PERSPECTIVE

https://doi.org/10.1038/s41586-019-0912-1

Deep learning and process understanding for data-driven Earth system science

Markus Reichstein[1,2]*, Gustau Camps-Valls[3], Bjorn Stevens[4], Martin Jung[1], Joachim Denzler[2,5], Nuno Carvalhais[1,6] & Prabhat[7]

Machine learning approaches are increasingly used to extract patterns and insights from the ever-increasing stream of geospatial data, but current approaches may not be optimal when system behaviour is dominated by spatial or temporal context. Here, rather than amending classical machine learning, we argue that these contextual cues should be used as part of deep learning (an approach that is able to extract spatio-temporal features automatically) to gain further process understanding of Earth system science problems, improving the predictive ability of seasonal forecasting and modelling of long-range spatial connections across multiple timescales, for example. The next step will be a hybrid modelling approach, coupling physical process models with the versatility of data-driven machine learning.

Humans have always striven to predict and understand the world, and the ability to make better predictions has given competitive advantages in diverse contexts (such as weather, diseases or financial markets). Yet the tools for prediction have substantially changed over time, from ancient Greek philosophical reasoning to non-scientific medieval methods such as soothsaying, towards modern scientific discourse, which has come to include hypothesis testing, theory development and computer modelling underpinned by statistical and physical relationships, that is, laws[1]. A success story in the geosciences is weather prediction, which has greatly improved through the integration of better theory, increased computational power, and established observational systems, which allow for the assimilation of large amounts of data into the modelling system[2]. Nevertheless, we can accurately predict the evolution of the weather on a timescale of days, not months. Seasonal meteorological predictions, forecasting extreme events such as flooding or fire, and long-term climate projections are still major challenges. This is especially true for predicting dynamics in the biosphere, which is dominated by biologically mediated processes such as growth or reproduction, and is strongly controlled by seemingly stochastic disturbances such as fires and landslides. Such predictive problems have not seen much progress in the past few decades[3].

At the same time, a deluge of Earth system data has become available, with storage volumes already well beyond dozens of petabytes and rapidly increasing transmission rates exceeding hundreds of terabytes per day[4]. These data come from a plethora of sensors measuring states, fluxes and intensive or time/space-integrated variables, representing fifteen or more orders of temporal and spatial magnitude. They include remote sensing from a few metres to hundreds of kilometres above Earth as well as in situ observations (increasingly from autonomous sensors) at and below the surface and in the atmosphere, many of which are further being complemented by citizen science observations. Model simulation output adds to this deluge; the CMIP-5 dataset of the Climate Model Intercomparison Project, used extensively for scientific groundwork towards periodic climate assessments, is over 3 petabytes in size, and the next generation, CMIP-6, is estimated to reach up to 30 petabytes[5]. The data from models share many of the challenges and statistical properties of observational data, including many forms of uncertainty. In summary, Earth system data are exemplary of all four of the 'four Vs' of 'big data': volume, velocity, variety and veracity (see Fig. 1). One key challenge is to extract interpretable information and knowledge from this big data, possibly almost in real time and integrating between disciplines.

Taken together, our ability to collect and create data far outpaces our ability to sensibly assimilate it, let alone understand it. Predictive ability in the last few decades has not increased apace with data availability. To get the most out of the explosive growth and diversity of Earth system data, we face two major tasks in the coming years: (1) extracting knowledge from the data deluge, and (2) deriving models that learn much more from data than traditional data assimilation approaches can, while still respecting our evolving understanding of nature's laws.

The combination of unprecedented data sources, increased computational power, and the recent advances in statistical modelling and machine learning offer exciting new opportunities for expanding our knowledge about the Earth system from data. In particular, many tools are available from the fields of machine learning and artificial intelligence, but they need to be further developed and adapted to geo-scientific analysis. Earth system science offers new opportunities, challenges and methodological demands, in particular for recent research lines focusing on spatio-temporal context and uncertainties (Box 1; see https://developers.google.com/machine-learning/glossary/ and http://www.wildml.com/deep-learning-glossary/ for more complete glossaries).

In the following sections we review the development of machine learning in the geoscientific context, and highlight how deep learning—that is, the automatic extraction of abstract (spatio-temporal) features—has the potential to overcome many of the limitations that have, until now, hindered a more wide-spread adoption of machine learning. We further lay out the most promising but also challenging approaches in combining machine learning with physical modelling.

State-of-the-art geoscientific machine learning

Machine learning is now a successful part of several research-driven and operational geoscientific processing schemes, addressing the atmosphere, the land surface and the ocean, and has co-evolved with data availability over the past decade. Early landmarks in classification of land cover and clouds emerged almost 30 years ago through the coincidence of high-resolution satellite data and the first revival of neural networks[6,7]. Most major machine learning methodological

[1]Department of Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena, Germany. [2]Michael-Stifel-Center Jena for Data-driven and Simulation Science, Jena, Germany. [3]Image Processing Laboratory (IPL), University of València, Valencia, Spain. [4]Max Planck Institute for Meteorology, Hamburg, Germany. [5]Computer Vision Group, Computer Science, Friedrich Schiller University, Jena, Germany. [6]CENSE, Departamento de Ciências e Engenharia do Ambiente, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa, Lisbon, Portugal. [7]National Energy Research Supercomputing Center, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. *e-mail: mreichstein@bgc-jena.mpg.de

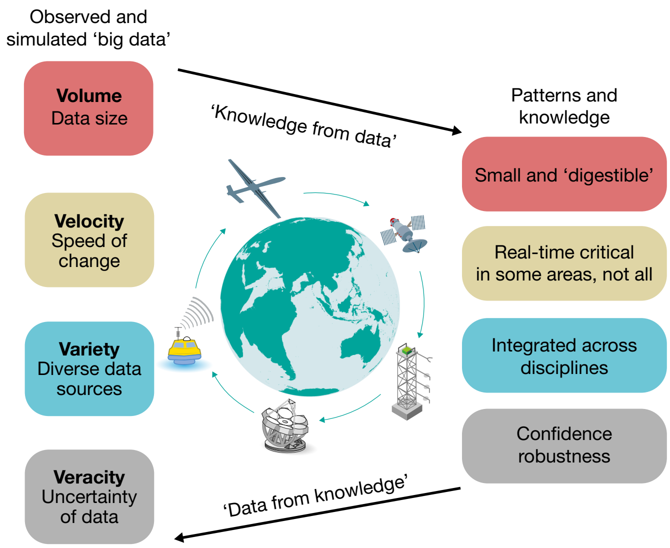

Fig. 1 | Big data challenges in the geoscientific context. Data size now exceeds 100 petabytes, and is growing quasi-exponentially (tapering of the figure to the right indicates decreasing data size.) The speed of change exceeds 5 petabytes a year; data are taken at frequencies of up to 10 Hz or more; reprocessing and versioning are common challenges. Data sources can be one- to four-dimensional, spatially integrated, from the organ level (such as leaves) to the global level. Earth has diverse observational systems, from remote sensing to in situ observation. The uncertainty of data can stem from observational errors or conceptual inconsistencies.

development (for example, kernel methods or 'random forests') has subsequently been applied to geoscience and remote sensing problems, often when data suitable for pertinent methods became available[8]. Thus, machine learning has become a universal approach in geoscientific classification, and change- and anomaly-detection problems[9–12]. In the past few years, geoscience has begun to use deep learning to better exploit spatial and temporal structures in the data, features that would normally be problematic for traditional machine learning to extract (see Table 1, and below).

Another class of problem where machine learning has been successful is regression. An example is soil mapping, where measurements of soil properties and covariates exist at points sparsely distributed in space, and where a 'random forest', a popular and efficient machine learning approach, is used to predict spatially dense estimates of soil properties or soil types[13,14]. In the past decade, machine learning has attained outstanding results in the regression estimation of biogeophysical parameters from remotely sensed reflectances at local and global scales[15–17]. These approaches emphasize spatial prediction, that is, prediction of properties that are relatively static over the observational time period.

Yet what makes the Earth system interesting is that it is not static, but dynamic. Machine learning regression techniques have also been used to study these dynamics by mapping temporally varying features onto temporally varying target variables in land, ocean and atmosphere domains. Since variables such as land–atmosphere or ocean–atmosphere carbon uptake cannot be observed everywhere, one challenge has been to infer continental or global estimates from point observations, by building models that relate climate and remote-sensing co-variates to the target variables. In this context, machine learning methods have proved more powerful and flexible than previous mechanistic or semi-empirical modelling approaches. For instance, an artificial neural network with one hidden layer was able to filter out noise, predict the diurnal and seasonal variation of carbon dioxide (CO_2) fluxes, and extract patterns such as an increased respiration in spring during root growth, which was formerly unquantified and not well represented in carbon cycle models[18]. Further developments have then allowed us to quantify global terrestrial photosynthesis and evapotranspiration of water in a purely data-driven way[19,20]. Spatial, seasonal, interannual or decadal variations of such machine-learning-predicted fluxes are even being used as important benchmarks for physical land-surface and climate model evaluation[21–24]. Similarly, ocean CO_2 concentrations and fluxes have been mapped spatio-temporally with neural networks, where classification and regression approaches have been combined, both for stratifying the data and for prediction[25]. Recently the random forest method has also been used to predict spatio-temporally varying precipitation[26]. Overall, we conclude that a diversity of influential machine learning approaches have already been applied across all the major sub-domains of Earth system science and are increasingly being integrated into operational schemes and being used to discover patterns, to improve our understanding and to evaluate comprehensive physical models.

Notwithstanding the success of machine learning in the geosciences, important caveats and limitations have hampered its wider adoption and impact. A few pitfalls such as the risk of naive extrapolation, sampling or other data biases, ignorance of confounding factors, interpretation of statistical association as causal relation, or fundamental flaws in multiple hypothesis testing ('*P*-fishing')[27–29] should be avoided by best practice and expert intervention. More fundamentally, there are inherent limitations of machine learning approaches as applied at present. It is in this realm that the techniques of deep learning promise breakthroughs.

Classical machine learning approaches benefit from domain-specific, hand-crafted features to account for dependencies in time or space (for example, cumulative precipitation derived from a daily time series), but rarely exploit spatio-temporal dependencies exhaustively. For instance, in ocean–atmosphere or land–atmosphere CO_2 flux prediction[19,25], mapping of instantaneous, local environmental conditions (such as radiation, temperature and humidity) to instantaneous fluxes is performed. In reality, processes at a certain point in time and space are almost always additionally affected by the state of the system, which is often not well observed and thus not available as a predictor. However, previous time steps and neighbouring grid cells contain hidden information on the state of the system (for example, a long period without rainfall combined with sustained sunny days implies a drought). One example where both spatial and temporal context are highly relevant is the prediction of fire occurrence and characteristics such as burnt area and trace gas emissions. Fire occurrence and spread depends not only on instantaneous climatic drivers and sources of ignition (such as humans, lightning or both) but also on state variables, such as the state and amount of available fuel[3]. Fire spread and thus the burned area depends not only on the local conditions of each pixel but also on the spatial arrangement and connectivity of fuel, its moisture, terrain properties, and of course wind speed and direction. Similarly, classifying a certain atmospheric situation as a hurricane or extratropical storm requires knowledge of the spatial context such as a storm's geometry as constituted by pixels, their values, and their topology. For instance, detecting symmetric outflow and a visible 'eye' is important for detecting hurricanes and assessing their strength, and this cannot be determined by localized, single-pixel values alone.

Certainly, temporally dynamic properties ('memory effects') can be represented by hand-designed and domain-specific features in machine learning. Examples are cumulative sums of daily temperature, which are used to predict phenological phases of vegetation, and the standardized precipitation index[30], which summarizes precipitation anomalies over the last months as a meteorological indicator of drought states. Very often, these approaches only consider memory in a single variable, ignoring the interactive effects of several variables, although exceptions exist[22,31].

Machine learning can also use hand-designed features, such as terrain shape and topographical or texture features from satellite images, to incorporate spatial context[6]. This is analogous to earlier approaches in computer vision where objects were often characterized by a set of features describing edges, textures, shapes and colours. Such features were then fed into a standard machine learning algorithm for localization, classification or detection of objects in images.

Box 1

Definition of terms

Term	Explanation
Artificial intelligence, machine learning and deep learning	Artificial intelligence is the capacity of an algorithm to assimilate information to perform tasks that are characteristic of human intelligence, such as recognizing objects and sounds, contextualizing language, learning from the environment, and problem solving. Machine learning is a field of statistical research for training computational algorithms that split, sort and transform a set of data to maximize the ability to classify, predict, cluster or discover patterns in a target dataset. Deep learning refers to machine learning algorithms that construct hierarchical architectures of increasing sophistication. Artificial neural networks with many layers are examples of deep learning algorithms.
Bayesian inference	Bayesian inference is a framework in statistics and machine learning that develops methods for data analysis in which observational evidence is used to update the probability that an hypothesis is true. The framework is mostly concerned about treating uncertainty, encoding prior beliefs and estimating error propagation when dealing with data and models.
Causal inference	Causal inference links events, processes or properties in a system via a cause-and-effect connection. Recent observational causal inference algorithms attempt to discover causal relationships in observational data.
Convolution	Convolution is one of the most important operations in signal and image processing, and it can operate in objects that are one-dimensional (for example, speech), two-dimensional (for example, images) or three-dimensional (for example, video). A convolutional filter is essentially a weighting vector/matrix/cube that uses a sliding-window approach. Depending on the kernel structure, the convolution enhances some features of the data, such as edges, trends or flat regions. Convolution is embedded in convolutional neural networks at the neuron level, which extracts useful features from the previous layers.
Differentiable programming	Differentiable programming refers to a programming paradigm to generate code that is automatically differentiated, such that its parameters can be seamlessly optimized. It generalizes current deep learning frameworks to arbitrary programs, which may include the hybrid modelling approaches that we discuss in 'Integration with physical modelling'.
Feedforward versus recurrent networks	An artificial neural network is a computational algorithm that simulates how signals are transferred between a network of neurons, via synapses. In an artificial neural network, information is transferred only in the forward direction, whereas in a recurrent artificial neural network the information can cycle or loop between the different nodes, creating complex dynamics such as memory, as seen in data.
Generative adversarial networks	This is a family of unsupervised machine learning methods widely used to generate realistic samples from an unknown probability density function. Generative adversarial networks are formed by a neural network that generates plausible examples; these examples are then used to attempt to fool a discriminator network that should discern real from fake examples.
Memory effects	This is a metaphorical term meaning that the current behaviour of a system cannot be explained without considering the effect of past states or forcing variables.
Nowcasting and forecasting	To forecast a certain variable means to establish a prediction of its value in the future, days to centuries from now. Nowcasting refers to making that prediction for a very near future (for example, predicting whether it is going to rain in a couple of hours).
Probabilistic programming	Probabilistic programming is a method of defining probabilistic models using a unified high-level programming language. Statistical inference is automatically achieved by built-in inference machines, freeing the developer from the difficulties of high-performance probabilistic inference.
Radiative transfer models	These are mathematical models that describe how radiation at different wavelengths (such as visible light) propagates through different media (such as the atmosphere or a vegetation canopy) by simulating absorption, emission, transmission and scattering processes.
Remote sensing	Most remote sensing deals with measuring the radiance at different wavelengths reflected or emitted from an object or surface. Remote sensing uses satellite or airborne sensors to detect and classify objects as well as to estimate geoscientific variables of interest (temperature, salinity or carbon dioxide concentrations), based on propagated reflectance signals (such as electromagnetic radiation).
Supervised and unsupervised learning	In supervised learning an algorithm learns the input-to-output relationship after being provided both the inputs and the respective outputs. For example, the input might be a set of photos and the output might be a set of corresponding labels. In unsupervised learning the algorithm does not have access to the output, so the goal is to infer the underlying structure of the data. For example, the algorithm could automatically separate pictures with different statistical or semantic properties (such as a set of images of cats and dogs).
Teleconnections	Teleconnections refer to climate anomalies related to each other at large distances (typically thousands of kilometres). Quantifying teleconnection patterns allows the prediction of key patterns on Earth, which are distant in space and time. For example, predicting an El Niño event enables prediction of North American rainfall, snowfall, droughts or temperature patterns with lead times of a few weeks to months.

Similar approaches have been followed for decades in remote-sensing image classification[8–10]. Hand-designed features can be seen both as an advantage (control of the explanatory drivers) and as a disadvantage (tedious, ad hoc process, probably non-optimal), but certainly the concerns related to the use of a restricted and subjective choice of features rather than an extensive and generic approach remain valid and important. New developments in deep learning, however, no longer limit us to such approaches.

Deep learning opportunities in Earth system science

Deep learning has achieved notable success in modelling ordered sequences and data with spatial context in the fields of computer vision, speech recognition and control systems[32], as well as in related scientific fields in physics[33–35], chemistry[36] and biology[37] (see also ref. [38]). Applications to problems in geosciences are in their infancy, but across the key problems (classification, anomaly detection, regression, space- or time-dependent state prediction) there are promising examples (see Table 1 and Supplementary Fig. 2)[39,40]. Two recent studies demonstrate the application of deep learning to the problem of extreme weather, for instance hurricane detection[41,42]—already mentioned as problematic for traditional machine learning to perform. The studies report success in applying deep learning architectures to objectively extract spatial features to define and classify extreme situations (for example, storms, atmospheric rivers) in numerical weather prediction model output. Such an approach enables rapid detection of such events and forecast simulations without using either subjective human annotation or methods that rely on predefined arbitrary thresholds for wind speed or other variables.

Table 1 | Conventional approaches and deep learning approaches to geoscientific tasks

Analytical task	Scientific task	Conventional approaches	Limitations of conventional approaches	Emergent or potential approaches
Classification and anomaly detection				
	Finding extreme weather patterns	Multivariate, threshold-based detection	Heuristic approach, ad hoc criteria used	Supervised and semi-supervised convolutional neural networks[41,42]
	Land-use and change detection	Pixel-by-pixel spectral classification	Shallow spatial context used, or none	Convolutional neural networks[43]
Regression				
	Predict fluxes from atmospheric conditions	Random forests, kernel methods, feedforward neural networks	Memory and lag effects not considered	Recurrent neural networks, long-short-term-memories (LSTMs)[89,99,100]
	Predict vegetation properties from atmospheric conditions	Semi-empirical algorithms (temperature sums, water deficits)	Prescriptive in terms of functional forms and dynamic assumptions	Recurrent neural networks[90], possibly with spatial context
	Predict river runoff in ungauged catchments	Process models or statistical models with hand-designed topographic features[91]	Consideration of spatial context limited to hand-designed features	Combination of convolutional neural network with recurrent networks
State prediction				
	Precipitation nowcasting	Physical modelling with data assimilation	Computational limits due to resolution, data used only to update states	Convolutional–LSTM nets short-range spatial context[92]
	Downscaling and bias-correcting forecasts	Dynamic modelling and statistical approaches	Computational limits, subjective feature selection	Convolutional nets[72], conditional generative adversarial networks (cGANs)[53,93,101]
	Seasonal forecasts	Physical modelling with initial conditions from data	Fully dependent on physical model, current skill relatively weak	Convolutional–LSTM nets with long-range spatial context
	Transport modelling	Physical modelling of transport	Fully dependent on physical model, computational limits	Hybrid physical–convolutional network models[68,94]

In particular, such an approach uses the information in the spatial shape of events, such as the typical spiral of hurricanes. Similarly, for classification of urban areas the automatic extraction of multi-scale features from remote-sensing data strongly improved the classification accuracy (to almost always greater than 95%)[43].

While deep learning approaches have classically been divided into spatial learning (for example, convolutional neural networks for object classification) and sequence learning (for example, speech recognition), there is a growing interest in blending these two perspectives. A prototypical example is video and motion prediction[44,45], a problem that has striking similarities to many dynamic geoscience problems. Here we are faced with time-evolving multi-dimensional structures, such as organized precipitating convection, which dominates patterns of tropical rainfall, and vegetation states that influence the flow of carbon and evapotranspiration. Studies are beginning to apply combined convolutional–recurrent approaches to geoscientific problems such as precipitation nowcasting (Table 1)[46]. Modelling atmospheric and ocean transport, fire spread, soil movements or vegetation dynamics are other examples of problems where spatio-temporal dynamics are important, but that have yet to benefit from a concerted effort to apply these new approaches.

In short, the similarities between the types of data addressed with classical deep learning applications and geoscientific data make a compelling argument for the integration of deep learning into the geosciences (Fig. 2). Images are analogous to two-dimensional data fields containing particular variables in analogy to colour triplets (RGB values) in photographs, while videos can be linked to a sequence of images and hence to two-dimensional fields that evolve over time. Similarly, natural language and speech signals share the same multi-resolution characteristics of dynamic time series of Earth system variables. Furthermore, classification, regression, anomaly detection, and dynamic modelling are typical problems in both computer vision and the geosciences.

Deep-learning challenges in Earth system science

The similarities between classical deep learning applications and geoscience applications outlined above are striking. Yet numerous differences exist. For example, while classical computer vision applications deal with photos which have three channels (red, green, blue) hyperspectral satellite images extend to hundreds of spectral channels well beyond the visible range, which often induce different statistical properties to those of natural images. This includes spatial dependence and interdependence of variables, violating the important assumption of identically, independently distributed data. Additionally, integrating multi-sensor data is not trivial since different sensors exhibit different imaging geometries, spatial and temporal resolution, physical meaning, content and statistics. Sequences of (multi-sensor) satellite observations also come with diverse noise sources, uncertainty levels, missing data and (often systematic) gaps (owing to the presence of clouds or snow, distortions in acquisition, storage and transmission, and so on).

In addition, spectral, spatial and temporal dimensionalities raise computational challenges. Data volume is increasing and soon it will be necessary to deal with petabytes per day globally. At present, the biggest meteorological agencies have to process terabytes per day in near real time, often at very high (32-bit, 64-bit) precision. Further, while typical computer vision applications have worked with image sizes of 512×512 pixels, a moderate-resolution (around 1 km) global field has sizes of approximately $40{,}000 \times 20{,}000$ pixels, that is, three orders of magnitude more.

Last but not least, unlike on ImageNet (a database of human-labelled images, with labels like, for example, ‘cat’ or ‘dog’[47]), large, labelled geoscientific datasets do not always exist in geoscience, not only because of the sizes of the datasets involved, but also owing to the conceptual difficulty in labelling datasets; for example, determining that an image depicts a cat is much easier than determining that a dataset reflects a drought, given that droughts are contingent on intensity and extent and can change according to the methods used to collect and analyse the data, and that there are not enough labelled cases for training a machine learning system. Besides the challenge of working with a limited training set, geoscientific problems are often under-constrained, leading to the possibility of models thought to be of high quality, which perform well in training and even test datasets, but deviate strongly for situations and data outside their valid domain (the extrapolation problem), which is true even for complex physical

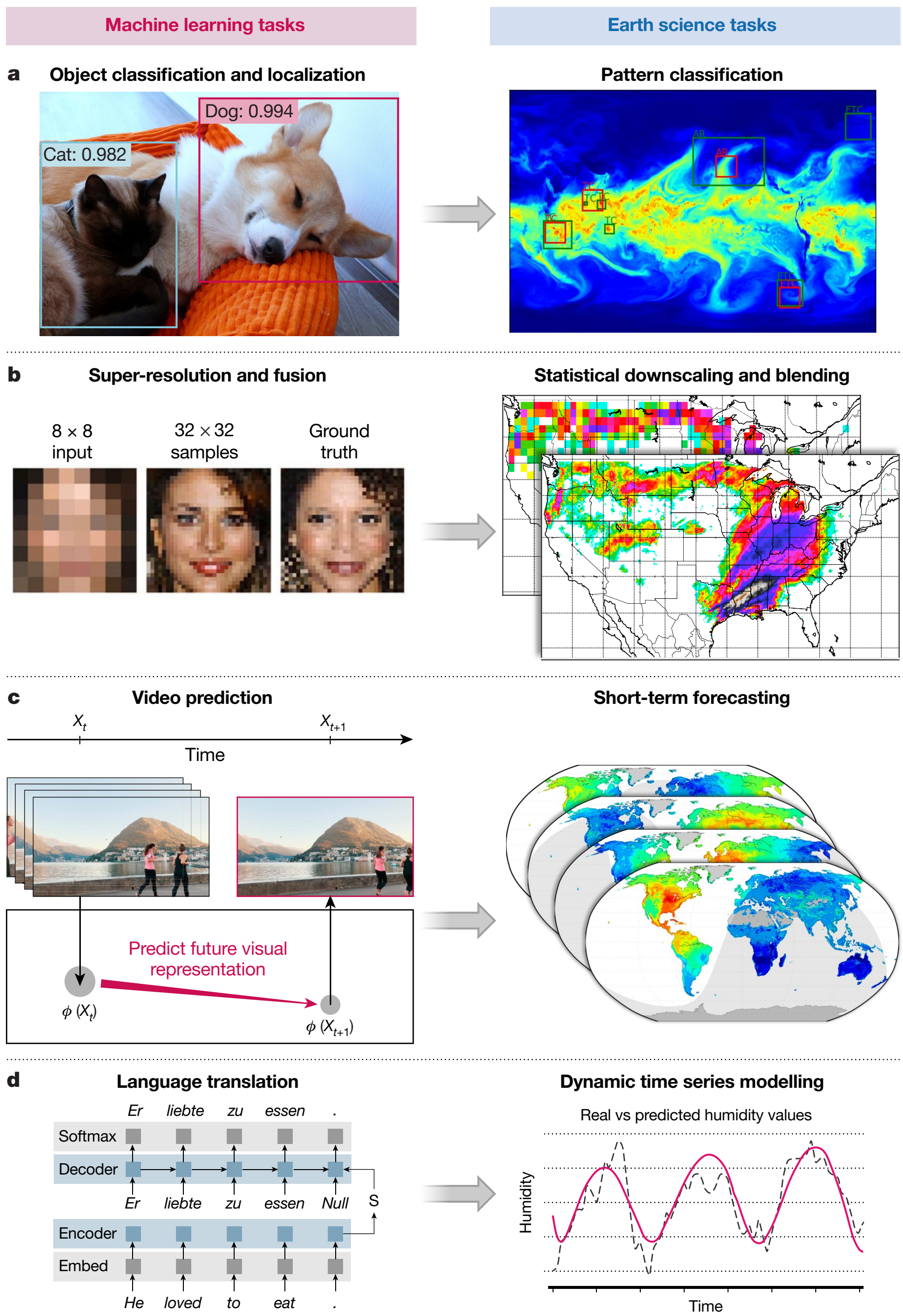

Fig. 2 | Four examples of typical deep learning applications (left panels) and the geoscientific problems they can be applied to (right panels). **a**, Object recognition in images links to classification of extreme weather patterns using a unified convolutional neural network on climate simulation data[41]. **b**, Super-resolution applications relate to statistical downscaling of climate model output[72]. **c**, Video prediction is similar to short-term forecasting of Earth system variables. Right image, courtesy of Sujan Koirala and Paul Bodesheim, Max Planck Institute for Biogeochemistry. **d**, Language translation links to modelling of dynamic time series (ref. [96] and figure 11 in ref. [97]). Left image, courtesy of Stephen Merity (figure 1 in https://smerity.com/articles/2016/google_nmt_arch.html).

Earth system models[48]. Overall, we identify five major challenges and avenues for the successful adoption of deep learning approaches in the geosciences, as follows.

(1) Interpretability

Improving predictive accuracy is important but insufficient. Certainly, interpretability and understanding are crucial, including visualization of the results for analysis by humans. Interpretability has been identified as a potential weakness of deep neural networks, and achieving it is a current focus in deep learning[49]. The field is still far from achieving self-explanatory models, and also far from causal discovery from observational data[50,51]. Yet we should note that, given their complexity, modern Earth system models are in practice often also not easily traceable back to their assumptions, limiting their interpretability too.

(2) Physical consistency

Deep learning models can fit observations very well, but predictions may be physically inconsistent or implausible, owing to extrapolation or observational biases, for example. Integration of domain knowledge and achievement of physical consistency by teaching models about the governing physical rules of the Earth system can provide very strong theoretical constraints on top of the observational ones.

(3) Complex and uncertain data

Deep learning methods are needed to cope with complex statistics, multiple outputs, different noise sources and high-dimensional spaces. New network topologies that not only exploit local neighbourhood (even at different scales), but also long-range relationships (for example, for teleconnections) are urgently needed, but the exact cause-and-effect relations between variables are not clear in advance and need to be discovered. Modelling uncertainties will be certainly an important aspect and will require concepts from Bayesian/probabilistic inference to be integrated, directly addressing such uncertainties (see Box 1 and ref. [52]).

(4) Limited labels

Deep learning methods are needed to learn from few labelled examples while exploiting the wealth of information in related unlabelled observations. These methods include unsupervised density modelling, feature extraction, semi-supervised learning and domain adaptation[53] (see Box 1).

(5) Computational demand

There is a huge technical challenge regarding the high computational cost of current geoscience problems—an excellent example of how to address this is Google's Earth Engine, which allowed the solution of real problems from deforestation[54] to lake[55] monitoring and is expected to follow up with deep learning applications in the future.

By addressing these challenges, deep learning could make an even bigger difference in the geosciences than in classical computer vision, because in computer vision hand-crafted features are derived from a clear understanding of the world (existence of surfaces, boundaries between objects, and so on), the mapping from the world to images, and assumptions about the (visual) appearance of world points (surface points) on two-dimensional images. Assumptions for successful processing include the assumption of Lambertian surfaces (that is, intensity does not depend on the angle between surface and light source), which results in the classical assumption of the constant intensity of the observation of a three-dimensional point over time. In addition, changes in the world (the motion of objects) are in most cases modelled as rigid transformations, or non-rigid transformations that arise from physical assumptions and that are only valid locally (such as in registration of brain structures, before and after removal of a tumour). Even complex problems in computer vision have been solved by hand-crafted features that reflect the assumptions and expectations that arise from common world knowledge. In geoscience and climate science, such global, general knowledge is still partly missing, and indeed, is exactly what we are seeking in research (hence, it cannot be an assumption). All problems, from segmentation in remote-sensing images to regression analysis of certain variables, have certain assumptions that are known to be valid or at least good approximations. Yet the less well processes are understood, the fewer high-quality hand-crafted features for modelling can be expected to exist. Thus, deep learning methods, particularly since they find a good representation from data, represent an opportunity with which to tackle geoscience and climate research problems.

The most promising near-future applications include nowcasting (that is, prediction of the very near future, up to two hours in meteorology) and forecasting applications, anomaly detection and classification based on spatial and temporal context information (see examples in Table 1). A longer-term vision includes data-driven seasonal forecasting, modelling of spatial long-range correlations across multiple timescales, modelling spatial dynamics where spatial context is important (for example, fires), and detecting teleconnections and connections between variables that a human may not have thought about.

We infer that deep learning will soon be the leading method for classifying and predicting space-time structures in the geosciences. More challenging is to gain understanding in addition to optimal prediction, and to achieve models that have maximally learned from data, while still taking into account physical and biological knowledge. One promising but largely uncharted approach to achieving this goal is the integration of machine learning with physical modelling, which we discuss next.

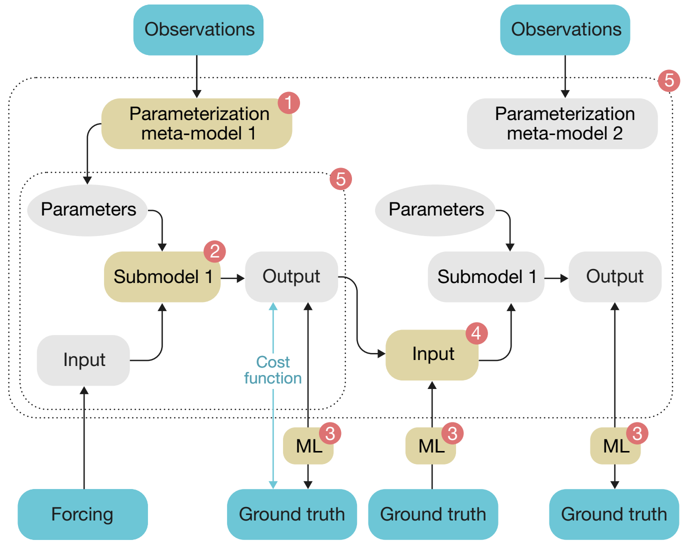

Fig. 3 | Linkages between physical models and machine learning. An abstraction of a part of a physical system—for example, an Earth system model—is depicted here. The model consists of submodels; each submodel has parameters and forcing variables as inputs and produces output, which can be input (forcing) to another sub-model. Data-driven learning approaches can be helpful in various instances, as indicated by the circled numbers. For example, the circle labelled 2 represents hybrid modelling. See the text for more detail. ML, machine learning.

Integration with physical modelling

Historically, physical modelling and machine learning have often been treated as two different fields with very different scientific paradigms (theory-driven versus data-driven). Yet, in fact these approaches are complementary, with physical approaches in principle being directly interpretable and offering the potential of extrapolation beyond observed conditions, whereas data-driven approaches are highly flexible in adapting to data and are amenable to finding unexpected patterns (surprises). The synergy between the two approaches has been gaining attention[56–58], expressed in benchmarking initiatives[59,60] and in concepts such as emergent constraints[27,61,62].

Here we argue that advances in machine learning and in observational and simulation capabilities within Earth sciences offer an opportunity to integrate simulation and data science approaches more intensively in multiple ways. From a systems modelling point of view there are five points of potential synergy (see Fig. 3, in which the numbered circles correspond to the following numbered list).

(1) Improving parameterizations

See Fig. 3 (circle 1). Physical models require parameters, but many of those cannot be easily derived from first principles. Machine learning can learn parameterizations to optimally describe the ground truth that can be observed or generated from detailed and high-resolution models through first principles. For example, instead of assigning parameters of the vegetation in an Earth system model to plant functional types (a common ad hoc decision in most global land surface models), one can allow these parameterizations to be learned from appropriate sets of statistical covariates, allowing them to be more dynamic, interdependent and contextual. A prototypical approach has been taken already in hydrology where the mapping of environmental variables (for example, precipitation and surface slope) to catchment parameters (such as mean, minimum and maximum streamflow) has been learned from a few thousand catchments and applied globally to feed hydrological models[63]. Another example from global atmospheric modelling is learning the effective coarse-scale physical parameters of precipitating convection (for example, the fraction of water that is precipitating out of a cloud during convection) from data or high-resolution models[64,65] (the high-resolution models are too expensive to run, which is why coarse-scale parametrizations are needed).

These learned parametrizations could lead to better representations of tropical convection[66,67].

(2) Replacing a 'physical' sub-model with a machine learning model

See Fig. 3 (circle 2). If formulations of a submodel are of semi-empirical nature, where the functional form has little theoretical basis (for example, biological processes), this submodel can be replaced by a machine learning model if a sufficient number of observations are available. This leads to a hybrid model, which combines the strengths of physical modelling (theoretical foundations, interpretable compartments) and machine learning (data-adaptiveness). For example, we could couple well established physical (differential) equations of diffusion for transport of water in plants with machine learning for the poorly understood biological regulation of water transport conductance. This results in a more 'physical' model that obeys accepted conservation of mass and energy laws, but its regulation (biological) is flexible and learned from data. Such principles have recently been taken to efficiently model motion of water in the ocean and specifically predict sea surface temperatures. Here, the motion field was learned via a deep neural network, and then used to update the heat content and temperatures via physically modelling the movement implied by the motion field[68]. Also, a number of atmospheric scientists have begun experimenting with related approaches to circumvent long-standing biases in physically based parameterizations of atmospheric convection[65,69].

The problem may become more complicated if physical model and machine learning parameters are to be estimated simultaneously while maintaining interpretability, especially when several sub-models are replaced with machine learning approaches. In the field of chemistry this approach has been used in calibration exercises and to describe changes in unknown kinetic rates while maintaining mass balance in biochemical reactor modelling[70], which, although less complex, bears many similarities to hydrological and biogeochemical modelling.

(3) Analysis of model–observation mismatch

See Fig. 3 (circle 3). Deviations of a physical model from observations can be perceived as imperfect knowledge causing model error, assuming no observational biases. Machine learning can help to identify, visualize and understand the patterns of model error, which allows us also to correct model outputs accordingly. For example, machine learning can extract patterns from data automatically and identify those which are not explicitly represented in the physical model. This approach helps to improve the physical model and theory. In practice, it can also serve to correct the model bias of dynamic variables, or it can facilitate improved downscaling to finer spatial scales compared to tedious and ad hoc hand-designed approaches[71,72].

(4) Constraining submodels

See Fig. 3 (circle 4). One can drive a submodel with the output from a machine learning algorithm, instead of another (potentially biased) submodel in an offline simulation. This helps to disentangle model error originating from the submodule of interest from errors of coupled submodules. As a consequence, this simplifies and reduces biases and uncertainties in model parameter calibration or the assimilation of observed system state variables.

(5) Surrogate modelling or emulation

See Fig. 3 (circle 5). Emulation of the full (or specific parts of) a physical model can be useful for computational efficiency and tractability reasons. Machine learning emulators, once trained, can achieve simulations orders of magnitude faster than the original physical model without sacrificing much accuracy. This allows for fast sensitivity analysis, model parameter calibration, and derivation of confidence intervals for the estimates. For example, machine learning emulators are used to replace computationally expensive, physics-based radiative-transfer models of the interactions between radiation, vegetation and atmosphere[57,73,74], which are critical for the interpretation and assimilation of land-surface remote sensing in models. Emulators are also used in dynamic modelling, where states are evolving, for example, in climate modelling[75] and more recently explored in vegetation dynamic models[76]. Further, given the complexity of physical models, emulation challenges are very good test beds in which to explore the potential of machine learning and deep learning approaches to extrapolate outside the range of training conditions.

Some of the concepts in Fig. 3 have already been adopted in a broad sense. For instance, linkage (3) relates to model benchmarking and statistical downscaling and model output statistics[77,78]. Here we argue that adopting a deep-learning approach will strongly improve the use of spatio-temporal context information for the modification of model output. Emulation (5) has been widely adopted in several branches of engineering and geosciences, mainly for the sake of efficient modelling, but tractability issues have not yet been explored in depth. Other paths, such as the hybrid modelling (linkage (2)), appear to be much less explored. Conceptually, the hybrid approaches discussed above can be interpreted as deepening a neural network (Fig. 4) to make it more physically realistic, where the physical model comes on top of a neural network layers (see examples in Fig. 4b, c). It contrasts with the reverse approach discussed above where physical model output is produced and then corrected using additional layers of machine learning approaches. We believe that it is worthwhile pursuing both avenues of integrating physical modelling and machine learning.

Figure 3 presents a system-modelling view that seeks to integrate machine learning into a system model. As an alternative perspective, system knowledge can be integrated into a machine learning framework. This may include design of the network architecture[36,79], physical constraints in the cost function for optimization[58], or expansion of the training dataset for undersampled domains (that is, physically based data augmentation)[80]. For instance, while usually a so-called cost function like ordinary least squares penalizes model–data mismatch, it can be modified to also avoid physically implausible predictions for lake temperature modelling[58]. The integration of physics and machine learning models may not only achieve improved performance and generalizations but, perhaps more importantly, incorporates the consistency and credibility of the machine learning models. As a byproduct, the hybridization has an interesting regularization effect, given that the physics discards implausible models. Therefore, physics-aware machine learning models should combat overfitting better, especially in low-to-medium sample sized datasets[81]. This notion is also related to the direction of attaining explainable and interpretable machine learning models ('explainable AI'[82]), and to combining logic rules with deep neural networks[83].

Recent advances in two fields of methodological approaches have potential in facilitating the fusion of machine learning and physical models in a sound way: probabilistic programming[52] and differentiable programming. Probabilistic programming allows for accounting of various uncertainty aspects in a formal but flexible way. A proper accounting for data and model uncertainty along with integration of knowledge by priors and constraints is critical for optimally combining the data-driven and theory-driven paradigms, including logical rules, as done in statistical relational learning. In addition, error propagation is conceptually seamless, facilitating well founded uncertainty margins for model output. This capability is largely missing so far but is crucial for scientific purposes, and in particular for management or policy decisions. Differentiable programming allows for efficient optimization owing to automated differentiation[84,85]. This helps in making the large, nonlinear and complex inversion problem computationally more tractable, and in addition allows for explicit sensitivity assessments, thus aiding interpretability.

Advancing science

There is no doubt, as exemplified in this Perspective, that modern machine learning methods greatly improve classification and prediction skills. This alone has great value. Yet, beyond statistical prediction, the question is how data-driven approaches can improve fundamental

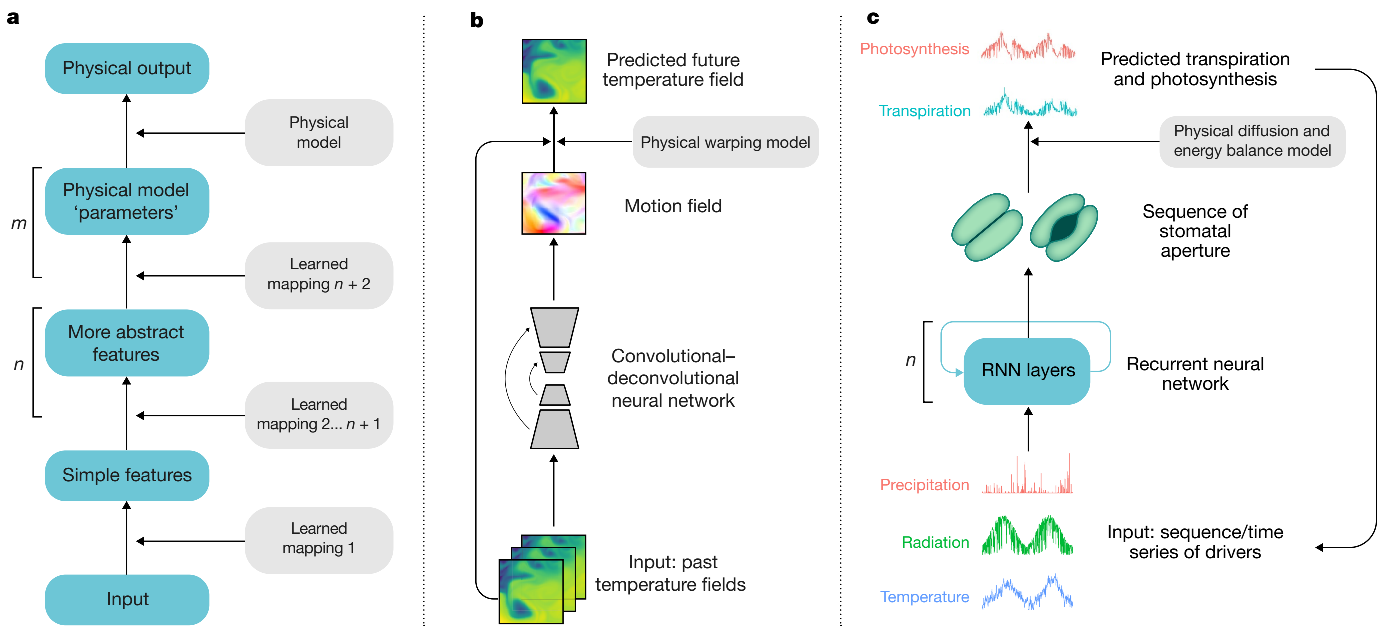

Fig. 4 | Interpretation of hybrid modelling as deepening a deep learning architecture by adding one or several physical layers after the multilayer neural network to make the model more physically realistic. a, The multilayer neural network, with *n* the number of neural layers and *m* the number of physical layers. **b** and **c** are concrete examples of hybrid modelling (circle 2 in Fig. 3). **b**, Prediction of sea-surface temperatures, where a motion field of the water is learned with a convolutional–deconvolutional neural network, and the motion field is further processed with a physical model to predict future states. Adapted from figure 1 of de Bezenac et al.[68]. **c**, A biological regulation process (opening of the stomatal 'valves' controlling water vapour flux from the leaves) is modelled with a recurrent neural network. Then a physical diffusion model is used to estimate transpiration, which in turn influences some of the drivers, such as soil moisture. The basic scheme in **a** is inspired by figure 1.5 in Goodfellow et al.[98] and redrawn.

scientific understanding, given that the outcome of complex statistical models, in particular, is often hard to interpret. One basic answer is that observations have almost always been the basis for scientific progress. For example, the Copernican discoveries were enabled by the precise observation of planetary trajectories to infer and test the laws governing them.

Now, although the general cycle of exploration, hypotheses generation and testing remains the same, modern data-driven science and machine learning can extract arbitrarily complex patterns in observational data to challenge complex theories and Earth system models (Supplementary Fig. 3). For instance, spatially explicit global data-driven estimates of photosynthesis based on machine learning have indicated an overestimation of photosynthesis in the tropical rainforest by climate models[86]. This mismatch has led scientists to develop hypotheses that enable a better description of the radiative transfer in vegetation canopies[23], which has led to better photosynthesis estimates in other regions, and better consistency with leaf-level observations. Related data-driven carbon cycle estimates have enabled the calibration of vegetation models and helped to explain the conundrum of the increasing seasonal amplitude of CO_2 concentration at high latitudes[87], which (according to these results) is caused by vegetation being more vigorous in the high latitudes.

In addition to data-driven theory and model building, such extracted patterns are increasingly being used as a way to explore improved parameterizations in Earth system models[65,69], and model emulators are increasingly being used as a basis for model calibration[88]. In this way, the scientific interplay between theory and observation, and between hypothesis generation and theory-driven hypothesis testing, will continue. Since the complexity of hypotheses and tests inferred from data and the pace of hypothesis generation are increasing by orders of magnitude via powerful machine learning techniques, we can expect unprecedented qualitative and quantitative progress in the science of the complex Earth system.

Conclusion

Earth sciences need to process large and rapidly increasing amounts of data to provide more accurate, less uncertain, and physically consistent inferences in the form of prediction, modelling and understanding the complex Earth system. Machine learning in general and deep learning in particular offer promising tools to build new data-driven models for components of the Earth system and thus to build our understanding of Earth. Challenges specific to the Earth system will further stimulate the development of methodologies, and we have four major recommendations, as follows.

(1) Recognition of the particularities of the data

Multi-source, multi-scale, high-dimensional, complex spatio-temporal relations, including non-trivial and lagged long-distance relationships (teleconnections) between variables need to be adequately modelled. Deep learning is well positioned to address these data challenges, and network architectures and algorithms need to be developed to produce approaches that address both spatial and temporal context at different scales (see Fig. 4).

(2) Plausibility and interpretability of inferences

Models should not only be accurate but also credible, incorporating the physics governing the Earth system. Wide adoption of machine learning in the Earth sciences will be facilitated if models become more transparent and interpretable: their parameters and feature rankings should have a minimal physical interpretation, and the model should be reducible to or explainable by a set of rules, descriptors and relations.

(3) Uncertainty estimation

Models should define their confidence and credibility. Bayesian/probabilistic inference should be integrated into models, because such inference allows for explicit representation and propagation of uncertainties. In addition, identifying and treating extrapolation is a priority.

(4) Testing against complex physical models

The spatial and temporal prediction ability of machine learning should be, at least, consistent with the patterns observed in physical models. Therefore we recommend testing the performance of machine learning methods against synthetic data derived from physical models of the Earth system. For instance, the models in Fig. 4b, c, which are applied to real data, should be tested across a broad range of dynamics as simulated by complex physical models. This is of particular relevance in conditions of limited training data and to assess extrapolation issues.

Overall, we suggest that future models should integrate process-based and machine learning approaches. Data-driven machine learning approaches to geoscientific research will not replace physical modelling, but strongly complement and enrich it. Specifically, we envision various synergies between physical and data-driven models, with the ultimate goal of hybrid modelling approaches: these should obey physical laws, feature a conceptualized and thus interpretable structure, and at the same time be fully data-adaptive where theory is weak. Importantly, machine learning research will benefit from plausible physically based relationships derived from the natural sciences. Among others, two major Earth system challenges in which little progress has been made recently—the parameterization of atmospheric convection and the description of the spatio-temporal dependency of ecosystems on climate and interacting geo-factors—could be addressed using the hybrid approaches discussed here.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-019-0912-1.

Received: 28 July 2017; Accepted: 5 December 2018;
Published online 13 February 2019.

1. Howe, L. & Wain, A. *Predicting the Future* Vol. V, 1–195 (Cambridge Univ. Press, 1993).
2. Bauer, P., Thorpe, A. & Brunet, G. The quiet revolution of numerical weather prediction. *Nature* **525**, 47–55 (2015).
3. Hantson, S. et al. The status and challenge of global fire modelling. *Biogeosciences* **13**, 3359–3375 (2016).
4. Agapiou, A. Remote sensing heritage in a petabyte-scale: satellite data and heritage Earth Engine© applications. *Int. J. Digit. Earth* **10**, 85–102 (2017).
5. Stockhause, M. & Lautenschlager, M. CMIP6 data citation of evolving data. *Data Sci. J.* **16**, 30 (2017).
6. Lee, J., Weger, R. C., Sengupta, S. K. & Welch, R. M. A neural network approach to cloud classification. *IEEE Trans. Geosci. Remote Sens*. **28**, 846–855 (1990).
7. Benediktsson, J. A., Swain, P. H. & Ersoy, O. K. Neural network approaches versus statistical methods in classification of multisource remote sensing data. *IEEE Trans. Geosci. Remote Sens*. **28**, 540–552 (1990).
8. Camps-Valls, G. & Bruzzone, L. *Kernel Methods for Remote Sensing Data Analysis* 434 (John Wiley & Sons, Chichester, 2009).
9. 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).
10. Camps-Valls, G., Tuia, D., Bruzzone, L. & Benediktsson, J. A. Advances in hyperspectral image classification: Earth monitoring with statistical learning methods. *IEEE Signal Process. Mag*. **31**, 45–54 (2014).
    This paper provides a comprehensive overview of machine learning for classification.
11. Gislason, P. O., Benediktsson, J. A. & Sveinsson, J. R. Random forests for land cover classification. *Pattern Recogn. Lett*. **27**, 294–300 (2006).
    This paper is one of the first machine learning papers for land-cover classification, a method now operationally used.
12. Muhlbauer, A., McCoy, I. L. & Wood, R. Climatology of stratocumulus cloud morphologies: microphysical properties and radiative effects. *Atmos. Chem. Phys*. **14**, 6695–6716 (2014).
13. Grimm, R., Behrens, T., Märker, M. & Elsenbeer, H. Soil organic carbon concentrations and stocks on Barro Colorado Island—digital soil mapping using Random Forests analysis. *Geoderma* **146**, 102–113 (2008).
14. Hengl, T. et al. SoilGrids250m: global gridded soil information based on machine learning. *PLoS ONE* **12**, e0169748 (2017).
    This paper describes machine learning used for operational global soil mapping.
15. Townsend, P. A., Foster, J. R., Chastain, R. A. & Currie, W. S. Application of imaging spectroscopy to mapping canopy nitrogen in the forests of the central Appalachian Mountains using Hyperion and AVIRIS. *IEEE Trans. Geosci. Remote Sens*. **41**, 1347–1354 (2003).
16. Coops, N. C., Smith, M.-L., Martin, M. E. & Ollinger, S. V. Prediction of eucalypt foliage nitrogen content from satellite-derived hyperspectral data. *IEEE Trans. Geosci. Remote Sens*. **41**, 1338–1346 (2003).
17. Verrelst, J., Alonso, L., Camps-Valls, G., Delegido, J. & Moreno, J. Retrieval of vegetation biophysical parameters using Gaussian process techniques. *IEEE Trans. Geosci. Remote Sens*. **50**, 1832–1843 (2012).
18. Papale, D. & Valentini, R. A new assessment of European forests carbon exchanges by eddy fluxes and artificial neural network spatialization. *Glob. Change Biol*. **9**, 525–535 (2003).
19. Jung, M. et al. Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite and meteorological observations. *J. Geophys. Res. Biogeo*. **116**, G00j07 (2011).
20. Tramontana, G. et al. Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. *Biogeosciences* **13**, 4291–4313 (2016).
21. Jung, M. et al. Recent decline in the global land evapotranspiration trend due to limited moisture supply. *Nature* **467**, 951–954 (2010).
    This paper describes the first data-driven machine-learning-based spatio-temporal estimation of global water fluxes on land.
22. Jung, M. et al. Compensatory water effects link yearly global land CO_2 sink changes to temperature. *Nature* **541**, 516–520 (2017).
23. Bonan, G. B. et al. Improving canopy processes in the Community Land Model version 4 (CLM4) using global flux fields empirically inferred from FLUXNET data. *J. Geophys. Res. Biogeosci*. **116**, G02014 (2011).
24. Anav, A. et al. Spatiotemporal patterns of terrestrial gross primary production: a review. *Rev. Geophys*. **53**, 785–818 (2015).
25. Landschützer, P. et al. A neural network-based estimate of the seasonal to inter-annual variability of the Atlantic Ocean carbon sink. *Biogeosciences* **10**, 7793–7815 (2013).
    This paper describes the first data-driven machine-learning-based carbon fluxes in the ocean.
26. Kühnlein, M., Appelhans, T., Thies, B. & Nauss, T. Improving the accuracy of rainfall rates from optical satellite sensors with machine learning—a random forests-based approach applied to MSG SEVIRI. *Remote Sens. Environ*. **141**, 129–143 (2014).
27. Caldwell, P. M. et al. Statistical significance of climate sensitivity predictors obtained by data mining. *Geophys. Res. Lett*. **41**, 1803–1808 (2014).
28. Reichstein, M. & Beer, C. Soil respiration across scales: the importance of a model-data integration framework for data interpretation. *J. Plant Nutr. Soil Sci*. **171**, 344–354 (2008).
29. Wright, S. Correlation and causation. *J. Agric. Res*. **20**, 557–585 (1921).
30. Guttman, N. B. Accepting the standardized precipitation index: a calculation algorithm. *J. Am. Water Resour. Assoc*. **35**, 311–322 (1999).
31. Vicente-Serrano, S. M., Beguería, S. & López-Moreno, J. I. A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. *J. Clim*. **23**, 1696–1718 (2010).
32. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
33. Lore, K. G., Stoecklein, D., Davies, M., Ganapathysubramanian, B. & Sarkar, S. Hierarchical feature extraction for efficient design of microfluidic flow patterns. *Proc. Machine Learning Res*. **44**, 213–225 (2015).
34. Baldi, P., Sadowski, P. & Whiteson, D. Searching for exotic particles in high-energy physics with deep learning. *Nat. Commun*. **5**, 4308 (2014).
35. Bhimji, W., Farrell, S. A., Kurth, T., Paganini, M. & Racah, E. Deep neural networks for physics analysis on low-level whole-detector data at the LHC. Preprint at https://arxiv.org/abs/1711.03573 (2017).
36. Schütt, K. T., Arbabzadah, F., Chmiela, S., Muller, K. R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun*. **8**, 13890 (2017).
37. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol*. **33**, 831–838 (2015).
38. Prabhat. A look at deep learning for science. *O'Reilly Blog* https://www.oreilly.com/ideas/a-look-at-deep-learning-for-science (2017).
39. Zhang, L. P., Zhang, L. F. & Du, B. Deep learning for remote sensing data: a technical tutorial on the state of the art. *IEEE Geosci. Remote Sens. Mag*. **4**, 22–40 (2016).
40. Ball, J. E., Anderson, D. T. & Chan, C. S. Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. *J. Appl. Remote Sens*. **11**, 042609 (2017).
41. Racah, E. et al. ExtremeWeather: a large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. *Adv. Neural Inform. Process. Syst*. **30**, 3405–3416 (2017).
42. Liu, Y. et al. Application of deep convolutional neural networks for detecting extreme weather in climate datasets. In *ABDA'16-International Conference on Advances in Big Data Analytics* 81–88 https://arxiv.org/abs/1605.01156 (2016).
    This paper is the first approach to detecting extreme weather automatically without any prescribed thresholds, using deep learning.
43. Zhao, W. Z. & Du, S. H. Learning multiscale and deep representations for classifying remotely sensed imagery. *ISPRS J. Photogramm. Remote Sens*. **113**, 155–165 (2016).
44. Mathieu, M., Couprie, C. & LeCun, Y. Deep multi-scale video prediction beyond mean square error. Preprint at https://arxiv.org/abs/1511.05440 (2015).
45. Oh, J., Guo, X., Lee, H., Lewis, R. L. & Singh, S. Action-conditional video prediction using deep networks in Atari games. *Adv. Neural Inf. Process. Syst*. **28**, 2863–2871 (2015).
46. Shi, X. et al. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst*. **28**, 802–810 (2015).
47. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
48. Friedlingstein, P. et al. Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks. *J. Clim*. **27**, 511–526 (2014).

49. Montavon, G., Samek, W. & Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process*. **73**, 1–15 (2017).
50. Runge, J. et al. Identifying causal gateways and mediators in complex spatio-temporal systems. *Nat. Commun*. **6**, 8502 (2015).
51. Chalupka, K., Bischoff, T., Perona, P. & Eberhardt, F. in *UAI'16 Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence* 72–81 (AUAI Press, 2016).
52. Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature* **521**, 452–459 (2015).
53. Goodfellow, I. J. et al. Generative Adversarial Nets. *Adv. Neural. Inf. Process. Syst*. **27**, 2672–2680 (2014).
This is a fundamental paper on a deep generative modelling approach, allowing possible futures to be modelled from data.
54. Hansen, M. C. et al. High-resolution global maps of 21st-century forest cover change. *Science* **342**, 850–853 (2013).
55. Pekel, J.-F., Cottam, A., Gorelick, N. & Belward, A. S. High-resolution mapping of global surface water and its long-term changes. *Nature* **540**, 418–422 (2016).
56. 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).
57. Camps-Valls, G. et al. Physics-aware Gaussian processes in remote sensing. *Appl. Soft Comput*. **68**, 69–82 (2018).
58. Karpatne, A., Watkins, W., Read, J. & Kumar, V. Physics-guided Neural Networks (PGNN): an application in lake temperature modeling. Preprint at https://arxiv.org/abs/1710.11431 (2017).
59. Luo, Y. Q. et al. A framework for benchmarking land models. *Biogeosciences* **9**, 3857–3874 (2012).
60. Eyring, V. et al. Towards improved and more routine Earth system model evaluation in CMIP. *Earth Syst. Dyn*. **7**, 813–830 (2016).
61. Klocke, D., Pincus, R. & Quaas, J. On constraining estimates of climate sensitivity with present-day observations through model weighting. *J. Clim*. **24**, 6092–6099 (2011).
62. Cox, P. M. et al. Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability. *Nature* **494**, 341–344 (2013).
63. Beck, H. E. et al. Global-scale regionalization of hydrologic model parameters. *Wat. Resour. Res*. **52**, 3599–3622 (2016).
64. Schirber, S., Klocke, D., Pincus, R., Quaas, J. & Anderson, J. L. Parameter estimation using data assimilation in an atmospheric general circulation model: from a perfect toward the real world. *J. Adv. Model. Earth Syst*. **5**, 58–70 (2013).
65. 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).
66. Becker, T., Stevens, B. & Hohenegger, C. Imprint of the convective parameterization and sea-surface temperature on large-scale convective self-aggregation. *J. Adv. Model. Earth Syst*. **9**, 1488–1505 (2017).
67. Siongco, A. C., Hohenegger, C. & Stevens, B. Sensitivity of the summertime tropical Atlantic precipitation distribution to convective parameterization and model resolution in ECHAM6. *J. Geophys. Res. Atmos*. **122**, 2579–2594 (2017).
68. de Bezenac, E., Pajot, A. & Gallinari, P. Deep learning for physical processes: incorporating prior scientific knowledge. Preprint at https://arxiv.org/abs/1711.07970 (2017).
69. Brenowitz, N. D. & Bretherton, C. S. Prognostic validation of a neural network unified physics parameterization. *Geophys. Res. Lett*. **45**, 6289–6298 (2018).
70. Willis, M. J. & von Stosch, M. Simultaneous parameter identification and discrimination of the nonparametric structure of hybrid semi-parametric models. *Comput. Chem. Eng*. **104**, 366–376 (2017).
71. McGovern, A. et al. Using artificial intelligence to improve real-time decision making for high-impact weather. *Bull. Am. Meteorol. Soc*. **98**, 2073–2090 (2017).
72. Vandal, T. et al. Generating high resolution climate change projections through single image super-resolution: an abridged version. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)* https://www.ijcai.org/proceedings/2018/0759.pdf (2018).
73. Verrelst, J. et al. Emulation of leaf, canopy and atmosphere radiative transfer models for fast global sensitivity analysis. *Remote Sens*. **8**, 673 (2016).
74. Chevallier, F., Chéruy, F., Scott, N. & Chédin, A. A neural network approach for a fast and accurate computation of a longwave radiative budget. *J. Appl. Meteorol*. **37**, 1385–1397 (1998).
75. Castruccio, S. et al. Statistical emulation of climate model projections based on precomputed GCM runs. *J. Clim*. **27**, 1829–1844 (2014).
76. Fer, I. et al. Linking big models to big data: efficient ecosystem model calibration through Bayesian model emulation. *Biogeosci. Disc*. **2018**, 1–30 (2018).
77. Glahn, H. R. & Lowry, D. A. The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteorol*. **11**, 1203–1211 (1972).
78. Wilks, D. S. Multivariate ensemble model output statistics using empirical copulas. *Q. J. R. Meteorol. Soc*. **141**, 945–952 (2015).
79. Tewari, A. et al. in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* 2549–2559 (IEEE, 2018).
80. Xie, Y., Franz, E., Chu, M. & Thuerey, N. tempoGAN: a temporally coherent, volumetric GAN for super-resolution fluid flow. Preprint at https://arxiv.org/abs/1801.09710 (2018).
81. Stewart, R. & Ermon, S. in *Proc. Thirty-First AAAI Conf. on Artificial Intelligence (AAAI-17)* 2576–2582 (2017).
82. Gunning, D. *Explainable Artificial Intelligence (XAI)* https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf (2017).
83. Hu, Z., Ma, X., Liu, Z., Hovy, E. & Xing, E. in *Proc. 54th Annual Meeting of the Association for Computational Linguistics* Vol. 1, 2410–2420 (Association for Computational Linguistics, 2016).
84. Pearlmutter, B. A. & Siskind, J. M. Reverse-mode AD in a functional framework: lambda the ultimate backpropagator. *ACM Trans. Progr. Lang. Syst*. **30**, 7 (2008).
85. Wang, F. & Rompf, T. in *ICLR 2018 Workshop* https://openreview.net/pdf?id=SJxJtYkPG (2018).
86. Beer, C. et al. Terrestrial gross carbon dioxide uptake: global distribution and covariation with climate. *Science* **329**, 834–838 (2010).
87. Forkel, M. et al. Enhanced seasonal CO_2 exchange caused by amplified plant productivity in northern ecosystems. *Science* **351**, 696–699 (2016).
88. Bellprat, O., Kotlarski, S., Lüthi, D. & Schär, C. Objective calibration of regional climate models. *J. Geophys. Res. Atmos*. **117**, D23115 (2012).
89. Reichstein, M. et al. in *AGU Fall Meeting Abstracts* 2016AGUFM.B2044A.2007R (AGU, 2016).
90. Rußwurm, M. & Körner, M. Multi-temporal land cover classification with long short-term memory neural networks. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci* **42**, 551–558 (2017).
This paper describes the first use of the LSTM deep learning model for multi-temporal land-cover classification.
91. Nash, J. E. & Sutcliffe, J. V. River flow forecasting through conceptual models. Part I—a discussion of principles. *J. Hydrol*. **10**, 282–290 (1970).
92. Shi, X. et al. Deep learning for precipitation nowcasting: a benchmark and a new model. *Adv. Neural. Inf. Process. Syst*. **30**, 5617–5627 (2017).
This paper describes the first approach to data-driven modelling of near-term precipitation using a combination of deep-learning concepts, that is, LSTMs and convolutional neural networks.
93. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. Preprint at https://arxiv.org/abs/1611.07004 (2016).
This paper is a geoscience-related extension application of ref. 53, in which, for example, remote sensing images are transferred to thematic maps.
94. Tompson, J., Schlachter, K., Sprechmann, P. & Perlin, K. Accelerating Eulerian fluid simulation with convolutional networks. *Proc. Machine Learning Res*. **70**, 3424–3433 (2017).
95. University Corporation for Atmospheric Research (UCAR). *Short-Term Explicit Prediction (STEP) Program Research Applications Laboratory 2013 Annual Report* https://nar.ucar.edu/2013/ral/short-term-explicit-prediction-step-program (NCAR/UCAR, 2013).
96. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst*. **28**, 91–99 (2015).
97. Zaytar, M. A. & El Amrani, C. Sequence to sequence weather forecasting with long short term memory recurrent neural networks. *Int. J. Comput. Appl*. **143**, 7–11 (2016).
98. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* Vol. xxii, 1–775 (MIT Press, Cambridge, 2016).
99. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput* **9**, 1735–1780 (1997).
100. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw*. **61**, 85–117 (2015).
101. Requena-Mesa, C., Reichstein, M., Mahecha, M., Kraft, B. & Denzler, J. Predicting landscapes as seen from space from environmental conditions. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* 1768–1771 (IEEE, 2018).

Acknowledgements We thank D. Frank and L. Maack for proofreading, help with the literature and technical help, and F. Gans for programming help in Julia. This research was supported by a grant by the Alexander von Humboldt Foundation (Max Planck Research Prize) to M.R. G.C.-V. was supported by the European Research Council (ERC) under the ERC Consolidator Grant ERC-CoG-2014 SEDAL (grant agreement 647423).

Author contributions M.R. conceived the work, created a first outline with Prabhat and ran the numerical experiment. All authors wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-0912-1.

Correspondence and requests for materials should be addressed to M.R.