https://doi.org/10.17221/94/2022-SWR

Past, present and future of the applications of machine learning in soil science and hydrology

Xiangwei Wang[1], Yizhe Yang[2], Jianglong Lv[1,3], Hailong He[1,3]*

[1]College of Natural Resources and Environment, Northwest A&F University, Yangling, Shaanxi, P.R. China
[2]Shaanxi Provincial Farmland Quality and Agricultural Environmental Protection Workstation, Department of Agriculture and Rural Affairs of Shaanxi Province, Xi'an, Shaanxi, P.R. China
[3]Key Laboratory of Plant Nutrition and the Agri-Environment in Northwest China (Ministry of Agriculture), Northwest A&F University, Yangling, P.R. China

*Corresponding author: hailong.he@hotmail.com

Citation: Wang X.W., Yang Y.Z., Lv J.L., He H.L. (2023): Past, present and future of the applications of machine learning in soil science and hydrology. Soil & Water Res., 18: 67–80.

Abstract: Machine learning can handle an ever-increasing amount of data with the ability to learn models from the data. It has been widely used in a variety of disciplines and is gaining increasingly more attention nowadays. As it is challenging to map soil and hydrological information that are characterised with high spatial and temporal variability, applications of machine learning in soil science and hydrology (AMLSH) have become popularised. To better understand the current state of AMLSH research, a scientific and quantitative approach was performed to statistically analyse publication information from 1973 to 2021 archived in the Scopus database using scientometric analysis tools, including VOSviewer, CiteSpace, and the open-source R package "bibliometrix". The results show a significant increase in the number of publications on AMLSH since 2006. The major contributions were identified based on country origins (China, the USA, and India), institutions (Hohai University, Islamic Azad University, and Wuhan University), and journals (Journal of Hydrology, Remote Sensing, and Geoderma). The keywords analysis of the AMLSH research demonstrates four research hotspots: neural network, artificial intelligence, machine learning, and soil. The most frequently utilised machine learning (ML) methods are neural networks, decision trees, random forests and other methods for image processing and predictive analysis. McBratney et al. 2003 is the most highly cited article. Our research sheds light on the research process on AMLSH and concludes with future research perspectives.

Keywords: machine learning; science mapping; scientometric analysis; soil

Abbreviations: AEEMD – adaptive ensemble empirical mode decomposition; AI – artificial intelligence; AMLSH – application of machine learning in soil science and hydrology; ANN – artificial neural network; ANFIS – neuro-fuzzy inference system; BPNNs – back propagation neural networks; CART – classification and regression tree; DEMs – digital elevation models; DL – deep learning; DSM – digital soil mapping; ELM – extreme learning machine; GIS – geographic information systems; GMDH – group method of data handling; GPS – global positioning system; IT – information technology; KNN – K-nearest neighbour; ML – machine learning; MnLR – multinomial logistic regression; NB – naive Bayes model; NN – neural network; RBF – radial basis function; REPTree – ensemble Reduced-error pruning trees; RF – random forest; RS – rough set theory; SAR(1) – seasonal first-order autoregressive; SOC – soil organic carbon; SSVM – smooth support vector machine; SVM – support vector machine; WGPM – Working Group in Pedometrics

Machine learning (ML), as a research branch of artificial intelligence (AI), shows full vitality with the advent of the big data era (Acker 2015; Plasek 2016). ML lies between computer science and statistical science, learning models from data to accomplish various tasks (Jordan & Mitchell 2015) with cluster-

https://doi.org/10.17221/94/2022-SWR

ing, classification, regression, and tagging. ML can be divided into supervised learning, semi-supervised learning, and unsupervised learning (Japkowicz 2001; Abraham et al. 2014; Huang et al. 2014; Usama et al. 2019; van Engelen & Hoos 2020; Liu et al. 2021) according to the learning process. The main difference is whether the training data contains known labels. Supervised learning determines labels for unknown data by learning from labelled training data, while unsupervised learning automatically learns unlabelled data and builds models, which often requires much more data to train than supervised learning. Semi-supervised learning is located in between supervised and unsupervised learning (Başkaya & Jurgens 2016). ML relies heavily on its powerful and varied algorithms to demonstrate its capabilities. Popular algorithms include linear regression, logistic regression, classification and regression tree (CART), naive Bayes model (NB), support vector machine (SVM), K-nearest neighbour (KNN), random forest (RF), and artificial neural networks (ANN) (Aha et al. 1991; Lee & Shin 1999; Le Gratiet & Garnier 2015; Mishina et al. 2015; Ao et al. 2019; Bonakdari et al. 2019; Choi et al. 2020; Xu et al. 2021). ML has been a research hotspot in various disciplines based on the powerful learning ability of ML, its black box learning method enables models to be defined according to people's needs. It also promotes the development of earth science, especially to investigate complex soil physical and hydrological processes and properties in difficult-to-sample areas (Govindaraju 2000b; McBratney et al. 2003). Goyal et al. (2017) demonstrates that ML-based models show good accuracy on a limited data set of interest, because ML can use multiple related remote sensing data sources to complement the data shortage (Kumar et al. 2021). In addition, ML can effectively supplement the world soil and hydrology related databases (Jafari et al. 2014). Moreover, soil science and hydrology are closely related disciplines and critical for the earth system science. Therefore, we focused on the application of ML in soil science and hydrology in this study.

Hydrology is the basic science of developing and controlling water resources, studying the occurrence, distribution, circulation and nature of water on Earth, and is related to multiple geophysical disciplines (Horton 1933; Harshbarger & Ferris 1963; Eagleson 1994). ML is used to build models in order to solve hydrological processes with high spatial-temporal variability such as runoff, precipitation, and pollutant concentrations (Govindaraju 2000a, b). Common ML algorithms in hydrological models include: ANN, SVM, extreme learning machine (ELM) (Gharib & Davies 2021). In fact, SVMs using kernel functions are alternative training methods for polynomials, radial basis functions (RBFs), and multilayer perceptron classifiers (Huang et al. 2010). While there are many challenges of using ML algorithms for data-driven modelling in hydrological modelling and prediction, including the inherently difficult to interpret properties of ML and the poor representation of hydrological data in some regions, these shortcomings are being tackled. Specifically, Besalatpour et al. (2012) pointed out that the accuracy of SVM results can be effectively improved by introducing a simulated annealing algorithm. Pappenberger et al. (2005) used a simplified approach aiming to reduce the computational burden of uncertainty estimation in flood modelling, because the impact of uncertainty from the model structure, parameters and inputs was too large in previous flood models, resulting in an increase in the cost of the model operation (Grimaldi et al. 2019).

In soil science, the growing power of tools, such as the geographic information system (GIS) (Burrough 1986), global positioning system (GPS), remote and proximal sensors, and data sources provided by digital elevation models (DEMs) is also boosting the application of ML (Wadoux et al. 2020). After the establishment of the Working Group in Pedometrics (WGPM) in 1990, pedometrics became a multidisciplinary field for soil mapping on a global scale (McBratney et al. 2019). By 2003, McBratney et al. (2003) proposed the digital soil mapping framework (DSM) based on Jenny's S = clorpt model (S, soil; cl, climate; or, organisms; p, plant; t, time) (Jenny 1941). Researchers have tried to use CART, neural networks (NNs) and other algorithms to fit the quantitative relationship between soil properties or categories and environmental factors (Maulik & Bandyopadhyay 2000; Fidêncio et al. 2001; Pachepsky et al. 2001; Lane 2002; McBratney et al. 2002; Moran & Bui 2002). ANN, RF, and Multinomial Logistic Regression (MnLR) have gradually become the most commonly used models for soil classification and soil prediction (Zeraatpisheh et al. 2020). ANN has good prediction accuracy for the soil enzyme activity (Tajik et al. 2012), SOC (Ayoubi & Karchegani 2012), soil aggregate stability (Besalatpour et al. 2013), soil hydraulic properties (Azadmard et al. 2020), and Atterberg consistency (Zolfaghari et al. 2015). Although some have questioned the credibility of ML (Rossiter 2018), it is widely acknowledged that modelling soil processes through ML improves our understanding of soil properties and processes (Rudin & Wagstaff

https://doi.org/10.17221/94/2022-SWR

2014; Brungard et al. 2015; Fajardo et al. 2016; Rossiter 2018; Ma et al. 2019; Mjolsness & Decoste 2001) when parallel genetic algorithms are applied. In addition, the introduction of remote sensing data is almost necessary whether it is soil science or hydrology, because ML requires a large amount of data and more covariates improves the model accuracy (Rindfuss et al. 2004; Hansen & Loveland 2012).

Previous studies have reviewed advances of ML in soil science or hydrology (Govindaraju 2000a, b; Xie et al. 2020; Zhang et al. 2020a). However, no overview on the applications of ML in both soil science and hydrology was found. Bibliometrics was already successfully used to investigate soil science and hydrology (He et al. 2020; Xie et al. 2020; Zhang et al. 2020a) to quantitatively calculate research trends on specific topics. Therefore, the bibliometric method was used in this study to investigate the current research status and research hotspots.

MATERIAL AND METHODS

Scopus is currently the world's largest database of abstracts and citations, containing nearly 50 million pieces of literature since 1823 (Vilchez-Roman 2014). The AMLSH research data used in this study was downloaded on February 4, 2022. The query sets are: TITLE-ABS-KEY ((soil* OR hydrolog* OR hydraulic OR hydropedolog*) AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network" OR "random forest" OR "gaussian process regression" OR "gradient descent" OR "decision tree" OR "back propagation" OR "support vector machine" OR "boosted regression trees" OR "classification and regression tree" OR "nearest neighbour" OR "multivariance linear regression" OR "ensemble learning method" OR "adaptive boosting" OR "extreme gradient boosting" OR "nonlinear regression" OR "extreme learning machine" OR "Group Method of Data Handling" OR "Fuzzy logic" OR "Multivariate quadratic equations")). TITLE-ABS-KEY is a combined field for retrieving papers that contain terms in the abstracts, keywords, and document titles meeting the query sets. The search results only retained articles, conference papers, reviews, book chapters, book, notes, data papers, letters, and reports, which returned 26 978 publications. In 2006, deep learning became so popular that mainstream ML algorithm research was largely on track (LeCun et al. 2015). Therefore, publications were manually screened for 2006 and before to ensure the accuracy and completeness of the returned data, thus 24 878 publications were retained for the scientometric analysis.

Data visualisation was performed using VOSviewer V1.6.18 (Van Eck & Waltman 2010), the open-source R package bibliometrix V3.1.4 (Aria & Cuccurullo 2017), and CiteSpace V5.8.R3 (Drexel University, USA) (Chen 2004). VOSviewer is used to generate Cluster density visualisation maps and overlay the visualisation of the co-authorship of countries. "TLS" refers to the total number of co-occurrences (including repeated co-occurrences) of an item with other items; "network" represents a set of items; "cluster" represents a set of items contained in a network map, where an item can belong to only one cluster. Bibliometrix was used to analyse the number of annual publications, major journals, institutions, and countries as well as to give important indices, such as the H-index (Hirsch 2005) and the G-index (Egghe 2006). A scientist has index H if h of his or her N papers have at least h citations each and the other $(N - h)$ papers have less than or equal to h citations each. The G-index g is the largest rank. Papers are sorted in descending order by the number of citations received, with the first g papers (added together) having at least g^2 citations. The G-index was proposed as an improvement to the H-index because the H-index may not be friendly enough for some young scientists. CiteSpace was used to present keyword burst times. Origin2021 (OriginLab Corporation, Northampton, MA, USA) and ArcMap10.1 (Environment System Research Institute, ESRI) were also used for data visualisation.

RESULTS AND DISCUSSION

Annual publication trend. The annual number of articles published is an important indicator of research trends in AMLSH research. This is shown in Figure 1A, the relevant publications first appeared in 1973. There are very few publications on AMLSH prior to 1995 (0.93% of AMLSH), a slow increase can be noticed between to 1995 and 2006 (> 50 per year) and a sharp increase can be noticed after 2006 (> 500 publications per year). Although AMLSH publications only account for a small portion of the overall volume of publications on ML, the share of AMLSH research is gradually increasing. Figure 1B shows that ML in soil science and hydrology is developing at the same time, and the number of publications in both disciplines is increasing. This may have something to do with the history of ML, as Hinton (LeCun et al.

https://doi.org/10.17221/94/2022-SWR

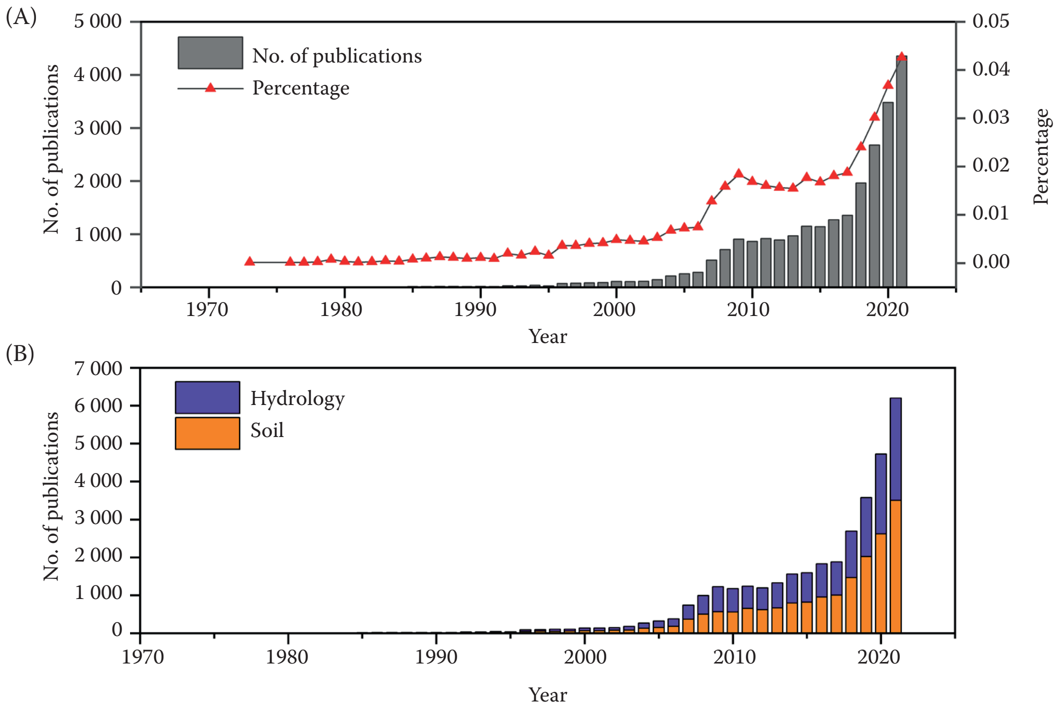

Figure 1. (A): Annual number and share of publications on the application of machine learning in soil science and hydrology (AMLSH) from 1973 to 2021 of Scopus (the red dots in the figure represent the percentage of AMLSH-related publications among all the soil science and hydrology-related publications in each year, and the grey bars represent the specific number of AMLSH publications per year), (B): soil science and hydrology publications in AMLSH (orange represents soil science publications in AMLSH, and purple represents hydrology publications in AMLSH; the sum of the two publications may be more than the total AMLSH publications in Figure A, because some publications have both soil science and hydrology)

2015) proposed a solution in 2006 to the gradient disappearance problem in deep network training allowing ML to process more data. The subsequent emergence of more open-source algorithms facilitates the wide application of ML (Hastie et al. 2001; McBratney et al. 2003). In addition, the increase in the publication volume may also be related to the increase in the number of related journals and the advancement of Information Technology (IT), especially the rapid development of computer hardware (Hastie et al. 2001).

Analysis of countries and organisations on publications. A total of 140 countries published research on AMLSH and 65 countries published at least 40. China (N = 7 335) and the United States of America (USA, N = 4 363) published the most publications (Table 1). Publications from the USA are cited much more than other countries (C = 108 694). Australian and British publications have greater per article citations, C/N = 30.23 and 28.99, respectively. Figure 2 shows that the USA, Italy, and France are pioneer countries conducting AMLSH research and have a large number of cooperative relationships with other countries. In recent years, countries such as Vietnam have also contributed to AMLSH research.

In bibliometrix's review of institutions' publications (Figure 3), particular attention is paid to the fact that all the co-author institutions were included at the time of the institutional review. Among the top ten institutions by the number of publications, Hohai University (384), Wuhan University (362), and Zhejiang University (360) from China publish more articles than any other universities. Table 2 summarises ten publications (first author) with the most global citations retrieved from the Scopus database from Hohai University, Wuhan University and Zhejiang University. The three institutions published more quality publications and these authors are excellent at combining traditional ML algorithms with other statistical or physical models. Most authors used the SVM algorithm and derived other improved algorithms. It also indicates that research

https://doi.org/10.17221/94/2022-SWR

Table 1. Top 10 productive organisations/countries with publications on the application of machine learning in soil science and hydrology (AMLSH) from 1973 to 2021

No.	Items	N	C	C/N	TLS
1	China	7 335	77 231	10.53	2 748
2	United States	4 363	108 694	24.91	3 106
3	India	2 167	30 208	13.94	967
4	Iran	2 041	30 306	19.30	1 915
5	Germany	1 103	26 036	23.60	1 422
6	Australia	1 074	32 472	30.23	1 344
7	Canada	1 038	24 726	23.82	1 123
8	United Kingdom	886	25 685	28.99	1 184
9	Italy	757	18 725	24.74	806
10	Turkey	745	17 921	24.06	424

N – number of publications; C – citations; C/N – the calculated average citations per publication; TLS – total link strength; this study included all the co-authors

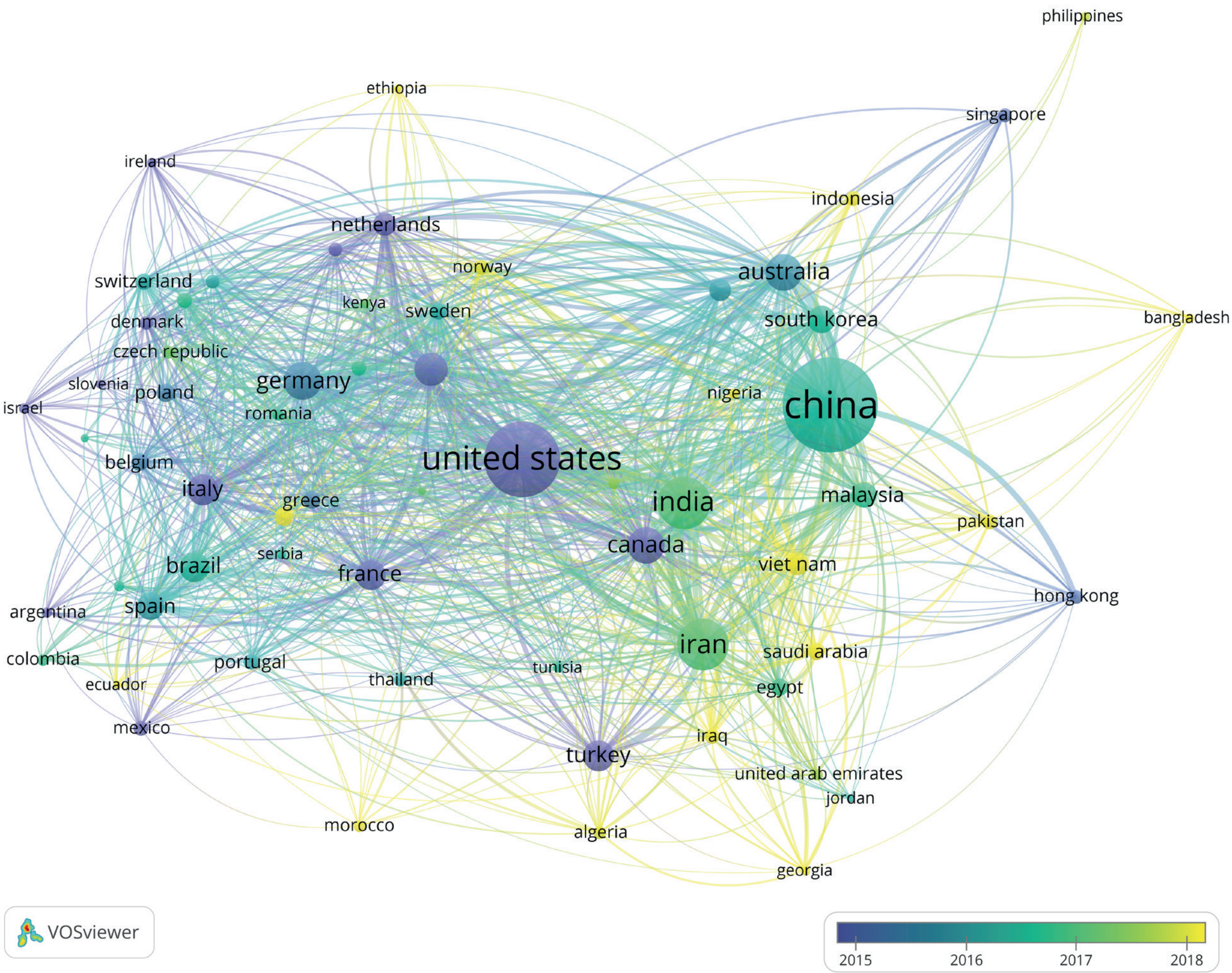

Figure 2. Overlay visualisation of the co-authorship for countries with a minimum of 40 publications on the application of machine learning in soil science and hydrology (AMLSH) between 1973 and 2021 (the size of the shapes and fonts in the diagram depends on the degree of nodes, the strength of the links, and the number of references)

https://doi.org/10.17221/94/2022-SWR

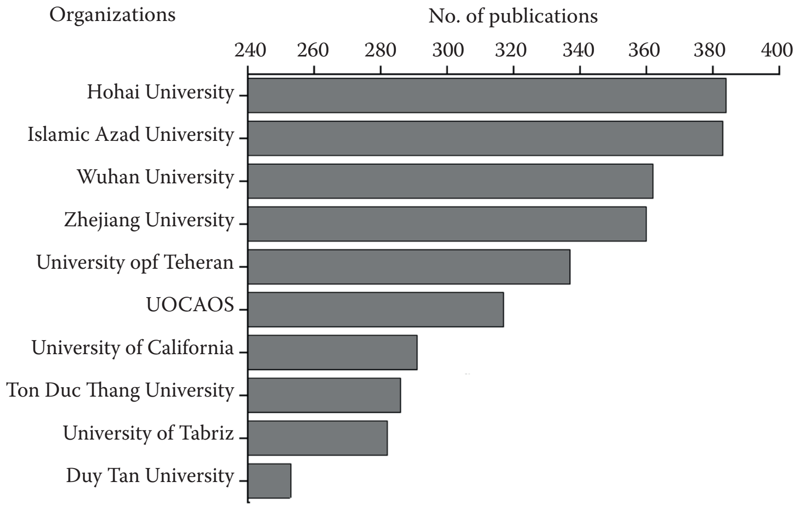

Figure 3. Top 10 institutions in the number of publications on the application of machine learning in soil science and hydrology (AMLSH)
UOCAOS – University of Chinese Academy of Sciences

on hydrological models and land use classification is very mature.

The establishment and analysis of mathematical models of hydrological data is an early research direction. After that, people began to study the geological model of Asia, Europe, and North America. In recent years, people have been more inclined to develop and utilise more complex ML algorithms. More work on climate, soil, and runoff studies, needs to be undertaken in Africa, Eastern Europe, and South America to improve the impact of the global climate change (Figure 4) (Jung et al. 2010). Moreover, the African region has richer mineral resources, and the study of geological models for Africa is also beneficial to the rational development of minerals and infrastructure construction (Nemmour & Chibani 2006).

The co-occurrence and burst time of keywords. VOSviewer is used for the network visualisation analysis of the keywords. For the 24 887 publications, there are 92 338 keywords in the title, abstract, and keywords provided by the authors, 182 keywords met the threshold of 300 occurrences. The filtered keywords are divided into four categories: red, green, yellow, and blue (Figure 5). Red is associated with artificial intelligence (AI), and the high-frequency words are climate change, hydrological modelling, and runoff (Mittermeier et al. 2019; Zhu et al. 2019; Sireesha Naidu et al. 2020; Loganathan & Mahindrakav 2021). The green portfolio themes include NN and ANN. Forecasting of soils and hydrology using NNs and ANNs is very popular (Sharifi et al. 2017; Zhang et al. 2020b; Li et al. 2021). As early as 1990 s, Hsu

Table 2. Top 10 cited publications (first authors) published by Hohai University, Wuhan University and Zhejiang University in the world

No.	Publications	Research	Algorithm	TC
1	Yuan et al. (2020)	environmental remote sensing (review)	DL	292
2	Huang and Zhang (2012)	land use identify (building detection)	SVM	271
3	Chen et al. (2012)	statistical downscaling and hydrological models	SSVM	214
4	Peng et al. (2014)	landslide susceptibility mapping	RS, SVM	185
5	Han et al. (2015)	winter wetland changes	SVM	177
6	Li et al. (2007)	precision agriculture (management zones)	fuzzy c-means clustering	160
7	Jingyi and Hall (2004)	regional flood frequency analysis	residuals method, Ward‘s cluster method, fuzzy c-means method, Kohonen neural network	140
8	Chen et al. (2019)	flood susceptibility modelling	REPTree	120
9	Wang et al. (2005)	slope stability	BPNN	120
10	Tan et al. (2018)	middle and long-term runoff forecast model	aEEMD-ANN, ANFIS, SVM, SAR	119

TC – total citations

https://doi.org/10.17221/94/2022-SWR

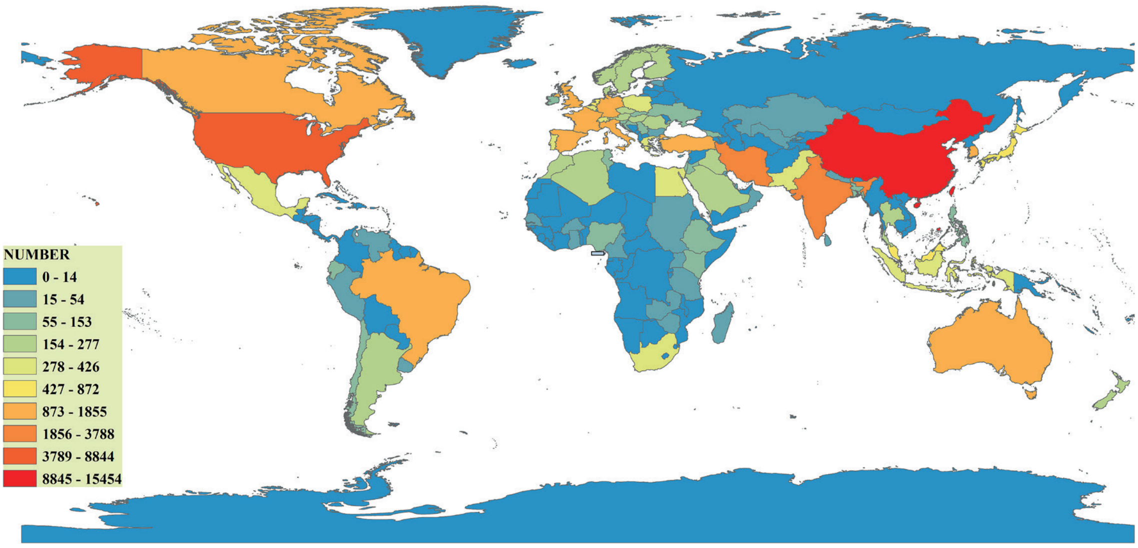

Figure 4. World map of publications on the application of machine learning in soil science and hydrology (AMLSH) between 1973 and 2021

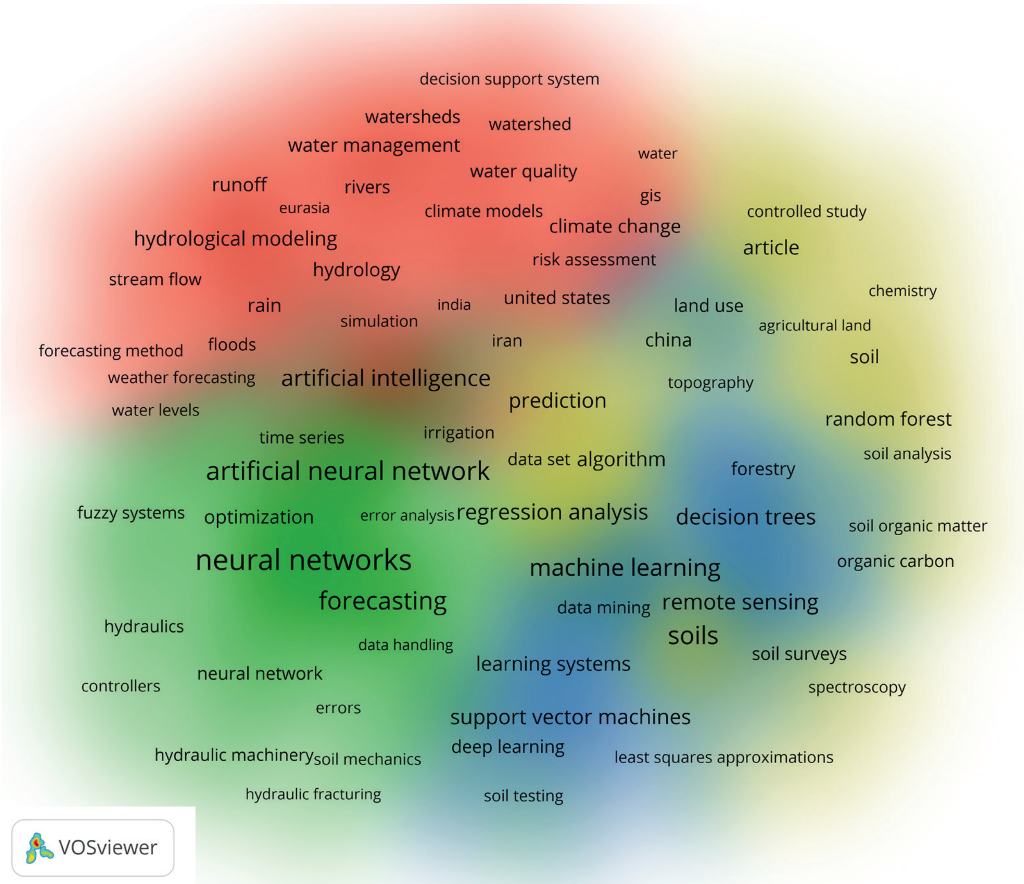

Figure 5. Cluster density visualisation map of 182 keywords, each having more than 300 occurrences

The size of the shapes and fonts in the diagram depends on the degree of nodes, the strength of the links, and the number of references; the colour represents the group to which it belongs and the group is represented by different colours; the colour of the points in the map depends on the number of items near the point and the importance of the adjacent items

https://doi.org/10.17221/94/2022-SWR

et al. (1995) suggested that non-linear ANNs could better represent the rainfall-runoff relationship, and proposed a rainfall-runoff modelling constructed by Multilayer Feed Forward Neural Networks. The blue group revolves around the combination of algorithms, such as decision trees and SVM, used in ML with remote sensing. The classifier function of ML is often used to process remote sensing images, which, in turn, is applied to DSM (Wang et al. 2020). The yellow cluster theme is soils, and hot words include regression analysis and random forest (Ließ et al. 2012; Qiu et al. 2016; Dai et al. 2022). For example, Zhang et al. (2017) found that RF is better than MnLR in establishing the non-linear and hierarchical relationship between soil organic carbon (SOC) and its impact factors.

One should be bear in mind that the choice of the ML algorithm is very critical as different algorithms have pros and cons and there are no universal algorithms that are applicable for all problems. For instance, the data volume hour tree model is more accurate in lake water level prediction, while the long short-

Figure 6. Top 50 keywords with the strongest citation bursts between 1973 and 2021

https://doi.org/10.17221/94/2022-SWR

Table 3. Top 10 most utilised journals with the application of machine learning in soil science and hydrology (AMLSH) research between 1973 and 2021.

Journals	N	H	G	TC
Journal of Hydrology	592	91	155	32 626
Remote Sensing	553	44	74	9 594
Geoderma	370	68	125	19 114
Water (Switzerland)	316	28	45	8 328
Science of the Total Environment	315	50	75	3 910
Water Resources Research	253	62	103	12 320
Proceedings of SPIE – TISFOE	252	8	11	481
Water Resources Management	226	47	68	6 922
IGARSS	216	12	15	637
TOTCSOAE	207	16	21	1347

N – number of publications; H – H-index; G – G-index; TC – total citations; TISFOE – the International Society for Optical Engineering; IGARSS – International Geoscience and Remote Sensing Symposium; TOTCSOAE – Nong Ye Gong Cheng Xue Bao/Transactions of The Chinese Society of Agricultural Engineering

term memory is more accurate when the data volume is larger (Zhu et al. 2020). In addition, when applying ML, it is necessary to pay attention to the selection of environmental covariates to improve the accuracy of the model. Prediction of soil pH with ML tends to be more accurate than most other soil properties, but its accuracy decreases with the soil depth (Chen et al. 2022). Therefore, it is necessary to further improve the accuracy of deeper soil property predictions.

CiteSpace is used to conduct the keyword burst time analysis (Figure 6), and the red line represents the time period with the most occurrences. Keywords began in 1977, probably because there were only two previous articles: 1973 (Duffy & Franklin 1973), 1976 (Ikeda et al. 1976). The hydrological data are processed using the group method of data handling (GMDH) proposed by Ivakhnenko (1971). There are four important nodes in the scientists' study of AMLSH: (1) The application of expert systems, mathematical models, and statistical methods in soil and hydrology was a research hotspot between 1997 and 2013 (Ikeda et al. 1976; Brillinger 1985; Certes & Hubert 1985; Cooley et al. 1986; Kinniburgh 1986). (2) After 1991, more attention is paid to NNs (Buszewski & Kowalkowski 2006) and GIS (Zhu & Band 1994; Moran & Bui 2002; Lee et al. 2003), and rainfall-runoff modelling (Minns & Hall 1996; Savic et al. 1999). (3) Hydraulic model studies in Eurasia and North America became popular in 2003 (Cronican & Gribb 2004; Mukerji et al. 2009). More ML algorithms were introduced that may sprang up the study of geological models (Gharahi Ghehi et al. 2012). (4) Since 2018, there has been a tendency to develop more complex networks (e.g., convolutional neural networks and deep neural networks) to solve research questions (Fang et al. 2017; Bui et al. 2020; Taghizadeh-Mehrjardi et al. 2020; Tien Bui et al. 2020; Yang et al. 2020).

Journals and most global cited documents. There are 3 938 journals that have published 24 878 AMLSH-related articles between 1973 and 2021, a total of 3 300 articles were published in the top ten most utilised journals (Table 3). The H-index balances the publication yield and quality (Hirsch 2005), and the G-index (Egghe 2006) complements the H-index. The Journal of Hydrology has the most publications (N = 592),

Table 4. Top 10 most cited publications on the application of machine learning in soil science and hydrology (AMLSH) research from 1973 to 2021

Publications	TC	TC per year	Normalised TC
McBratney A.B., 2003, Geoderma (McBratney et al. 2003)	1 870	93.50	30.87
Schaap M.G., 2001, Journal of Hydrology (Schaap et al. 2001)	1 621	73.68	21.96
Weng Q., 2004, Remote Sensing Environment (Weng et al. 2004)	1 457	76.68	32.40
Jung M., 2010, Nature (Jung et al. 2010)	1 296	99.69	51.23
Govindaraju R.S., 2000, Journal of Hydrologic Engineering (Govindaraju 2000a)	1 278	55.57	18.05
Hengl T., 2017, PLoS ONE (Hengl et al. 2017)	1 225	204.17	68.96
Govindaraju R.S., 2000, Journal of Hydrologic Engineering (Govindaraju 2000b)	1 197	52.04	16.90
Shadbolt N., 2006, IEEE Intelligent Systems (Shadbolt et al. 2006)	1 150	67.65	28.45
Hsu K., 1995, Water Resources Research (Hsu et al. 1995)	1 094	39.07	11.33
Wen L., 2018, IEEE Transactions on Industrial Electronics (Wen et al. 2018)	726	145.20	48.96

TC – total citations

https://doi.org/10.17221/94/2022-SWR

followed by Remote Sensing (N = 553). It is worth noting that Geoderma (N = 370), Water Resources Research (N = 253) have fewer publications, but have a more important role in the field of AMLSH research, with TC = 19 114 and 12 320, respectively. Their indices are high, with H = 68 and 62; G = 125 and 103, respectively.

The most frequently cited AMLSH publication is McBratney et al. (2003) entitled "On digital soil mapping" (TC = 1 870). It reviews various methods of soil mapping based on GIS, and proposes DSM methods, which better promotes the application of ML in soil science. The highest average annual citation is "250 m resolution Global gridded soil information database" released in 2017 by Hengl et al. (2017). The database provides global projections of soil characteristics (i.e., SOC, bulk density, CEC, pH, soil texture fractions, and coarse fragments) at seven depths (0, 5, 15, 30, 60, 100, and 200 cm). In addition, Govindaraju (2000a, b) presented "A comprehensive introduction to the application of artificial neural networks in hydrology", which is a seminal article for ML in hydrology (Table 4).

CONCLUSION AND PERSPECTIVES

Three bibliometric methods were used to analyse the research applications of machine learning in soil science and hydrology (AMLSH). The results show that the number of publications increased from 1973 to 2021, with a sharp increase in the number of annual publications after 2006 (annual publication reserves of > 500) and peaked in 2021 (N = 4352) and we can foresee a continuous increase thereafter. China and the USA have conducted the most AMLSH research. Hohai University (N = 384), Wuhan University (N = 362), and Zhejiang University (N = 360) are the three institutions with the largest number of publications, and have published a large number of high-quality publications on Hydrological Models and Land Use Identify. The Journal of Hydrology, Remote Sensing, Geoderma are most widely utilised journals in the field of AMLSH. The top cited AMLSH publications are related to digital soil mapping and a gridded soil information database. The keyword analysis shows that neural networks, artificial intelligence, machine learning most frequently appear. It is expected that AMLSH research will continue to boom in the next decades to solve complex questions or to make predictions when observations are not enough. The combination of ML with physical models is encouraged to meet the research needs with more confidence.

REFERENCES

Abraham A., Pedregosa F., Eickenberg M., Gervais P., Mueller A., Kossaifi J., Gramfort A., Thirion B., Varoquaux G. (2014): Machine learning for neuroirnaging with scikit-learn. Frontiers in Neuroinformatics, 8: 14.

Acker A. (2015): Toward a hermeneutics of data. Ieee Annals of the History of Computing, 3: 70–75.

Aha D.W., Kibler D., Albert M.K. (1991): Instance-based learning algorithms. Machine Learning, 1: 37–66.

Ao Y.L., Li H.Q., Zhu L.P., Ali S., Yang Z.G. (2019): The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. Journal of Petroleum Science and Engineering, 174: 776–789.

Aria M., Cuccurullo C. (2017): Bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics, 4: 959–975.

Ayoubi S., Karchegani P.M. (2012): Determination the factors explaining variability of physical soil organic carbon fractions using Artificial Neural Network. International Journal of Soil Science, 1: 1–14.

Azadmard B., Mosaddeghi M.R., Ayoubi S., Chavoshi E., Raoof M. (2020): Estimation of near-saturated soil hydraulic properties using hybrid genetic algorithm-artificial neural network. Ecohydrology and Hydrobiology, 3: 437–449.

Başkaya O., Jurgens D. (2016): Semi-supervised learning with induced word senses for state of the art word sense disambiguation. Journal of Artificial Intelligence Research, 55: 1025–1058.

Besalatpour A., Hajabbasi M.A., Ayoubi S., Gharipour A., Jazi A.Y. (2012): Prediction of soil physical properties by optimized support vector machines. International Agrophysics, 2: 109–115.

Besalatpour A.A., Ayoubi S., Hajabbasi M.A., Mosaddeghi M.R., Schulin R. (2013): Estimating wet soil aggregate stability from easily available properties in a highly mountainous watershed. Catena, 111: 72–79.

Bonakdari H., Ebtehaj I., Samui P., Gharabaghi B. (2019): Lake water-level fluctuations forecasting using Minimax Probability Machine Regression, Relevance Vector Machine, Gaussian Process Regression, and Extreme Learning Machine. Water Resources Management, 11: 3965–3984.

Brillinger D.R. (1985): Fourier inference: Some methods for the analysis of array and nongaussian series data. JAWRA Journal of the American Water Resources Association, 5: 743–756.

https://doi.org/10.17221/94/2022-SWR

Brungard C.W., Boettinger J.L., Duniway M.C., Wills S.A., Edwards T.C. (2015): Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma, 239: 68–83.

Bui D.T., Tsangaratos P., Nguyen V.T., Liem N.V., Trinh P.T. (2020): Comparing the prediction performance of a Deep Learning Neural Network model with conventional machine learning models in landslide susceptibility assessment. Catena, 188: 104426.

Burrough P.A. (1986): Principles of geographical information systems for land resources assessment, Geocarto International, 1: 54.

Buszewski B., Kowalkowski T. (2006): A new model of heavy metal transport in the soil using nonlinear artificial neural networks. Environmental Engineering Science, 4: 589–595.

Certes C., Hubert P. (1985): Application de la programmation logique en hydrologie. Definition d'un programme d'interpretation automatique des pompages d'essai. Journal of Hydrology, 1–2: 137–155.

Chen C. (2004): Searching for intellectual turning points: Progressive knowledge domain visualization. Proceedings of the National Academy of Sciences of the United States of America, 101 (Suppl 1): 5303–5310.

Chen H., Xu C.Y., Guo S. (2012): Comparison and evaluation of multiple GCMs, statistical downscaling and hydrological models in the study of climate change impacts on runoff. Journal of Hydrology, 434–435: 36–45.

Chen S., Arrouays D., Leatitia Mulder V., Poggio L., Minasny B., Roudier P., Libohova Z., Lagacherie P., Shi Z., Hannam J., Meersmans J., Richer-De-Forges A.C., Walter C. (2022): Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. Geoderma, 409: 115567.

Chen W., Hong H., Li S., Shahabi H., Wang Y., Wang X., Ahmad B.B. (2019): Flood susceptibility modelling using novel hybrid approach of reduced-error pruning trees with bagging and random subspace ensembles. Journal of Hydrology, 575: 864–873.

Choi R.Y., Coyner A.S., Kalpathy-Cramer J., Chiang M.F., Campbell J.P. (2020): Introduction to machine learning, neural networks, and deep learning. Translational Vision Science & Technology, 2: 12.

Cooley R.L., Konikow L.F., Naff R.L. (1986): Nonlinear-regression groundwater flow modeling of a deep regional aquifer system. Water Resources Research, 13: 1759–1778.

Cronican A.E., Gribb M.M. (2004): Hydraulic conductivity prediction for sandy soils. Ground Water, 3: 459–464.

Dai L., Ge J., Wang L., Zhang Q., Liang T., Bolan N., Lischeid G., Rinklebe J. (2022): Influence of soil properties, topography, and land cover on soil organic carbon and total nitrogen concentration: A case study in Qinghai-Tibet plateau based on random forest regression and structural equation modeling. Science of the Total Environment, 821: 153440.

Duffy J., Franklin M. (1973): Case study of environmental system modeling with the group method od data handling. Joint Automatic Control Conference, 11: 101–111.

Eagleson P.S. (1994): The evolution of modern hydrology (from watershed to continent in 30 years). Advances in Water Resources, 1–2: 3–18.

Egghe L. (2006): Theory and practise of the g-index. Scientometrics, 1: 131–152.

Fajardo M., McBratney A., Whelan B. (2016): Fuzzy clustering of Vis-NIR spectra for the objective recognition of soil morphological horizons in soil profiles. Geoderma, 263: 244–253.

Fang K., Shen C., Kifer D., Yang X. (2017): Prolongation of SMAP to spatiotemporally seamless coverage of Continental U.S. using a deep learning neural network. Geophysical Research Letters, 44: 11030–011039.

Fidêncio P.H., Ruisánchez I., Poppi R.J. (2001): Application of artificial neural networks to the classification of soils from São Paulo state using near-infrared spectroscopy. Analyst, 12: 2194–2200.

Gharahi Ghehi N., Nemes A., Verdoodt A., Van Ranst E., Cornelis W.M., Boeckx P. (2012): Nonparametric techniques for predicting soil bulk density of tropical rainforest topsoils in Rwanda. Soil Science Society of America Journal, 4: 1172–1183.

Gharib A., Davies E.G.R. (2021): A workflow to address pitfalls and challenges in applying machine learning models to hydrology. Advances in Water Resources, 152: 103920.

Govindaraju R.S. (2000a): Artificial neural networks in hydrology. I: Preliminary concepts. Journal of Hydrologic Engineering, 2: 115–123.

Govindaraju R.S. (2000b): Artificial neural networks in hydrology. II: Hydrologic applications. Journal of Hydrologic Engineering, 2: 124–137.

Goyal M.K., Sharma A., Katsifarakis K.L. (2017): Prediction of flow rate of karstic springs using support vector machines. Hydrological Sciences Journal, 13: 2175–2186.

Grimaldi S., Schumann G.J.P., Shokri A., Walker J.P., Pauwels V.R.N. (2019): Challenges, opportunities, and pitfalls for global coupled hydrologic-hydraulic modeling of floods. Water Resources Research, 7: 5277–5300.

Han X., Chen X., Feng L. (2015): Four decades of winter wetland changes in Poyang Lake based on Landsat observations between 1973 and 2013. Remote Sensing of Environment, 156: 426–437.

Hansen M.C., Loveland T.R. (2012): A review of large area monitoring of land cover change using Landsat data. Remote Sensing of Environment, 122: 66–74.

https://doi.org/10.17221/94/2022-SWR

Harshbarger J.W., Ferris J.G. (1963): Interdisciplinary training program in scientific hydrology. Groundwater, 2: 11–14.

Hastie T., Friedman J., Tibshirani R. (2001): The Elements of Statistical Learning. New York, Springer.

He H.L., Dyck M., Lv J.L. (2020): The heat pulse method for soil physical measurements: A bibliometric analysis. Applied Sciences-Basel, 18: 15.

Hengl T., De Jesus J.M., Heuvelink G.B.M., Gonzalez M.R., Kilibarda M., Blagotic A., Shangguan W., Wright M.N., Geng X.Y., Bauer-Marschallinger B., Guevara M.A., Vargas R., Macmillan R.A., Batjes N.H., Leenaars J.G.B., Ribeiro E., Wheeler I., Mantel S., Kempen B. (2017): SoilGrids250 m: Global gridded soil information based on machine learning. PLoS ONE, 12: e0169748 .

Hirsch J.E. (2005): An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America, 46: 16569–16572.

Horton R.E. (1933): The relation of hydrology to the botanical sciences. Eos, Transactions American Geophysical Union, 1: 23–25.

Hsu K.L., Gupta H.V., Sorooshian S. (1995): Artificial neural-network modeling of the rainfall-runoff process. Water Resources Research, 10: 2517–2530.

Huang G., Song S.J., Gupta J.N.D., Wu C. (2014): Semi-supervised and unsupervised extreme learning machines. IEEE Transactions on Cybernetics, 12: 2405–2417.

Huang X., Zhang L. (2012): Morphological building/shadow index for building extraction from high-resolution imagery over urban areas. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 1: 161–172.

Huang Y., Lan Y., Thomson S.J., Fang A., Hoffmann W.C., Lacey R.E. (2010): Development of soft computing and applications in agricultural and biological engineering. Computers and Electronics in Agriculture, 2: 107–127.

Ikeda S., Sawaragi Y., Ochiai M. (1976): Sequential GMDH algorithm and its application to river flow prediction. IEEE Transactions on Systems, Man and Cybernetics, 7: 473–479.

Ivakhnenko A.G. (1971): Polynomial theory of complex systems. IEEE Transactions on Systems, Man and Cybernetics, 4: 364–378.

Jafari A., Khademi H., Finke P.A., Van De Wauw J., Ayoubi S. (2014): Spatial prediction of soil great groups by boosted regression trees using a limited point dataset in an arid region, southeastern Iran. Geoderma, 232–234: 148–163.

Japkowicz N. (2001): Supervised versus unsupervised binary-learning by feedforward neural networks. Machine Learning, 1–2: 97–122.

Jenny H. (1941): Factors of Soil Formation, A System of Quantitative Pedology. New York, Dover Publications.

Jingyi Z., Hall M.J. (2004): Regional flood frequency analysis for the Gan-Ming River basin in China. Journal of Hydrology, 1–4: 98–117.

Jordan M.I., Mitchell T.M. (2015): Machine learning: Trends, perspectives, and prospects. Science, 6245: 255–260.

Jung M., Reichstein M., Ciais P., Seneviratne S.I., Sheffield J., Goulden M.L., Bonan G., Cescatti A., Chen J., De Jeu R., Dolman A.J., Eugster W., Gerten D., Gianelle D., Gobron N., Heinke J., Kimball J., Law B.E., Montagnani L., Mu Q., Mueller B., Oleson K., Papale D., Richardson A.D., Roupsard O., Running S., Tomelleri E., Viovy N., Weber U., Williams C., Wood E., Zaehle S., Zhang K. (2010): Recent decline in the global land evapotranspiration trend due to limited moisture supply. Nature, 7318: 951–954.

Kinniburgh D.G. (1986): General purpose adsorption isotherms. Environmental Science and Technology, 9: 895–904.

Kumar A., Ramsankaran R., Brocca L., Muñoz-Arriola F. (2021): A simple machine learning approach to model real-time streamflow using satellite inputs: Demonstration in a data scarce catchment. Journal of Hydrology, 595: 126046.

Lane P.W. (2002): Generalized linear models in soil science. European Journal of Soil Science, 2: 241–251.

Le Gratiet L., Garnier J. (2015): Asymptotic analysis of the learning curve for Gaussian process regression. Machine Learning, 3: 407–433.

LeCun Y., Bengio Y., Hinton G. (2015): Deep learning. Nature, 7553: 436–444.

Lee C.H., Shin D.G. (1999): Using Hellinger distance in a nearest neighbour classifier for relational databases. Knowledge-Based Systems, 7: 363–370.

Lee S., Ryu J.H., Min K., Won J.S. (2003): Landslide susceptibility analysis using GIS and artificial neural network. Earth Surface Processes and Landforms, 12: 1361–1376.

Li Y., Shi Z., Li F., Li H.Y. (2007): Delineation of site-specific management zones using fuzzy clustering analysis in a coastal saline land. Computers and Electronics in Agriculture, 2: 174–186.

Li Y., Zhao Z., Wei S., Sun D., Yang Q., Ding X. (2021): Prediction of regional forest soil nutrients based on gaofen-1 remote sensing data. Forests, 12: 1430.

Ließ M., Glaser B., Huwe B. (2012): Uncertainty in the spatial prediction of soil texture. Comparison of regression tree and Random Forest models. Geoderma, 170: 70–79.

Liu Z., Huang S.L., Jin W., Mu Y. (2021): Broad learning system for semi-supervised learning. Neurocomputing, 444: 38–47.

Loganathan P., Mahindrakar A.B. (2021): Intercomparing the robustness of machine learning models in simulation and forecasting of streamflow. Journal of Water and Climate Change, 5: 1824–1837.

https://doi.org/10.17221/94/2022-SWR

Ma Y.X., Minasny B., Malone B.P., McBratney A.B. (2019): Pedology and digital soil mapping (DSM). European Journal of Soil Science, 2: 216–235.

Maulik U., Bandyopadhyay S. (2000): Genetic algorithm-based clustering technique. Pattern Recognition, 9: 1455–1465.

McBratney A., De Gruijter J., Bryce A. (2019): Pedometrics timeline. Geoderma, 338: 568–575.

McBratney A.B., Minasny B., Cattle S.R., Vervoort R.W. (2002): From pedotransfer functions to soil inference systems. Geoderma, 1–2: 41–73.

McBratney A.B., Mendonça Santos M.L., Minasny B. (2003): On digital soil mapping. Geoderma, 1–2: 3–52.

Minns A.W., Hall M.J. (1996): Artificial neural networks as rainfall-runoff models. Hydrological Sciences Journal, 3: 399–417.

Mishina Y., Murata R., Yamauchi Y., Yamashita T., Fujiyoshi H. (2015): Boosted random forest. IEICE Transactions on Information and Systems, 9: 1630–1636.

Mittermeier M., Braun M., Hofstätter M., Wang Y., Ludwig R.. (2019): Detecting climate change effects on Vb cyclones in a 50-member single-model ensemble using machine learning. Geophysical Research Letters, 24: 14653–14661.

Mjolsness E., Decoste D. (2001): Machine learning for science: State of the art and future prospects. Science, 5537: 2051–2055.

Moran C.J., Bui E.N. (2002): Spatial data mining for enhanced soil map modelling. International Journal of Geographical Information Science, 6: 533–549.

Mukerji A., Chatterjee C., Singh Raghuwanshi N. (2009): Flood forecasting using ANN, neuro-fuzzy, and neuro-GA models. Journal of Hydrologic Engineering, 6: 647–652.

Nemmour H., Chibani Y. (2006): Multiple support vector machines for land cover change detection: An application for mapping urban extensions. ISPRS Journal of Photogrammetry and Remote Sensing, 2: 125–133.

Pachepsky Y.A., Timlin D.J., Rawls W.J. (2001): Soil water retention as related to topographic variables. Soil Science Society of America Journal, 6: 1787–1795.

Pappenberger F., Beven K.J., Hunter N.M., Bates P.D., Gouweleeuw B.T., Thielen J., De Roo A.P.J. (2005): Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS). Hydrology and Earth System Sciences, 4: 381–393.

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.

Plasek A. (2016): On the cruelty of really writing a history of machine learning. IEEE Annals of the History of Computing, 4: 6–8.

Qiu L., Wang K., Long W., Wang K., Hu W., Amable G.S. (2016): A comparative assessment of the influences of human impacts on soil Cd concentrations based on stepwise linear regression, classification and regression tree, and random forest models. PLoS ONE, 11: e0151131.

Rindfuss R.R., Walsh S.J., Turner Ii B.L., Fox J., Mishra V. (2004): Developing a science of land change: Challenges and methodological issues. Proceedings of the National Academy of Sciences of the United States of America, 39: 13976–13981.

Rossiter D.G. (2018): Past, present & future of information technology in pedometrics. Geoderma, 342: 131–137.

Rudin C., Wagstaff K.L. (2014): Machine learning for science and society. Machine Learning, 1: 1–9.

Savic D.A., Walters G.A., Davidson J.W. (1999): A genetic programming approach to rainfall-runoff modelling. Water Resources Management, 3: 219–231.

Schaap M.G., Leij F.J., van Genuchten M.T. (2001): ROSETTA: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. Journal of Hydrology, 3–4: 163–176.

Shadbolt N., Hall W., Berners-Lee T. (2006): The semantic Web revisited. IEEE Intelligent Systems, 3: 96–101.

Sharifi A., Dinpashoh Y., Mirabbasi R. (2017): Daily runoff prediction using the linear and non-linear models. Water Science and Technology, 4: 793–805.

Sireesha Naidu G., Pratik M., Rehana S. (2020): Modelling hydrological responses under climate change using machine learning algorithms – Semi-arid river basin of peninsular India. H_2Open Journal, 1: 481–498.

Taghizadeh-Mehrjardi R., Schmidt K., Amirian-Chakan A., Rentschler T., Zeraatpisheh M., Sarmadian F., Valavi R., Davatgar N., Behrens T., Scholten T. (2020): Improving the spatial prediction of soil organic carbon content in two contrasting climatic regions by stacking machine learning models and rescanning covariate space. Remote Sensing, 12: 1095.

Tajik S., Ayoubi S., Nourbakhsh F. (2012): Prediction of soil enzymes activity by digital terrain analysis: Comparing artificial neural network and multiple linear regression models. Environmental Engineering Science, 8: 798–806.

Tan Q.F., Lei X.H., Wang X., Wang H., Wen X., Ji Y., Kang A.Q. (2018): An adaptive middle and long-term runoff forecast model using EEMD-ANN hybrid approach. Journal of Hydrology, 567: 767–780.

Tien Bui D., Hoang N.D., Martínez-Álvarez F., Ngo P.T.T., Hoa P.V., Pham T.D., Samui P., Costache R. (2020): A novel deep learning neural network approach for predicting flash flood

https://doi.org/10.17221/94/2022-SWR

susceptibility: A case study at a high frequency tropical storm area. Science of the Total Environment, 701: 134413.
Usama M., Qadir J., Raza A., Arif H., Yau K.L.A, Elkhatib Y., Hussain A., Al-Fuqaha A. (2019): Unsupervised machine learning for networking: techniques, applications and research challenges. IEEE Access, 78713992: 65579–65615.
Van Eck N.J., Waltman L. (2010): Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 2: 523–538.
Van Engelen J.E., Hoos H.H. (2020): A survey on semi-supervised learning. Machine Learning, 2: 373–440.
Vilchez-Roman C. (2014): Bibliometric factors associated with h-index of Peruvian researchers with publications indexed on Web of Science and Scopus databases. Transinformacao, 2: 143–154.
Wadoux A.M.J.C., Minasny B., McBratney A.B. (2020): Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews, 210: 103359.
Wang H.B., Xu W.Y., Xu R.C. (2005): Slope stability evaluation using Back Propagation Neural Networks. Engineering Geology, 3–4: 302–315.
Wang N., Xue J., Peng J., Biswas A., He Y., Shi Z. (2020): Integrating remote sensing and landscape characteristics to estimate soil salinity using machine learning methods: A case study from Southern Xinjiang, China. Remote Sensing, 12: 1–21.
Weng Q.H., Lu D.S., Schubring J. (2004): Estimation of land surface temperature-vegetation abundance relationship for urban heat island studies. Remote Sensing of Environment, 4: 467–483.
Xie H.L., Zhang Y.W., Wu Z.L., Lv T.G. (2020): A bibliometric analysis on land degradation: Current status, development, and future directions. Land, 9: 28.
Xu Y.Y., Zhou Y., Sekula P., Ding L.Y. (2021): Machine learning in construction: From shallow to deep learning. Developments in the Built Environment, 6: 100045.
Yang J., Wang X., Wang R., Wang H. (2020): Combination of convolutional neural networks and recurrent neural networks for predicting soil properties using Vis-NIR spectroscopy. Geoderma, 380: 114616.
Yuan Q., Shen H., Li T., Li Z., Li S., Jiang Y., Xu H., Tan W., Yang Q., Wang J., Gao J., Zhang L. (2020): Deep learning in environmental remote sensing: Achievements and challenges. Remote Sensing of Environment, 241: 111716.
Zeraatpisheh M., Jafari A., Bagheri Bodaghabadi M., Ayoubi S., Taghizadeh-Mehrjardi R., Toomanian N., Kerry R., Xu M. (2020): Conventional and digital soil mapping in Iran: Past, present, and future. Catena, 188: 104424.
Zhang H., Wu P., Yin A., Yang X., Zhang M., Gao C. (2017): Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: A comparison of multiple linear regressions and the random forest model. Science of the Total Environment, 592: 704–713.
Zhang H.L., Liu X.Y., Yi J., Yang X.F., Wu T.I., He Y., Duan H., Liu M.X., Tian P. (2020a): Bibliometric analysis of research on soil water from 1934 to 2019. Water, 12: 1631.
Zhang T., Wang C.J., Liu S.Y., Zhang N., Zhang T.W. (2020b): Assessment of soil thermal conduction using artificial neural network models. Cold Regions Science and Technology, 169: 102907.
Zhu A.X., Band L.E. (1994): A knowledge-based approach to data integration for soil mapping. Canadian Journal of Remote Sensing, 4: 408–418.
Zhu R., Yang L., Liu T., Wen X., Zhang L., Chang Y. (2019): Hydrological responses to the future climate change in a data scarce region, northwest China: Application of machine learning models. Water (Switzerland), 11: 1588.
Zhu S.N., Lu H.F., Ptak M., Dai J.Y., Ji Q.F. (2020): Lake water-level fluctuation forecasting using machine learning models: A systematic review. Environmental Science and Pollution Research, 36: 44807–44819.
Zolfaghari Z., Mosaddeghi M.R., Ayoubi S. (2015): ANN-based pedotransfer and soil spatial prediction functions for predicting Atterberg consistency limits and indices from easily available properties at the watershed scale in western Iran. Soil Use and Management, 1: 142–154.

Received: June 26, 2022
Accepted: February 9, 2023
Online first: March 22, 2023