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

Postprint version

Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais & Prabhat

Published in: Nature

Reference: Reichstein, M., Camps-Valls, G., Stevens, B. et al. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204 (2019). https://doi.org/10.1038/s41586-019-0912-1

Web link: https://www.nature.com/articles/s41586-019-0912-1

This paper has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 647423

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)], Mr Prabhat[7)]

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, 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, Caparica, Portugal
7. National Energy Research Supercomputing Center, Lawrence Berkeley National Laboratory, Berkeley, California, USA

Summary paragraph

Machine learning approaches are increasingly used to extract patterns and insights from the exploding universe of geospatial data, but current approaches may not be an optimal approach when system behavior is dominated by spatial or temporal context. Rather than amending classical machine learning, however, we argue that these contextual cues should be at the core of a modified approach – termed deep learning – to extract novel understanding and predictive ability for topics such as seasonal forecasting and modeling of long-range spatial connections across multiple time-scales. A critical further step will be a hybrid modeling approach coupling physical processes with deep learning versatility.

1. Introduction

Humans have always been striving to predict and understand the world, and the ability to make better predictions has given competitive advantages in diverse contexts (e.g., weather, diseases, or more recently financial markets). Yet the tools for prediction have substantially changed over time, from ancient Greek philosophical reasoning to non-scientific medieval methods like soothsaying, toward modern scientific discourse, which has come to include hypothesis testing, theory development and computer modelling underpinned by statistical and/or physical relationships, i.e., laws[1]. A success story in the geosciences is weather prediction, which has greatly improved through integration of better theory, increased computational power, and established observational systems which allow for the assimilation of large amounts of data into the modeling system[2]. Nevertheless, we can only accurately predict the evolution of the weather on a time-scale 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 strongly controlled by the seemingly stochastic disturbances such as fires and landslides. Such problems have been rather resistant to progress in the past 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 with rapidly increasing transmission rates beyond hundreds of terabytes per day[4]. These data come from a plethora of sensors measuring states, fluxes, and intensive or time/space integrated variables, and representing fifteen or more orders of temporal and spatial magnitude. They include remote sensing from meters to hundreds kilometers above the 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 (Climate Model Intercomparison Project), used extensively by the scientific community for scientific groundwork towards periodic climate assessments, is over 3PB in size, and the next generation, CMIP-6, is estimated to reach up to 30PB[5]. While not observations, the model data 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 V's” of Big Data: volume, velocity, variety, and veracity (Figure 1). One key challenge is to extract interpretable information and knowledge from this Big Data, possibly in near-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 which learn maximally from data, beyond traditional data assimilation approaches, 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 modeling 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 (see Glossary).

[Place Glossary around here]

In the following sections we review the development of machine learning in the geoscientific context, and highlight how deep learning, i.e. 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.

2. State-of-the-art in geoscientific machine learning

Machine learning is now a successful part of several research-driven andoperational geoscientific processing schemes, addressing the atmosphere, the land surface and the ocean, but has co-evolved with data availability over the last 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 development (e.g. 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 geo-scientific classification, and change and anomaly detection problems[9,10-12]. In the last few years, the field has begun to use deep learning to better exploit spatial and temporal structure in the data, features that would normally be problematic for traditional machine learning (e.g. Table 1, and next section).

Another class of problems where machine learning has been successful is regression problems. 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 last decade, machine learning has attained outstanding results in the regression estimation of bio-geo-physical parameters from remotely sensed reflectances at local and global scales[15,16,17]. These approaches emphasize spatial prediction, i.e. prediction of properties which 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 utilized 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- 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, which relate climate and remote sensing co-variates to the target variables. In this context, machine learning methods have proven to be more powerful and flexible than previous mechanistic or semi-empirical modelling approaches.. For instance an ANN with one hidden layer was able to filter out noise, predict the diurnal and seasonal variation of 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 for the first time to quantify global terrestrial photosynthesis and evapotranspiration of water in a purely data-driven way[19,20]. Spatial, seasonal, interannual or decadal variation 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 random forests have 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 new patterns, advance understanding and evaluate comprehensive physical models.

Notwithstanding the success of machine learning in the geosciences, important caveats and limitations have hampered a wider adoption and impact of such methods. A few pitfalls such as the risk of naïve 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 practices and expert intervention. More fundamentally, there are inherent limitations of currently-applied machine learning approaches. It is in this realm that the techniques of deep learning promise breakthroughs, as we explain in the paragraphs below.

Classical machine learning approaches benefit from domain-specific, hand-crafted features to account for dependencies in time or space (e.g. 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 (e.g. radiation, temperature, 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 neighboring grid cells contain hidden information on the state of the system (e.g. a long period without rain-fall 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 (e.g. humans, lightning, or both) but also on state variables, such as the state and amount of available fuel[3]. Fire spread and thus the burnt 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 size and shape of a geometry 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 which cannot be determined alone by localized, single pixel values.

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 (SPI[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 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 colors. Such features were then fed into a standard machine learning for localization, classification or detection of objects in images. 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, likely non-optimal), but certainly the concern of a restricted, and subjective choice of features rather than an extensive and generic approach remains a valid and important one. New developments in deep learning, however, no longer limit us to such approaches.

3. 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 arising (Table 1, Supplementary Box 1)[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 a problematic question for traditional machine learning" They report success in applying deep-learning architectures to objectively extract spatial features define and classify extreme situations (e.g. storms, atmospheric rivers) in numerical weather prediction model output. Such approach enables rapid detection of such events and forecast simulations without using either subjective human annotation or methods that rely on predefined somewhat arbitrary thresholds for wind speed or other variables. In particular, such approach uses the information in the spatial shape of respective events such as the typical spiral for 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 (e.g. convolutional neural networks for object classification) and sequence learning (e.g. speech recognition), there is a growing interest in blending these two perspectives. A prototypical example is video and motion prediction[44,45], which is strikingly similar 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, vegetation states which 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 where spatio-temporal dynamics are important, but which 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 geosciences (Figure 2): Images are analogous to two-dimensional data fields containing particular variables in analogy to color-triplets (RGB values) in photographs, while videos can be likened to a sequence of images and hence of 2D fields that evolve in time. Similarly, natural language and speech signals share the same multiresolution characteristics of dynamic time-series of Earth system variables. Furthermore, classification, regression, anomaly detection, and dynamic modeling are typical problems in both computer vision and geosciences.

4. 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, independent 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 (due to the presence of clouds or snow, distortions in the acquisition, storage and transmission, etc.).

In addition, spectral, spatial, and temporal dimensionalities raise computational challenges. The data volume is increasing geometrically and soon it will be necessary to deal with Petabytes/day globally. Currently, the biggest meteorological agencies have to process Terabytes per day in near real time. often at very high precision (32-bit, 64-bit). Further, while typical computer vision applications have worked with image sizes of 512 x 512 pixels, a moderate resolution (ca. 1km) global field has sizes of approximately 40000 x 20000 pixels, i.e. three orders of magnitude more.

Last but not least, unlike the ImageNet benchmark (a data base of images with labels, e.g. “cat” or “dog”[47]) in the computer vision community, large, labeled geoscientific datasets do not always exist in geo-science, not only due to the sizes of the datasets involved, but also due to the conceptual difficulty in labeling data sets, e.g. determining “it’s a cat” vs “it’s a drought”, given that the second label is contingent on intensity and extent and can change according to methods, and there are not enough labeled cases for training. These aspects raise the challenge of working with a limited training set. More generally, geo-scientific problems are often underconstrained, leading to the possibility of models thought to be of high quality, which perform well in training and even test data sets, but deviate strongly for situations and data outside their valid domain (extrapolation problem), which is even true for complex physical Earth system models[48]. Overall, we identify at least five major challenges and avenues for the successful adoption of deep learning approaches in the geosciences:

1. <u>Interpretability:</u> Improving predictive accuracy is important but insufficient. Certainly, interpretability and understanding are crucial in this arena, 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 from causal discovery from observational data[50,51]. Yet, we should note that, given their complexity, also modern Earth system models are in practice often not easily traceable back to their assumptions, limiting their interpretability as well.
2. <u>Physical consistency:</u> Deep learning models can fit observations very well, but predictions may be physically inconsistent or implausible, e.g. owing to extrapolation or observational biases. 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: New 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 neighborhood (even at different scales), but also long-range relationships (e.g., for teleconnections) are urgently needed, but the exact cause-effect relations between variables are even not clear in advance and need to be discovered. Modelling uncertainties will be certainly an important aspect and will require to integrate concepts from Bayesian/probabilistic inference, which are directly addressing that (Glossary and [52]).
4. Limited labels: Methods need to be further developed which can learn from few labelled examples, by utilizing the information in related unlabeled observations, so-called unsupervised density modeling, feature extraction and semi-supervised learning[53] (cf. glossary).
5. Computational demand: There is a huge technical challenge regarding the high computational cost of current geoscience problems - good examples to address this includes Google Earth Engine, which allowed solving real problems from deforestation[54] to lake[55] monitoring, yet still without deep learning application.

By addressing these challenges, deep learning could make an even bigger difference in the geosciences in comparison to 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, etc.), the mapping from the world to images, and assumptions about the (visual) appearance of world points (surface points, the state in 3D) on 2D images. Assumptions for successful processing include the assumption of Lambertian surfaces (i.e. intensity does not depend on the angle between surface and light source) which results in the classical assumption of constant intensity of the observation of a 3D point over time. In addition, changes in the world (the motion of objects) are in most cases modeled as rigid transformations, or non-rigid transformations that arise from physical assumptions and that are only valid locally (like in registration of brain structures, before and after removal of a tumor). Even complex problems in computer vision have been solved by hand-crafted features that reflect the assumptions and expectations arising from common world knowledge. In geoscience and climate science, such global, general assumptions are still partly missing. In fact, these assumptions and expectations are exactly the models we are looking for! 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 processes are understood, the fewer high-quality hand-crafted features for modeling are expected to exist. Thus, deep learning methods, particularly since they find a good representation from data, represent an opportunity to tackle geoscience and climate research problems.

The most promising near-future applications include nowcasting, (i.e. 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 time-scales, modelling spatial dynamics where spatial context plays an important role (e.g. fires), and detecting teleconnections and connections between variables that a human may not have thought about.

Overall, 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 respecting and taking advantage of the 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 explore in the following section.

5 Integration with physical modelling

Historically, physical modelling and machine learning have been often treated as “two different worlds” 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, while data-driven approaches are highly flexible in adapting to the 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 more intensively integrate simulation and data science approaches in multiple fashions. From a systems modelling point of view there are five points of potential synergy (Figure 3) [the numbers in the following list correspond to the circles in the figure]:

1) Improving parameterizations *(Fig. 3, linkage 1).* Physical models require parameters but many of those cannot be easily derived from first principles. Here, machine learning can learn parameterizations to optimally describe the ground truth which 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 (e.g. precipitation, surface slope) to catchment parameters (e.g. mean, minimum, maximum streamflow) has been learned from a few thousands 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 (e.g. 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 *(Fig. 3, linkage 2).* If formulations of a submodel are of semi-empirical nature where the functional form has little theoretical basis (e.g. 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 modeling (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 the regulation (biological) is flexible and learned from data. Such principle has 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 reactors modeling[70], which, albeit less complex, bears many similarities to hydrological and biogeochemical modelling.

3) Analysis of model-observation mismatch (Fig. 3, linkage 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 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 improving the physical model and theory. In practice, it can also serve to correct 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 sub-models (Fig. 3, linkage 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 in disentangling 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 modeling or emulation: 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 orders of magnitude faster simulations than the original physical model without sacrificing significant 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 (RTMs) 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, e.g. in climate modeling[75] and more recently explored in vegetation dynamic models[76]. Further, given the complexity of physical models, emulation challenges are very good test beds to explore the potential of machine learning and deep learning approaches to extrapolate outside the ranges of training conditions.

Some of the concepts in Figure 3 have already been adopted in a broad sense. For instance, point 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 (Fig. 3, link 2), appear to be much less explored. Conceptually the hybrid approaches discussed before can be interpreted as deepening and “physicizing” a neural network (Figure 4), where the physical model comes on top of a neural network layers (see examples Fig. 4b-c). It contrasts 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 started from a system-modelling view and seeks to integrate machine learning. As an alternative perspective system knowledge can be integrated into a machine learning framework. This may include respective design of the network architecture[36,79], physical constraints in the cost function for optimization[58], or expansion of the training data set for under-sampled domains (i.e. 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 consistency and credibility of the machine learning models. As a by-product, the hybridization has an interesting regularization effect as physics discards implausible models. Therefore, physics-aware machine learning models should better combat overfitting, 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 advancements 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 crucial for scientific purposes, and in particular for management, or policy decisions. Differentiable programming allows for efficient optimization due to automated differentiation[84,85]. This greatly helps in making the large, non-linear and complex inversion problem computationally more tractable, and in addition allows for explicit sensitivity assessments, thus aiding in interpretability.

6. Advancing science

There is no doubt and there are numerous examples as discussed in this manuscript, that modern machine learning methods significantly improve classification and prediction skills. This alone has great value. Yet, how do they improve fundamental scientific understanding, given that in particular the outcome of complex statistical models remains hard to grasp? The answer can be found in the observations which have virtually always been the basis for scientific progress. The Copernican revolution was possible by precisely observing planetary trajectories to infer and test the laws governing them. While 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 machine learning based estimates of photosynthesis, has 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 also in other regions, and better consistency with leaf level observations.. Related data-driven carbon cycle estimates have helped calibrating vegetation models and explain the conundrum of the increasing seasonal amplitude of the CO_2 concentration in high latitudes[87], which according to these results is caused by more vigorous vegetation in the high latitudes. In addition to data-driven theory and model building, extracted patterns are increasingly being used as a way to explore improved parameterizations in Earth system models[65,69], and emulators are increasingly being used as a basis for model calibration[88]. In other words, the scientific interplay between theory and observation, of hypothesis generation and theory-driven hypothesis testing will prevail, but the complexity of hypotheses and tests inferred from data and the pace of this generation are changing by orders of magnitude, implying unprecedented, qualitative and quantitative progress of the science of the complex Earth system.

7. Conclusion

Earth sciences face the 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, modeling 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 for understanding of the Earth. The Earth system specific challenges shall further stimulate the development of methodologies, where we have four major recommendations.

Recognition of the particularities of the data: multi-source, multi-scale, high dimensional, complex spatial-temporal relations, including non-trivial, and lagged long-distance relationships (teleconnections) between variables need to be adequately modelled. While the deep learning approach is well-positioned to address these data challenges, this may stimulate development of new network architectures, algorithms and approaches, in particular deep-learning approaches which address both spatial and temporal context at different scales (cf. Figure 4).

Plausibility and interpretability of inferences: models should not only be accurate but also credible and aware of 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/explainable in a set of rules, descriptors, and relations.

Uncertainty estimation: Models should speak about their confidence and credibility. A strong integration of Bayesian/probabilistic inference will be an avenue to follow here, because they allow for explicit representation and propagation of uncertainties. In addition, identifying and treating extrapolation is a priority.

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. Thus 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 and 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 geo-scientific 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: they obey physical laws, feature a conceptualized and thus interpretable structure, and at the same time are fully data-adaptive where theory is weak. Importantly, the other way around also holds: machine learning research will benefit from plausible physically based relationships derived from the natural sciences. Among others, two major Earth system challenges resistant to past progress, the parameterization of atmospheric convection and the description of spatio-temporal dependency of ecosystems on climate and interacting geo-factors, are open to be addressed with the approaches discussed here.

Author information

The authors declare no competing interests.

References

1 Howe, L. & Wain, A. *Predicting the future*. V, 195 p. (Cambridge University Press, 1993).

2 Bauer, P., Thorpe, A. & Brunet, G. The quiet revolution of numerical weather prediction. *Nature* **525**, 47, doi:10.1038/nature14956 (2015).

3 Hantson, S. *et al.* The status and challenge of global fire modelling. *Biogeosciences* **13**, 3359-3375, doi:10.5194/bg-13-3359-2016 (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**, doi:10.5334/dsj-2017-030 (2017).

6 Lee, J., Weger, R. C., Sengupta, S. K. & Welch, R. M. A neural network approach to cloud classification. *IEEE T Geosci. Remote.* **28**, 846-855, doi:10.1109/36.58972 (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 T Geosci. Remote.* **28**, 540-552, doi:10.1109/Tgrs.1990.572944 (1990).

8 Camps-Valls, G. & Bruzzone, L. *Kernel methods for remote sensing data analysis*. 434 p. (John Wiley & Sons, 2009).

9 Gómez-Chova, L., Tuia, D., Moser, G. & Camps-Valls, G. Multimodal classification of remote sensing images: A review and future directions. *P IEEE* **103**, 1560-1584, doi:10.1109/Jproc.2015.2449668 (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 Proc. Mag.* **31**, 45-54 (2014). **[Comprehensive overview of machine learning for classification]**

11 Gislason, P. O., Benediktsson, J. A. & Sveinsson, J. R. Random Forests for land cover classification. *Pattern Recog. Lett.* **27**, 294-300, doi:10.1016/j.patrec.2005.08.011 (2006). **[One of the first machine learning papers for land cover classification, 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, doi:10.5194/acp-14-6695-2014 (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, doi:10.1371/journal.pone.0169748 (2017). **[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 T Geosci. Remote.* **41**, 1347-1354, doi:10.1109/Tgrs.2003.813205 (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 T Geosci. Remote.* **41**, 1338-1346, doi:10.1109/Tgrs.2003.813135 (2003).

17 Verrelst, J., Alonso, L., Camps-Valls, G., Delegido, J. & Moreno, J. Retrieval of vegetation biophysical parameters using Gaussian process techniques. *IEEE T Geosci. Remote.* **50**, 1832-1843, doi:10.1109/Tgrs.2011.2168962 (2012).

18 Papale, D. & Valentini, R. A new assessment of European forests carbon exchanges by eddy fluxes and artificial neural network spatialization. *Global Change Biol.* **9**, 525-535, doi:10.1046/j.1365-2486.2003.00609.x (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, doi:10.1029/2010jg001566 (2011).

20 Tramontana, G. *et al.* Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. *Biogeosciences* **13**, 4291 - 4313, doi:10.5194/bg-13-4291-2016 (2016).

21 Jung, M. *et al.* Recent decline in the global land evapotranspiration trend due to limited moisture supply. *Nature* **467**, 951 - 954, doi:10.1038/nature09396 (2010). **[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, doi:10.1038/nature20780 (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. - Biogeo.* **116**, G02014, doi:10.1029/2010jg001593 (2011).

24 Anav, A. *et al.* Spatiotemporal patterns of terrestrial gross primary production: A review. *Rev. Geophys.* **53**, 785-818, doi:10.1002/2015rg000483 (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, doi:10.5194/bg-10-7793-2013 (2013). **[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, doi:10.1002/2014gl059205 (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, doi:10.1002/jpln.200700075 (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. As.* **35**, 311-322, doi:10.1111/j.1752-1688.1999.tb03592.x (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, doi:10.1175/2009jcli2909.1 (2010).

32 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444, doi:10.1038/nature14539 (2015).

33 Lore, K. G., Stoecklein, D., Davies, M., Ganapathysubramanian, B. & Sarkar, S. in *JMLR: Workshop and Conference Proceeding / The 1st International Workshop "Feature Extraction: Modern Questions and Challenge"* Vol. 44 213-225 (Montreal, Canada, 2015).

34 Baldi, P., Sadowski, P. & Whiteson, D. Searching for exotic particles in high-energy physics with deep learning. *Nat. Commun.* **5**, 4308, doi:10.1038/ncomms5308 (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. *arXiv.org e-Print archive*, arXiv:1711.03573 (2017).

36 Schutt, K. T., Arbabzadah, F., Chmiela, S., Muller, K. R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **8**, 13890, doi:10.1038/ncomms13890 (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, doi:10.1038/nbt.3300 (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 Geosc. Rem. Sen. M* **4**, 22-40, doi:10.1109/Mgrs.2016.2540798 (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, doi:10.1117/1.Jrs.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. Inf. Process. Syst.*, 3405-3416 (2017).

42 Liu, Y. *et al.* in *ABDA'16 - Internationall Conference on Advances in Big Data Analytics* 81-88 (2016). **[First approach to automatically detect extreme weather 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, doi:10.1016/j.isprsjprs.2016.01.004 (2016).

44 Mathieu, M., Couprie, C. & LeCun, Y. Deep multi-scale video prediction beyond mean square error. *arXiv.org e-Print archive*, arXiv: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.*, 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.* in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248-255 (IEEE, Miami, FL, 2009).

48 Friedlingstein, P. *et al.* Uncertainties in CMIP5 Climate Projections due to Carbon Cycle Feedbacks. *J. Clim.* **27**, 511-526, doi:10.1175/jcli-d-12-00579.1 (2014).

49 Montavon, G., Samek, W. & Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* (2017).

50 Runge, J. *et al.* Identifying causal gateways and mediators in complex spatio-temporal systems. *Nat. Commun.* **6**, 8502, doi:10.1038/ncomms9502 (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 Arlington, Jersey City, New Jersey, USA, 2016).

52 Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature* **521**, 452, doi:10.1038/nature14541 (2015).

53 Goodfellow, I. J. *et al.* Generative Adversarial Nets. *Adv. Neural. Inf. Process. Syst.* **27**, 2672-2680 (2014). **[Fundamental paper on a deep generative modelling approach, allowing one to model e.g. possible futures from data]**

54 Hansen, M. C. *et al.* High-resolution global maps of 21st-century forest cover change. *Science* **342**, 850-853, doi:10.1126/science.1244693 (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, doi:10.1038/nature20584 (2016).

56 Karpatne, A. *et al.* Theory-guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE T Knowl. Data En.* **29**, 2318-2331, doi:10.1109/TKDE.2017.2720168 (2017).

57 Camps-Valls, G. *et al.* Physics-aware Gaussian processes in remote sensing. *Appl. Soft Comput.* **68**, 69-82, doi:10.1016/j.asoc.2018.03.021 (2018).

58 Karpatne, A., Watkins, W., Read, J. & Kumar, V. Physics-guided Neural Networks (PGNN): An Application in Lake Temperature Modeling. *arXiv.org e-Print archive*, arXiv:1710.11431 (2017).

59 Luo, Y. Q. *et al.* A framework for benchmarking land models. *Biogeosciences* **9**, 3857 - 3874, doi:10.5194/bg-9-3857-2012 (2012).

60 Eyring, V. *et al.* Towards improved and more routine Earth system model evaluation in CMIP. *Earth Syst. Dynam.* **7**, 813-830, doi:10.5194/esd-7-813-2016 (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, doi:10.1175/2011jcli4193.1 (2011).

62 Cox, P. M. *et al.* Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability. *Nature* **494**, 341-344, doi:10.1038/nature11882 (2013).

63 Beck, H. E. *et al.* Global-scale regionalization of hydrologic model parameters. *Water 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 Systems* **5**, 58-70, doi:10.1029/2012ms000167 (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, doi:10.1029/2018GL078202 (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 Systems* (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, doi:10.1002/2016jd026093 (2017).

68 de Bezenac, E., Pajot, A. & Gallinari, P. Deep Learning for Physical Processes: Incorporating Prior Scientific Knowledge. *arXiv.org e-Print archive*, arXiv: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, doi:10.1029/2018gl078510 (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, doi:10.1016/j.compchemeng.2017.05.005 (2017).

71 McGovern, A. *et al.* Using artificial intelligence to improve real-time decision making for high-impact weather. *B Am. Meteorol. Soc.* **98**, 2073-2090, doi:10.1175/Bams-D-16-0123.1 (2017).

72 Vandal, T. *et al.* in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)* 5389-5393 (Stockholm, Sweden, 2018).

73 Verrelst, J. *et al.* Emulation of Leaf, Canopy and Atmosphere Radiative Transfer Models for Fast Global Sensitivity Analysis. *Remote Sens.* **8**, 673, doi:10.3390/rs8080673 (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, doi:10.1175/1520-0450(1998)037<1385:Annafa>2.0.Co;2 (1998).

75 Castruccio, S. *et al.* Statistical emulation of climate model projections based on precomputed GCM runs. *J. Clim.* **27**, 1829-1844, doi:10.1175/Jcli-D-13-00099.1 (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, doi:10.5194/bg-2018-96 (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 Roy. Meteor. Soc.* **141**, 945-952, doi:10.1002/qj.2414 (2015).

79 Tewari, A. *et al.* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2549-2559 (2018).

80 Xie, Y., Franz, E., Chu, M. & Thuerey, N. tempoGAN: A Temporally Coherent, Volumetric GAN for Super-resolution Fluid Flow. *arXiv.org e-Print archive*, arXiv:1801.09710 (2018).

81 Stewart, R. & Ermon, S. in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)* 2576-2582 (San Francisco, California USA, 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 *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* Vol. 1: Long Papers 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 T Progr. Lang. Sys.* **30**, 7, doi:10.1145/1330017.1330018 (2008).

85 Wang, F. & Rompf, T. in *ICLR 2018 Workshop* (2018).

86 Beer, C. *et al.* Terrestrial Gross Carbon Dioxide Uptake: Global Distribution and Covariation with Climate. *Science* **329**, 834 - 838, doi:10.1126/science.1184984 (2010).

87 Forkel, M. *et al.* Enhanced seasonal CO_2 exchange caused by amplified plant productivity in northern ecosystems. *Science* **351**, 696-699, doi:10.1126/science.aac4971 (2016).

88 Bellprat, O., Kotlarski, S., Lüthi, D. & Schär, C. Objective calibration of regional climate models. *J Geophys. Res. - Atmos.* **117**, D23115, doi:10.1029/2012jd018262 (2012).

89 Reichstein, M. *et al.* in *AGU Fall Meeting Abstracts* 2016AGUFM.B2044A..2007R (2016).

90 Rußwurm, M. & Körner, M. Multi-temporal land cover classification with long short-term memory neural networks. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* **XLII-1/W1**, 551-558, doi:10.5194/isprs-archives-XLII-1-W1-551-2017 (2017). **[First Use of 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). **[First approach to data-driven modelling of near-term precipitation using a combination of deep-learning concepts, i.e. LSTMs and convolutional neural networks]**

93 Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. *arXiv.org e-Print archive*, arXiv:1611.07004v07001 (2016). **[A geo-science related extension application of Goodfellow et al., where e.g. remote sensing images are transferred to thematic maps]**

94 Tompson, J., Schlachter, K., Sprechmann, P. & Perlin, K. in *Proceedings of the 34th International Conference on Machine Learning* Vol. 70 (eds Doina Precup & Yee Whye Teh) 3424-3433 (PMLR, Proceedings of Machine Learning Research, 2017).

95 https://nar.ucar.edu/2013/ral/short-term-explicit-prediction-step-program

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.*, 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** (2016).

98 Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning*. xxii, 775 p. (MIT press, 2016).

99 May, R. M. Simple mathematical models with very complicated dynamics. *Nature* **261**, 459-467, doi:10.1038/261459a0 (1976).

100 Siegelmann, H. T. & Sontag, E. D. On the computational power of neural nets. *J Comput. Syst. Sci.* **50**, 132-150, doi:10.1006/jcss.1995.1013 (1995).

101 Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735-1780 (1997).

102 Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **61**, 85-117, doi:10.1016/j.neunet.2014.09.003 (2015).

Tables

Table 1: Geoscientific tasks, conventional approaches, their limitations and potential of deep learning approaches

Analytical Task	Scientific Task	Conventional approaches	Limitations	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	No or only shallow spatial context used	Convolutional Neural Networks[43]
Regression				
	Predict fluxes from atmospheric conditions	Random forests Kernel methods Feedforward NNs	Memory and lag effects not considered	Recurrent neural networks, LSTMs [89]
	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 only used 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], cGANs[53,93]
	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[94, 68],

Figures with captions

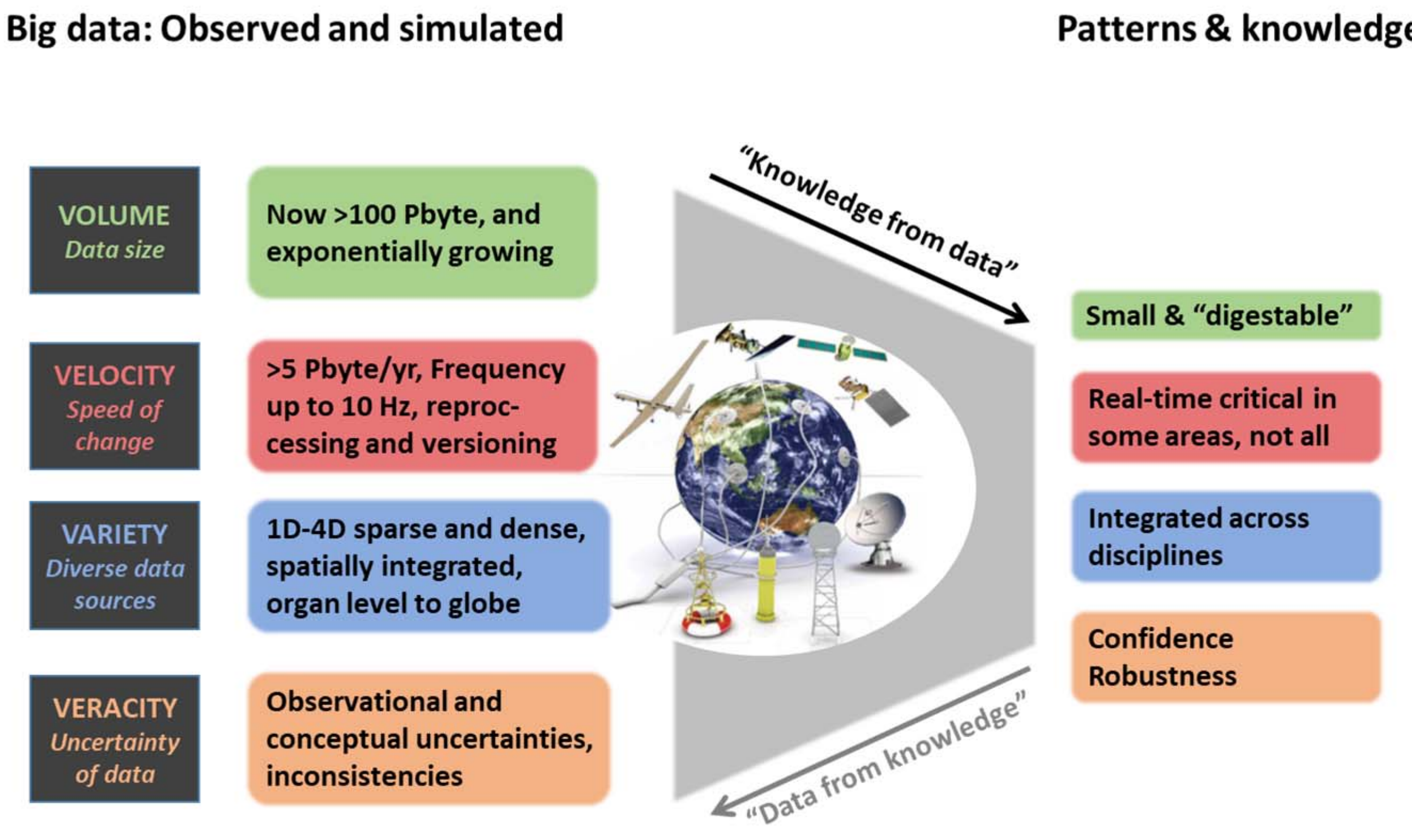


Figure 1: Big data challenges in the geoscientific context (Earth picture from https://nosc.noaa.gov/tpio/images/ObsSys.jpg)

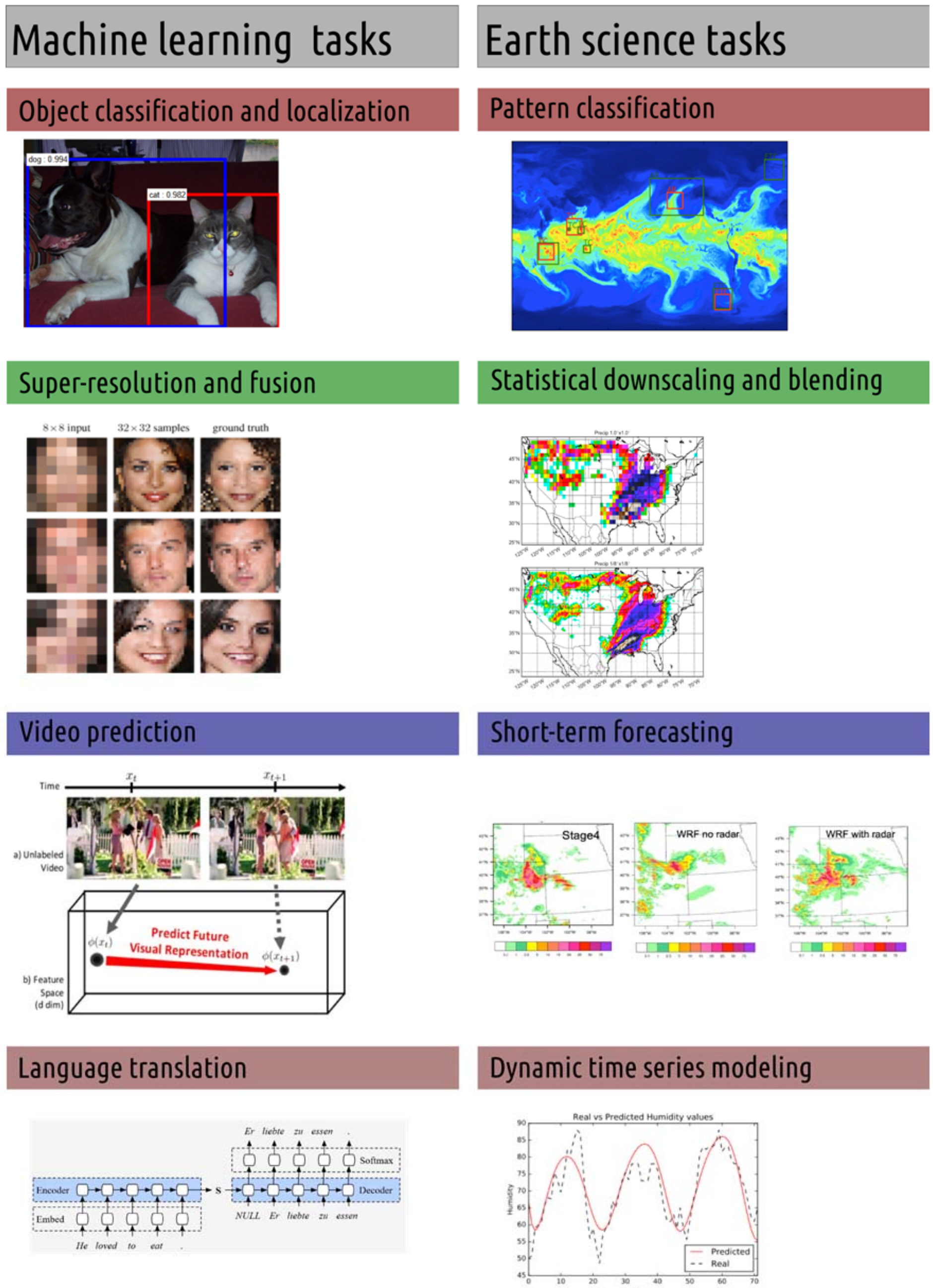


Figure 2: Typical deep learning applications (left) and the geo-scientific problems they apply to (right). From top to bottom: 1) classification of extreme weather patterns using a unified convolutional neural network on climate simulation data [42, 41], 2) statistical downscaling of climate model output [72], 3) short-term forecasting of climate variables[95], and 4) modelling of dynamic time-series.[96], [97] Image sources: https://smerity.com/articles/2016/google_nmt_arch.html; https://arxiv.org/abs/1612.02095;.

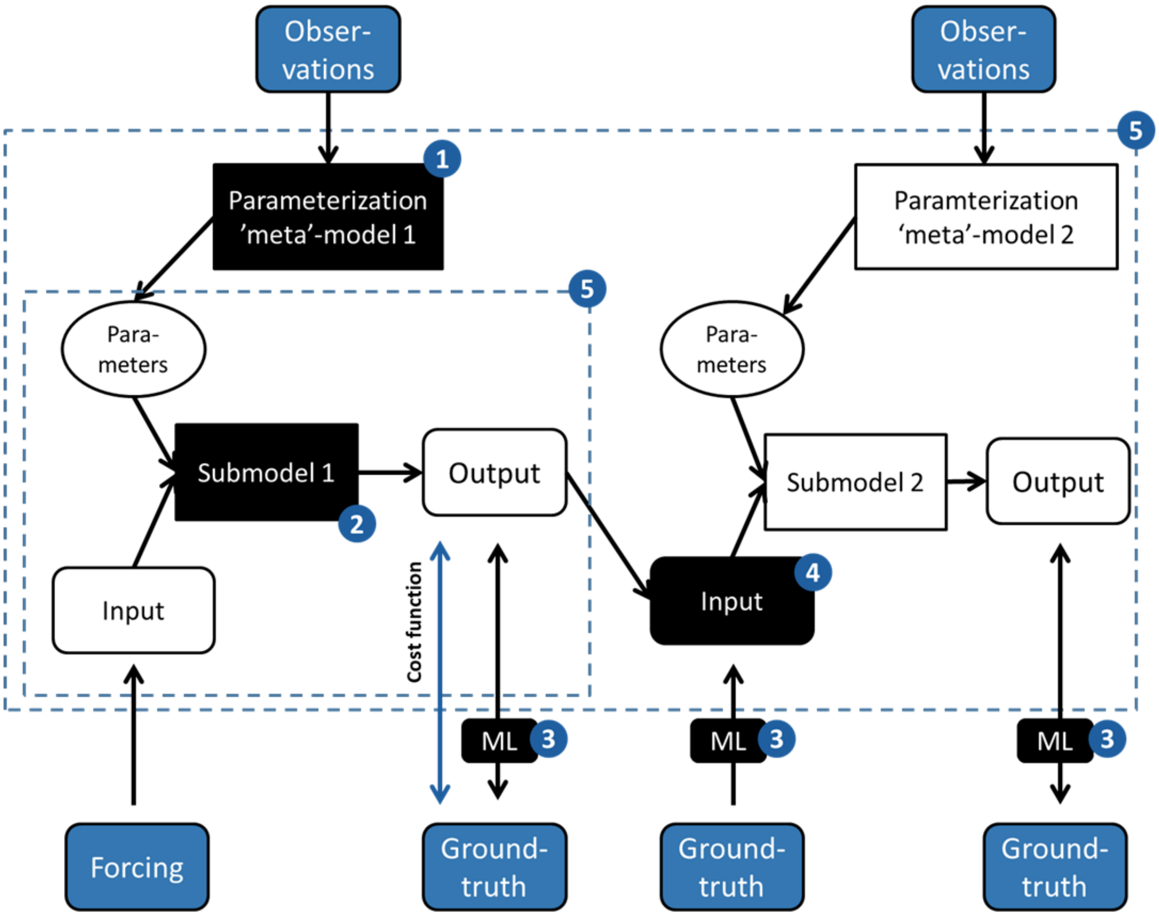


Figure 3: Linkages between physical models and machine learning: Depicted here is an abstraction of a part of a physical system, e.g. a climate model. The model consists of submodels which each have parameters, and forcing variables as inputs, and produce output, which can be input (forcing) to another sub-model. Data-driven learning approaches can be helpful in various instances, cf. the black-boxes and numbers. More detail in the text. ML = Machine Learning

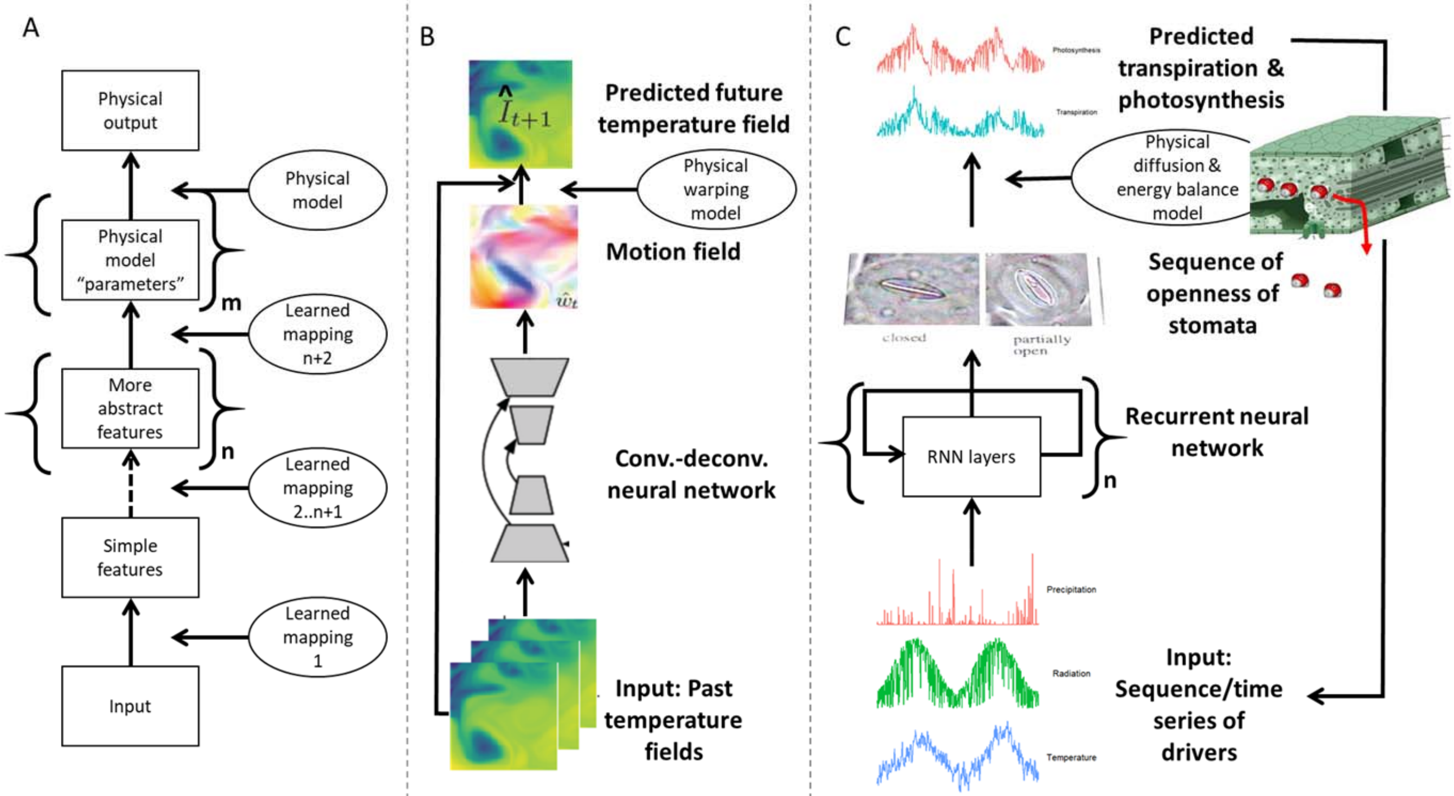


Figure 4: Interpretation of hybrid modelling (circle 2 in Figure 3) as deepening and "physicsizing" a deep learning architecture by adding one or several (m) physical layers after the multilayer neural network (A). (B) and (C) are concrete examples, where (B) is from de Bezenac et al.[68], where a motion field is learned with a convolutional-deconvolutional neural network, and the motion field further processed with a physical model. (C) models a biological regulation process (opening of the stomatal "valves" controlling water vapor flux from the leaves) with a recurrent neural network and processes this further with a physical diffusion model to estimate transpiration, which in turn influences some of the drivers, e.g. soil moisture. Basic scheme (A) modified after Goodfellow et al. [98].

Supplementary material and glossary

Efficient modelling a dynamic non-linear system with recurrent neural networks

Aforementioned state-of-the-art examples of mapping sequences of driving variables (e.g. meteorological conditions) onto target variables such as CO_2 fluxes from ocean or land have considered instantaneous mapping without representation of state dynamics. Dynamic effects have either been considered by directly using observed states as predictors (e.g. vegetation state represented by reflectance) or by introducing hand-designed features. The general problem is depicted in the figure below, where the input acts on an unknown, unobservable system state, while the observable is both influenced by the past state and the current input. It is not a problem of forecasting a time series a few steps ahead, because the whole output sequence has to be predicted by the model.

As an example, in the synthetic dynamic system below (one realization in Figure Box 1) we have three forcing variables x_1, x_2, x_3 where two of them influence one (unobserved) state *r* according to

$$r_{t+1} = f\left(x_{1,t}, x_{2,t}, r_t\right), \text{ with}$$

$$f\left(x_{1,t}, x_{2,t}, r_t\right) = \tau \cdot x_{1,t} \cdot x_{2,t} \cdot e^{x1,t} + (1-\tau) \cdot r_t,$$

τ being a parameter determining the inertia of the dynamics of *r*, here set to 0.05. A target state *y* to be predicted evolves as a logistic map well known from ecology and chaos theory[99]:

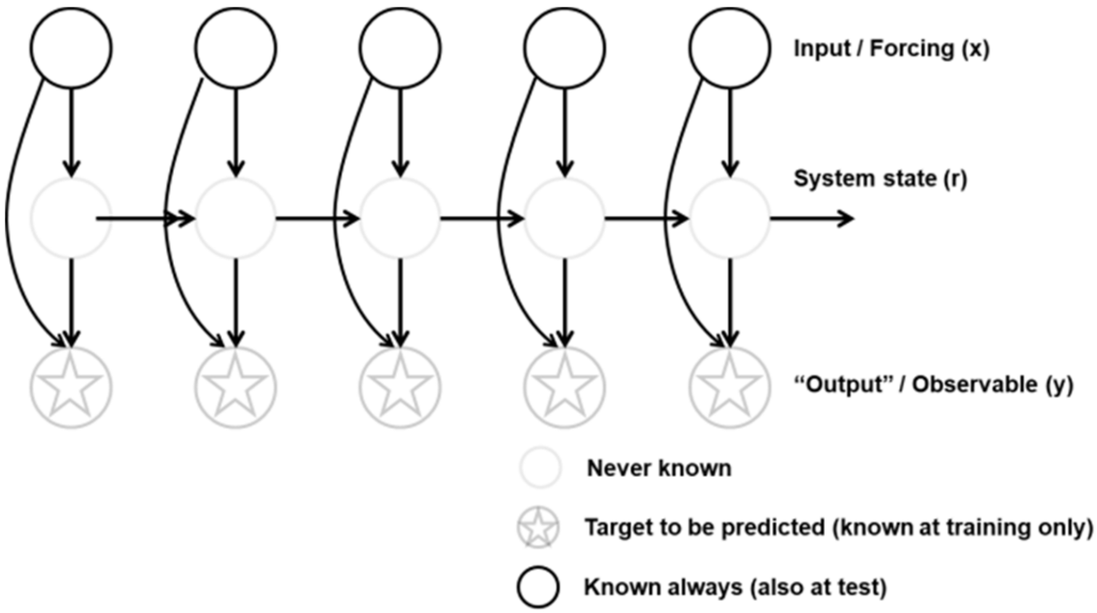
$$y_{t+1} = \widetilde{r_t} \cdot y_t \cdot (1 - y_t),$$

where (contrary to the standard logistic map) the parameter $\widetilde{r_t}$ is not fixed but dynamic and dependent on *r* as

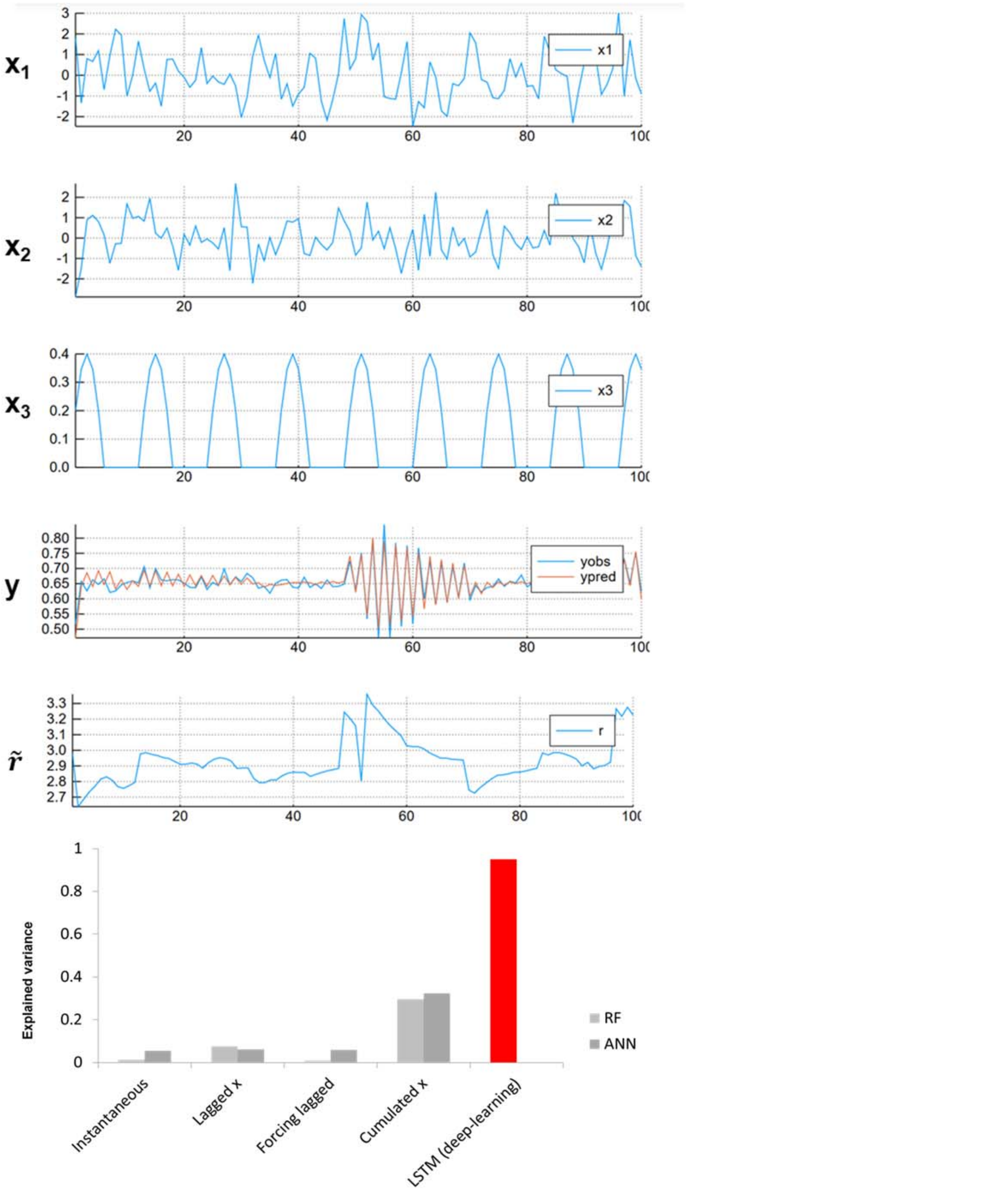
$$\widetilde{r_t} = g(\, r_t + x_{3,t}),$$

where *g* simply scales $\widetilde{r_t}$ onto the interval [2.5, 4] which implies dynamics varying with time between dampened oscillations, limit cycles and chaos. In the synthetic example 500 realizations of x_1 and x_2 as Gaussian i.i.d. variables are generated, while x_3 is always a seasonal variable as in the Figure below. Obviously x_1 and x_2 are mimicking a stochastic forcing, whereas x_3 represents a deterministic forcing (e.g. solar radiation varying diurnally and seasonally).

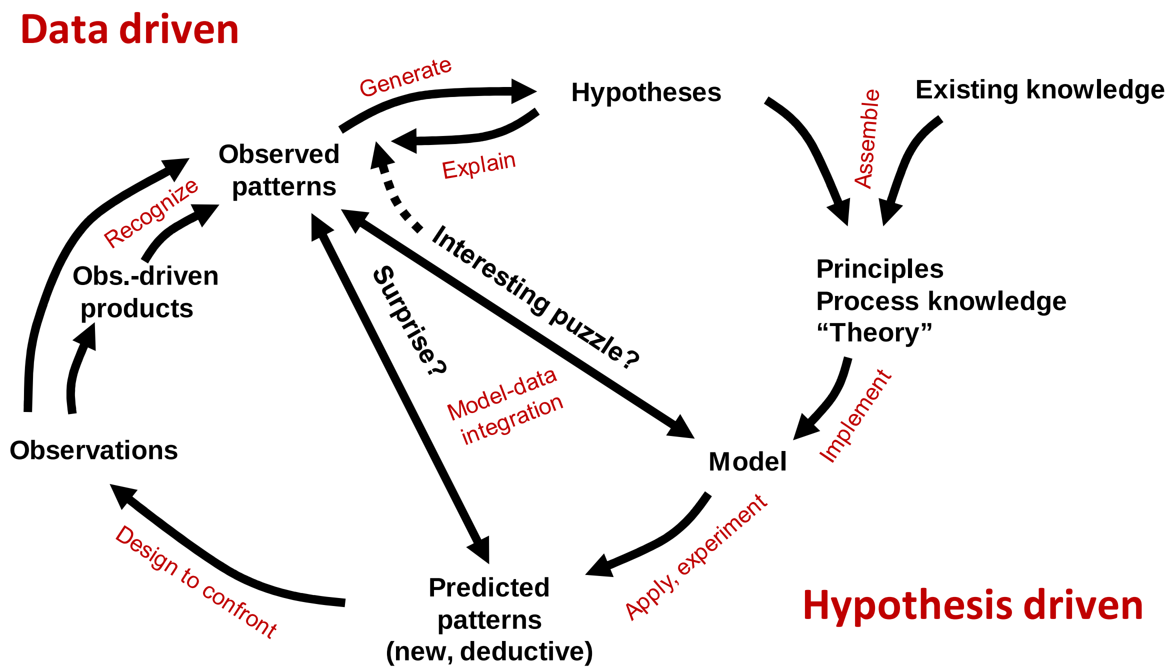
The lower panel shows the performance of different approaches to model the y_t sequence given the sequences of $x_1 \ldots x_3$. With a feed-forward ANN or random forests it is hard to model the sequence y_t, even with including intuitive features which represent lagged or memory effects, such as lagged or cumulated x variables over the last 25 time steps. On the contrary, being turing-complete[100] a recurrent NN has the potential to describe any dynamic system, and the challenge is the parameter estimation or training. In the specific case a simple LSTM[101] with 8 cells was trained on 80% of the realizations and the results are shown here for the test set. Certainly, other modelling approaches such as dynamic Bayesian approaches (e.g. hidden Markov models) exist as well for state estimation, and the relation to recurrent neural networks and deep learning is under research[102].



Supplementary Figure 1: Concept of modelling a dynamic system, i.e. mapping an input sequence to an output sequence, where a (hidden) dynamic state is involved.



Supplementary Figure 2: Data-driven modelling of a synthetic geoscientific time-series depicted in (a) with dynamic effects. Shown are predictor variables x_1, x_2, x_3, the resulting time-series of the system state ("observed" and modelled with an LSTM), and the parameter $\tilde{r}_t$ of the logistic map. (b) While classical approaches including typical feature design fail to explain the dynamics (grey bars, RF = Random Forest, ANN = feedforward ANN), a deep learning approach, long-short-term-memory neural network (LSTM) is able to explain almost all variance (red), without designing any features.



Supplementary Figure 3: Cycle of hypothesis generation and testing in complex systems involving process-based models and extraction of patterns from observations. Such patterns are only a surprise, and constitute an puzzle, if state-of-the-art theory and models do not predict them. Machine learning allows to extract hidden and complex patterns, which should be confronted with modelled patterns.

Glossary

Term	**Explanation**
Artificial Intelligence, Machine Learning & Deep Learning	Artificial intelligence (AI) is the capacity of an algorithm for assimilating 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 (ML) is a field of statistical research for training computational algorithms that split, sort, transform a set of data in order to maximize the ability to classify, predict, cluster or discover new patterns in target datasets. Deep learning refers to ML 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 field in statistics and machine learning that develop methods for data analysis based on updating the probability for an hypothesis based on observational evidence. 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-effect connection. Recent observational causal inference try to discover causal relations from data.
Convolution	Convolution is one of the most important operations in signal and image processing, and it can operate in 1D (e.g. speech), 2D (e.g. images) or 3D (e.g. video) objects. A convolutional filter is essentially a weight vector/matrix/cube that operates in a sliding window approach on the data. Depending on the kernel structure, the operation enhances some features of the data, such as edges, trends, or flat regions. The operation 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 we discuss in section 5.
Feedforward vs Recurrent networks	An artificial neural network (ANN) is a computational algorithm that simulates how signals are transferred between a network of neurons, via synapses. In aan ANN, informationis transferred only in the forward direction while in a recurrent ANN the information can cycle/loop between the different nodes, creating complex dynamics, like memory, as seen in data.
Generative Adversarial Networks	Family of unsupervised ML methods widely used to generate

(GAN)	realistic samples from an unknown probability density function. GANs are formed by a neural network that generates plausible examples that try to fool a discriminator network that should discern real from fake examples.
Memory effects	Metaphoric term, meaning that the current behavior of a system cannot be explained without considering the effect of past states or forcing variables.
Nowcasting & Forecasting	To forecast a certain variable refers to establish a prediction of its value in the future, from days to centuries. Nowcasting refers to making that prediction in a very near future (e.g. predicting if it is going to rain in a couple of hours).
Probabilistic programming	Probabilistic programming is an approach to define probabilistic models with 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 (RTMs)	Mathematical models that describe how radiation at different wavelengths (e.g. visible light) propagates through different mediums (e.g. atmosphere, vegetation canopy) by simulating absorption, emission, transmission and scattering processes.
Remote sensing	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 geo-scientific variables of interest (temperature, salinity or carbon dioxide), based on propagated reflectance signals (e.g. electromagnetic radiation).
Supervised & Unsupervised learning	In supervised learning an algorithm learns the input-to-output relationship by being provided both the inputs and the respective outputs, e.g. a set of photos (inputs) and a set of corresponding labels (outputs). In unsupervised learning the algorithms do not have access to the labels, so the goal is to infer the underlying structure of the data (e.g. the algorithm automatically separates pictures with different statistical or even semantic properties, e.g. images of cats and dogs).
Teleconnections	Teleconnections refer to climate anomalies related to each other at large distances (typically thousands of kilometers). Quantifying teleconnection patterns allows predicting key patterns on Earth, which are distant in space and time: e.g. predicting El Niño enables prediction of North American rainfall, snowfall, droughts or temperature patterns with a few weeks to months lead time.

See https://developers.google.com/machine-learning/glossary/ and http://www.wildml.com/deep-learning-glossary/ for more complete glossaries.