Iterative integration of deep learning in hybrid Earth surface system modelling

Min Chen 1,2,3,28, Zhen Qian 1,3,4,28, Niklas Boers5,6,7, Anthony J. Jakeman8, Albert J. Kettner9, Martin Brandt 10, Mei-Po Kwan 11,12, Michael Batty13,14, Wenwen Li15, Rui Zhu16, Wei Luo17, Daniel P. Ames 18, C. Michael Barton19, Susan M. Cuddy20, Sujan Koirala 21, Fan Zhang22, Carlo Ratti23, Jian Liu 1,3,4, Teng Zhong1,3,4, Junzhi Liu24, Yongning Wen1,3,4, Songshan Yue1,3,4, Zhiyi Zhu 1,3,4, Zhixin Zhang 25, Zhuo Sun1,3,4, Jian Lin11, Zaiyang Ma 1,3,4, Yuanqing He 1,3,4, Kai Xu1,3,4, Chunxiao Zhang26, Hui Lin27 & Guonian Lü 1,3,4

Abstract

Earth system modelling (ESM) is essential for understanding past, present and future Earth processes. Deep learning (DL), with the data-driven strength of neural networks, has promise for improving ESM by exploiting information from Big Data. Yet existing hybrid ESMs largely have deep neural networks incorporated only during the initial stage of model development. In this Perspective, we examine progress in hybrid ESM, focusing on the Earth surface system, and propose a framework that integrates neural networks into ESM throughout the modelling lifecycle. In this framework, DL computing systems and ESM-related knowledge repositories are set up in a homogeneous computational environment. DL can infer unknown or missing information, feeding it back into the knowledge repositories, while the ESM-related knowledge can constrain inference results of the DL. By fostering collaboration between ESM-related knowledge and DL systems, adaptive guidance plans can be generated through question-answering mechanisms and recommendation functions. As users interact iteratively, the hybrid system deepens its understanding of their preferences, resulting in increasingly customized, scalable and accurate guidance plans for modelling Earth processes. The advancement of this framework necessitates interdisciplinary collaboration, focusing on explainable DL and maintaining observational data to ensure the reliability of simulations.

Introduction

Earth is a complex, dynamic and adaptive system, with diverse interactions driven by energy, matter and organisms[1]. Human activities are increasingly disturbing the Earth system, for example through releasing greenhouse gases and pollutants and destroying habitats[2]. Understanding these multifaceted and interconnected Earth processes requires integration of observational measurements with physical models of the environment[3]. In particular, identification of the underlying mechanisms and anticipation of potential feedback cycles in Earth systems is essential to more fully understand the impacts of human pressures and how they can be mitigated[4].

Earth system modelling (ESM, Fig. 1) represents a primary tool for characterizing and quantifying the spatiotemporal variations and internal interactions of the Earth across the past, present and future[5–7]. ESMs are composed of a set of physics-based equations that simulate physical, chemical and biological processes within the Earth system, such as carbon and nitrogen cycles, solar radiation dynamics and terrestrial ecosystem dynamics[8–10]. These process-based models merge all aspects of the Earth system together, unlike their predecessors (such as global climate models) that just focused on the atmosphere and oceans. However, owing to their complexity, ESMs are very computationally demanding, time-consuming and expensive. The vast volume of data available has created analytical barriers to ESM research and necessitates the adoption of sophisticated machine learning technologies to streamline processing times and overcome computational bottlenecks[11,12].

Deep learning (DL; Fig. 1) has advanced many research fields, including computer vision, natural language processing (such as ChatGPT[13]) and protein structure prediction[14], as it has the ability to improve the prediction accuracy and computational efficiency of other computational models. It also has the beneficial ability to process multimodal data, which is especially important in Earth sciences, where vast quantities of heterogeneous and noisy raw observational data are gathered on a daily basis (for example satellite data from different sensors, ground-based observations and socioeconomic data)[15,16]. As such, geoscientific applications of deep learning[17–19] have shown potential to address the analytical and computational challenges faced by ESM research[15,20]. However, the data-intensive nature of DL has underlying abstract formulations that are often not visible to the user, and insufficient quantities of labelled and preprocessed machine-readable data[21] can make it challenging for DL models to recognize patterns and generate trustworthy trends. Hence it must be stressed that any deep or machine learning model is only as good as the quality of the input data[22,23], except when they are trained with prior domain expertise and physical principles[20,24].

Hybrid ESM, which combines the strengths of ESM and DL, is a current research trend that is leading to improved emulation of Earth surface processes in high resolution[25,26] (Fig. 1). DL enhances the analysing efficiency of observational data into ESMs to accelerate discovery[27,28]. In addition, hybrid ESMs have also broadened the application scope of DL, such as extraction of information from remote-sensing imagery and prediction of climate variables[15]. However, existing research has primarily focused on combining approaches at the initial model-integration level. Such approaches can have potential for subjective bias towards one system over the other leading to an imbalance between the two systems, potentially impeding their successful integration. Integration over the modelling lifecycle could help to build compatible model deployment that can better understand and solve given tasks.

In this Perspective, we review the development of hybrid Earth surface system modelling (ESSM) and propose a conceptual framework for intelligent ESSM in which DL and human insights are integrated throughout the modelling lifecycle. We focus on the Earth surface system, that is, the interacting system of processes occurring at or near the Earth's surface, such as hydrological, geological, (near-surface) atmospheric, biological and social subsystems[29,30] (Fig. 1). To align with this focus, we narrow the broad concept of ESM to a more specific subset of ESSM. The proposed framework is primarily designed for ESSM, but has the potential for broader applicability in the overall field of ESM. Finally, we envisage future directions toward advancing ESM research through its integration with DL.

Challenges of current Earth surface system modelling

Numerous process-based models have been developed and applied in Earth surface system science throughout the evolution of the geosciences. To analyse more comprehensive issues involving numerous processes, communities have developed a series of integrated ESSMs that are able to depict complicated interactions among multiple subsystems[31]. The scientific lifecycle of ESSM generally has five methodological stages, namely: problem definition and contextualization; data preparation and processing; model development and integration; model evaluation and optimization; and model simulation and application[32,33]. These stages need iterative fine-tuning to ensure that the key modelling processes are incorporated and the purposes or objectives are sufficiently considered[34,35]. Table 1 lists examples of prominent modelling applications in distinct domains. As indicated below, we have identified four notable challenges facing ESSM.

Completeness of understanding problems

To understand the dynamics of the Earth surface system, which exhibit self-organization, emergent and hierarchical properties, we should consider the intrinsic interactions and feedbacks among different subsystems[36,37].

In ESSM, macroscopic problems are often hierarchically decomposed into less complex and more manageable pieces to aid analysis and problem-solving, while underlining the importance of interactions and emergent properties across multiple scales[38,39]. Yet some current methodologies in ESSM, particularly those designed for large-scale simulations, might not fully capture the intrinsic connections among related subsystems, potentially resulting in a reductionist approach[40–42]. Furthermore, these methods could lead to incomplete understanding and computational challenges. Specifically, decomposed subproblems with too few geographical objects (such as landforms, vegetation or rivers) in subsystems might not provide a comprehensive view of the relevant Earth surface states[5,43]. By contrast, those with many geographical objects might not necessarily address the nonlinearity problem effectively and could introduce additional formulating complexities[44,45].

Capability of handling Big Data

A plethora of sensors continue to produce unstructured observational data that capture states, fluxes and interactions of the Earth's surface[46]. These include Earth observation satellites, the global positioning system, in situ observations and social media, and they generate quintillions of bytes every day[47,48]. Although this data availability has created numerous opportunities for ESSM, it has also led to unprecedented technological obstacles, namely, volume, variety, veracity, velocity and value[49,50]. It is generally difficult to fully process the various data sources and further extract deep-level patterns, let alone discover knowledge from them, through conventional ESSM approaches[51,52].

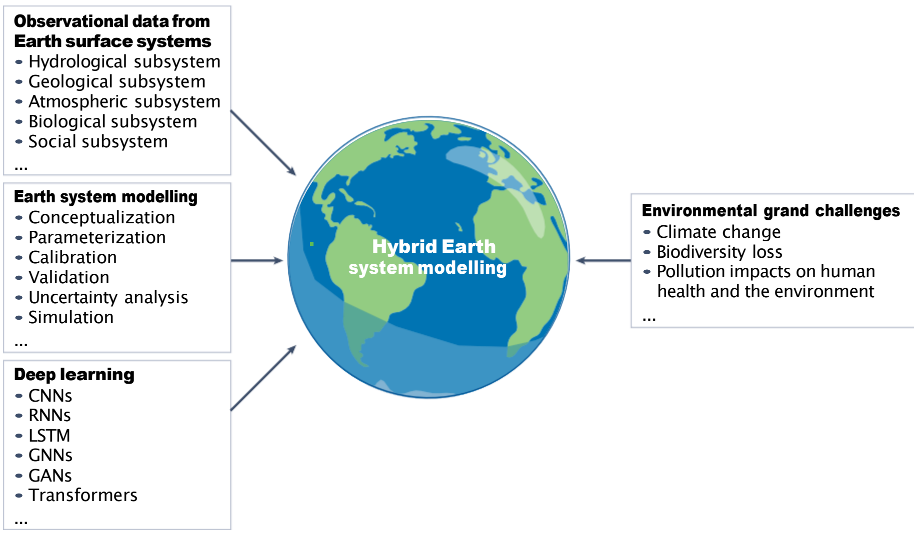

Fig. 1 | Answering the grand environmental challenges through integrating Earth system modelling and deep learning. Earth surface system dynamics and their interactions can be interpreted and simulated through Earth system modelling (ESM). Integration of ESM with deep learning methods increases the predictive power, accuracy and interpretability. These features make hybrid ESMs promising tools to better understand and mitigate the environmental grand challenges such as climate change, biodiversity loss, and pollution impacts on human health and the environment. CNNs, convolutional neural networks; GANs, generative adversarial networks; GNNs, graph neural networks; LSTM, long short-term memory network; RNNs, recurrent neural networks.

Precision of modelling dynamics

The construction of ESSM needs to be fundamentally grounded in established physical principles, such as heat transfer or fluid dynamics, and adheres to well-established scientific knowledge. Nevertheless, expert bias can still have a role in the process[53]. Hence, model architecture and configuration are potentially affected by subjectivity and are prone to bias, errors and unexpected simulation results[54]. This issue could be aggravated when the derived models consist of physical or (semi-)empirical models also encounter the challenging of effectively addressing complex nonlinear dynamics[42,55,56]. Although data assimilation strategies can improve the performance of these models, the pace of creating data frequently far exceeds the ability of models to assimilate it sensibly[20].

Efficiency of computational technology

The computational efficiency of process-based models is crucial, particularly for high-resolution or (near-) real-time modelling (for example, natural disaster assessment), where delays in results and knowledge production could narrow the window of opportunity in decision-making processes[57,58]. In terms of hardware, current ESSM research often relies on multiple central-processing-unit (CPU)-based computers or supercomputers, which have been outperformed by expanding computational demands[59,60]. A three-year study of fine-grained climate simulations on supercomputers shows that graphics processing units (GPUs) outperform CPUs by at least an order of magnitude during high-resolution simulations[61]. Regarding software, ESSM lifecycle processes typically require manual operations or intermediate data transfers, which can impede the computing pipeline. In addition, some models with computationally expensive modules, such as the solution of optimization problems and partial differential equations, necessitate time-intensive iterative simulations.

These four challenges have the potential to disrupt a more complete understanding of Earth system dynamics, dilute the insights gleaned from ever-expanding data reserves, introduce discrepancies and inaccuracies in model constructions and slow down crucial, time-sensitive decisions. To transcend these barriers, it is imperative that current ESSM frameworks are enhanced with forward-thinking methodologies, such as deep learning.

Neural strengths of deep learning

As a specific subfield of artificial intelligence, DL comprises a large class of approaches based on different variations of deep neural network architectures. For example, convolutional neural networks, architectures that focus on local connections through multidimensional convolutions, are often used to extract patterns from various data modalities (for instance, 1D convolutions for sequences, 2D convolutions for images and 3D convolutions for videos)[14]. Recurrent neural networks, particularly those equipped with memory cells known as long short-term memory networks[62], are commonly adept at learning features and long-term dependencies from sequential inputs[63]. More sophisticated networks, such as graph neural networks, generative adversarial networks and transformers, expand the applicability of neural networks beyond relatively specific uses and demonstrate greater flexibility and adaptability for various tasks;[50,64,65] in particular, transformers have been shown to be applicable across diverse purposes with outstanding performance in geoscientific applications, such as modelling spatiotemporal patterns of climate variables[66] and tectonic plate movement[67].

Compared with conventional process-based models, deep neural networks generally exhibit superior prediction performance in terms of fitting observational data[14]. Although it is important to acknowledge that these networks typically have limited interpretability for understanding decision processes[68], with the research community actively working to address these shortcomings, the characteristics of deep learning still pave the way for data-driven discovery of patterns in Earth surface system dynamics. Table 1 contains some existing examples of DL-integrated ESSM options for the different domains. On a broader note, the opportunities that DL brings to mitigate the challenges of ESSM can be seen from four perspectives, as described in the following sections.

Maximum use of multimodal data

Data derived across space and time are often characterized by multimodalities; that is, they are multisource, heterogeneous, unstructured or multitemporal[69]. Integrating information from various modalities into a homogeneous space helps to uncover distinctive characteristics and explain the observed processes[70,71]. Techniques for multimodal data

Table 1 | Examples of conventional Earth system modelling approaches and integrated deep learning options

Domain	Scientific challenge	Conventional ESM examples	DL-integrated options
Hydrological system	Rainfall–runoff simulation	SAC-SMA[154]	MLP[155]
	Groundwater modelling	MODFLOW[156]	CNN-BiLSTM[157] (CNN- and LSTM-based)
Geomorphological system	Soil erosion modelling	WEPP[158]	ANFIS[159] (MLP-based)
	Sediment estimation	SEDD[160]	CNN[161]
Atmospheric system	Air quality assessment	Gaussian plume model[162]	CNN[163]
	Weather prediction	WRF[164]	GNN[65]
Biological system	Forest carbon estimation	SEIB-DGVM[165]	MLP[166]
	Wetland monitoring	WSM[167]	MLP[168]
Social system	Epidemic spread modelling	Susceptible–infected–susceptible model[169]	LSTM[112]
	Human migration simulation	Gravity model[170]	Deep gravity model[171] (MLP-based)

ANFIS, adaptive network-based fuzzy inference system; BiLSTM, bidirectional long short-term memory network; CNN, convolutional neural network; DL, deep learning; ESM, Earth system modelling; GNN, graph neural network; LSTM, long short-term memory network; MLP, multilayer perceptron; MODFLOW, US Geological Survey modular finite-difference flow model; SAC-SMA, Sacramento soil moisture accounting model; SEDD, sediment delivery distributed model; SEIB-DGVM, spatially explicit individual-based dynamic global vegetation model; WEPP, water erosion prediction project model; WRF, weather research and forecasting model; WSM, wetland shrinkage monitoring model.

fusion are numerous. Those techniques that rely heavily on manual encoding with domain-specific expertise inevitably impair the fusion results[72]. In contrast, deep neural networks can adapt to unstructured multimodal data and uncover complicated correlations among them[73]. The ability to tackle the challenges of ESSM using this aspect of DL is a major advantage. For instance, DL-based approaches can fuse the various multimodal data derived from decomposed problems, thereby affording an efficient and comprehensive way to understand Earth's surface processes.

Self-adaptive feature representation

Even data generated by natural laws exhibit considerable uncertainty and high dimensionality[20,74]. To extract information from and understand such data, scientific communities have a strong interest in representing their features. Traditional methods, such as scale-invariant feature transform, term frequency–inverse document frequency and principal component analysis, commonly extract low- or mid-level features and are only suitable for certain workloads[75]. In contrast, DL-based approaches have received considerable attention in geoscientific applications because of the self-adaptive learning mechanism (commonly based on supervised learning and labelled data). Specifically, deep neural networks can reveal patterns and relationships from data, such as interpreting various objects within complex backgrounds in observed images, which can be challenging to formulate using traditional methods and a priori knowledge[76,77]. The feature representation ability of DL aids the extraction of deep-level features without tedious feature engineering. Furthermore, unsupervised or self-supervised approaches can automatically adapt to latent domains in heterogeneous data at a fraction of manual and computational cost[78,79]. Modellers can use pretrained models on public datasets like ImageNet[80] to transition to geoscientific applications, reducing time-consuming labelling efforts.

Superior fitting precision

DL-based approaches perform well in complex Earth system dynamics as universal functional approximators[81]. For example, DL-based forecasting or nowcasting of climate variables (such as precipitation, temperature and humidity) can achieve better results, spatially and temporally, including the exact timing, location and intensity[79,82]. By contrast, traditional models such as optical flow frequently struggle to effectively capture nonlinear climate dynamics (such as moist convection and cloud formation)[82,83], which can be attributed to the separation of internal processes and the presence of non-optimizable parameters[84]. DL has been used in some tasks, such as visual question-answering for geographic scenes[85], synthetic spatiotemporal data generation[86] and extreme weather prediction[87] and has notably improved these tasks' simulation accuracy, which seems impossible for traditional process-based models. All of the preceding examples rely on the ability of deep neural networks to fit with superior precision. There is, however, one caveat to recognize here in that, as with all modelling, the parameterization of deep neural networks depends on the training dataset(s), which greatly affects fitting performance[88,89]. Biases embedded in training data can be encoded into a model, making it essential to consider data quality and the conditions that affect their parameterizations and extracted patterns[90,91].

High inferencing speed

It is undeniable that training deep neural networks requires a substantial amount of time[92], ranging from several hours to multiple weeks. However, the inferencing speed of trained networks can be orders of magnitude faster than conventional process-based models[93], such as numerical methods, which frequently require lengthy simulation durations to yield reliable outcomes on simulating complex dynamics[61,94]. The computational efficiency of these conventional models can be substantially improved with trained networks as a substitute[17]. End-to-end network architecture and parallel computing explain the computational advantage of inferencing. First, end-to-end setups enable networks to learn complex representations of data, from inputs to targets, by feeding given data directly without manual manipulations, thereby being highly beneficial for large-scale simulation[95]. Second, the data in deep neural networks are usually structured as a couple of tensors or matrices, which is suitable for parallel computation[96]. The resulting inferencing speed can be increased by several orders of magnitude with GPUs and TPUs[97].

Leveraging DL-based approaches provides a transformative approach to processing abundant observational data and modelling Earth system dynamics. By integrating DL's strengths, the scientific community can enhance its comprehension of the fundamental mechanisms driving Earth's surface processes, paving the way towards surmounting the outlined challenges in ESSM more effectively.

Integrating ESSM and DL

The integration of ESSM and DL offers a promising avenue for advancing our understanding of Earth surface system dynamics. Although these two approaches are distinct — theory-simulation-driven and data-driven — they complement each other in principle[28]. ESSM offers a strong theoretical foundation for interpreting and representing Earth surface processes but can struggle with handling complex dynamics and feedbacks in the context of big observational data. By contrast, DL excels at extracting information and identifying trends in large datasets. However, it lacks interpretive equations and physical constraints, and its forecasting capabilities for new scenarios are limited, as it relies entirely on previously observed relationships (even complex ones) among variables. Hence, by leveraging the integration of both approaches, hybrid ESSM demonstrates enhanced prediction and interpretability capacities, potentially expediting the discovery of underlying Earth surface system dynamics and interactions[98–100] (see Fig. 2).

Existing hybrid ESSM research primarily focuses on integrating process-based models with deep neural networks during the initial stage of model development and integration in the modelling lifecycle. The main integration modes can be categorized into three fundamental modes: the cascading mode, the parallel mode and the embedding mode (Fig. 3). It is worth noting that complex tasks often require a combination of these fundamental modes.

Cascading mode

The cascading mode consists of a computational pipeline that sequentially runs process-based models and deep neural networks, transmitting intermediate results between them. This cascading mode can be categorized into two cases.

In the first case, the process-based model is executed before the deep neural network (diagram 1 in Fig. 3). Common functions include using a process-based model to generate training data or perform feature engineering for a deep neural network or using the latter to downscale the output variables of the former. For instance, process-based models can filter high-quality samples based on physics-based criteria or construct simulated datasets to train deep neural networks for achieving high prediction accuracy with less ground-truth data[101,102]. Deep neural networks can statistically downscale the coarse outputs of process-based models, crucial for predicting climate variables[103] and reconstructing real-world landscapes[104]. Moreover, deep neural networks could identify attractor states and characterize uncertainty in complex, multidimensional output from process model ensembles, for example from parameter sweeps[105].

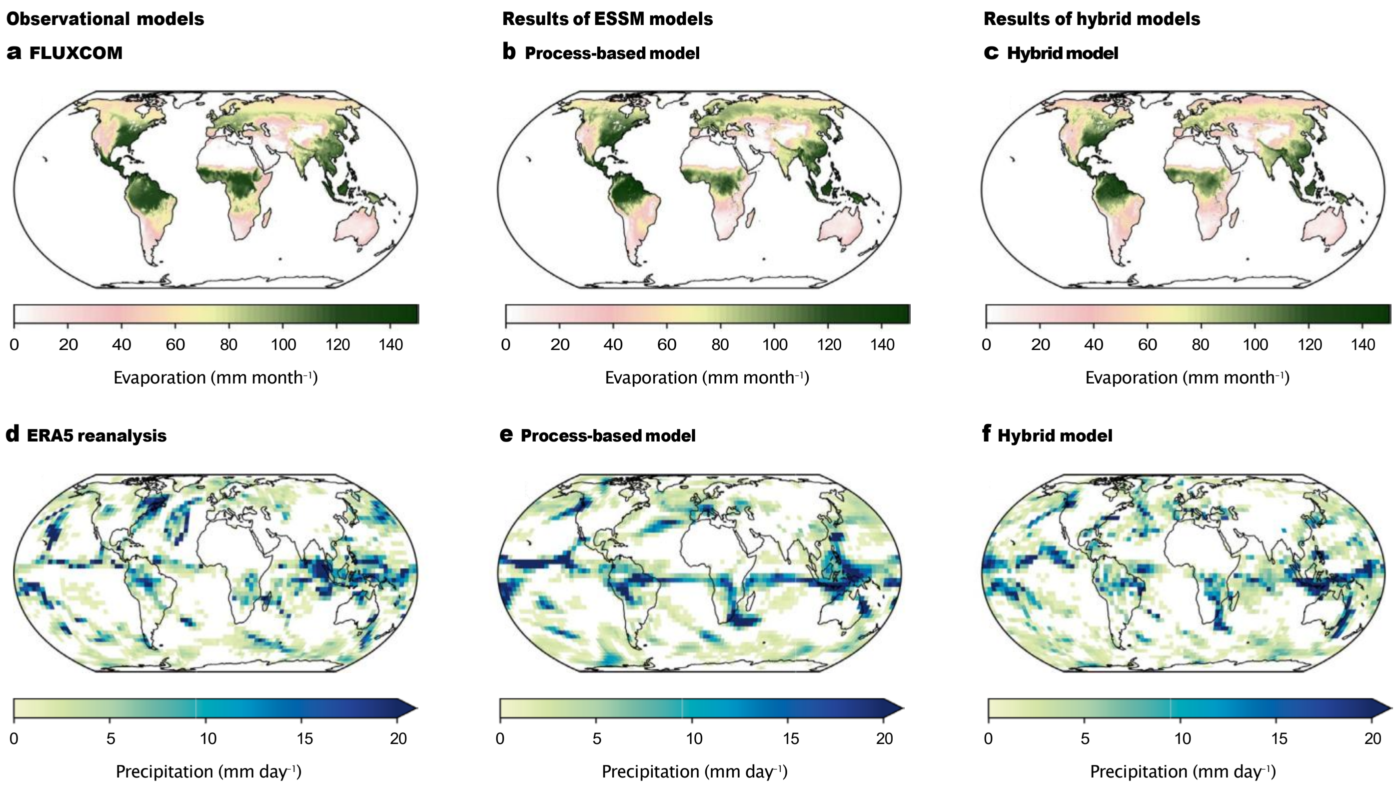

Fig. 2 | Comparison between outputs from Earth surface system models and hybrid models. a, The seasonal aggregates of terrestrial evaporation from a model trained directly on evaporation from FLUXNET sites[100]. FLUXNET sites refer to a global network of measurement stations equipped with high-frequency sensors, which collect data on carbon dioxide, water vapour and energy exchanges between the biosphere and atmosphere across diverse ecosystems. **b**, Evaporation predicted by a process-based Earth surface system model[100]. **c**, Evaporation predicted by hybrid model[100]. **d**, Daily precipitation above 1 mm day^{-1} from ERA5 reanalysis[79]. The ERA5 reanalysis, produced by the European Centre for Medium-Range Weather Forecasts (ECMWF), is a comprehensive dataset of past global climate conditions, providing hourly estimates of various atmospheric, land and oceanic variables from 1950 to near real-time. **e**, Results from CM2Mc-LPJmL model based on quantile mapping-based post-processing[79]. CM2Mc-LPJmL is the result of coupling the well-validated dynamic global vegetation model LPJmL5 (Lund–Potsdam–Jena managed land) with the coupled climate model CM2Mc, which is founded on the atmosphere model AM2 and the ocean model MOM5 (Modular Ocean Model 5). **f**, Results from CM2Mc-LPJmL model based on physically constrained GAN-based post-processing[79]. GAN, generative adversarial network. Owing to the incorporation of deep learning mechanisms, the hybrid model simulates Earth's surface processes more effectively than traditional process-based models.

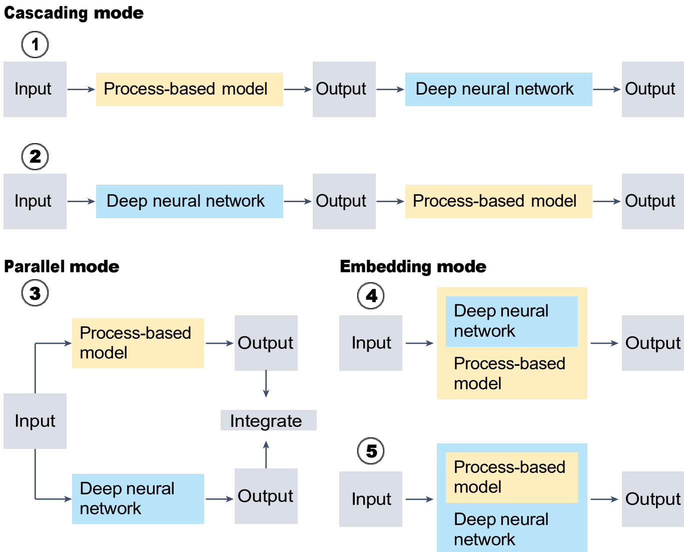

Fig. 3 | Computational logics of hybrid models. a, Cascading mode is a computational pipeline consisting of process-based models and deep neural networks that runs sequentially and transmits intermediate results. There are two cases according to the sequential order of the models. **b**, Parallel mode is when both types of models are run simultaneously. **c**, The embedding mode is when the two types of models are embedded into each other as plug-in modules. According to the embedding relationship, they can also be divided into two cases depending on which model is embedded into the other. Among the three integration modes, the cascading approach is the most simplistic, whereas the embedded mode provides greater flexibility and broader potential in simulating Earth surface processes.

In the second case, the deep neural network is used first, followed by the process-based model (diagram 2 in Fig. 3). Process-based models can, for example, constrain or refine deep neural network outputs to adhere to physical mechanisms[106], such as the law of conservation of energy[51]. In addition, deep neural networks can be used to calibrate process-based models, reducing parameterization complexity when solving partial differential equations[107,108]. Deep neural networks could also propose process model algorithms to minimize developer biases or clarify non-intuitive relationships that could be incorporated into process models[109].

Parallel mode
In parallel mode, process-based models and deep neural networks run simultaneously (diagram 3 in Fig. 3). This parallel mode offers three practical advantages due to its concurrent nature: addressing complex issues using a divide-and-conquer approach; processing multimodal data; and facilitating parallel computing. Specifically, the divide-and-conquer strategy, generally built for decomposed sub-problems, leverages both the process-based model and deep neural network simultaneously to tackle challenges in their respective areas of expertise[110,111]. For instance, a modified susceptible–exposed–infectious–removed model can be used to derive the COVID-19 epidemic curve based on population migration data, while a long short-term memory network trained on SARS data predicts the epidemic[112].

Furthermore, process-based models tend to handle specific file formats (for example Shapefile and NetCDF) more efficiently than deep neural networks in terms of preprocessing and encoding these raw datasets. Consequently, using process-based models or deep neural networks to process data they can handle most efficiently while working with heterogeneous data sources can improve computational efficiency[113,114]. Parallel computing not only uses supercomputer technology to enhance computational performance[115] but also divides the modelling environment, preventing incompatibilities caused by heterogeneous computing resources between process-based models and deep neural networks[59,116].

Embedding mode
The embedding mode enables process-based models and deep neural networks to function as plug-and-play components[117–119]. Specifically, these two approaches serve as complementary plug-ins. The embedding mode can be further subdivided into two cases.

The first case involves integrating deep neural networks as surrogate modules into process-based models (diagram 4 in Fig. 3). Trained deep neural networks can act as neural surrogates or solvers using emulation-style algorithms for computationally intensive process submodules, such as those based on partial differential equations[120], optimization procedures[121] and high-dimensional tasks[107]. These neural surrogates or solvers allow for the automatic parameterization and modification of local modules in process-based models[122], improving computational efficiency with less resources needed in simulation complex dynamics[123].

The second case entails incorporating process-based models into deep neural networks (diagram 5 in Fig. 3) to include physical mechanisms and principles, thereby constructing physics-informed architectures[124], such as Physics-Informed Neural Networks (PINNs)[81]. For example, designing specific loss functions for network optimization is a straightforward and effective way for constraining inferred results to adhere to domain-specific understanding[125]. Methods for determining the network's structure, like hidden layers, have been explored based on domain laws or physical techniques. Although challenging, groundbreaking results have been achieved, such as neural ordinary differential equations[126] and geographically weighted artificial neural network[127]. Another promising area of research is incorporating physical constraints into deep neural networks to derive new equations that characterize Earth surface dynamics[128].

The integration of ESSM and DL, achieved through diverse strategies such as cascading, parallel and embedding modes, opens new doors in the geosciences. This fusion of theory-simulation-based and data-driven techniques offers a more expansive and clearer perspective for predicting and understanding Earth surface processes.

Shortcomings of hybrid ESSM
Despite years of continuous research and development, hybrid ESSM is still in its infancy. Highly heterogeneous data, insufficient ground-truth data and low interpretability of outcomes have been previously described as the main challenges[20]. This section examines further theoretical and practical shortcomings in existing hybrid ESSM, with the aim to identify opportunities for improvements in hybrid ESSM capabilities.

Restricted integration scenarios
Existing hybrid ESSM research concentrates mainly on model-level integration. Nevertheless, the ESSM lifecycle is more comprehensive and generally includes the five indispensable stages in the scientific methodology summarized in the introduction. These stages are all essential for determining the quality and relevance of the solution so that the modelling is suitable for the purpose, such as is captured by the notions of usability, feasibility and reliability[34].

Subjectivity in the modelling lifecycle

Subjective factors can be the primary obstacles to achieving highly accurate outcomes in hybrid ESSM. As noted previously, researchers are prone to use their expertise or criteria, likely making the modelling logic less precise and potentially biased. For example, modellers can favour physical or numerical models, whereas others with a strong background in DL prefer a more data-driven approach. Both can lead to suboptimal hybrid models for a specific task[129]. Another underlying challenge is that numerous innovative ideas and techniques about DL continue to inundate scientific communities, necessitating researchers to comprehend the most current technical advancements[130]. When it comes to choosing configurations (for example, architectures or hyperparameters) for deep neural networks, many experienced ESSM researchers might be at a loss.

Incompatible computational environment

Incompatibilities between ESSM and DL in terms of hardware, software stack and operating environment could impair computational efficiency. Specifically, process-based models are often executed on multi-CPU computers or high-performance computing facilities[131], whereas the training and inferencing phases of deep neural networks are typically deployed in GPU-based and container-based (like Docker) environments[132]. Further, process-based models, particularly mechanistic ones, are often constructed using Fortran and C++, whereas deep neural networks in specific environments use Python and packages such as Tensorflow and PyTorch. This latter distinction has become less problematic as many scientists are starting to embrace Python and the emerging technique of scientific machine learning (SciML) developed with the Julia programming packages[133]. But these discrepancies in development and deployment methodologies generally result in separating DL and ESSM workloads. This substantially affects data and message transmission and limits the computing capacity of hybrid ESSM.

Despite notable progress, hybrid ESSM still has substantial ground to cover. Overcoming these hurdles calls for expansion of integration approaches, deeper scrutiny of modelling biases, and harmony between the computational environments of ESSM and DL.

Towards iterative hybrid ESSM

Constructing appropriate and effective solutions to ascertain the dynamics of the Earth surface system is generally challenging. An initial undertaking is to fully understand the problem contexts and associated geographic objects. Handling big and multimodal data, especially extracting useful information or knowledge from it, is also a laborious task. Further, it is essential to focus on the trade-offs between model complexity and computational efficiency, as well as to calibrate the derived models and quantify or at least indicate model performance including uncertainty aspects. Finally, when applying constructed models, computational environments and software stacks are not easy to comprehend for those domain experts who are often not also experts in computation. Given the possible challenges and shortcomings analysed earlier, our goal is to put forward a hybrid ESSM-deep learning framework that holds the potential to deliver customized, scalable and accurate solutions to given ESSM tasks so as to lower technical barriers.

Conceptual framework

We propose a conceptual framework designed to intelligently guide the entire modelling lifecycle, ultimately generating accurate solutions for specific tasks through collaboration between human insights and the strengths of neural networks (Fig. 4). Furthermore, this framework not only possesses self-renewal capabilities through its internal mechanisms, but also continuously improves its performance by incorporating user feedback from practical applications. For example, our iterative hybrid-ESM framework can generate customized responses based on user queries and prompts in a similar way to ChatGPT, but it is specifically designed for the ESSM field and has potential for broader applications in ESM. Notable differences between our framework and ChatGPT include the output form (multimodal outputs and modelling resource assignment versus pure-text outputs), technical ESM foundation (knowledge-constrained inference versus inference by large-scale deep neural networks) and learning strategy (online self-learning versus periodic background updates)[13,134].

The framework consists of three major components. First, there is a modelling-related knowledge repository (representing human insights), which organizes diverse knowledge established by previous modelling of Earth surface processes. For example, there should be knowledge about formulated geophysical mechanisms, data, methods, models (both conventional and DL-based approaches) and computational infrastructure (software and hardware basis for running and analysing developed models). Second, there is a DL computing system that encompasses various deep neural networks (for example, transformers for language understanding, convolutional neural networks for image processing and graph neural networks for knowledge reasoning), enhanced by effective learning strategies such as self-taught learning, online learning and Bayesian learning. Third, there are adaptive guidance plans, which result from interactions between the first two components and improve the modelling of lifecycle processes.

Ideally, the modelling-related knowledge repository and the DL computing system are built within a homogeneous environment, sharing cloud-based services and high-performance computing support. This configuration offers potential practical advantages. For example, a homogeneous environment enables efficient interaction and communication between the two components. In addition, it aids data acquisition from the Internet and crowdsourcing, online updates of deep neural networks and dynamic extension of the modelling-related knowledge repository. Finally, it allows generated models or computational solutions to be deployed and run in separate environments, avoiding incompatibility issues.

The knowledge repository not only organizes various types of knowledge but also serves as a knowledge graph, providing a priori constraints to improve the inference performance of the DL computing system. Consequently, the generated output aligns with the current scope of knowledge or logically inferred derivations from existing information, rather than being arbitrarily generated by deep neural networks (which can contradict domain principles or modelling logic). In constructing the knowledge repository, both 'top-down' and 'bottom-up' strategies are used. Top-down strategies depend on domain experts or communities for the structure and items of the repository, resulting in a more scientific but potentially biased approach that typically involves manual manipulation. Bottom-up strategies use natural language processing and computer vision methods to automatically extract knowledge about Earth surface process concepts, entities and relationships from publicly available authoritative data (for example peer-reviewed research literature and web corpora). This approach generates more comprehensive and up-to-date knowledge through an automated process, but it relies on existing information and knowledge extraction technologies.

The deep neural networks within the DL computing system are ideally pretrained using authoritative datasets, which require further preprocessing, such as converting them into question–answer pairs

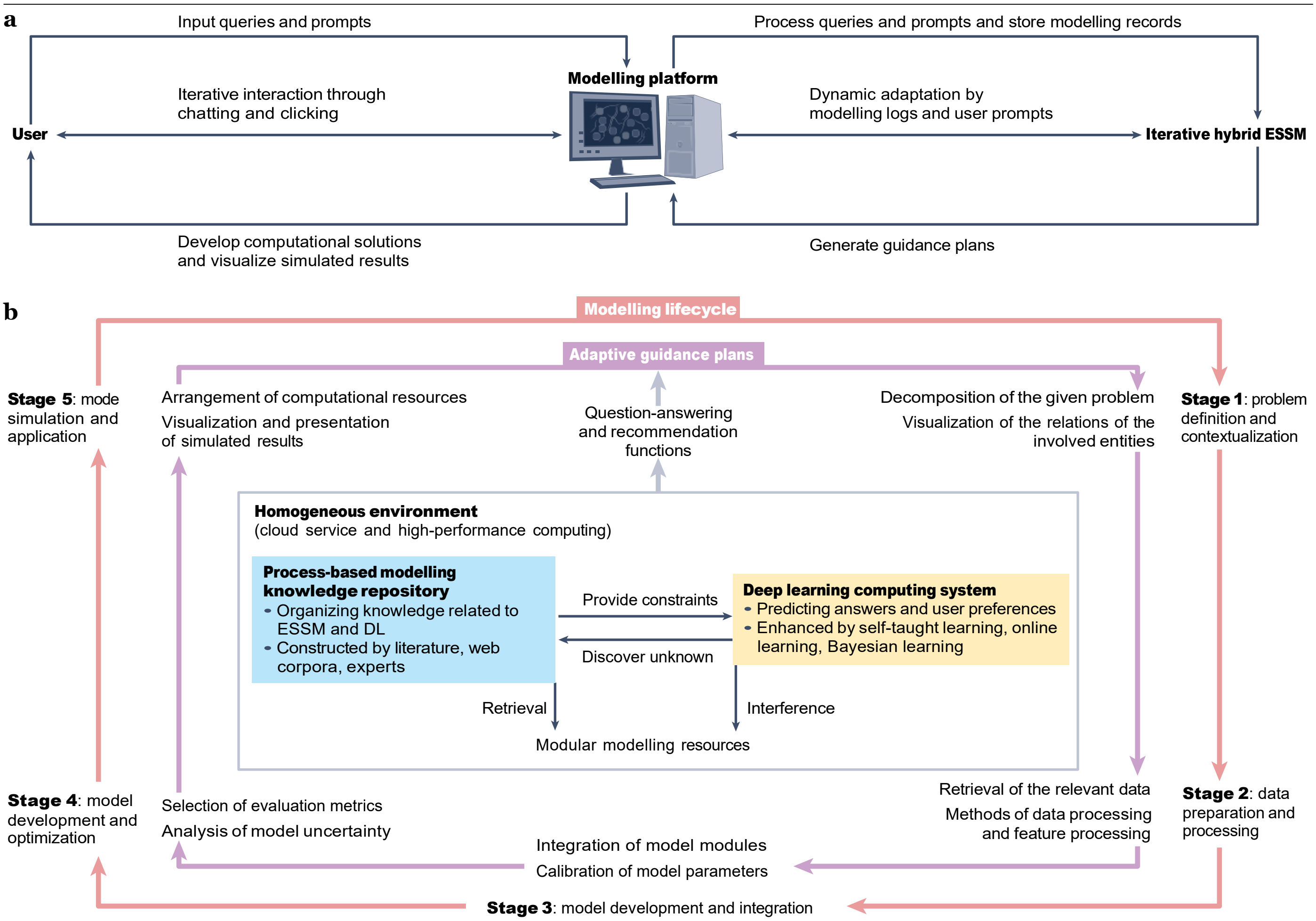

Fig. 4 | A framework for iterative integration of Earth surface system modelling and deep learning. a, Interactions between the hybrid Earth surface system model (ESSM) and the user allow the system to deepen its understanding of their questions or preferences, resulting in increasingly customized, scalable and accurate guidance plans. **b**, Adaptive guidance plans are intelligently generated within a homogeneous environment (central box) to guide modelling tasks and allocate modular resources throughout the modelling lifecycle. These plans result from question-answering mechanisms and recommendation functions, powered by a modelling-related knowledge repository and a deep learning (DL) computing system. The guidance plans can direct modellers to build customized solutions to solve given environmental or Earth process problems. This guidance can be divided into five stages according to the ESSM lifecycle (outer ring). When users interact iteratively with a modelling platform that adopts this framework, the outcomes of the guidance plans can be more precise and tailored to specific tasks.

and supplementing them with user–item interaction datasets (such as click-throughs or historical modelling logs) available from open ESSM platforms. The developed DL computing system can help complete the existing knowledge repository by inferring unknown or missing information (entities or their relations) using graph-based networks and Bayesian learning strategies. Furthermore, it can extract subjects from users' inputs (questions or preferences) and feed them into neural networks constrained by the knowledge graph to generate answers and recommendations. However, data deficiency, or the lack of high-quality labelled data, might pose a challenge for the DL computing system after some time. In such cases, semi- or unsupervised learning or self-taught learning offer promising solutions. Deep neural networks can update their parameters based on unlabelled data and maintain online learning on the backend, facilitating self-renewal of both the deep neural networks and the knowledge repository.

Adaptive modelling guidance plans, a comprehensive term for inference results derived from question-answering mechanisms and recommendation functions, are generated through the collaboration between the modelling-related knowledge repository and the DL computing system. These plans are multimodal in nature, producing outputs in various forms, such as text, images and video. Additionally, given the homogeneous deployment environment of the framework, modular modelling resources such as data, methods and models can be allocated accordingly. As users engage in iterative interactions with a platform implementing our proposed framework, the system deepens its understanding of their questions or preferences, resulting in increasingly customized, scalable and accurate guidance plans. Generated throughout the entire modelling lifecycle, these adaptive guidance plans serve as clear guidelines and modelling resource allocators for arising modelling tasks. In the end, they aid complete and effective computational solutions by integrating all modular resources within separate spaces in the deployment environment.

The guidance plans provided by the iterative hybrid ESSM can be divided into five stages. Stage 1 involves the problem definition and contextualization. A given topic can be hierarchically decomposed into multiple sub-analyses or interactions of the subsystems involved. Geographic objects and their relations in space and time can be recognized and visualized automatically, giving modellers a clearer understanding of what to analyse before investigation commences. Stage 2 forms the data preparation and processing. Relevant data is retrieved, with adaptive recommendations for processing techniques. Therefore, beneficial patterns and interior knowledge can be acquired from various types of data rather than through manual manipulation. Stage 3 is the model development and integration step. A modular strategy organizes model modules, allowing knowledge-based reasoning methods to build customized models for specific needs. Automatic methods, including calibration and uncertainty estimation, improve prediction results and computational efficiency. Stage 4 is where the model is evaluated and optimized. Suitable metrics for evaluating performance will also be selected at this stage. The probabilistic inferencing and other methods will be used to estimate the statistical confidence in the models and other ways to represent uncertainties. Finally, stage 5 is where the model simulation and application are produced. Considering the data and model characteristics, an efficient computational infrastructure will be developed by allocating resources to create separate environments. Simulated results will be effectively visualized and communicated to relevant stakeholders or the general public.

Potential application case

The COVID-19 pandemic has had unprecedented impacts on human behaviour and decision-making, leading to effects on Earth surface processes at local to global scales[135]. As we navigate the post-COVID-19 era, with all countries adapting to living with the virus, it is crucial to reflect on the pandemic's impacts over the past three years and anticipate future trends. Understanding the questions related to COVID-19 impacts will aid in analysing the pandemic's effects on the socioeconomic environment, as well as discerning causality in the Earth surface system response, ultimately promoting sustainable development in the post-COVID-19 era[136].

The ensuing discussion assumes that this framework has been implemented and integrated in a modelling platform. The aim is to elucidate the application of our framework to this macroscopic and multifaceted problem and to explain the mechanisms of operation and interaction between the modelling-related knowledge repository and the DL computing system contained within the framework. In addition, the potential generation of modelling guidance plans is explored, contributing to a comprehensive understanding of the underlying principles and goals of our framework.

Upon receiving a user's enquiry, such as "What are the various impacts of the changed travel behaviours adapted to the post-COVID-19 era?", the platform processes the input using natural language processing techniques. Specifically, it extracts key terms and concepts, including "changed travel behaviours", "post-COVID-19 era" and "various impacts." The DL computing system then consults the knowledge graph maintained by the modelling-related knowledge repository, searching for relevant concepts, entities and relationships associated with the extracted key terms. Graph neural networks within the computing system infer potential answers and recommendations based on user preferences, using connections and patterns in the knowledge graph. Simultaneously, if necessary, other deep learning models such as transformers for language understanding and convolutional neural networks for image processing are used to generate coherent and meaningful responses. Inference processes take place throughout the entire modelling lifecycle stages, and the DL computing system's output is validated by the knowledge repository, ensuring that the generated guidance plans adhere to the current scope of knowledge and domain principles.

Generated guidance plans will progress through modelling lifecycle stages, starting with the decomposition of the primary question into finer subtopics. Initially, the overarching question is divided into intermediate-level concerns, such as impacts on the natural environment (emissions, air quality and climate) and the socioeconomic environment (poverty, food and globalization). These intermediate topics are further divided into finer details, including spatiotemporal variations in local travel behaviours[137], energy supply and consumption[138], and air pollution changes[139]. These issues span local to global scales and short to long-term timescales. By visualizing geographical objects and their relationships, the decomposed subproblems enable users to gain a comprehensive understanding of the causal relationships and interactions among changed human behaviours, their impacts on the economy and society, and the resulting Earth surface system responses.

Furthermore, the guidance plans will include recommendations for data sources and their associated processing methods. For instance, the suggested data can encompass a broad range of Earth observation data, such as satellite remote-sensing products, measurement networks, and ground-based sensor data. Additionally, long-term socioeconomic data related to energy, trade, and transportation are valuable, but their time lags might hinder real-time modelling and analysis, necessitating the use of licensed and shared private sector data[135]. The generated plans will also propose suitable data processing methods. These methods can include data cleaning to eliminate inconsistencies and errors, normalization to standardize values across different scales, feature extraction to identify relevant variables, and data fusion or integration techniques to merge disparate data types into a coherent dataset. Moreover, the guidance plans could suggest appropriate techniques for handling missing data, such as interpolation or imputation methods, and for reducing dimensionality through methods such as principal component analysis or t-distributed stochastic neighbour embedding.

During the model development and integration stage, adaptive guidance plans will recommend a cohesive and potentially hybrid approach, combining process-based and data-driven models. These plans emphasize the significance of model integration, comparing and calibrating various models to achieve a comprehensive understanding of the problem while evaluating their performance against observational data or benchmarks. For instance, process-based models, such as those in the Coupled Model Intercomparison Project (CMIP), could be suggested to simulate complex interactions between different subsystems, while data-driven models like machine learning or statistical models can analyse correlations between travel behaviours and environmental or socioeconomic impacts. The integrated hybrid approach aids simultaneous simulation of both human and natural environments, ensuring a thorough investigation of post-COVID-19 travel behaviour impacts by leveraging the strengths of process-based and data-driven models while enhancing parameterization or developing simulators to reduce computational costs.

In stage 4, model evaluation and optimization, the guidance plans will recommend suitable metrics and methodologies, such as cross-validation, mean squared error and precision-recall curves, to assess model performance and conduct uncertainty analysis, ensuring trustworthiness. Furthermore, optimization strategies such as parameter tuning and model ensembles can be suggested and performed to enhance model accuracy and generalization capabilities. For model simulation and application, human travel data and Earth observation data are often heterogeneous and large, and the models built are often complex. The guidance plans will therefore emphasize the effectiveness and efficiency of data processing and model computation, optimizing the compatibility and performance of the computational pipeline by dynamically allocating and configuring computational resources with appropriate operating systems and software applications. The plans will also propose visualization and communication tools for effectively presenting and sharing results with stakeholders, while highlighting potential avenues for further research or real-world implementation.

During the modelling guidance generation process, users can offer feedback and prompts on the plans, which can be used to refine the plans further and potentially enhance the performance of the DL computing system while fine-tuning the modelling-related knowledge base. This iterative interaction loop enables the platform to learn from and adapt to users' needs and preferences, ensuring increasingly customized and precise guidance. Additionally, the platform can foster collaboration among users, allowing them to exchange insights, experiences, and best practices for modelling the impacts of altered travel behaviours in the post-COVID-19 context. This collaborative element contributes to the ongoing growth and refinement of the knowledge repository and the platform as a whole.

In summary, this iterative hybrid ESSM framework, consisting of a modelling-related knowledge repository (which represents human insights), a DL computing system (embodying the strengths of neural networks) and adaptive guidance plans (derived from the synergy of the two), could help to enhance the intricacies of ESSM.

Summary and future perspectives

ESSM is confronted at present with substantial technical challenges, primarily in discerning the complexities of the problem space and in executing efficient, precise analysis of voluminous observational data. DL, noted for its potent data-driven proficiencies in processing and representing features inherent in large-scale data, has the potential to augment the performance of ESSM in simulating Earth surface processes, and displays considerable promise in integrating with ESSM, thus contributing to the development of a hybrid ESSM. However, integration of ESSM and DL approaches is an emerging technique for understanding Earth surface system dynamics. Most research focuses on integrating process-based models and deep neural networks during the initial development of hybrid models, rather than exploring the advantages of a comprehensive approach that covers all modelling lifecycle stages. We propose a conceptual framework in which DL is iteratively integrated into ESSM throughout the modelling lifecycle. This framework aims to enable customized, scalable and accurate solutions for modelling Earth surface processes by iteratively integrating ESSM-related knowledge with the data-harnessing strengths of DL. Such integration could reduce subjective biases in the modelling processes and provide compatible computational environments between the approaches.

ESM's interdisciplinarity nature necessitates an open knowledge sharing community, open resources (for example, datasets, codes and models) and open research cooperation. These requirements underscore the importance of practices such as adhering to the 'FAIR' principles, which ensure data findability, accessibility, interoperability, and reusability[140]. Research organizations, such as the OMF (Open Modeling Foundation)[141], the OpenGMS (Open Geographic Modeling and Simulation)[142], the CSDMS (Community Surface Dynamics Modeling System)[143] and the CoMSES Net (Network for Computational Modeling in Social and Ecological Sciences)[144,145], have already contributed to collaboration and sharing. In conjunction with similar scientific entities, it is hoped that these environments will encourage the collaboration of scientists from various disciplines to address complex problems. Building a virtual online platform for researchers to experiment and discuss will also enhance the transparency and reproducibility of modelling.

The underlying abstract formulations of DL networks present a unique challenge for geoscientific applications, as they are not easily interpretable despite producing precise simulation results. Explainable or interpretable artificial intelligence using explanatory approaches (such as layer-wise relevance propagation, integrated gradients and occlusion analysis) do however allow users to understand internal mechanics of deep neural networks[146]. Merging process-based models with domain-specific knowledge as surrogates in deep neural networks can enhance the transparency of these typically abstract systems[147]. Related research projects are still evolving, but there remains a substantial trade-off between model performance in terms of explainability and simulation accuracy of model outputs.

Moving forward, the intelligent development of customized ESMs could be achieved through the iterative ESM framework. However, the pathway in this framework might not align entirely with geoscientists' thought processes, as generated solutions predominantly depend on the inference results of deep neural networks. Therefore, our framework advocates for enhancing the accuracy of the DL computing system based on specific objective functions and also implementing contextually appropriate logic constraints that are compliant with the mindset of the geoscience community. These considerations should be taken into account throughout the entire modelling lifecycle, ultimately enhancing the trustworthiness of results and outcomes.

Moreover, two typical characteristics reduce confidence in the predictive accuracy of ESM integrated with DL. The first is the difficulty in accurately simulating certain extreme events due to the highly dynamic character of the Earth's surface system[148]. For instance, hybrid models can struggle to precisely predict heavy rainfall or landslides triggered by a combination of geophysical factors such as seismic activities, soil saturation and steep topography[149]. Second, disruptions in observed data due to climate change, which alters weather patterns and environmental conditions, along with human-induced alterations such as sediment redistribution, pose additional challenges to the efficacy of models created to predict Earth surface processes[150,151]. To mitigate these issues, maintaining regular updates of models and software is crucial, as is using data assimilation, lifelong learning techniques, and explainable or interpretable artificial intelligence. In addition, acquiring up-to-date and widespread data and processing the vulnerable observations using hybrid models can effectively improve modelling performance. Ultimately, the community needs to recognize that uncertainty will always be present, but adherence to good modelling practices can enhance the trustworthiness and credibility of results[34,152]. These include deliberating on fitness for purpose, applying systematic procedures, characterizing and discussing uncertainties, justifying choices, and clearly stating assumptions and limitations[32]. Ensuring transparency through thorough documentation further strengthens the reliability of the outcomes[34,153].

In conclusion, we believe the integration of ESSM and DL is a considerable step towards more accurate and reliable models of Earth's surface processes. As this synergy spans across multiple disciplinaries and is still an evolving field of science, the capability and capacity should be increased through the collaboration of an open scientific community and the adherence of good modelling practices.

Published online: 11 July 2023

References

1. Steffen, W. et al. The emergence and evolution of Earth system science. *Nat. Rev. Earth Environ.* **1**, 54–63 (2020).
2. Bowman, D. M. et al. Vegetation fires in the Anthropocene. *Nat. Rev. Earth Environ.* **1**, 500–515 (2020).
3. Zipkin, E. F. et al. Addressing data integration challenges to link ecological processes across scales. *Front. Ecol. Environ.* **19**, 30–38 (2021).
4. Knight, J. & Harrison, S. The impacts of climate change on terrestrial Earth surface systems. *Nat. Clim. Change* **3**, 24–29 (2013).
5. Chen, M. et al. Geographic modeling and simulation systems for geographic research in the new era: some thoughts on their development and construction. *Sci. China Earth Sci.* **64**, 1207–1223 (2021).
6. Luttge, A., Arvidson, R. S., Fischer, C. & Kurganskaya, I. Kinetic concepts for quantitative prediction of fluid-solid interactions. *Chem. Geol.* **504**, 216–235 (2019).
7. Pelletier, J. D. *Quantitative Modeling of Earth Surface Processes* (Cambridge Univ. Press, 2008).
8. Meir, P., Cox, P. & Grace, J. The influence of terrestrial ecosystems on climate. *Trends Ecol. Evol.* **21**, 254–260 (2006).
9. Zhang, Z. Carbon mitigation potential afforded by rooftop photovoltaic in China. *Nat. Commun.* **14**, 2347 (2023).
10. Zhu, R. et al. GIScience can facilitate the development of solar cities for energy transition. *Adv. Appl. Energy* **10**, 100129 (2023).
11. Lee, C. A., Gasster, S. D., Plaza, A., Chang, C.-I. & Huang, B. Recent developments in high performance computing for remote sensing: a review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **4**, 508–527 (2011).
12. Li, S. et al. Geospatial Big Data handling theory and methods: a review and research challenges. *ISPRS J. Photogramm. Remote Sens.* **115**, 119–133 (2016).
13. Thorp, H. H. ChatGPT is fun, but not an author. *Science* **379**, 313 (2023).
14. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
15. Camps-Valls, G., Tuia, D., Zhu, X. X. & Reichstein, M. *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science and Geosciences* (Wiley, 2021).
16. Chen, M. et al. Artificial intelligence and visual analytics in geographical space and cyberspace: research opportunities and challenges. *Earth Sci. Rev.* **241**, 104438 (2023).
17. Goldstein, E. B., Coco, G. & Plant, N. G. A review of machine learning applications to coastal sediment transport and morphodynamics. *Earth Sci. Rev.* **194**, 97–108 (2019).
18. Qian, Z. et al. Deep Roof Refiner: a detail-oriented deep learning network for refined delineation of roof structure lines using satellite imagery. *Int. J. Appl. Earth Obs. Geoinf.* **107**, 102680 (2022).
19. Li, W. & Hsu, C.-Y. GeoAI for large-scale image analysis and machine vision: recent progress of artificial intelligence in geography. *ISPRS Int. J. Geoinf.* **11**, 385 (2022).
20. Reichstein, M. et al. Deep learning and process understanding for data-driven Earth system science. *Nature* **566**, 195–204 (2019).
21. 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).
22. Sutton, R. The bitter lesson. *Incomplete Ideas* (13 March 2019); http://www.incompleteideas.net/IncIdeas/BitterLesson.html.
23. Zhang, Q., Yang, L. T., Chen, Z. & Li, P. A survey on deep learning for Big Data. *Inf. Fusion.* **42**, 146–157 (2018).
24. Razavi, S. et al. Coevolution of machine learning and process-based modelling to revolutionize Earth and environmental sciences: a perspective. *Hydrol. Process.* **36**, e14596 (2022).
25. Bolton, T. & Zanna, L. Applications of deep learning to ocean data inference and subgrid parameterization. *J. Adv. Model. Earth Syst.* **11**, 376–399 (2019).
26. Sillmann, J. et al. Understanding, modeling and predicting weather and climate extremes: challenges and opportunities. *Weather Clim. Extremes* **18**, 65–74 (2017).
27. Kadow, C., Hall, D. M. & Ulbrich, U. Artificial intelligence reconstructs missing climate information. *Nat. Geosci.* **13**, 408–413 (2020).
28. Razavi, S. Deep learning, explained: fundamentals, explainability, and bridgeability to process-based modelling. *Environ. Model. Softw.* **144**, 105159 (2021).
29. Murray, A. B. et al. Geomorphology, complexity, and the emerging science of the Earth's surface. *Geomorphology* **103**, 496–505 (2009).
30. Phillips, J. D. Amplifiers, filters and geomorphic responses to climate change in Kentucky rivers. *Clim. Change* **103**, 571–595 (2010).
31. Fisher, R. A. et al. Vegetation demographics in Earth system models: a review of progress and priorities. *Glob. Change Biol.* **24**, 35–54 (2018).
32. Jakeman, A. J., Letcher, R. A. & Norton, J. P. Ten iterative steps in development and evaluation of environmental models. *Environ. Model. Softw.* **21**, 602–614 (2006).
33. Ma, Z. et al. Activity-based process construction for participatory geo-analysis. *GIScі. Remote Sens.* **58**, 180–198 (2021).
34. Hamilton, S. H., Pollino, C. A., Stratford, D. S., Fu, B. & Jakeman, A. J. Fit-for-purpose environmental modeling: targeting the intersection of usability, reliability and feasibility. *Environ. Model. Softw.* **148**, 105278 (2022).
35. Chen, M. et al. Position paper: open web-distributed integrated geographic modelling and simulation to enable broader participation and applications. *Earth Sci. Rev.* **207**, 103223 (2020).
36. Werner, B. T. Complexity in natural landform patterns. *Science* **284**, 102–104 (1999).
37. Anderson, P. W. More is different: broken symmetry and the nature of the hierarchical structure of science. *Science* **177**, 393–396 (1972).
38. Werner, B. T. & McNamara, D. E. Dynamics of coupled human–landscape systems. *Geomorphology* **91**, 393–407 (2007).
39. Klir, G. J. & Simon, H. A. *The Architecture of Complexity* (Springer, 1991).
40. Heymann, M. & Dahan Dalmedico, A. Epistemology and politics in Earth system modeling: historical perspectives. *J. Adv. Model. Earth Syst.* **11**, 1139–1152 (2019).
41. Shen, C. & Phanikumar, M. S. A process-based, distributed hydrologic model based on a large-scale method for surface–subsurface coupling. *Adv. Water Resour.* **33**, 1524–1541 (2010).
42. Robinson, D. T. et al. Modelling feedbacks between human and natural processes in the land system. *Earth Syst. Dynam.* **9**, 895–914 (2018).
43. Lü, G. et al. Geographic scenario: a possible foundation for further development of virtual geographic environments. *Int. J. Digit. Earth* **11**, 356–368 (2018).
44. Haarsma, R. J. et al. High resolution model intercomparison project (HighResMIP v1. 0) for CMIP6. *Geosci. Model. Dev.* **9**, 4185–4208 (2016).
45. Worley, P. H. et al. Performance of the Community Earth System Model. In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis* 1–11 (IEEE, 2011).
46. Salcedo-Sanz, S. et al. Machine learning information fusion in Earth observation: a comprehensive review of methods, applications and data sources. *Inf. Fusion.* **63**, 256–272 (2020).
47. Lee, J.-G. & Kang, M. Geospatial Big Data: challenges and opportunities. *Big Data Res.* **2**, 74–81 (2015).
48. Ansari, S. et al. Unlocking the potential of NEXRAD data through NOAA's Big Data partnership. *Bull. Am. Meteorol. Soc.* **99**, 189–204 (2018).
49. Ge, Y. et al. Progress of big geodata. *Sci. Bull.* **67**, 1739–1742 (2022).
50. Oussous, A., Benjelloun, F.-Z., Lahcen, A. A. & Belfkih, S. Big Data technologies: A survey. *J. King Saud Univ. Comput Inf. Sci.* **30**, 431–448 (2018).
51. 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).
52. Qian, Z. et al. Vectorized dataset of roadside noise barriers in China using street view imagery. *Earth Syst. Sci. Data* **14**, 4057–4076 (2022).
53. Clark, M. P., Kavetski, D. & Fenicia, F. Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resour. Res.* **47**, WR009827 (2011).
54. Beven, K. & Freer, J. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *J. Hydrol.* **249**, 11–29 (2001).
55. Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G. & Yacalis, G. Could machine learning break the convection parameterization deadlock? *Geophys. Res. Lett.* **45**, 5742–5751 (2018).
56. Tang, Y., Reed, P., Wagener, T. & Van Werkhoven, K. Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation. *Hydrol. Earth Syst. Sci.* **11**, 793–817 (2007).
57. Di Baldassarre, G., Schumann, G. & Bates, P. Near real time satellite imagery to support and verify timely flood modelling. *Hydrol. Process.* **23**, 799–803 (2009).
58. Kucera, P. A. et al. Precipitation from space: advancing Earth system science. *Bull. Am. Meteor. Soc.* **94**, 365–375 (2013).
59. See, S. & Adie, J. Challenges and opportunities for a hybrid modelling approach to Earth system science. *CCF Trans. HPC* **3**, 320–329 (2021).
60. Zhang, K. et al. Quantifying the photovoltaic potential of highways in China. *Appl. Energy* **324**, 119600 (2022).
61. Fuhrer, O. et al. Near-global climate simulation at 1 km resolution: establishing a performance baseline on 4888 GPUs with COSMO 5.0. *Geosci. Model. Dev.* **11**, 1665–1681 (2018).
62. Kratzert, F., Klotz, D., Brenner, C., Schulz, K. & Herrnegger, M. Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrol. Earth Syst. Sci.* **22**, 6005–6022 (2018).
63. Wei, X., Zhang, L., Yang, H.-Q., Zhang, L. & Yao, Y.-P. Machine learning for pore-water pressure time-series prediction: application of recurrent neural networks. *Geosci. Front.* **12**, 453–467 (2021).
64. Bihlo, A. A generative adversarial network approach to (ensemble) weather prediction. *Neural Netw.* **139**, 1–16 (2021).
65. Peng, X., Li, Q. & Jing, J. CNGAT: A graph neural network model for radar quantitative precipitation estimation. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2021).
66. Zhou, L. & Zhang, R.-H. A self-attention–based neural network for three-dimensional multivariate modeling and its skillful ENSO predictions. *Sci. Adv.* **9**, eadf2827 (2023).

67. Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y. & Beroza, G. C. Earthquake transformer — an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nat. Commun.* **11**, 3952 (2020).
68. Montavon, G., Samek, W. & Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digital Signal. Process.* **73**, 1–15 (2018).
69. Gómez-Chova, L., Tuia, D., Moser, G. & Camps-Valls, G. Multimodal classification of remote sensing images: a review and future directions. *Proc. IEEE* **103**, 1560–1584 (2015).
70. Dalla Mura, M. et al. Challenges and opportunities of multimodality and data fusion in remote sensing. *Proc. IEEE* **103**, 1585–1601 (2015).
71. Zhu, R. et al. Deep solar PV refiner: a detail-oriented deep learning network for refined segmentation of photovoltaic areas from satellite imagery. *Int. J. Appl. Earth Obs. Geoinf.* **116**, 103134 (2023).
72. 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).
73. Fan, R. et al. Fine-scale urban informal settlements mapping by fusing remote sensing images and building data via a transformer-based multimodal fusion network. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–16 (2022).
74. Ives, A. R. et al. Statistical inference for trends in spatiotemporal data. *Remote Sens. Environ.* **266**, 112678 (2021).
75. Li, X., Zhang, C. & Li, W. Building block level urban land-use information retrieval based on Google Street View images. *GISci. Remote Sens.* **54**, 819–835 (2017).
76. Zhang, K. et al. Using street view images to identify road noise barriers with ensemble classification model and geospatial analysis. *Sustain. Cities Soc.* **78**, 103598 (2022).
77. Zhong, T. et al. Assessment of solar photovoltaic potentials on urban noise barriers using street-view imagery. *Renew. Energy* **168**, 181–194 (2021).
78. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D. & Krishnan, D. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3722–3731 (IEEE, 2017).
79. Hess, P., Drüke, M., Petri, S., Strnad, F. M. & Boers, N. Physically constrained generative adversarial networks for improving precipitation fields from Earth system models. *Nat. Mach. Intell.* **4**, 828–839 (2022).
80. He, X., Chen, Y. & Ghamisi, P. Heterogeneous transfer learning for hyperspectral image classification based on convolutional neural network. *IEEE Trans. Geosci. Remote Sens.* **58**, 3246–3263 (2019).
81. Raissi, M., Perdikaris, P. & Karniadakis, G. E. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378**, 686–707 (2019).
82. Shi, X. et al. Deep learning for precipitation nowcasting: a benchmark and a new model. *Adv. Neural. Inf. Process. Syst.* **30**, 5617–5627 (2017).
83. Stevens, B. & Bony, S. What are climate models missing? *Science* **340**, 1053–1054 (2013).
84. Gao, Z. et al. in *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences* 218–239 (Wiley, 2021).
85. Lobry, S., Marcos, D., Murray, J. & Tuia, D. RSVQA: visual question answering for remote sensing data. *IEEE Trans. Geosci. Remote Sens.* **58**, 8555–8566 (2020).
86. Chai, S., Xu, Z., Jia, Y. & Wong, W. K. A robust spatiotemporal forecasting framework for photovoltaic generation. *IEEE Trans. Smart Grid* **11**, 5370–5382 (2020).
87. Boers, N. et al. Complex networks reveal global pattern of extreme-rainfall teleconnections. *Nature* **566**, 373–377 (2019).
88. Sambasivan, N. et al. 'Everyone wants to do the model work, not the data work': data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* 1–15 (ACM, 2021).
89. Zhang, Z. et al. Vectorized rooftop area data for 90 cities in China. *Sci. Data* **9**, 66 (2022).
90. Goldstein, E. B. et al. Labeling poststorm coastal imagery for machine learning: measurement of interrater agreement. *Earth Space Sci.* **8**, e2021EA001896 (2021).
91. Geiger, R. S. et al. 'Garbage In, Garbage Out' revisited: what do machine learning application papers report about human-labeled training data? *Quant. Sci. Stud.* **2**, 795–827 (2021).
92. Samsi, S., Mattioli, C. J. & Veillette, M. S. Distributed deep learning for precipitation nowcasting. In *2019 IEEE High Performance Extreme Computing Conference (HPEC)* 1–7 (IEEE, 2019).
93. Rasp, S., Pritchard, M. S. & Gentine, P. Deep learning to represent subgrid processes in climate models. *Proc. Natl Acad. Sci. USA* **115**, 9684–9689 (2018).
94. Flato, G. M. Earth system models: an overview. *WIREs Clim. Change* **2**, 783–800 (2011).
95. Brandt, M. et al. An unexpectedly large count of trees in the West African Sahara and Sahel. *Nature* **587**, 78–82 (2020).
96. Tung, F. & Mori, G. Deep neural network compression by in-parallel pruning-quantization. *IEEE Trans. pattern Anal. Mach. Intell.* **42**, 568–579 (2018).
97. Jouppi, N. P., Young, C., Patil, N. & Patterson, D. A domain-specific architecture for deep neural networks. *Commun. ACM* **61**, 50–59 (2018).
98. Shen, H. & Zhang, L. Mechanism-learning coupling paradigms for parameter inversion and simulation in Earth surface systems. *Sci. China Earth Sci* https://doi.org/10.1007/s11430-022-9999-9 (2023).
99. Hunter, J. M. et al. Framework for developing hybrid process-driven, artificial neural network and regression models for salinity prediction in river systems. *Hydrol. Earth Syst. Sci.* **22**, 2987–3006 (2018).
100. Koppa, A., Rains, D., Hulsman, P., Poyatos, R. & Miralles, D. G. A deep learning-based hybrid model of global terrestrial evaporation. *Nat. Commun.* **13**, 1912 (2022).
101. Lv, X. et al. BTS: a binary tree sampling strategy for object identification based on deep learning. *Int. J. Geogr. Inf. Sci.* **36**, 822–848 (2022).
102. Sun, Z. et al. Improving the performance of automated rooftop extraction through geospatial stratified and optimized sampling. *Remote Sens.* **14**, 4961 (2022).
103. Vandal, T. et al. Deepsd: generating high resolution climate change projections through single image super-resolution. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1663–1672 (ACM, 2017).
104. Lanaras, C., Bioucas-Dias, J., Galliani, S., Baltsavias, E. & Schindler, K. Super-resolution of Sentinel-2 images: learning a globally applicable deep neural network. *ISPRS J. Photogramm. Remote Sens.* **146**, 305–319 (2018).
105. Schneider, T., Lan, S., Stuart, A. & Teixeira, J. Earth System Modeling 2.0: a blueprint for models that learn from observations and targeted high-resolution simulations. *Geophys. Res. Lett.* **44**, 12–396 (2017).
106. Eslami, E., Choi, Y., Lops, Y., Sayeed, A. & Salman, A. K. Using wavelet transform and dynamic time warping to identify the limitations of the CNN model as an air quality forecasting system. *Geosci. Model. Dev.* **13**, 6237–6251 (2020).
107. Han, J., Jentzen, A. & E, W. Solving high-dimensional partial differential equations using deep learning. *Proc. Natl Acad. Sci.* **115**, 8505–8510 (2018).
108. Gagne, D. J., Christensen, H. M., Subramanian, A. C. & Monahan, A. H. Machine learning for stochastic parameterization: generative adversarial networks in the Lorenz'96 model. *J. Adv. Model. Earth Syst.* **12**, e2019MS001896 (2020).
109. Jia, X. et al. Physics-guided machine learning for scientific discovery: an application in simulating lake temperature profiles. *ACM/IMS Trans. Data Sci.* **2**, 1–26 (2021).
110. Zhu, X. et al. Comparison of two optimized machine learning models for predicting displacement of rainfall-induced landslide: a case study in Sichuan Province, China. *Eng. Geol.* **218**, 213–222 (2017).
111. Rasp, S. Coupled online learning as a way to tackle instabilities and biases in neural network parameterizations: general algorithms and Lorenz 96 case study (v1. 0). *Geosci. Model. Dev.* **13**, 2185–2196 (2020).
112. Yang, Z. et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *J. Thorac. Dis.* **12**, 165 (2020).
113. Qian, Z., Liu, X., Tao, F. & Zhou, T. Identification of urban functional areas by coupling satellite images and taxi GPS trajectories. *Remote Sens.* **12**, 2449 (2020).
114. Amini, A., Dolatshahi, M. & Kerachian, R. Adaptive precipitation nowcasting using deep learning and ensemble modeling. *J. Hydrol.* **612**, 128197 (2022).
115. Li, G. & Choi, Y. HPC cluster-based user-defined data integration platform for deep learning in geoscience applications. *Comput. Geosci.* **155**, 104868 (2021).
116. Sun, A. Y. & Scanlon, B. R. How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. *Environ. Res. Lett.* **14**, 073001 (2019).
117. Venkatakrishnan, S. V., Bouman, C. A. & Wohlberg, B. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing* 945–948 (IEEE, 2013).
118. Shen, H. et al. Coupling model- and data-driven methods for remote sensing image restoration and fusion: improving physical interpretability. *IEEE Geosci. Remote Sens. Mag.* **10**, 231–249 (2022).
119. Goldstein, E. B. & Coco, G. Machine learning components in deterministic models: hybrid synergy in the age of data. *Front. Environ. Sci.* **3**, 00033 (2015).
120. Gelbrecht, M., Boers, N. & Kurths, J. Neural partial differential equations for chaotic systems. *New J. Phys.* **23**, 043005 (2021).
121. Mahjoubi, S., Barhemat, R., Guo, P., Meng, W. & Bao, Y. Prediction and multi-objective optimization of mechanical, economical, and environmental properties for strain-hardening cementitious composites (SHCC) based on automated machine learning and metaheuristic algorithms. *J. Clean. Prod.* **329**, 129665 (2021).
122. Goldstein, E. B., Coco, G., Murray, A. B. & Green, M. O. Data-driven components in a model of inner-shelf sorted bedforms: a new hybrid model. *Earth Surf. Dynam.* **2**, 67–82 (2014).
123. Bar-Sinai, Y., Hoyer, S., Hickey, J. & Brenner, M. P. Learning data-driven discretizations for partial differential equations. *Proc. Natl Acad. Sci. USA* **116**, 15344–15349 (2019).
124. Kraft, B., Jung, M., Körner, M., Koirala, S. & Reichstein, M. Towards hybrid modeling of the global hydrological cycle. *Hydrol. Earth Syst. Sci.* **26**, 1579–1614 (2022).
125. Zubov, K. et al. NeuralPDE: automating physics-informed neural networks (PINNs) with error approximations. https://doi.org/10.48550/arXiv.2107.09443 (2021).
126. Chen, T. Q., Rubanova, Y., Bettencourt, J. & Duvenaud, D. K. Neural ordinary differential equations. *Adv. Neural Inf. Process. Syst.* **31**, 6571–6583 (2018).
127. Hagenauer, J. & Helbich, M. A geographically weighted artificial neural network. *Int. J. Geogr. Inf. Sci.* **36**, 215–235 (2022).
128. Li, H. & Weng, Y. Physical equation discovery using physics-consistent neural network (PCNN) under incomplete observability. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* 925–933 (ACM, 2021).
129. Ma, Z. et al. Customizable process design for collaborative geographic analysis. *GISci. Remote Sens.* **59**, 914–935 (2022).
130. Zhang, W. et al. Application of machine learning, deep learning and optimization algorithms in geoengineering and geoscience: comprehensive review and future challenge. *Gondwana Res.* **109**, 1–17 (2022).
131. Massonnet, F. et al. Replicability of the EC-Earth3 Earth system model under a change in computing environment. *Geosci. Model. Dev.* **13**, 1165–1178 (2020).
132. Warnat-Herresthal, S. et al. Swarm learning for decentralized and confidential clinical machine learning. *Nature* **594**, 265–270 (2021).

133. Rackauckas, C. & Nie, Q. Differentialequations.jl — a performant and feature-rich ecosystem for solving differential equations in Julia. *J. Open Res. Softw.* **5**, 151 (2017).
134. Stokel-Walker, C. AI bot ChatGPT writes smart essays — should academics worry? *Nature* https://doi.org/10.1038/d41586-022-04397-7 (2022).
135. Diffenbaugh, N. S. et al. The COVID-19 lockdowns: a window into the Earth System. *Nat. Rev. Earth Environ.* **1**, 470–481 (2020).
136. Zhu, R. et al. The effects of different travel modes and travel destinations on COVID-19 transmission in global cities. *Sci. Bull.* **67**, 588–592 (2022).
137. Wen, W., Yang, S., Zhou, P. & Gao, S. Impacts of COVID-19 on the electric vehicle industry: evidence from China. *Renew. Sustain. Energy Rev.* **144**, 111024 (2021).
138. Hoang, A. T. et al. Impacts of COVID-19 pandemic on the global energy system and the shift progress to renewable energy: opportunities, challenges, and policy implications. *Energy Policy* **154**, 112322 (2021).
139. Wu, X., Nethery, R. C., Sabath, M. B., Braun, D. & Dominici, F. Air pollution and COVID-19 mortality in the United States: strengths and limitations of an ecological regression analysis. *Sci. Adv.* **6**, eabd4049 (2020).
140. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 1–9 (2016).
141. Barton, C. M. et al. How to make models more useful. *Proc. Natl Acad. Sci. USA* **119**, e2202112119 (2022).
142. Chen, M. et al. Teamwork-oriented integrated modeling method for geo-problem solving. *Environ. Model. Softw.* **119**, 111–123 (2019).
143. Tucker, G. E. et al. CSDMS: a community platform for numerical modeling of Earth surface processes. *Geosci. Model. Dev.* **15**, 1413–1439 (2022).
144. Rollins, N. D., Barton, C. M., Bergin, S., Janssen, M. A. & Lee, A. A computational model library for publishing model documentation and code. *Environ. Model. Softw.* **61**, 59–64 (2014).
145. Janssen, M. A., Pritchard, C. & Lee, A. On code sharing and model documentation of published individual and agent-based models. *Environ. Model. Softw.* **134**, 104873 (2020).
146. Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J. & Müller, K.-R. Explaining deep neural networks and beyond: a review of methods and applications. *Proc. IEEE* **109**, 247–278 (2021).
147. Irrgang, C. et al. Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. *Nat. Mach. Intell.* **3**, 667–674 (2021).
148. Trapp, R. J. et al. Changes in severe thunderstorm environment frequency during the 21st century caused by anthropogenically enhanced global radiative forcing. *Proc. Natl Acad. Sci. USA* **104**, 19719–19723 (2007).
149. Ma, Z. & Mei, G. Deep learning for geological hazards analysis: data, models, applications, and opportunities. *Earth Sci. Rev.* **223**, 103858 (2021).
150. Diffenbaugh, N. S. & Field, C. B. Changes in ecologically critical terrestrial climate conditions. *Science* **341**, 486–492 (2013).
151. Syvitski, J. P., Vörösmarty, C. J., Kettner, A. J. & Green, P. Impact of humans on the flux of terrestrial sediment to the global coastal ocean. *Science* **308**, 376–380 (2005).
152. Goswami, B. et al. Abrupt transitions in time series with uncertainties. *Nat. Commun.* **9**, 48 (2018).
153. Zhu, Z. et al. Documentation strategy for facilitating the reproducibility of geo-simulation experiments. *Environ. Model. Softw.* **163**, 105687 (2023).
154. Burnash, R. in *Computer Models of Watershed Hydrology* (ed. Singh, V.) 311–366 (Water Resources Publication, 1995).
155. Tikhamarine, Y. et al. Rainfall-runoff modelling using improved machine learning methods: Harris hawks optimizer vs. particle swarm optimization. *J. Hydrol.* **589**, 125133 (2020).
156. Harbaugh, A. W. *MODFLOW-2005, the US Geological Survey Modular Ground-water Model: The Ground-water Flow Process* Vol. 6 (US Department of the Interior, US Geological Survey Reston, 2005).
157. Ali, A. S. A., Ebrahimi, S., Ashiq, M. M., Alasta, M. S. & Azari, B. CNN-Bi LSTM neural network for simulating groundwater level. *Environ. Eng.* **8**, 1–7 (2022).
158. Laflen, J. M., Lane, L. J. & Foster, G. R. WEPP: a new generation of erosion prediction technology. *J. Soil Water Conserv.* **46**, 34–38 (1991).
159. Senanayake, S. & Pradhan, B. Predicting soil erosion susceptibility associated with climate change scenarios in the Central Highlands of Sri Lanka. *J. Environ. Manag.* **308**, 114589 (2022).
160. Ferro, V. & Porto, P. Sediment delivery distributed (SEDD) model. *J. Hydrol. Eng.* **5**, 411–422 (2000).
161. Buscombe, D. SediNet: a configurable deep learning model for mixed qualitative and quantitative optical granulometry. *Earth Surf. Process. Landf.* **45**, 638–651 (2020).
162. Turner, D. B. *Workbook of Atmospheric Dispersion Estimates: An Introduction to Dispersion Modeling* (CRC, 2020).
163. Zhang, Q., Fu, F. & Tian, R. A deep learning and image-based model for air quality estimation. *Sci. Total Environ.* **724**, 138178 (2020).
164. Skamarock, W. C. et al. *A Description of the Advanced Research WRF Version 2* (National Center for Atmospheric Research, Mesoscale and Microscale Meteorology Division, 2005).
165. Sato, H., Itoh, A. & Kohyama, T. SEIB–DGVM: a new dynamic global vegetation model using a spatially explicit individual-based approach. *Ecol. Model.* **200**, 279–307 (2007).
166. Jung, M. et al. The FLUXCOM ensemble of global land–atmosphere energy fluxes. *Sci. Data* **6**, 74 (2019).
167. Mondal, B. et al. Urban expansion and wetland shrinkage estimation using a GIS-based model in the East Kolkata wetland, India. *Ecol. Indic.* **83**, 62–73 (2017).
168. Wang, X. et al. Estimation of soil salt content (SSC) in the Ebinur Lake Wetland National Nature Reserve (ELWNNR), northwest China, based on a Bootstrap-BP neural network model and optimal spectral indices. *Sci. Total Environ.* **615**, 918–930 (2018).
169. Bailey, N. T. et al. *The Mathematical Theory of Infectious Diseases and Its Applications* (Griffin, 1975).
170. Karemera, D., Oguledo, V. I. & Davis, B. A gravity model analysis of international migration to North America. *Appl. Econ.* **32**, 1745–1755 (2000).
171. Simini, F., Barlacchi, G., Luca, M. & Pappalardo, L. A deep gravity model for mobility flows generation. *Nat. Commun.* **12**, 6576 (2021).

[1]Key Laboratory of Virtual Geographic Environment (Ministry of Education of PRC), Nanjing Normal University, Nanjing, Jiangsu, China. [2]International Research Center of Big Data for Sustainable Development Goals, Beijing, China. [3]Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, Jiangsu, China. [4]State Key Laboratory Cultivation Base of Geographical Environment Evolution, Nanjing, Jiangsu, China. [5]School of Engineering and Design, Technical University of Munich, Munich, Germany. [6]Potsdam Institute for Climate Impact Research, Potsdam, Germany. [7]Global Systems Institute and Department of Mathematics, University of Exeter, Exeter, UK. [8]Fenner School of Environment and Society, Australian National University, Canberra, Australia. [9]Institute of Arctic and Alpine Research (INSTAAR), University of Colorado, Boulder, CO, USA. [10]Department of Geosciences and Natural Resource Management, University of Copenhagen, Copenhagen, Denmark. [11]Department of Geography and Resource Management, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China. [12]Institute of Space and Earth Information Science, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China. [13]The Bartlett Centre for Advanced Spatial Analysis (CASA), University

College London (UCL), London, UK. [14]The Alan Turing Institute, The British Library, London, UK. [15]School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, USA. [16]Systems Science Department, Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [17]Department of Geography, National University of Singapore, Singapore, Singapore. [18]Department of Civil and Construction Engineering, Brigham Young University, Provo, UT, USA. [19]School of Human Evolution and Social Change, Arizona State University, Tempe, AZ, USA. [20]CSIRO Environment, Canberra, Australian Capital Territory, Australia. [21]Department of Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena, Germany. [22]Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China. [23]Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, USA. [24]Center for the Pan-Third Pole Environment, Lanzhou University, Lanzhou, China. [25]College of Geography and Marine, Nanjing University, Nanjing, Jiangsu, China. [26]School of Information Engineering, China University of Geosciences in Beijing, Beijing, China. [27]Key Laboratory of Poyang Lake Wetland and Watershed Research, Ministry of Education, Jiangxi Normal University, Nanchang, Jiangxi, China. [28]These authors contributed equally: Min Chen, Zhen Qian.