalla, S. G. Sparse model selection via integral terms. *Phys. Rev. E* **96**, 1–7 (2017).
82. Kang, S. H., Liao, W. & Liu, Y. IDENT: Identifying Differential Equations with Numerical Time Evolution. *J. Sci. Comput.* **87**, 1–27 (2021).
83. Wentz, J. & Doostan, A. Derivative-based SINDy (DSINDy): Addressing the challenge of discovering governing equations from noisy data. *Comput. Methods Appl. Mech. Eng.* **413**, 116096 (2023).
84. Messenger, D. A. & Bortz, D. M. Weak SINDy for partial differential equations. *J. Comput. Phys.* **443**, 110525 (2021).
85. Gurevich, D. R., Reinbold, P. A. K. & Grigoriev, R. O. Robust and optimal sparse regression for nonlinear PDE models. *Chaos* **29**, 103113 (2019).
86. Li, J., Sun, G., Zhao, G. & Lehman, L. H. Robust Low-Rank Discovery of Data-Driven Partial Differential Equations. *Proc. AAAI Conf. Artif. Intell.* **34**, 767–774 (2020).
87. Dorigo, W. et al. The International Soil Moisture Network: serving Earth system science for over a decade. *Hydrol. Earth Syst. Sci*. 5749–5804 https://doi.org/10.5194/hess-25-5749-2021 (2021).
88. Vivien, M. The big challenges of big data. *Nature* **498**, 255–260 (2013).
89. Kim, S. et al. Integration of Neural Network-Based Symbolic Regression in Deep Learning for Scientific Discovery. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 4166–4177 (2021).
90. Berg, J. & Nyström, K. Data-driven discovery of PDEs in complex datasets. *J. Comput. Phys.* **384**, 239–252 (2019).
91. Kemeth, F. P. et al. Learning emergent partial differential equations in a learned emergent space. *Nat. Commun.* **13**, 1–13 (2022).
92. Champion, K., Lusch, B., Nathan Kutz, J. & Brunton, S. L. Data-driven discovery of coordinates and governing equations. *Proc. Natl. Acad. Sci. USA.* **116**, 22445–22451 (2019).
93. Bakarji, J., Champion, K., Kutz, J. N. & Brunton, S. L. Discovering Governing Equations from Partial Measurements with Deep Delay Autoencoders. *Proc. R. Soc. Math. Phys. Eng. Sci*. https://doi.org/10.1098/rspa.2023.0422 (2023).
94. Mars Gao, L. & Nathan Kutz, J. Bayesian autoencoders for data-driven discovery of coordinates, governing equations and fundamental constants. *Proc. R. Soc. Math. Phys. Eng. Sci.* **480**, 20230506 (2024).
95. Song, W., Shi, L., Wang, L., Wang, Y. & Hu, X. Data-Driven Discovery of Soil Moisture Flow Governing Equation: A Sparse Regression Framework. *Water Resour. Res.* **58**, 1–24 (2022).
96. Meng, X. & Karniadakis, G. E. A composite neural network that learns from multi-fidelity data: Application to function approximation and inverse PDE problems. *J. Comput. Phys.* **401**, 109020 (2020).
97. Foster, D. *Generative Deep Learning*. (O'Reilly Media, Inc., 2022).
98. Chang, H. & Zhang, D. Identification of physical processes via combined data-driven and data-assimilation methods. *J. Comput. Phys.* **393**, 337–350 (2019).
99. Bach, F., Jenatton, R., Mairal, J. & Obozinski, G. Optimization with sparsity-inducing penalties. *Found. Trends® Mach. Learn.* **4**, 1–106 (2012).
100. Yuan, M. & Lin, Y. Model Selection and Estimation in Regression with Grouped Variables. *Tech. Rep. Dep. Stat. Univ. Wis*. (2004).
101. Luo, Y., Liu, Q., Chen, Y., Hu, W. & Zhu, J. Physics-Guided Discovery of Highly Nonlinear Parametric Partial Differential Equations. *NeurIPS* **1**, 22 (2022).
102. Bloom, A. A., Exbrayat, J. F., Van Der Velde, I. R., Feng, L. & Williams, M. The decadal state of the terrestrial carbon cycle: Global retrievals of terrestrial carbon allocation, pools, and residence times. *Proc. Natl. Acad. Sci. USA.* **113**, 1285–1290 (2016).
103. Pinder, G. F. & Celia, M. A. *Subsurface Hydrology*. *Subsurface Hydrology*. https://doi.org/10.1002/0470044209 (2006).
104. Xu, H., Zeng, J. & Zhang, D. Discovery of partial differential equations from highly noisy and sparse data with physics-informed information criterion. *Research* 1–30 https://doi.org/10.34133/research.0147 (2023).
105. Ma, W., Zhang, J., Feng, K., Xing, H. & Wen, D. Dimensional homogeneity constrained gene expression programming for discovering governing equations. *J. Fluid Mech.* **985**, A12 (2024).
106. Zhang, J. & Ma, W. Data-driven discovery of governing equations for fluid dynamics based on molecular simulation. *J. Fluid Mech.* **892**, 1–18 (2020).
107. Reinbold, P. A. K., Kageorge, L. M., Schatz, M. F. & Grigoriev, R. O. Robust learning from noisy, incomplete, high-dimensional experimental data via physically constrained symbolic regression. *Nat. Commun.* **12**, 1–8 (2021).
108. Bonan, G. B. & Doney, S. C. Climate, ecosystems, and planetary futures: The challenge to predict life in Earth system models. *Science* **359**, eaam8328 (2018).
109. Floryan, D. & Graham, M. D. Data-driven discovery of intrinsic dynamics. *Nat. Mach. Intell.* **4**, 1113–1120 (2022).
110. Chen, B. et al. Automated discovery of fundamental variables hidden in experimental data. *Nat. Comput. Sci.* **2**, 433–442 (2022).
111. Lu, P. Y., Dangovski, R. & Soljačić, M. Discovering conservation laws using optimal transport and manifold learning. *Nat. Commun.* **14**, 4744 (2023).
112. Cohrs, K.-H., Varando, G., Sales-Pardo, M., Guimera, R. & Camps-Valls, G. Semiparametric inference and equation discovery with the bayesian machine scientist. (2024).
113. Guimerà, R. et al. A Bayesian machine scientist to aid in the solution of challenging scientific problems. *Sci. Adv*. **6**, eaav6971 (2020).
114. More, K., Tripura, T., Nayek, R. & Chakraborty, S. A Bayesian Framework for learning governing Partial Differential Equation from Data. *Phys. Nonlinear Phenom.* **456**, 133927 (2023).
115. Tripura, T. & Chakraborty, S. A sparse Bayesian framework for discovering interpretable nonlinear stochastic dynamical systems with Gaussian white noise. *Mech. Syst. Signal Process.* **187**, 109939 (2023).
116. Mathpati, Y. C., Tripura, T., Nayek, R. & Chakraborty, S. Discovering stochastic partial differential equations from limited data using variational Bayes inference. *Comput. Methods Appl. Mech. Eng.* **418**, 116512 (2023).
117. Naozuka, G. T., Silva, R. S. & Almeida, R. C. SINDy-SA: Enhancing Nonlinear System Identi cation with Sensitivity Analysis sensitivity analysis. *Nonlinear Dyn*. https://doi.org/10.1007/s11071-022-07755-2 (2022).
118. Fasel, U., Kutz, J. N., Brunton, B. W. & Brunton, S. L. Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control. *Proc. R. Soc. Math. Phys. Eng. Sci*. https://doi.org/10.1098/rspa.2021.0904 (2021).
119. Genkin, M., Hughes, O. & Engel, T. A. Learning non-stationary Langevin dynamics from stochastic observations of latent trajectories. *Nat. Commun.* **12**, 1–9 (2021).
120. Kutz, J. N. & Brunton, S. L. Parsimony as the ultimate regularizer for physics-informed machine learning. *Nonlinear Dyn*. https://doi.org/10.1007/s11071-021-07118-3 (2022).
121. Famiglietti, C. A. et al. Optimal model complexity for terrestrial carbon cycle prediction. *Biogeosciences* **18**, 2727–2754 (2021).
122. Akaike, H. Information theory and an extension of the maximum likelihood principle. in *Selected papers of hirotugu akaike* 199–213 (Springer, 1998).
123. Schwarz, G. Estimating the dimension of a model. *Ann. Stat*. **6**, 461–464 (1978).

124. Xie, X., Liu, W. K. & Gan, Z. Data-driven discovery of dimensionless numbers and scaling laws from experimental measurements. *Nat. Commun*. 1–11 https://doi.org/10.1038/s41467-022-35084-w (2022).

Acknowledgements

This Perspective was funded by the National Natural Science Foundation of China (No. 52179038 and No. 51979200).

Author contributions

W.S., S.J., and L.S. led the conceptualization, writing and figure drafting. G.C., L.Z., M.W., M.R., H.V., L.H., and X.H. supported in equal parts of the conceptualization, writing and reviewing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43247-024-01760-6.

Correspondence and requests for materials should be addressed to Liangsheng Shi.

Peer review information *Communications Earth & Environment* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Alireza Bahadori. A peer review file is available.