Artificial intelligence and machine learning in earth system sciences with special reference to climate science and meteorology in South Asia

Manmeet Singh[1,5,6,*], Bipin Kumar[1], Rajib Chattopadhyay[1], K. Amarjyothi[2], Anup K. Sutar[3], Sukanta Roy[3], Suryachandra A. Rao[1] and Ravi S. Nanjundiah[1,4,7]

[1]Indian Institute of Tropical Meteorology, Ministry of Earth Sciences, Pune 411 008, India
[2]National Centre for Medium Range Weather Forecasting, Ministry of Earth Sciences, Noida 201 309, India
[3]Borehole Geophysics Research Laboratory, Ministry of Earth Sciences, Karad 415 114, India
[4]Centre for Atmospheric and Oceanic Sciences, Indian Institute of Science, Bengaluru 560 012, India
[5]Jackson School of Geosciences, The University of Texas at Austin, Austin 78712, USA
[6]IDP in Climate Studies, Indian Institute of Technology Bombay, Powai, Mumbai 400 076, India
[7]Divecha Centre for Climate Change, Indian Institute of Science, Bengaluru 560 012, India

This study focuses on the current problems in earth system science (ESS), where machine learning (ML) algorithms can be applied. It provides an overview of previous studies, ongoing work at the Ministry of Earth Sciences, Government of India, and future applications of ML algorithms to some significant earth science problems. We compare previous studies, a mind map of multidimensional areas related to ML and Gartner's hype cycle for ML in ESS. We mainly focus on the critical components in earth sciences, including studies on the atmosphere, oceans, biosphere, hydrogeology, human health and seismology. Various artificial intelligence (AI)/ML applications to problems in the core fields of earth sciences are discussed, in addition to gap areas and the potential for AI techniques.

Keywords: Artificial intelligence, climate science, earth sciences, machine learning, meteorology, mind map.

THE recent increase in computational power has promoted the application of novel artificial intelligence (AI) and machine learning (ML) techniques. In the last few decades, there has been a significant improvement in forecasts at various scales using numerical methods in conjunction with increasing computational power. The advent of satellites, modern instruments and advanced global/regional modelling capabilities has helped amass large amounts of data surpassing petabytes per day. Hence the need of the hour is to exploit these data innovatively. The datasets have been collected using sensors that monitor the magnitude of states, fluxes and more intensive or time/space-integrated variables. The earth system data exemplify all 'four vs of big data', namely volume, velocity, variety and veracity. The big picture shows that our capacity to gather and store data vastly outpaces our ability to access them, leave alone comprehending them meaningfully. The power to make accurate predictions has not kept pace with abundant data generation/accumulation. We need to undertake two significant endeavours to maximize the wealth of earth system data growth and diversity. These are (1) identifying and utilizing data insights, and (2) developing predictive models that can discover previously unknown laws of nature without neglecting the physical understanding that has been developed so far.

Enhanced data availability and advances in computing capacity provide exceptional new prospects. For example, ML and AI technologies are now accessible, but they require additional development and adaptation to geoscientific studies. In both spatial and temporal domains, new methods present new opportunities, new problems, and ethical demands for contemporary fields of study in earth system science (ESS)[1]. ML algorithms have grown with data availability. They are being successfully applied to many geoscientific processes in the atmosphere, on the land surface and in the ocean. Land cover and cloud classifications have been possible due to Geographic Information Systems (GIS) and the resurgence of neural networks, thanks to the availability of very high-resolution satellite data. The majority of ML research methodologies (for example, kernel techniques or random forests) have since been applied to geoscience and remote sensing problems. ML has emerged as a versatile method for geoscientific data analysis, prediction and quality control.

Need for ML in ESS

ML aims to uncover the transformation functions which map the fields of enormous interest, such as precipitation,

*For correspondence. (e-mail: manmeet.cat@tropmet.res.in)

Table 1. Comprehensive summary of previous surveys on machine learning in earth system science and comparison with this survey

Previous reviews	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Rolnick *et al.*[2]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓									
Reichstein *et al.*[3]							✓		✓	✓	✓			✓		✓	✓					✓	✓	✓		✓
Shen *et al.*[4]							✓		✓	✓	✓			✓	✓	✓	✓			✓			✓	✓		✓
Sit *et al.*[5]			✓	✓	✓		✓			✓		✓	✓	✓	✓	✓	✓						✓	✓	✓	
Ball *et al.*[6]		✓	✓		✓		✓			✓				✓	✓	✓	✓			✓			✓	✓	✓	
Fang *et al.*[7]	✓	✓	✓				✓		✓	✓		✓		✓	✓					✓			✓			✓
The present study	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

A, Electricity systems; B, Transportation systems; C, Buildings and cities/urban climate; D, Industrial systems; E, Farms and forests; F, Climate change mitigation; G, Weather and climate prediction; H, Climate finance; I, Causality; J, Computer vision, K, Interpretable machine learning; L, Natural language processing; M, Reinforcement learning; N, Time series; O, Transfer learning; P, Uncertainty estimation; Q, Unsupervised learning; R, Seismology; S, South Asian monsoon; T, Short-range weather prediction; U, Extended range weather forecasting; V, Seasonal weather prediction; W, Hydrology; X, Oceanography; Y, Transformers or generative adversarial networks; Z, Weather and climate extremes.

temperature, etc. The developments in physical sciences associated with simple statistical methodologies have left a large grey area in uncovering the relationships leading to complex, nonlinear variables. Hence, there is a need to dedicate resources to using advanced ML-based tools to decipher the links between physical fields which are still out of our reach and improve their predictability. The developments in deep learning, deep reinforcement learning, transformers, nonlinear science, and recent advances in interpretable ML are the areas that can help solve crucial research problems in ESS. Recognizing this need, to effectively utilize the extensive data, the Ministry of Earth Sciences (MoES), Government of India (GoI) has recently set up a virtual centre for AI and ML devoted to earth sciences, which is anchored at the Indian Institute of Tropical Meteorology (IITM), Pune.

Related surveys

Table 1 summarizes previous surveys on the use of ML in ESS[2–7]. These reviews have primarily focused on the broad applications of ML in earth science problems. Rolnick *et al.*[2], in the most detailed assessment yet on the topic, focused in general on solutions to tackle the issues associated with climate change using ML. Others focused more on hydrology or remote sensing problems. The survey by Reichstein *et al.*[3] is close to that we have done in the present study.

Motivation for this study

The previous surveys have only addressed problems within ESS in general. There is a need for a review focusing on studies and issues addressing the South Asia region using ML. For example, the Indian monsoon is one of the most complex climate phenomena, which is not fully understood. It requires particular focus and attention to address the challenges in accurately predicting the various spatio-temporal scales of the monsoon. We also focus on using ML methods for extended range predictions.

The studies summarized in Table 1 have not considered the latest state-of-the-art algorithms, such as the attention-based transformers and generative adversarial networks. The advancements brought about by these models in the computer vision and natural language processing community make them excellent candidates to be explored in the domain of ESS.

This study outlines all the previous reviews on the subject, delineates the tools required, the materials needed by interested researchers to gain hands-on experience in ML and can be used to further the applications of ML in ESS.

Background

This section discusses the algorithms, data, problems, tools, educational materials, feature engineering and the emerging areas related to ML in earth sciences. These have been summarized in the mind map depicted in Figure 1, taking the case of weather and climate sciences as an example.

ML algorithms for ESS

Various algorithms that have shown remarkable performance in computer vision, natural language processing, reinforcement learning, etc. can be directly applied to ESS problems. For example, the super-resolution methodology (SRCNN, DeepSD) developed by Dong *et al.*[8] to enhance the resolution of image datasets has been used to downscale the precipitation datasets from coarser resolution to high resolution[9,10]. Seasonal forecast of various aspects of the monsoon has been studied using single and stacked encoder-based techniques[11,12]. Prediction of solar irradiance using convolutional neural network (CNN)

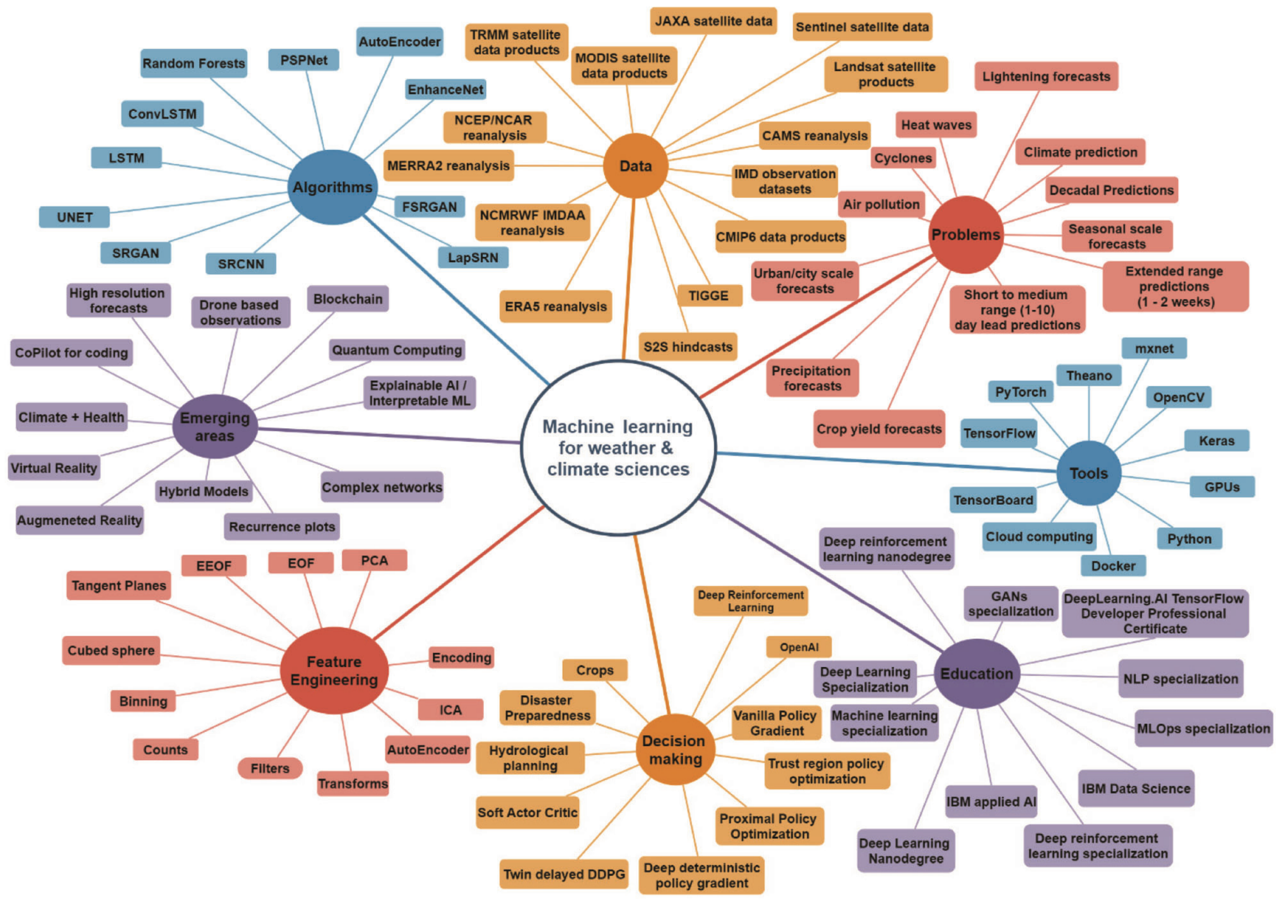

Figure 1. Mind map of multidimensional areas related to machine learning (ML) in weather and climate sciences.

with added attention has been recently done[12]. Recent advances in computer vision show that algorithms such as SRGAN, LapSRN, FSRGAN and UNET outperform the standard SRCNN. Long short-term memory (LSTM) networks, sequence-to-sequence networks and the recent attention-based transformer models have improved the accuracies in natural language processing. Some of these algorithms have also been used or can be applied to the time-series forecasting problems in ESS. A survey on these applications can be found in Lim and Zohren[13]. Weather and climate data are so massive that they have not been explored exhaustively by the community working on big data. The spatio-temporal nature of the datasets, i.e. three-dimensional fields at each temporal dimension, makes it a complex problem to solve. The patterns in this four-dimensional data cannot be deciphered manually, and ML offers a perfect opportunity. Models that have shown good performance on video datasets such as ConvLSTM, can build large-scale, deep learning-based systems that can predict the information in high spatial and temporal resolution[14,15]. Sequence-to-sequence and LSTM networks have been used to predict and forecast active-break cycles of Indian monsoon[16]. Before starting any analysis, traditional algorithms such as random forest, support vector machines and multivariate linear regression should be the first go-to methods. EnhanceNeT and PSPNet are algorithms that can be used to classify the objects in images and spatially locate them. They have shown excellent results in computer vision applications. They can be used for problems such as identifying floods from satellite imagery.

ESS datasets

The understanding of ESS datasets is important while developing ML models. These datasets primarily come in three classes: (i) observational data, (ii) reanalysis products which are merged model outputs and observations created invariably and (iii) dynamical model simulated outputs (such as climate change data from models). For the South Asian domain, long-period ground-based observations made by India Meteorological Department (IMD) are available. These datasets can now be obtained from the website https://dsp.imdpune.gov.in. Satellite-based products are available from the Tropical Rainfall Measuring Mission (TRMM), Landsat, Sentinel and MODIS. Reanalysis products are gridded products which are developed through blending models and observational data products using data assimilation techniques, and are useful for the fields which are not/cannot be measured directly by instruments. They offer insights into the information which is closet to reality. Various reanalysis products are available for the South Asian region, such as IMDAA reanalysis, NCEP/NCAR reanalysis, CAMS reanalysis, ERA5 reanalysis, MERRA-2 reanalysis and JRA55 reanalysis.

With regard to model products, TIGGE (short- and medium-range forecasts), CMIP5/CMIP6 (past and future climate scenarios), and seasonal to sub-seasonal (S2S) hindcasts are available. The model outputs are based on the integration of partial differential equations of dynamical systems. ML offers an innovative methodology to improve these dynamical model estimates by combining them with the observed or reanalysis products.

The archive of seismic waveform data, global positioning system (GPS) data, oceanographic and other geoscience datasets in India is increasing exponentially every year, calling for fast and efficient processing and dissemination of information to the public service systems.

Research problems in ESS

South Asia is home to more than two billion people who are largely dependent on natural climate variability for their livelihood. For example, the Indian monsoon feeds agricultural lands over the region, thus directly impacting its economic well-being. Monsoon is a complex, multiscale and nonlinear problem. Hence linear methods cannot unravel the fundamental processes, especially the feedback processes leading to its variability. Forecasts at various temporal scales such as short to medium range (1–10 days), extended range (2–3 weeks), seasonal scale (for the coming season) and climate scale (hundreds of years) are essential for planning hydrological resources of the region. It has been known that the crop yields are dependent on meteorological variables; ML can be used to accurately forecast the spatial crop yield a season in advance and thus economically benefit the society. The demographics in the South Asian region have considerably changed in the past decades, and many people now live in the cities. This demographic shift could be attributed to the agricultural variability arising from the modulations in rainfall patterns (and other factors such as new opportunities in various sectors).

The population density in South Asian countries is also very high. Hence, locally accurate urban forecasts are a need of the hour. These locations are also sources of chemical species harmful to the environment and all living beings. Hence air-pollution prediction is a significant task. Identifying localities with high air pollution is essential for city planning; for example, deciding the number of electric buses to be introduced in a city. ML-based algorithms can be used to improve the cyclone forecasts of dynamical models. Extreme weather events such as heatwaves and cloud bursts are causing havoc in recent times. It is challenging to predict them accurately. Other important problems of interest to the ESS community are flood forecasting and disaster management using AI/ML-based techniques.

In seismology, AI/ML-based techniques are being used for earthquake detection, phase-picking (measurement of arrival times of distinct seismic phases), event classification, early warning of earthquake, ground motion prediction, tomography and earthquake geodesy. They are also useful to determine and predict tsunami inundation and heights.

Popular tools to perform ML for ESS

The open-source software packages have provided a bridge to the domain experts to avoid reinventing the wheel while applying ML to their problems. Python is the most popular language for ML, and various libraries such as TensorFlow, PyTorch, Theano, MXNet, OpenCV, Keras and PyTorch Lightning are available freely. Visualization software such as TensorBoard and Tableau assist in communicating the results from ML models. In addition to the software requirements, deep learning needs graphical processing units (GPUs) to perform tensor computations in neural networks. Tensor processing units (TPUs) are a step ahead of GPUs, wherein the neural network is encoded on the chip to perform fast calculations. However, TPUs are only available over the cloud, and each individual cannot buy a personal GPU for deep learning. Hence free and paid cloud computing services, such as Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP), Paperspace, Digital Ocean, Google Earth Engine, etc. provide an option to build machines over the cloud to perform deep learning and data analysis in ESS[17]. A step further, the concept of Jupiter notebooks as a service has become popular, and there are several free and paid vendors providing notebooks as a service. Notable amongst them are the free services offered by Kaggle, Google Colab and others. Readers can find information on more cloud vendors at https://github.com/binga/cloud-gpus, https://github.com/zszazi/Deep-learning-in-cloud, https://github.com/discdiver/deep-learning-cloud-providers/blob/master/list.md, etc. 'Docker containers' have also become an essential part of the ecosystem, helping us to deploy end-to-end packages for deep learning.

Educational materials for learning earth system data science

A key component in the ML cycle is the educational resources to build knowledge and apply it to ESS. The avenues to learn data science and use ML for earth sciences applications are the Coursera specializations, courses, professional certificates, Udacity nanodegrees, Udemy courses and other free and paid materials available as massive open online courses (MOOCs). The Development of Skilled Manpower in Earth System Sciences (DESK), Ministry of Earth Sciences (MoES), GoI regularly holds training programmes for young researchers on ML applications in earth sciences. DESK conducted one such training workshop in 2021, and the video recordings of the sessions can be found at https://tinyurl.com/448t8yb4.

Decision-making for ML in ESS

Once the weather/hydrological forecasts are generated, they must be used to make decisions for the benefit of society. Deep reinforcement learning is an excellent method for this. State-of-the-art algorithms such as Deep-Q-networks, vanilla policy gradient, trust region policy optimization, proximal policy optimization, deep deterministic policy gradient (DDPG), soft actor-critic, twin delayed DDPG, etc. can be used to train agents who can guide in decision-making. The most crucial aspect of deep reinforcement learning is the design of the environment, action(s) and reward(s). The authorities can use these tools in decision-making for disaster preparedness/mitigation, hydrological planning and other associated tasks.

Feature engineering for ML in ESS

Feature engineering is the generation of meaningful predictors or parameters to improve the performance of a ML model. It is performed after cleaning the data and preparing them in a format that can train statistical models. It has been noted that removing redundant variables improves the performance of ML systems. Various methods can be used to find the most valuable predictors; some of them are principal component analysis (PCA), empirical orthogonal functions (EOF) and independent component analysis (ICA). Binning, counting, transforming or filtering can extract the predictive signal from the data to improve the models. Unsupervised learning techniques, such as autoencoder, can also assist in finding valuable predictors from raw datasets. The deep learning-based models are, however, coded for image-based input datasets. To overcome this limitation, strategies such as transforming the spherical global data to a cubed sphere or tangent planes mapping can effectively reduce spherical distortions in the data.

Emerging areas in ML for ESS

While the previous decade has seen the hype of deep learning overshadow other ML methodologies, numerous emerging and innovative ML methods can be used for ESS. Graph ML is training neural networks on graphs and is becoming increasingly popular. Complex networks and recurrence plots fall in the category of nonlinear methodologies and are suitable for specific applications. While using ML physical sciences, one primary concern is that these could be considered as black-box models. Interpretable ML aims to address this concern, and analysis of deep learning model weights reveals the patterns learned. Active research is being done in this area, and it is crucial for the increasing acceptability of deep learning models at the production scale in ESS. The emerging fields of augmented reality, virtual reality, improved remote sensing measurements, crowd-sourcing and drone technology offer excellent potential to advance observation data collection and improve ML models.

Applications of AI and ML in earth sciences

The AI/ML algorithms have vast applications in earth sciences problems. Figure 2 depicts a few such applications in areas such as atmosphere/biosphere, seismology and ocean.

Statistical downscaling

Downscaling of data is necessary to obtain a local projection of the information. The present-day models and observations generated from weather stations (or other instruments) are available at a coarser resolution. They are irregularly spaced, which may often lead to misrepresentation (or absence) of precipitation, temperature or other variables at local levels. Downscaling the Indian summer monsoon (ISM) rainfall is a difficult task involving a multi-scale spatio-temporal dynamical process with significant variance[18]. Further, regional variations of ISM rainfall are often quite substantial, varying from a few millimetres to thousands of millimetres within a few hundred kilometres. The ISM rainfall can be classified into different coherently fluctuating zones, linked to complex multi-scale processes[19–21].

Statistical downscaling is a low-cost method to obtain information at the local scale and provide it to the stakeholders. AI and ML techniques are used for statistical downscaling[8,22]. Recently, development in the single image super-resolution using deep learning has proved to be one of the best methods used for this purpose[8–10]. Another method that has shown promising results in statistical downscaling is ConvLSTM documented by Harilal *et al.*[23].

Seismological events

The growing volume of seismological and other geoscience-related datasets acquired from surface and borehole studies requires efficient analysis and trend recognition techniques to extract valuable signals. AI/ML tools have been applied in different fields in seismology, from event

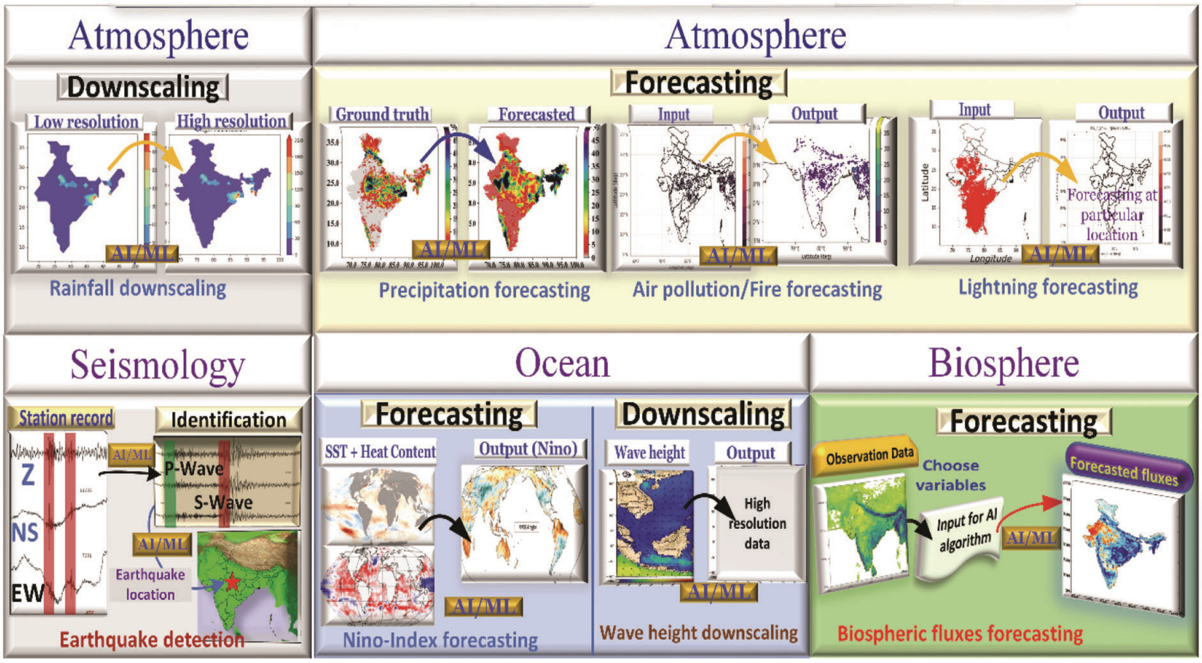

Figure 2. An overview of the application of artificial intelligence (AI)/ML algorithms in some earth sciences problems. The precipitation forecasting can include data from short-range, medium-range and extended-range forecasting.

identification to earthquake prediction, with varying degrees of success[24–29]. The case studies also highlight the need for further research and development to refine the existing techniques, and develop new tools that could be utilized in the processing and analyses of large datasets and identification of different geophysical signals. AI/ML techniques in geoscience/seismology could be employed gainfully to analyse other seismological datasets that MoES, GoI and its affiliated institutions routinely acquire. Identifying seismic phases accurately is one of the primary requirements in seismological data analysis to determine earthquake source parameters. ML helps identify different seismic phases in the data.

In many earthquake detection algorithms, short-term average (STA)/long-term average (LTA) criteria are used to detect possible arrival times of *P* and *S* waves[30]. Therefore, matched filtering or template matching technique is used for event detection. In this method, waveforms of known events are used as templates to scan through continuous waveforms to detect new events[31]. Recently, ML has been utilized to improve earthquake detection and phase-picking capabilities[25,32]. Fingerprinting and similarity thresholding (FAST) is the latest algorithm using ML techniques that have been used to identify earthquakes without prior knowledge of seismicity. FAST would facilitate the automated processing of large and voluminous datasets by being computationally more efficient than template matching. Similarly, the generalized phase detection (GPD) algorithm searches for near-identical waveforms from millions of seismograms, which is used to classify windowed data as *P*, *S* or noise. GPD can be applied to datasets not only encompassed by training sets, but also to complex cases such as clipped seismograms. Kong *et al.*[33] used neural networks to detect *P*-wave onset and *P*-wave polarity. ML techniques have important applications in detecting small-magnitude local earthquakes in areas characterized by sparsity of receivers. AI/ML algorithms may play an essential role in the identification of events and in locating earthquakes with recordings of the events at fewer stations[33,34]. Other applications in earth sciences such as hydrology, show that AI/ML can estimate and predict streamflow in ungauged basins[35–37].

Short- and medium-range data-driven weather forecasting

Currently, the highest global resolution ensemble prediction system at ~12.5 km horizontal resolution (with 21 members) is being used for providing ten-day probabilistic forecast based on the Global Ensemble Forecast System (GEFS@T1534) by IMD. IITM has implemented the high-resolution GEFS for operational application since June 2018. While the deterministic GFS model[38] at 12.5 km horizontal resolution provides a better skill up to ~five days compared to the earlier coarser resolution (~25 km resolution GFST574)[39], the ensemble prediction system has shown much better skill than the control member (the deterministic GFS model), particularly for predicting extreme rainfall events[40,41]. The model forecast inaccuracies mainly arise from initial conditions and improper physical parameterization. The uncertainties of initial conditions are resolved primarily by the perturbed initial states in the ensemble prediction system. However, the uncertainty arising from deterministic closures of the physical parameterization still adds many errors due to unrealistic constraints, namely the quasi-equilibrium[42]. Under the AI/ML paradigms, the use of sub-grid-scale tendencies generated by the cloud-resolving models within each climate model grid would be used as the input of a deep learning model. The inputs would be mapped for training to target the heat and moisture tendencies and this framework holds promise in improving the model fidelity[43–45].

ML for extended range forecasts

AI/ML methods have recently found applications in climate forecast models. There are two basic applications that show promise for near-future climate applications. The first is the bias correction and improvement of the numerical model forecasts. The second relates to the methods attempting the sub-seasonal low-frequency predictions. The bias correction and model post-processing applications are helpful to the stakeholders using climate forecasts. The climate forecasts from dynamical models show substantial bias when the forecast is considered over scales lower than the balanced flow, mainly arising due to unknown physics or unresolved dynamics. When sufficient observations are available over a location, some of the systematic errors arising due to unresolved scale dynamics or physics can be corrected[46]. Sub-seasonal forecasting using ML methods are now under active research[12,47–49].

ML for seasonal and climate-scale forecasting

Seasonal forecasting is one of the most challenging problems in forecasting. As pointed out by Lorenz[50], the weather forecasts are highly dependent on initial conditions (today's weather determines tomorrow's weather). In contrast, climate projections/decadal predictions (an average of weather for a few decades) are less sensitive to the initial conditions. However, they depend on boundary conditions. When we try to make seasonal forecasts, the distinction is somewhat blurred, and the seasonal forecasts still depend on initial conditions[51]. Chattopadhyay *et al.*[51] have shown that model hindcasts initialized with February initial conditions exhibit better prediction skills for the Indian summer monsoon rainfall (ISMR). Further complexities such as resolving ocean processes also become essential at a seasonal scale. Hence, extracting predictive information (which changes from event to event) across both space and timescales is vital to significantly improve seasonal forecasts[52]. Therefore, the use of AI/ML methods for improving seasonal forecasts is imperative, and the research community has started using these methods extensively in seasonal forecasts[53–55]. Some researchers also consider that AI/ML methods can outperform conventional prediction systems for seasonal forecasts[54,55]. Currently, they outperform statistical models.

One of the long-standing seasonal prediction problems is the ISMR prediction. Blandford started seasonal forecasting of ISMR using empirical methods in 1886. Since then, numerous attempts have been made to predict seasonal mean monsoon over India using empirical and dynamical models (atmosphere and coupled ocean–atmosphere models; see Rao *et al.*[39] for more details). Empirical models showed very high skills (>0.9) during the development stages and during the actual operational phase, while they showed weak skills (<0.5). On the other hand, dynamical models showed moderate skill during the hindcast and operational forecast phase[39]. The primary reason for the failure of empirical models in providing high skills during the operational phase is that the relationship between predictors and predictands undergoes secular changes from the time the model has been developed to the stage when it is made operational. To avoid such a situation, AI/ML models can be used efficiently to identify new predictors[53]. Using autoencoders, Saha *et al.*[53] have developed an AI/ML model to predict ISMR with two months lead time and an absolute mean error of less than 3%. On the other hand, the dynamical models exhibit systematic biases in precipitation that arise due to parametrization schemes used in these models[39] and therefore underestimate the extremes. To avoid such systematic errors, AI/ML models will be useful.

ML for improving the physical processes in dynamical models

Dynamical models work on the principle of solving partial differential equations over the area of interest with the necessary initial and boundary conditions. They consist of various components such as atmosphere, ocean, land surface, etc. and a correct representation of physical processes in the numerical models is highly essential for accurate simulations of the coupled climate systems. For example, various researchers have tried to understand the relationship between the Indian monsoon and the global and regional teleconnections such as El Niño-Southern Oscillation (ENSO)[56,57], Indian Ocean dipole (IOD)[58], North Atlantic Oscillation[59], Pacific Decadal Oscillation[60], volcanic eruptions[61] and aerosols[62,63]. Recent studies have attempted to use deep learning to develop models that better represent the physical processes. For example, de Witt and Hornigold[64] used deep reinforcement learning-based approach to test the stratospheric aerosol injection on climate. Volcanic eruptions have been used as an analogue for stratospheric aerosol injection, and deep learning can assist in addressing the nonlinear nature of the problem. Recently, Lamb and Gentine[43] used graph neural networks to study the aerosol optical properties. Seifert and Rasp[65] discusses the role of ML in estimating cloud microphysics. The uncertainties in the simulation of the Indian monsoon arise from the missing or erroneous physics in the dynamical systems. ML to improve the understanding of physical processes can lead to cascading returns by enhancing the hydrological outputs from the numerical weather prediction (NWP) models[66–70].

ML for nowcasting weather and tracking storms cells

There is a need for a high-resolution early warning system with reliable nowcasts in the regions of steep topography and urban areas during severe weather. Traditionally, nowcasting is performed by carrying out extrapolation, probabilistic nowcasting[71], semi-Lagrangian advection scheme[72] and using algorithms like optical flow, etc. The state-of-the-art, data-driven approach plays a pivotal role in weather nowcasting. Doppler weather radar provides extremely high geographical and temporal resolution weather information. Agarwal *et al.*[73] utilized radar images to forecast the weather using the U-Net algorithm, demonstrating that it outperformed the optical flow technique. Su *et al.*[74] have shown that ML approaches have a high learning capacity, and enhance echo position and intensity forecast accuracy in convective cells. The temporal precision of such convective cells varies from 30 to 60 min during a relatively short period. Estimating precipitation in complicated orography regions is a well-known problem. Arulraj and Barros[75] used detection and classification ML algorithms to improve the estimation of orographic precipitation across the Southern Appalachian Mountains. Human lives, ecosystems, manmade structures, and landscapes are at risk when snow avalanches occur in mountainous locations. The International Commission for Alpine Rescue anticipates an increase in the frequency of deadly occurrences caused by snow avalanches, with an average of 138 recorded cases per year in 2015 across Alpine nations and North America. A recent study used ML to simulate the hazards due to snow avalanches[76]. Important precursors for modelling snow avalanche hazards were found to be slope, topographic location, surface wetness and precipitation.

ML for numerical weather prediction

Satellite remote sensing and NWP groups are ripe for rapid advancement in the application of ML. NWP relies heavily on integrating fields generated by satellites and other remote sensing devices. Both spatially and temporally, gaps are a common occurrence in such data. The existence of spatial and temporal gaps is a typical issue in such observations. Alleviating uncertainties arising due to these data gaps is necessary before performing ML. The time series of satellite ocean fields are constructed using an ensemble of neural networks with varying weights[77] and a deep learning method to reconstruct the optical images[78]. While modelling and deploying systems and issuing warnings, the ML method can give a post-forecast correction to account for the uncertainties after learning from all previous failures[79].

ML for hydrogeological modelling

Rajaee *et al.*[80] use 67 published studies to assess the AI approaches towards groundwater level (GWL) modelling. They found that ML could accurately simulate and forecast GWL time series in various aquifers. This type of modelling uses data science to unravel physical relationships between GWL and various hydrological factors. Due to the lack of mathematical/physical representations of the processes, AI models are beneficial in groundwater modelling, where knowledge-driven simulation is challenging to design. Research and methods in hydrogeology have evolved in response to global challenges[81]. Hydrogeologists are now working to find solutions to a wide range of issues, including the long-term supply of potable water, geothermal energy production, preservation of the natural environment and the impact of climate change on groundwater. These challenges can be solved by hydrogeologists using numerical modelling. Identifying piezometric risk zones and calculating groundwater recharge are two examples of simple hydrogeological issues that are routinely treated using simpler models. Iterative discrete forms of the equations driving the hydrogeological process are solved using numerical models to handle complex difficulties. The Internet of Things and other recent technological advancements have allowed hydrogeologists to acquire large amounts of real-time data. Traditional modelling approaches have difficulty extracting useful features, quantifying uncertainty or establishing correlations between diverse factors. At least four issues impede the broad adoption of ML in hydrogeology as a complement to the numerical models. The first constraint is that most ML models are opaque black boxes. Using a black-box model, one does not know the laws that govern the system's operation or the causal relationships between the variables. Hence hydrogeologists cannot explain or justify the model results, either for improved understanding of the phenomena or to support high-stakes judgements. A second issue is that generalization is challenging in hydrogeology data-driven models even with high simulation fidelity. Another drawback of the ML models is that they may not converge and cannot be automatically extended to respond to new events in a system under study. Extensive and dedicated research efforts are needed at the intersection of hydrogeology and ML.

Tsunami evacuations helped by early warnings can considerably reduce the number of casualties. However, incorrect danger predictions and warnings might have the opposite impact. To limit the number of casualties in

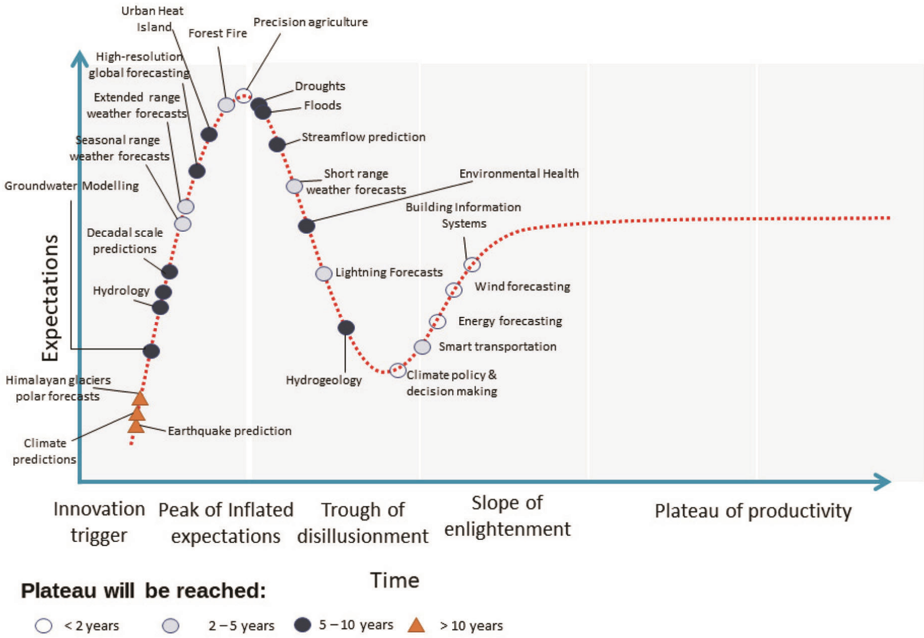

Figure 3. Gartner's hype cycle for ML in earth system science (ESS) with a focus on research problems associated with South Asia.

Figure 4. Word cloud incorporating the crucial aspects of ML in ESS.

future tsunamis, it is vital to develop tsunami forecasting systems based on real-time tsunami observation data and provide early warnings. Using an advanced CNN, researchers were able to accurately forecast tsunamis based on data from extensive tsunami and geodetic monitoring networks[82], which is the first effort at AI-enabled end-to-end tsunami inundation predictions.

AI for climate and human health

Using supervised ML, topic modelling and geoparsing, Berrang-Ford *et al.*[82] identified mapped all climate change and health research published between 1 January 2013 and 9 April 2020. Their analysis included only the studies published in English, with 15,963 climate and health studies published between 2013 and 2019. They found an overwhelming focus on the effects of climate change on human health, with little attention paid to mitigation and adaptation. Causal mortality and infectious disease incidence due to heat and air pollution were most frequently studied. Seasonality, harsh weather, heat and weather variability were the most researched weather exposures. Mental health, undernutrition, and maternal and child health were the areas of climate health study that received less attention. Low-income countries, which often bear the brunt of health consequences due to climate change, were underrepresented in the studies. Climate change and human health must be mapped using automated ML in the era of big data. With the lack of data guidance on climate and health, policymakers may be hesitant to make decisions on how to mitigate the health effects of climate change. ML to generate the datasets can lead to transformational benefits for society.

Summary and future directions

In this study, a review of ML applications in ESS has been done. The future directions especially relevant to solutions for the South Asian region have been summarized as a Gartner's curve (Figure 3). Hard AI problems such as earthquake prediction and climate-scale predictions require long lead times of several years to centuries. They will take more than a decade of development to be fully solved by ML and allied techniques. Such a long development time is expected because of data sparsity; for example, over the Himalayan region, for earthquake prediction. Significant uncertainties in dynamical models to project end-of-century estimates of climate are also expected to be resolved after extensive research and development. Recent developments in ML, particularly in deep learning, are expected to lead to transformative improvements in the short to extended-range forecast, intelligent transportation, precision agriculture, policymaking, wind and energy forecasts during this decade. These advancements would be driven by the critical nature of such problems and the availability of high spatio-temporal drones, ground-based observations and satellite datasets.

We have discussed various AI/ML techniques that have been used and those with high potential for improving the state-of-the-art in ESS. Figure 4 is a word cloud showing all the critical components required for ML in ESS. An exhaustive literature survey on AI/ML/DL applications in the South Asian domain, a mind map incorporating all the essential components of data science applications in ESS and a Gartner's curve for future directions are the main contributions of this review. It can be used as a starting point to understand the existing research problems, applicable algorithms, educational resources, hardware/software stacks and other vital aspects essential to data science for ESS. This work aims to further ESS over South Asia using ML applications as an end goal.

1. Chantry, M., Christensen, H., Dueben, P. and Palmer, T., Opportunities and challenges for machine learning in weather and climate modelling: hard, medium and soft AI. *Philos. Trans. R. Soc. London, Ser. A*, 2021, **379**, 20200083.
2. Rolnick, D. *et al.*, Tackling climate change with machine learning. arXiv.org, 2019; doi:https://arxiv.org/pdf/1906.05433.pdf.
3. Reichstein, M. *et al.*, Deep learning and process understanding for data-driven earth system science. *Nature*, 2019, **566**, 195–204.
4. Shen, C., A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resour. Res.*, 2018, **54**, 8558–8593.
5. Sit, M. *et al.*, A comprehensive review of deep learning applications in hydrology and water resources. *Water Sci. Technol.*, 2020, **82**, 2635–2670.
6. Ball, J. E., Anderson, D. T. and Chan, C. S., A comprehensive survey of deep learning in remote sensing: theories, tools and challenges for the community. *J. Appl. Remote Sensing*, 2017, **11**, 042609.
7. Fang, W., Xue, Q., Shen, L. and Sheng, V. S., Survey on the application of deep learning in extreme weather prediction. *Atmosphere*, 2021, **12**.
8. Dong, C., Chen, Loy, C. C., He, K. and Tang, X., Image super-resolution using deep convolutional networks. CoRR, abs/1501.00092, 2015.
9. Kumar, B. *et al.*, Deep learning-based downscaling of summer monsoon rainfall data over Indian region. *Theor. Appl. Climatol.*, 2021, **143**, 1145–1156.
10. Vandal, T. *et al.*, DeepSD: generating high resolution climate change projections through single image super-resolution. arXiv.org, 2017, 1–9; doi:https://arxiv.org/abs/1703.03126.
11. Saha, M., Mitra, P. and Nanjundiah, R. S., Autoencoder-based identification of predictors of Indian monsoon. *Meteorol. Atmos. Phys.*, 2016, **128**, 613–628.
12. Saha, M. and Nanjundiah, R. S., Prediction of the ENSO and EQUINOO indices during June–September using a deep learning method. *Meteorol. Appl.*, 2020, **27**, e1826.
13. Lim, B. and Zohren, S., Time-series forecasting with deep learning: a survey. *Philos. Trans. R. Soc. London, Ser. A*, 2021, **379**, 20200209.
14. Kumar, B. *et al.*, Deep learning based forecasting of Indian summer monsoon rainfall. 2021; arXiv:2107.04270.
15. Shi, X. *et al.*, Convolutional LSTM network: a machine learning approach for precipitation nowcasting. arXiv.org, 1506.04214, 2015.
16. Viswanath, S., Saha, M., Mitra, P. and Nanjundiah, R. S., Deep learning based LSTM and SeqToSeq models to detect monsoon spells of India. In International Conference on Computational Science – ICCS 2019, Springer, Cham, 2019, pp. 204–218.
17. Singh, M., Singh, B. B., Singh, R., Upendra, B., Kaur, R., Gill, S. S. and Biswas, M. S., Quantifying COVID-19 enforced global changes in atmospheric pollutants using cloud computing based remote sensing. *Remote Sensing Appl.: Soc. Environ.*, 2021, **22**, 100489.
18. Chang, C.-P. *et al.*, The multiscale global monsoon system: research and prediction challenges in weather and climate. *Bull. Am. Meteorol. Soc.*, 2018, **99**, ES149–ES153.
19. Gadgil, S., Yadumani and Joshi, N. V., Coherent rainfall zones of the Indian region. *J. R. Meteorol. Soc.*, 1993, **13**, 546–566.
20. Gadgil, S., The Indian monsoon and its variability. *Annu. Rev. Earth Planet. Sci.*, 2003, **31**, 429–467.
21. Moron, V., Robertson, A. W. and Pai, D. S., On the spatial coherence of sub-seasonal to seasonal Indian rainfall anomalies. *Climate Dyn.*, 2017, **49**, 3403–3423.

22. Tripathi, S., Srinivas, V. V. and Nanjundiah, R. S., Downscaling of precipitation for climate change scenarios: a support vector machine approach. *J. Hydrol.*, 2006, **330**, 621–640.
23. Harilal, N., Singh, M. and Bhatia, U., Augmented convolutional LSTMs for generation of high-resolution climate change projections. *IEEE Access*, 2021, **9**, 25208–25218.
24. Bergen Karianne, J., Johnson Paul, A., de Hoop Maarten, V. and Beroza Gregory, C., Machine learning for data-driven discovery in solid earth geoscience. *Science*, 2019, **363**, eaau0323.
25. Thibaut, P., Michaël, G. and Marine, D., Convolutional neural network for earthquake detection and location. *Sci. Adv.*, 2018, **4**, e1700578.
26. Rouet-Leduc, B., Hulbert, C. and Johnson, P. A., Continuous chatter of the Cascadia subduction zone revealed by machine learning. *Nature Geosci.*, 2019, **12**, 75–79.
27. Reynen, A. and Audet, P., Supervised machine learning on a network scale: application to seismic event classification and detection. *Geophys. J. Int.*, 2017, **210**, 1394–1409.
28. Qingkai, K., Allen Richard, M., Louis, S. and Young-Woo, K., MyShake: a smartphone seismic network for earthquake early warning and beyond. *Sci. Adv.*, 2016, **2**, e1501055.
29. Reddy, R. and Nair, R. R., The efficacy of support vector machines (SVM) in robust determination of earthquake early warning magnitudes in central Japan. *J. Earth Syst. Sci*., 2013, **122**, 1423–1434.
30. Allen, R. V., Automatic earthquake recognition and timing from single traces. *Bull. Seismol. Soc. Am*., 1978, **68**, 1521–1532.
31. Gibbons, S. J. and Ringdal, F., The detection of low magnitude seismic events using array-based waveform correlation. *Geophys. J. Int*., 2006, **165**, 149–166.
32. Wiszniowski, J., Plesiewicz, B. M. and Trojanowski, J., Application of real time recurrent neural network for detection of small natural earthquakes in Poland. *Acta Geophys.*, 2014, **62**, 469–485.
33. Kong, Q. *et al.*, Machine learning in seismology: turning data into insights. *Seismol. Res. Lett*., 2018, **90**, 3–14.
34. Zhu, L. *et al.*, Deep learning for seismic phase detection and picking in the aftershock zone of 2008 M_w 7.9 Wenchuan earthquake. *Phys. Earth Planet. Inter.*, 2019, **293**, 106261.
35. Besaw, L. E., Rizzo, D. M., Bierman, P. R. and Hackett, W. R., Advances in ungauged streamflow prediction using artificial neural networks. *J. Hydrol*., 2010, **386**, 27–37.
36. Mudashiru, R. B., Sabtu, N., Abustan, I. and Balogun, W., Flood hazard mapping methods: a review. *J. Hydrol.*, 2021, **603**, 126846.
37. Zhang, D. *et al.*, Intensification of hydrological drought due to human activity in the middle reaches of the Yangtze River, China. *Sci. Total Environ.*, 2018, **637–638**, 1432–1442.
38. Mukhopadhyay, P. *et al.*, Performance of a very high-resolution global forecast system model (GFS T1534) at 12.5 km over the Indian region during the 2016–2017 monsoon seasons. *J. Earth Syst. Sci*., 2019, **128**, 155.
39. Rao, S. A. *et al.*, Monsoon mission: a targeted activity to improve monsoon prediction across scales. *Bull. Am. Meteorol. Soc.*, 2019, **100**, 2509–2532.
40. Deshpande, N. R. and Kulkarni, J. R., Spatio-temporal variability in the stratiform/convective rainfall contribution to the summer monsoon rainfall in India. *Int. J. Climatol*., 2021.
41. Mukhopadhyay, P. *et al.*, Unraveling the mechanism of extreme (more than 30 sigma) precipitation during August 2018 and 2019 over Kerala, India. *Weather Forecast.*, 2021, **36**, 1253–1273.
42. Tirkey, S., Mukhopadhyay, P., Krishna, R. P., Dhakate, A. and Salunke, K., Simulations of monsoon intraseasonal oscillation using Climate Forecast System Version 2: insight for horizontal resolution and moist processes parameterization. *Atmosphere*, 2019, **10**.
43. Lamb, K. D. and Gentine, P., Zero-shot learning of aerosol optical properties with graph neural networks. 2021; doi:arXiv:2107.10197.
44. Rasp, S., Pritchard, M. S. and Gentine, P., Deep learning to represent subgrid processes in climate models. *Proc. Natl. Acad. Sci. USA*, 2018, **115**, 9684.
45. Brajard, J., Carrassi, A., Bocquet, M. and Bertino, L., Combining data assimilation and machine learning to infer unresolved scale parametrization. *Philos. Trans. R. Soc. London, Ser. A*, 2021, **379**, 20200086.
46. Chattopadhyay, R., Sahai, A. K. and Goswami, B. N., Objective identification of nonlinear convectively coupled phases of monsoon intraseasonal oscillation: implications for prediction. *J. Atmos. Sci*., 2008, **65**, 1549–1569.
47. Martin, Z., Barnes, E. and Maloney, E., Predicting the MJO using interpretable machine-learning models. Earth and Space Science Open Archive, 2021; doi:https://doi.org/10.1002/essoar.10506356.1.
48. Borah, N., Sahai, A. K., Chattopadhyay, R., Joseph, S. and Goswami, B. N., A self-organizing map-based ensemble forecast system for extended range prediction of active/break cycles of Indian summer monsoon. *J. Geophys. Res.* (*Atmos.*), 2013, **118**, 9022–9034.
49. Giffard-Roisin, S. *et al.*, Tropical cyclone track forecasting using fused deep learning from aligned reanalysis data. *Front. Big Data*, 2020, **3**, 1.
50. Lorenz, E. N., Deterministic nonperiodic flow. *J. Atmos. Sci*., 1963, **20**, 130–141.
51. Chattopadhyay, R. *et al.*, Large-scale teleconnection patterns of Indian summer monsoon as revealed by CFSv2 retrospective seasonal forecast runs. *Int. J. Climatol.*, 2016, **36**, 3297–3313.
52. Hoskins, B., The potential for skill across the range of the seamless weather–climate prediction problem: a stimulus for our science. *Q. J. R. Meteorol. Soc.*, 2013, **139**, 573–584.
53. Saha, M., Santara, A., Mitra, P., Chakraborty, A. and Nanjundiah, R. S., Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model. *Int. J. Forecast.*, 2021, **37**, 58–71.
54. Ham, Y.-G., Kim, J.-H. and Luo, J.-J., Deep learning for multi-year ENSO forecasts. *Nature*, 2019, **573**, 568–572.
55. Nooteboom, P. D., Feng, Q. Y., López, C., Hernández-García, E. and Dijkstra, H. A., Using network theory and machine learning to predict El Niño. *Earth Syst. Dyn.*, 2018, **9**, 969–983.
56. Sikka, D. R., Some aspects of the large scale fluctuations of summer monsoon rainfall over India in relation to fluctuations in the planetary and regional scale circulation parameters. *Proc. Indian Acad. Sci. – Earth Planet. Sci.*, 1980, **89**, 179–195.
57. Ashok, K., Behera, S. K., Rao, S. A., Weng, H. and Yamagata, T., El Niño Modoki and its possible teleconnection. *J. Geophys. Res.: Oceans*, 2007, **112**.
58. Ashok, K., Guan, Z., Saji, N. H. and Yamagata, T., Individual and combined influences of ENSO and the Indian Ocean dipole on the Indian summer monsoon. *J. Climate*, 2004, **17**, 3141–3155.
59. Goswami, B. N., Venugopal, V., Sengupta, D., Madhusoodanan, M. S. and Xavier, P. K., Increasing trend of extreme rain events over India in a warming environment. *Science*, 2006, **314**, 1442.
60. Krishnan, R. and Sugi, M., Pacific decadal oscillation and variability of the Indian summer monsoon rainfall. *Climate Dyn.*, 2003, **21**, 233–242.
61. Singh, M. *et al.*, Fingerprint of volcanic forcing on the ENSO–Indian monsoon coupling. *Sci. Adv.*, 2020, **6**, eaba8164.
62. Ayantika, D. C. *et al.*, Understanding the combined effects of global warming and anthropogenic aerosol forcing on the South Asian monsoon. *Climate Dyn.*, 2021, **56**, 1643–1662.
63. Fadnavis, S. *et al.*, Atmospheric aerosols and trace gases. In *Assessment of Climate Change over the Indian Region* (eds Krishnan, R. *et al.*), A Report of the Ministry of Earth Sciences (MoES), Government of India, Springer, Singapore, 2020, pp. 93–116; doi:10.1007/978-981-15-4327-2_5.
64. de Witt, C. S. and Hornigold, T., Stratospheric aerosol injection as a deep reinforcement learning problem. arXiv.org, 2019; doi: arXiv:1905.07366.

65. Seifert, A. and Rasp, S., Potential and limitations of machine learning for modeling warm-rain cloud microphysical processes. *J. Adv. Model. Earth Syst.*, 2020, **12**, e2020MS002301.
66. Singh, B. B. *et al.*, Linkage of water vapor distribution in the lower stratosphere to organized Asian summer monsoon convection. *Climate Dyn*., 2021; doi:10.1007/s00382-021-05772-2.
67. Geer, A. J., Learning earth system models from observations: machine learning or data assimilation? *Philos. Trans. R. Soc. London, Ser. A*, 2021, **379**, 20200089.
68. Grönquist, P. *et al.*, Deep learning for post-processing ensemble weather forecasts. *Philos. Trans. R. Soc. London, Ser. A*, 2021, **379**, 20200092.
69. Kashinath, K. *et al.*, Physics-informed machine learning: case studies for weather and climate modelling. *Philos. Trans. R. Soc. London, Ser. A*, 2021, **379**, 20200093.
70. Balaji, V., Climbing down Charney's ladder: machine learning and the post-Dennard era of computational climate science. *Philos. Trans. R. Soc. London, Ser. A*, 2021, **379**, 20200085.
71. Pulkkinen, S. *et al.*, Pysteps: an open-source Python library for probabilistic precipitation nowcasting (v1.0). *Geosci. Model Dev.*, 2019, **12**, 4185–4219.
72. Kim, T.-J. and Kwon, H.-H., Development of tracking technique for the short term rainfall field forecasting. *Proc. Eng*., 2016, **154**, 1058–1063.
73. Agarwal, S. *et al.*, Machine learning for precipitation nowcasting from radar images. arXiv.org, 2019; doi:https://arxiv.org/abs/1912.12132.
74. Su, A., Li, H., Cui, L. and Chen, Y., A convection nowcasting method based on machine learning. *Adv. Meteorol.*, 2020, **2020**, 5124274.
75. Arulraj, M. and Barros, A. P., Automatic detection and classification of low-level orographic precipitation processes from space-borne radars using machine learning. *Remote Sensing Environ.*, 2021, **257**, 112355.
76. Choubin, B., Borji, M., Mosavi, A., Sajedi-Hosseini, F., Singh, V. P. and Shamshirband, S., Snow avalanche hazard prediction using machine learning methods. *J. Hydrol.*, 2019, **577**, 123929.
77. Sarafanov, M., Kazakov, E., Nikolay, N. O. and Kalyuzhnaya, A. V., A machine learning approach for remote sensing data gap-filling with open-source implementation: an example regarding land surface temperature, surface albedo and NDVI. *Remote Sensing*, 2020, **12**, 3865.
78. Cresson, R., Ienco, D., Gaetano, R., Ose, K. and Tong Minh, D. H., Optical image gap filling using deep convolutional autoencoder from optical and radar images. In IEEE International Geoscience and Remote Sensing Symposium, IGARSS, 2019, pp. 218–221; doi:10.1109/IGARSS.2019.8900353.
79. Boukabara, S.-A. *et al.*, Leveraging modern artificial intelligence for remote sensing and NWP: benefits and challenges. *Bull. Am. Meteorol. Soc.*, 2019, **100**, ES473–ES491.
80. Rajaee, T., Ebrahimi, H. and Nourani, V., A review of the artificial intelligence methods in groundwater level modeling. *J. Hydrol*., 2019, **572**, 336–351.
81. Adombi, Adoubi Vincent De Paul, Chesnaux, R. and Boucher, Marie-Amélie, Theory-guided machine learning applied to hydrogeology – state of the art, opportunities and future challenges. *Hydrogeology*, 2021, **29**, 2671–2683.
82. Berrang-Ford, L. *et al.*, Systematic mapping of global research on climate and health: a machine learning review. *Lancet Planet. Health*, 2021, **5**(8), e514–e525.

Received 1 December 2021; revised accepted 9 March 2022

doi: 10.18520/cs/v122/i9/1019-1030