Machine Learning for the Geosciences: Challenges and Opportunities

Anuj Karpatne, Imme Ebert-Uphoff, Sai Ravela, Hassan Ali Babaie, and Vipin Kumar

Abstract—Geosciences is a field of great societal relevance that requires solutions to several urgent problems facing our humanity and the planet. As geosciences enters the era of big data, machine learning (ML)—that has been widely successful in commercial domains—offers immense potential to contribute to problems in geosciences. However, geoscience applications introduce novel challenges for ML due to combinations of geoscience properties encountered in every problem, requiring novel research in machine learning. This article introduces researchers in the machine learning (ML) community to these challenges offered by geoscience problems and the opportunities that exist for advancing both machine learning and geosciences. We first highlight typical sources of geoscience data and describe their common properties. We then describe some of the common categories of geoscience problems where machine learning can play a role, discussing the challenges faced by existing ML methods and opportunities for novel ML research. We conclude by discussing some of the cross-cutting research themes in machine learning that are applicable across several geoscience problems, and the importance of a deep collaboration between machine learning and geosciences for synergistic advancements in both disciplines.

Index Terms—Machine learning, earth science, geoscience, earth observation data, physics-based models

1 Introduction

Momentous challenges facing our society require solutions to problems that are geophysical in nature [1], [2], [3], [4], such as predicting impacts of climate change, measuring air pollution, predicting increased risks to infrastructures by disasters such as hurricanes, modeling future availability and consumption of water, food, and mineral resources, and identifying factors responsible for earthquake, landslide, flood, and volcanic eruption. The study of such problems is at the confluence of several disciplines such as physics, geology, hydrology, chemistry, biology, ecology, and anthropology that aspire to understand the Earth system and its various interacting components, collectively referred to as the field of geosciences.

As the deluge of big data continues to impact practically every commercial and scientific domain, geosciences has also witnessed a major revolution from being a data-poor field to a data-rich field. This has been possible with the advent of better sensing technologies (e.g., remote sensing satellites and deep sea drilling vessels), improvements in computational resources for running large-scale simulations of Earth system models, and Internet-based democratization of data that has enabled the collection, storage, and processing of data on crowd-sourced and distributed environments such as cloud platforms. Most geoscience data sets are publicly available and do not suffer from privacy issues that have hindered adoption of data science methodologies in areas such as health-care and cyber-security. Furthermore, there are large new efforts to standardize geoscience data sets [5] to make it easier to mine them, and to create data sets specifically for machine learning, e.g., Microsoft's new effort on *AI for Earth*. The growing availability of big geoscience data offers immense potential for machine learning (ML)—that has revolutionized almost all aspects of our living (e.g., commerce, transportation, and entertainment)—to significantly contribute to geoscience problems of great societal relevance. For instance, a 2018 report by the World Economic Forum [6] identifies machine learning as a key discipline for solving many of these problems.

There are several communities working on the emerging field of ML for geosciences. These include, but are not limited to, Climate Informatics [7], Climate Change Expeditions [8], and ESSI [9]. More recently, NSF has funded a research coordination network on Intelligent Systems for Geosciences (IS-GEO) [10], with the intent of forging stronger connections between the ML and geoscience communities. On the educational side, the NSF is now funding three related NSF Research Trainee (NRT) programs, namely *Data Science for Energy and Environmental Research* at the University of Chicago; *Environment and Society: Data Sciences for the 21st Century* at UC Berkeley; and the *Computational Geoscience Program* at Stanford University.

Furthermore, a number of leading conferences in machine learning and data mining (e.g., SIGKDD, ICDM, SDM, and NIPS) have included workshops or tutorials on topics related

- *A. Karpatne and V. Kumar are with the University of Minnesota, Minneapolis, MN 55455. E-mail: {karpa009, kumar001}@umn.edu.*
- *I. Ebert-Uphoff is with Colorado State University, Fort Collins, CO 80523. E-mail: iebert@colostate.edu.*
- *S. Ravela is with the Massachusetts Insititue of Technology, Cambridge, MA 02139. E-mail: ravela@mit.edu.*
- *H. Babaie is with Georgia State University, Atlanta, GA 30302. E-mail: hbabaie@gsu.edu.*

Manuscript received 2 Sept. 2017; revised 7 May 2018; accepted 8 July 2018. Date of publication 31 July 2018; date of current version 3 July 2019.
(Corresponding author: Anuj Karpatne.)
Recommended for acceptance by X. Lin.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2018.2861006

to geosciences. The role of big data in geosciences has also been recognized in recent perspective articles (e.g., [11], [12]) and special issues of journals and magazines (e.g., [13]). Most importantly, in the past couple of years we are witnessing a paradigm shift in the geoscience community. Thanks to the increasing number of success stories, machine learning has finally earned its place in mainstream geoscience conferences and journals, and geoscience-related departments are actively seeking to hire or team up with faculty with significant ML knowledge. This opens up huge opportunities for collaboration for machine learning researchers.

Despite the need and growing attention in ML for geosciences, there are several properties of geoscience problems that have to be taken into account while designing ML algorithms. Some of these properties are also shared by other application domains, e.g., spatio-temporal structure of data—a prevalent feature in geoscience applications—is also experienced in applications such as computer vision, bio-medical imaging, neuroscience, and transportation [14]. However, it is the combination of these properties in the context of geoscience problems that presents novel challenges and opportunities for research in machine learning.

A unique aspect of geoscience problems that differentiate it from mainstream applications in the commercial sector is that geoscience phenomena are governed by physical laws and principles, developed over centuries of systematic research by the scientific community. For example, the motion of water in the lithosphere, or of air in the atmosphere, is governed by principles of fluid dynamics such as the Navier–Stokes equation. While these physical principles underlie all geoscience observations collected via different sources, they also form the basis of physics-based models that can *simulate* geoscience phenomena. This availability of scientific knowledge, either in the form of heuristic rules, closed-form equations, or as complex numerical models of dynamical systems, presents unprecedented opportunities for designing, learning, and applying ML algorithms that can address the over-arching challenges encountered in gesocience problems.

The purpose of this article is to introduce researchers in ML to (a) geoscience data and their properties, (b) geoscience problems where ML can have an impact, (c) the challenges faced by existing ML approaches in these problems due to combinations of geoscience properties, and (d) opportunities for novel research in ML. We hope that this paper helps the ML community to identify solutions to socially relevant problems and create novel foundations in ML, e.g., the paradigm of theory-guided data science, that may be broadly applicable to problems even outside the scope of geosciences. The remainder of this article is organized as follows. Section 2 provides an overview of the types and origins of geoscience data. Section 3 describes some of the common gesocience properties that have to be accounted for during ML research. Section 4 outlines four categories of geoscience problems where machine learning can yield major advances, highlighting the challenges and opportunities for ML in every problem. Section 5 discusses two cross-cutting themes of research in ML that are generally applicable across all gesocience problems. Section 6 provides concluding remarks by briefly discussing the best practices for collaboration between machine learning researchers and geoscientists.

2 Overview of Geoscience Data

Data about the Earth system components and processes are generally obtained from three broad data sources: (a) remotely sensed data acquired by Earth observing satellites, (b) data collected by *in-situ* sensors in the sea, land, or the air, and (c) simulation data generated from physics-based models of the Earth system. We briefly describe these categories of gesocience data sources in the following. A detailed review of Earth science data sets and sources for obtaining them can be found in [15].

2.1 Remote Sensing Data

There is a growing body of space research organizations such as the National Aeronautics and Space Administration (NASA), European Space Agency (ESA), and Japan Aerospace Exploration Agency (JAXA) and private companies such as Planet.org and Digital Globe that are together contributing to the huge volume and variety of remote sensing data about our Earth, many of which are freely available (e.g., see [16]). We have a growing collection of Earth observing satellites in space that monitor several geoscience variables such as surface temperature, humidity, optical reflectance, and chemical compositions of the atmosphere. Remote sensing data provides a global picture of the history of geoscience variables at fine spatial scales (1 km to 10 m, and less) and at regular time intervals (monthly to daily) for long periods, sometimes starting from the 1970s (e.g., Landsat archives [17]). For targeted studies in specific geographic regions of interest, remote sensing data can also be collected using imaging sensors on-board flying devices such as unmanned aerial vehicles (drones) or airplanes, e.g., to detect and classify sources of methane (a powerful greenhouse gas) being emitted into the atmosphere [18]. Remote sensing data sets are commonly available as rasters over regularly-spaced grid cells in space and time, and can be represented as geo-registered images over individual time points or as time series data at individual spatial locations

2.2 Sensor Data

Another major source of geoscience observations is the collection of *in-situ* sensors placed on ground (e.g., weather stations) or moving in the atmosphere (e.g., weather balloons) or the ocean (e.g, ships and ocean buoys). Sensor data constitute some of the most reliable and direct sources of information about the Earth's weather and climate systems and are actively maintained by public agencies such as the National Oceanic and Atmospheric Administration (NOAA) [19]. Sensor-based measurements from rain and river gauges are also central for understanding hydrological processes such as surface water discharge [20]. Land-based seismic sensors, Global Positioning System (GPS)-enabled devices, and other geophysical instruments also continuously measure the Earth's geological structure and processes [21]. In addition, we also have proxy measurements such as paleoclimatic records that are sparsely available at a select few locations but go back several thousands of years. Sensor data are generally available over non-uniform grids in space and at irregular intervals of time, sometimes even over moving bodies such as balloons, ships, or buoys. Sensor data are commonly represented as point reference data (also termed as geostatistical data in the spatial statistics literature) of continous spatio-temporal fields.

2.3 Model Simulation Data

Physics-based models, that use physical principles to *simulate* the evolution of the states of the Earth system using numerical methods, are the standard workhorse for studying a majority of geoscience processes. Physics-based models generate large volumes of simulation data of different components of the Earth system, using inputs such as initial and boundary conditions or values of internal parameters in physics-based equations. They are developed and maintained by a number of centers constituting of diverse groups of researchers around the world. For example, the World Climate Reserach Programme (WCRP) develops and distributes simulations of General Circulation Models (GCM) of climate variables such as sea surface temperature and pressure under the Coupled Model Intercomparison Project (CMIP) [22]. Simulations of terrestrial processes related to the lithosphere and biosphere are produced by the Community Land Model (CLM) [23], developed by a number of international agencies collaborating with the National Center for Atmospheric Research (NCAR).

3 Geoscience Properties

There are several properties of geoscience data that are common across many applications. Some of these properties arise from the nature of geoscience processes, others are due to the procedures used for collecting geoscience observations and ground truth. In the following, we review some of these properties using illustrative examples.

3.1 Spatio-Temporal Structure

Since almost every geoscience phenomenon occurs in the realm of space and time, geoscience observations generally show spatio-temporal auto-correlation when observed at appropriate resolutions. For example, a location that is covered by a certain land cover label (e.g., forest, shrubland, urban) is generally surrounded by locations that have similar land cover labels. This type of spatial auto-correlation is known as Tobler's First Law of Geography [24]. Land cover labels are also consistent along time, i.e., the label at a certain time is related to the labels in its immediate temporal vicinity. While spatio-temporal auto-correlation ensures strong connectivity among observations in close vicinity of space and time, a unique aspect of geoscience processes is that they also show long-range spatial and temporal dependencies. For example, a commonly studied phenomenon in climate science is teleconnections [25], [26], where two distant regions in the world show strongly coupled activity in climate variables such as temperature or pressure. As another example, geoscience processes also show long-memory characteristics in time, e.g., the effect of climate indices such as the El Niño Southern Oscillation (ENSO) and Atlantic Multidecadal Oscillation (AMO) on global floods, droughts, and forest fires [27], [28].

3.2 Heterogeneity in Space and Time

Geoscience processes show a high degree of variability in space and time, leading to a rich heterogeneity in geoscience data sets. For example, due to the presence of varying geographies, vegetation types, rock formations, and climatic conditions in different regions of the Earth, the characteristics of geoscience variables vary significantly from one location to the other. Furthermore, the Earth system is not stationary in time and goes through many cycles, ranging from seasonal and decadal cycles to long-term geological changes (e.g., glaciation, polarity reversals) and even climate change phenomena, that impact all local processes. This heterogeneity of geoscience processes needs to be accounted for while modeling the joint distribution of geoscience variables across all points in space and time.

3.3 High Dimensionality

The Earth system is incredibly complex with a huge number of potential variables that interact with each other. Furthermore, geoscience phenomena are not limited to the Earth's surface, but extend beneath the Earth's surface (e.g., in the study of groundwater, faults, or petroleum) and across multiple layers in the mantle or the atmosphere. Capturing the effects of these multiple variables at fine resolutions of space and time renders geoscience data inherently high dimensional. Section 4.4 provides an example of a problem involving one million variables.

3.4 Lack of Concise Object Definitions

Geoscience objects and events, e.g., cyclones, atmoshpheric rivers, and ocean eddies, generally have amorphous boundaries in space and time that are not as crisply defined as objects in other domains, such as users on a social networking website, or products in a retail store. Hence, there can be multiple ways to define a geoscience object or event in continous spatio-temporal fields. Further, the form, structure, and patterns of geoscience objects and events are much more complex than those found in discrete spaces that ML algorithms typically deal with, such as items in market basket data. For example, storms and hurricanes dynamically deform in complicated ways over very short durations of time, which has to be kept in mind during their detection and tracking.

3.5 Rare Classes

In a number of geoscience problems, we are interested in studying objects, processes, and events that occur infrequently in space and time but have major impacts on our society and the Earth's ecosystem. For example, extreme weather events such as cyclones, flash floods, and heat waves can result in huge losses of human life and property, thus making it vital to monitor them for adaptation and mitigation requirements. These processes may relate to emergent (or anomalous) states of the Earth system, or other features of complex systems such as anomalous state trajectories and basins of attractions [29]. As another example, detecting rare changes in the Earth's biosphere such as deforestation, insect damage, and forest fires can be helpful in assessing the impact of human actions and informing decisions to promote ecosystem sustainability.

3.6 Multi-Source Multi-Resolution Data

Information about the Earth system is collected via different data sources at varying scales of space and time, which given their complementary strengths often have to be analyzed together for a variety of geoscience objectives. These data sets may exhibit varying characteristics, such as sampling rate, accuracy, and uncertainty. For example, in-situ sensors, such as buoys in the ocean and hydrological and weather measuring stations, are often irregularly spaced. As another example, collecting high-resolution data of ecosystem processes, such as forest fires, sometimes require using aerial imageries from planes flying over the region of interest, which may need to be combined with coarser resolution satellite imageries available at frequent time intervals. The analysis of multi-resolution geoscience data sets can help us characterize processes that occur at varying scales of space and time. For example, processes such as plate tectonics and gravity occur at a global scale, while local processes include volcanism, earthquakes, and landslides. In some cases, it is possible to convert one data type to another and across different spatial and temporal resolutions using simple interpolation methods or more advanced methods based on physical understanding such as reanalysis techniques [30].

3.7 Poor Quality of Data

Many geoscience data sets (e.g., those collected by Earth observing satellite sensors) are plagued with noise and missing values. For example, sensors may temporarily fail due to malfunctions or severe weather conditions, resulting in missing data. Additionally, changes in measuring equipment, e.g., replacing a faulty sensor or switching from one satellite generation to the next, may change the interpretation of sensor values over time, making it difficult to deploy a consistent methodology of analysis across different time periods. Furthermore, many sensor properties can increase noise, such as sensor interference, e.g., in the case of remotely sensed land surface data, where atmospheric (clouds and other aerosols) and surface (snow and ice) interference are constantly encountered. Even data generated from model outputs have uncertainties because of our imperfect knowledge of the initial and boundary conditions of the system or the parametric forms of approximations used in the model.

3.8 Small Sample Size

Even though many geoscience applications involve large amounts of data, e.g., global observations of ecosystem variables at high spatial and temporal resolutions using Earth observing satellites, the number of samples available for solving some of the pressing challenges in geosciences is often limited. For example, most satellite products are only available since the 1970s, and when monthly (yearly) processes are considered, this means that less than 600 (50) samples are available. The spatial and temporal resolution of some geoscience variables is also limited by the nature of observation methodology. For example, paleo-climate data are derived from coral, lake sediments (varves), tree rings, and deep ice core samples, which are only available at a few places around the Earth. Similarly, early records of precipitation only exist in areas covered by land.

3.9 Paucity of Ground Truth

In supervised learning problems in geosciences, we often encounter paucity of labeled samples that have gold-standard ground truth. This is because high-quality measurements of several geoscience variables can only be taken by expensive apparatus such as low-flying airplanes, or tedious and time-consuming operations such as field-based surveys, which severely limit the collection of ground truth samples. Other geoscience processes (e.g., subsurface flow of water) do not have ground truth at all, since, due to the complexity of the system, the exact state of the system is never fully known. This is in contrast to commercial applications involving Internet-scale data, e.g., text mining or object recognition, where large volumes of labeled data have been one of the major factors behind the success of machine learning methodologies such as supervised deep learning methods.

4 Geoscience Applications for ML

In the following, we discuss four broad categories of applications in geosciences that provide promising directions for ML research. For every application, we present a brief description of the problem from an ML perspective using illustrative examples, the challenges faced by traditional ML algorithms, and the opportunities for novel research in ML, providing illustrative examples of recent successes wherever possible.

4.1 Detecting Objects and Events

Problem Description. Detecting objects and events in geoscience data is important for a number of reasons. For example, detecting spatial and temporal patterns in climate data can help in tracking the formation and movement of objects such as cyclones, weather fronts, atmospheric rivers, and ocean eddies, which are responsible for the transfer of precipitation, energy, and nutrients in the atmosphere and ocean. While traditional approaches for characterizing geoscience objects and events are primarily based on the use of hand-coded features (e.g., ad-hoc rules on size and shape constraints for finding ocean eddies [31]), machine learning algorithms can enable their automated detection from data with improved performance using pattern mining techniques.

ML Challenges. A major challenge in using ML approaches for this problem is the lack of concise definitions of objects and events in geoscience data, that appear with amorphous boundaries in continuous spatio-temporal fields. Furthermore, geoscience objects and events are highly dynamic in nature, e.g., floods and forest fires show sudden changes in contrast to slow and persistent progression of objects in mainstream ML applications such as computer vision. This challenge gets compounded when the objects or events are rare and thus occur infrequently, making their detection further difficult.

ML Opportunities. There is a need to develop novel pattern mining approaches that can account for the spatial and temporal properties of objects and events, e.g., spatial coherence and temporal persistence, that can work with amorphous boundaries. One such approach has been successfully used for finding spatio-temporal patterns in sea surface height data [32], [33], resulting in the creation of a global catalogue of mesoscale ocean eddies [34]. The use of topic models has also been explored for finding extreme events from climate time series data [35]. A promising research direction for ML is to develop techniques that can consider both the pattern and dynamics of coherent objects, as explored for simple geoscience applications in [36], [37]. Finally, recent work to detect extreme weather events using deep learning are discussed in Section 5.1.

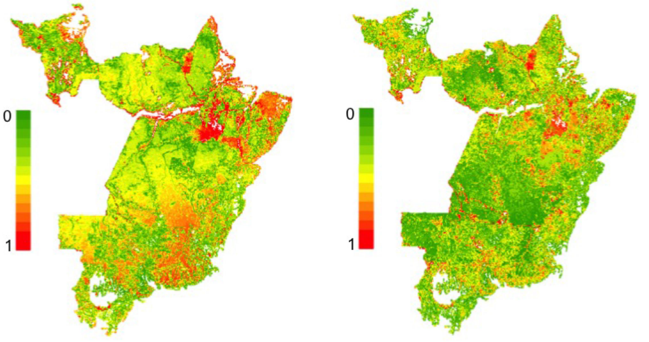
4.2 Estimating Geoscience Variables

Problem Description. An important category of geoscience problems where ML can contribute is to estimate physical variables that are difficult to monitor directly, e.g., methane concentrations in air or groundwater seepage in soil, using information about other observed or simulated variables. For example, supervised learning algorithms can be used to produce estimates of ecosystem variables such as forest cover, health of vegetation, water quality, and surface water availability, using remote sensing data sets that are available at fine resolutions of space and time. Such estimates can help in informing management decisions and enabling scientific studies of changes occurring on the Earth's surface.

ML Challenges. One of the major challenges in using ML algorithms for estimating geoscience variables is the rich heterogeneity in the characteristics of geoscience process across space and time. Further, the source of heterogeneity in the data is often unknown and attributed to a number of physical factors, e.g., changes in topography, land cover, climatic zone, season, and temporal regime. This heterogeneity makes it desirable to learn a different model for every homogeneous partition in the data, using the training data only for that partition. However, partitioning the training data into smaller portions further exacerbates the impact of paucity of labeled data, especially when some partitions suffer from paucity of training samples. This challenge is even more serious when we are interested in mapping rare phenomena such as forest fires, because we often have an inadequate number of data samples from the rare class. Hence, the combined effects of heterogeneity, paucity of labeled data, and rare classes make it difficult for standard ML algorithms to achieve good prediction performance.

Another challenge for ML algorithms is the poor quality of geoscience data due to noise and missing values, that increases the risks of spurious estimates. While noise and missing values are common in many other application domains, particularly in spatio-temporal applications, geoscience data show novel types of structured noise and missing values that are uncommon in other domains. For example, in remote sensing applications, the presence of cloud and aerosols results in spatially and temporally auto-correlated structures of noise, going against the common salt-and-pepper noise assumption in spatio-temporal problems. Hence, common approaches developed for handling noise and missing values in other spatio-temporal applications, e.g., using data imputation or post-processing methods such as Markov random fields, are not directly applicable in many geoscience problems.

ML Opportunities. To address the combined effects of heterogeneity and small sample size, there is an opportunity to explore recent advances in ML such as multi-task learning frameworks [38], [39], where the learning of a model at every homogeneous partition of the data is considered as a separate task, and the models are shared across similar tasks. This sharing of learning can help in regularizing the models across all tasks and avoid the problem of overfitting. An example of a multi-task learning based approach for handling heterogeneity can be found in a recent work in [40], where the learning of a forest cover model at every vegetation type (discovered by clustering vegetation time-series at locations) was treated as a separate task, and the similarity among vegetation types (extracted using hierarchical clustering techniques) was used to share the learning at related tasks. Fig. 1 shows the improvement in prediction performance of forest cover in Brazil using a multi-task learning approach. While this example shows the promise in using multi-task learning methods for geoscience problems, existing formulations of multi-task learning require explicit knowledge of the construction of the tasks and the relationships between them, which is not always available in geoscience problems. Novel methods are thus required to automatically extract the form of heterogeneity in the data using our physical understanding of geoscience processes, and use this knowledge in the construction of multi-task learning frameworks.



(a) Absolute resdiual errors of the baseline method.

(b) Absolute resdiual errors of the multi-task learning method presented in [40].

Fig. 1. Performance improvement in estimation of forest cover in four states of Brazil using a multi-task learning method. Figure courtesy: Karpatne et al. [53].

To address the heterogeneity (or non-stationary nature) of climate data, online learning algorithms have been developed to combine the ensemble outputs of expert predictors (climate models) and produce robust estimates of climate variables such as temperature [41], [42]. In this line of work, weights over experts were updated in an adaptive way across space and time, to capture the right structure of non-stationarity in the data. This was shown to significantly outperform the baseline technique used in climate science, which is the non-adaptive mean over experts (multi-model mean). Another approach for addressing non-stationarity was explored in [43] for downscaling climate variables, where a Bayesian mixture of models for every homogeneous cluster of locations in space was learned. In a recent work, adaptive ensemble learning methods [44], [45] have been developed to address the challenge of heterogeneity for mapping the dynamics of surface water bodies using remote sensing data [46]. This has enabled the creation of a global surface water monitoring system (publicly available at [47]) that is able to discover a variety of changes in surface water such as shrinking lakes due to droughts, melting glacial lakes, migrating river courses, and constructions of new dams and reservoirs.

To address the paucity of labeled data, novel learning frameworks such as semi-supervised learning, that leverages the structure in the unlabeled data for improving classification performance [48], and active learning, where an expert annotator is actively involved in the process of model building [49], have huge potential for improving the state-of-the-art in estimation problems encountered in gesocience applications [50], [51]. In a recent line of work, attempts to build a machine learning model to predict forest fires in the tropics using remote sensing data led to a novel methodology for building predictive models for rare phenomena [52] that can be applied in any setting where it is not possible to get high quality labeled data even for a small set of samples, but poor quality labels (perhaps in the form of heuristics) are available for all samples.

4.3 Long-Term Forecasting of Geoscience Variables

Problem Description. Forecasting long-term trends of geoscience variables such as temperature and greenhouse gas concentrations ahead in time, can help in modeling future scenarios and devising early resource planning and adaptation policies. The standard approach for generating forecasts of geoscience variables is to run physics-based model simulations, which basically encode geoscience processes using state-based dynamical systems where the current state of the system is influenced by previous states and observations using physical laws and principles. From a machine learning perspective, the problem of forecasting can be treated as a time-series regression problem where the future conditions of a geoscience variable have to be predicted based on present and past conditions. Some of the existing methods in ML for time-series forecasting include exponential smoothing techniques [54], autoregressive integrated moving average (ARIMA) models [55], state-space models [56], and probabilistic models such as hidden Markov models and Kalman filters [57], [58]. Machine learning methods for forecasting climate variables using the spatial and temporal structure of geoscience data have been explored in recent works such as [59], [60], [61], [62].

ML Challenges. A key challenge in predicting the long-term trends of geoscience variables is to develop approaches that can represent and propagate prediction uncertainties, which is particularly difficult given the small sample size (limited number of years with reliable historical records) faced in geoscience applications and the non-stationary nature of geoscience processes [63], [64]. Further, the heavy-tailed nature of extreme events such as cyclones and floods aggravates the challenges in producing their long-term forecasts.

ML Opportunities. There is a need to develop novel ML methods that can capture long-range temporal dependencies while producing robust uncertainty estimates of its predictions, even with longer forecasting horizons. In a recent work [59], regression models based on extreme value theory have been developed to automatically discover sparse temporal dependencies and make predictions in multivariate extreme value time series. Other approaches for predicting extreme weather events such as abnormally high rainfall, floods, and tornadoes using climate data have also been explored in [65], [66], [67]. Novel methods are also needed to quantify uncertainty in the results of ML methods such as deep learning, particularly given the scientific relevance of the discovered results and societal implications of incorrect conclusions.

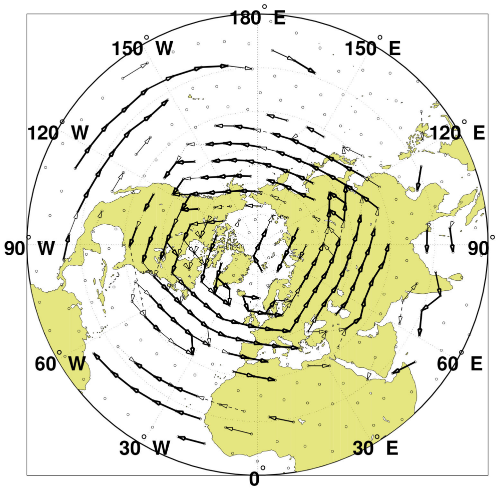


Fig. 2. Network plot for Northern hemisphere generated from daily geopotential height data using constraint-based structure learning of graphical models. The resulting arrows represent the pathways of storm tracks, see [77].

4.4 Mining Relationships in Geoscience Data

Problem Description. An important problem in geoscience applications is to understand how different physical processes are related to each other, e.g., periodic changes in the sea surface temperature over eastern Pacific Ocean—also known as the El Niño-Southern Oscillation—and their impact on several terrestrial events such as floods, droughts, and forest fires [27], [28]. Identifying such relationships from geoscience data can help us capture vital signals of the Earth system and advance our understanding of geoscience processes.

A common class of relationships that is studied in the climate domain is *teleconnections*, which are pairs of distant regions that are highly correlated in climate variables such as sea level pressure or temperature. One widely-studied category of teleconnections is dipoles [26], [68], which are pairs of regions with strong negative correlations (e.g., the ENSO phenomena).

One of the first ML attempts in discovering relationships from climate data is a seminal work by Steinbach et al. [69]. In this work, graph-based representations of global climate data were constructed in which each node represents a location on the Earth and an edge represents the similarity (e.g., correlation) between the climate time series observed at a pair of locations. Dipoles and other higher-order relationships (e.g., tripoles involving triplets of regions) could then be discovered from climate graphs using clustering and pattern mining approaches. Another family of approaches for mining relationships in climate science is based on representing climate graphs as complex networks [70]. This includes approaches for examining the structure of the climate system [71], studying hurricane activity [72], and finding communities in climate networks [73], [74]. Finally, causality-based networks, based typically on either the framework of *Granger causality* [75] or of *Pearl causality* [76] can be used to detect potential cause-effect relationships in geoscience applications. Such methods can be used, for example, to track interactions between different locations. Fig. 2, for example, shows the pathways of storm tracks obtained using Pearl's causality framework [77].

ML Challenges. Formidable challenges for ML arise in the problem of relationship mining due to the high dimensionality of geoscience data, leading to massive search space of candidate patterns. To illustrate this challenge, consider the problem of mining teleconnections, e.g., the impact of rising temperatures in eastern Pacific ocean (termed as the El Niño phenomena) on forest fires in California. For this problem, the number of variables to be considered for evaluating relationships is proportional to the number of locations considered in both the regions of the teleconnection, which can be in the order of thousands even at coarse spatial resolutions of 2.5^o. As another example, to analyze interactions in the atmosphere between three different atmospheric fields, using a grid with 1 degree resolution in both longitude and latitude, and including the Earth' surface and four higher altitudes, involves the study of interactions between $3 \times 360 \times 180 \times 5$-roughly one million-observed time series variables. Further, given the multiple ways that are possible to define geoscience objects and their relationships, we need to simultaneously extract spatio-temporal objects, their relationships, and their dynamics. An additional challenge in relationship mining is the small sample size due to the limited number of years available for capturing relationships among objects.

Finally, causality-based networks can be applied to many applications using existing methods. For other geoscience applications it is necessary to devise suitable adaptations to handle the challenge of high dimensionality, deal with the prevalence of hidden common causes in geoscience applications, and deal with effects due to spatial auto-correlations [77]

ML Opportunities. There is a need for novel approaches that can directly discover the relationships as well as the interacting objects in geoscience data [26], [78]. For example, recent work on the development of such approaches have led to the discovery of previously unknown climate phenomena [79], [80], [81]. Methods for handling the high dimensionality of geoscience data along with small sample sizes have also been explored in [82], where sparsity-inducing regularizers such as sparse group Lasso were developed to model the domain characteristics of climate variables.

As for causality analysis, the most common tool used to date in the geosciences is bivariate Granger analysis [83], followed by multi-variate Granger analysis using vector autoregression (VAR) models [84], but the latter is still not commonly used. Pearl's framework based on probabilistic graphical models has only rarely been used in the geosciences to date [85], [86]. The fact that multi-variate causality tools based on VAR/LASSO or Pearl analysis, which have yielded tremendous breakthroughs in biology and medicine over the past decade, are still not commonly used in the geosciences, is in stark contrast to the huge potential these methods have for tackling numerous geoscience problems. Thus there is abundant opportunity for ML experts to collaborate with geoscientists to generate new scientific insights in numerous geoscience applications.

Finally, many components of the Earth system are affected by human actions, thus introducing the need for causal attribution [84], [86] in decision making. Causal attribution based on Pearl causality is another well developed framework in the ML community, that most geoscientists are not familiar with, opening up huge potential for collaborative work to yield advances in causal attribution in the geosciences.

5 Cross-cutting Research Themes

In this section, we discuss two emerging themes of ML research that are applicable across many of the geoscience problems discussed in Section 4. This includes deep learning and the paradigm of theory-guided data science, as described below.

5.1 Deep Learning

Deep learning methods, that use several hidden layers in artificial neural network (ANN) architectures, have revolutionized the scale and impact of ML in several applications of commercial significance, e.g., computer vision, speech recognition, and language translation. The power of deep learning can be attributed to its use of a deep hierarchy of latent features (learned at hidden nodes) where complex features are represented as compositions of simpler features. This, in conjunction with the availability of large volumes of labeled samples, e.g., the 1-million ImageNet data set [87] for computer vision problems, has made it possible for deep learning methods to extract complex and highly non-linear features automatically from the data that show remarkably better accuracy than contemporary ML methods, sometimes even beating human-level performance [88].

Given the growing success of deep learning methods in mainstream ML applications, there is a huge opportunity to develop and apply deep learning methods for a number of problems encountered in geosciences. This is especially true given the spatio-temporal nature of geoscience data that is well-suited for deep learning developments such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). For example, deep learning methods can be used to detect geoscience objects and events, e.g., atmospheric rivers and weather fronts, automatically from the data as spatio-temporal patterns, without needing to build hand-coded features. Indeed, the use of CNNs for detecting extreme weather events from climate model simulations has recently been demonstrated in [89], [90]. Deep learning methods can also be used for estimating geoscience variables from observations using supervised or semi-supervised techinques. For example, a recent work explored the use of deep learning methods to model the temperature of water in lakes across depth and over time using data for lake drivers such as air temperature, wind speed, and solar radiation, in conjunction with the outputs of physics-based models of lake temperature [91]. Deep learning based frameworks have also been explored for downscaling outputs of Earth system models and generating climate change projections at local scales [92], and classifying objects such as trees and buildings in high-resolution satellite images [93]. In another development, RNN-based frameworks such as long-short-term-memory (LSTM) models have been explored for mapping plantations in Southeast Asia from remote sensing data, using spatial as well as temporal properties of the dynamics of plantation conversions [94], [95], [96]. Given the ability of LSTM methods to extract the right length of memory needed for making future predictions in time, deep learning methods can also be useful for long-term forecasting of geoscience variables with appropriate lead times.

While the success of deep learning methods in commercial domains has been greatly enabled by the availability of large volumes of labeled data, a key challenge for deep learning in geoscience problems is the paucity of labeled samples that limit the effectiveness of conventional supervised deep learning frameworks. Hence, there is a need to develop novel frameworks in deep learning that can work with limited number of labeled samples while leveraging the abundance of unlabeled samples available in geoscience problems, e.g., through remote sensing data. As an example of a recent success in this direction, transfer learning approaches have been explored in [97], [98] for mapping poverty using high-resolution satellite imagery, by using nighttime light intensities as a data-rich proxy signal for training deep learning models, while simultaneously learning features that are useful for poverty prediction. Similar approaches can be explored in other geoscience problems where we have scarcity of gold-standard ground truth, but we may have access to proxy labels with some inaccuracies. Another major limitation of deep learning models that limit their usefulness in goescience problems is their inability to produce *explainable* theories from data that can lead to advancement of scientific knowledge. While improving predictive performance is an important and necessary consideration, a common end-goal in geoscience problems is to discover the physical pathways affecting the patterns or relationships in the data, that can be used as building blocks in scientific models of geoscience phenomena. Since conventional deep learning models are inherently "black-box" in nature and the prevailing practice in deep learning is to avoid expert-designed features and solely rely on the information contained in the data, their utility is currently limited to applications where we are mostly concerned only with improving prediction performance.

5.2 Theory-Guided Data Science

Theory-guided data science is an emerging paradigm of research that aims to combine scientific knowledge (or theory) with ML methods to advance knowledge discovery in a number of scientific disciplines including the geosciences [99]. While both physics-based models and ML methods show unique strengths in scientific problems, they also suffer from complementary weaknesses. For example, physics-based models, that are founded on core scientific principles, are unable to make effective use of data to infer unknown parameters or system states in the models, sometimes even suffering from unavoidable biases in the form of missing or incomplete physics (e.g., see recent debate papers in hydrology [100], [101], [102]). On the other hand, ML methods, that solely rely on the information contained in the data, can only be as good as the data they are fed with, and in scientific problems where we have paucity of labeled samples, it is common for black-box ML methods to discover spurious and physically inconsistent patterns. Theory-guided data science attempts to alleviate the challenges of physics-only and data-only methods by using physics and data at an equal footing.

It is important to adopt the paradigm of theory-guided data science as an overarching philosophy in all applications of ML methods in geoscience problems, so that we discover patterns and relationships that are not only generalizable but also consistent with our existing scientific knowledge. This is especially necessary given the limited size of labeled data in geoscience applications, where solely relying on the data while ignoring the underlying physics is not adequate for ensuring generalizability. There are several ways scientific knowledge can be brought together with ML methods, as discussed in a recent perspective article [99] using illustrative examples from diverse disciplines such as climate science, hydrology, fluid dynamics, material science, chemistry, neuroscience, and biomedicine. Indeed, the paradigm of theory-guided data science has already begun to show promise in a number of geoscience problems. For example, a label-free approach for detecting geoscience objects (namely ocean eddies) was presented in [33], where the physical properties of geoscience objects such as spatio-temporal consistency was used to provide supervision to the ML methods. In another line of work to create global maps of surface water dynamics using high-resolution remote sensing data [47], [103], a key physical constraint governing the dynamics of surface water bodies, i.e., locations at lower elevations in lakes and reservoirs get filled with water first before locations at higher elevations, was used to refine the outputs of ML methods. This was instrumental in overcoming the poor quality of remote sensing data due to clouds, aerosols, etc., that show spatially and temporally correlated noise structures that cannot be handled by traditional spatio-temporal smoothing techniques that assume salt-pepper noise.

In a recent work by Karpatne et al. [91], a novel framework combining physics-based models and deep learning methods, termed as physics-guided neural networks (PGNN), was proposed for modeling lake temperature. In this work, a joint-physics-ML model that used the output of physics-based models in ANN architectures was developed, which showed significant improvements in performance over pure physics-based models and black-box ANN models. Further, by using loss functions that measure the consistency of the model predictions w.r.t. physics-based equations, the proposed PGNN model was able to generate generalizable as well as physically meaningful results, even in paucity of labeled data. Similar lines of research can be explored in other geoscience applications that employ physics-based models and have some availablity of data. By pursuing the continuum between physics-based models and ML, the paradigm of theory-guided data science can unlock the full potential of ML methods in scientific applications of great societal relevance such as geoscience.

6 Conclusions

The Earth system is a place of great scientific interest that impacts every aspect of life on this planet and beyond. The survey of geoscience properties, applications, and promising machine learning directions provided in this article is clearly not exhaustive, but it illustrates the great emerging possibilities of future machine learning research in this important area.

Successful application of machine learning techniques in the geosciences is generally driven by a science question arising in the geosciences, and the best recipe for success tends to be for a machine learning researcher to collaborate very closely with a geoscientist during all phases of research. This is because the geoscientists are in a better position to understand which science question is novel and important, which variables and data set to use to answer that question, the strengths and weaknesses inherent in the data collection process that yielded the data set, and which pre-processing steps to apply, such as smoothing or removing seasonal cycles. Likewise, the machine learning researchers are better placed to decide which data analysis methods are available and appropriate for the data, the strengths and weaknesses of those methods, and what they can realistically achieve. Interpretability is also an important end goal in geosciences and hence, choosing methods that are inherently transparent are generally preferred in most geoscience applications. Further, the end results of a study need to be translated into geoscience language so that it can be related back to the original science questions. Hence, frequent communication between the researchers avoids long detours and ensures that the outcome of the analysis is indeed rewarding for both machine learning researchers and geoscientists [104].

Acknowledgments

The authors of this paper are supported by inter-disciplinary projects at the interface of machine learning and geoscience, including the NSF Expeditions in Computing grant on "Understanding Climate Change: A Data-driven Approach" (Award #1029711), the NSF-funded 2015 IS-GEO workshop (Award #1533930), and subsequent Research Collaboration Network (EarthCube RCN IS-GEO: Intelligent Systems Research to Support Geosciences, Award #1632211). The vision outlined in this paper has been greatly influenced by the collaborative works in these projects. In particular, the description of geoscience properties in Section 3 has been motivated by the initial discussions at the 2015 IS-GEO workshop (see workshop report [105]). Dr. Ravela's work was funded in part by an MIT Environmental Solutions Initiative seed fund award, the MIT MISTI program, and a Seaver Institute award.

References

[1] R. W. Kates, W. C. Clark, R. Corell, J. M. Hall, C. C. Jaeger, I. Lowe, J. J. McCarthy, H. J. Schellnhuber, B. Bolin, N. M. Dickson, et al., "Sustainability science," *Sci.*, vol. 292, no. 5517, pp. 641–642, 2001.

[2] F. Press, "Earth science and society," *Nature*, vol. 451, no. 7176, 2008, Art. no. 301.

[3] W. V. Reid, D. Chen, L. Goldfarb, H. Hackmann, Y. T. Lee, K. Mokhele, E. Ostrom, K. Raivio, J. Rockström, H. J. Schellnhuber et al., "Earth system science for global sustainability: Grand challenges," *Sci.*, vol. 330, no. 6006, pp. 916–917, 2010.

[4] Intergovernmental Panel on Climate Change, *Climate Change 2014–Impacts, Adaptation and Vulnerability: Regional Aspects*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[5] S. D. Peckham, "The CSDMS standard names: Cross-domain naming conventions for describing process models, data sets and their associated variables," in *Proc. Int. Environ. Modelling Softw. Soc.*, 2014, Art. no. 67.

[6] The World Economic Forum, "Harnessing artificial intelligence for the earth," Tech. Rep., Jan. 2018. [Online]. Available: http://www3.weforum.org/docs/Harnessing_Artificial_Intelligence_for_the _Earth_report_2018.pdf

[7] I. Ebert-Uphoff, et al., "Climate informatics," 2017. [Online]. Available: http://climateinformatics.org

[8] NSF Expeditions in Computing, "Understanding climate change: A data-driven approach," 2017. [Online]. Available: http://climatechange.cs.umn.edu/

[9] American Geophysical Union, "Earth & space sciences informatics," 2017. [Online]. Available: http://essi.agu.org/

[10] NSF-funded Research Collaboration Network, "Intelligent systems for geosciences," 2017. [Online]. Available: https://is-geo.org/

[11] J. H. Faghmous and V. Kumar, "A big data guide to understanding climate change: The case for theory-guided data science," *Big Data*, vol. 2, no. 3, pp. 155–163, 2014.

[12] C. Monteleoni, G. A. Schmidt, and S. McQuade, "Climate informatics: accelerating discovering in climate science with machine learning," *Comput. Sci. Eng.*, vol. 15, no. 5, pp. 32–40, 2013.

[13] J. H. Faghmous, V. Kumar, and S. Shekhar, "Computing and climate," *Comput. Sci. Eng.*, vol. 17, no. 6, pp. 6–8, 2015.

[14] G. Atluri, A. Karpatne, and V. Kumar, "Spatio-temporal data mining: A survey of problems and methods," *arXiv: 1711.04710*, 2017.

[15] A. Karpatne and S. Liess, "A guide to earth science data: Summary and research challenges," *Comput. Sci. Eng.*, vol. 17, no. 6, pp. 14–18, 2015.

[16] U.S. Geological Survey, "Land processes distributed active archive center," 2017. [Online]. Available: https://lpdaac.usgs.gov/

[17] NASA and USGS, "Landsat data archive," 2017. [Online]. Available: https://landsat.gsfc.nasa.gov/data/

[18] C. Frankenberg, A. K. Thorpe, D. R. Thompson, G. Hulley, E. A. Kort, N. Vance, J. Borchardt, T. Krings, K. Gerilowski, C. Sweeney, et al.,"Airborne methane remote measurements reveal heavy-tail flux distribution in four corners region," *Proc. Nat. Acad. Sci. United States America*, vol. 113, pp. 9734–9739, 2016.

[19] National Oceanic and Atmospheric Administration, "National centers for environmental information," 2017. [Online]. Available: https://www.ncdc.noaa.gov/

[20] World Meteorological Organisation, "Global runoff data centre," 2017. [Online]. Available: http://www.bafg.de/GRDC/

[21] National Science Foundation, "EarthScope," 2017. [Online]. Available: http://www.earthscope.org/

[22] World Climate Research Programme, "Coupled model intercomparison project," 2017. [Online]. Available: http://cmip-pcmdi.llnl.gov/

[23] National Corporation for Atmospheric Research (NCAR), "Community land model," 2017. [Online]. Available: http://www.cesm.ucar.edu/models/clm/

[24] H. J. Miller, "Tobler's first law and spatial analysis," *Ann. Assoc. Amer. Geographers*, vol. 94, no. 2, pp. 284–289, 2004.

[25] J. M. Wallace, J. M. Gutzler, and S. David, "Teleconnections in the geopotential height field during the northern hemisphere winter," *Monthly Weather Rev.*, vol. 109, no. 4, pp. 784–812, 1981.

[26] J. Kawale, S. Liess, A. Kumar, M. Steinbach, P. Snyder, V. Kumar, A. R. Ganguly, N. F. Samatova, and F. Semazzi, "A graph-based approach to find teleconnections in climate data," *Statistical Anal. Data Mining: The ASA Data Sci. J.*, vol. 6, no. 3, pp. 158–179, 2013.

[27] F. Siegert, G. Ruecker, A. Hinrichs, and A. A. Hoffmann, "Increased damage from fires in logged forests during droughts caused by EL niño," *Nature*, vol. 414, no. 6862, pp. 437–440, 2001.

[28] P. J. Ward, B. Jongman, M. Kummu, M. D. Dettinger, F. C. S. Weiland, and H. C. Winsemius, "Strong influence of EL niño southern oscillation on flood risk around the world," *Proc. Nat. Acad. Sci. United States America*, vol. 111, no. 44, pp. 15 659–15 664, 2014.

[29] H. Babaie and A. Davarpanah, "Ontology of earth's nonlinear dynamic complex systems," in *Proc. EGU General Assem. Conf. Abstracts*, 2017, Art. no. 11198.

[30] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, et al., "The NCEP/NCAR 40-year reanalysis project," *Bulletin Amer. Meteorological Soc.*, vol. 77, no. 3, pp. 437–471, 1996.

[31] D. B. Chelton, M. G. Schlax, R. M. Samelson, and R. A. de Szoeke, "Global observations of large oceanic eddies," *Geophysical Res. Lett.*, vol. 34, no. 15, 2007.

[32] J. H. Faghmous, M. Le, M. Uluyol, V. Kumar, and S. Chatterjee, "A parameter-free spatio-temporal pattern mining model to catalog global ocean dynamics," in *Proc. IEEE 13th Int. Conf. Data Mining*, 2013, pp. 151–160.

[33] J. H. Faghmous, H. Nguyen, M. Le, and V. Kumar, "Spatio-temporal consistency as a means to identify unlabeled objects in a continuous data field," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 410–416.

[34] J. H. Faghmous, I. Frenger, Y. Yao, R. Warmka, A. Lindell, and V. Kumar, "A daily global mesoscale ocean eddy dataset from satellite altimetry," *Nature Sci. Data*, vol. 2, 2015, Art. no. 150028.

[35] C. Tang and C. Monteleoni, "Can topic modeling shed light on climate extremes?" *Comput. Sci. Eng.*, vol. 17, no. 6, pp. 43–52, 2015.

[36] S. Ravela, "A statistical theory of inference for coherent structures," *Lecture Notes Comput. Sci.*, vol. 8964, pp. 121–133, 2015.

[37] S. Ravela, "Quantifying uncertainty for coherent structures," *Procedia Comput. Sci.*, vol. 9, pp. 1187–1196, 2012.

[38] J. Baxter, et al., "A model of inductive bias learning," *J. Artif. Intell. Res.*, vol. 12, no. 1, pp. 149–198, 2000.

[39] R. Caruana, "Multitask learning: A knowledge-based source of inductive bias," in *Proc. 10th Int. Conf. Mach. Learn.*, 1993, pp. 41–48.

[40] A. Karpatne, A. Khandelwal, S. Boriah, S. Chatterjee, and V. Kumar, "Predictive learning in the presence of heterogeneity and limited training data," in *Proc. SIAM Int. Conf. Data Mining*, 2014, pp. 253–261.

[41] C. Monteleoni, G. A. Schmidt, S. Saroha, and E. Asplund, "Tracking climate models," *Statistical Anal. Data Mining: ASA Data Sci. J.*, vol. 4, no. 4, pp. 372–392, 2011.

[42] S. McQuade and C. Monteleoni, "Global climate model tracking using geospatial neighborhoods," in *Proc. 26th AAAI Conf. Artif. Intell.*, 2012.

[43] D. Das, J. Dy, J. Ross, Z. Obradovic, and A. Ganguly, "Nonparametric Bayesian mixture of sparse regressions with application towards feature selection for statistical downscaling," *Nonlinear Processes Geophysics*, vol. 21, no. 6, pp. 1145–1157, 2014.

[44] A. Karpatne and V. Kumar, "Adaptive heterogeneous ensemble learning using the context of test instances," in *Proc. IEEE Int. Conf. Data Mining*, 2015, pp. 787–792.

[45] A. Karpatne, A. Khandelwal, and V. Kumar, "Ensemble learning methods for binary classification with multi-modality within the classes," in *Proc. SIAM Int. Conf. Data Mining*, 2015, pp. 730–738.

[46] A. Khandelwal, A. Karpatne, M. E. Marlier, J. Kim, D. P. Lettenmaier, and V. Kumar, "An approach for global monitoring of surface water extent variations in reservoirs using MODIS data," *Remote Sens. Environ.*, vol. 202, pp. 113–128, 2017.

[47] University of Minnesota, "Global surface water monitoring system," 2017. [Online]. Available: http://z.umn.edu/watermonitor/

[48] X. Zhu, "Semi-supervised learning literature survey," Univ. Wisconsin, Madison, Wisconsin, Tech. Rep. TR 1530, Dec. 2007.

[49] B. Settles, "Active learning literature survey," Univ. Wisconsin, Madison, Wisconsin, Tech. Rep. TR 1648, 2010.

[50] R. R. Vatsavai, S. Shekhar, and T. E. Burk, "A semi-supervised learning method for remote sensing data mining," in *Proc. 17th IEEE Int. Conf. Tools Artif. Intell.*, 2005, pp. 5 pp.–211

[51] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2218–2232, Jul. 2009.

[52] V. Mithal, G. Nayak, A. Khandelwal, V. Kumar, N. C. Oza, and R. Nemani, "RAPT: Rare class prediction in absence of true labels," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 11, pp. 2484–2497, Nov. 2017.

[53] A. Karpatne, Z. Jiang, R. R. Vatsavai, S. Shekhar, and V. Kumar, "Monitoring land-cover changes," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 8–21, Jun. 2016.

[54] E. S. Gardner, "Exponential smoothing: The state of the art–part II," *Int. J. Forecasting*, vol. 22, no. 4, pp. 637–666, 2006.

[55] G. E. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*. San Francisco, CA, USA: Holden-Day, 1976.

[56] M. Aoki, *State Space Modeling of Time Series*. Berlin, Germany: Springer, 2013.

[57] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *IEEE ASSP Mag.*, vol. 3, no. 1, pp. 4–16, Jan. 1986.

[58] A. C. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[59] Y. Liu, T. Bahadori, and H. Li, "Sparse-GEV: Sparse latent space model for multivariate extreme value time serie modeling," *arXiv:1206.4685*, 2012.

[60] A. McGovern, D. J. Gagne, J. Basara, T. M. Hamill, and D. Margolin, "Solar energy prediction: An international contest to initiate interdisciplinary research on compelling meteorological problems," *Bulletin Amer. Meteorological Soc.*, vol. 96, no. 8, pp. 1388–1395, 2015.

[61] D. J. Gagne II, A. McGovern, J. Brotzge, M. Coniglio, J. Correia Jr, and M. Xue, "Day-ahead hail prediction integrating machine learning with storm-scale numerical weather models," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 3954–3960.

[62] D. J. Gagne, A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, "Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles," *Weather Forecasting*, vol. 32, no. 5, pp. 1819–1840, 2017.

[63] Y. Gel, A. E. Raftery, and T. Gneiting, "Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation method," *J. Amer. Statistical Assoc.*, vol. 99, no. 467, pp. 575–583, 2004.

[64] Y. R. Gel, V. Lyubchich, and S. E. Ahmed, "Catching uncertainty of wind: A blend of sieve bootstrap and regime switching models for probabilistic short-term forecasting of wind speed," in *Advances in Time Series Methods and Applications*. New York, NY, USA: Springer, 2016, pp. 279–293.

[65] Y. Zhuang, K. Yu, D. Wang, and W. Ding, "An evaluation of big data analytics in feature selection for long-lead extreme floods forecasting," in *Proc. IEEE 13th Int. Conf. Netw. Sens. Control*, 2016, pp. 1–6.

[66] D. Wang and W. Ding, "A hierarchical pattern learning framework for forecasting extreme weather events," in *Proc. IEEE Int. Conf. Data Mining*, 2015, pp. 1021–1026.

[67] K. Yu, D. Wang, W. Ding, J. Pei, D. L. Small, S. Islam, and X. Wu, "Tornado forecasting with multiple Markov boundaries," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 2237–2246.

[68] J. Kawale, M. Steinbach, and V. Kumar, "Discovering dynamic dipoles in climate data," in *Proc. SIAM Int. Conf. Data Mining*, 2011, pp. 107–118.

[69] M. Steinbach, P.-N. Tan, V. Kumar, S. Klooster, and C. Potter, "Discovery of climate indices using clustering," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2003, pp. 446–455.

[70] A. A. Tsonis and P. J. Roebber, "The architecture of the climate network," *Physica A: Statistical Mech. Appl.*, vol. 333, pp. 497–504, 2004.

[71] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, "Complex networks in climate dynamics," *Eur. Phys. J.-Special Topics*, vol. 174, no. 1, pp. 157–179, 2009.

[72] J. Elsner, T. Jagger, and E. Fogarty, "Visibility network of united states hurricanes," *Geophysical Res. Lett.*, vol. 36, no. 16, 2009.

[73] K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly, "An exploration of climate data using complex networks," *ACM SIGKDD Explorations Newslett.*, vol. 12, no. 1, pp. 25–32, 2010.

[74] K. Steinhaeuser, N. Chawla, and A. Ganguly, "Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science," *Statistical Anal. Data Mining: The ASA Data Sci. J.*, vol. 4, no. 5, pp. 497–511, 2011.

[75] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: J. Econometric Soc.*, vol. 37, pp. 424–438, 1969.

[76] J. Pearl and T. Verma, "A theory of inferred causation," in *Proc. 2nd Int. Conf. Principles Knowl. Representation Reasoning*, Apr. 1991, pp. 441–452.

[77] I. Ebert-Uphoff and Y. Deng, "Causal discovery from spatio-temporal data with applications to climate science," in *Proc. 13th Int. Conf. Mach. Learn. Appl.*, 2014, pp. 606–613.

[78] S. Agrawal, G. Atluri, A. Karpatne, W. Haltom, S. Liess, S. Chatterjee, and V. Kumar, "Tripoles: A new class of relationships in time series data," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 697–706.

[79] S. Liess, S. Agrawal, S. Chatterjee, and V. Kumar, "A teleconnection between the west siberian plain and the ENSO region," *J. Climate*, vol. 30, no. 1, pp. 301–315, 2017.

[80] M. Lu, U. Lall, J. Kawale, S. Liess, and V. Kumar, "Exploring the predictability of 30-day extreme precipitation occurrence using a global SST–SLP correlation network," *J. Climate*, vol. 29, no. 3, pp. 1013–1029, 2016.

[81] S. Liess, A. Kumar, P. K. Snyder, J. Kawale, K. Steinhaeuser, F. H. Semazzi, A. R. Ganguly, N. F. Samatova, and V. Kumar, "Different modes of variability over the tasman sea: Implications for regional climate," *J. Climate*, vol. 27, no. 22, pp. 8466–8486, 2014.

[82] S. Chatterjee, K. Steinhaeuser, A. Banerjee, S. Chatterjee, and A. Ganguly, "Sparse group lasso: Consistency and climate applications," in *Proc. SIAM Int. Conf. Data Mining*, 2012, pp. 47–58.

[83] M. C. McGraw and E. A. Barnes, "Memory matters: A case for Granger causality in climate variability studies," *J. Climate*, 2018.

[84] A. C. Lozano, H. Li, A. Niculescu-Mizil, Y. Liu, C. Perlich, J. Hosking, and N. Abe, "Spatial-temporal causal modeling for climate change attribution," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 587–596.

[85] I. Ebert-Uphoff and Y. Deng, "Causal discovery for climate research using graphical models," *J. Climate*, vol. 25, no. 17, pp. 5648–5665, 2012.

[86] A. Hannart, J. Pearl, F. E. L. Otto, P. Naveau, and M. Ghil, "Causal counterfactual theory for the attribution of weather and climate-related events," *Bulletin Amer. Meteorological Soc.*, vol. 97, pp. 99–110, 2016.

[87] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[88] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[89] Y. Liu, E. Racah, J. Correa, A. Khosrowshahi, D. Lavers, K. Kunkel, M. Wehner, W. Collins et al., "Application of deep convolutional neural networks for detecting extreme weather in climate datasets," *arXiv:1605.01156*, 2016.

[90] E. Racah, C. Beckham, T. Maharaj, C. Pal, et al., "Semi-supervised detection of extreme weather events in large climate datasets," *arXiv:1612.02095*, 2016.

[91] A. Karpatne, W. Watkins, J. Read, and V. Kumar, "Physics-guided neural networks (PGNN): An application in lake temperature modeling," *arXiv: 1710.11431*, 2017.

[92] T. Vandal, E. Kodra, S. Ganguly, A. Michaelis, R. Nemani, and A. R. Ganguly, "DeepSD: Generating high resolution climate change projections through single image super-resolution," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 1663–1672.

[93] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. Nemani, "DeepSat: A learning framework for satellite imagery," in *Proc. 23rd SIGSPATIAL Int. Conf. Advances Geographic Inf. Syst.*, 2015, Art. no. 37.

[94] X. Jia, A. Khandelwal, G. Nayak, J. Gerber, K. Carlson, P. West, and V. Kumar, "Incremental dual-memory LSTM in land cover prediction," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 867–876.

[95] X. Jia, A. Khandelwal, G. Nayak, J. Gerber, K. Carlson, P. West, and V. Kumar, "Predict land covers with transition modeling and incremental learning," in *Proc. SIAM Int. Conf. Data Mining*, 2017, pp. 171–179.

[96] X. Jia, A. Khandelwal, J. Gerber, K. Carlson, P. West, and V. Kumar, "Learning large-scale plantation mapping from imperfect annotators," in *Proc. IEEE Int. Conf. Big Data*, 2016, pp. 1192–1201.

[97] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon, "Transfer learning from deep features for remote sensing and poverty mapping," *arXiv:1510.00098*, 2015.

[98] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Sci.*, vol. 353, no. 6301, pp. 790–794, 2016.

[99] A. Karpatne, G. Atluri, J. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar, "Theory-guided data science: A new paradigm for scientific discovery," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 10, pp. 2318–2331, Oct. 2017.

[100] H. V. Gupta and G. S. Nearing, "Debates–the future of hydrological sciences: A (common) path forward? using models and data to learn: A systems theoretic perspective on the future of hydrological science," *Water Resources Res.*, vol. 50, no. 6, pp. 5351–5359, 2014.

[101] U. Lall, "Debates–the future of hydrological sciences: A (common) path forward? one water. one world. many climes. many souls," *Water Resources Res.*, vol. 50, no. 6, pp. 5335–5341, 2014.

[102] J. J. McDonnell and K. Beven, "Debates–the future of hydrological sciences: A (common) path forward? a call to action aimed at understanding velocities, celerities and residence time distributions of the headwater hydrograph," *Water Resources Res.*, vol. 50, no. 6, pp. 5342–5350, 2014.

[103] A. Khandelwal, V. Mithal, and V. Kumar, "Post classification label refinement using implicit ordering constraint among data instances," in *Proc. IEEE Int. Conf. Data Mining*, 2015, pp. 799–804.

[104] I. Ebert-Uphoff and Y. Deng, "Three steps to successful collaboration with data scientists," *EOS Trans. Amer. Geophysical Union*, 2017.

[105] Y. Gil et al., "Final workshop report," *Workshop Intell. Inf. Syst. Geosci.*, 2015. [Online]. Available: https://isgeodotorg.files.wordpress.com/2015/12/

Anuj Karpatne received the BTech-MTech degrees in mathematics & computing from the Indian Institute of Technology (IIT) Delhi. He is working toward the PhD degree in the Department of Computer Science and Engineering (CSE), University of Minnesota (UMN). He works in the area of data mining with applications in scientific problems related to the environment.

Imme Ebert-Uphoff received the BS and MS degrees in mathematics from the University of Karlsruhe, Karlsruhe, Germany, and the PhD degree in mechanical engineering from Johns Hopkins University, Baltimore, MD. She is a research faculty member with the Department of Electrical and Computer Engineering, Colorado State University. Her research interests include causal discovery and other machine learning methods, applied to geoscience applications.

Sai Ravela received the PhD degree in computer science from the University of Massachusetts at Amherst, in 2003. He directs the Earth Signals and Systems Group (ESSG) with the Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology. His primary research interests are in dynamic data-driven stochastic systems theory and machine intelligence methodology with application to earth, atmospheric, and planetary Sciences.

Hassan Ali Babaie received the PhD degree in structural geology from Northwestern University. He is an associate professor with the Department of Geosciences, with a joint appointment in the Computer Science Department, Georgia State University. His research interest includes geoinformatics, Semantic Web, representing the knowledge of structural geology, applying ontologies, and machine learning.

Vipin Kumar received the BE degree in electronics & communication engineering from IIT Roorkee, the ME degree in electrical engineering from Philips International Institute Eindhoven, and the PhD degree in computer science from the University of Maryland. He is a regents professor and William Norris chair in Large Scale Computing with the Department of CSE, UMN. His research interests include data mining, high-performance computing, and their applications in climate/ecosystems and biomedical domains.