nature reviews earth & environment

https://doi.org/10.1038/s43017-023-00409-w

Review article

Check for updates

Big Data in Earth system science and progress towards a digital twin

Xin Li[1,2,7] ✉, Min Feng[1,2,7] ✉, Youhua Ran[2,3], Yang Su[2,3], Feng Liu[3], Chunlin Huang[3], Huanfeng Shen[4], Qing Xiao[2,5], Jianbin Su[1], Shiwei Yuan[1] & Huadong Guo[2,5,6]

Abstract

The concept of a digital twin of Earth envisages the convergence of Big Earth Data with physics-based models in an interactive computational framework that enables monitoring and prediction of environmental and social perturbations for use in sustainable governance. Although computational advances are rapidly progressing, digital twins of Earth have not yet been produced. In this Review, we summarize the methodological and cyberinfrastructure advances in Big Data that have advanced the progress towards a digital Earth twin. Data assimilation provides the framework for incorporation of high-resolution observations into Earth system models but lacks the decision-making interface and learning ability needed for the digital twin. Machine learning (and particularly deep learning) in Earth system science is now more capable of reaching the high dimensionality, complexity and nonlinearity of real-life Earth systems and is expanding the learning ability from Big Data. Progress in causal inference and reinforcement learning are, respectively, increasing the interpretability of Big Data and the ability of simulations to solve sequential decision-making problems. Social sensing data could provide inputs for multiagent deep reinforcement learning via feedback loops between agents and the environment, enabling large-scale applications in human system modelling. Future research must focus on finding the optimal way to integrate these individual methodologies to achieve digital twins.

Sections

[1]National Tibetan Plateau Data Center, State Key Laboratory of Tibetan Plateau Earth System, Environment and Resources (TPESER), Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing, China. [2]University of Chinese Academy of Sciences, Beijing, China. [3]Northwest Institute of Eco-Environment and Resources, University of Chinese Academy of Sciences, Lanzhou, China. [4]School of Resource and Environmental Sciences, Wuhan University, Wuhan, China. [5]Aerospace Information Research Institute, University of Chinese Academy of Sciences, Beijing, China. [6]International Research Center of Big Data for Sustainable Development Goals, Beijing, China. [7]These authors contributed equally: Xin Li, Min Feng. ✉e-mail: xinli@itpcas.ac.cn; mfeng@itpcas.ac.cn

Key points

- The volume of Big Earth Data is increasing year on year across all categories (remote sensing, in situ, social sensing, and simulation and reanalysis), with the addition of social sensing data contributing the largest increase since the 2010s.
- Big Data assimilation encapsulates the strengths of data-driven approaches and incorporates them into ultrahigh-resolution Earth system models, allowing the assimilation of multisource observations.
- Combining machine learning with process-based models and causal inference can enhance the transferability, interpretability and predictability of Earth system science.
- Deep reinforcement learning integrated with agent-based modelling provides a promising framework to address complex governance decision-making problems.
- These advances, plus technological innovations in computer infrastructure, are allowing Earth system research to evolve towards a digital twin of Earth, a replication of the Earth system constrained by physical laws and available Big Earth Data.
- Big Data and the development of the digital twin are helping the scientific community to comprehensively model the coevolution of humans and nature, and to address sustainable development issues at a planetary scale.

Introduction

Vast quantities of observational data are gathered on Earth every day, including from satellites[1], sensors[2], drones and non-traditional social sensing data[3]. This sheer amount of data is big in volume, variety, format and source (Box 1) and is allowing increased understanding of Earth systems and processes at ever finer scales[4]. However, although Earth scientists are experienced at dealing with large amounts of data (such as in numerical weather prediction[5], Earth system modelling[6] and long-term climate forecasts), many data processing methods were designed to handle discipline-specific or structured data with limited volume and/or variables. The speed of collecting and creating data is currently far outpacing the ability to effectively assimilate and perceive it[5,7,8]. New approaches are needed that can effectively extract information and knowledge from multisource large-volume Big Earth Data[9].

The concept of a digital twin was introduced in the early 2000s in response to the increasing need to accurately design and operate a simulation of our complex world[10]. A digital twin is a dynamic simulation of a process, system or environment that identically replicates the physical or real-life counterpart[11–13]. By integrating observational data with physical laws, these information systems allow possible real-life outcomes and issues to be digitally monitored and predicted[14–16]. In essence, a digital twin of Earth would help enable scientists and policymakers alike to assess environmental change and human impacts in support of sustainable development[17]. Considering the huge amount of resources needed for a digital twin of Earth, the progress of implementation requires long-term continuous efforts from extensive international collaborations between Earth system scientists, computer scientists, industry and governing bodies[18]. Ambitious projects have been established to explore the digital twin of Earth by both governmental organisations and large private corporations, such as the European Union[17] and NVIDIA.

Progress in Big Data, Earth system science (ESS), machine learning and cyberinfrastructure is enabling rapid advancement toward Earth's digital twin[17,19,20]. Technological innovations in ESS[21] have enabled efficient modelling of chaotic complexity, integration of human dynamics, and the ability to discover future Earth governance pathways that operate within safe planetary boundaries[22]. Additionally, advances in machine learning have been marked by the development of deep learning and other methods that are more capable of manipulating the high dimensionality, complexity and nonlinearity of Earth systems[23,24]. In particular, progress in deep reinforcement learning allows intelligent agents (for example decision-makers) to learn from interaction with the environment rather than given sets of data[25,26], even in nonlinear and high-dimensional environments[25]. This will be a critical feature of a digital twin.

In this Review, we provide an overview of advances in Big Data assimilation and machine learning that are enabling progress towards a digital twin of Earth. We aim to provide a broad assessment of Big Data analytics that goes beyond previous thorough reviews that provided perspectives from deep learning[9,27] or ESS[28]. We identify the remaining and upcoming challenges in Big Earth Data, including quantification of social tipping points in human–environment interactions, digitalization of the deep Earth and deep geological time where limited data are available, and cultivation of an open and fair data culture.

Towards a digital twin

A digital twin of Earth would be an indispensable technology to achieve bidirectional communication, dynamic interaction and real-time connection between the physical and digital worlds[14–16]. The value of the digital twin has been quickly recognized by those in ESS, owing to the urgent need for accurately modelling the Earth and simulating and forecasting complex natural events and phenomena[18]. The digital twin of Earth is envisaged to be a system that builds a digital replication of the state and temporal evolution of the Earth system, constrained by available Earth observations and the laws of physics[17,29,30]. The development has its roots in the rapid advancement of Earth observations, high-performance computing and long-term Earth modelling research, especially data assimilation[18,30,31].

A digital twin is also an open and interactive system, and it can help to inform us of the potential impacts of human activities. Therefore, the digital twin of Earth could fully integrate observations, Earth system models and data from human systems to finally deliver the knowledge base needed for guiding the actions of human society in addressing crises such as climate change and its adaptation and mitigation. In response to the growing threat of climate change, researchers and policymakers are increasingly considering actions to address the changes, such as reducing greenhouse gas emissions[32–34] and keeping Earth systems within safe planetary boundaries[35,36]. However, because of the large-scale nature of the problems, many actions are costly and/or risky to engage in, and often they have unknown impacts on other parts of the Earth system. The digital twin of Earth could be a simulation tool to develop and test the effects and trade-offs of interventions or different choices, resulting in better-informed decisions for sustainable development.

As such, many efforts have been initialized worldwide to implement a digital twin of Earth. One of the leading efforts is from the European Union on finalizing plans for an ambitious digital twin of planet

Box 1

Big Earth Data

The term 'Big Data' was initially introduced to capture the scale and variety of large-volume datasets, in addition to the processing, organization, policies and challenges of handling a large amount of data[189,190], and from there, the term 'Big Earth Data' was brought forward to encompass all data related to Earth systems[191]. Big Earth Data involve a wide range of data sources, which can be categorized as follows.

- **Remote sensing**: observations obtained from satellite, airborne, unmanned aerial vehicles and ground-based instruments, providing measurements of the spatiotemporal variability in the Earth system. The accumulated volume of remote sensing data by 2020 was ~1.3 EB, and this number will keep increasing with the launch of new satellites and the expansion of observation band capacities and spatiotemporal resolution. For example, the weather satellite Himawari-8 obtains observations at a frequency of one per minute and produces ~100 TB of data yearly, approximately 25 times the data produced by the previous-generation satellites (Himawari-6 and -7).
- **In situ and laboratory analyses**: includes data collected from observation stations, networks, laboratory analyses, surveys, expeditions, field experiments and so on. As these types of sample are usually taken close to the observed subjects, they can acquire real-time fidelity measurements of the subjects, but this does limit the spatial representation and variation across large regions. Networking sites, such as Particulate Matter (PM) 2.5 sites in China, FluxNet[2] and eLTER, could make it possible for regions to collect sufficient observations. With the support of the cutting-edge Internet of Things and sensor techniques[192], deploying a large amount of in situ sensors and establishing real-time data transmission have become more practical over time[193].
- **Social sensing**: broadly refers to data relating to human activities[3]. Social sensing data are quickly accumulating and now represent three-quarters of all the data currently generated[194]. This category of data is highly diverse, covering human behaviour, population, economic and other perspectives[42,195]. Analysis of social sensing data is benefiting from the advancements in text mining and deep learning, which have attracted increasing attention in Earth sciences[196], in particular to better understand human–environment interactions.
- **Simulation and reanalysis**: data produced from simulations and reanalysis (data assimilation) of Earth systems and their interactions. Computer-based physical and theoretical models have been developed to simulate different parts of Earth (for example, atmosphere, oceans, deep Earth, land surface and cryosphere) and coupled together to perform integrated simulations, such as the Coupled Model Intercomparison Project (CMIP)[6] and Copernicus services, which provide essential information for Earth system simulations. The improvements in temporal and spatial resolutions of these models have led to rapidly increasing data volumes. For example, over 5.6 million datasets more than 21.9 petabytes in size have been published from CMIP6 activities. Additionally, state-of-the-art numerical models with assimilation of data observations from different sources produce reanalysis datasets with long time records and high consistency in both spatial and temporal domains.

Each data category provides insight into the Earth system from different aspects, but each has limitations regarding data quality versus coverage. Integration of various datasets from across these categories can provide a more spatiotemporally and physically consistent representation of the Earth system.

Earth named 'Destination Earth'[17]. This system would simulate the atmosphere, ocean, ice and land with unrivalled precision, such as real time at 1-km resolution, providing forecasts of floods, droughts and fires from days to years in advance (Fig. 1). In addition, NASA has plans to develop the Earth System Digital Twin (ESDT), which would dynamically integrate state-of-the-art Earth and human activity models, observations from Big Data and analytical tools to provide interactive multidomain, multiscale, digital replicas of Earth systems[37]. The private corporation NVIDIA has also revealed plans to build a digital twin of Earth on an artificial intelligence (AI) supercomputer named Earth-2, dedicated to predicting climate change. Those projects represent the latest exploration of the digital twin of Earth from diverse perspectives, attracting attention and action from academical and industrial communities. Although these projects are still at an early stage, Destination Earth is setting a clearer pathway towards the realization of the digital twin of Earth, which is expected by 2030.

Cyberinfrastructure is a combination of networked data, computing and other information technology resources that provide high-performance computing for data-rich applications[38]. Considering the vast cyberinfrastructure resources needed for digital twins of Earth, such as data compression methods[39], graphics processing units and high-performance computing[19,40], their progress and implementation requires long-term continuous efforts from extensive international collaborations[18]. Upgrades in these cyberinfrastructures have empowered the advancement in Earth science modelling and reanalysis, from small-scale analysis on personal computers to extensive computing and data demands that can only be met by high-performance computing[19]. Data-use scenarios in Earth sciences have transitioned from a 'data supermarket' phase (simply connecting data users with different providers for accessing data) to a 'data library' phase (better-organized data archiving and interfacing). The emergence of Big Data has been pushing data use into the data laboratory phase, which involves comprehensive data integration, interactive interfaces, and powerful analytics and processing capabilities, allowing scientists to efficiently collaborate, and quickly design and evaluate ideas without being concerned about the details of data handling and processing (Fig. 1). The digital twin has been advocated as the next critical step in the evolution of the Big Data era[18], because of its exceptional ability to fuse diversified observations and models.

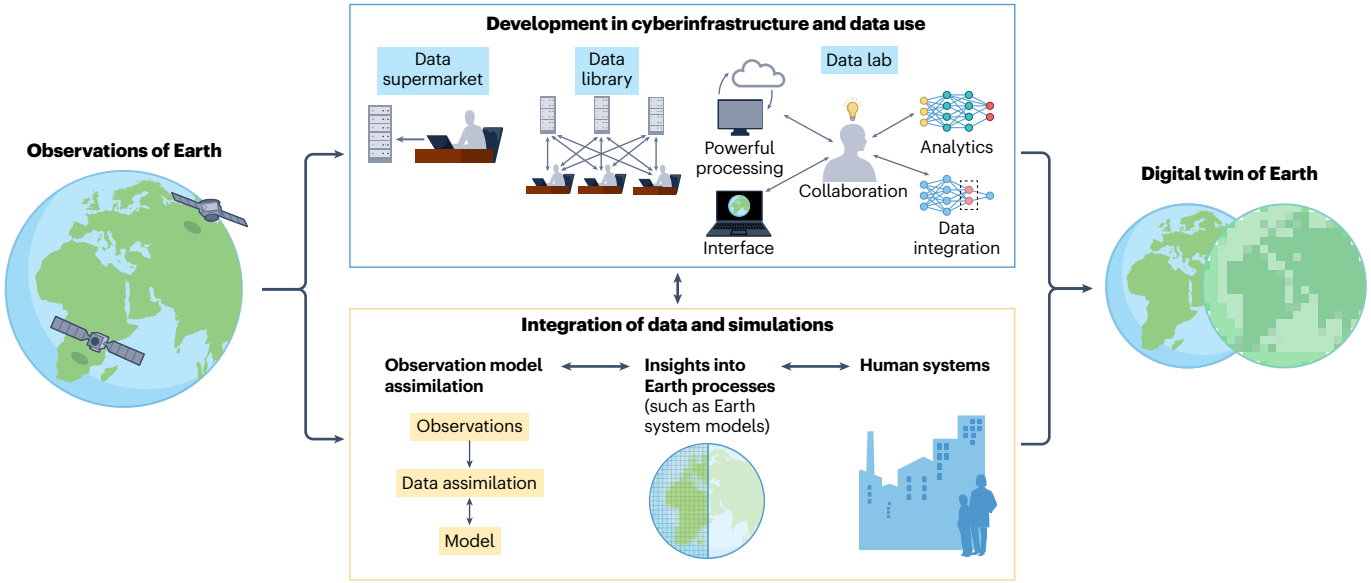

Fig. 1 | Transition of data use in Earth system science. Observational data of the physical world are processed with support from rapidly improving cyberinfrastructure. The data-use phase is transitioning from a data supermarket to a data library, a data laboratory and finally towards a digital twin of Earth. Big Earth Data and social sensing data are assimilated into process-based models, such as Earth system models, to give insights into Earth processes. The digital twin of Earth will include both dynamic models and data with high spatiotemporal resolution, to provide a virtual representation and digital counterpart of the real world[17,19].

This next generation of cyberinfrastructure would become a key component of the Earth metaverse, a single, universal and immersive virtual world[41]. More importantly, it would provide a critical tool for fighting climate change and advancing towards sustainable governance[42]. However, achieving the digital twin of Earth requires innovations across a wide range of analytical methods, which we discuss in the following sections, to provide the ability to fuse all categories of Big Earth Data.

Big Data assimilation

Data assimilation refers to the mathematical techniques used to optimally combine theory and observational data together to estimate possible evolving states of a system over time[43,44]. In Earth sciences, the classic example of data assimilation is the generation of initial conditions used in numerical weather prediction. Since then, data assimilation has evolved to be a backbone methodology in almost all branches of ESS[43,45]. The theoretical element usually takes the form of an Earth system model, climate system model, physical seismic model or numerical model, or some combination of these. Observational elements are typically data from remote sensing, in situ sensors, monitoring stations, sometimes also the Internet of Things and, recently, social sensing.

In the Big Data era, data assimilation could integrate all these sources of Big Earth Data (Box 1) to achieve a more spatiotemporally and physically consistent representation of the Earth system. However, challenges remain in various aspects, including quantification of random errors of models and observations, correction of bias, managing computational cost and assimilating social data. 'Big Data assimilation' (BDA) was proposed to tackle these challenges. This is a data assimilation scheme that combines machine learning methods with ultrahigh-resolution Earth system models and observations (Fig. 2a), which can include non-mainstream data and social data of human systems. Primarily, BDA is manifested in the following four aspects.

First, BDA is capable of synthesizing models and observational data at higher resolutions than traditional data assimilation. For example, a BDA system for numerical weather prediction was developed at resolutions of 100 m (Fig. 2b), outperforming mainstream weather prediction approaches, which typically use a resolution coarser than 10 km (refs. 46,47) or ~5 km in advanced Earth system models[48]. The observations were acquired from a phased-array weather radar with 100-m spatial resolution, 30-s temporal resolution and 1.2° angular resolution. The ensemble size in one of these BDA experiments was set to 10,240 (ref. 49), much larger than those of traditional systems, whose ensemble size is usually smaller than 100. As such, this system helps to grasp finer-scale climate dynamics (for example cloud physics), which cannot be captured in coarse-scale simulation and observations. The non-Gaussian error distributions were reduced by increasing assimilation frequency and observation number[50]. This BDA system was supported by a supercomputer with 10 petaflops (10×10^{15} floating-point operations per second) to provide 30-min nowcasting of sudden heavy precipitation[51]. The system could expand the volume and variety of adopted observations in practice, to move towards being an operational system for global weather forecasting services.

Second, BDA is capable of assimilating non-mainstream data, which are usually multidisciplinary, unstructured and characterized by larger observation errors. For example, smartphone air-pressure observations collected from 350 smartphone locations were assimilated into high-resolution numerical weather predictions[52] (Fig. 2c), allowing more accurate short-term mesoscale (~4 km) numerical weather prediction by resolving convective-scale features. An update of this

type of BDA is to assimilate the smartphone observations with higher spatiotemporal resolution (10 m and 6 minutes) to analyse fine-scale patterns of hailstorms[53]. Traditional surface pressure observations are limited by the number of weather monitoring sites, so the addition of smartphone pressure observations could increase the data volume considerably, potentially generating billions of extra observations. Using these smartphone observations worldwide could resolve finer-scale convective features and enable prediction of extreme weather, which is critical for the digital twin of Earth.

Third, assimilating social data is of unique added value because these data complement traditional Earth observations by capturing human dynamics, which is considered an indispensable component in the next generation of Earth system models[22]. However, obstacles exist, as social data can be categorial (for example ratings, choices and decisions) or unstructured (for example videos, photos and natural languages), and often have larger errors than natural system data[54]. Early-stage experiments on social data assimilation have explored the incorporation of social data derived from surveys or interviews into a socio-hydrological model to estimate parameters of flood awareness and preparedness, improving understanding of human–water interactions[55,56]. In addition, footfall count data collected by surveillance cameras have been successfully assimilated into agent-based models of urban population dynamics to forecast the real-time flows of people arriving and departing from a city[57]. However, social data assimilation is still in its infancy. Its maturation relies on developing observation operators that can map states or parameters of human dynamic models to social Big Data. Methodological maturation in assimilating social data and quantifying the representative errors could help to capture human dynamics and support the development of the human component in a digital twin of Earth.

Fourth, BDA is manifested in the integration of data assimilation and machine learning. Data assimilation explicitly includes physics in both model dynamics and observation operators. Thus, the integration of data assimilation and machine learning can maintain physical consistency while learning from Big Earth Data. So far, this integration has succeeded in uncertainty estimation of model and observations, improving parameterization and reducing computational costs by orders-of-magnitude[58–60]. For example, machine learning methods including kernel conditional density estimation[61] and Gaussian process regression[62] have been used to characterize model error with limited prior knowledge of error structure, thereby improving the forecast accuracy of streamflow by ~25–50%. In addition, the random forest algorithm was used to correct the bias in a snow data assimilation system, reducing the absolute mean error of snow water equivalent from 16 mm to 0.2 mm (ref. 63). Additionally, neural networks combined with low-resolution data assimilation were demonstrated to perform equally to a high-resolution system for only 55% of the computational cost[64]. Low-dimensional surrogate models for wildfire forecasting in a data assimilation system were developed to replace high-dimensional models and observation operators[65], reducing the computational cost to 0.1%[66]. Overall, unifying data assimilation and machine learning is an important aspect of BDA. However, implantation of such prototype studies into practical applications in numerical weather prediction or Earth system reanalysis requires further development. The next step is to explore more effective ways for integrating machine learning into BDA to provide a cost-effective drive layer for building a digital twin of Earth.

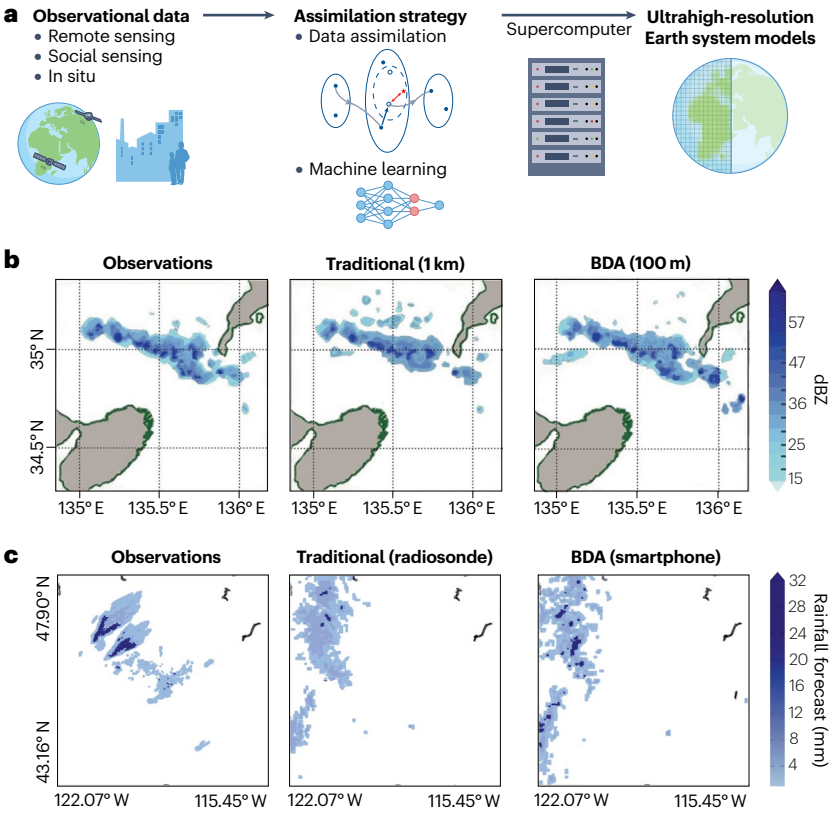

Fig. 2 | Big Data assimilation into ultrahigh-resolution models. a, Assimilation of observational data into ultrahigh-resolution models, such as Earth system models. The data assimilation strategy combines the strengths of traditional data assimilation methods and machine learning algorithms. Supercomputing infrastructures are required for the intensive computing of Big Data assimilation (BDA). **b**, One-kilometre-resolution assimilation results of precipitation from a traditional model (centre), compared with the 100-m-resolution BDA result (right), obtained by assimilating phased-array weather radar observations (left) into numerical weather models. dbZ is a measure of radar reflectivity. **c**, Assimilation of smartphone observations of pressure, a type of non-mainstream data, into high-resolution rainfall forecast models. The rainfall estimated using observations from a weather service radar (left) can be compared with rainfall forecast estimations by assimilating conventional radiosonde surface observations (centre) and 350 smartphone pressure observations (right), with the latter being a demonstration of BDA. The case studies demonstrate that BDA can outperform traditional data assimilation and could aid prediction and analysis in Earth system sciences. Data in panels **b** and **c** are from refs. 52,188.

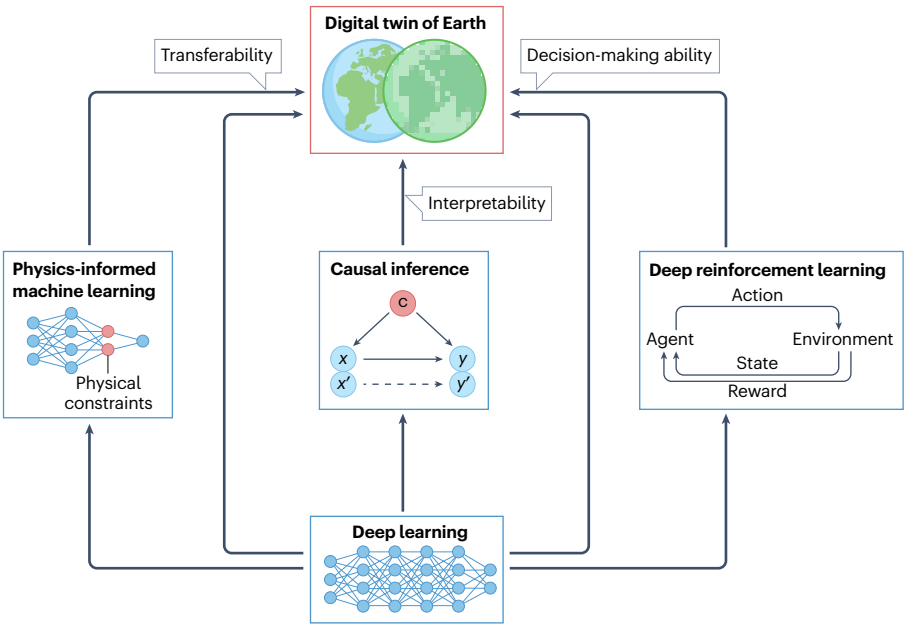

Fig. 3 | Interactions between deep learning, physics-informed machine learning, causal inference and reinforcement learning in Earth system science. Deep learning uses more layers than classic neural networks to mine deeper and more abstract features from the data and improve the learning ability of models. Physics-informed machine learning incorporates physical constraints to improve the transferability of models. Causal inference integrates the causal relationship between variables into machine learning models to improve the interpretability. Deep reinforcement learning performs interactions between the agent (human system) and the environment (natural system) to improve the application of Earth system models for decision-making purposes. These four frontier methods would complement each other to help to produce the digital twin of Earth.

To summarize, BDA is evolving from traditional data assimilation, with the main priorities being to incorporate global-scale social sensing data and to integrate data assimilation with machine learning. BDA has the potential to provide the real-time control layer in digital twins[19], through sequentially assimilating Big Data[13], adjusting to ultrahigh-resolution Earth system models, and producing a higher-resolution simulation of the Earth system. This progress urgently demands revolutionary yet unknown supercomputing infrastructures[19].

Machine and deep learning

The Earth system is an open and complex system, characterized by nonlinearity and chaos[22,67]. Insights into these complex systems can be gained through applying machine learning[28]. Classical machine learning methods, such as graphical models, support vector machines, random forests, dictionary learning and neural networks, have achieved success in Earth science[68–71], especially in extracting relevant information, such as land cover, from remote sensing data[72]. This progress has been well reviewed[9,27,28,73]. Machine learning is entering new stages marked by deep learning and other methods that are more capable of manipulating the high dimensionality, complexity and nonlinearity of Earth systems[23,24]. However, deep learning family algorithms still have limitations, including model interpretation, demand for large amounts of labelled data, extrapolation beyond training data, long-term predictability and decision-making in complex Earth systems[9]. New developments in physics-informed machine learning, causal inference and reinforcement learning are demonstrating a promising ability to overcome the limits of transferability, interpretability and decision-making (Fig. 3), with potential to deepen the application of Big Data in ESS and promote the implementation, accuracy and intelligence of the digital twin.

Deep learning

Deep learning performs learning directly from data with multiple processing layers, generally more than three layers, while nonlinear transformation of the representation occurs from one layer into the next higher, more abstract layer[28,74,75]. Multiple deep learning architectures, such as deep belief networks[23], recurrent neural networks[76], convolutional neural networks[77] and generative countermeasure networks[78], have achieved great success in addressing problems including classification, regression and prediction in Earth sciences[9,28,79–81].

The application of deep learning has increased in various Earth science fields, including but not limited to remote sensing, atmosphere and solid Earth. In the remote sensing field, deep learning has been successfully used in land use and cover classification[82,83], object detection[84] and quantitative inversion of Earth system parameters[62,85]. Deep learning has also outperformed classical statistical methods in short-term predictions, such as the prediction of landslides, crop yield, ice flow, sand dune migration, and runoff[86–89]. In atmospheric science, deep learning has been successfully used to deal with weather forecasting[81] and climate downscaling[90]. Its accuracy and efficiency for short-term rainfall predictions are higher than those of traditional methods[81]. In the solid Earth field, deep learning has shown the ability to outperform classical approaches, particularly in seismology, including seismic tomography, ground motion prediction, earthquake detection and early warning, and inverse and forward modelling[27,28,69,91–93]. For example, deep neural networks have been proven effective for solving inverse problems such as inferring Earth's subsurface properties that are often highly nonlinear and ill-posed[27,28].

From the perspective of data utilization, these preliminary successes of deep learning mainly refer to supervised learning, which benefited from the availability of a large volume and high quality of labelled data. However, it is costly and even unfeasible to obtain large and high-quality labelled Earth data for broader applications. Therefore, there is a strong interest in deep learning with weakly labelled data (incomplete, inexact and inaccurate)[94] and even nonlabelled data. A series of weakly supervised learning techniques, such as active learning, semisupervised learning, transductive learning and multi-instance learning, might potentially be combined with a deep learning architecture to promote high-accuracy spatiotemporal prediction with less

ground-truth labelled data and a large amount of inaccurate, crowd-sourced geographical data[94]. Self-supervised learning, such as contrastive learning, as an unsupervised learning architecture has received great attention[95] and has been applied to land cover classification and object recognition with unlabelled data[96,97].

Despite increasing interest in deep learning in Earth science applications, the mathematical reasons for the empirical success of deep network models remain elusive[98]. A deep learning model is defined in a latent space without the direct incorporation of known mechanisms or knowledge for Earth system processes[99]. The complexities of deep learning architectures also make the models difficult to interpret. These challenges, together with other limitations, such as the inability to deal with a hierarchical structure, struggle with open-ended inference, identification of causation from correlation[100], and potential risk of overfitting[101], are the technical bottlenecks of deep learning. Therefore, the architecture of deep learning needs to be expanded, for example by incorporating physics knowledge to break those limitations to address more complex ESS problems. This progress in deep learning would help to improve the intelligent evolution of digital twins of Earth.

Physics-informed machine learning

As mentioned above, the lack of labelled data, interpretability and physical consistency undermines the transferability of machine learning models and limits their applications in Earth sciences. The concept of physics-informed machine learning (also known as theory-guided machine learning or physics-constrained machine learning) incorporates machine learning and physical models that represent interpretable theories to improve the transferability of machine learning models[102,103]. Moreover, this type of incorporation could introduce domain theories to enhance the effectiveness of addressing complex prediction, inverse problems, scientific discovery and decision-making[9,80,93,104,105]. Unlike BDA, which is based on a dynamical model of the Earth system, physics-informed machine learning mainly couples physical knowledge (for example differential equations, causal relationship) within the machine learning framework. Physics-informed machine learning can be achieved through three main approaches to make a learning algorithm adhere to physical knowledge and generate physically consistent solutions[102,106].

The first approach is to incorporate physical knowledge through carefully crafted training samples that embody underlying physics. These samples, generated from physical models, can be used for pre-training of machine learning, acting as a weak constraint mechanism or enhancements to the original training samples to overcome data sparsity challenge and improve the models' generalization[107]. For example, observed data on lake temperature might be limited to a specific depth, location and time, but combining these observations with simulated data produced by a physics-based General Lake Model can overcome the scarcity of observed data[107]. Physics-based models such as the General Lake Model usually perform poorly without calibrations in unobserved lakes. Combining observed data from instrumented lakes produces better accuracy and generalization than the physics-based model[107,108].

The second approach is to customize a specialized machine learning architecture, embedding dynamics constraints and/or other prior knowledge in machine learning models. For example, one or more physical layers can be added on top of a multilayer neural network to make the model more physically realistic[9]. These layers can represent the physical laws of human or natural systems. Incorporating these layers in a neural network can ensure the conservation of important quantities (for example mass, momentum, energy) by imposing flux continuity constraints within the neural network architecture[109]. The approach has been explored in applications such as convective parametrization of climate modelling[110] and ozone simulation[109]. This approach involves imposing physical constraints on machine learning and requires a clear understanding of both the internal structure of machine learning and the physical processes.

The third approach is to embed physical knowledge via cost functions, by introducing constraints that are based on physical knowledge to improve the physical consistency and model performance, such as narrowing the search space for parameter solving and/or overcoming the overfitting problem. For example, a physics-based loss function is introduced in the learning objective of neural networks by considering the conservation of energy for modelling the dynamics of temperature in lakes[111] or retrieving latent heat flux[105]. Pilot studies demonstrate that this physics-informed approach outperforms pure machine learning models, reducing errors against training sets. Overall, the physics-informed approaches can introduce different kinds of physical constraints into machine learning models, and this can be easier to implement than the second approach because it involves modifications only to the loss function, avoiding altering the structures of machine learning models.

Deep integration of physical models and machine learning models is promising for addressing the limitations of pure machine learning and pure physical process models. Hybrid models can be interpreted as deepening a machine learning model to make it more physically realistic[9]. However, their implementations are challenging, because of the fundamental differences originating from their distinct fields. The successes of these early applications are starting a trend in embedding physics into machine learning in both communities. A key challenge is bridging the knowledge gap between members of the two individual communities, who are mostly either experts in machine learning or experts in physical modelling[112]. Overcoming these challenges will provide a key ability for building digital replications of the Earth system constrained by physical laws.

Causal inference

It is well known that "correlation does not imply causation"[113], meaning that the causal knowledge behind a phenomenon is not legitimately deduced from its observed association or correlation[114]. Although machine learning has been successful in Big Data applications, its statistical basis is still correlation relationships between variables[115]. Confounding factors and selection bias can result in spurious correlations, which are ubiquitous in the universal relation network of machine learning, affecting their robustness, stability and interpretability[116]. In addition, owing to a lack of consideration of temporal structure and causal direction, the algorithmic processes and predictions of most off-the-shelf machine learning cannot be easily explained mathematically or physically[117].

Traditional approaches to discovering causal relations from Big Earth Data, such as interventions and randomized experiments, are either infeasible or ethically problematic when studying the dynamical processes of the Earth system, such as detecting and attributing climate change[118,119]. Therefore, causal inference, which acts to establish covariation of cause and effect between phenomena and ambiguous factors from observational Big Earth Data or Earth system model outputs, has attracted much attention in ESS[120].

In ESS, causal inference mainly focuses on finding causal associations between Earth system variables, identifying their pathway

and quantifying causal effect from Big Earth Data, and understanding whether an intervention produces differences. Granger causality (GC) is a widely used qualitative method for discovering the causal association between pairs of variables, such as soil moisture feedback on precipitation[121,122] and mechanisms of climate–vegetation dynamics[123,124]. However, the method still struggles in complex multivariable systems, such as climate interactions related to the influence of the El Niño–Southern Oscillation (ENSO) on the Arctic stratosphere and the sea-level pressure in central Asia, which involve at least five variables[125,126]. More sophisticated causal inference methods, such as state-space methods[127] and mode decomposition algorithms[128], have been applied to address the aforementioned multivariate system and infer their causal association[129]; however, these methods still cannot identify the causal pathway. The latest developed methods, such as the causal network learning algorithm[130], have more ability to handle high-dimensional causal links via causal graphs[131,132]. These methods succeeded in analysing the complex spatiotemporal causal network structure in the weakly coupled Earth system, and obtained a causal graph of the variables by distinguishing pathways and eliminating the effect of confounding factors[133]. To address the challenge of quantifying the causal effect, the do-operator, $P(Y|\mathrm{Do}(X))$, was introduced to quantitatively represent the intervention of cause (X) on effect (Y)[134]. These methods have the potential to comprehensively evaluate the effect of physical intervention experiments, which could be impractical for the Earth system in real life[135].

A typical application of causal inference in ESS is in teleconnections[125,136,137], which are recurrent climate effects related to spatial climatic interactions[138] and play a crucial role in improving climate predictability[125]. As exemplified in the temperature anomalies of the tropical Pacific in the Walker circulation, correlation analysis only provided a completely connected graph of considered variables (Fig. 4a), while the standard bivariate GC identified the causal association between each pair of variables (Fig. 4b). PCMCI[139], a causal network learning algorithm, successfully reconstructed the Walker circulation (Fig. 4c). Benefiting from cooperation with a counterfactual framework[140] and the do-operator, the causal pathway and average causal effect of the considered variables can be acquired to evaluate their global climatic teleconnections (Fig. 4d).

In summary, these methods and applications mainly centre on helping to understand Earth processes, providing prior knowledge to diagnose and interpret Earth models, and optimizing the model structure[141]. Causal inference can confirm the appropriate model structure for machine learning, increasing their mathematical or physical interpretability[142]. Additionally, discovering causal association in Earth systems can provide constraints to machine learning models by adopting expert knowledge about the physical process to improve the transferability and stability of modelling the Earth system.

Deep reinforcement learning

Decision-making ability is one of the most critical aspects in ESS to identify pathways towards sustainable governance and achievement of the United Nations Sustainable Development Goals[36,143]. However, the machine learning methods mentioned above have limited decision-making ability. Reinforcement learning is a type of machine learning for solving sequential decision-making problems. Unlike supervised and unsupervised learning, in reinforcement learning, intelligent agents (for example decision-makers) learn from interaction with the environment rather than being given sets of data[25,26]. However, learning from higher-dimensional environment feedbacks has been one of the challenges in reinforcement learning, and most classical applications of reinforcement learning use linear value functions or policy representations. Therefore, deep reinforcement learning was proposed for more powerful feature expression and nonlinear modelling in high-dimensional environments[25]. The success in adversarial computer games has stimulated the deep reinforcement learning revolution[144–146].

The ability to learn from complex, heterogeneous and high-dimensional agent–environment interactions enables deep reinforcement learning to address decision-making problems in ESS. Typical applications of deep reinforcement learning in Earth sciences lie in natural disaster mitigation[147], natural resource management (carbon[148], farmland nitrogen[149] and water resources[150–152]) and sustainability governance[153]. For example, in flood mitigation and preparedness, deep reinforcement learning is used to guide dam operation[154] and for automated real-time stormwater system control[155,156]. In water resource management, it enables optimized irrigation scheduling to bring about considerable benefits in water saving[150,151]. For sustainability governance,

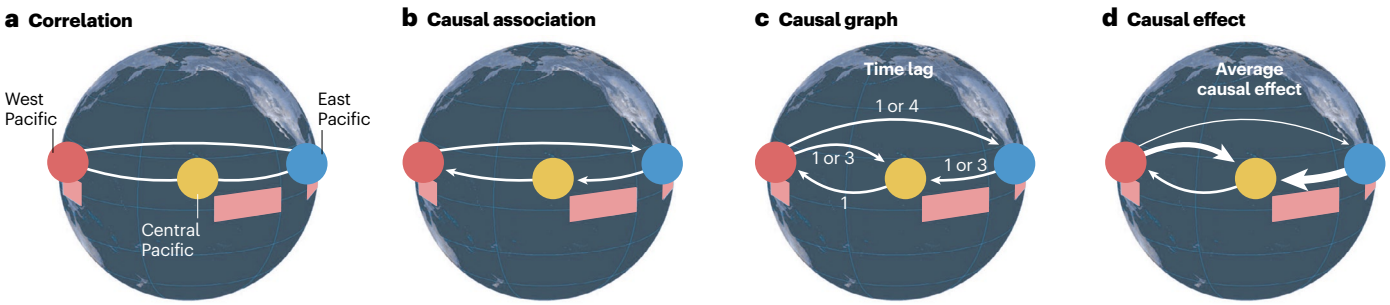

Fig. 4 | Causal inference to determine causation, causal pathway and causal effect of the Walker circulation. The Walker circulation is an air-flow model for the tropical Pacific, caused by the sharp contrast in sea surface temperatures and pressures between the east (cooler temperatures, higher pressures) and west (warmer temperatures, lower pressures). The monthly sea surface pressure anomalies in the West Pacific (red circle), and sea surface air-temperature anomalies in the Central Pacific (yellow circle) and East Pacific (blue circle), respectively, can be considered as node variables for causal inference. The pink boxes indicate sampling regions. **a**, Correlation analysis for the Walker circulation. **b**, Causal association can be acquired through antecedents and consequences (arrow) from each pair of climatic variables. **c**, The causal graph can be reconstructed by the time-lagged variables representing the common drivers and direct or indirect links of the Walker circulation. Numbers represent the time lag between antecedents and consequences. **d**, The causal effect of multiple climatic variables acquired by the 'do-operator'. The thickness of the arrow generalizes the potential averaged causal effect. Figure adapted from Runge et al.[118], Springer Nature Limited.

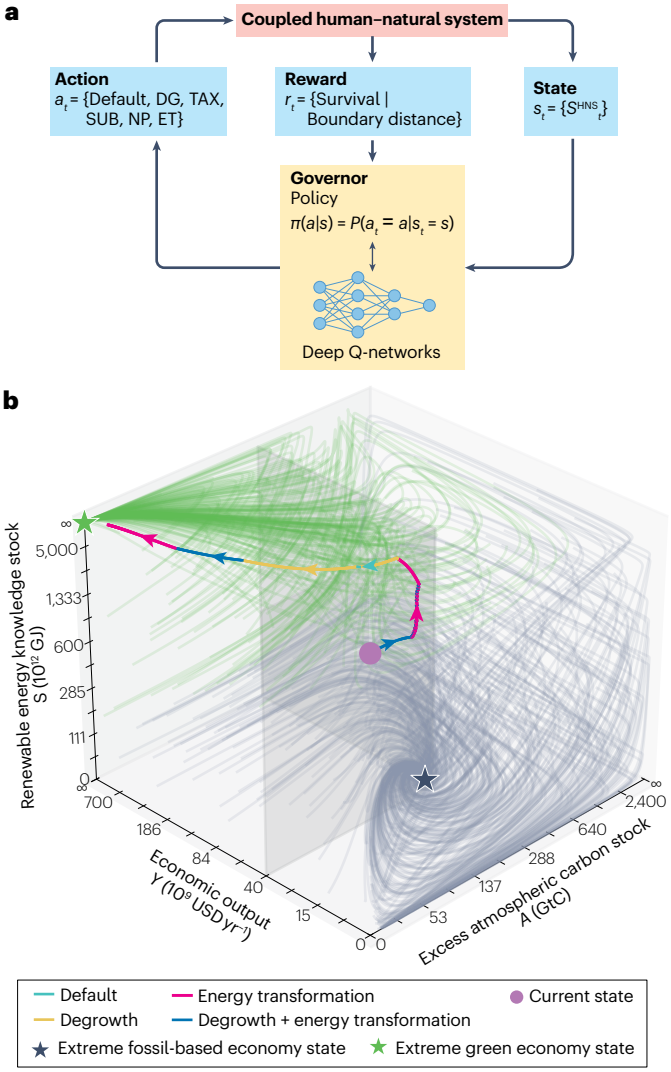

Fig. 5 | Identification of policy pathways on sustainable development using deep reinforcement learning. a, A summary diagram displaying the deep reinforcement learning (DRL) coupled human–natural system (HNS) model framework used to generate panel **b**. The environment is represented by the HNS model, and the agent is the governor and/or policymaker of the HNS. s_t is a certain state at time t of HNS that responds to the governor's action a_t. The action is a set of governance policy options. In this case, the action set includes degrowth (DG), tax on fossils (TAX), subsidies on renewables (SUB), nature protection during land use (NP), non-action (Default), and combinations such as energy transformation (ET) based on TAX and SUB. The governor's step-by-step governance decision-making is related to the system status s_t, corresponding reward r_t and the governor's learning ability. In this example, the learning process is implemented with a deep Q-network algorithm. **b**, Application of DRL-HNS using the AYS (A, excess atmospheric carbon stock; Y, economic output; S, renewable energy knowledge stock) model to identify trajectories inside the sustainability boundaries (grey surfaces). The green and black lines represent the default dynamics of the AYS model moving towards extreme green (green star) and fossil-based economy (black star) states, respectively. The coloured thick line is an example trajectory identified by DRL to guide the HNS to move from the estimated current state towards a green future within the sustainability boundaries. The DRL-HNS framework provides a promising tool to discover and analyse Earth governance pathways and to obtain a deeper understanding of the impact of governance policies. GtC, gigatonnes carbon; USD, US dollars; GJ, gigajoule. The data for part **b** were acquired from Strnad et al.[153]. The visualization for part **b** was performed using Python scripts on GitHub provided by Felix M. Strnad.

deep reinforcement learning combined with a human–natural system model provides a promising framework to identify and analyse policy portfolios in the context of global sustainability, and to obtain a better understanding of the impact of governance policies through an explicitly expressed state–reward–action transformation cycle between decision-maker and human–natural system (Fig. 5a; ref. 153). To illustrate, deep reinforcement learning was embedded with a simplified Earth system model (the AYS model), which can simulate the coupling among climate change, welfare growth and energy transformation. This integration successfully identifies the sustainable development trajectory under advisable governance actions to navigate the Earth in safe and just spaces (Fig. 5b).

However, the above applications are limited to a single-agent scenario. When the number of agents grows, as would be the case for a digital twin of Earth, the behaviours and decision-making in spatiotemporally heterogeneous human–natural systems become much more complex and highly nonlinear because of the exponential growth of the state and action spaces and computational loads. Benefiting from developments in high-performance computing and deep learning, multiagent deep reinforcement learning has been developed to tackle this problem[157–159]. Its effectiveness has been proven by a preliminary application in watershed water resource management, in which multiple agents are considered with different behaviour orientations and communication[160]. However, large-scale and sophisticated applications remain scarce, challenged by extracting the behaviours, values, norms and interactions of complex and large spatially heterogeneous multi-agents. Big Earth Data, particularly social sensing data, can potentially provide perceptual inputs for multiagent deep reinforcement learning via agent–environmental feedback loops (state–reward–action) and enable large-scale applications in human system modelling[161,162]. Other obstacles to large-scale multiagent deep reinforcement learning applications in Earth sciences, such as non-stationarity, high dimensionality and partial observability, can be tackled by new breakthroughs in Big Data and machine learning[157,159] and deep integration of complex adaptive systems theory with multiagent modelling[163] in human–natural systems. These advances should enhance the ability to addressing complex governance and decision-making problems.

Critical challenges

The progress of in situ, remote sensing, reanalysis and social sensing has contributed to a deluge of data towards ubiquitous sensing of the Earth system (Fig. 6). These advances are enabling computational and data-intensive research to deepen understanding of Earth system functions, discover new natural and societal processes, and pursue sustainable development in the Anthropocene epoch. However, embracing these advantages will require broad and long-term interdisciplinary and transdisciplinary efforts, technological advances in computing infrastructure, and an innovative and open Big Data culture.

Expanding into deep space and deep time

For a digital twin of Earth to be fully encompassing, it must include data and understanding of the deep seas, deep Earth and the evolution

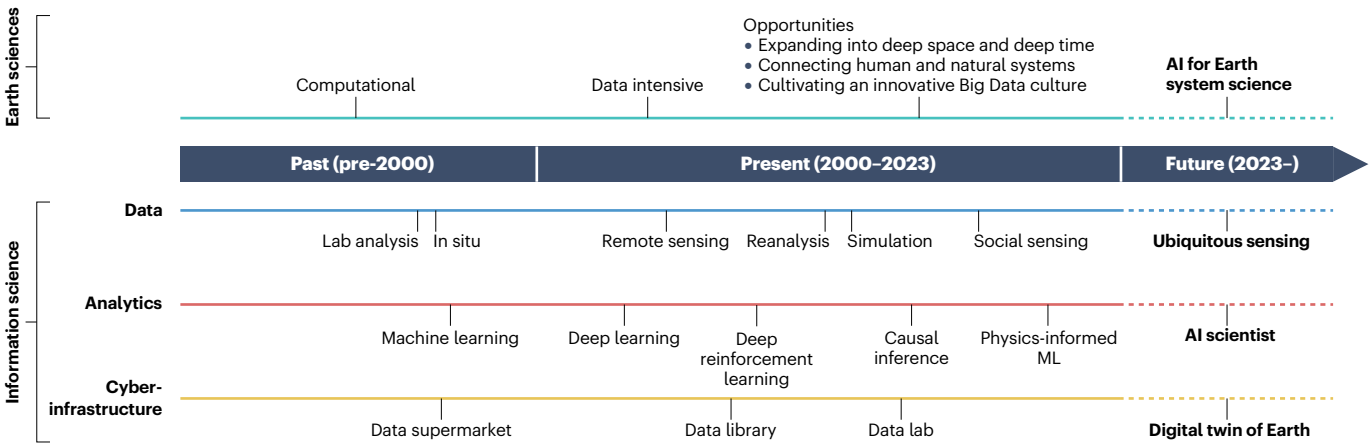

Fig. 6 | Grand challenges of Big Data use in Earth system science. Advances in data, analytics and cyberinfrastructure are providing opportunities for overcoming challenges in bridging human–natural systems, exploring deep space and time, and in developing an open-sharing data culture. Overcoming these challenges will boost Earth science from mostly computational and data-intensive research to the realization of artificial intelligence (AI) for Earth system science. Specifically, the progress of in situ, remote sensing, reanalysis and social sensing has contributed towards ubiquitous sensing of the Earth system. Advances in machine learning, deep learning, physics-informed machine learning (ML) and other analytics are enabling possibilities of AI scientists. The rapid development of cyberinfrastructure has moved through the data supermarket, data library and data laboratory phases towards a digital twin of Earth.

over geological time[22,164]. However, observational data are limited for most of Earth's interior and for ~4.5 billion years of its history, owing to sampling limitations[165,166].

Advances in Earth observation and Big Data analytics could provide a promising ability to explore and understand the complex Earth, but grand challenges are anticipated in transforming these data resources and analytical abilities to explore the full extent of the Earth system, particularly regarding its deep-time history and interior. To overcome data limitations[166], it is critical to construct and improve cyberinfrastructure to explore the deep-time and deep-space Earth by collecting, sharing and integrating data from various sources. An example would be the Deep-Time Digital Earth programme, which aims to harmonize deep-time Earth data to aid in data-driven discovery in the understanding of Earth's evolution[167]. Second, the high dimensionality, lack of causality and higher uncertainty of deep-time and deep-space Earth is likely to be beyond the current abilities of the Big Data analytics reviewed here. As such, combining deep-time and internal understanding of Earth with the transferability learned from data-rich times and spaces is likely to be the best path forward. Integrating physical deep-time or deep-Earth models and machine learning models could lead to methodological innovations by uniting deductive, inductive and abductive reasoning[102], which could enable transformative discoveries in the Earth sciences.

Connecting human and natural systems

Earth has entered a new human-dominated geological epoch, the Anthropocene[168]. In the Anthropocene, human activities have been exerting increasing impacts on the environment on all scales, triggering tipping points and surpassing some planetary boundaries[22,35,169,170]. To mitigate and undo human damage to the environment, urgent economic, political and social action is needed to keep Earth's environment within safe planetary boundaries[171].

Big Earth Data, especially social sensing data, have brought new opportunities to effectively capture and quantify human dynamics and their impacts on the environment[172]. However, quantifying, analysing and discovering the commonly unstructured and highly uncertain social sensing data are bringing about unprecedented complexity and resource demands. Innovative solutions, such as the human–natural system[173,174], need to be explored to effectively use the data for modelling the Anthropocene[174]. Moreover, metasynthesis or soft systems methodology are qualitative–quantitative frameworks that have emerged to address complex unstructured problems[175], such as ecosystem management[176] and climate change policy[177]. Exploring and further developing these innovative methods to seamlessly integrate hard data mainly from natural components and soft data from social components could help to solve the methodological dilemma of unstructured or ill-structured social data[178,179]. These advances will contribute to capturing human dynamics in the digital twin of Earth.

Cultivating an innovative Big Data culture

Broad adaptation to Big Data in ESS requires knowledge and ability beyond individual scientists. A smooth transition involves promoting awareness of open data and sharing principles, and close collaboration between geoscientists, data scientists, computational scientists and engineers.

As the influence of Big Data and artificial intelligence continues to grow, the community needs to recognize the challenges of promoting Big Data across different fields of Earth sciences and address them by promoting talent training and cultivation. Agencies such as the National Science Foundation (NSF) in the United States and funders in China[180] have begun to take action by funding Big Data projects featuring interdisciplinary research. Projects dedicated to the digital twin of Earth, such as the European Union's Destination Earth[17] and the Earth-2 simulation, have also been established. These pioneer projects are ambitious and still far from realizing a digital replica of the Earth system, but they are exploring the pathway towards it.

Big Data is known for its high demand for computational infrastructure and resources. Powerful and sophisticated cyberinfrastructures, such as dedicated data centres and exascale machines capable of

1 billion billion calculations per second, need to be built to support Big Data applications by aiding the sharing, discovery and integration of data and computational resources[28]. Data science platforms, such as Google Earth Engine[181] and Kaggle[182], with their capacities for data and analysis exposed through user-friendly interfaces, can have a valuable role in promoting scientific problems and bringing all machine learning communities together to work on global problems.

The effectiveness of these cyberinfrastructures would be maximized by adopting international open standards and principles, such as the Findability, Accessibility, Interoperability and Reuse (FAIR) principle[183], which ensures that data are usable by both humans and machines by enduringly preserving openly accessible data. A concise and measurable set of these principles can promote harmonized and standardized Big Earth Data to improve data quality and usefulness, and to transform the state-of-practice for data governance of Big Earth Data to benefit Earth system modelling and reanalysis[184]. The adoption of these data principles will also advocate open science ecosystems[28], benefiting Big Data practice in the long term[185,186]. These efforts will build a sustainable foundation of talent and cyberinfrastructure for Big Data innovation in ESS, which is important for the long journey towards digital twins of Earth.

Summary and future directions

Emerging Big Data are changing scientific research. Earth system science in particular is responding strongly to the increasing volume of data and rapid advance of Big Data analytics. Developments in Big Earth Data are fundamental for driving Earth science into the next phase of the digital era, marked by the implementation of a digital twin of Earth. Big Earth Data will empower simulation of complex and interconnected Earth systems, including human dynamics, eventually providing the ability to predict and monitor the impacts of sustainable decision-making. With support from rapidly developing physical knowledge, machine intelligence and cyberinfrastructure, the Earth's digital twin has been introduced as a computational tool with the ability to simulate and forecast the Earth system as a whole at real time and fine scales, promoting unprecedented accuracy in simulating complex Earth events and phenomena, particularly extreme weather.

BDA is a promising technique to enhance reanalysis and prediction precisions by assimilating Big Earth Data into ultrahigh-resolution Earth system models. By deeply fusing all categories of Big Earth Data to support coupled modelling of nature–human systems, BDA can act as a control layer of the digital twin of Earth. However, current BDA techniques are time-consuming and computationally expensive, meaning that most BDA studies are currently restricted to local-scale prototypes. The synergy of advances in supercomputing technology and integration with deep learning could reduce the computational cost and allow expansion to regional or global scales.

From a data-driven perspective, four frontier Big Data analytics methods – deep learning, physics-informed machine learning, causal inference and deep reinforcement learning – are particularly promising for ESS applications. The intelligence of ESS could be enhanced by machine learning, including the models' learning ability, transferability, interpretability and decision-making ability. To overcome the limitations of current machine learning models, new learning architectures should be designed specifically for the uniqueness of Earth science applications – in particular, new deep learning frameworks that can accommodate the complex, interactive and multiscale characteristics of the Earth system and also be able to integrate physics-based models, to eventually become the AlphaFold[187] for ESS. Advances in the interpretability of machine learning could promote AI-aided hypotheses in ESS, enabling not only intelligent learning from data but also creativity derived from new theorizations, which will aid the realization of Earth's digital twin.

Many Big Data techniques are becoming mainstream methodologies in ESS. Future research must pursue comprehension of human–natural coevolution system modelling at a planetary scale, so that simulations of socioeconomic activities, sustainable development and climate targets can be tested virtually before they are implemented in the real world. To aid implementation of these digital twins that can examine scenarios of prediction and decision-making at different timescales and spatial scales, we suggest that researchers should prioritize organizing and normalizing social sensing to improve data collection on human dynamics, and as part of this, building a core digital platform to help explore integration of Big Data assimilation and advance its analytics. These actions will help the realization of an operational Earth–human observation system.

The shift towards digital twins of Earth will be a long and difficult journey. Embracing the deluge of data and recognizing the challenges of Big Data analytical techniques can only be addressed through extensive interdisciplinary and transdisciplinary collaborations, and a fair data culture.

Published online: 2 May 2023

References

1. Yang, C. et al. Big Earth Data analytics: a survey. *Big Earth Data* **3**, 83–107 (2019).
2. Baldocchi, D. et al. FLUXNET: a new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. *Bull. Am. Meteorol. Soc.* **82**, 2415–2434 (2001).
3. Liu, Y. et al. Social sensing: a new approach to understanding our socioeconomic environments. *Ann. Assoc. Am. Geogr.* **105**, 512–530 (2015).
4. Whitcraft, A. K. et al. No pixel left behind: toward integrating Earth observations for agriculture into the United Nations Sustainable Development Goals framework. *Remote Sens. Environ.* **235**, 111470 (2019).
5. Graham, M. & Shelton, T. Geography and the future of Big Data, Big Data and the future of geography. *Dialogues Hum. Geogr.* **3**, 255–261 (2013).
6. Eyring, V. et al. Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geosci. Model. Dev.* **9**, 1937–1958 (2016).
7. Hey, T., Tansley, S., Tolle, K. & Gray, J. *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Microsoft Research, 2009).
8. Kitchin, R. Big Data, new epistemologies and paradigm shifts. *Big Data Soc.* **1**, 2053951714528481 (2014).
9. Reichstein, M. et al. Deep learning and process understanding for data-driven Earth system science. *Nature* **566**, 195–204 (2019).
   Provides a comprehensive overview of deep learning for Earth system science.
10. Grieves, M. Digital twin: manufacturing excellence through virtual factory replication. *White Paper* **1**, 1–7 (2014).
11. Barricelli, B. R., Casiraghi, E. & Fogli, D. A survey on digital twin: definitions, characteristics, applications, and design implications. *IEEE Access.* **7**, 167653–167671 (2019).
12. Raj, P. in *Advances in Computers* Vol. 121, 267–283 (Elsevier, 2021).
13. Rasheed, A., San, O. & Kvamsdal, T. Digital twin: values, challenges and enablers from a modeling perspective. *IEEE Access.* **8**, 21980–22012 (2020).
14. Abdeen, F. N. & Sepasgozar, S. M. E. City digital twin concepts: a vision for community participation. *Environ. Sci. Proc.* **12**, 19 (2022).
15. Liu, Y. K., Ong, S. K. & Nee, A. Y. C. State-of-the-art survey on digital twin implementations. *Adv. Manuf.* **10**, 1–23 (2022).
16. Tao, F., Zhang, H., Liu, A. & Nee, A. Y. C. Digital twin in industry: state-of-the-art. *IEEE Trans. Ind. Inform.* **15**, 2405–2415 (2019).
17. Bauer, P., Stevens, B. & Hazeleger, W. A digital twin of Earth for the green transition. *Nat. Clim. Chang.* **11**, 80–83 (2021).
    Provided a conceptual framework of the digital twin of Earth.
18. Voosen, P. Europe builds 'digital twin' of Earth to hone climate forecasts. *Science* **370**, 16–17 (2020).
19. Bauer, P. et al. The digital revolution of Earth-system science. *Nat. Comput. Sci.* **1**, 104–113 (2021).
    Discussed the revolution in digital Earth systems and proposed the concept of an efficient software infrastructure for the Earth-system digital twin.

20. Latif, M. The roadmap of climate models. *Nat. Comput. Sci.* **2**, 536–538 (2022).
21. Schellnhuber, H. J. 'Earth system' analysis and the second Copernican revolution. *Nature* **402**, C19–C23 (1999).
22. Steffen, W. et al. The emergence and evolution of Earth system science. *Nat. Rev. Earth Environ.* **1**, 54–63 (2020).
23. Hinton, G. E., Osindero, S. & Teh, Y.-W. A fast learning algorithm for deep belief nets. *Neural Comput.* **18**, 1527–1554 (2006).
24. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507 (2006).
25. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
26. Kaelbling, L. P., Littman, M. L. & Moore, A. W. Reinforcement learning: a survey. *J. Artif. Int. Res.* **4**, 237–285 (1996).
27. Mousavi, S. M. & Beroza, G. C. Deep-learning seismology. *Science* **377**, eabm4470 (2022).
28. 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).
Gave a comprehensive overview of the state of machine learning in the solid Earth geosciences and solutions to broaden and accelerate these capabilities.
29. Herman, L. et al. A comparison of monoscopic and stereoscopic 3D visualizations: Effect on spatial planning in digital twins. *Remote Sens.* **13**, 2976 (2021).
30. Jiang, P. et al. Digital twin Earth — Coasts: developing a fast and physics-informed surrogate model for coastal floods via neural operators. Preprint at https://doi.org/10.48550/arXiv.2110.07100 (2021).
31. Tao, F. et al. Digital twin-driven product design, manufacturing and service with Big Data. *Int. J. Adv. Manuf. Technol.* **94**, 3563–3576 (2018).
32. Keith, D. W. Geoengineering. *Nature* **409**, 420–420 (2001).
33. Lawrence, M. G. et al. Evaluating climate geoengineering proposals in the context of the Paris Agreement temperature goals. *Nat. Commun.* **9**, 3734 (2018).
34. Parson, E. A. Geoengineering: symmetric precaution. *Science* **374**, 795–795 (2021).
35. Armstrong McKay, D. I. et al. Exceeding 1.5 °C global warming could trigger multiple climate tipping points. *Science* **377**, eabn7950 (2022).
36. Rockström, J. et al. A safe operating space for humanity. *Nature* **461**, 472–475 (2009).
37. Oza, N. et al. *NASA Earth Science Technology for Earth System Digital Twins (ESDT)* https://essopenarchive.org/doi/full/10.1002/essoar.10509965.1 (ESS Open Archive, 2022).
38. Yang, C., Raskin, R., Goodchild, M. & Gahegan, M. Geospatial cyberinfrastructure: past, present and future. *Comput. Environ. Urban. Syst.* **34**, 264–277 (2010).
39. Dax, G., Nagarajan, S., Li, H. & Werner, M. Compression supports spatial deep learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **16**, 702–713 (2023).
40. Reed, D. A. & Dongarra, J. Exascale computing and Big Data. *Commun. ACM* **58**, 56–68 (2015).
41. Mystakidis, S. Metaverse. *Encyclopedia* **2**, 486–497 (2022).
42. Guo, H., Chen, F., Sun, Z., Liu, J. & Liang, D. Big Earth Data: a practice of sustainability science to achieve the sustainable development goals. *Sci. Bull.* **66**, 1050–1053 (2021).
43. Li, X., Liu, F. & Fang, M. Harmonizing models and observations: data assimilation in Earth system science. *Sci. China Earth Sci* **63**, 1059–1068 (2020).
44. Gettelman, A. et al. The future of Earth system prediction: advances in model–data fusion. *Sci. Adv.* **8**, eabn3488 (2022).
45. Carrassi, A., Bocquet, M., Bertino, L. & Evensen, G. Data assimilation in the geosciences: an overview of methods, issues, and perspectives. *WIREs Clim. Change* **9**, e535 (2018).
46. Hewitt, H., Fox-Kemper, B., Pearson, B., Roberts, M. & Klocke, D. The small scales of the ocean may hold the key to surprises. *Nat. Clim. Chang.* **12**, 496–499 (2022).
47. Schneider, T. et al. Climate goals and computing the future of clouds. *Nat. Clim. Change* **7**, 3–5 (2017).
48. Stevens, B. et al. DYAMOND: the DYnamics of the Atmospheric general circulation modeled on non-hydrostatic domains. *Prog. Earth Planet. Sci.* **6**, 61 (2019).
49. Miyoshi, T., Kondo, K. & Imamura, T. The 10,240-member ensemble kalman filtering with an intermediate agcm. *Geophys. Res. Lett.* **41**, 5264–5271 (2014).
50. Ruiz, J., Lien, G.-Y., Kondo, K., Otsuka, S. & Miyoshi, T. Reduced non-Gaussianity by 30 s rapid update in convective-scale numerical weather prediction. *Nonlinear Process Geophys.* **28**, 615–626 (2021).
51. Honda, T. et al. Development of the real-time 30-s-update Big Data assimilation system for convective rainfall prediction with a phased array weather radar: description and preliminary evaluation. *J. Adv. Model. Earth Syst.* **14**, e2021MS002823 (2022).
52. Mass, C. F. & Madaus, L. E. Surface pressure observations from smartphones: a potential revolution for high-resolution weather prediction? *Bull. Am. Meteorol. Soc.* **95**, 1343–1349 (2014).
53. Li, R. et al. Smartphone pressure data: quality control and impact on atmospheric analysis. *Atmos. Meas. Tech.* **14**, 785–801 (2021).
54. Avellaneda, P. M., Ficklin, D. L., Lowry, C. S., Knouft, J. H. & Hall, D. M. Improving hydrological models with the assimilation of crowdsourced data. *Water Resour. Res.* **56**, e2019WR026325 (2020).
55. Sawada, Y. & Hanazaki, R. Socio-hydrological data assimilation: analyzing human–flood interactions by model–data integration. *Hydrol. Earth Syst. Sci.* **24**, 4777–4791 (2020).
56. Barendrecht, M. H. et al. The value of empirical data for estimating the parameters of a sociohydrological flood risk model. *Water Resour. Res.* **55**, 1312–1336 (2019).
57. Jonathan, W., Evans, A. J. & Malleson, N. S. Dynamic calibration of agent-based models using data assimilation. *R. Soc. Open Sci.* **3**, 150703 (2016).
58. Boukabara, S.-A. et al. Outlook for exploiting artificial intelligence in the Earth and environmental sciences. *Bull. Am. Meteorol. Soc.* **102**, 1–53 (2021).
59. Geer, A. J. Learning earth system models from observations: machine learning or data assimilation? *Phil. Trans. R. Soc. A* **379**, 20200089 (2021).
60. Buizza, C. et al. Data learning: integrating data assimilation and machine learning. *J. Comput. Sci.* **58**, 101525 (2022).
61. Pathiraja, S., Moradkhani, H., Marshall, L., Sharma, A. & Geenens, G. Data-driven model uncertainty estimation in hydrologic data assimilation. *Water Resour. Res.* **54**, 1252–1280 (2018).
62. Zhang, Q. et al. A dynamic data-driven method for dealing with model structural error in soil moisture data assimilation. *Adv. Water Resour.* **132**, 103407 (2019).
63. King, F., Erler, A. R., Frey, S. K. & Fletcher, C. G. Application of machine learning techniques for regional bias correction of snow water equivalent estimates in Ontario, Canada. *Hydrol. Earth Syst. Sci.* **24**, 4887–4902 (2020).
64. Barthélémy, S., Brajard, J., Bertino, L. & Counillon, F. Super-resolution data assimilation. *Ocean Dyn.* **72**, 661–678 (2022).
65. Cheng, S. et al. Generalised latent assimilation in heterogeneous reduced spaces with machine learning surrogate models. *J. Sci. Comput.* **94**, 11 (2022).
66. Cheng, S. et al. Data-driven surrogate model with latent data assimilation: application to wildfire forecasting. *J. Comput. Phys.* **464**, 111302 (2022).
67. Lorenz, E. N. Designing chaotic models. *J. Atmos. Sci.* **62**, 1574–1587 (2005).
68. Bonavita, M. et al. Machine learning for Earth system observation and prediction. *Bull. Am. Meteorol. Soc.* **102**, E710–E716 (2021).
69. Kong, Q. et al. Machine learning in seismology: turning data into insights. *Seismol. Res. Lett.* **90**, 3–14 (2018).
70. Lary, D. J., Alavi, A. H., Gandomi, A. H. & Walker, A. L. Machine learning in geosciences and remote sensing. *Geosci. Front.* **7**, 3–10 (2016).
71. Tahmasebi, P., Kamrava, S., Bai, T. & Sahimi, M. Machine learning in geo- and environmental sciences: from small to large scale. *Adv. Water Resour.* **142**, 103619 (2020).
72. Feng, M. & Li, X. Land cover mapping toward finer scales. *Sci. Bull.* **65**, 1604–1606 (2020).
73. Yu, S. & Ma, J. Deep learning for geophysics: current and future trends. *Rev. Geophys.* https://doi.org/10.1029/2021RG000742 (2021).
74. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
75. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
76. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
77. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
78. Goodfellow, I. et al. Generative adversarial networks. *Commun. ACM* **63**, 139–144 (2020).
79. Scher, S. Toward data-driven weather and climate forecasting: approximating a simple general circulation model with deep learning. *Geophys. Res. Lett.* **45**, 616–12,622 (2018).
80. Ma, L. et al. Deep learning in remote sensing applications: a meta-analysis and review. *ISPRS J. Photogramm. Remote Sens.* **152**, 166–177 (2019).
81. Ravuri, S. et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature* **597**, 672–677 (2021).
Proposed a deep generative adversarial network model for faster and more accurate precipitation nowcasting from historical radar data.
82. Zhong, Y. et al. WHU-Hi: UAV-borne hyperspdectral with high spatial resolution (H2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF. *Remote Sens. Environ.* **250**, 112012 (2020).
83. Hong, D. et al. More diverse means better: multimodal deep learning meets remote sensing imagery classification. *IEEE Trans. Geosci. Remote Sens.* **59**, 4340–4354 (2020).
84. Huang, L., Luo, J., Lin, Z., Niu, F. & Liu, L. Using deep learning to map retrogressive thaw slumps in the Beiluhe region (Tibetan Plateau) from CubeSat images. *Remote Sens. Environ.* **237**, 111534 (2020).
85. Chi, J., Kim, H., Lee, S. & Crawford, M. M. Deep learning based retrieval algorithm for Arctic sea ice concentration from AMSR2 passive microwave and MODIS optical data. *Remote Sens. Environ.* **231**, 111204 (2019).
86. Crane-Droesch, A. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environ. Res. Lett.* **13**, 114003 (2018).
87. Korup, O. & Stolle, A. Landslide prediction from machine learning. *Geol. Today* **30**, 26–33 (2014).
88. Shen, C. A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resour. Res.* **54**, 8558–8593 (2018).
89. Kochanski, K., Mohan, D., Horrall, J., Rountree, B. & Abdulla, G. Deep learning predictions of sand dune migration. Preprint at https://doi.org/10.48550/arXiv.1912.10798 (2019).
90. Leinonen, J., Nerini, D. & Berne, A. Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. *IEEE Trans. Geosci. Remote Sens.* **59**, 7211–7223 (2021).
91. Li, Z., Meier, M.-A., Hauksson, E., Zhan, Z. & Andrews, J. Machine learning seismic wave discrimination: application to earthquake early warning. *Geophys. Res. Lett.* **45**, 4773–4779 (2018).
92. Wang, B., Zhang, N., Lu, W. & Wang, J. Deep-learning-based seismic data interpolation: a preliminary result. *Geophysics* **84**, V11–V20 (2019).
93. Wang, N., Zhang, D., Chang, H. & Li, H. Deep learning of subsurface flow via theory-guided neural network. *J. Hydrol.* **584**, 124700 (2020).

94. Zhou, Z.-H. A brief introduction to weakly supervised learning. *Natl Sci. Rev.* **5**, 44–53 (2018).
95. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *Proc. of the 37th International Conference on Machine Learning* 1597–1607 (ICML, 2020).
96. Chen, Y. & Bruzzone, L. Self-supervised change detection in multi-view remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–12 (2022).
97. Jung, H., Oh, Y., Jeong, S., Lee, C. & Jeon, T. Contrastive self-supervised learning with smoothed representation for remote sensing. *IEEE Geosci. Remote Sens. Lett.* **19**, 1–5 (2022).
98. Vidal, R., Bruna, J., Giryes, R. & Soatto, S. Mathematics of deep learning. Preprint at https://doi.org/10.48550/arXiv.1712.04741 (2017).
99. Rackauckas, C. et al. Universal differential equations for scientific machine learning. Preprint at https://doi.org/10.48550/arXiv.2001.04385 (2021).
100. Marcus, G. Deep learning: a critical appraisal. Preprint at https://doi.org/10.48550/arXiv.1801.00631 (2018).
101. Rice, L., Wong, E. & Kolter, J. Z. Overfitting in adversarially robust deep learning. In *Proc. of the 37th International Conference on Machine Learning* 8093–8104 (ICML, 2020).
102. Karniadakis, G. E. et al. Physics-informed machine learning. *Nat. Rev. Phys.* **3**, 422–440 (2021).
Provides a comprehensive overview for embedding physics-based knowledge into machine learning.
103. 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).
104. Kashinath, K. et al. Physics-informed machine learning: case studies for weather and climate modelling. *Phil. Trans. R. Soc. A* **379**, 20200093 (2021).
105. Zhao, W. L. et al. Physics-constrained machine learning of evapotranspiration. *Geophys. Res. Lett.* **46**, 14496–14507 (2019).
106. Huanfeng, S. & Liangpei, Z. Mechanism-learning coupling paradigms for parameter inversion and simulation in Earth surface systems. *Sci. China Earth Sci.* **66**, 568–582 (2023).
107. Jia, X. et al. Physics-guided machine learning for scientific discovery: an application in simulating lake temperature profiles. *ACM/IMS Trans. Data Sci.* **2**, 1–20 (2021).
108. Daw, A., Karpatne, A., Watkins, W., Read, J. & Kumar, V. Physics-guided neural networks (PGNN): an application in lake temperature modeling. Preprint at https://doi.org/10.48550/arXiv.1710.11431 (2021).
109. Sturm, P. O. & Wexler, A. S. Conservation laws in a neural network architecture: enforcing the atom balance of a Julia-based photochemical model (v0.2.0). *Geosci. Model. Dev.* **15**, 3417–3431 (2022).
110. Beucler, T. et al. Enforcing analytic constraints in neural networks emulating physical systems. *Phys. Rev. Lett.* **126**, 098302 (2021).
111. Read, J. S. et al. Process-guided deep learning predictions of lake water temperature. *Water Resour. Res.* **55**, 9173–9190 (2019).
112. Karniadakis, G. E. et al. Physics-informed machine learning. *Nat. Rev. Phys.* **3**, 422–440 (2021).
113. Aldrich, J. Correlations genuine and spurious in Pearson and Pule. *Stat. Sci.* **10**, 364–376 (1995).
114. Altman, N. & Krzywinski, M. Association, correlation and causation. *Nat. Methods* **12**, 899–900 (2015).
115. Schölkopf, B. in *Probabilistic and Causal Inference: The Works of Judea Pearl* Vol. 36, 765–804 (Association for Computing Machinery, 2022).
116. Pearl, J. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM* **62**, 54–60 (2019).
117. Cui, P. & Athey, S. Stable learning establishes some common ground between causal inference and machine learning. *Nat. Mach. Intell.* **4**, 110–115 (2022).
118. Runge, J. et al. Inferring causation from time series in Earth system sciences. *Nat. Commun.* **10**, 2553 (2019).
119. van Nes, E. H. et al. Causal feedbacks in climate change. *Nat. Clim. Change* **5**, 445–448 (2015).
120. Zhang, K., Schölkopf, B., Spirtes, P. & Glymour, C. Learning causality and causality-related learning: Some recent progress. *Natl Sci. Rev.* **5**, 26–29 (2018).
121. Salvucci, G. D., Saleem, J. A. & Kaufmann, R. Investigating soil moisture feedbacks on precipitation with tests of Granger causality. *Adv. Water Resour.* **25**, 1305–1312 (2002).
122. Tuttle, S. E. & Salvucci, G. D. Confounding factors in determining causal soil moisture–precipitation feedback. *Water Resour. Res.* **53**, 5531–5544 (2017).
123. Jiang, B., Liang, S. & Yuan, W. Observational evidence for impacts of vegetation change on local surface climate over northern China using the Granger causality test. *J. Geophys. Res. Biogeosci.* **120**, 1–12 (2015).
124. Papagiannopoulou, C. et al. A non-linear Granger-causality framework to investigate climate–vegetation dynamics. *Geosci. Model. Dev.* **10**, 1945–1960 (2017).
125. Kretschmer, M. et al. Quantifying causal pathways of teleconnections. *Bull. Am. Meteorol. Soc.* **102**, E2247–E2263 (2021).
126. Kretschmer, M., Coumou, D., Donges, J. F. & Runge, J. Using causal effect networks to analyze different arctic drivers of midlatitude winter circulation. *J. Clim.* **29**, 4069–4081 (2016).
127. Sugihara, G. et al. Detecting causality in complex ecosystems. *Science* **338**, 496–500 (2012).
128. Yang, A. C., Peng, C.-K. & Huang, N. E. Causal decomposition in the mutual causation system. *Nat. Commun.* **9**, 3378 (2018).
129. Wang, J.-Y., Kuo, T.-C. & Hsieh, C. Causal effects of population dynamics and environmental changes on spatial variability of marine fishes. *Nat. Commun.* **11**, 2635 (2020).
130. An, W., Beauvile, R. & Rosche, B. Causal network analysis. *Annu. Rev. Sociol.* **48**, 23–41 (2022).
131. Moraffah, R. et al. Causal inference for time series analysis: problems, methods and evaluation. *Knowl. Inf. Syst.* **63**, 3041–3085 (2021).
132. Runge, J. et al. Identifying causal gateways and mediators in complex spatio-temporal systems. *Nat. Commun.* **6**, 8502 (2015).
133. Bareinboim, E. & Pearl, J. Causal inference and the data-fusion problem. *Proc. Natl Acad. Sci. USA* **113**, 7345–7352 (2016).
134. Rubin, D. B. Causal inference using potential outcomes: design, modeling, decisions. *J. Am. Stat. Assoc.* **100**, 322–331 (2005).
135. Crutzen, P. J. Albedo enhancement by stratospheric sulfur injections: a contribution to resolve a policy dilemma? *Clim. Change* **77**, 211 (2006).
136. Gupta, V. & Jain, M. K. Unravelling the teleconnections between ENSO and dry/wet conditions over India using nonlinear Granger causality. *Atmos. Res.* **247**, 105168 (2021).
137. Silva, F. N. et al. Detecting climate teleconnections with granger causality. *Geophys. Res. Lett.* **48**, e2021GL094707 (2021).
138. Wallace, J. M. & Gutzler, D. S. Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Mon. Weather. Rev.* **109**, 784–812 (1981).
139. Runge, J., Nowack, P., Kretschmer, M., Flaxman, S. & Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Sci. Adv.* **5**, eaau4996 (2019).
Ilustrated the capabilities of multivariate causal discovery techniques in a large-scale analysis of the nonlinear global climatic system.
140. Hannart, A., Pearl, J., Otto, F. E. L., Naveau, P. & Ghil, M. Causal counterfactual theory for the attribution of weather and climate-related events. *Bull. Am. Meteorol. Soc.* **97**, 99–110 (2016).
141. Nowack, P., Runge, J., Eyring, V. & Haigh, J. D. Causal networks for climate model evaluation and constrained projections. *Nat. Commun.* **11**, 1415 (2020).
142. Luo, Y., Peng, J. & Ma, J. When causal inference meets deep learning. *Nat. Mach. Intell.* **2**, 426–427 (2020).
143. Degai, T. S. & Petrov, A. N. Rethinking Arctic sustainable development agenda through indigenizing UN sustainable development goals. *Int. J. Sustain. Dev. World Ecol.* **28**, 518–523 (2021).
144. Schrittwieser, J. et al. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature* **588**, 604–609 (2020).
145. Silver, D. et al. Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).
146. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
147. Sun, W., Bocchini, P. & Davison, B. D. Applications of artificial intelligence for disaster management. *Nat. Hazards* **103**, 2631–2689 (2020).
148. Sun, A. Y. Optimal carbon storage reservoir management through deep reinforcement learning. *Appl. Energy* **278**, 115660 (2020).
149. Wu, J., Tao, R., Zhao, P., Martin, N. F. & Hovakimyan, N. Optimizing nitrogen management with deep reinforcement learning and crop simulations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* 1711–719 (CVPRW, 2022).
150. Alibabaei, K., Gaspar, P. D., Assunção, E., Alirezazadeh, S. & Lima, T. M. Irrigation optimization with a deep reinforcement learning model: case study on a site in Portugal. *Agric. Water Manag.* **263**, 107480 (2022).
151. Chen, M. et al. A reinforcement learning approach to irrigation decision-making for rice using weather forecasts. *Agric. Water Manag.* **250**, 106838 (2021).
152. Zhou, N. Intelligent control of agricultural irrigation based on reinforcement learning. *J. Phys. Conf. Ser.* **1601**, 052031 (2020).
153. Strnad, F. M., Barfuss, W., Donges, J. F. & Heitzig, J. Deep reinforcement learning in World-Earth system models to discover sustainable management strategies. *Chaos* **29**, 123122 (2019).
Demonstrated the first attempt to identify sustainable management strategies by combining deep reinforcement learning with Earth system models.
154. Wang, X. et al. Efficient reservoir management through deep reinforcement learning. Preprint at https://doi.org/10.48550/arXiv.2012.03822 (2020).
155. Mullapudi, A., Lewis, M. J., Gruden, C. L. & Kerkez, B. Deep reinforcement learning for the real time control of stormwater systems. *Adv. Water Resour.* **140**, 103600 (2020).
156. Tian, W., Liao, Z., Zhi, G., Zhang, Z. & Wang, X. Combined sewer overflow and flooding mitigation through a reliable real-time control based on multi-reinforcement learning and model predictive control. *Water Resour. Res.* **58**, e2021WR030703 (2022).
157. Gronauer, S. & Diepold, K. Multi-agent deep reinforcement learning: a survey. *Artif. Intell. Rev.* **55**, 895–943 (2022).
158. Hernandez-Leal, P., Kartal, B. & Taylor, M. E. A survey and critique of multiagent deep reinforcement learning. *Auton. Agent. Multi-Agent Syst.* **33**, 750–797 (2019).
159. Nguyen, T. T., Nguyen, N. D. & Nahavandi, S. Deep reinforcement learning for multiagent systems: a review of challenges, solutions, and applications. *IEEE Trans. Cybern.* **50**, 3826–3839 (2020).
160. Hung, F. & Yang, Y. C. E. Assessing adaptive irrigation impacts on water scarcity in nonstationary environments — a multi-agent reinforcement learning approach. *Water Resour. Res.* **57**, e2020WR029262 (2021).

161. Galesic, M. et al. Human social sensing is an untapped resource for computational social science. *Nature* **595**, 214–222 (2021).
162. Shmueli, E., Singh, V. K., Lepri, B. & Pentland, A. Sensing, understanding, and shaping social behavior. *IEEE Trans. Comput. Soc. Syst.* **1**, 22–34 (2014).
163. An, L. Modeling human decisions in coupled human and natural systems: review of agent-based models. *Ecol. Model.* **229**, 25–36 (2012).
164. Zhu, R., Hou, Z., Guo, Z. & Wan, B. Summary of "The past, present and future of the habitable Earth: development strategy of Earth science". *Chin. Sci. Bull.* **66**, 4485–4490 (2021).
165. Zhu, R., Zhao, G., Xiao, W., Chen, L. & Tang, Y. Origin, accretion, and reworking of continents. *Rev Geophys.* **59**, e2019RG000689 (2021).
166. Fan, J. et al. A high-resolution summary of Cambrian to Early Triassic marine invertebrate biodiversity. *Science* **367**, 272 (2020).
167. Wang, C. et al. The deep-time digital Earth program: data-driven discovery in geosciences. *Natl Sci. Rev.* **8**, nwab027 (2021).
A review of the current fundamental challenges of data-driven discoveries in the understanding of Earth's evolution in deep time.
168. Lewis, S. L. & Maslin, M. A. Defining the Anthropocene. *Nature* **519**, 171–180 (2015).
169. Ritchie, P. D. L., Clarke, J. J., Cox, P. M. & Huntingford, C. Overshooting tipping point thresholds in a changing climate. *Nature* **592**, 517–523 (2021).
170. Keys, P. W. et al. Anthropocene risk. *Nat. Sustain.* **2**, 667–673 (2019).
171. Otto, I. M. et al. Social tipping dynamics for stabilizing Earth's climate by 2050. *Proc. Natl Acad. Sci. USA* **117**, 2354–2365 (2020).
172. Guo, H. et al. Measuring and evaluating SDG indicators with Big Earth Data. *Sci. Bull.* **67**, 1792–1801 (2022).
173. Fu, B. & Li, Y. Bidirectional coupling between the Earth and human systems is essential for modeling sustainability. *Natl Sci. Rev.* **3**, 397–398 (2016).
174. Liu, J. et al. Complexity of coupled human and natural systems. *Science* **317**, 1513–1516 (2007).
175. Cheng, G. & Li, X. Integrated research methods in watershed science. *Sci. China Earth Sci* **58**, 1159–1168 (2015).
176. DeFries, R. & Nagendra, H. Ecosystem management as a wicked problem. *Science* **356**, 265–270 (2017).
177. Grundmann, R. Climate change as a wicked social problem. *Nat. Geosci.* **9**, 562–563 (2016).
178. Li, X., Zheng, D., Feng, M. & Chen, F. Information geography: the information revolution reshapes geography. *Sci. China Earth Sci* **65**, 379–382 (2022).
179. Rittel, H. W. J. & Webber, M. M. Dilemmas in a general theory of planning. *Policy Sci.* **4**, 155–169 (1973).
180. Huang, Y., Zhang, Y., Youtie, J., Porter, A. L. & Wang, X. How does national scientific funding support emerging interdisciplinary research: a comparison study of Big Data research in the US and China. *PLoS ONE* **11**, e0154509 (2016).
181. Gorelick, N. et al. Google Earth engine: planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **202**, 18–27 (2017).
182. Bojer, C. S. & Meldgaard, J. P. Kaggle forecasting competitions: an overlooked learning opportunity. *Int. J. Forecast.* **37**, 587–603 (2021).
183. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
184. Cannon, M., Kelly, A. & Freeman, C. Implementing an Open & FAIR data sharing policy — a case study in the Earth and environmental sciences. *Learned Publ.* **35**, 56–66 (2022).
185. Li, X. et al. Boosting geoscience data sharing in China. *Nat. Geosci.* **14**, 541–542 (2021).
186. National Academies of Sciences, Engineering, and Medicine. *Open Science by Design: Realizing a Vision for 21st* Century Research (National Academies Press, 2018).
187. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
188. Miyoshi, T. et al. "Big Data assimilation" revolutionizing severe weather prediction. *Bull. Am. Meteorol. Soc.* **97**, 1347–1354 (2016).
Exemplified the ability of Big Data assimilation for faster weather prediction with ultrahigh spatial–temporal resolution.
189. Fan, J., Han, F. & Liu, H. Challenges of Big Data analysis. *Natl Sci. Rev.* **1**, 293–314 (2014).
190. Guo, H. Big Earth Data: A new frontier in Earth and information sciences. *Big Earth Data* **1**, 4–20 (2017).
191. Guo, H. et al. Big Earth Data: a new challenge and opportunity for digital Earth's development. *Int. J. Digital Earth* **10**, 1–12 (2017).
192. Liang, J. & Gamarra, J. G. P. The importance of sharing global forest data in a world of crises. *Sci. Data* **7**, 424 (2020).
193. Klopper, K. B., de Witt, R. N., Bester, E., Dicks, L. M. T. & Wolfaardt, G. M. Biofilm dynamics: linking in situ biofilm biomass and metabolic activity measurements in real-time under continuous flow conditions. *npj Biofilms Microbiomes* **6**, 1–10 (2020).
194. Madaan, A., Sharma, V., Pahwa, P., Das, P. & Sharma, C. in *Big Data Analytics* (eds. Aggarwal, V. B. et al.) 47–54 (Springer, 2018).
195. Li, J. et al. Social media: new perspectives to improve remote sensing for emergency response. *Proc. IEEE* **105**, 1900–1912 (2017).
196. Huang, Z., Qi, H., Kang, C., Su, Y. & Liu, Y. An ensemble learning approach for urban land use mapping based on remote sensing imagery and social sensing data. *Remote Sens.* **12**, 3254 (2020).

Acknowledgements

The authors thank Y. Zen and G. Zhang for comments on the manuscript, X. Tian for suggestions on data assimilation, Y. Bai for suggestions on simulation and reanalysis data, C. Wang and K. Zhang for assistance in preparing the manuscript, Y. Ge and J. Qin for inspiring and improving figures, J. Runge for the PCMCI dataset, P. Bauer for sharing the Destination Earth figure, C. F. Mass and T. Miyoshi for permission to use their data in Fig. 2, and F. M. Strnad for providing the code and data in Fig. 5b. This work was jointly supported by the Strategic Priority Research Program of Chinese Academy of Sciences (XDA19070104) and the National Natural Science Foundation of China (41988101 and 42171140).

Author contributions

X.L. conceptualized the Review. X.L. and M.F. led the discussions and coordinated inputs. X.L. and F.L. contributed the section on Big Data assimilation. Y.R., H.S., J.S., S.Y., Y.S. and C.H. contributed the section on machine and deep learning. M.F. and Q.X. contributed the digital twin section. All authors reviewed the manuscript before submission.

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Reviews Earth & Environment* thanks the anonymous reviewers for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Related links

Copernicus services: https://www.copernicus.eu/en/copernicus-services
Destination Earth: https://digital-strategy.ec.europa.eu/en/policies/destination-earth
Earth-2: https://blogs.nvidia.com/blog/2021/11/12/earth-2-supercomputer
eLTER: https://elter-ri.eu
National Science Foundation of the United States of America: https://www.nsf.gov/cise/bigdata/
Particulate Matter (PM) 2.5 sites in China: https://aqicn.org