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, problems in geosciences have several unique challenges that are seldom found in traditional applications, requiring novel problem formulations and methodologies 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 properties that make it challenging to use traditional machine learning techniques. We then describe some of the common categories of geoscience problems where machine learning can play a role, and discuss some of the existing efforts and promising directions for methodological development in machine learning. We conclude by discussing some of the emerging research themes in machine learning that are applicable across all problems in the geosciences, 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. 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.

Given the variety of disciplines participating in geoscience research and the diverse nature of questions being investigated, the analysis of geoscience data has several unique aspects that are strikingly different from standard data science problems encountered in commercial domains. For example, geoscience phenomena are governed by physical laws and principles and involve objects and relationships that often have amorphous boundaries and complex latent variables. Challenges introduced by these characteristics motivate the development of new problem formulations and methodologies in machine learning that may be broadly applicable to problems even outside the scope of geosciences.

Thus, there is a great opportunity for machine learning researchers to closely collaborate with geoscientists and cross-fertilize ideas across disciplines for advancing the frontiers of machine learning as well as geosciences. There are several communities working on this emerging field of inter-disciplinary collaboration at the intersection of geosciences and machine learning. These include, but are not limited to, Climate Informatics: a community of researchers conducting annual workshops to bridge problems in climate science with methods from statistics, machine learning, and data mining [5]; Climate Change Expeditions: a multi-institution multi-disciplinary collaboration funded by the National Science Foundation (NSF) Expeditions in Computing grant on “Understanding Climate Change: A
data-driven Approach" [6]; and ESSI: a focus group of the American Geophysical Union (AGU) on Earth & Space Sciences Informatics [7]. More recently, NSF has funded a research coordination network on Intelligent Systems for Geosciences (IS-GEO) [8], with the intent of forging stronger connections between the two communities. Furthermore, a number of leading conferences in machine learning and data mining such as Knowledge Discovery and Data Mining (KDD), IEEE International Conference on Data Mining (ICDM), SIAM International Conference on Data Mining (SDM), and Neural Information Processing Systems (NIPS) have included workshops or tutorials on topics related to geosciences. The role of big data in geosciences has also been recognized in recent perspective articles (e.g., [9], [10]) and special issues of journals and magazines (e.g., [11]).

The purpose of this article is to introduce researchers in the machine learning (ML) community to the opportunities and challenges offered by geoscience problems. 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 the challenges for machine learning arising from both the underlying geoscience processes and their data collection. Section 4 outlines important geoscience problems where machine learning can yield major advances. Section 5 discusses two cross-cutting themes of research in machine learning that are generally applicable across all areas of geoscience. Section 6 provides concluding remarks by briefly discussing the best practices for collaboration between machine learning researchers and geoscientists.

2 Sources of Geoscience Data

The Earth and its major interacting components (e.g., lithosphere, biosphere, hydrosphere, and atmosphere) are complex dynamic systems in which the states of the system perpetually keep changing in space and time, in order to create a balance of mass and energy. The elements of the Earth system (e.g., layers in oceans, ions in air, minerals and grains in rock, and land covers on the ground) interact with each other through complex and dynamic geoscience processes (e.g., rain falling on Earth’s surface and nourishing the biomass, sediments depositing on river banks and changing river course, and magma erupting on sea floor and forming islands).

Data about these Earth system components and geoscience processes can generally be obtained from two broad categories of data sources: (a) observational data collected via sensors in space, in the sea, or on the land, and (b) simulation data from physics-based models of the Earth system. We briefly describe both these categories of geoscience data sources in the following. A detailed review of Earth science data sets and their properties can be found in [14].

2.1 Geoscience Observations

Information about the Earth system is collected via different acquisition methods at varying scales of space and time and for a variety of geoscience objectives. For example, there is a nexus of Earth observing satellites in space that are tasked to monitor a number of geoscience variables such as surface temperature, humidity, optical reflectance, and chemical compositions of the atmosphere. There is a growing body of space research organizations ranging from public agencies such as the National Aeronautics and Space Administration (NASA), European Space Agency (ESA), and Japan Aerospace Exploration Agency (JAXA) to private companies such as SpaceX that are together contributing to the huge volume and variety of remote sensing data about our Earth, many of which are publicly available (e.g., see [15]). Remote sensing data provides a global picture of the history of geoscience variables at fine spatial scales (1km to 10m, and less) and at regular time intervals (monthly to daily) for long periods, sometimes starting from the 1970s (e.g., Landsat archives [16]). For targeted studies in specific geographic regions of interest, geoscience observations can also be collected using 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 [17].

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-based observations of geoscience processes 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. They 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) [18]. Sensor-based measurements from rain and river gauges are also central for understanding hydrological processes such as surface water discharge [19]. Land-based seismic sensors, Global Positioning System (GPS)-enabled devices, and other geophysical instruments also continuously measure the Earth’s geological structure and processes [20]. 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.

Given the huge variety in the characteristics of data for different geoscience processes, it is important to identify the type and properties of a given geoscience data set to make utmost use of relevant data analytics methodologies. For example, remote sensing data sets, that are commonly available as rasters over regularly-spaced grid cells in space and time, can be represented as geo-registered images over individual time points or as time series data at individual spatial locations. On the other hand, sensor measurements from ships and ocean buoys can be represented as point reference data (also termed as geostatistical data in the spatial statistics literature) of continuous spatio-temporal fields. Indeed, 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 [21].

2.2 Earth System Model Simulations

A unique aspect of geoscience processes is that the relationships among variables or the evolution of states of the
system are deeply grounded in physical laws and principles, discovered by the scientific community over multiple centuries of systematic research. 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. Although such physics-based equations can sometimes be solved in closed form for small-scale experiments, most often it is difficult to obtain their exact solutions for complex real-world systems encountered in the geosciences. However, the underlying physical principles can still be used to simulate the evolution of the states of the Earth system using numerical models referred to as physics-based models. Such models are the standard workhorse for studying a majority of geoscience processes where the state of the dynamical system can be time-stepped back in the past or forward in the future using inputs such as initial and boundary conditions or values of internal parameters in physical equations. Physics-based models generate large volumes of simulation data of different components of the Earth system, which can be used in data-driven analyses. 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) \cite{22}. Simulations of terrestrial processes related to the lithosphere and biosphere are produced by the Community Land Model (CLM) \cite{23}, developed by a number of international agencies collaborating with the National Center for Atmospheric Research (NCAR).

3 Geoscience Challenges

There are several characteristics of geoscience applications that limit the usefulness of traditional machine learning algorithms for knowledge discovery. Firstly, there are some inherent challenges arising from the nature of geoscience processes. For example, geoscience objects 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. Geoscience phenomena also have spatio-temporal structure, are highly multi-variate, follow non-linear relationships (e.g., chaotic), show non-stationary characteristics, and often involve rare but interesting events. Secondly, apart from the inherent challenges of geoscience processes, the procedures used for collecting geoscience observations introduce more challenges for machine learning. This includes the presence of data at multiple resolutions of space and time, with varying degrees of noise, incompleteness, and uncertainties. Thirdly, for supervised learning approaches, there are additional challenges due to the small sample size (e.g., small number of historical years with adequate records) and lack of gold-standard ground truth in geoscience applications. In the following, we describe these three categories of geoscience challenges, namely (a) inherent challenges of geoscience processes, (b) geoscience data collection challenges, and (c) paucity of samples and ground truth, in detail.

3.1 Inherent Challenges of Geoscience Processes

Property 1: Objects with Amorphous Boundaries

Geoscience objects include waves, flows, and coherent structures in all phases of matter. Hence, the form, structure, and patterns of geoscience objects that can exist at multiple scales in continuous spatio-temporal fields are much more complex than those found in discrete spaces that machine learning algorithms typically deal with, such as items in market basket data. For example, eddies, storms and hurricanes dynamically deform in complicated ways from a purely object-oriented perspective. New techniques to consider both the pattern and dynamical information of coherent objects and their uncertainties are being developed \cite{24}, \cite{25}, but new methods for capturing other features of geoscience objects, e.g., fluid segmentation and fluid feature characterization, are needed.

Property 2: Spatiotemporal Structure

Since almost every geoscience phenomena occurs in the realm of space and time, geoscience observations are generally auto-correlated in both space and time when observed at appropriate spatial and temporal 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. 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. Furthermore, if the land cover at a certain location changes (e.g., from forest to croplands), the change generally persists for some temporal duration instead of switching back and forth.

Although spatio-temporal autocorrelation dictates stronger connectivity among nearby observations in space and time, geoscience processes can also show long-range spatial dependencies. For example, a commonly studied phenomenon in climate science is teleconnections \cite{26}, \cite{27}, where two distant regions in the world show strongly coupled activity in climate variables such as temperature or pressure. Geoscience processes can 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 \cite{28}, \cite{29}.

The inherent spatio-temporal structure of geoscience data has several implications on machine learning methods. This is because many of the widely used machine learning methods are founded on the assumption that observed variables are independent and identically distributed (i.i.d). However, this assumption is routinely violated in geoscience problems, where variables are structurally related to each other in the context of space and time, unless there is a discontinuity, such as a fault, across which autocorrelation ceases to persist. Cognition of the spatio-temporal autocorrelation in geoscience data collected in continuous media is crucial for the effective modeling of geophysical phenomena.

Property 3: High Dimensionality

The Earth system is incredibly complex, with a huge number of potential variables, which may all impact each other, and thus many of which may have to be considered si-
multaneously \[30\]. For example, the robust and complete
detection of land cover changes, such as forest fires, requires
the analysis of multiple remote sensing variables, such as
vegetation indices and thermal anomaly signals. Capturing
the effects of these multiple variables at fine resolutions
of space and time renders geoscience data inherently high
dimensional, where the number of dimensions can easily
reach orders of millions.

As an example, in order to study processes occurring
on the Earth’s surface, even a relatively coarse resolution
data set (e.g. at 2.5° spatial resolution) may easily result in
more than 10,000 spatial grid points, where every grid point
has multiple observations in time. 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 atmosphere or the mantle, additionally increasing the
dimensionality of the data in 3D spatial resolutions. Hence,
there is a need to scale existing machine learning methods
to handle tens of thousands, or millions, of dimensions for
global analysis of geoscience phenomena.

**Property 4: Heterogeneity in Space and Time**

An interesting characteristic of geoscience processes is their
degree of variability in space and time, leading to a rich
heterogeneity in geoscience data across space and time.
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 geo-
science 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 phe-
nomena, that impact all local processes. This heterogeneity
of geoscience processes makes it difficult to study the joint
distribution of geoscience variables across all points in space
and time. Hence, it is difficult to train machine learning
models that have good performance across all regions in
space and across all time-steps. Instead, there is a need
to build local or regional models, each corresponding to a
homogeneous group of observations.

**Property 5: Interest in Rare Phenomena**

In a number of geoscience problems, we are interested in
studying objects, processes, and events that occur infre-
cently 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 prop-
erty, 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 \[31\]. As another exam-
ple, 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. Identifying
such rare classes of changes and events from geoscience
data and characterizing their behavior is challenging. This
is because we often have an inadequate number of data
samples from the rare class due to the skew (imbalance)
between the classes, making their modeling and characteri-
Zation difficult.

### 3.2 Geoscience Data Collection Challenges

**Property 6: Multi-resolution Data**

Geoscience data sets are often available via different sources
(e.g., satellite sensors, in-situ measurements and model-
based simulations) and at varying spatial and temporal
resolutions. These data sets may exhibit varying charac-
teristics, 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,
may require using aerial imagers from planes flying over
the region of interest, which may need to be combined with
coarser resolution satellite imagers 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. To handle multi-resolution data, a common ap-
proach is to build a bridge between data sets at disparate
scales (e.g., using interpolation techniques), so that they
can be represented at the same resolution. We also need to
develop algorithms that can identify patterns at multiple
resolutions without upsampling all the data sets to the
highest resolution.

**Property 7: Noise, Incompleteness, and Uncertainty in 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, result-
ing 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 inter-
pretation of sensor values over time, making it difficult to
deploy a consistent methodology of analysis across different
time periods. In many geoscience applications, the signal
of interest can be small in magnitude compared to the mag-
nitude of noise. 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.

Many geoscience variables cannot even be measured
directly, but can only be inferred from other observations
or model simulations. For example, one can use airborne
imaging spectrometers to detect sources of methane (e.g.,
pipeline leaks), an important greenhouse gas. These instru-
ments fly overhead surveys and map the ground-reflected
sunlight arriving at the sensor. Methane plumes can then
be identified from excess sunlight absorption \[32\]. But to
determine the leak rate (flux) and the resulting greenhouse
gas impact, one must also know how fast the excess methane
mass is dispersing. This requires considering the influence
of air transport, which in turn requires steady-state physical assumptions, morphology-based plume modeling, or direct in-situ measurements of the wind speed. 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.

3.3 Paucity of Samples and Ground Truth

Property 8: Small sample size
The number of samples in geoscience data sets is often limited in both space and time. Factors that limit sample size include history of data collection and the nature of phenomenon being measured. For example, most satellite products are only available since the 1970s, and when monthly (yearly) processes are considered, this means that there are less than 600 (50) samples are available. Furthermore, there are many events in geosciences that are important to monitor but occur very infrequently, thus resulting in small sample sizes. For example, a majority of land cover changes, landslide, tsunami, and forest fire are rare events, and only occur for short temporal durations mostly over small spatial regions. With less than 80 years of reliable sensor-based data, only a few dozens of rare events are available as training data.

The limited 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.

This is in contrast to commercial applications involving Internet-scale data, e.g., text mining or object recognition, where large volumes of labeled or unlabeled data have been one of the major factors behind the success of machine learning methodologies such as deep learning. The limited number of samples in geoscience applications along with the large number of physical variables result in problems that are under-constrained in nature, requiring novel machine learning advances for their robust analyses.

Property 9: Paucity of Ground Truth
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, a common feature of geoscience problems is the paucity of labeled samples with 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 complexity of the system, the exact state of the system is never fully known.

The paucity of representative training samples can result in poor performance for many machine learning methods, either due to underfitting where the model is too simple, or due to overfitting where the model is overly complex relative to the dimensionality of features and the limited number of training samples. Hence, there is a need to develop machine learning methods that can learn parsimonious models even in the paucity of labeled data. Another possibility is to construct synthetic data sets through simulations or perturbation that can be used for training, to make the most of the few observations.

4 Geoscience Problems and ML Directions

Geoscientists constantly strive to develop better approaches for modeling the current state of the Earth system (e.g., how much methane is escaping into the atmosphere right now, which parts of Earth are covered by what kind of biomass) and its evolution, as well as the connections within and between all of its subsystems (e.g., how does a warming ocean affects specific ecosystems). This is aimed at advancing our scientific understanding of geoscience processes. This can also help in providing actionable information (e.g., extreme weather warnings) or informing policy decisions that directly impact our society (e.g., adapting to climate change and progressing towards sustainable lifestyles). The boundaries between these goals often blur in practice, e.g., an improved tornado model may simultaneously lead to a better science model as well as a more effective warning system.

Viewing from the lens of geosciences, many methods from machine learning are a natural fit for the problems encountered in geoscience applications. For example, classification and pattern recognition methods are useful for characterizing objects such as extreme weather events or swarms of foreshocks or aftershocks (tremors preceding or following an earthquake), estimating geoscience variables, and producing long-term forecasts of the state of the Earth system. As another example, approaches for mining relationships and causal attribution can provide insights into the inner workings of the Earth system and support policy making. In the following, we briefly describe five broad categories of geoscience problems, and discuss promising machine learning directions and examples of some recent successes that are relevant for each problem.

4.1 Characterizing Objects and Events
Machine learning algorithms can help in characterizing objects and events in geosciences that are critical for understanding the Earth system. For example, we can analyze patterns in geoscience data sets to detect climate events such as cyclogenesis and tornadogenesis, and discover their precursors for predicting them with long leads in time. Analyzing spatial and temporal patterns in geoscience data can also help in studying the formation and movement of climate objects such as weather fronts, atmospheric rivers, and ocean eddies, which are major drivers of vital geoscience processes such as 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), machine learning algorithms can enable their automated detection from data with improved performance using pattern mining techniques. However, in the presence of spatio-temporal
objects with amorphous boundaries and their associated uncertainties, there is a need to develop pattern mining approaches that can account for the spatial and temporal properties of geoscience data while characterizing objects and events. One such approach has been successfully used for finding spatio-temporal patterns in sea surface height data, resulting in the creation of a global catalogue of mesoscale ocean eddies. Another approach for finding anomalous objects buried under the surface of the Earth (e.g., land mines) from radar images was explored in using unsupervised techniques that can work with mediums of varying properties. The use of topic models has also been explored for finding extreme events from climate time series data.

4.2 Estimating Geoscience Variables from Observations

There is a great opportunity for machine learning methods to infer critical geoscience variables that are difficult to monitor directly, e.g., methane concentrations in air or groundwater seepage in soil, using information about other variables collected via satellites and ground-based sensors, or simulated using Earth system models. For example, supervised machine learning algorithms can be used to analyze remote sensing data and produce estimates of ecosystem variables such as forest cover, health of vegetation, water quality, and surface water availability, at fine spatial scales and at regular intervals of time. Such estimates of geoscience variables can help in informing management decisions and enabling scientific studies of changes occurring on the Earth’s surface.

A major challenge in the use of supervised learning approaches for estimating geoscience variables is the heterogeneity in the characteristics of variables across space and time. One way to address this challenge of heterogeneity is to explore multi-task learning frameworks, 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 to regularize their learning and avoid the problem of overfitting, especially when some tasks suffer from paucity of training samples. An example of a multi-task learning based approach for handling heterogeneity can be found in a recent work in 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. Figure shows the improvement in prediction performance of forest cover in Brazil using a multi-task learning approach. A detailed review of promising machine learning advances such as multi-task learning, multi-view learning, and multi-instance learning for addressing the challenges in supervised monitoring of land cover changes from remote sensing data is presented in.

To address the 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. 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 where a Bayesian mixture of models was learned for downscaling climate variables, where a different model was learned for every homogeneous cluster of locations in space.

In a recent work, adaptive ensemble learning methods, in conjunction with physics-based label refinement techniques, have been developed to address the challenge of heterogeneity and poor data quality for mapping the dynamics of surface water bodies using remote sensing data. This has enabled the creation of a global surface water monitoring system (publicly available at) 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.

Another challenge in the supervised estimation of geoscience variables is the small sample size and paucity of ground-truth labels. Methods for handling the problem of high dimensions and small sample sizes have been explored in, where sparsity-inducing regularizers such as sparse group Lasso were developed to model the domain characteristics of climate variables. To address the paucity of labels, novel learning frameworks such as semi-supervised learning, that leverages the structure in the unlabeled data for improving classification performance, and active learning, where an expert annotator is actively involved in the process of model building, have huge potential for improving the state-of-the-art in estimation problems encountered in geoscience applications. 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 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.

In addition to supervised learning approaches, given the plentiful availability of unlabeled data in geoscience applications such as remote sensing, there are several opportunities for unsupervised learning methods in estimating geoscience variables. For example, changes in time series of vegetation data, collected by satellite instruments on fixed time intervals at every spatial location on the Earth’s surface, have been extensively studied using unsupervised learning approaches for mapping land cover changes such as deforestation, insect damage, farm conversions, and forest fires.

4.3 Long-term Forecasting of Geoscience Variables

Predicting long-term trends of the state of the Earth system, e.g., forecasting geoscience variables ahead in time, can help in modeling future scenarios and devising early resource planning and adaptation policies. One 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, this 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 for time-series forecasting include exponential smoothing techniques [62], autoregressive integrated moving average (ARIMA) models [63], state-space models [64], and probabilistic models such as hidden Markov models and Kalman filters [65], [66]. Machine learning methods for forecasting climate variables using the spatial and temporal structure of geoscience data have been explored in recent works such as [67], [68], [69], [70].

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 due to the high-dimensional and non-stationary nature of geoscience processes [71], [72]. In the climate scenario, there is limited long-term predictability at fine spatial scales necessary for implementing policy decisions. Some advances have been made in downscaling future projections to high spatial resolutions, e.g., using physics-based Markov Chain and Random Field models [73], but much remains to be done. Further, the data is sparse, and the uncertainty distributions remain poorly sampled [74], [75]. The heavy-tailed nature of extreme events such as cyclones and floods further exacerbates the challenges in producing their long-term forecasts. In a recent work [67], 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 [76], [77], [78]. Effective prediction of geoscience variables can benefit from recent advances in machine learning such as transfer learning [79], where the model trained on a present task (with sufficient number of training samples) is used to improve the prediction performance on a future task with limited number of training samples.

4.4 Mining Relationships in Geoscience Data
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 (ENSO)—and their impact on several terrestrial events such as floods, droughts, and forest fires [28], [29]. Identifying such relationships from geoscience data can help us capture vital signs 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 of the widely-studied category of teleconnections is dipoles [27], [80], which are pairs of regions with strong negative correlations (e.g., the ENSO phenomena). There is a huge potential in discovering such relationships using data-driven approaches, that can sift through vast volumes of observational and model-based geoscience data and discover interesting patterns corresponding to geoscience relationships.

One of the first attempts in discovering relationships from climate data is a seminal work by Steinbach et al. [61]. 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 [82]. This includes approaches for examining the structure of the climate system [83], studying hurricane activity [84], and finding communities in climate networks [85], [86].

Formidable challenges arise in the problem of relationship mining due to the enormous search space of candidate relationships, and the need to simultaneously extract spatio-temporal objects, their relationships, and their dynamics, from noisy and incomplete geoscience data. Hence, there is a need for novel approaches that can directly discover the relationships as well as the interacting objects [27], [87]. For example, recent work on the development of such approaches have led to the discovery of previously unknown climate phenomena [88], [89], [90].

4.5 Causal Discovery and Causal Attribution
Discovering cause-effect relationships is an important task in the geosciences, closely related to the task of learning relationships in geoscience data, discussed in Subsection 4.4. The two primary frameworks for analyzing cause-effect relationships are based on the concept of Granger causality [91], which defines causality in terms of predictability, and on the concept of Pearl causality [92], which defines causality in terms of changes resulting from intervention. Currently,
the most common tool for causality analysis in the geosciences is bivariate Granger analysis, followed by multi-variate Granger analysis using vector autoregression (VAR) models [93], 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 [94], [95]. The fact that such multi-variate causality tools, 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. These range from variable selection for estimation and prediction tasks to identifying causal pathways of interactions around the globe, (see Figure 2), and causal attribution [93], [95]. The latter is discussed in more detail below.

Many components of the Earth system are affected by human actions, thus introducing the need for integrating policy actions in the modeling approaches. The outputs produced by geoscience models can help inform policy and decision making. The science of causal attribution is an essential tool for decision making that helps scientists determine the causes of events. The framework of causal calculus [27] provides a concise terminology for causal attribution of extreme weather and climate events [95]. Methods based on Graphical Granger models have also been proposed [93], but neither framework has been widely used. Of great interest is the development of decision methodology with uncertain prediction probabilities, producing ambiguous risk with poorly resolved tails representing the most interesting extreme, rare, and transient events produced by models. The application of reinforcement learning and other stochastic dynamic programming approaches that can solve decision problems with ambiguous risk [98] are promising directions that need to be pursued.

5 CROSs-cutting RESEARCH THEMES

In this section, we discuss two emerging themes of machine learning research that are generally applicable across all problems of geosciences. This includes deep learning and the paradigm of theory-guided data science, as described below.

5.1 Deep Learning

Artificial neural networks have had a long and winding history spanning more than six decades of research, starting from humble origins with the perceptron algorithm in 1960s to present-day “deep” architectures consisting of several layers of hidden nodes, dubbed as deep learning [99]. 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 big labeled data sets [100], computational advancements for training large networks, and algorithmic improvements for back-propagating errors across deep layers of hidden nodes have revolutionized several areas of machine learning, such as supervised, semi-supervised, and reinforcement learning. Deep learning has resulted in major success stories in a wide range of commercial applications, such as computer vision, speech recognition, and natural language translation.

Given the ability of deep learning methods to extract relevant features automatically from the data, they have a huge potential in geoscience problems where it is difficult to build hand-coded features for objects, events, and relationships from complex geoscience data. Owing to the space-time nature of geoscience data, geoscience problems share some similarity with problems in computer vision and speech recognition, where deep learning has achieved major accomplishments using frameworks such as convolutional neural networks (CNN) and recurrent neural networks (RNN), respectively. For example, if a CNN can learn to recognize objects such as cats in images, it could also be used to recognize objects and events such as tornadoes, hurricanes, and atmospheric rivers, which show structural features (e.g., sinkholes) in geoscience data. Indeed, the use of CNNs for detecting extreme weather events from climate model simulations has recently been explored in [101], [102]. Similarly, 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 [103], [104], [105]. Such frameworks are able to extract the right length of memory needed for making predictions in time, and thus can be useful for forecasting geoscience variables with appropriate lead times. Deep learning based frameworks have also been explored for downscaling outputs of Earth system models and generating climate change projections at local scales [106], and classifying objects such as trees and buildings in high-resolution satellite images [107]. These efforts highlight the promise of using deep learning to obtain similar accomplishments in geosciences as in the commercial arena, by incorporating the characteristics of geoscience processes (e.g., spatio-temporal structure) in deep learning frameworks. While the availability of large volumes of labeled data have been one of the major factors behind the success of deep learning in commercial domains, a key challenge in geoscience problems is the paucity of labeled samples.
thus limiting the effectiveness of traditional deep learning methods. There is thus a need to develop novel deep learning frameworks for geoscience problems, that can overcome the paucity of labeled data, for example, by using domain-specific information of physical processes.

5.2 Theory-Guided Data Science

Given the complexity of problems in geoscience applications and the limitations of current methodological frameworks in geosciences (e.g., see recent debate papers in hydrology [108, 109, 110]), neither a data-only, nor a physics-only, approach can be considered sufficient for knowledge discovery. Instead, there is an opportunity to pursue an alternate paradigm of research that explores the continuum between physics (or theory)-based models and data science methods by deeply integrating scientific knowledge in data science methodologies, termed as the paradigm of theory-guided data science [111]. For example, scientific consistency can be weaved in the learning objectives of predictive learning algorithms, such that the learned models are not only less complex and show low training errors, but are also consistent with existing scientific knowledge. This can help in pruning large spaces of models that are inconsistent with our physical understanding, thus reducing the variance without likely affecting the bias. Hence, by anchoring machine learning frameworks with scientific knowledge, the learned models can stand a better chance against overfitting, especially when training data are scarce. For example, a recent work explored the use of physics-guided loss functions for tracking objects in sequences of images [112], where elementary knowledge of laws of motion was solely used for constraining outputs and learning models, without the help of training labels. Another motivation for learning physically consistent models and solutions is that they can be easily understood by domain scientists and ingested in existing knowledge bases, thus translating to scientific advancements.

The paradigm of theory-guided data science is beginning to be pursued in several scientific disciplines ranging from material science to hydrology, turbulence modeling, and biomedicine. A recent paper [111] builds the foundation of this paradigm and illustrates several ways of blending scientific knowledge with data science models, using emerging applications from diverse domains. There is a great opportunity for exploring similar lines of research in geoscience applications, where machine learning methods can play a major role in accelerating knowledge discovery by automatically learning patterns and models from the data, but without ignoring the wealth of knowledge accumulated in physics-based model representations of geoscience processes [75]. This can complement existing efforts in the geosciences on integrating data in physics-based models, e.g., in model calibration, where parametric forms of approximations used in models are learned from the data by solving inverse problems, or in data assimilation, where the sequence of state transitions of the system are informed by measurements of observed variables wherever available [113].

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 challenges, problems, 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. That 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 because if we can understand the basic reasoning behind the patterns, models, or relationships extracted from the data, they can be used as building blocks in scientific knowledge discovery. 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 [114].

Acknowledgements

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 challenges in Section 3 has been motivated by the initial discussions at the 2015 IS-GEO workshop (see workshop report [115]).

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,” Science, vol. 292, no. 5517, pp. 641–642, 2001.
[2] F. Press, “Earth science and society,” Nature, vol. 451, no. 7176, p. 301, 2008.
[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,” Science, vol. 330, no. 6006, pp. 916–917, 2010.
Intergovernmental Panel on Climate Change, Climate Change 2014–Impacts, Adaptation and Vulnerability: Regional Aspects. Cambridge University Press, 2014.

Imme Ebert-Uphoff et al., “Climate Informatics,” [http://climateinformatics.org/]

NSF–Expeditions in Computing, “Understanding Climate Change: A data-driven Approach,” [http://climatechange.es.umn.edu/]

American Geophysical Union, “Earth & Space Sciences Informatics,” [http://essi.agu.org/]

NSF-funded Research Collaboration Network, “Intelligent Systems for Geosciences,” [https://isgeo.org/]

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. 153–163, 2014.

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,” Proceedings of the National Academy of Sciences, p. 201605617, 2016.

National Oceanic and Atmospheric Administration, “National Centers for Environmental Information,” [https://www.nco.ncep.noaa.gov/]

World Meteorological Organisation, “Global Runoff Data Centres,” [http://www.bage.dk/GRDC/]

NSF-funded Research Collaboration Network, “Intelligent Systems for Geosciences,” [https://isgeo.org/]

J. M. Wallace, J. M. Gutzler, and S. David, “Teleconnections in the geopotential height field during the northern hemisphere winter,” Monthly Weather Review, vol. 109, no. 4, pp. 784–812, 1981.

J. Kaylor, S. Liess, A. Kumar, M. Steinbach, P. Snyder, V. Kumar, A. R. Ganguly, N. F. Samatova, and F. Senazmi, “A graph-based approach to find teleconnections in climate data,” Statistical Analysis and Data Mining: The ASA Data Science Journal, Wiley Online Library, vol. 6, no. 3, pp. 158–179, 2013.

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.

P. J. Ward, B. Jorgensen, M. Kummel, M. D. Dettinger, F. C. S. Welian, and H. C. Winsemius, “Strong influence of el niño in the Southern Oscillation on flood risk around the world,” Proceedings of the National Academy of Sciences, vol. 111, no. 44, pp. 15659–15664, 2014.
[99] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT press, 2016.

[100] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009, pp. 248–255.

[101] 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 preprint arXiv:1605.01156, 2016.

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

[103] X. Jia, A. Khandelwal, G. Nayak, J. Gerber, K. Carlson, P. West, and V. Kumar, “Incremental dual-memory lstm in land cover prediction,” in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017, pp. 867–876.

[104] 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 Proceedings of the 2017 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2017, pp. 171–179.

[105] X. Jia, A. Khandelwal, J. Gerber, K. Carlson, P. West, and V. Kumar, “Learning large-scale plantation mapping from imperfect annotators,” in Big Data (Big Data), 2016 IEEE International Conference on. IEEE, 2016, pp. 1192–1201.

[106] T. Vandal, E. Kodra, S. Ganguly, A. Michaelis, R. Nemani, and A. R. Ganguly, “Deepsat: Generating high resolution climate change projections through single image super-resolution,” ACM SIGKDD, 2017.

[107] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. Nemani, “Deepsat: a learning framework for satellite imagery,” in Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2015, p. 37.

[108] 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 Research, vol. 50, no. 6, pp. 5351–5359, 2014.

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

[110] 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 Research, vol. 50, no. 6, pp. 5342–5350, 2014.

[111] 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 TKDE, 2017.

[112] R. Stewart and S. Ermon, “Label-free supervision of neural networks with physics and domain knowledge.” in AAAI, 2017, pp. 2576–2582.

[113] G. Evensen, “Introduction,” in Data Assimilation. Springer, 2009, pp. 1–4.

[114] I. Ebert-Uphoff and Y. Deng, “Three steps to successful collaboration with data scientists,” EOS, Transactions American Geophysical Union, 2017, (in press).

[115] Y. Gil and et al., “Final workshop report,” Workshop on Intelligent and Information Systems for Geosciences, 2015. [Online]. Available: https://isgeodotorg.files.wordpress.com/2015/12/