Towards a Collective Agenda on AI for Earth Science Data Analysis

Devis Tuia, *Senior Member, IEEE*, Ribana Roscher, *Member, IEEE*, Jan Dirk Wegner, Nathan Jacobs, *Senior Member, IEEE*, Xiao Xiang Zhu, *Senior Member, IEEE*, Gustau Camps-Valls, *Fellow, IEEE*.

Abstract

This is the pre-acceptance version, to read the final version published in the Geoscience and Remote Sensing Magazine, please go to: 10.1109/MGRS.2020.3043504
In the last years we have witnessed the fields of geosciences and remote sensing and artificial intelligence to become closer. Thanks to both the massive availability of observational data, improved simulations, and algorithmic advances, these disciplines have found common objectives and challenges to advance the modeling and understanding of the Earth system. Despite such great opportunities, we also observed a worrying tendency to remain in disciplinary comfort zones applying recent advances from artificial intelligence on well resolved remote sensing problems. Here we take a position on research directions where we think the interface between these fields will have the most impact and become potential game changers. In our declared agenda for AI on Earth sciences, we aim to inspire researchers, especially the younger generations, to tackle these challenges for a real advance of remote sensing and the geosciences.

Index Terms

AI, machine learning, causal inference, interpretability, hybrid modeling, physics, domain knowledge, geosciences, climate science, reasoning, new challenges.

INTRODUCTION

Artificial intelligence promises to change the way we do science. Nowadays it is widely accepted, almost a mantra, that data along with faster computers and advanced machine learning algorithms can solve any data science problem. Approaches issued from machine learning, computer vision, applied mathematics, or big data in general, are undoubtedly revolutionizing the way we tackle challenges in remote sensing and geosciences. This is particularly visible since deep learning has entered the arena [1]: the promise of a technology able to process large amounts of data and to learn the complex structures of environmental processes, thus leading to an improved modeling, is making machine learning an unavoidable approach. For a multidisciplinary overview, see the recent book 'Deep learning for the Earth Sciences' [2].

We are convinced that the rise of data-driven approaches in the geosciences is beneficial and will lead to important discoveries. However, we want to raise awareness of pitfalls and fundamental questions that are largely obviated by the community.

DT was with Wageningen University, the Netherlands. He is now with Ecole Polytechnique Fédérale de Lausanne, Sion, Switzerland. E-mail: devis.tuia@epfl.ch (corresponding author). RR is with the University of Bonn, Germany. JDW is with ETH Zurich NJ is with the University of Kentucky (USA), XXZ is with the Technical University of Munich and the German Aerospace Center, GCV is with the Universitat de València, Spain.

TABLE I
A SUMMARY OF THE SIX RESEARCH DIRECTIONS PRESENTED IN THIS POSITION PAPER.

	In a nutshell	Refs.	Current issues	10 years from now	Page
1	Going beyond recognition towards induction, deduction, spatial and temporal reasoning, and structural inference.	[6], [7]	Missing of or very limited benchmarks, novel tasks, as well as reasoning models, interpretability unsolved.	Intelligent systems linking meaningful transformation of entities, e.g. over space or time, and deriving knowledge, as the way people understand visual world and processes.	3
2	Think beyond the raster, consider all possible inputs and sources of supervision, in particular geo-tagged social media data.	[8]–[10]	Presence of dataset biases, presence of label noise. Spatio-temporal mismatch between data sources. Scalability with increasing number of sources.	Systems that use a wide variety of sources to enable fine-grained understanding of the world, all with minimal human effort required for dataset building and system design.	4
3	Query the world by asking questions to images, create descriptions.	[11], [12]	Underexploited language part. Limited choice of thematic interactions. Lack of large scale infrastructure.	Visual search engines understanding questions about images, able to adapt to different types of requests and usable for everyone.	8
4	Make models learned with deep neural networks consistent with domain-specific knowledge like equations from physics.	[1], [13]–[16]	Networks' outputs are not physically consistent. Networks are often used as emulators of simulations, but don't explore beyond current simulators constraints: they can't discover new physical rules.	Systems trainable with much less data, because they constrain output space via physical knowledge. Systems that learn new hypothesis for new science generation.	9
5	Enhancing interpretability and explainability to understand processes in ML models in a better way.	[17], [18]	Lack of human-understandable interpretation. Tendency towards confirmation bias (e.g. with attention maps)	Models that are more understandable, and therefore more reliable and trustworthy. Models that can be queried (and challenged) by humans about their inner reasoning.	13
6	Learn cause-effect relations, not just correlations, from observations and assumptions about the underlying generating process and system.	[19]–[22]	Models can't work with unevenly sampled time series or non-stationary/noisy process. They extrapolate poorly.	Machines that automatically blend domain knowledge, observational data and assumptions to learn the causal graph and generate causal narrative explanations of the problem.	16

A first risk is that of seeing everything as an opportunity of applying machine learning and to then deploy massive technology regardless of whether such technology is necessary and adapted to the problem at hand. Would that become a common practice, one would miss the opportunity of using (and improving) the new technology to tackle new challenges which were impossible before. For example, remote sensing is a data hungry discipline that has embraced machine learning very early: after an exploration phase (see the review papers [3]–[5]), we see now the need for an impulse to embrace the technology to unlock new, difficult problems , which in turn will *create value for these geo-spatial data*. Part of the concepts presented below will sketch some of these directions (see Table I): first, the question on introducing reasoning in the modeling processing as a way to mimic cognitive processes about space and time (direction 1, page 3). Second, the need for exploring unconventional data modalities to capture the complexity of the visual world (direction 2, page 4). And third, new ways of human interaction with remote sensing models, for instance via question answering as a way to retrieve image content on demand (direction 3, page 8).

A second pitfall is the blind faith in data science[1]. Driven by the impressive results obtained in machine learning and computer vision, it would be tempting to believe that everything can be solved with data and algorithms only. We believe that domain knowledge and model assumptions is of prime importance and that models must be challenged i) to respect the reality of the physical/biological/chemical processes governing the system under study and ii) to be accountable - by the transparency of their internal reasoning - of the decisions they lead to. This

---

[1]Data science has been regrettably a misleading term. Shouldn't be replaced with 'data for science' as some researchers suggest? Science is about contrasting hypotheses, understanding physical phenomena and validating causal and explanatory models. If those objectives were achieved through data analysis, a new Science would be born.

is important, especially when models are intended to be used for actual decision making and can affect balances of power or society changing decisions. In the second part of the paper, we will present ideas along these directions (see Table I) centered on the injection of domain knowledge, but for different purposes: first we will discuss physics-aware machine learning, which has the goal of using domain knowledge to restrict solution spaces of the models so that the outcome is physically plausible (direction 4, page 9). This will grant physically consistency of the solutions and avoid aberrant outcomes that break physics (e.g. mass and energy conservation). Then, we will discuss how to obtain human understandable interpretations and explanations of the inner functioning of the models, in order to understand why and how models make decisions (direction 5, page 13). This has the advantage to make the model trustable and non-falsifiable and to avoid that the right conclusions are reached for the wrong reasons. Explainability also enhances the potential of testing new hypothesis and learn new scientific knowledge by the analysis of the model's functioning. These two first directions can be combined and use domain knowledge in different ways with different goals. Transparency of the weights of the models is not absolutely necessary at this stage, as interpretations can be achieved by analyses of the inputs (e.g. LIME [23]) and physics awareness by modified loss functions. Yet, as argued before, science is about understanding the world we live in, not just to approximate it. We argue that without learning causal relations from observational data and assumptions, this ambitious goal of understanding the Earth system will not be possible (direction 6, page 16). In this case, the learning of cause-effect relations is a mix of the previous ingredients, as domain knowledge is needed to design the model in such a way that it can reveal (maybe novel) cause-effect relations, which can be then explained using domain knowledge. Here transparency of the model is achieved by construction, as the product is a self-explanatory causal graph. All three fields (explainability, hybrid modeling and causal inference) also share in common the need of an active and tight interaction between domain experts and computer scientists to make a decisive, non-incremental leap in Earth sciences.

With this position paper we present six research directions that we subjectively think hold particular promise for the future of Earth observation data analysis. In the following, we argue their potential and relevance, and provide some pointers to relevant resources. Our goal is to trigger curiosity and to foster successful research to truly advance the field of Earth sciences with Artificial Intelligence.

I. REASONING AND HUMAN-MACHINE DIALOGUES

Current research at the interface between machine learning and remote sensing largely focuses on the direct recognition of materials, objects or on estimating geo-physical parameters. Reasoning goes beyond the concept of recognition and aims at mimicking how people think and learn. It is centered around tasks such as induction, deduction, spatial and temporal reasoning, and structural inference [24].

To date, only a few pioneering studies are published on reasoning for remote sensing tasks. In computer vision, reasoning is mostly interpreted as the capability to link meaningful transformations of entities over space or time. This is a fundamental property of intelligent species and also the way people understand visual data. Recently, papers implementing reasoning in CNNs started to appear: Santoro et al. proposed a relational reasoning network as a simple plug-and-play module to solve problems requiring the understanding of arbitrary relations between

objects (ordering or comparisons of relative positions/sizes) and applied it to the problem of visual question answering (VQA) [25] (more on VQA on direction 3, page 8). A second pioneering work concerns temporal relations in video sequences. When it comes to understand what takes place between two sampled video frames, humans can easily infer the temporal relations and transformations between observations, unlike neural netowrks. In [26], authors proposed a temporal relation network, which learns intuitive and interpretable common sense knowledge in videos.

Why Should Relational Reasoning Matter in Remote Sensing?

Earth observation images carry strong spatial and temporal information, since each pixel is precisely referenced and connected to neighbors in space and time. When considering land processes (and in particular the geophysical ones), the relevant relationships can be learned by models. In [6], authors explicitly modeled long-range relations for semantic segmentation in aerial scenes. With the aim of increasing the representation capacity of a fully convolutional network (FCN), two tailored relation modules were used: one describing relationships between observations in convolved images and another producing relation-augmented feature representations. Given that convolutions operate by blending spatial and cross-channel information together, they captured relations in both spatial and channel domains.

Perspectives

The work mentioned above showcases how spatial relational reasoning helps in improving semantic understanding of remote sensing images, and many other problems may also benefit from visual reasoning. One exciting example is temporal reasoning for the analysis of multi-temporal data/aerial videos, e.g. for event recognition. This is a new exciting field, where one is concerned by understanding complex events being imaged or filmed, such as cultural events, manifestations, or locating people in distress. Using reasoning enables to understand if a person on a roof during a flood is in actual need of help, or if a video of a crowd is related to a pacific or violent manifestation. This could be of interest to various stake holders including local authorities. To foster this research direction, authors in [7] introduce a dataset named ERA (Event Recognition in Aerial videos), consisting of three thousand UAV videos manually annotated into dozens of types of events (Fig. 1).

Beyond videos, other interesting examples include VQA (see Direction 3, page 8), captioning [27] and audiovisual reasoning, i.e., linking remote sensing images to in-situ audio signals [28]. In the long run, we hope that reasoning Earth observation systems would be capable of deduce clues and make structural inference, in order to explain processes (see direction 5, page 13) and understand causal structures in Earth Systems (see direction 6, page 16).

II. EXTREMELY MULTI-MODAL REMOTE SENSING

Remote sensing is not restricted anymore to observation with airborne or satellite sensors. Nowadays, we can monitor our planet's health and status with social media data, socio-economic indicators, all kind of imagery, audio, and text, in addition to satellite imagery [29]. This direction raises several questions related to the importance of the different sources and their adequateness to specific tasks. In this research direction, we will discuss some of these aspects at the crossroads between sometimes extremely different data sources.

Fig. 1. Overview of the ERA dataset – A benchmark for event recognition from aerial videos [7]. For each class, the first (left) and last (right) frames of an example video are shown.

The traditional approach

The first step in creating a traditional remote sensing system is to identify a property of interest, y (e.g. land use or snow depth). We typically restrict to a set of locations, l, often described as a grid of points within a polygon, and times, t. We can formalize this as modeling a conditional distribution, $P(y|l,t)$, over the property, y, for a given set of locations and times. In the traditional approach, one would resort to a pixel or object based classifier learning input-output relationships from a labeled set of image pixels. The goal is to provide an accurate estimate of the uncertainty over the property, y, and to have this generalize to new locations and times.

The traditional approach is limited in several ways: (1) It requires a person to be able to label the training satellite imagery, either in the field or through manual image interpretation. This can be expensive and limiting, especially when asking the annotator to label difficult, fine-grained label spaces. (2) It is only able to make distinctions between phenomena that are easily visible from an overhead perspective. It cannot, for example, see inside buildings. (3) It is tightly coupled to the geographic region, the task, and source of data. This approach has led to a profusion of remote sensing papers use minor variations of the same computational methods.

Multimodal approaches with social media

Social media data can be used to address our fundamental task, the estimation of $P(y|l,t)$. We begin by considering how social media can be incorporated as an input, by using l and t to query for nearby media content, and then how it can be used to expand the types of properties, y, that can be estimated.

Fig. 2. Example of a system providing likelihood of presence of an object with multiple modalities. Ground and satellite images provide hard and weak (picture-based) labels respectively to create a heatmap of the presence of objects in urban areas. Location information is used to perform the fusion.

a) As an Alternative Input Modality: Traditional remote sensing largely ignores social media, and only uses l and t to index the image. However, today there are countless photographic, audio, and textual data being collected using cellphones. These data are often associated with the location and time of capture, making each a potentially useful source of information about the state of the world (see Fig. 2 where a system based on both remote sensing and ground level images uses both modalities in synergy to provide likelihoods about the presence of objects). People make use of these data to make decisions on a daily basis. For example, when reading reviews and looking at photographs while trying to make travel plans. With the rapid advances in automatic interpretation of imagery, audio, and text we can start thinking of these data as potential inputs for remote sensing. Recent work [30]–[32] has begun to explore the use of social media imagery, especially for fine-grained landuse classification. A new methodological framework of information fusion with machine learning is actually emerging [33]. They tend to use blackbox models to extract vector-valued features from the ground images and combine them with the remote sensing features. They also tend to ignore the rich geometric information the images contain.

b) As a Source of Supervision: An untrained person can easily interpret a wide array of properties from a single social media object. The interpretation will be fine grained and include subjective properties such as dangerousness or scenicness, which are generally not visible in overhead images. With convolutional neural networks approaching, and sometimes exceeding

human-level performance for ground-level image interpretation, we can now consider using the output of CNNs as a semantic description of a given place and time. We can then use this description to train a remote sensing model, which might only take satellite imagery as input. In this way, we can extend the information learned from social media to areas where these media are absent, and simultaneously reduce the need to manually annotate satellite images. This approach has been applied for a variety of tasks, including mapping scenes categories [34], and time-varying visual attributes [35]. However, there remain significant issues to address, including:

- The data to be included: each source must add value, being correlated to the task and independent from each other. In an extremely multi-modal setting, the volume and velocity of data acquired from each source cannot be directly controlled, since they depend on completely independent acquisition systems. Given that, it becomes important to understand how the spatial coverage and quality of each source varies. For example, clouds have a significant impact on optical imagery but do not affect SAR. Similarly, the coverage and quality of various social media sources depend on a variety of conditions, including: the proximity to tourist landmarks, population density, and differences in the culture of a given social network[2]. Therefore, we must choose between satellite sources, but also social media types, social media platforms, and often must apply further filtering. For example, only including data collected from cellphone cameras [38] or of certain types of scenes [34].
- The quality of the matching between sources: it needs to accurately relate the ground and overhead perspectives. The satellite/aerial to ground matching at the scene level is highly challenging due to the large semantic gap between the ground and overhead scene, but still resolvable. For example, the authors in [39] proposed an dual adversarial solution for unsupervised satellite/aerial to ground scene adaptation solution. However, it becomes very crucial when object level matching is concerned, e.g., when approaching automatic geolocalization [40], [41], and in particular when considering image synthesis [42]–[44], where strong geometric models of the various modalities, with the ability to model uncertainty, are strongly needed.

Perspectives

Considering social media data as extra sources for remote sensing analysis is gaining momentum. Besides classification and automatic geolocalization, using these additional data could unlock new applications, like modeling soundscapes [45], landscape scenicness [46] or place perception analysis [47]–[49]. Moving even further, one could study phenomena closer to the consumer providing the data. For example, one could think of a system finding the most visually appealing driving route.

This will inevitably raise the question of dataset biases, since social media data are personalized views of space: photographers tend to take pictures from places that are easy to reach, they are biased in the types of subjects they prefer and tend to take more pictures in the day time and in good weather conditions[3]. This problem was recently considered when using observations collected by citizen scientists for species distribution mapping [50]. In general, biases in learning

---

[2]We refer the interested reader to discussions about social media data quality in [36], [37].

[3]See for instance the oversampling of particular photographic forms and scenes in the Instagram account insta_repeat (https://www.instagram.com/insta_repeat/).

models is a growing topic of study both in the machine learning (see, for example [51], [52]) and the social media (see reviews in [53], [54]) communities. There is wide room for such studies in remote sensing and biases issued by fusion in multimodal settings or hallucinations when using GANs.

Using social media also implies developing models that are robust to differences in appearance of classes, which becomes critical when predicting in new geographies or time moments. Since acquiring new labeled data is not always an option, one could envision to use these alternative sources as a form of weakly supervised training data, or even as unsupervised supervisory signal for knowledge discovery, as authors in [55] proposed to explore the urban latent space of the streetscape of London. Considering data acquired by autonomous vehicles and IoT devices will push this need even further, but will also unlock the potential of mapping on demand with extremely multi-modal remote sensing.

Finally, further integration could make it possible to use social media as an early detection system for events, such as natural disasters [56]. For example, the social media imagery could be used to detect damaged structures or people in distress [57]. Such a system could even be used to cue satellite image acquisition over areas of interest based on image content, location data densities, or Tweets.

III. INTERACTIVE & SEMANTIC MACHINE LEARNING

With the massive increase in availability, remote sensing images are now used beyond scientific research. Firstly, images are available worldwide and with a high update rate. But they are also way more accepted by the general public: nobody is surprised anymore when shown a satellite view from Google maps; consumer level drones can be used by virtually anyone and for all kinds of tasks: farmers monitoring crops, ecologists surveying animals or architects keeping track of construction sites are just a few examples.

But despite the massive potential for image acquisition and updating, the usage of images remains static, in the sense that the images are mostly used for visualization, or at best to compute standard indices such as the NDVI, which is then assumed to represent vegetation status. Moreover, models answering specific needs of users are scarce and the more often limited to classical processing tasks (e.g. cars detection or land cover mapping) and cannot cover the variety of tasks different users could be interested in. Another limitation is that end users rarely have the technical skills to design and run machine learning models, and would like to be able to receive an answer to a specific question of interest asked in natural language (e.g. in english).

Fortunately, a large variety of these questions boils down to the presence of objects, to counting or some kind of relational attribute (e.g. whether there was an increase of forest area or whether there are buildings in risk zones): a model able to pursue some kind of reasoning about the image content (see direction 1, page 3), but taking into account a specific question (in english) by a user could open the door to a new type of interaction with remote sensing. Similarly to what search engines do on the Internet, a Remote Sensing Visual Question Answering (RSVQA [12]) engine could allow anyone (from scientists, to laymen and journalists) to retrieve the relevant information contained in the images.

Research in VQA is a vivid topic in computer vision [11], where it has had a lot of impact in creating systems to support vision impaired people in everyday tasks [58]. A traditional VQA system in this context can indeed be used to help people when buying groceries, crossing the street, etc.
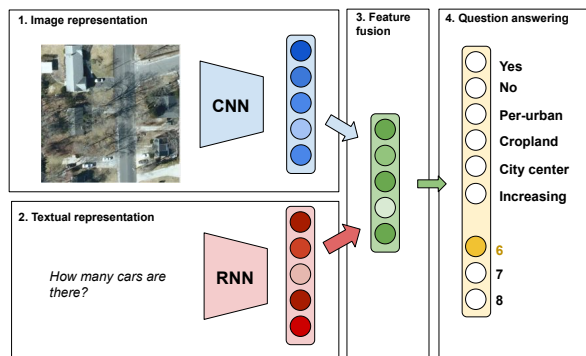
Fig. 3. Example of a RSVQA system (modified from [12]): a remote sensing image (left, top) and a question in natural language (left, bottom) pair enter two source-specific neural nets, and outputting a vector representing their information; (middle) both vectors are combined and become the input of (right) a classifier that outputs possible answers as separate classes.

Dialogues between users and Earth observation images requires both remote sensing and natural language processing

In remote sensing, the first VQA system was proposed in [12] and is summarized in Fig. 3. To become truly general-purpose, such model needs to be trained with a large quantity of data from several areas and different thematic objectives: in [12] two models were designed, one for Sentinel-2 data and another for subdecimeter resolution aerial images; the models were trained with large sets of image/answers pairs spanning tasks of classification, relative position reasoning and objects counting. Since a large quantity of labels was necessary, OpenStreetMap (OSM) vector data were used to automatically generate labels: following the CLEVR protocol [59], 100 questions per image involving objects occurring in the image (as informed by OSM) were generated. For each image/question pair, the answer (i.e. the label) was automatically obtained by querying OSM directly. Data and models are openly available at https://rsvqa.sylvainlobry.com/. Examples of predictions of the RSVQA model are reported in Fig. 4 for both resolution images. Note that for a single image, several questions are possible and the same model is used to answer them.

Perspectives

This first work opens a wide range of possibilities for a new line of research towards the next level of human/image interactions. Nonetheless, all the blocks of the model can (and must) be improved: for example, the automatic data generation has its flaws, especially due to the very simplistic language model used, for which new models from NLP could help improving the performance greatly. Also, less classical tasks (i.e. not reducible to classification, regression or detection) should be imagined, for instance allowing more complex output spaces: lessons learned from image captioning in remote sensing [60] show that it is possible to move towards models that generating descriptions of the image content, which could be used, for instance, in image retrieval [61].

IV. PHYSICS-AWARE MACHINE LEARNING

Seen from the eyes of a practitioner, a major drawback of deep learning models is that they can lead to implausible results with scores that indicate high confidence in the outputs if no
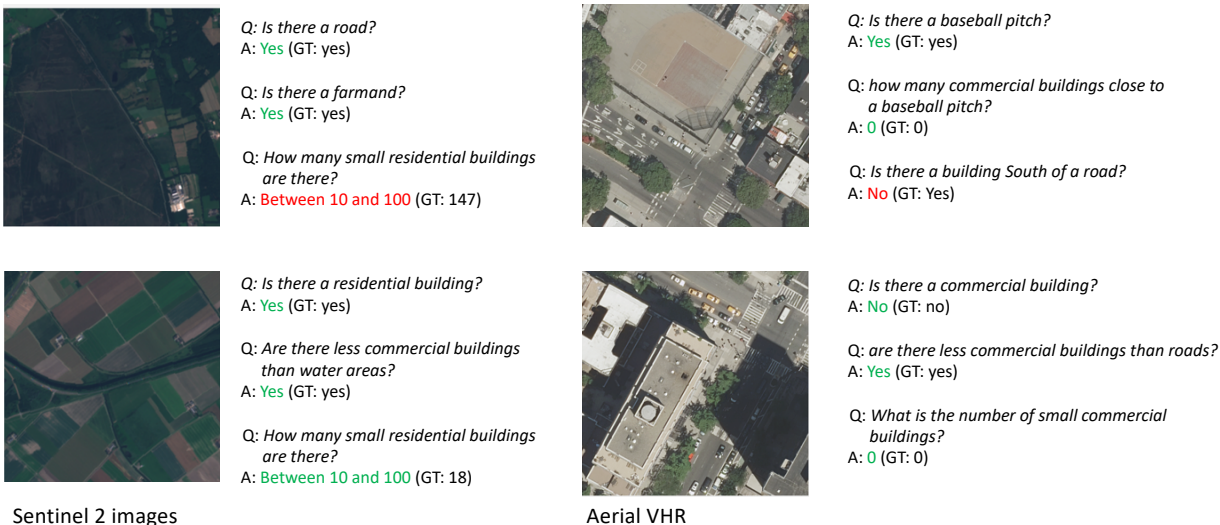
Q: *Is there a road?*
A: Yes (GT: yes)

Q: *Is there a farmand?*
A: Yes (GT: yes)

Q: *How many small residential buildings are there?*
A: Between 10 and 100 (GT: 147)

Q: *Is there a baseball pitch?*
A: Yes (GT: yes)

Q: *how many commercial buildings close to a baseball pitch?*
A: 0 (GT: 0)

Q: *Is there a building South of a road?*
A: No (GT: Yes)

Q: *Is there a residential building?*
A: Yes (GT: yes)

Q: *Are there less commercial buildings than water areas?*
A: Yes (GT: yes)

Q: *How many small residential buildings are there?*
A: Between 10 and 100 (GT: 18)

Q: *Is there a commercial building?*
A: No (GT: no)

Q: *are there less commercial buildings than roads?*
A: Yes (GT: yes)

Q: *What is the number of small commercial buildings?*
A: 0 (GT: 0)

Sentinel 2 images   Aerial VHR

Fig. 4. Predictions of the RSVQA system on two Sentinel and aerial images, respectively, and different questions each. The same model is used to answer all questions related to one resolution imagery.

high-level constraints are imposed that check for consistency with theory.

One possibility for compensating this shortcoming is integrating domain knowledge into the modeling procedure. Particularly in the environmental and geosciences, the laws of physics, chemistry, or biology govern the underlying processes and a lot of theory exists. An interesting direction of research is thus how to tightly couple machine learning and especially deep learning with physical laws. The hope is that this introduction of domain knowledge can help to reduce the manual labeling effort for supervised learning, counter dataset biases, reduce the influence of label noise, lead to good generalization capabilities, and eventually result in plausible outputs that adhere to the underlying physical principles. An advantage of physics equations, and a difference to more generally explainable machine learning (see direction 5, page 13) and causal models (see direction 6, page 16), is that they can often be simplified into differentiable representations, which makes them amenable to backpropagation. In addition, emphasis in physics-consistent machine learning approaches for remote sensing is on modelling natural phenomena with higher accuracy, which is not necessarily the case for the two other research directions. We will present some first ideas below, clustered into three lines of thought: constrained optimization, physics layers in deep neural networks, and encoding and learning differential equations. A recent overview of the main families and approaches to the general field of the interaction between physics and machine learning for Earth observation is available in [62].

Constrained Optimization

A first idea to design physics-consistent machine learning approaches is imposing constraints in the loss function [13], [63]. Loss functions that encode the physical principles of a particular problem while using otherwise mostly unchanged model architectures can ensure that the learned

model respects the laws of physics, see Fig. 5 for an example when including a dependence-based regularizer [52]. In addition, this strategy can significantly reduce the amount of necessary labels for training, up to practically zero in some cases [14].

Designing custom-tailored loss functions and possibly combining them with models that are trained on simulated data is another promising direction of research. However, this approach calls for very specific designs of loss functions that are not always straightforward and may simply not exist for many problems in remote sensing. For example, it seems very hard to design a corresponding loss function for semantic segmentation of cars in aerial images or detection of building facades in street-level panoramas, because the large intra-class variability of the appearances would require a very large set of constraints.

Physics Layers in Deep Neural Networks

An interesting idea to make use of well-established deep neural networks, but still learn and constrain the underlying physics, is adding additional layers that encode physics [1], [64], see Fig. 6 for an example. General background knowledge gained from physics can be encoded in the deeper network layers. Together with a custom-tailored loss function, this approach enables end-to-end training of common deep networks that comply with physical constraints.

Although the idea of adding physical layers on top of common deep network architectures seems intuitive, implementing it for a wide range of remote sensing tasks is far from trivial. One idea could be to start with simplified versions that do not encode physics directly, but some related, simplified constraints, for example, for imposing maximum values for vegetation height mapping [65], tree stress [66], or flood water depth [57].

Encoding and Learning Differential Equations

Probably the biggest step towards deep neural networks that incorporate physics are so-called physics-informed neural networks (PINN) that directly encode nonlinear ordinary differential equations (ODE) and partial differential equations (PDE) in deep learning architectures while allowing for end-to-and training [15], [67]. Instead of using standard network layers, the authors propose a framework to directly encode nonlinear differential equations in the network that is fully end-to-end trainable. This idea allows to learn yet unknown correlations and to come up with novel research hypothesis in a data-driven way, a central point also raised in the previous research direction on interpretability. Probabilistic models like Gaussian processes also allow encoding ODEs as a form of convolutional process [68], and report additional advantages: besides the uncertainty quantification and propagation, they also learn the explicit form of the driving force and the ODE parameters, offering a solid ground for model understanding and interpretability (see next research direction V, on explainable machine learning).

Perspectives

Translated to remote sensing, physics-informed machine learning models allow to encode and learn radiative transfer equations and further physical laws like the backscattering of synthetic aperture radar (SAR) signal. Although directly starting with a full set of forward modeling equations like for simulation engines seems very hard, one could start with simplified versions and a subset of the most important components. Another idea would be encoding a simplified version of the change of spectral properties of vegetation as a function of seasonality. Similar to
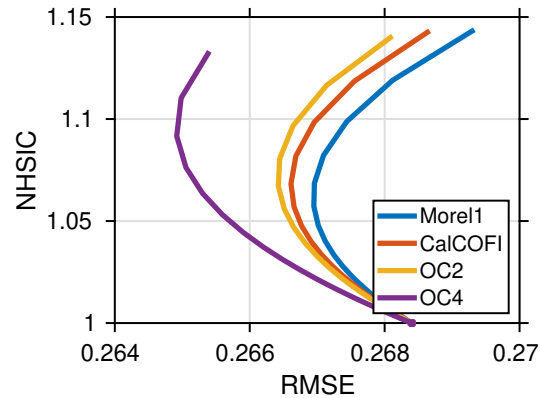
Fig. 5.   A standard family of hybrid modeling can be framed as a constrained optimization problem, where the physical rules are included as a particular form of regularizer [69]. The fair kernel learning (FKL) [52] method enforces model predictions to be not only accurate but also *statistically dependent* on a physical model, simulations, or ancillary observations. In this example, we forced dependence of a data-driven model with respect to four standard ocean color parametric models (Morel1, CalCOFI 2-band linear, OC2 and OC4) and trained our constrained model to estimate ocean chlorophyll content from input radiances. We did so with increased dependency (as estimated by the NHSIC metric) between the machine learning and the physical model. Results show that including the dependence regularizer (i.e. for higher NHSIC values) helps to reduce the RMSE and reveals that the OC2 and OC4 physical models in particular improve error and consistency of the data-driven model.
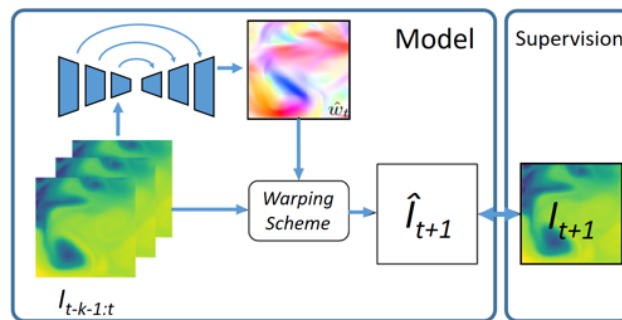

Fig. 6.   Another approach to hybrid ML modeling is that of including layers with physics motivation into a deep neural network, which are learned from data end-to-end. The shown architecture learns a motion field with a convolution-deconvolution net and the motion field is further processed with a warping physical model [70]. The error is used to adjust the network weights, and after training the model can produce multiple time-step predictions recursively. Credits: Figure adapted from [70].

the warping model proposed in [64], one could encode change of spectral canopy properties in the infrared domain to ease domain transfer between summer and winter scenes. In the broader context of geosciences and climate sciences, learning ODEs/PDEs from observational data and simulations is the direct way to explain the problem and variable relations mechanistically, while still resorting to empirical data. The main challenges are the needed simplicity of the ODEs/PDEs that scientists can understand (so sparsity gets involved here) as well as validating the plausibility of such equations (so domain experts and computer scientists should work together). This links strongly physics-based deep learning to the explainable machine learning discussed in the next section.

V. INTERPRETABLE AND EXPLAINABLE MACHINE LEARNING

Using machine learning for scientific applications aims at acquiring new scientific knowledge from observational data. Additionally to the accuracy of the results, their scientific consistency, reliability, and explainability are of central importance. A prerequisite to achieve those is to design models that can be challenged; in other words to create models whose inner functioning can be visualized, queried or interpreted. In this section, we will discuss the foundations of explainable AI (Fig. 7), its exciting perspectives, and make links with physics-aware/informed machine learning that were discussed in the previous section on physics-based ML (page 9).

From Transparency to Explainability

Explainable machine learning has various definitions (see [17]), but they all revolve around the properties of (1) transparency, (2) interpretability and (3) explainability:

1) A transparent model allows us to access its components and ideally motivate why certain model components were chosen. This is in contrast to black box models as traditional neural networks, for which one could indeed write the mathematical relationships explicitly (they are transparent in this sense), but their complexity makes it inaccessible for users.

2) An interpretable model counteracts the lack of transparency by presenting complex facts like the processes in a neural network in a space that can be understood by humans. Sorting by increasing interpretability power, such space can be made of localized image coordinates [71], semantic concepts [72] or understandable text [73].

3) To achieve explainability, domain knowledge is exploited, which is used in combination with the interpretable model and its components to understand, for example, why the model came to a certain decision. Therefore, explanations become application-dependent and identical interpretations can lead to different explanations when linked to different domain knowledge.

How Explainable AI may help remote sensing?

Recently, many tools have been proposed for increasing interpretability and explainability when combined with domain knowledge [18]. Two major groups emerge:

1) *Post-hoc interpretability*. In this group, outcomes and decisions of the model are interpreted and explained by looking at the input. The most common visualizations for interpretations are heatmaps and prototypes. Heatmaps highlight parts of the input data that are prominent, important, or occlusion-sensitive. For example, they are created using the gradients flows in the neural network. Prototypes are optimized input data that, given a model, maximize the targeted output. Both these approaches help to understand what a model bases its decisions on, what influences the output, or what is a typical input for the learned input-output relationships. In all cases, attention must be paid to the confirmation bias. This is defined as the tendency to try to explain interpretations that are consistent with our existing knowledge, even if the explanation does not apply to the given case (see [74] for an example of overinterpretation of saliency maps).

2) *Interpretability by design*. In this case, the model is inherently designed so that it can be interpreted. Interpretability is achieved by representing model components or obtained latent variables in a way that they can be explained with knowledge from a certain application domain.
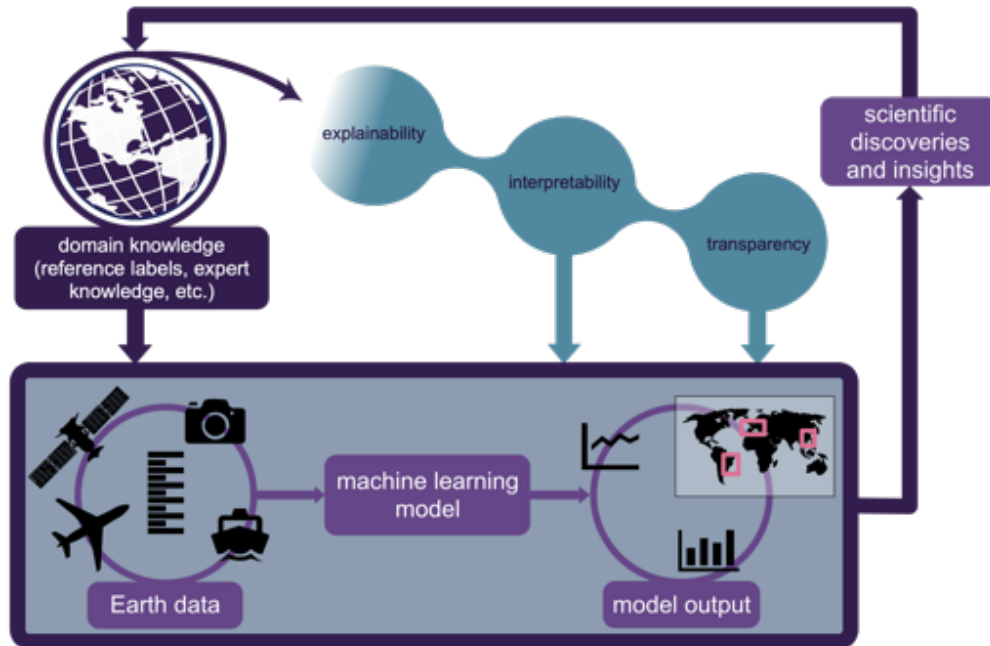
Fig. 7. Explainable Machine learning can be used to gain scientific discoveries and insights by explaining a learned model and/or results (shown in the light gray box). Prerequisites are interpretability and potentially transparency that lead to scientific explanations when combined with domain knowledge. A feedback loop allows to extend and improve known domain knowledge. One application would be the derivation of improved definitions, for example for certain land use classes, which are currently only vaguely, incompletely, or not uniformly described.

For example, units in hidden layers can be designed in such a way that the underlying factors of variation such as driving forces in Earth system data become disentangled and are captured in separate units. This could be seen, for example, by the fact that simple correlations exist between variations of the input and the activation of the neurons. Interesting applications of this idea are proposed in [75], where authors disentangle physical forces applying between objects in videos or in [76], discussed below, for explaining human perception of beauty in landscapes.

To ensure scientific value of the output, interpretation tools can be used to check its reliability. Besides the inherently existing output score of the neural network, for example, visualizations of the processes within the neural network can be used to check whether correct decisions have been made for the wrong reasons (so-called Clever-Hans-effect, [78]). This can be seen as an additional test for the reliability of the output, due to the fact that a high score of the network does not always mean a correct result. In summary, these tools can increase confidence by improving traceability as estimates are generated, and reveal biases in the data through human-understandable visualization.

Perspectives

Explainable machine learning has so far received comparatively little attention in remote sensing, partly because of the still predominant opinion that explainability is tightly coupled with the complexity of a model and thus, an increase in explainability leads directly to a decrease in
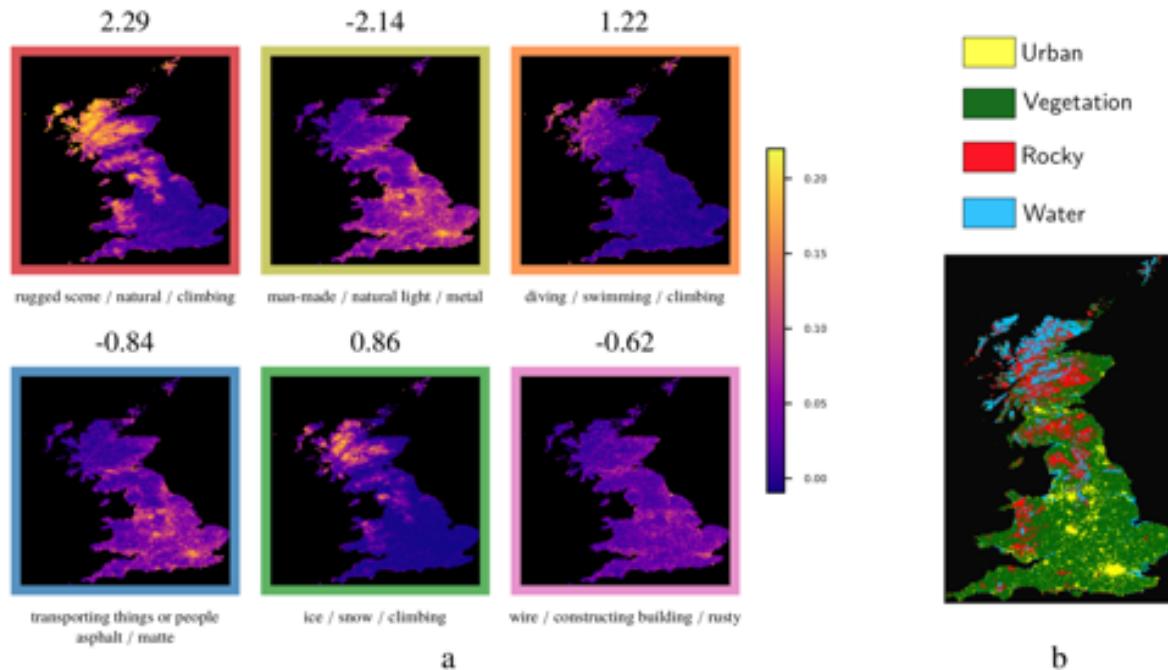
Fig. 8. Explaining factors behind landscape beauty in the UK (from [77]): (a) Maps of landscape attributes contributing to the estimation of beauty in a series of landscape images; these maps are learned from ground based pictures and used to predict landscape attributes observed in the single images. Such attributes are then combined end-to-end in a neural network that predicts the beauty scores. The number above the single factor maps correspond to the contribution of the factor to landscape beauty on average. (b) Land cover map of the UK, used for visual comparison of the single factor maps learned automatically.

accuracy (e.g., [79]). In the meantime, however, several applications have shown that this is not the case anymore.

Most approaches consider so far are post-hoc interpretations, but first approaches considering interpretability by design are appearing. In [76], for example, the model is forced to predict human-interpretable concepts before predicting the final task (Figure 8). Such approaches have the potential to provide both reliability checks and human understandable explanations, and could be used to go towards physics-explicit models as those discussed in the next section.

As a further step, not only explanations can be allowed that are already understood with today's knowledge. New insights could be gained by using the explainable machine learning model and outcomes to formulate hypotheses for explanations that are not yet known to us. These hypotheses could then be tested with simulation software, for example, and confirmed or rejected. In order to show the potential, previous neural network approaches discover presumed scientific laws from observations without providing the complete underlying prior knowledge to the learning method (e.g., [80]). As illustrated in Fig. 7, the next promising step would be to reveal new hypotheses from remote sensing data. This can, for example, be accomplished by searching patterns in interpretations, which can be assigned to novel discoveries and insights when combined with domain knowledge. This approach may point us to previously undiscovered spatio-temporal input-output relationships or data biases. For example, interpretation tools and the resulting insight could use how certain land-use classes are recognized to derive improved

class definitions and help with a targeted acquisition of training data.

VI. Learning cause-effect relations from data

The Earth is a highly complex, dynamic, and networked system where very different physical, chemical and biological processes interact in several spheres, and at diverse spatio-temporal scales. Despite the great predictive capabilities of current machine and deep learning methods, there is still little actual *learning*. Understanding is harder than predicting. Machine learning algorithms excel in fitting arbitrary functional data relations but do not have a clear notion of the underlying causal relations. Machine learning is far from problem understanding and even more from machine intelligence[4].

Earth system data analysis aims to extract information from multivariate non-grided datasets, where missing data, nonlinearities and nonstationarities are present in the wild. Variables and physical processes are coupled in space and time, and (tele)connections can be of large-range, discontinuous are variant in strength and intensity. Addressing this problem will allow us to identify the right set of predictors, develop robust models and to avoid getting the right answer for the wrong reasons. The links with physics (direction 4, page 9) and interpretability (direction 5, page 13) are very strong.

Causality as the way forward

Causal inference aims at discovering and explaining the causal structure of the system [20], [81], [82]. Very often, interventions in the system are not possible because of ethical, practical or economical reasons. Then *observational* causal inference comes into play to extract cause-effect relationships from multivariate datasets, going beyond the commonly adopted correlation approach, which merely captures associations between variables.

Today the science of "causal inference" [19], [83] is fast advancing and, under reasonable assumptions, can unravel *causal* relations between two or more coupled variables even in the presence of non-linearities and non-stationarities, and even when time is not even involved. Several rigorous algorithms have been developed in the last decade that allow us to make inferences across multiple variables to discover plausible causal relations from observations. Causal inference is of course very relevant for the scientific endeavour, but it also has impactful practical implications. For example, learning causal structures allows us to build more parsimonious and robust models and that means faster, more fault-tolerant, and interpretable models.

A taxonomy of causal discovery methods

Causal discovery methods can be divided into four main families. First, *Granger Causality (GC)* [84] is the most widely used approach in Earth and climate sciences to quantitatively identify relations between time series. It tests if including past states of a variable X improves the prediction of an output variable Y more than considering other covariates. GC is a linear test, but nonlinear (kernel) versions have been proposed [85]. In [86] a generalized kernel GC is presented able to discover footprints of El Niño-Southern Oscillation (ENSO) on soil moisture (SM) and vegetation optical depth (VOD) records (see Fig. 9a). However, GC approaches have

---

[4]See critical perspectives in the blogs by Gary Marcus and Michael Jordan, and the perspective papers [1], [22]

problems in nonstationary, nonlinear and deterministic relations, especially in dynamic systems with weak to moderate coupling. The second family considers nonlinear state-space methods, such as the *Convergent Cross-Mapping (CCM)* [87]. CCM tries to address GC problems by reconstructing the variable's state spaces (M_x, M_y) using time embeddings, and conclude on $X \rightarrow Y$ if points on M_x can be predicted using nearest neighbors in M_y more accurately as more points are used. However, CCM is very sensitive to noise and time series length. Recent works included bootstrap resampling to alleviate such problems, and showed good results in identifying causal links in long global records of carbon and water fluxes [88], see Fig. 9(b). The third family, collectively known as *causal network learning algorithms*, heavily relies on conditional independence tests. Methods iteratively remove links between pairs of variables (X, Y) if they are found independently conditioned on any subset of the other variables. The PC algorithm allows to identify parents and can be flexibly implemented with different kinds of conditional independence tests, which can handle nonlinear dependencies and variables that are discrete or continuous, and univariate or multivariate. Finally, *structural causal models (SCM)* are used when *time* is not involved or the sampling frequency is too low. SCMs search for the causal direction within Markov equivalence classes by exploiting asymmetries between cause and effect. Additive noise models rely on the principle of independence between the cause and the generating mechanism, and have recently shown good results in remote sensing and geosciences in cases where time is not involved and only two variables are observed [21].

Perspectives

While the field of machine and deep learning has traditionally progressed very rapidly, we observe that this is not the case in tackling the challenge of learning causal relations from Earth observation data. The role that deep learning will play on causal discovery is at best uncertain, since deep learning models mostly focus on fitting and are largely overparameterized, which is (apparently) against the causal, sparse, reasoning. Only very recently we witnessed efforts towards either incorporating or understanding deep models *causally*: [89], implements a meta-learning objective that maximizes the speed of domain transfer, which under certain assumptions can be seen as a way to localize changes in causal mechanisms. In [90] authors learn individual-level causal effects from observational data that can efficiently handle confounding (hidden) factors. Both methods are in principle well suited to the problems in remote sensing and geoscience datasets, which exhibit spatio-temporal relations to be exploited, but have not (so far) been considered.

Yet, we will have to face a more important challenge, the *cognitive barriers*. Domain knowledge is elusive and difficult to encode, interaction between computer scientists and physicists is still a barrier, and education in synergistic concepts still needs to become a reality in the coming years. Causal inference is named to be the way to develop Earth sciences, but this will only be possible with a strong and continuous interaction between domain knowledge experts and computer scientists.

CONCLUSION

Six ideas, six directions where the geosciences, Earth observation and artificial intelligence still have a lot to achieve if synergystically combined. With this position paper we have provided our appreciation of research avenues that are new, refreshing and exciting for scientists willing

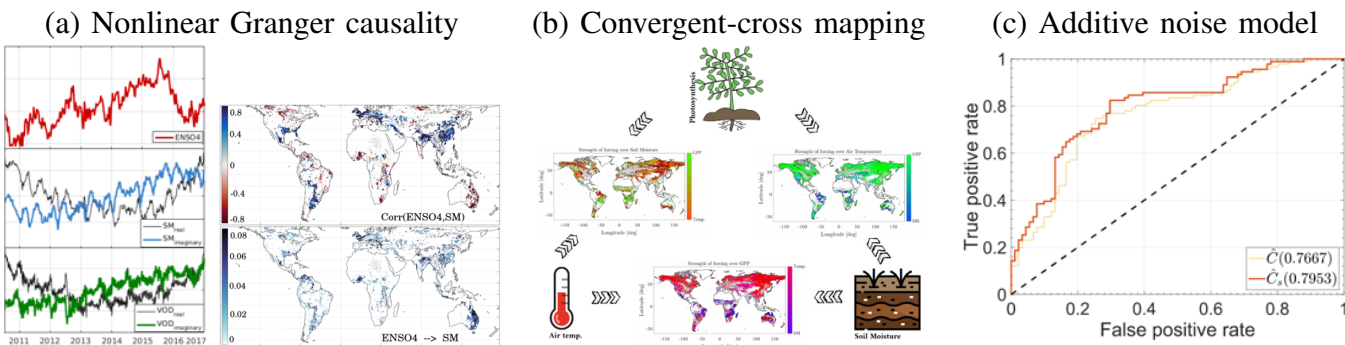(a) Nonlinear Granger causality   (b) Convergent-cross mapping   (c) Additive noise model

Fig. 9. Examples of causal inference approaches in remote sensing and the geosciences. **(a)** On the left, we show time series of ENSO4 (ENSO in region 4), which captures sea surface temperature anomalies in the central equatorial Pacific, SM and VOD, in order to explore their causal relations extracted with nonlinear PCA in [91]; on the right we show the 5-day lagged correlation (top) and and causal (bottom) maps of ENSO4 and SM inter-annual components using the kernel Granger causality method in [86]. Results show that many of the correlations are not causal, even the highest ones ($\rho \sim 0.8$), thus suggesting mere spurious associations. **(b)** Example of the application of the unbiased CCM in [88] to derive causal relations between variables accounting for photosynthesis (gross primary productivity, GPP from FLUXCOM), temperature (T_{air} from from ERA Interim) and soil moisture (SM from ESA's CCI (v 2.0)). Datacubes at $0.5°$ and 8 day spatial and temporal resolutions respectively, spanning $2001 - 2012$ were used. Reasonable spatial causal patterns are observed for SM and Tair on GPP; GPP drives T_{air} mostly in cold ecosystems (probably due to changes in land surface albedo such as snow/ice to vegetation changes); SM is mostly controlled by T_{air}, which partially drives evaporation in water-limited regions; and GPP dominates SM. **(c)** Structural equation models in the form of additive noise model with kernels in [21] for hypothesis testing. Assessing cause-effect relations is also possible when time is not involved. We here rely on a look-up-table (LUT) generated with by and RTM which gives the right direction of causation: state vectors (parameters) cause radiances. The algorithms accurately detect this from pairs of data, and can be used for retrieval model-data intercomparison and RTM assessment.

to evolve at the interface between AI and the geosciences. We hope they will sparkle curiosity and that the community, especially the younger generations, will embrace them.

REFERENCES

[1] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, and J. Denzler, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, pp. 195–204, 2019.

[2] G. Camps-Valls, D. Tuia, X. X. Zhu, and M. Reichstein, *Deep learning for Earth Sciences - A comprehensive approach to remote sensing, climate science and geosciences*. UK: Wiley & Sons, 2021.

[3] X. Zhu, D. Tuia, L. Mou, G. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[4] N. Audebert, B. Le Saux, and S. Lefevre, "Deep learning for classification of hyperspectral data: A comparative review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 2, pp. 159–173, 2019.

[5] Q. Yuan, H. Shen, T. Li, Z. Li, S. Li, Y. Jiang, H. Xu, W. Tan, Q. Yang, J. Wang, J. Gao, and L. Zhang, "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sensing of Environment*, vol. 241, p. 111716, 2020.

[6] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 12 416–12 425.

[7] L. Mou, Y. Hua, P. Jin, and X. X. Zhu, "ERA: A dataset and deep learning benchmark for event recognition in aerial videos," *IEEE Geoscience and Remote Sensing Magazine*, pp. 1–6, in press.

[8] Y. Aytar, C. Vondrick, and A. Torralba, "See, hear, and read: Deep aligned representations," *arXiv preprint arXiv:1706.00932*, 2017.

[9] J.-M. Perez-Rua, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "Mfas: Multimodal fusion architecture search," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[10] A. R. Zamir, A. Sax, N. Cheerla, R. Suri, Z. Cao, J. Malik, and L. J. Guibas, "Robust learning through cross-task consistency," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 197–11 206.

[11] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra, and D. Parikh, "VQA: Visual Question Answering," in *Int. Conf. Comp. Vis. (ICCV)*, 2015.

[12] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "RSVQA: visual question answering for remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, in press.

[13] A. Karpatne, W. Watkins, J. Read, and V. Kumar, "Physics-guided neural networks (pgnn): An application in lake temperature modeling," 2017.

[14] R. Stewart and S. Ermon, "Label-free supervision of neural networks with physics and domain knowledge," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, ser. AAAI'17.   AAAI Press, 2017, p. 2576–2582.

[15] M. Raissi, P. Perdikaris, and G. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019.

[16] G. Camps-Valls, L. Martino, D. H. Svendsen, M. Campos-Taberner, J. Muñoz-Marí, V. Laparra, D. Luengo, and F. J. García-Haro, "Physics-aware Gaussian processes in remote sensing," *Applied Soft Computing*, vol. 68, pp. 69–82, 2018.

[17] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *IEEE Access*, vol. 8, pp. 42 200–42 216, 2020.

[18] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Múller, "Toward interpretable machine learning: Transparent deep neural networks and beyond," *arXiv:2003.07631*, 2020.

[19] J. Pearl and D. Mackenzie, *The Book of Why*.   New York: Basic Books, 2018.

[20] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference - Foundations and Learning Algorithms*, ser. Adaptive Computation and Machine Learning Series.   Cambridge, MA, USA: MIT, 2017.

[21] A. Pérez-Suay and G. Camps-Valls, "Causal inference in geoscience and remote sensing from observational data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 3, pp. 1502–1513, 2019.

[22] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Clymour, M. Kretschmer, M. Mahecha, J. Muñoz-Marí, E. van Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, and J. Zscheischler, "Inferring causation from time series with perspectives in Earth system sciences," *Nature Communications*, vol. 10, no. 2553, 2019.

[23] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[24] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building Machines That Learn and Think Like People," *arXiv:1604.00289 [cs, stat]*, Nov. 2016, arXiv: 1604.00289. [Online]. Available: http://arxiv.org/abs/1604.00289

[25] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," in *Advances in neural information processing systems*, 2017, pp. 4967–4976.

[26] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 803–818.

[27] X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, "Description generation for remote sensing images using attribute attention mechanism," *Remote Sensing*, vol. 11, no. 6, p. 612, Mar 2019. [Online]. Available: http://dx.doi.org/10.3390/rs11060612

[28] D. Hu, X. Li, L. Mou, P. Jin, D. Chen, L. Jing, X. Zhu, and D. Dou, "Cross-task transfer for multimodal aerial scene recognition," *Proceedings of the ECCV*, 2020. [Online]. Available: arXivpreprintarXiv:2005.08449

[29] S. Lefèvre, D. Tuia, J. D. Wegner, T. Produit, and A. S. Nassar, "Towards seamless multi-view scene analysis from satellite to street-level," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1884–1899, 2017.

[30] S. Workman, M. Zhai, D. Crandall, and N. Jacobs, "A unified model for near/remote sensing," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[31] J. Kang, M. Körner, Y. Wang, H. Taubenböck, and X. X. Zhu, "Building instance classification using street view images," vol. 145, pp. 44–59. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271618300352

[32] S. Srivastava, J. E. Vargas-Muñoz, and D. Tuia, "Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution," *Remote sensing of environment*, vol. 228, pp. 129–143, 2019.

[33] S. Salcedo-Sanz, P. Ghamisi, M. Piles, M. Werner, L. Cuadra, A. Moreno-Martínez, E. Izquierdo-Verdiguier, J. Muñoz-Marí, A. Mosavi, and G. Camps-Valls, "Machine learning information fusion in earth observation: A comprehensive review of methods, applications and data sources," *Information Fusion*, 2020.

[34] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1–9.

[35] T. Salem, S. Workman, and N. Jacobs, "Learning a Dynamic Map of Visual Appearance," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, acceptance rate: 25%.

[36] R. Kitchin, "Big data and human geography: opportunities, challenges and risks," *Dialogues in Human Geography*, vol. 3, no. 3, pp. 262–267, 2013.

[37] J. E. Vargas, S. Srivastava, D. Tuia, and A. X. Falcão, "OpenStreetMap: Challenges and opportunities in machine learning and remote sensing," *IEEE Geosci. Remote Sens. Mag.*, in press.

[38] M. Zhai, T. Salem, C. Greenwell, S. Workman, R. Pless, and N. Jacobs, "Learning geo-temporal image features," in *British Machine Vision Conference (BMVC)*, 2018.

[39] J. Lin, L. Mou, T. Yu, X. Zhu, and Z. J. Wang, "Dual adversarial network for unsupervised ground/satellite-to-aerial scene adaptation," in *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, pp. 10–18. [Online]. Available: https://dl.acm.org/doi/10.1145/3394171.3413893

[40] Y. Shi, X. Yu, L. Liu, T. Zhang, and H. Li, "Optimal feature transport for cross-view image geo-localization." in *AAAI*, 2020, pp. 11 990–11 997.

[41] Y. Shi, X. Yu, D. Campbell, and H. Li, "Where am i looking at? joint location and orientation estimation by cross-view matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4064–4072.

[42] X. Deng, Y. Zhu, and S. Newsam, "What is it like down there? generating dense ground-level views and image features from overhead imagery using conditional generative adversarial networks," in *Proc. SIGSPATIAL*, 2018.

[43] K. Regmi and M. Shah, "Bridging the domain gap for ground-to-aerial image matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 470–479.

[44] X. Lu, Z. Li, Z. Cui, M. R. Oswald, M. Pollefeys, and R. Qin, "Geometry-aware satellite-to-ground image synthesis for urban areas," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 859–867.

[45] T. Salem, M. Zhai, S. Workman, and N. Jacobs, "A multimodal approach to mapping soundscapes," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2018.

[46] S. Workman, R. Souvenir, and N. Jacobs, "Understanding and mapping natural beauty," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[47] Y. Zhu and S. Newsam, "Spatio-temporal sentiment hotspot detection using geotagged photos," in *Proc. SIGSPATIAL*, 2016.

[48] F. Zhang, B. Zhou, L. Liu, Y. Liu, H. H. Fung, H. Lin, and C. Ratti, "Measuring human perceptions of a large-scale urban region using machine learning," *Landscape Urban Plan.*, 2018.

[49] F. Zhang, B. Zhou, C. Ratti, and Y. Liu, "Discovering place-informative scenes and objects using social media photos," *R. Soc. Open. Sci.*, 2019.

[50] T. J. Bird, A. E. Bates, J. S. Lefcheck, N. A. Hill, R. J. Thomson, G. J. Edgar, R. D. Stuart-Smith, S. Wotherspoon, M. Krkosek, J. F. Stuart-Smith, G. T. Pecl, N. Barrett, and S. Frusher, "Statistical solutions for error and bias in global citizen science datasets," *Biological Conservation*, vol. 173, pp. 144 – 154, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0006320713002693

[51] H. Bahng, S. Chun, S. Yun, J. Choo, and S. J. Oh, "Learning de-biased representations with biased representations," 2020.

[52] A. Pérez-Suay, V. Laparra, G. Mateo-García, J. Muñoz-Marí, L. Gómez-Chova, and G. Camps-Valls, "Fair Kernel Learning," in *Machine Learning and Knowledge Discovery in Databases*, ser. Lecture Notes in Computer Science, M. Ceci, J. Hollmén, L. Todorovski, C. Vens, and S. Džeroski, Eds. Springer International Publishing, 2017, pp. 339–355.

[53] P. Cihon and T. Yasseri, "A biased review of biases in twitter studies on political collective action," *Frontiers in Physics*, vol. 4, p. 34, 2016.

[54] F. Morstatter and H. Liu, "Discovering, assessing, and mitigating data bias in social media," *Online Social Networks and Media*, vol. 1, pp. 1 – 13, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2468696416300040

[55] S. Law and M. Neira, "An unsupervised approach to geographical knowledge discovery using street level and street network images," in *SIGSPATIAL workshop GEOAI*, 2019.

[56] M. Imran, F. Ofli, D. Caragea, and A. Torralba, "Using AI and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions," *Information Processing and Management*, vol. 57, no. 5, p. 102261, 2020.

[57] P. Chaudhary, S. D'Aronco, J. Leitão, K. Schindler, and J. D. Wegner, "Water level prediction from social media images with a multi-task ranking approach," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 167, pp. 252–262, 2020.

[58] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people," in *Comp. Vis. Pattern Rec. (CVPR)*, 2019.

[59] J. Johnson, L. Fei-Fei, B. Hariharan, C. L. Zitnick, L. van der Maaten, and R. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," in *Comp. Vis. Pattern Rec. (CVPR)*, 2017.

[60] Z. Shi and Z. Zou, "Can a machine generate humanlike language descriptions for a remote sensing image?" *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3623–3634, 2017.

[61] G. Sumbul, S. Nayak, and B. Demir, "SD-RSIC: Summarization driven deep remote sensing image captioning," 2020.

[62] G. Camps-Valls, D. H. Svendsen, J. Cortes-Andres, J. Moreno-Martinez, A. Perez-Suay, J. Adsuara, I. Martin, M. Piles, J. Munoz-Mari, and L. Martino, "Living in the physics and machine learning interplay for earth observation," in *AAAI Fall Series 2020 Symposium on Physics-guided AI for Accelerating Scientific Discovery*, 2020.

[63] A. Karpatne, G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar, "Theory-guided data science: A new paradigm for scientific discovery from data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 10, pp. 2318–2331, 2017.

[64] E. de Bézenac, A. Pajot, and P. Gallinari, "Deep learning for physical processes: incorporating prior scientific knowledge," in *6th International Conference on Learning Representations (ICLR)*, 2018.

[65] N. Lang, K. Schindler, and J. D. Wegner, "Country-wide high-resolution vegetation height mapping with Sentinel-2," *Remote Sensing of Environment*, vol. 233, 2019.

[66] U. Kälin, N. Lang, C. Hug, A. Gessler, and J. D. Wegner, "Defoliation estimation of forest trees from ground-level images," *Remote Sensing of Environment*, vol. 223, pp. 143–153, 2019.

[67] M. Raissi, "Deep Hidden Models: Deep Learning of Nonlinear Partial Differential Equations," *Journal of Machine Learning Research*, vol. 19, pp. 1–24, 2018.

[68] D. Svendsen, M. Piles, J. Muñoz-Marí, D. Luengo, L. Martino, and G. Camps-Valls, "Integrating domain knowledge in data-driven earth observation with process convolutions," Submitted.

[69] L. Von Rueden, S. Mayer, J. Garcke, C. Bauckhage, and J. Schuecker, "Informed machine learning–towards a taxonomy of explicit integration of knowledge into machine learning," *Learning*, vol. 18, pp. 19–20, 2019.

[70] E. de Bézenac, A. Pajot, and P. Gallinari, "Deep learning for physical processes: incorporating prior scientific knowledge," *Journal of Statistical Mechanics: Theory and Experiment*, 2019.

[71] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[72] M. M. Losch, M. Fritz, and B. Schiele, "Interpretability beyond classification output: Semantic bottleneck networks," *CoRR*, vol. abs/1907.10882, 2019. [Online]. Available: http://arxiv.org/abs/1907.10882

[73] D. Huk Park, L. Anne Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, "Multimodal explanations: Justifying decisions and pointing to the evidence," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[74] J. Adebayo, J. Gilmer, M. Muelly, I. J. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Neural Information Processing Systems*, 2018.

[75] T. Ye, X. Wang, J. Davidson, and A. Gupta, "Interpretable intuitive physics model," in *Proc. ECCV*, 2018.

[76] D. Marcos, S. Lobry, and D. Tuia, "Semantically interpretable activation maps: What-where-how explanations within CNNs," *Int. Conf. Computer Vision Workshop*, pp. 4207–4215, 2019.

[77] D. Marcos, S. Lobry, R. Fong, N. Courty, R. Flamary, and D. Tuia, "Contextual semantic interpretability," in *Asian Conference on Computer Vision (ACCV)*, 2020.

[78] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking Clever Hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, no. 1, p. 1096, 2019.

[79] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[80] R. Iten, T. Metger, H. Wilming, L. del Rio, and R. Renner, "Discovering physical concepts with neural networks," *Phys. Rev. Lett.*, vol. 124, p. 010508, Jan 2020.

[81] K. Zhang, B. Schölkopf, P. Spirtes, and C. Glymour, "Learning causality and causality-related learning: Some recent progress," *National Science Review*, vol. 5, no. 1, pp. 26–29, 2018.

[82] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Clymour, M. Kretschmer, M. Mahecha, E. van Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, and J. Zscheischler, "Inferring causation from time series with perspectives in Earth system sciences," *Nature Communications*, 2019.

[83] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed. New York, NY, USA: Cambridge University Press, 2009.

[84] C. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, pp. 424–438, 1969.

[85] D. Marinazzo, M. Pellicoro, and S. Stramaglia, "Kernel method for nonlinear granger causality," *Phys. Rev. Lett.*, vol. 100, p. 144103, Apr 2008.

[86] D. Bueso, M. Piles, and G. Camps-Valls, "Cross-information kernel causality: Revisiting global teleconnections of enso over soil moisture and vegetation," in *Climate Informatics 2019*, Paris, France, 2-4 October 2019 2019.

[87] G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch, "Detecting causality in complex ecosystems," *science*, vol. 338, no. 6106, pp. 496–500, 2012.

[88] G. Camps-Valls, E. Diaz, J. Adsuara, M. Piles, A. Moreno, P. Gentine, M. Jung, M. Reichstein, and S. W. Running, "Inferring causal graphs from observational long-term carbon and water fluxes records," in *AGU Fall Meeting*, San Francisco, USA, 9-13 December 2019. 2019.

[89] Y. Bengio, T. Deleu, N. Rahaman, R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, and C. Pal, "A meta-transfer objective for learning to disentangle causal mechanisms," 01 2019.

[90] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 6446–6456. [Online]. Available: http://papers.nips.cc/paper/7223-causal-effect-inference-with-deep-latent-variable-models.pdf

[91] D. Bueso, M. Piles, and G. Camps-Valls, "Nonlinear PCA for Spatio-Temporal Analysis of Earth Observation Data," *IEEE Transactions on Geoscience and Remote Sensing*, 2019.