Computers & Geosciences 159 (2022) 105034

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

A review of Earth Artificial Intelligence

Ziheng Sun [a,*], Laura Sandoval [b,**], Robert Crystal-Ornelas [c], S. Mostafa Mousavi [d], Jinbo Wang [e], Cindy Lin [f], Nicoleta Cristea [g,h], Daniel Tong [a], Wendy Hawley Carande [b], Xiaogang Ma [i], Yuhan Rao [j], James A. Bednar [k], Amanda Tan [h], Jianwu Wang [l], Sanjay Purushotham [l], Thomas E. Gill [m], Julien Chastang [n], Daniel Howard [o], Benjamin Holt [e], Chandana Gangodagamage [p,q], Peisheng Zhao [a], Pablo Rivas [r], Zachary Chester [a], Javier Orduz [r], Aji John [s]

[a] *Center for Spatial Information Science and Systems/Department of Geography and Geoinformation Science/Department of Atmospheric, Oceanic and Earth Sciences, George Mason University, Fairfax, VA, USA*
[b] *Laboratory for Atmospheric and Space Physics, University of Colorado Boulder, Boulder, CO, USA*
[c] *Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, Berkeley, CA, USA*
[d] *Geophysics Department, Stanford University, Stanford, CA, USA*
[e] *Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA*
[f] *School of Information, University of Michigan, Ann Arbor, MI, USA*
[g] *Department of Civil and Environmental Engineering, University of Washington, Seattle, WA, USA*
[h] *eScience Institute, University of Washington, Seattle, WA, USA*
[i] *Department of Computer Science, University of Idaho, Moscow, ID, USA*
[j] *North Carolina Institute for Climate Studies, North Carolina State University, Asheville, NC, USA*
[k] *Anaconda Inc., Austin, TX, USA*
[l] *Department of Information Systems, University of Maryland Baltimore County, Baltimore, MD, USA*
[m] *Department of Earth, Environmental and Resource Sciences, University of Texas at El Paso, El Paso, TX, USA*
[n] *University Corporation for Atmospheric Research, Boulder, CO, USA*
[o] *National Center for Atmospheric Research (NCAR), Boulder, CO, USA*
[p] *Hydrological Sciences Laboratory, Goddard Space Flight Center, NASA, Greenbelt, MD, USA*
[q] *Earth System Science Interdisciplinary Center (ESSIC), University of Maryland, College Park, MD, USA*
[r] *Department of Computer Science, Baylor University, Waco, TX, USA*
[s] *Department of Biology, University of Washington, Seattle, WA, USA*

ARTICLE INFO

Keywords:
Geosphere
Hydrology
Atmosphere
Artificial intelligence/machine learning
Big data
Cyberinfrastructure

ABSTRACT

In recent years, Earth system sciences are urgently calling for innovation on improving accuracy, enhancing model intelligence level, scaling up operation, and reducing costs in many subdomains amid the exponentially accumulated datasets and the promising artificial intelligence (AI) revolution in computer science. This paper presents work led by the NASA Earth Science Data Systems Working Groups and ESIP machine learning cluster to give a comprehensive overview of AI in Earth sciences. It holistically introduces the current status, technology, use cases, challenges, and opportunities, and provides all the levels of AI practitioners in geosciences with an overall big picture and to “blow away the fog to get a clearer vision” about the future development of Earth AI. The paper covers all the majorspheres in the Earth system and investigates representative AI research in each domain. Widely used AI algorithms and computing cyberinfrastructure are briefly introduced. The mandatory steps in a typical workflow of specializing AI to solve Earth scientific problems are decomposed and analyzed. Eventually, it concludes with the grand challenges and reveals the opportunities to give some guidance and pre-warnings on allocating resources wisely to achieve the ambitious Earth AI goals in the future.

* Corresponding author. 4400 University Drive, MSN 6E1, Fairfax, VA, 22030, USA.
** Corresponding author. Laboratory for Atmospheric and Space Physics, Space Science Building (SPSC) W280E, University of Colorado, 3665 Discovery Drive, Boulder, CO, 80303, USA.
E-mail addresses: zsun@gmu.edu (Z. Sun), laura.sandoval@lasp.colorado.edu (L. Sandoval).

https://doi.org/10.1016/j.cageo.2022.105034
Received 10 August 2021; Received in revised form 17 December 2021; Accepted 3 January 2022
Available online 5 January 2022

1. Introduction

With countless sensors deployed all over the globe, human knowledge about earth systems is growing explosively. Every day these sensors capture huge amounts of geolocated data to help us gain a deeper understanding of the natural environment, human society, and outer space. The information is critical to (1) learn and understand natural systems, (2) foresee trends and consequences of human activities, and (3) assess hazards to human society and the Earth. Despite numerous tools, methods, and theories, we are still incapable of efficiently and fully utilizing this huge data mine. Current theories about how the earth will respond to global change are full of unrealistic and subjective assumptions due to the manual configuration and handling of data.

Artificial intelligence (AI) models have outperformed conventional data handling in many cases, like recognizing street views, extracting roads, and comprehending medical images. The first generation of AI research in the 1980s resulted in many classic theories and methods, but the earliest models took too long to train due to computing limitations. With the recent rapid development of hardware and software, AI has accelerated scientific advances and discoveries in medicine, biology, and economics.

Nowadays, AI is no longer a lab concept but used practically in many daily scenarios such as banking, camera object identification, telecommunications, household robot cleaners, recommendation systems, autonomous driving, self-checkout, etc. All of these applications depend on computer algorithms that digest information and solve problems by mimicking brain nervous systems. However, unlike human brains that can differentiate many objects by only deductively learning one object, AI algorithms must learn thousands of patterns before making accurate decisions (Qiu et al., 2016). Owing to the vital role big data plays in building AI, manipulating big data is critical to designing reliable AI-based workflows (Mayer-Schönberger and Cukier, 2013).

Geoscientists led the development of tools bridging gaps between geoscientific data and AI models (Fig. 1). Here, we probed the modern computing workflows, storage needs, and revolutionary cyberinfrastructure for conducting AI research in geosciences. The breakthroughs in both theory and infrastructure will carry geoscience into the next phase: Earth Artificial Intelligence (Earth AI). We envision Earth AI to be a huge combination of systems to automatically monitor and forecast nature, help adapt human society to environmental changes, guide humans to make planet-wise policies and decisions, and protect us from geohazards. Earth AI will be a significant tool to confront grand challenges such as exploding population, food security, and climate change. This paper will overview the current status of Earth AI, list the grand challenges, and foresee the big opportunities in Earth sciences. Section 2 describes the popular AI techniques at present, and their applications in geosciences will be introduced in Section 3. Section 4 summarizes the generic steps in Earth AI workflows, and section 5 talks about the useful tools and services. Section 6 discusses the primary challenges Earth AI practitioners face and the opportunities coming along, and it is concluded in section 7.

2. AI techniques

The term AI, a buzzword used in so many different places, can be confusing for geoscientists. The scope of AI techniques is vastly bigger than the popular ones like machine learning (ML) and deep learning (DL). Generally, ML is a subset of AI, and DL is a subset of ML. Since it is impractical to cover the entire AI universe, this section will briefly introduce the milestone techniques that are widely used in geosciences.

2.1. Knowledge-based system

Before ML became viral, rule-based systems dominated data digesting and decision support techniques, and still perform critical data analysis today. Rule-based approaches rely on a set of rules, each depicting some contextual knowledge (Clancey, 1983), typically appearing as IF/THEN expressions. For example, if the river reaches an action (flood) stage, the weather agency must take mitigation action in preparation for possible significant hydrologic activity (NWS, 2021). As

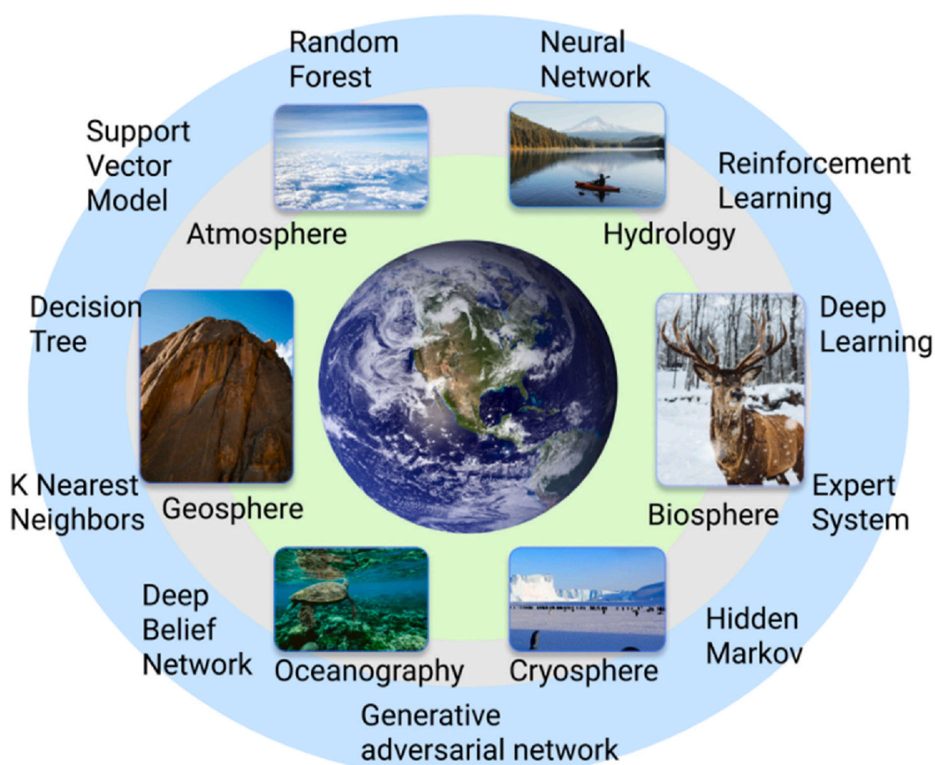

Fig. 1. Earth AI overview.

the rules are common knowledge and contain less ambiguous judgment, rule-based systems have very good stability and certainty and are commonly seen in many industries.

2.2. Probabilistic machine learning

Probabilistic ML offers a practical method for engineering machines that can evolve by learning realistic data (Ghahramani, 2019). Most ML models are using probabilistic theory to tackle uncertainty challenges. Probability theory can be utilized to express many forms of variances and noises and prevent excessive errors in prediction (Ghahramani, 2019). In ML, a probabilistic reasoner can infer the probability function given input data and eventually make predictions with control over uncertainty (Pearl, 1988).

2.3. Unsupervised learning

Unsupervised learning searches for hidden patterns in a dataset with neither annotations nor intervention (Ferran et al., 2013). Different from supervised learning heavily subject to manual labels, unsupervised learning probes the general probability densities simply based on the inputs. One of the common examples is clustering analysis entrenched in Earth scientific analysis, e.g., geochemical sample grouping (Templ et al., 2008). The clusters are automatically grouped using distance metrics like Euclidean distance in a feature space and algorithms like K Means, Hidden Markov, etc.

2.4. Supervised learning

Most current AI applications involve supervised learning which builds a transformer connecting outputs with inputs. It can be further categorized into two subtypes: regression and classification. Regression could output any continuous number in a range (such as atmospheric pressure, surface temperature, precipitation). Classification model outputs are limited to a collection of pre-fixed numbers. Supervised learning has an extensive method collection including K nearest neighbor (KNN) (Henley and Hand, 1996), Decision Tree (DT) (Safavian and Landgrebe, 1991), Support Vector Machine (SVM), Random Forest (RF) (Breiman, 2001), Artificial Neural Network (ANN) (Gurney, 2014), etc. Meta algorithms like Bagging (bootstrapping) (Breiman, 1996) or Boosting (i.e., AdaBoost) can be used to further advance accuracy and stability (Freund and Schapire, 1997).

2.5. Deep learning

Deep learning (DL) refers to a powerful group of neural networks with more hidden layers and complex architecture compared to their ancestors (i.e. Multilayer Perceptron). DL can be used in supervised, unsupervised, and semi-supervised fashion. Deep convolutional neural networks (DCNN) are commonly used for feature extraction and dimensionality reduction (Krizhevsky et al., 2012). The power of CNNs in learning representation usually results in a better performance on prediction. However, superior performance comes with a limitation that DL is more data-hungry and its application is often limited to cases when large amounts of high-quality labeled data are available (Mousavi et al., 2019).

According to data flows, DL can be generally bifurcated into two main branches: feedforward neural networks (FNN), and recurrent neural networks (RNN). The former is simple with information moving in one single forward direction. The latter has information moving in a circle, meaning the output of the previous step shall be inputted to the ongoing step. Each branch has numerous variants and forms a wide variety of advanced networks such as ResNet (He et al., 2016), U-Net (Ronneberger et al., 2015), PSP (Zhao et al., 2017), SegNet (Badrinarayanan et al., 2017), VGG-16, DenseNet (Iandola et al., 2014), YOLO (Redmon and Farhadi, 2018), R–CNN (Girshick, 2015), Mask RCNN (He et al., 2017), DeepLab (Chen et al., 2017).

2.6. Reinforcement learning

Reinforcement learning finds an optimal way to maximize a numerical reward signal (Sutton and Barto, 2018). The learning module must select actions by its own decisions to find the best path (not unique) with the most reward. It differs from supervised and unsupervised learning requiring neither training dataset nor finding hidden structure in collections of unlabeled data. A key feature is that it explicitly considers goal-directed problems by agents interacting with an uncertain environment and countless potential solutions. The term “agent” is not necessarily a real robot but could be a virtual program to explore data. Reinforcement learning is suitable for situations where it is unrealistic to retrieve data of desired behaviors that are both correct and holistic for all the possibilities that the agents might act.

3. Existing Earth AI research

3.1. Geosphere

Human population growth raises daunting challenges in requiring natural resources to sustain the population but increases vulnerability by exposing more people to natural (e.g. tectonic earthquakes, volcanos, landslides) and anthropogenic (e.g. induced earthquakes, dam failures) geohazards. Sustaining infrastructure in the face of these challenges requires a deeper understanding of these phenomena and the physical mechanisms behind them, provided by earth scientists. Although it is far from becoming fully realized, AI is now becoming widespread in all areas of geology, including the search for minerals (Saliu et al., 2020) and energy (Koroteev and Tekic, 2021).

Here is an overview of major practices in applying AI toward this goal (Table 1).

(1) Earthquake

Despite their frequency and devastating consequences, much remains unknown about earthquake generation mechanisms and effects. Earthquake forecasting, the Grail of Seismology, has been a topic of interest for extensive applications of AI techniques. Feedforward (Lin and Chiou, 2019) and recurrent neural networks (Adeli and Panakkat, 2009) are among the most used ML approaches for this task. In these approaches, neural networks predict the magnitude and location of future earthquakes (Karasözen and Karasözen, 2020)- in a time or spacetime window - often based on the time series of previous earthquake characteristics such as occurrence time, magnitude, or focus location. Despite recent progress in developing advanced DL, there are still challenges as to how it will be effectively applied to AI-based earthquake predictions (Mignan and Broccardo, 2020). This is caused by the fact that most earthquake catalogs are recorded in plain tabular format and limited features are available for training more complex models. However, DL methodologies have accelerated the development of more reliable and efficient algorithms for earthquake monitoring (Mousavi et al., 2020). AI-based earthquake monitoring methods can result in advancing seismic hazard safety in two folds: by empowering Earthquake Early Warning (EEW) systems (Bose et al., 2008) with faster and more reliable estimations of earthquake parameters and by providing more complete and precise earthquake catalogs used for improving long-term seismic hazard assessments (Mousavi and Beroza, 2018).

(2) Volcano

In volcanology, manual analyses of gas emissions, deformation measurement, and seismic signals have been used for decades to monitor, mitigate, and minimize risks associated with volcanic hazards

Table 1
Literature review summary.

Earth Spheres	AI Techniques	Research Topics
Atmosphere (39)	SVM (10) RF (7) BRT (1) ANN (12) DL (17) Cubist (1)	Ozone (5) Hurricane (4) Dust (3) Wildfire (5) Drought (4) Air Quality and Pollutants (6) Precipitation (11) Dew Point (1)
Geosphere (15)	ANN (6) Hidden Markov (1) DT (1) DL (3) SVM (2) Logistic Regression (2)	Earthquake (7) Volcano (2) Mineral (1) Landslide (4) Soil Erosion (1)
Hydrology (22)	DL (5) ANN (13) SVM (5) RF (4) Cubist (2)	Water forecasting (7) Water quality (3) Groundwater (7) Rainfall-runoff (4) River sediment (1) River discharge (2)
Cryosphere (14)	DL (5) RF (4) SVM (3) DT (2)	Glacier (2) Sea ice (9) Snow (3)
Oceanography (15)	DL (10) ANN (5)	Sea surface temperature (4) Surface process (2) Eddy (7) Deep current (1) Subsurface temperature (1)
Biosphere (16)	DL (15) SVM (1)	Animal behavior (6) Microorganism (6) Plant disease (1) Agriculture (5)

(Tilling, 1989). A major application of AI in volcano monitoring is discriminating between seismic volcanic tremors and similar events including earthquakes, landslides, lava fountains, wind, and thunder. The successfully tested ML techniques include ANN (Scarpetta et al., 2005), SVM (Masotti et al., 2006), Hidden Markov models (Beyreuther et al., 2008), and Fuzzy Logic (Hibert et al., 2014). Short-period forecasting of sudden steam-driven eruptions can also be done using AI/ML by detecting precursors from the streaming seismic data (Dempsey et al., 2020). The capability of AI in identifying the energy bursts happening from a few hours to several days ahead of large eruptions is enlightening and has proven that ML could issue life-saving short-term volcano alerts in future.

(3) Landslides

Landslides in mountainous areas cause billions of dollars in losses annually. AI applications in landslide studies have been mainly devoted to risk estimation efforts (Mousavi et al., 2011). Landslide susceptibility mapping has experimented with ML approaches like logistic regression (Umar et al., 2014), ANN (Nefeslioglu et al., 2008), and SVM (Peng et al., 2014). A set of control variables like land slope, vegetation cover, precipitation, soil mass, and hydrologic setting, are measured and used as ML inputs to calculate landslide likelihoods. Another group of AI applications is the automation of landslide identification on remote sensing (RS) imagery. For instance, CNN is evaluated in accomplishing automatic landslide detection in Nepal, concluding that CNN is “still in its infancy” for landslide detection (Ghorbanzadeh et al., 2019). Accurately predicting the place and time of landslides remains a vital challenge (Korup and Stolle, 2014). Although our knowledge about the underlying mechanism of slope failure could be weaved into physics models, the inadequate high-resolution observation of soil and groundwater restricts us from effectively running the models or enhancing the precision. Input data quality and potential overfitting remain major issues influencing the accuracy of models in real-world forecasting scenarios. Nevertheless, data mining and ML methods are increasingly popular in addressing landslide forecasting.

3.2. Hydrosphere

Hydrosphere research has greatly benefitted from AI methods and applications (Hu et al., 2018; Kratzert et al., 2018; Mo et al., 2019; Mohajerani et al., 2019; Naganna et al., 2019; Shen, 2018). This section will elaborate on three aspects: rainfall, surface water, and groundwater.

(1) Rainfall

Rainfall forecasting involves learning complex nonlinear patterns in the data. Methods proposed for rainfall forecasting include using the combinations of RNNs and SVMs (Hong, 2008; Lin et al., 2009) or Singular Spectrum Analysis (SSA) and SVMs (Sivapragasam et al., 2001). This multi-model approach was extended to include ANN, KNN, and radial basis SVM to forecast the daily or monthly precipitation (Sumi et al., 2012). Other examples include the use of convolutional LSTMs (Shi et al., 2015), RF to retrieve rainfall rates from optical satellite images (Kühnlein et al., 2014), and the combination of ANN, SVM, and DT for short-term rainfall prediction (Ingsrisawang et al., 2008).

(2) Surface water

AI-based methods have been frequently exercised on modeling nonlinear hydrological problems (Fathian et al., 2019; Yaseen et al., 2015). ML-based approaches like neuron-wavelet hybrid systems show similar performances for predicting streamflow (Anctil and Tape, 2004), monitoring coastal water quality (Kim et al., 2014), and discovering complex relationships between water level and discharge (Bhattacharya and Solomatine, 2005). FNN, generalized regression NN, and Fuzzy Logic are also helpful to populate the under-measured water-level data (Turan and Yurdusev, 2009). River researchers use ANN, adaptive network-based fuzzy inference system (ANFIS), and wavelet-coupled NN for predicting sediment load (Olyaie et al., 2015) and water level (Seo et al., 2015), and finding that ML techniques are more efficient. Coupled approaches like the ensemble of ANN, Bayesian, and Genetic Algorithms (GA) are tested and yield improvement (by 3–11%) (Perea et al., 2019). RNN like LSTM was used in discovering polluting substances in water (Wang et al., 2019b). Remote sensing data like Landsat 8 images provide rich data sources for ML to quantify concentrations of different surface water quality parameters (Sharaf El Din et al., 2017). Considering water-society research, ML models have been utilized successfully in forecasting water consumption around Indianapolis (Shah et al., 2018) and many other scenarios.

(3) Groundwater

As groundwater is hard to measure at scale, AI-based algorithms are useful in deriving information and making predictions crucial for groundwater management. ML has successfully created ground water management maps (Barzegar et al., 2018), assessing risks of nitrate contamination (Nolan et al., 2015; Sajedi-Hosseini et al., 2018) and predicting groundwater levels (Sahoo et al., 2017). ML models including SVM, RF, and GA optimized random forest, can assess groundwater potential by locations (Naghibi et al., 2017). It noticed that RF outperforms classification and regression trees (CART) in large-scale nitrate concentration prediction (Knoll et al., 2019). Ensembled ML models are practical alternatives to sophisticated conventional models to perceive the subsurface water patterns. Regarding city underground water networks, ML (e.g., extreme learning machine - ELM) (Sattar et al., 2019) can help in estimating the potential failures on individual pipes to prevent future tragic events.

3.3. Atmosphere

This section highlights the progress of AI development in atmospheric phenomena. In addition to addressing the specific atmospheric geohazards below, AI is of growing importance in essentially all aspects of meteorology, especially for improving the skill and efficiency of numerical weather forecasting, and in assimilating and interpreting the huge amounts of data contained in weather satellite observations (Boukabara et al., 2021).

(1) Hurricane

Tropical cyclones (hurricanes, typhoons, etc.) are amongst the most costly of all the disasters (Klotzbach et al., 2018). ML was used to predict hurricane path and assess damage using reanalysis data (Giffard-Roisin et al., 2018) and satellite images (Cao and Choe, 2020; Yu et al., 2019). The damage annotation ML model achieved >97% accuracy for Hurricane Harvey. Time-series forecasting models like RNN and ConvLSTM can learn hurricane behavior and calculate trajectories (Alemany et al., 2019; Kim et al., 2019). Extensive experiments using 20 years of climate reanalysis data show that ConvLSTM has higher accuracy than other approaches. Other data sources like passive microwave satellite data are also used together with DL for monitoring tropical cyclones (Wimmers et al., 2019). To simplify the problem by removing small-scale low-impact events, DL has successfully detected only severe storms (Maskey et al., 2018). From the social impact perspective, some researchers used ML to rapidly identify hurricane-critical Tweets (Shams et al., 2019).

(2) Meteorological Drought

Drought is a complex natural hazard causing tremendous global economic, social and environmental damages every year (Wilhite, 2016). Efforts have applied ML for drought prediction in Africa (Belayneh et al., 2016), Australia (Deo and Şahin, 2015), the USA (Agana and Homaifar, 2018), and China (Chen et al., 2012). Some studies used ML to predict drought indicators (Sutanto et al., 2019), such as SPEI and SPI (Belayneh and Adamowski, 2012; Maca and Pech, 2016) and estimate drought severity at ungauged sites (Sadri and Burn, 2012). ML-powered a high-resolution drought forecasting model using remote sensing data (Rhee and Im, 2017). On product processing, different ML methods are compared in downscaling hourly reanalysis precipitation to monthly data, and relevance vector machines work best (Sachindra et al., 2018).

(3) Wildfire

Wildfires are increasing in many countries, imposing adverse effects on human health and the economy. Early fire detection and intervention are vital for wildfire damage minimization. Various AI/ML methods have been applied to improve fire detection and prediction (Jain et al., 2020), classify and map wildfire severity (Brewer et al., 2005), and automatically detect wildfires on UAVs or satellite images (Zhao et al., 2018). High-profile studies used AI in improving smoke plume forecasting combining ML with satellite (e.g., CALIPSO) observations (Yao et al., 2018) and infer ozone expansion and distribution (Watson et al., 2019). Other applications include identifying wildfires on RS images (Sayad et al., 2019) and assessing human health issues connected to poor air quality (Reid et al., 2016). Meanwhile, scientists use ML to trace human-caused wildfires and found RF is currently the most accurate among those tested (Rodrigues and de la Riva, 2014).

(4) Dust storm

Dust sources are associated with multiple health effects and socioeconomic impacts, including infectious diseases (Tong et al., 2017) and highway safety (Ashley et al., 2015). ML is increasingly used to detect dust sources, transport, and wind erosion susceptibility at various scales (Boloorani et al., 2022; Gholami et al., 2021; Lin et al., 2020). ML was utilized in inverse emission modeling to improve accuracy and outperformed a traditional chemical transport model (Jin et al., 2020). A Dust Source Susceptibility Map (DSSM) was developed using RS and ML to show dust sources 2005–2016 in Iran (Boroughani et al., 2020). Various ML models were benchmarked to investigate soil susceptibility to dust, finding RF performs best (Gholami et al., 2021). On a global scale ML is still applicable (Lee et al., 2021).

(5) Anthropogenic Air Pollutants

Air pollution is associated with over seven million premature deaths each year (WHO, 2021). A majority of them stem from exposure to O_3 (ozone) and $PM_{2.5}$ (fine particles). However, the ever-changing dynamics make it extremely difficult for computer models to predict air quality. AI has been involved to address these challenges, particularly for predicting O_3, $PM_{2.5,}$ and nitrogen oxides, a precursor chemical that contributes to the formation of O_3 and $PM_{2.5}$ (Nowack et al., 2018; Wang et al., 2003; Wu et al., 2017; Zhang et al., 2012). Earlier works often utilize neural network methods to improve air quality forecasting (Abdul-Wahab and Al-Alawi, 2002; Kolehmainen et al., 2001; Ruiz-Suarez et al., 1995). Recently, more advanced ML algorithms are used to enhance O_3 and NO_2 prediction and SVM is better than NN in predicting daily maximum O_3 concentrations (Chelani, 2010). For small-grain air quality forecasting, DL can complete common tasks like mosaicking, inserting missing values, or selecting features (Du et al., 2018; Fan et al., 2017; Qi et al., 2018).

3.3.1. Biosphere

The biosphere represents the living parts of the Earth system. This section briefly introduces the status of AI in life sciences under three themes: plant, animal, and microorganism.

(1) Plant (Botany)

Phytogeography, the study of plant distribution, is an active area in Earth AI research, and using RS imagery and ML, especially DL, has become the mainstream technique due to the low cost and the high accuracy of ML classification. AI-derived maps are proliferating in biogeographic studies. DCNN, trained on a public dataset of leaves to distinguish fourteen crops and twenty-six diseases, can achieve 99.35% accuracy (Mohanty et al., 2016; Sun et al., 2019a). Agriculture has many profound use scenarios for AI like disease detection, crop yield prediction, and irrigation recommendation (Kamilaris et al., 2017). Coupled RNN-CNN model can predict corn yield in the midwestern U.S (Sun et al., 2019b) and can be a low-cost reliable alternative in guiding irrigation (Vij et al., 2020).

(2) Animal (Zoology)

Advances in sensing technologies provide big data of animals like GPS and video surveillance. Together with data manually collected by professionals and citizen scientists, a huge dataset exists on wild animals' location, movements, behaviors, and well-being. Similarly, big data is becoming a norm in animal agriculture (Neethirajan, 2020). Based on these datasets, the application of AI in zoology focuses on detecting, counting, and describing animals and their behavior from images. DL has been proven efficient in recognizing wild animals on camera-trap imagery (Chen et al., 2014), attributing wildlife behaviors (Norouzzadeh et al., 2018), detecting ultrasonic calls of bats (Mac Aodha et al., 2018), and projecting diving of cormorants (Browning et al., 2018). For urban animals, DL can analyze city audio data

(Fairbrass et al., 2019) and animal trajectories (Maekawa et al., 2020). However, despite progress, AI in zoology is still in an experimental phase and hasn't fully penetrated the zoological community.

(3) Microorganisms

Similar to zoology, AI is intensively studied in microbiology (Egli et al., 2020). DL has identified 30 common bacterial pathogens (Ho et al., 2019), detected pathogenic bacteria in food and water on time-lapse holograms (Wang et al., 2020a), and achieved an overall accuracy of 99% for 80-diatom classification (Kloster et al., 2020; Pedraza et al., 2017). DL-driven workflow can automatically recognize microscopic images of viruses, bacteria, fungi, and parasites (Zhang et al., 2021). Scientists also use AI in predicting the evolution of microorganisms, estimating optimal growth temperature for bacteria, archaea, and microbial eukaryotes, (Li et al., 2019), and predicting sgRNA activity in *Escherichia coli* (Wang and Zhang, 2019). However, since ML requires a lot of work to obtain adequate training labels, pre-trained models can be repurposed to classify environmental microorganisms to lower the cost (Kosov et al., 2018).

3.3.2. *Cryosphere*

Polar science studies the Earth's frozen zones, which are more highly subject to environmental changes than the planet as a whole. Despite years of efforts on modeling, precisely forecasting changes and consequences is still an unsolved challenge for the cryosphere community.

(1) Sea Ice

AI/ML has been used to map the ice shelves in Antarctica from Sentinel-1 (Baumhoer et al., 2019), estimate Arctic sea ice thickness (Tiemann et al., 2018), and evaluate its melting speed on SAR images (Lee et al., 2016; Wang et al., 2016) and distinguish water from ice (Leigh et al., 2013). It can help identify ages/types of sea ice as radar backscattering signals of sea ice are composed of scattering from both the rough surface as well from underneath ice according to radar signal penetration (Ghanbari et al., 2019; Lohse et al., 2019; Park et al., 2020). GNSS images and ML can be harmonized for sea ice detection (Yan and Huang, 2018). The ambiguous connections between microseisms and sea ice activities are also suitable for AI/ML (Cannata et al., 2019).

(2) Snow

Snow research has two main indicators: snow water equivalent (SWE) and snow depth; both can be monitored and forecasted by AI/ML with decent reliability (Holt et al., 2015; Wang and Zhang, 2019). SVM-derived snow depth products from microwave satellites can pass the validation tests by stationary observation with higher precision while effectively suppressing the saturation effects (Xiao et al., 2018). Advanced DL methods such as deep residual networks show excellence over RF, SVM, and NN in snow detection from satellite imagery (Xia et al., 2019). Meantime, AI/ML is intensively experimented to differentiate snow from cloud at the pixel level (Zhan et al., 2017).

3.4. *Oceanography*

The turbulent ocean contains small-scale eddies that imprint on oceanographic observables like sea surface height (SSH), color, roughness, and temperature (SST). Identifying these features with ML is a hot study area. Oceanic mesoscale eddies (~300 km diameter) are usually identified by physics-based algorithms and previous seminal work produced an eddy database (Chelton et al., 2011) as a robust benchmark for ML. So far, CNN has been used in eddy identification with SSH (Franz et al., 2018; Santana et al., 2020), SAR images (Du et al., 2019; Huang et al., 2017), high-frequency radar (HFR) data (Liu et al., 2021) and SST images (Moschos et al., 2020).

SAR provides unprecedented detail of ocean surface roughness at fine resolutions (~10–25 m). With higher-quality Sentinel-1 succeeding the earlier Radarsat-1 and Envisat missions, ML efforts are increasing on SAR ocean imagery to identify and map many surface features beyond eddies (Wang et al., 2019a). Submesoscale eddies (on the order of 5–30 km diameter) are more fully captured on standard SAR imagery in coastal regions under low to moderate wind speeds due to multiple dark, curvilinear slicks within each eddy. An early application of ML in SAR ocean detection was mapping oil spills arising out of petroleum seeps (Garcia-Pineda et al., 2009, 2013).

Satellite ocean surface observations are intrinsically gappy due to cloud cover or sparse ground tracks such as the conventional nadir altimeters and upcoming SWOT (Durand et al., 2010) altimeter mission. AI/ML can address the gappy issue in synthetic SWOT SSH data demonstrating the feasibility of AI-based interpolation algorithms in filling gaps containing small-scale ocean eddies (Manucharyan et al., 2020). One step further, the CNN-based algorithm can be applied to reconstruct fluxes induced by those eddies (Bolton and Zanna, 2019; George et al., 2021). These algorithms will be useful in parameterizing eddy fluxes not resolved in coarse resolution climate models.

Since ocean circulation is three-dimensional, AI-based algorithms can also retrieve deep-ocean information based on surface satellite fields (Ali et al., 2007; Cheng et al., 2021; Wang et al., 2021). Other methods include a self-organizing map (Chapman and Charantonis, 2017; Wu et al., 2012), CCN (Han et al., 2019), neural net with fruit fly optimization algorithm (Bao et al., 2019), and RF (Su et al., 2018).

Oceanography is transiting from a state of data scarcity to a state of extreme data abundance. How to utilize a sea of data on a scale of petabytes and distill useful information for either new scientific discoveries or applications of a direct societal impact on the "blue economy" is a new challenge for the community (Watson-Wright and Snelgrove, 2021). AI-based algorithms will foreseeably play a compelling role in the transition.

4. Workflow

4.1. *Data preparation*

In most supervised ML research, a training dataset includes two components: input observations and associated labels. Inputs are fully observed and cyclic data sources like RS images, stationary data, model simulations, etc. Output variables are usually less-observed but critical for understanding Earth system processes, such as emissions, land cover, soil moisture, etc. Several problems arise in the process:

(1) Time Series

The time axis is a fundamental characteristic of Earth data for trend analysis and forecasting. Earth observations are discrete sequences of numbers (e.g., samples per second, minute, hour, etc) in which data gaps and time-varying noise are common. Bandpass filtering, down sampling, up sampling, detrending, interpolation, and smoothing are commonly applied to preprocess time series data.

(2) Format

Almost every major data provider or professional software has an exclusive self-defined format. For example, HDF is the official format in NASA, NetCDF is commonly used in NOAA and climate communities, and GeoTiff is popular for georeferenced imagery. Furthermore, each format has various versions that might cause compatibility issues in I/O programs. Libraries like GDAL/OGR and NCO could address these problems. However, disparate formats still create a headache in aggregating multiple source datasets, requiring extra effort.

(3) Projection & Grid

Multisource datasets usually have various coordinate systems. NASA products use Sinusoidal projection, netCDF uses a 4-D Grid space system, OpenStreetMap uses EPSG:5070, and many public datasets use WGS84 (EPSG:4326). To integrate data from different sources within the same region/location, data needs to be re-projected or re-gridded into the same coordinate system. Any displacement can result in erroneous misleading conclusions. GDAL, Proj4, ESMF Regridding Toolkit, are common tool solutions for re-projection and re-gridding.

(4) Metadata

Metadata is an important part of data acquisition and sharing. By providing information like naming conventions, variable units, resolution, projection, observation time, contact information, and data file versions in a comprehensive and standardized manner, can potentially enable more efficient reuse of datasets. However, if metadata is not standardized, the underlying datasets may be misused if users are unfamiliar with the data or don't fully understand the provenance of the data contained within files (e.g., that precipitation is reported in inches or centimeters) (Mons 2020). A recent survey suggests that most researchers do not use or are unfamiliar with metadata standardization protocols for their disciplines (Tenopir et al., 2020).

4.2. Model building

Building an appropriate ML model for a specific problem in Earth sciences is tricky, requiring much comparison and experimentation. Specialists must gain expertise with several models and compare their performance characteristics before choosing one best meeting their objectives.

As an example, given a problem description, there is no generic methodology to assess *a priori* about the optimal setup of neurons and layers for an ANN model. A common approach starts with a rough guess based on prior experience about networks employed on similar problems. This supposition could be user's experience, or second/third-hand experience learned from a training course, blog, or research paper. At that point, the researcher may try some variations and carefully assess the performance of the model before deciding on a strategy. The size and depth of neural networks interact with other hyperparameters and changing one variable can affect the other hyperparameters. A simple stepwise guide is:

- Create a network with hidden layers of similar size to the input.
- Try varying network widths and depths.
- Try dropping out some nodes and other solutions (e.g., dropout, learning rate decay, regularization, optimization algorithm, loss function, etc).
- After a few adjustments, settle on an overall better model.

Users shouldn't get lost in tuning ML models as there will always be better models. Exploring the data helps form a reasonable expectation of accuracy. Attempt simple linear approaches first to create benchmarks to surpass. Considering a different ML algorithm may be mind-changing, faster, and more effective than your original pick.

4.3. Training, testing & validation

Most ML models need three datasets: training, validation, and testing. In practice, the overall dataset is first fractionated into the learning dataset and test dataset. The learning data is further split into a training dataset and validation dataset. Training datasets are used to fit the model. Validation datasets provide a real-time evaluation of the model during training. Test datasets provide an out-of-box evaluation of the final model. There is no fixed optimal ratio to allocate the three datasets. To ensure the model is unbiased, the splitting is repeated N times, and the accuracy is averaged, which is called N-fold cross-validation.

4.4. Sensitivity analysis

Sensitivity analysis is a series of methods used to quantify ML uncertainty. It studies the feature importance of each input variable for the outputs. To measure the influence of each input variable, a comparison is made on model outputs with all variables in place and the model with one variable excluded or fixing the values of all other variables, only tuning the weight of one input factor to discover how the model output changes. Sensitivity analysis is mandated for practical use of ML in the real world; it explicitly reveals the level of dependence of model output on each variable, and hands more control to practitioners, especially when the new observations are extreme events and could be extra outliers exceeding the prediction capacity of models.

5. Tooling and services

The big data nature of Earth science and the high complexity of AI algorithms demand powerful computing. This section overviews popular hardware and software for Earth AI.

5.1. Computing device

Commonly used ML devices are Central Processing Unit (CPU), Graphics Processing Unit (GPU), Field-Programmable Gate Array (FPGA), and specialized accelerators (e.g., TPU - Tensor Processing Unit). GPUs are dominant due to their performance in speeding up the calculation of convolution and matrix operation. In DL the weights are updated in every cycle and are stored in a memory or local cache to be carried over from iteration to iteration. GPUs have higher memory bandwidths than CPUs and are optimized for more intensive workloads and streaming memory models.

In addition, scientists actively explore the next revolution in AI computing. After R. Feynman proposed the idea of a quantum computer, quantum computing is believed to be the next potential big breakthrough by producing the statistical patterns that are computationally difficult for a classical computer to produce (Biamonte et al., 2017; Deutsch, 1985; Feynman, 2018). Edge computing is another way around by leveraging the Internet of Things (e.g., endpoints, gateways, smart watches, smartphones, sensors, etc) with embedded AI techniques to process data locally without transmitting much data, which can reduce reliance on networks and increase the AI's resilience and practicality (Li et al., 2018).

Individual researchers can set up their workstations by assembling GPUs into a computer. Research groups and institutes can purchase more powerful pre-built servers configured by professionals. Self-maintaining workstations cost less if the experiments will last long. However, maintainers are required to build and sustain the rig. They need to find appropriate GPUs, compatible motherboard, CPU, and memory and fix any problems observed like GPU collapse, memory leak, disk failures, etc. This solution is suggested for people with experiences on servers.

5.2. Cyberinfrastructure

Manipulating large-scale high-resolution Earth datasets requires massive computational power beyond the capacity of personal computers or even self-built DL workstations. Private companies with large computing power have developed some public cyberinfrastructure as the ultimate solutions. One typical example is Google Earth Engine (GEE) (Gorelick et al., 2017), which has digested petabyte-scale archives of publicly available RS imagery and model-simulated data. It optimizes Google's computational infrastructure for the parallel processing of geospatial data. Utilizing provided APIs with basic ML algorithms in Javascript and Python, GEE has powered many breakthroughs in

RS-based Earth scientific researches like natural resource management, climate change monitoring, and disaster prediction and evaluation (Amani et al., 2020; Campos-Taberner et al., 2018; Tamiminia et al., 2020).

To exercise AI techniques on GEE, Colab (Bisong, 2019), a Jupyter-notebook-like interactive coding environment, can be used to program deep neural networks or other complicated ML models. Colab allows people to write and execute Python in web browsers with zero configuration required and easy sharing. With Colab, Earth scientists now can work with large datasets, build complex AI models, train them at lower cost and share the results seamlessly with others.

As a major competitor to the GEE ecosystem, Amazon is developing AI capability rooted in its AWS (Amazon Web Service) ecosystem. SageMaker (Januschowski et al., 2018) is their recent product and advertised as a managed web service to create and deploy ML models faster. SageMaker could be considered as an AutoML solution for scientists who are less technical and want less coding.

5.3. Software

The recommended operating system is Linux-derived systems with active long-term technical support. At present Ubuntu is the bellwether with many built-in dependencies for AI. It is easy for users to install GPU drivers, like CUDA (a software allowing coding for NVIDIA GPUs), and Python package manager (i.e., Conda, Pip) can facilitate the package installation. To interact with the machines remotely, Jupyter server (Kluyver et al., 2016) (either of notebook, Lab, or Hub) is highly recommended. It allows Earth scientists to create and share their experiments, from codes to full result reports in one single document to streamline their work and enable more productivity and easy collaboration.

The dominance of Python in the AI world is largely credited to its thriving, openly accessible, and pro-collaborative library ecosystem. Table 2 lists some widely used open-source libraries. Generally, those tools can be categorized into six types: DL, non-DL ML, non-ML AI, data manipulation, parallel computing, and visualization. These tools play a significant role in recent scientific breakthroughs, i.e., plotting the first blackhole photo (Numpy, 2020), confirming the existence of the gravitational waves (Biwer et al., 2019), the mission to fly a helicopter on Mars (Vaughan-Nichols, 2021), etc. Many tools are for processing Earth scientific datasets, such as Rasterio, Shapely, Geopandas, ESMPy, which make the infusion between Earth science and AI techniques possible.

6. Challenges and opportunities

This section highlights some major challenges and potential opportunities (shown in Fig. 2).

6.1. Model development

Model development is the process of choosing one suitable model or customizing a coupled model for one or multiple training datasets. Candidate off-the-shelf models include single models such as Neural Network, SVM, and Decision Tree, as well as ensemble models like RF, XGBoost, and most DL models. Finding optimal models or coupling new models is time-consuming and might never be satisfactory, which created a strong demand for AutoML that does not require expert knowledge or manual tuning. For example, OptiML, AutoScikit-learn, and AutoWeka use Bayesian parameter optimization for predicting the model's performance on a given dataset, assuming the performance of an ML algorithm is data-dependent. For instance, OptiML, after automatically trying a few models, can learn a regression model to predict the performance of other not yet tested models to save time. Auto-sklearn's hyperparameter tuning also uses Bayesian optimization, meta-learning, and ensemble construction. However, unsolved serious issues remain. First, the best metrics used for selecting models should be different according to various use cases. Second, the cross-validation technique performs poorly on big data training. Third, performance on accuracy should not be the only factor: stability, reliability, computational cost, and generalizability are all very important and often overlooked in seeking solutions.

A good AutoML solution should automatically produce a model addressing all the concerns on scenario adaption, big data, and comprehensive metrics besides accuracy performance. The shortage of ML experts in industry and academia has been widely acknowledged, yet highly skillful ML experts are rare to find and hard to train. AutoML can bridge that gap and could derive many new opportunities in the AI job market, including Earth science. With AutoML, model selection would be easy and quick, and the barrier of shopping around ML models will be greatly reduced. AI-powered value-added services would no longer be the privilege of tech giants. Small groups will also be able to quickly put solid models together to simulate the real world, extract actionable

Table 2
Python ecosystem for earth AI.

Category	Name	Description	License	Github Repo
DL	**Keras**	A friendly API running on top of Tensorflow	MIT	keras-team/keras
	PyTorch	Multidimensional array (tensor) computation with strong GPU acceleration, for deep neural networks	BSD	pytorch/pytorch
	Tensorflow	A powerful open-source platform for ML	Apache-2.0	tensorflow
	Chainer	DL framework aiming at flexibility	MIT	chainer/chainer
	Caffe	Fast DL	BSD	BVLC/caffe
	Mxnet	Efficient and flexible DL	Apache 2.0	apache/incubator-mxnet
ML	**Scikit-learn**	ML built on SciPy	BSD	scikit-learn/scikit-learn
	OpenCV	Computer vision and ML	BSD	opencv/opencv
Non-ML AI	**PyKe**	Knowledge-based inference engine	MIT	e-loue/pyke
Data I/O	**Numpy**	A basic package to provide N-d arrays, and linear algebra methods, and mathematical transforms for conveniently manipulating N-d arrays.	BSD	numpy/numpy
	Pandas	Support various data operations like reshape, merge, slice, extract, clean, etc.	BSD	pandas-dev/pandas
	Xarray	Simple labeled multi-dimensional arrays	Apache	pydata/xarray
	Zarr	Chunked, compressed, N-dimensional arrays	MIT	zarr-developers/zarr-python
	Shapely	Manipulation and analysis of planar geometric objects	BSD	Toblerity/Shapely
	Geopandas	Support for geographic data in pandas	BSD	geopandas/geopandas
	Rasterio	Read and write gridded or raster datasets, with API based on N-D arrays	BSD	mapbox/rasterio
Parallel Computing	**Dask**	Parallel computing with task scheduling	BSD	dask/dask
	Ray	Building and running fast distributed applications	Apache-2.0	ray-project/ray
Visualization	**Matplotlib**	Static, animated, and interactive visualizations	PSF	matplotlib/matplotlib
	Plotly.py	Interactive, open-source, and browser-based graphing and apps	MIT	plotly/plotly.py
	hvPlot	Interactive plotting and apps directly from your xarray, pandas, dask, or geopandas data	BSD	holoviz/hvplot

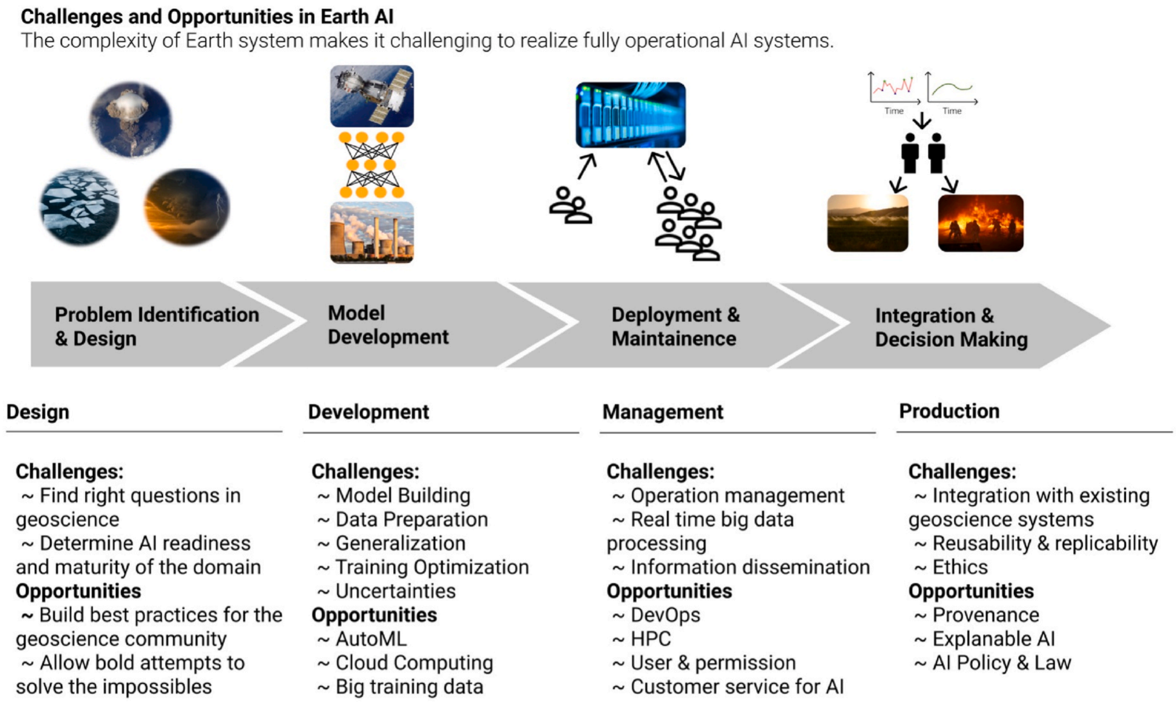

Fig. 2. Challenges and opportunities.

information, and guide climate and environmental policy making. New doors will be opened for the next generation of Earth AI.

6.2. Data preparation

The majority of an Earth AI project is typically spent on data preparation. Acquiring a large-scale labeled dataset in Earth science is very costly as labeling is usually manually done by in-house labor. A popular tactic is to crowdsource hand labeling tasks using services like Amazon Mechanical Turk. Despite unprecedented amounts of data to analyze, a lack of openly available curated and labeled training data is an obstacle to realizing efficient AI in the earth sciences (Maskey et al., 2018; Reichstein et al., 2019). Standardized training and testing datasets launched the AI revolution in other disciplines (e.g., imagnet, MNIST), yet training datasets capturing the diversity of geoscience data are being developed, and where they exist are heavily used. For example, Spacenet, an online hub for satellite imagery, algorithms, and tools, provides RS data with labeled information for ML. This leaves inexperienced modelers with the time-consuming and difficult task of locating, integrating, and labeling disparate datasets. Other times earth scientists must go outside of their domains to train their models. Right now, the incentive structure has scientists focused on ‘building a better algorithm’ rather than curating datasets (Hutchinson et al., 2021).

With more data producers, repositories, and publishers embracing calls for FAIR data, community-developed data standards (Sansone et al., 2019) are being developed where no international standards exist. OGC standards have been developed by international members to make geospatial information and services FAIR. There is a movement in the Earth and Environmental Sciences to create libraries of standardized and benchmark datasets (ESIP, 2021). These benchmark datasets can be used to efficiently evaluate how newly developed algorithms perform compared to already existing models on a common, standardized dataset. Standardization of benchmark datasets can lift data curation burdens by offering ready-to-use data for modelers (Reichstein et al., 2019).

6.3. Training optimization

Tuning AI models is an essential but painful experience step for many beginners. It is a process of adjusting hyperparameters to minimize the cost function. Optimizers are algorithms for changing the attributes like weights and learning rates to lower the losses. Commonly used optimizers include Gradient Descent, Nesterov’s Accelerated Gradient, Adaptive Moment Estimation, AdaDelta, etc. One common challenge of gradient-based optimizers is that most found minimum points are local minima. The global minima are hard to locate as the gradient becomes smaller when the training goes further and the learning rate is too large to get closer to the right answer. Another way is the genetic algorithm that applies the theory of evolution to ML. The process is repeated many times and only the best models survive at the end of the process. All the optimization methods have flaws. No one-size-for-all method can adapt to any dataset and speed up the learning to reach minima faster. An ideal ultimate solution should make the training quickly converge to the point with minimum loss within fewer iterations/epochs. The gradient vanishing problem (the gradient is too small to update the weight in the next loop) should be well addressed.

6.4. Parallel computing

Parallel computing, which improves the efficiency of AI training and running, is a valuable tool in Earth AI. The first reason is the ever-increasing size of available Earth data due to advances in both RS techniques and numerical Earth simulation. For example, total available climate data may increase exponentially from 100 PB in 2020 to about 350 PB in 2030 (Overpeck et al., 2011). The second reason is the increasing complexity of AI models. Advances in ML models, especially DL models, are more and more complex to achieve prediction accuracy. For instance, the Turing natural language generation model from Microsoft has 17 billion parameters. Because of the two reasons, it could take weeks or even months for a complex AI model to train without parallelization (Johnsirani Venkatesan et al., 2019).

There have been many efforts on studying how to support parallel ML from different perspectives. (Verbraeken et al., 2020; Wang et al., 2020b). We summarize three opportunities for parallel ML below, the first general to all ML tasks, the second and third unique to Earth AI. The first opportunity is the requirement of developing a unified system combining parallel hyperparameter tuning and parallel deep model training. Currently, these two tasks are often done via different systems, for example, Spark for parallel hyperparameter tuning and Tensorflow supporting parallel DL. A more integrative way/platform supporting both efficiently is still needed. The second opportunity is the support for parallel learning on top of array-based Earth system datasets including HDF and NetCDF. Xarray and Dask are recent community efforts on accessing/processing HDF and NetCDF datasets efficiently. But it is still not clear how to integrate these techniques with machine/DL. The last opportunity is parallel ML support for spatiotemporal data, typifying Earth system datasets. Unlike traditional independent and identically distributed (IID) datasets, partitioning spatiotemporal data would break their spatial/temporal correlation and dependence. Therefore, special attention should be considered for parallel ML with spatiotemporal Earth data.

6.5. *Explainable AI*

Compared to basic or tree-structured ML models (e.g. linear regression, DT, Bayesian, RF), complex ML models (e.g., DNN, SVM) cannot provide a self-explainable theory for their results. Many Earth scientists called for adding explanation into ML models to facilitate understanding of the ML models and build user trust. Explainable AI (XAI) tools provide a way to look into the original “black-box” model with “explanations” providing a qualitative understanding of relationships between model features and predictions. This process answers questions about the model, such as what features are the most important and why some features are more responsible for driving decisions than others. It also provides insights allowing for meaningful changes to the models. An overview of common explainable methods can be found in Molnar et al. (2020). Decisive factors in selecting XAI methods may include the need for model-agnostic or model-specific methods, the extent of the explanation required, and spatiotemporal or computational constraints.

Limitations of current XAI methods include that they cannot tell the problems in the training dataset, and they focus on RGB images and are user-friendly for high dimensional images (Krishnan, 2019). Despite the problems, opportunities exist with XAI for improving geoscientific models. Artifacts that create errors in numeric models could be revealed by XAI.

6.6. *Generalization*

The conventional goal of generalization is to make trained AI models perform better on the test data. However, it becomes complicated as the Earth dataset is tremendous and the training dataset is only a tiny portion. In Earth AI, it is no longer simply finding a balancing point between overfitting and underfitting: models trained in one place at one time may not apply in another place at another time. However, a root cause of common AI failures is that current empirically trained models do not generalize well on new samples with different distributions. Finding a good generalization strategy to make models fit beyond the training dataset is a major bottleneck for applying AI in Earth science. The developing field of generalization theory may hold promise in solving these problems.

AI generalization has been studied for decades. Ockham’s Razor principle (Ariew, 1976) proves the less complex a model is, the more likely a good empirical result is not just due to the peculiarities of the chosen samples. The edges between under-learning and over-learning the training samples are obscure. One of the classic methods to detect underfitting or overfitting is to separate samples into two parts: training subset and testing subset. During each iteration in the training, the program will run the trained model on the testing subset to calculate the prediction accuracy on samples that are outside the original training pool. If the accuracy of testing data starts to gradually decrease, it means the model is overfitting. On the contrary, if the testing accuracy hasn’t reached the peak, it means the model is still underfitting. A method is needed to find a balance between bias (underfitting) and variance (overfitting). One common solution is cross-validation to ensure no coincidental training bias is in place. Regularization is another technique used to make the learning algorithm generalizes better. It focuses on reducing the impacts of noise samples that don’t reflect the real characteristics of the dataset, but random errors and coincidences. It discourages training a more sophisticated model to reduce the risk of poor generalization. Dropout is a recently proposed approach dedicated for neural networks to randomly drop units to force the subsequent layers to rely on all their connections to previous layers. However, no method can avoid intensive endless tuning to optimize the model with better generalization.

An attractive feature of Artificial Intelligence is that model performance will improve when a model is fed with larger datasets. However, it will eventually reach some limits posed by the model capacity that is capable of learning. Many DL models are over-parameterized and likely to become biased after learning more noise samples. Addressing the generalization problem will make AI models of the Earth system much more stable and noise-proof in a long-time operational run. A future solution would be to run an automatic algorithm to self-adjust in adopting samples by judging their quality. Those samples which might destabilize the model should be automatically given less consideration in the propagation and their impacts on the future updating should be reduced.

6.7. *Uncertainties*

ML models are fundamentally algorithms composed of a set of rules, which involve random number generation and optimization to determine model parameters. Therefore, ML models developed on the same dataset are almost always different. The uncertainty of ML applications is a combination of uncertainties from two sources: data and knowledge. The uncertainty associated with the inherent noise of the real data is also known as aleatory uncertainty, which is not caused by the model but irreducible (Hüllermeier and Waegeman, 2021). The uncertainty caused by inadequate knowledge and data is also called epistemic uncertainty, which is often a result of the mismatch between the data in model training and prediction.

To quantify the aleatory uncertainty, we need to estimate the uncertainty of all the inputted data of ML models and understand how uncertainty propagates through the model. This can be challenging for DL models because of the high model complexity. A small permutation in input data for a DL model can lead to notable changes in final model outputs. The epistemic uncertainty is related to the issue of generalization. Most ML applications are developed based on a specific set of data, thus the model may not be easily generalized to other conditions that are not covered in the original dataset. Because of the lack of representation in the original data set, it can be very challenging to accurately quantify the uncertainty related to generalization.

Accurate uncertainty quantification is essential to enhance users’ trust and increase the usability of ML applications. To address uncertainty quantification (UQ), many statistical and computational methods have been proposed. The most commonly used methods can be grouped into two categories – Bayesian UQ and ensemble UQ. Bayesian UQ approaches focus on approximating the posterior probability distribution given the training dataset (Abdar et al., 2021). Ensemble UQ means training multiple models, calculating their synthesized prediction (e.g., mean), and measuring uncertainty using deviation. Recently, there have been different variations of Monte Carlo (MC) simulation (Ferrenberg and Swendsen, 1989) for UQ, such as MC Dropout (Gal, 2016), to characterize prediction uncertainty more efficiently.

6.8. *Integration with physics-based models*

Model-driven solutions based on known physical laws have long been the main trend in applied sciences. Numerical modeling plays a dominant role in Earth system science on scales ranging from performing density functional theory (DFT) calculations to predict properties of molecules, to studying the climate using general circulation models (Han and Zhang, 2020). However, difficulties remain within developing efficient and accurate models. Unlike traditional physics-based Earth science models requiring high flops and massive CPU cores, ML, especially DL, can parallelize its processing by simply using GPU, or custom processing units like TPU to achieve the same effects as a stack of massive CPUs. Currently, there are two main trends in approaching this problem: 1) partial use of AI or AI-platforms (like Tensorflow and PyTorch) within traditional modeling frameworks to improve computational efficiency and performance accuracy (Xu et al., 2020); 2) incorporating physics laws into ML-based approaches to improve the interpretability of data-driven models (Raissi et al., 2020). In both cases, ML offers unprecedented opportunities for empowering modeling capability in approximating complex functions. The emergence of the physics-informed ML model (Kashinath et al., 2021) underscores the importance of advancing cutting-edge algorithms.

6.9. Provenance, reproducibility, replicability, & reusability

Four broad and interrelated concerns for Earth AI research include:

- *Provenance:* Where did the training data, AI model, software, and hardware originate, and what transformations have the data undergone before the findings were reported?
- *Reproducibility:* Can an independent party replicate the precise AI workflow and reported results, using the same data and algorithms?
- *Replicability:* Can an independent party run similar (but not identical) ML analyses on similar (but not necessarily the same) data and come to the same conclusions?
- *Reusability:* How easily can the trained AI models be applied to new data or other new situations?

Earth scientists have proposed standards to document the provenance of both data and scientific workflows (Sun et al., 2020a) including ISO 19115:2003 and ISO 19115–2:2009, the Open Provenance Model (Moreau et al., 2008), the data service standards of the Open Geospatial Consortium, and the Provenance Ontology of W3C (Hills et al., 2015; Lebo et al., 2013; Sun et al., 2013; Tilmes et al., 2013; Zhang et al., 2020).

Software like Docker, Helm, Conda/Anaconda-project, Prov, MetaClip, and Geoweaver can be used to record the AI workflow being used so that it can be made available for later retrieval to understand, replicate, reproduce, and reuse the trained AI models. As Earth scientists increasingly embrace open data and managed workflow platforms, the topics of provenance, reusability, replicability, and reproducibility have received increased attention (Gil et al., 2019; Kedron et al., 2021). Provenance is critical for Earth AI models to be understood and trusted by the public, and a standardized provenance framework for AI would be an ideal solution to address these concerns. Another challenge for reusability is ensuring that data used for training and evaluating algorithms are openly accessible (Neylon, 2012; Tenopir et al., 2020). As a step toward more open data, researchers should archive their data in a long-term repository (Duerr et al., 2018). Many of these repositories provide templates and tools to enable the submission of metadata that describes the data being archived and AI practitioners may benefit from guidelines that suggest which files are most important to submit to long-term repositories.

6.10. Full-stack workflow automation

AI engineering is an inclusive discipline involving many technologies, algorithms, tools, libraries, and its product pipelines are composed of a series of links ranging from hardware to software, from a raw data repository to actionable information dissemination, from web services to endpoint software. Manually managing all the portions is unrealistic. Automating all the processing steps are required to make Earth AI practical in real-world scenarios. However, full-stack automation of Earth AI workflows is still under development. To maintain AI adoption and scale, the Earth science community needs a better way to holistically deploy and manage the lifecycle of deployed ML models.

MLOps (ML DevOps) is the process of deploying an experimental ML model into a production web system. It manages the deployment, monitoring, managing, and governing of production-level ML models. There are many opportunities ahead for open source software developers to take on this task. Ongoing projects within the NASA Earth community like Geoweaver (Sun et al., 2020b) already realized this challenge and are working to deliver practically stable software as a solution.

Running efficient and productive Earth AI models requires the collaboration of various entities and resources, and involves various programs, scripts, libraries, software, and platforms from automation of data preparation, indigestion, training, validation, testing, deployment, and production. It requires building a workflow, meaning a logically chained flow of multiple processes to complete a big mission. Workflow orchestration could be conducted in many ways, e.g., writing a Python notebook, a Shell script, or using workflow management software like Cylc. The basic components of workflows are similar. All the workflows have atomic processes and connections among them. Once the workflow is started, all the atomic processes will be executed automatically without asking, which is called workflow automation. There are many workflow management software (WfMS) developing to enable automation, i.e., Apache Airflow, Cylc, Galaxy, Pegasus-WMS, Geoweaver (Sun et al., 2020b), and so forth. These WfMS can not only automate the process but also record the provenance to improve the replicability and reproducibility of Earth AI discovery.

6.11. AI ethics

Earth AI is designed to protect us with an unseen powerful capability of forecasting the Earth's future and navigating natural hazards and resources in advance to save people and conserve the environment. However, the power has a limit and it cannot save everyone equally, for instance, in a geohazard or disruptive event. What if Earth AI miscalculates the situation, misses a region/group, underestimates the harm, and results in more fatalities or greater damage? Earth AI is intelligent but still a lifeless system, which is not a legal entity. Yet its decisions impact society, and it behaves on a certain level of self-will.

There is a wealth of research focused on the ethical problems caused by AI when it is in operation (Jobin et al., 2019). Critics have examined the relationship between the role cultural bias plays in algorithmic inequality (Eubanks, 2018) and how AI systems oppress racial minorities and reinforce existing discrimination (Buolamwini and Gebru, 2018). We can foresee many regulations and laws regarding Earth AI ethics soon. Here, we outline several of the many paths toward more ethical AI in the earth and environmental sciences that include more open datasets and unbiased algorithms. Engineers should develop Earth AI ethics-related logic by partnering with social scientists, ethicists, and philosophers who have been studying the social implications of AI in the domain of policing, law, finance. This includes developing a guideline for ML researchers to engage with ethics as not only a philosophical project but also a pragmatic one where the collection of data and the use of particular models over others have direct impacts on ecosystems and humans. Last but not least, we believe that communicating one's application of any ML or AI application to the broader community it impacts (for example, if an automated method for developing land cover maps will directly impact on representations of Indigenous land) will be necessary for achieving a fair and ethical movement in AI in geosciences.

6.12. Operation management

Operationalizing AI service cannot be simply fulfilled by one scientist or one small Earth research group. AI products need maintainers and customer service after being deployed. A large corporation could produce gigantic volumes of business and log data. Transition to and sustainment of AI operations is complicated by the rapid pace of technological evolution. However, DevOps practices, which emphasize close coordination between developers and operations, can mitigate these difficulties, and even provide useful AI feedback from operations into model evolution in some cases. Another potentially effective technique is internal capacity building, such as training the operations staff in the basics of the AI technology in use, so that they can better recognize issues and provide support to customers.

7. Summary

Focusing on applications to geosciences, this paper overviews the cutting-edge technology and progress of AI research. Breakthroughs in Earth AI theory and infrastructure will carry geoscience into the next phase: Earth AI. The geoscience community must catch up with the pace of exploding observational datasets and quickly build useable AI models at an affordable cost promptly with adequate accuracy. The research and development of Earth AI are still at the infancy stage, and all the grand challenges ranging from data to model to operation can derive numerous opportunities in all sectors from academia to government and industry. The future of Earth AI is bright and dramatically beneficial to the entire human society and Earth system and should advance our civilization into its next epic phase and transform the Earth into a more sustainable, healthy planet.

Funding sources

This work is sponsored by NASA ACCESS (#80NSSC21M0028 and #80NSSC21M0027, 2020), DOE (DE-AC02-05CH11231), NSF EPSCoR (#2019609, 2020), NSF Geoinformatics program (EAR-1947893 & EAR-1947875, 2020), NASA Health and Air Quality project (#80NSSC21K0512, 2018), NASA PO program (#80NM0018D0004), NSF Cybertraining (#2117834, 2021), NSF CSSI (#1835717, 2018), NSF EarthCube (#2126315, 2021). **This paper has gone through NASA internal review and been cleared for publication.**

8. Authorship contribution statement

Ziheng Sun drafted the initial version, coordinated and contributed to the edits of all the sections, and prepared the manuscript; Laura Sandoval contributed to section 6.5, 6.1 to 6.4, coordinated the edits to all the sections and prepared the manuscript; Rob Crystal-Ornelas contributed to section 4.1, 6.2, 6.5, 6.9, 6.11, and did major editorial changes to the paper; S. Mostafa Mousavi contributed to section 2.1, 3.5, 4.1, and 6.8; Thomas E. Gill contributed to section 2.3 and 6.8 and did a thorough editorial improvement on the entire paper; Jinbo Wang contributed to section 2.2 and 2.6; Cindy Lin contributed to section 2.3, 6.2, 6.3, and 6.11; Javier Orduz contributed to section 2.1, 5.5, and 6.8; Peisheng Zhao contributed to section 4.1, 5.1, 6.1, 6.2, and 6.4; Chandana Gangodagemage contributed to section 2.1, 2.2, 5.1, and 6.1; James Bednar contributed to section 5.1 and 6.9, Pablo Rivas contributed to section 6.5 and 6.11; Wendy Carande contributed to section 6.2, 6.3, 6.5, 6.6, 6.7, 6.8, 6.9, 6.10, and 6.11; Amanda Tan contributed to section 6.8, 6.10, and 6.11; Jianwu Wang contributed to section 5.3 and 6.4; Benjamin Holt contributed to section 2.6; Sanjay Purushotham contributed to section 6.5 and 2.3; Nicoleta Cristea contributed to section 1, 2.2, 2.5 and 3.4; Daniel Tong contributed to section 2.3, 3.1, 4.4, and 6.8; Daniel Howard contributed to section 5.2, 5.3, 5.5, and 6.4; Julien Chastang contributed to section 4.2 and 6.9; Yuhan Rao contributed to section 3.2, 3.3, 6.6, and 6.11, Xiaogang Ma contributed to section 6.5 and 6.9; Zack Chester contributed to section 2.3 and 5.5; Aji John contributed to section 2.2, 4.1, and 4.3.

Computer code availability

Not applicable.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This is a community effort led by ESIP Machine Learning Cluster and NASA Earth Science Data System Working Groups. Great thanks to all the colleagues in ESDSWG and ESIP who have provided their valuable time and advice and editing work during the writing of this paper. Special thanks to Dr. Christopher Lynnes, Dr. Justin Rice, Mr. Stephen Olding, Dr. Annie Burgess, Ms. Megan Carter, Mr. Dave Jones, Ms. Karen Moe, Dr. Karthik Kashinath, for all the kind support. The contribution by J. Wang and B. Holt was done at Caltech JPL under NASA contract (80NM0018D0004, PO.DAAC). RCO is funded through the ESS-DIVE repository by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Earth and Environmental Sciences Division, Data Management program under contract number DE-AC02-05CH11231.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., 2021. A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges. Information Fusion.

Abdul-Wahab, S.A., Al-Alawi, S.M., 2002. Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. Environ. Model. Software 17, 219–228.

Adeli, H., Panakkat, A., 2009. A probabilistic neural network for earthquake magnitude prediction. Neural Network. 22, 1018–1024.

Agana, N., Homaifar, A., 2018. EMD-based predictive deep belief network for time series prediction: an application to drought forecasting. Hydrology 5, 18.

Alemany, S., Beltran, J., Perez, A., Ganzfried, S., 2019. Predicting hurricane trajectories using a recurrent neural network. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 468–475.

Ali, M., Jagadeesh, P.V., Jain, S., 2007. Effects of eddies on Bay of Bengal cyclone intensity. Eos, Transactions American Geophysical Union 88, 93–95.

Amani, M., Ghorbanian, A., Ahmadi, S.A., Kakooei, M., Moghimi, A., Mirmazloumi, S.M., Moghaddam, S.H.A., Mahdavi, S., Ghahremanloo, M., Parsian, S., 2020. Google earth engine cloud computing platform for remote sensing big data applications: a comprehensive review. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens.

Anctil, F., Tape, D.G., 2004. An exploration of artificial neural network rainfall-runoff forecasting combined with wavelet decomposition. J. Environ. Eng. Sci. 3, S121–S128.

Ariew, R., 1976. OCKHAM'S RAZOR: A HISTORICAL AND PHILOSOPHICAL ANALYSIS OF OCKHAM'S PRINCIPLE OF PARSIMONY. University of Illinois at Urbana-Champaign.

Ashley, W.S., Strader, S., Dziubla, D.C., Haberlie, A., 2015. Driving blind: weather-related vision hazards and fatal motor vehicle crashes. Bull. Am. Meteorol. Soc. 96, 755–778.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39, 2481–2495.

Bao, S., Zhang, R., Wang, H., Yan, H., Yu, Y., Chen, J., 2019. Salinity profile estimation in the Pacific Ocean from satellite surface salinity observations. J. Atmos. Ocean. Technol. 36, 53–68.

Barzegar, R., Moghaddam, A.A., Deo, R., Fijani, E., Tziritis, E., 2018. Mapping groundwater contamination risk of multiple aquifers using multi-model ensemble of machine learning algorithms. Sci. Total Environ. 621, 697–712.

Baumhoer, C.A., Dietz, A.J., Kneisel, C., Kuenzer, C., 2019. Automated extraction of antarctic glacier and ice shelf fronts from sentinel-1 imagery using deep learning. Rem. Sens. 11, 2529.

Belayneh, A., Adamowski, J., 2012. Standard precipitation index drought forecasting using neural networks, wavelet neural networks, and support vector regression. Applied computational intelligence and soft computing 6, 2012.

Belayneh, A., Adamowski, J., Khalil, B., Quilty, J., 2016. Coupling machine learning methods with wavelet transforms and the bootstrap and boosting ensemble approaches for drought prediction. Atmos. Res. 172, 37–47.

Beyreuther, M., Carniel, R., Wassermann, J., 2008. Continuous hidden Markov models: application to automatic earthquake detection and classification at Las Canãdas caldera, Tenerife. J. Volcanol. Geoth. Res. 176, 513–518.
Bhattacharya, B., Solomatine, D.P., 2005. Neural networks and M5 model trees in modelling water level–discharge relationship. Neurocomputing 63, 381–396.
Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., Lloyd, S., 2017. Quantum machine learning. Nature 549, 195–202.
Bisong, E., 2019. Google Colaboratory, Building Machine Learning and Deep Learning Models on Google Cloud Platform. Springer, pp. 59–64.
Biwer, C.M., Capano, C.D., De, S., Cabero, M., Brown, D.A., Nitz, A.H., Raymond, V., 2019. PyCBC Inference: A Python-Based Parameter Estimation Toolkit for Compact Binary Coalescence Signals, 131. Publications of the Astronomical Society of the Pacific, 024503.
Boloorani, A.D., Samany, N.N., Papi, R., Soleimani, M., 2022. Dust source susceptibility mapping in Tigris and Euphrates basin using remotely sensed imagery. Catena 209, 105795.
Bolton, T., Zanna, L., 2019. Applications of deep learning to ocean data inference and subgrid parameterization. J. Adv. Model. Earth Syst. 11, 376–399.
Boroughani, M., Pourhashemi, S., Hashemi, H., Salehi, M., Amirahmadi, A., Asadi, M.A. Z., Berndtsson, R., 2020. Application of remote sensing techniques and machine learning algorithms in dust source detection and dust source susceptibility mapping. Ecol. Inf. 56, 101059.
Bose, S., Das, K., Arima, M., 2008. Multiple stages of melting and melt-solid interaction in the lower crust: new evidence from UHT granulites of Eastern Ghats Belt, India. J. Mineral. Petrol. Sci. 103, 266–272.
Boukabara, S.-A., Krasnopolsky, V., Penny, S.G., Stewart, J.Q., McGovern, A., Hall, D., Ten Hoeve, J.E., Hickey, J., Allen Huang, H.-L., Williams, J.K., 2021. Outlook for exploiting artificial intelligence in the earth and environmental sciences. Bull. Am. Meteorol. Soc. 102, E1016–E1032.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Brewer, C.K., Winne, J.C., Redmond, R.L., Opitz, D.W., Mangrich, M.V., 2005. Classifying and mapping wildfire severity. Photogramm. Eng. Rem. Sens. 71, 1311–1320.
Browning, E., Bolton, M., Owen, E., Shoji, A., Guilford, T., Freeman, R., 2018. Predicting animal behaviour using deep learning: GPS data alone accurately predict diving in seabirds. Methods in Ecology and Evolution 9, 681–692.
Buolamwini, J., Gebru, T., 2018. In: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, Conference on Fairness, Accountability and Transparency. PMLR, pp. 77–91.
Campos-Taberner, M., Moreno-Martínez, Á., García-Haro, F.J., Camps-Valls, G., Robinson, N.P., Kattge, J., Running, S.W., 2018. Global estimation of biophysical variables from Google Earth Engine platform. Rem. Sens. 10, 1167.
Cannata, A., Cannavò, F., Moschella, S., Gresta, S., Spina, L., 2019. Exploring the link between microseism and sea ice in Antarctica by using machine learning. Sci. Rep. 9, 1–15.
Cao, Q.D., Choe, Y., 2020. Building damage annotation on post-hurricane satellite imagery based on convolutional neural networks. Nat. Hazards 103, 3357–3376.
Chapman, C., Charantonis, A.A., 2017. Reconstruction of subsurface velocities from satellite observations using iterative self-organizing maps. Geosci. Rem. Sens. Lett. IEEE 14, 617–620.
Chelani, A.B., 2010. Prediction of daily maximum ground ozone concentration using support vector machine. Environ. Monit. Assess. 162, 169–176.
Chelton, D.B., Schlax, M.G., Samelson, R.M., 2011. Global observations of nonlinear mesoscale eddies. Prog. Oceanogr. 91, 167–216.
Chen, G., Han, T.X., He, Z., Kays, R., Forrester, T., 2014. In: Deep Convolutional Neural Network Based Species Recognition for Wild Animal Monitoring, 2014 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 858–862.
Chen, J., Jin, Q., Chao, J., 2012. Design of deep belief networks for short-term prediction of drought index using data in the Huaihe river basin. Math. Probl Eng. 2012.
Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40, 834–848.
Cheng, H., Sun, L., Li, J., 2021. Neural network approach to retrieving ocean subsurface temperatures from surface parameters observed by satellites. Water 13, 388.
Clancey, William J., 1983. The epistemology of a rule-based expert system—a framework for explanation. Artif. Intell. 20, 215–251.
Dempsey, D., Cronin, S.J., Mei, S., Kempa-Liehr, A.W., 2020. Automatic precursor recognition and real-time forecasting of sudden explosive volcanic eruptions at Whakaari, New Zealand. Nat. Commun. 11, 1–8.
Deo, R.C., Şahin, M., 2015. Application of the extreme learning machine algorithm for the prediction of monthly Effective Drought Index in eastern Australia. Atmos. Res. 153, 512–525.
Deutsch, D., 1985. Quantum theory, the Church–Turing principle and the universal quantum computer. Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences 400, 97–117.
Du, S., Li, T., Yang, Y., Horng, S.-J., 2018. Deep Air Quality Forecasting Using Hybrid Deep Learning Framework arXiv preprint arXiv:1812.04783.
Du, Y., Song, W., He, Q., Huang, D., Liotta, A., Su, C., 2019. Deep learning with multi-scale feature fusion in remote sensing for automatic oceanic eddy detection. Inf. Fusion 49, 89–99.
Duerr, R., Kinkade, D., Witt, M., Yarmey, L., 2018. Data Repository Selection Decision Tree for Researchers in the Earth, Space, and Environmental Sciences. Enabling FAIR Data Community.
Durand, M., Fu, L.-L., Lettenmaier, D.P., Alsdorf, D.E., Rodriguez, E., Esteban-Fernandez, D., 2010. The surface water and ocean topography mission: observing terrestrial surface water and oceanic submesoscale eddies. Proc. IEEE 98, 766–779.
Egli, A., Schrenzel, J., Greub, G., 2020. Digital Microbiology. Clinical Microbiology and Infection.
ESIP, 2021. Awesome Earth Artificial Intelligence.
Eubanks, V., 2018. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. s Press, St. Martin.
Fairbrass, A.J., Firman, M., Williams, C., Brostow, G.J., Titheridge, H., Jones, K.E., 2019. CityNet—deep learning tools for urban ecoacoustic assessment. Methods in Ecology and Evolution 10, 186–197.
Fan, J., Li, Q., Hou, J., Feng, X., Karimian, H., Lin, S., 2017. A spatiotemporal prediction framework for air pollution based on deep RNN. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 4, 15.
Fathian, F., Mehdizadeh, S., Sales, A.K., Safari, M.J.S., 2019. Hybrid models to improve the monthly river flow prediction: integrating artificial intelligence and non-linear time series models. J. Hydrol. 575, 1200–1213.
Ferran, A., Bernabe, S., Rodriguez, P.G., Plaza, A., 2013. A web-based system for classification of remote sensing data. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 6, 1934–1948.
Ferrenberg, A., Swendsen, R., 1989. Optimized Monte Carlo data analysis. Comput. Phys. 3, 101–104. https://doi.org/10.1063/1.4822862.
Feynman, R.P., 2018. Simulating Physics with Computers, Feynman and Computation. CRC Press, pp. 133–153.
Franz, K., Roscher, R., Milioto, A., Wenzel, S., Kusche, J., 2018. In: Ocean Eddy Identification and Tracking Using Neural Networks, IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 6887–6890.
Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 119–139.
Gal, Y., 2016. Uncertainty in Deep Learning. PhD Thesis. University of Cambridge.
Garcia-Pineda, O., MacDonald, I.R., Li, X., Jackson, C.R., Pichel, W.G., 2013. Oil spill mapping and measurement in the Gulf of Mexico with textural classifier neural network algorithm (TCNNA). IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 6, 2517–2525.
Garcia-Pineda, O., Zimmer, B., Howard, M., Pichel, W., Li, X., MacDonald, I.R., 2009. Using SAR images to delineate ocean oil slicks with a texture-classifying neural network algorithm (TCNNA). Can. J. Rem. Sens. 35, 411–421.
George, T.M., Manucharyan, G.E., Thompson, A.F., 2021. Deep learning to infer eddy heat fluxes from sea surface height patterns of mesoscale turbulence. Nat. Commun. 12, 1–11.
Ghahramani, Z., 2019. Probabilistic Modelling, Machine Learning, and the Information Revolution. MIT Computer Science and Artificial Intelligence Lab. http://mlg.eng.cam.ac.uk/zoubin/talks/mit12csail.pdf.
Ghanbari, M., Clausi, D.A., Xu, L., Jiang, M., 2019. Contextual classification of sea-ice types using compact polarimetric SAR data. IEEE Trans. Geosci. Rem. Sens. 57, 7476–7491.
Gholami, V., Sahour, H., Amri, M.A.H., 2021. Soil erosion modeling using erosion pins and artificial neural networks. Catena 196, 104902.
Ghorbanzadeh, O., Blaschke, T., Gholamnia, K., Meena, S.R., Tiede, D., Aryal, J., 2019. Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. Rem. Sens. 11, 196.
Giffard-Roisin, S., Yang, M., Charpiat, G., Kégl, B., Monteleoni, C., 2018. In: Fused Deep Learning for Hurricane Track Forecast from Reanalysis Data. Proceedings of 8th International Workshop on Climate Informatics, Boulder, Colorado, USA.
Gil, Y., Honaker, J., Gupta, S., Ma, Y., D'Orazio, V., Garijo, D., Gadewar, S., Yang, Q., Jahanshad, N., 2019. Towards human-guided machine learning. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 614–624.
Girshick, R., 2015. Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448.
Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google earth engine: planetary-scale geospatial analysis for everyone. Rem. Sens. Environ. 202, 18–27.
Gurney, K., 2014. An Introduction to Neural Networks. CRC press.
Han, J., Zhang, L., 2020. Integrating Machine Learning with Physics-Based Modeling arXiv preprint arXiv:2006.02619.
Han, M., Feng, Y., Zhao, X., Sun, C., Hong, F., Liu, C., 2019. A convolutional neural network using surface data to predict subsurface temperatures in the Pacific Ocean. IEEE Access 7, 172816–172829.
He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. In: Mask R-Cnn, Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, pp. 2980–2988.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.
Henley, W., Hand, D.J., 1996. A k-nearest-neighbour classifier for assessing consumer credit risk. The Statistician 77–95.
Hibert, C., Mangeney, A., Grandjean, G., Baillard, C., Rivet, D., Shapiro, N.M., Satriano, C., Maggi, A., Boissier, P., Ferrazzini, V., 2014. Automated identification, location, and volume estimation of rockfalls at Piton de la Fournaise volcano. J. Geophys. Res.: Earth Surf. 119, 1082–1105.
Hills, D., Downs, R.R., Duerr, R., Goldstein, J.C., Parsons, M.A., Ramapriyan, H.K., 2015. The importance of data set provenance for science. Eos 96.
Ho, C.-S., Jean, N., Hogan, C.A., Blackmon, L., Jeffrey, S.S., Holodniy, M., Banaei, N., Saleh, A.A., Ermon, S., Dionne, J., 2019. Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. Nat. Commun. 10, 1–8.
Holt, B., Johnson, M.P., Perkovic-Martin, D., Panzer, B., 2015. Snow depth on Arctic sea ice derived from radar: in situ comparisons and time series analysis. J. Geophys. Res.: Oceans 120, 4260–4287.
Hong, W.-C., 2008. Rainfall forecasting by technological machine learning models. Appl. Math. Comput. 200, 41–57.

Hu, C., Wu, Q., Li, H., Jian, S., Li, N., Lou, Z., 2018. Deep learning with a long short-term memory networks approach for rainfall-runoff simulation. Water 10, 1543.
Huang, D., Du, Y., He, Q., Song, W., Liotta, A., 2017. DeepEddy: a simple deep architecture for mesoscale oceanic eddy detection in SAR images. In: 2017 IEEE 14th International Conference on Networking, Sensing and Control (ICNSC). IEEE, pp. 673–678.
Hüllermeier, E., Waegeman, W., 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach. Learn. 110, 457–506.
Hutchinson, D.K., Coxall, H.K., Lunt, D.J., Steinthorsdottir, M., de Boer, A.M., Baatsen, M., von der Heydt, A., Huber, M., Kennedy-Asser, A.T., Kunzmann, L., 2021. The Eocene–Oligocene transition: a review of marine and terrestrial proxy data, models and model–data comparisons. Clim. Past 17, 269–315.
Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., Keutzer, K., 2014. Densenet: Implementing Efficient Convnet Descriptor Pyramids arXiv preprint arXiv: 1404.1869.
Ingsrisawang, L., Ingsriswang, S., Somchit, S., Aungsuratana, P., Khantiyanan, W., 2008. Machine learning techniques for short-term rain forecasting system in the northeastern part of Thailand. In: Proceedings of World Academy of Science, Engineering and Technology. Citeseer.
Jain, P., Coogan, S.C., Subramanian, S.G., Crowley, M., Taylor, S., Flannigan, M.D., 2020. A review of machine learning applications in wildfire science and management. Environ. Rev. 28, 478–505.
Januschowski, T., Arpin, D., Salinas, D., Flunkert, V., Gasthaus, J., Stella, L., Vazquez, P., 2018. Now Available in Amazon Sagemaker: Deepar Algorithm for More Accurate Time Series Forecasting.
Jin, J., Segers, A., Liao, H., Heemink, A., Kranenburg, R., Lin, H.X., 2020. Source backtracking for dust storm emission inversion using an adjoint method: case study of Northeast China. Atmos. Chem. Phys. 20, 15207–15225.
Jobin, A., Ienca, M., Vayena, E., 2019. The global landscape of AI ethics guidelines. Nature Machine Intelligence 1, 389–399.
Johnsirani Venkatesan, N., Nam, C., Shin, D.R., 2019. Deep learning frameworks on Apache spark: a review. IETE Tech. Rev. 36, 164–177.
Kamilaris, A., Kartakoullis, A., Prenafeta-Boldú, F.X., 2017. A review on the practice of big data analysis in agriculture. Comput. Electron. Agric. 143, 23–37.
Karasözen, E., Karasözen, B., 2020. Earthquake location methods. GEM-International Journal on Geomathematics 11, 1–28.
Kashinath, K., Mustafa, M., Albert, A., Wu, J., Jiang, C., Esmaeilzadeh, S., Azizzadenesheli, K., Wang, R., Chattopadhyay, A., Singh, A., 2021. Physics-informed machine learning: case studies for weather and climate modelling. Philosophical Transactions of the Royal Society A 379, 20200093.
Kedron, P., Frazier, A.E., Trgovac, A.B., Nelson, T., Fotheringham, A.S., 2021. Reproducibility and replicability in geographical analysis. Geogr. Anal. 53, 135–147.
Kim, S., Kim, H., Lee, J., Yoon, S., Kahou, S.E., Kashinath, K., Prabhat, M., 2019. Deep-hurricane-tracker: tracking and forecasting extreme climate events. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 1761–1769.
Kim, Y.H., Im, J., Ha, H.K., Choi, J.-K., Ha, S., 2014. Machine learning approaches to coastal water quality monitoring using GOCI satellite data. GIScience Remote Sens. 51, 158–174.
Kloster, M., Langenkämper, D., Zurowietz, M., Beszteri, B., Nattkemper, T.W., 2020. Deep learning-based diatom taxonomy on virtual slides. Sci. Rep. 10, 1–13.
Klotzbach, P.J., Bowen, S.G., Pielke, R., Bell, M., 2018. Continental US hurricane landfall frequency and associated damage: observations and future risks. Bull. Am. Meteorol. Soc. 99, 1359–1376.
Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B.E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J.B., Grout, J., Corlay, S., 2016. Jupyter Notebooks-A Publishing Format for Reproducible Computational Workflows, ELPUB, pp. 87–90.
Knoll, L., Breuer, L., Bach, M., 2019. Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. Sci. Total Environ. 668, 1317–1327.
Kolehmainen, M., Martikainen, H., Ruuskanen, J., 2001. Neural networks and periodic components used in air quality forecasting. Atmos. Environ. 35, 815–825.
Koroteev, D., Tekic, Z., 2021. Artificial intelligence in oil and gas upstream: trends, challenges, and scenarios for the future. Energy and AI 3, 100041.
Korup, O., Stolle, A., 2014. Landslide prediction from machine learning. Geol. Today 30, 26–33.
Kosov, S., Shirahama, K., Li, C., Grzegorzek, M., 2018. Environmental microorganism classification using conditional random fields and deep convolutional neural networks. Pattern Recogn. 77, 248–261.
Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall–runoff modelling using long short-term memory (LSTM) networks. Hydrol. Earth Syst. Sci. 22, 6005–6022.
Krishnan, M., 2019. Against Interpretability: a Critical Examination of the Interpretability Problem in Machine Learning. Philosophy & Technology, pp. 1–16.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 1097–1105.
Kühnlein, M., Appelhans, T., Thies, B., Nauss, T., 2014. Improving the accuracy of rainfall rates from optical satellite sensors with machine learning—a random forests-based approach applied to MSG SEVIRI. Rem. Sens. Environ. 141, 129–143.
Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J., 2013. Prov-o: the Prov Ontology.
Lee, J., Shi, Y.R., Cai, C., Ciren, P., Wang, J., Gangopadhyay, A., Zhang, Z., 2021. Machine learning based algorithms for global dust aerosol detection from satellite images: inter-comparisons and evaluation. Rem. Sens. 13, 456.
Lee, S., Im, J., Kim, J., Kim, M., Shin, M., Kim, H.-c., Quackenbush, L., 2016. Arctic sea ice thickness estimation from CryoSat-2 satellite data using machine learning-based lead detection. Rem. Sens. 8, 698.
Leigh, S., Wang, Z., Clausi, D.A., 2013. Automated ice–water classification using dual polarization SAR satellite imagery. IEEE Trans. Geosci. Rem. Sens. 52, 5529–5539.
Li, G., Rabe, K.S., Nielsen, J., Engqvist, M.K., 2019. Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima. ACS Synth. Biol. 8, 1411–1420.
Li, H., Ota, K., Dong, M., 2018. Learning IoT in edge: deep learning for the Internet of Things with edge computing. IEEE network 32, 96–101.
Lin, G.F., Chen, G.R., Wu, M.C., Chou, Y.C., 2009. Effective forecasting of hourly typhoon rainfall using support vector machines. Water Resour. Res. 45, W08440.
Lin, J.-W., Chiou, J.-S., 2019. Active probability backpropagation neural network model for monthly prediction of probabilistic seismic hazard analysis in Taiwan. IEEE Access 7, 108990–109014.
Lin, X., Chang, H., Wang, K., Zhang, G., Meng, G., 2020. Machine learning for source identification of dust on the Chinese Loess Plateau. Geophys. Res. Lett. 47, e2020GL088950.
Liu, F., Zhou, H., Wen, B., 2021. DEDNet: offshore eddy detection and location with HF radar by deep learning. Sensors 21, 126.
Lohse, J., Doulgeris, A.P., Dierking, W., 2019. An optimal decision-tree design strategy and its application to sea ice classification from sar imagery. Rem. Sens. 11, 1574.
Mac Aodha, O., Gibb, R., Barlow, K.E., Browning, E., Firman, M., Freeman, R., Harder, B., Kinsey, L., Mead, G.R., Newson, S.E., 2018. Bat detective—deep learning tools for bat acoustic signal detection. PLoS Comput. Biol. 14, e1005995.
Maca, P., Pech, P., 2016. Forecasting SPEI and SPI drought indices using the integrated artificial neural networks. Comput. Intell. Neurosci. 14, 2016.
Maekawa, T., Ohara, K., Zhang, Y., Fukutomi, M., Matsumoto, S., Matsumura, K., Shidara, H., Yamazaki, S.J., Fujisawa, R., Ide, K., 2020. Deep learning-assisted comparative analysis of animal trajectories with DeepHL. Nat. Commun. 11, 1–15.
Manucharyan, G.E., Siegelman, L., Klein, P., 2020. A Deep Learning approach to spatiotemporal SSH interpolation and estimation of deep currents in geostrophic ocean turbulence. J. Adv. Model. Earth Syst., e2019MS001965
Maskey, M., Ramachandran, R., Miller, J.J., Zhang, J., Gurung, I., 2018. In: Earth Science Deep Learning: Applications and Lessons Learned, IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 1760–1763.
Masotti, M., Falsaperla, S., Langer, H., Spampinato, S., Campanini, R., 2006. Application of support vector machine to the classification of volcanic tremor at etna, Italy. Geophys. Res. Lett. 33, L20304.
Mayer-Schönberger, V., Cukier, K., 2013. Big Data: A Revolution that Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt.
Mignan, A., Broccardo, M., 2020. Neural network applications in earthquake prediction (1994–2019): meta-analytic and statistical insights on their limitations. Seismol Res. Lett. 91, 2330–2342.
Mo, S., Zabaras, N., Shi, X., Wu, J., 2019. Deep autoregressive neural networks for high-dimensional inverse problems in groundwater contaminant source identification. Water Resour. Res. 55, 3856–3881.
Mohajerani, Y., Wood, M., Velicogna, I., Rignot, E., 2019. Detection of glacier calving margins with convolutional neural networks: a case study. Rem. Sens. 11, 74.
Mohanty, S.P., Hughes, D.P., Salathé, M., 2016. Using deep learning for image-based plant disease detection. Front. Plant Sci. 7, 1419.
Molnar, C., Casalicchio, G., Bischl, B., 2020. Interpretable Machine Learning–A Brief History, State-Of-The-Art and Challenges arXiv preprint arXiv:2010.09337.
Mons, Barend, 2020. Invest 5% of research funds in ensuring data are reusable. Nature 578, 491-491.
Moreau, L., Freire, J., Futrelle, J., McGrath, R.E., Myers, J., Paulson, P., 2008. The Open Provenance Model: an Overview, International Provenance and Annotation Workshop. Springer, pp. 323–326.
Moschos, E., Stegner, A., Schwander, O., Gallinari, P., 2020. Classification of eddy sea surface temperature signatures under cloud coverage. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 13, 3437–3447.
Mousavi, S.M., Beroza, G.C., 2018. Evaluating the 2016 one-year seismic hazard model for the central and eastern United States using instrumental ground-motion data. Seismol Res. Lett. 89, 1185–1196.
Mousavi, S.M., Ellsworth, W.L., Zhu, W., Chuang, L.Y., Beroza, G.C., 2020. Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. Nat. Commun. 11, 1–12.
Mousavi, S.M., Zhu, W., Sheng, Y., Beroza, G.C., 2019. CRED: a deep residual network of convolutional and recurrent units for earthquake signal detection. Sci. Rep. 9, 1–14.
Mousavi, S.Z., Kavian, A., Soleimani, K., Mousavi, S.R., Shirzadi, A., 2011. GIS-based spatial prediction of landslide susceptibility using logistic regression model. Geomatics, Nat. Hazards Risk 2, 33–50.
Naganna, S.R., Deka, P.C., Ghorbani, M.A., Biazar, S.M., Al-Ansari, N., Yaseen, Z.M., 2019. Dew point temperature estimation: application of artificial intelligence model integrated with nature-inspired optimization algorithms. Water 11, 742.
Naghibi, S.A., Ahmadi, K., Daneshi, A., 2017. Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. Water Resour. Manag. 31, 2761–2775.
Neethirajan, S., 2020. The Role of Sensors, Big Data and Machine Learning in Modern Animal Farming. Sensing and Bio-Sensing Research, p. 100367.
Nefeslioglu, H.A., Gokceoglu, C., Sonmez, H., 2008. An assessment on the use of logistic regression and artificial neural networks with different sampling strategies for the preparation of landslide susceptibility maps. Eng. Geol. 97, 171–191.
Neylon, C., 2012. Open access must enable open use. Nature 492, 348–349.

Nolan, B.T., Fienen, M.N., Lorenz, D.L., 2015. A statistical learning framework for groundwater nitrate models of the Central Valley, California, USA. J. Hydrol. 531, 902–911.
Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C., Clune, J., 2018. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proc. Natl. Acad. Sci. Unit. States Am. 115, E5716–E5725.
Nowack, P., Braesicke, P., Haigh, J., Abraham, N.L., Pyle, J., Voulgarakis, A., 2018. Using machine learning to build temperature-based ozone parameterizations for climate sensitivity simulations. Environ. Res. Lett. 13, 104016.
Numpy, 2020. Case Study: First Image of a Black Hole.
NWS, 2021. NWS High Water Level Terminology.
Olyaie, E., Banejad, H., Chau, K.-W., Melesse, A.M., 2015. A comparison of various artificial intelligence approaches performance for estimating suspended sediment load of river systems: a case study in United States. Environ. Monit. Assess. 187, 1–22.
Overpeck, J.T., Meehl, G.A., Bony, S., Easterling, D.R., 2011. Climate data challenges in the 21st century. Science 331, 700–702.
Park, J.-W., Korosov, A.A., Babiker, M., Won, J.-S., Hansen, M.W., Kim, H.-C., 2020. Classification of sea ice types in Sentinel-1 synthetic aperture radar images. Cryosphere 14, 2629–2645.
Pearl, J., 1988. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan kaufmann.
Pedraza, A., Bueno, G., Deniz, O., Cristóbal, G., Blanco, S., Borrego-Ramos, M., 2017. Automated diatom classification (Part B): a deep learning approach. Appl. Sci. 7, 460.
Peng, L., Niu, R., Huang, B., Wu, X., Zhao, Y., Ye, R., 2014. Landslide susceptibility mapping based on rough set theory and support vector machines: a case of the Three Gorges area, China. Geomorphology 204, 287–301.
Perea, R.G., Poyato, E.C., Montesinos, P., Díaz, J.A.R., 2019. Optimisation of water demand forecasting by artificial intelligence with short data sets. Biosyst. Eng. 177, 59–66.
Qi, Z., Wang, T., Song, G., Hu, W., Li, X., Zhang, Z., 2018. Deep air learning: interpolation, prediction, and feature analysis of fine-grained air quality. IEEE Trans. Knowl. Data Eng. 30, 2285–2297.
Qiu, J., Wu, Q., Ding, G., Xu, Y., Feng, S., 2016. A survey of machine learning for big data processing. EURASIP J. Appl. Signal Process. 67, 2016.
Raissi, M., Yazdani, A., Karniadakis, G.E., 2020. Hidden fluid mechanics: learning velocity and pressure fields from flow visualizations. Science 367, 1026–1030.
Redmon, J., Farhadi, A., 2018. Yolov3: an Incremental Improvement arXiv preprint arXiv:1804.02767.
Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195.
Reid, C.E., Brauer, M., Johnston, F.H., Jerrett, M., Balmes, J.R., Elliott, C.T., 2016. Critical review of health impacts of wildfire smoke exposure. Environ. Health Perspect. 124, 1334–1343.
Rhee, J., Im, J., 2017. Meteorological drought forecasting for ungauged areas based on machine learning: using long-range climate forecast and remote sensing data. Agric. For. Meteorol. 237, 105–122.
Rodrigues, M., de la Riva, J., 2014. An insight into machine-learning algorithms to model human-caused wildfire occurrence. Environ. Model. Software 57, 192–201.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.
Ruiz-Suarez, J., Mayora-Ibarra, O., Torres-Jimenez, J., Ruiz-Suarez, L., 1995. Short-term ozone forecasting by artificial neural networks. Adv. Eng. Software 23, 143–149.
Sachindra, D., Ahmed, K., Rashid, M.M., Shahid, S., Perera, B., 2018. Statistical downscaling of precipitation using machine learning techniques. Atmos. Res. 212, 240–258.
Sadri, S., Burn, D., 2012. Nonparametric methods for drought severity estimation at ungauged sites. Water Resour. Res. 48.
Safavian, S.R., Landgrebe, D., 1991. A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics 21, 660–674.
Sahoo, S., Russo, T.A., Elliott, J., Foster, I., 2017. Machine learning algorithms for modeling groundwater level changes in agricultural regions of the US. Water Resour. Res. 53, 3878–3895.
Sajedi-Hosseini, F., Malekian, A., Choubin, B., Rahmati, O., Cipullo, S., Coulon, F., Pradhan, B., 2018. A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. Sci. Total Environ. 644, 954–962.
Saliu, O., Curilla, D., Lennon, M., Chung, A., 2020. In: Lessons Learned: Deep Learning for Mineral Exploration, First EAGE Conference on Machine Learning in Americas. European Association of Geoscientists & Engineers, 1-1.
Sansone, S.-A., McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A.L., Thurston, M., 2019. FAIRsharing as a community approach to standards, repositories and policies. Nat. Biotechnol. 37, 358–367.
Santana, O.J., Hernández-Sosa, D., Martz, J., Smith, R.N., 2020. Neural network training for the detection and classification of oceanic mesoscale eddies. Rem. Sens. 12, 2625.
Sattar, A.M., Ertuğrul, Ö.F., Gharabaghi, B., McBean, E.A., Cao, J., 2019. Extreme learning machine model for water network management. Neural Comput. Appl. 31, 157–169.
Sayad, Y.O., Mousannif, H., Al Moatassime, H., 2019. Predictive modeling of wildfires: a new dataset and machine learning approach. Fire Saf. J. 104, 130–146.
Scarpetta, S., Giudicepietro, F., Ezin, E.C., Petrosino, S., Del Pezzo, E., Martini, M., Marinaro, M., 2005. Automatic classification of seismic signals at Mt. Vesuvius volcano, Italy, using neural networks. Bull. Seismol. Soc. Am. 95, 185–196.
Seo, Y., Kim, S., Kisi, O., Singh, V.P., 2015. Daily water level forecasting using wavelet decomposition and artificial intelligence techniques. J. Hydrol. 520, 224–243.
Shah, S., Hosseini, M., Miled, Z.B., Shafer, R., Berube, S., 2018. A water demand prediction model for central Indiana. In: Proceedings of the AAAI Conference on Artificial Intelligence.
Shams, S., Goswami, S., Lee, K., 2019. In: Deep Learning-Based Spatial Analytics for Disaster-Related Tweets: an Experimental Study, 2019 20th IEEE International Conference on Mobile Data Management (MDM). IEEE, pp. 337–342.
Sharaf El Din, E., Zhang, Y., Suliman, A., 2017. Mapping concentrations of surface water quality parameters using a novel remote sensing and artificial intelligence framework. Int. J. Rem. Sens. 38, 1023–1042.
Shen, C., 2018. A transdisciplinary review of deep learning research and its relevance for water resources scientists. Water Resour. Res. 54, 8558–8593.
Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-c., 2015. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. Adv. Neural Inf. Process. Syst. 802–810.
Sivapragasam, C., Liong, S.-Y., Pasha, M., 2001. Rainfall and runoff forecasting with SSA–SVM approach. J. Hydroinf. 3, 141–152.
Su, H., Li, W., Yan, X.H., 2018. Retrieving temperature anomaly in the global subsurface and deeper ocean from satellite observations. J. Geophys. Res.: Oceans 123, 399–410.
Sumi, S.M., Zaman, M.F., Hirose, H., 2012. A rainfall forecasting method using machine learning models and its application to the Fukuoka city case. Int. J. Appl. Math. Comput. Sci. 22, 841–854.
Sun, J., Di, L., Sun, Z., Shen, Y., Lai, Z., 2019b. County-level soybean yield prediction using deep CNN-LSTM model. Sensors 19, 4363.
Sun, Z., et al., 2020a. Advanced cyberinfrastructure for intercomparison and validation of climate models. Environ. Modell. Softw. 123, 104559. https://doi.org/10.1016/j.envsoft.2019.104559.
Sun, Z., Di, L., Burgess, A., Tullis, J.A., Magill, A.B., 2020b. Geoweaver: advanced cyberinfrastructure for managing hybrid geoscientific AI workflows. ISPRS Int. J. Geo-Inf. 9, 119.
Sun, Z., Di, L., Fang, H., 2019a. Using long short-term memory recurrent neural network in land cover classification on Landsat and Cropland data layer time series. Int. J. Rem. Sens. 40, 593–614.
Sun, Z., Yue, P., Hu, L., Gong, J., Zhang, L., Lu, X., 2013. GeoPWProv: interleavinghspace map and faceted metadata for provenance visualization and navigation. IEEE Trans. Geosci. Rem. Sens. 51, 5131–5136.
Sutanto, S.J., van der Weert, M., Wanders, N., Blauhut, V., Van Lanen, H.A.J., 2019. Moving from drought hazard to impact forecasts. Nat. Commun. 10, 4945.
Sutton, R.S., Barto, A.G., 2018. Reinforcement Learning: an Introduction. MIT press.
Tamiminia, H., Salehi, B., Mahdianpari, M., Quackenbush, L., Adeli, S., Brisco, B., 2020. Google Earth Engine for geo-big data applications: a meta-analysis and systematic review. ISPRS J. Photogrammetry Remote Sens. 164, 152–170.
Templ, M., Filzmoser, P., Reimann, C., 2008. Cluster analysis applied to regional geochemical data: problems and possibilities. Appl. Geochem. 23, 2198–2213.
Tenopir, C., Rice, N.M., Allard, S., Baird, L., Borycz, J., Christian, L., Grant, B., Olendorf, R., Sandusky, R.J., 2020. Data sharing, management, use, and reuse: practices and perceptions of scientists worldwide. PLoS One 15, e0229003.
Tiemann, L., Nicolaus, M., Hoppmann, M., Huntemann, M., Haas, C., 2018. Sea-ice properties derived from ice mass-balance buoys using machine learning. In: Proceedings of Polar 2018 Open Science Conference.
Tilling, R.I., 1989. Volcanic hazards and their mitigation: progress and problems. Rev. Geophys. 27, 237–269.
Tilmes, C., Fox, P., Ma, X., McGuinness, D.L., Privette, A.P., Smith, A., Waple, A., Zednik, S., Zheng, J.G., 2013. Provenance representation for the national climate assessment in the global change information system. IEEE Trans. Geosci. Rem. Sens. 51, 5160–5168.
Tong, D.Q., Wang, J.X., Gill, T.E., Lei, H., Wang, B., 2017. Intensified dust storm activity and Valley fever infection in the southwestern United States. Geophys. Res. Lett. 44, 4304–4312.
Turan, M.E., Yurdusev, M.A., 2009. River flow estimation from upstream flow records by artificial intelligence methods. J. Hydrol. 369, 71–77.
Umar, Z., Pradhan, B., Ahmad, A., Jebur, M.N., Tehrany, M.S., 2014. Earthquake induced landslide susceptibility mapping using an integrated ensemble frequency ratio and logistic regression models in West Sumatera Province, Indonesia. Catena 118, 124–135.
Vaughan-Nichols, S.J., 2021. Flying on Mars Fueled with Open-Source Software.
Verbraeken, J., Wolting, M., Katzy, J., Kloppenburg, J., Verbelen, T., Rellermeyer, J.S., 2020. A survey on distributed machine learning. ACM Comput. Surv. 53, 1–33.
Vij, A., Vijendra, S., Jain, A., Bajaj, S., Bassi, A., Sharma, A., 2020. IoT and machine learning approaches for automation of farm irrigation system. Procedia Comput. Sci. 167, 1250–1257.
Wang, C., Tandeo, P., Mouche, A., Stopa, J.E., Gressani, V., Longepe, N., Vandemark, D., Foster, R.C., Chapron, B., 2019a. Classification of the global Sentinel-1 SAR vignettes for ocean surface process studies. Rem. Sens. Environ. 234, 111457.
Wang, H., Koydemir, H.C., Qiu, Y., Bai, B., Zhang, Y., Jin, Y., Tok, S., Yilmaz, E.C., Gumustekin, E., Rivenson, Y., 2020a. Early detection and classification of live bacteria using time-lapse coherent imaging and deep learning. Light Sci. Appl. 9, 1–17.
Wang, H., Song, T., Zhu, S., Yang, S., Feng, L., 2021. Subsurface temperature estimation from sea surface data using neural network models in the western pacific ocean. Mathematics 9, 852.
Wang, L., Scott, K.A., Xu, L., Clausi, D.A., 2016. Sea ice concentration estimation during melt from dual-pol SAR scenes using deep convolutional neural networks: a case study. IEEE Trans. Geosci. Rem. Sens. 54, 4524–4533.

Wang, L., Zhang, J., 2019. Prediction of sgRNA on-target activity in bacteria by deep learning. BMC Bioinf. 20, 1–14.
Wang, M., Fu, W., He, X., Hao, S., Wu, X., 2020b. A survey on large-scale machine learning. IEEE Trans. Knowl. Data Eng.
Wang, P., Yao, J., Wang, G., Hao, F., Shrestha, S., Xue, B., Xie, G., Peng, Y., 2019b. Exploring the application of artificial intelligence technology for identification of water pollution characteristics and tracing the source of water quality pollutants. Sci. Total Environ. 693, 133440.
Wang, W., Xu, Z., Weizhen Lu, J., 2003. Three improved neural network models for air quality forecasting. Eng. Comput. 20, 192–210.
Watson-Wright, W., Snelgrove, P., 2021. Technical Advancements to Improve Ocean Understanding, Preparing a Workforce for the New Blue Economy. Elsevier, pp. 17–32.
Watson, G.L., Telesca, D., Reid, C.E., Pfister, G.G., Jerrett, M., 2019. Machine learning models accurately predict ozone exposure during wildfire events. Environ. Pollut. 254, 112792.
WHO, 2021. Air Pollution.
Wilhite, D.A., 2016. Droughts: A Global Assesment. Routledge.
Wimmers, A., Velden, C., Cossuth, J.H., 2019. Using deep learning to estimate tropical cyclone intensity from satellite passive microwave imagery. Mon. Weather Rev. 147, 2261–2282.
Wu, X., Yan, X.-H., Jo, Y.-H., Liu, W.T., 2012. Estimation of subsurface temperature anomaly in the North Atlantic using a self-organizing map neural network. J. Atmos. Ocean. Technol. 29, 1675–1688.
Wu, Y.-C., Shiledar, A., Li, Y.-C., Wong, J., Feng, S., Chen, X., Chen, C., Jin, K., Janamian, S., Yang, Z., 2017. Air quality monitoring using mobile microscopy and machine learning. Light Sci. Appl. 6, e17046.
Xia, M., Liu, W.a., Shi, B., Weng, L., Liu, J., 2019. Cloud/snow recognition for multispectral satellite imagery based on a multidimensional deep residual network. Int. J. Rem. Sens. 40, 156–170.
Xiao, X., Zhang, T., Zhong, X., Shao, W., Li, X., 2018. Support vector regression snow-depth retrieval algorithm using passive microwave remote sensing data. Rem. Sens. Environ. 210, 48–64.
Xu, K., Zhu, W., Darve, E., 2020. Distributed Machine Learning for Computational Engineering Using MPI arXiv preprint arXiv:2011.01349.
Yan, Q., Huang, W., 2018. Sea ice sensing from GNSS-R data using convolutional neural networks. Geosci. Rem. Sens. Lett. IEEE 15, 1510–1514.
Yao, J., Raffuse, S.M., Brauer, M., Williamson, G.J., Bowman, D.M., Johnston, F.H., Henderson, S.B., 2018. Predicting the minimum height of forest fire smoke within the atmosphere using machine learning and data from the CALIPSO satellite. Rem. Sens. Environ. 206, 98–106.
Yaseen, Z.M., El-Shafie, A., Jaafar, O., Afan, H.A., Sayl, K.N., 2015. Artificial intelligence based models for stream-flow forecasting: 2000–2015. J. Hydrol. 530, 829–844.
Yu, M., Huang, Q., Qin, H., Scheele, C., Yang, C., 2019. Deep learning for real-time social media text classification for situation awareness–using Hurricanes Sandy, Harvey, and Irma as case studies. International Journal of Digital Earth 12, 1230–1247.
Zhan, Y., Wang, J., Shi, J., Cheng, G., Yao, L., Sun, W., 2017. Distinguishing cloud and snow in satellite images via deep convolutional network. Geosci. Rem. Sens. Lett. IEEE 14, 1785–1789.
Zhang, M., Jiang, L., Zhao, J., Yue, P., Zhang, X., 2020. Coupling OGC WPS and W3C PROV for provenance-aware geoprocessing workflows. Comput. Geosci. 138, 104419.
Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., Baklanov, A., 2012. Real-time air quality forecasting, part II: state of the science, current research needs, and future prospects. Atmos. Environ. 60, 656–676.
Zhang, Y., Jiang, H., Ye, T., Juhas, M., 2021. Deep learning for imaging and detection of microorganisms. Trends Microbiol. 29, 569–572.
Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890.
Zhao, Y., Ma, J., Li, X., Zhang, J., 2018. Saliency detection and deep learning-based wildfire identification in UAV imagery. Sensors 18, 712.