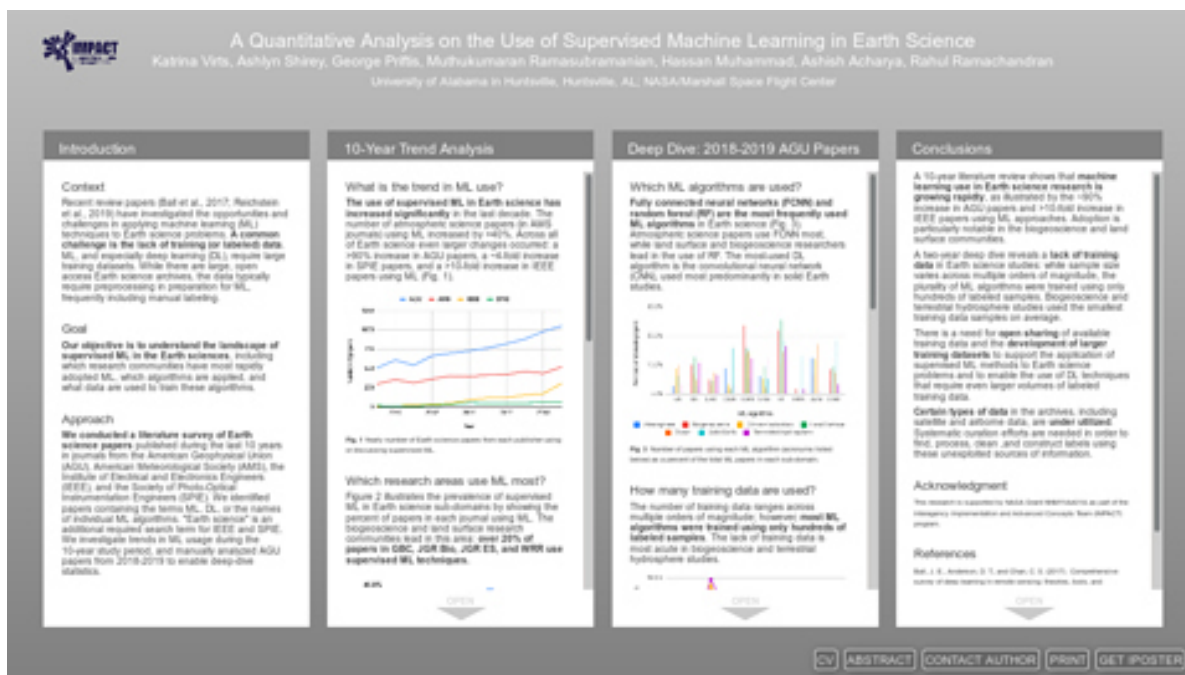


A Quantitative Analysis on the Use of Supervised Machine Learning in Earth Science



Katrina Virts, Ashlyn Shirey, George Priftis, Muthukumaran Ramasubramanian, Hassan Muhammad, Ashish Acharya, Rahul Ramachandran

University of Alabama in Huntsville, Huntsville, AL; NASA/Marshall Space Flight Center

PRESENTED AT:



INTRODUCTION

Context

Recent review papers (Ball et al., 2017; Reichstein et al., 2019) have investigated the opportunities and challenges in applying supervised machine learning (ML) techniques to Earth science problems. **A common challenge is the lack of training (or labeled) data.** Supervised ML, and especially deep learning (DL), require large training datasets. While there are large, open access Earth science archives, the data typically require preprocessing in preparation for supervised ML, frequently including manual labeling.

Goal

Our objective is to understand the landscape of supervised ML in the Earth sciences, including which research communities have most rapidly adopted supervised ML, which algorithms are applied, and what data are used to train these algorithms.

Approach

We conducted a literature survey of Earth science papers published during the last 10 years in journals from the American Geophysical Union (AGU), American Meteorological Society (AMS), the Institute of Electrical and Electronics Engineers (IEEE), and the Society of Photo-Optical Instrumentation Engineers (SPIE). We identified papers containing the terms ML, DL, or the names of individual supervised ML algorithms. "Earth science" is an additional required search term for IEEE and SPIE. We investigate trends in supervised ML usage during the 10-year study period, and manually analyzed AGU papers from 2018-2019 to enable deep-dive statistics.

10-YEAR TREND ANALYSIS

What is the trend in the use of supervised ML?

The use of supervised ML in Earth science has increased significantly in the last decade. The number of atmospheric science papers (in AMS journals) using supervised ML increased by >40%. Across all of Earth science even larger changes occurred: a >90% increase in AGU papers, a ~4-fold increase in SPIE papers, and a >10-fold increase in IEEE papers using supervised ML (Fig. 1).

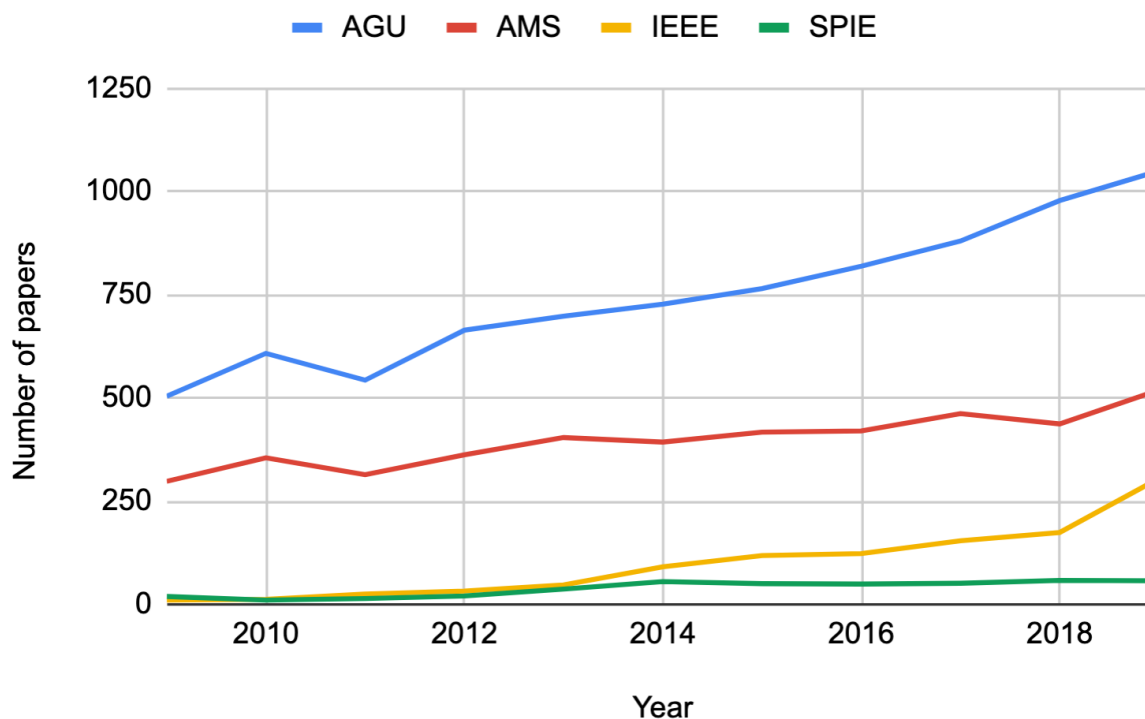


Fig. 1 Yearly number of Earth science papers from each publisher using or discussing supervised ML.

Which research areas use supervised ML most?

Figure 2 illustrates the prevalence of supervised ML in Earth science sub-domains by showing the percent of papers in each journal using supervised ML. The biogeoscience and land surface research communities lead in this area: **over 20% of papers in GBC, JGR Bio, JGR ES, and WRR use supervised ML techniques.**

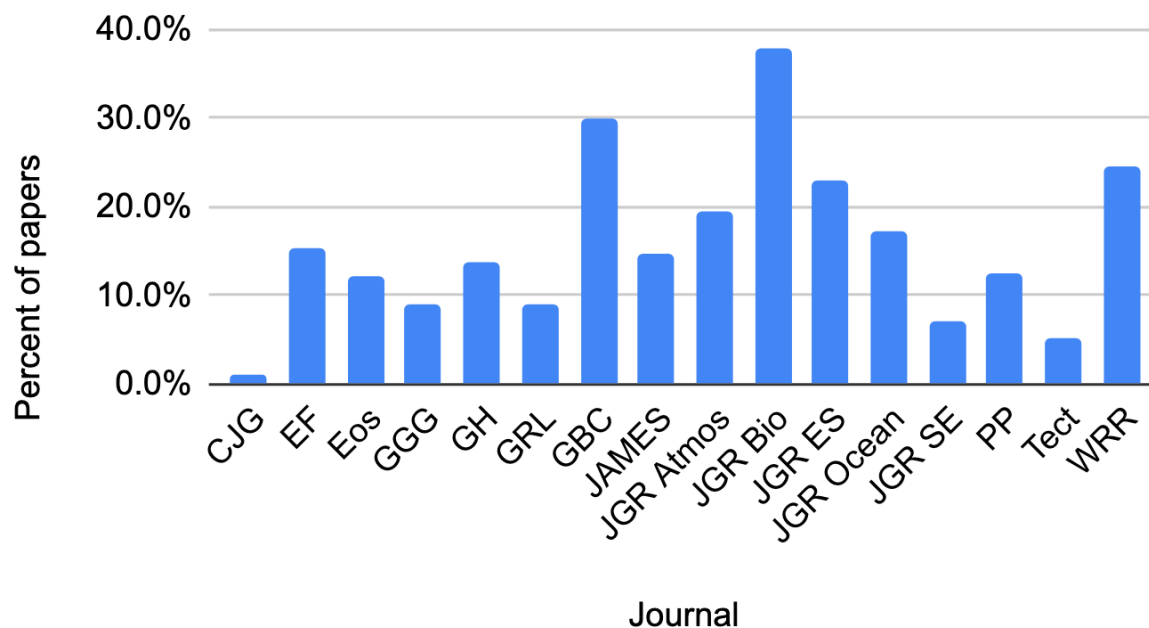


Fig. 2 Percent of papers in each journal (acronyms listed below) from 2009-2019 using supervised ML.

The animation below shows the leading keywords selected by authors of AGU articles using supervised ML techniques. **Climate change was the leading keyword each year**, illustrating the application of these emerging techniques to this significant research area. Other prominent keywords include remote sensing, precipitation, ENSO, evapotranspiration, and Arctic.

[VIDEO] <https://www.youtube.com/embed/raEHGMpEac?rel=0&fs=1&modestbranding=1&rel=0&showinfo=0>

Acronyms

CJG = Chinese Journal of Geophysics

EF = Earth's Future

GGG = Geochemistry, Geophysics, Geosystems

GH = GeoHealth

GRL = Geophysical Research Letters

GBC = Global Biogeochemical Cycles

JAMES = Journal of Advances in Modeling Earth Systems

JGR Atmos = Journal of Geophysical Research: Atmospheres

JGR Bio = Journal of Geophysical Research: Biogeosciences

JGR ES = Journal of Geophysical Research: Earth Surface

JGR Ocean = Journal of Geophysical Research: Oceans

JGR SE = Journal of Geophysical Research: Solid Earth

PP = Paleoceanography and Paleoclimatology

Tect = Tectonics

WRR = Water Resources Research

DEEP DIVE: 2018-2019 AGU PAPERS

Which supervised ML algorithms are used?

Fully connected neural networks (FCNN) and random forest (RF) are the most frequently used ML algorithms in Earth science (Fig. 3). Atmospheric science papers use FCNN most, while land surface and biogeoscience researchers lead in the use of RF. **The most-used DL algorithm is the convolutional neural network (CNN)**, used most predominantly in solid Earth studies.

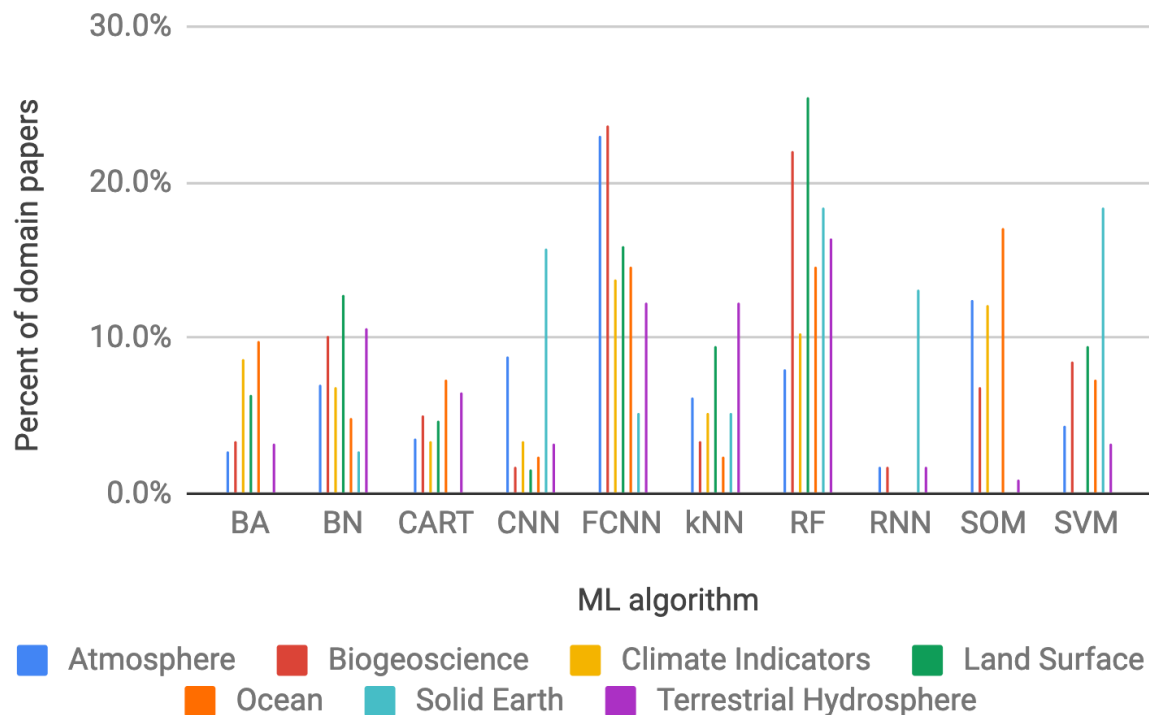


Fig. 3 Number of papers using each supervised ML algorithm (acronyms listed below) as a percent of the total supervised ML papers in each sub-domain.

How many training data are used?

The number of training data ranges across multiple orders of magnitude; however, **most supervised ML algorithms were trained using only hundreds of labeled samples**. The lack of training data is most acute in biogeoscience and terrestrial hydrosphere studies.

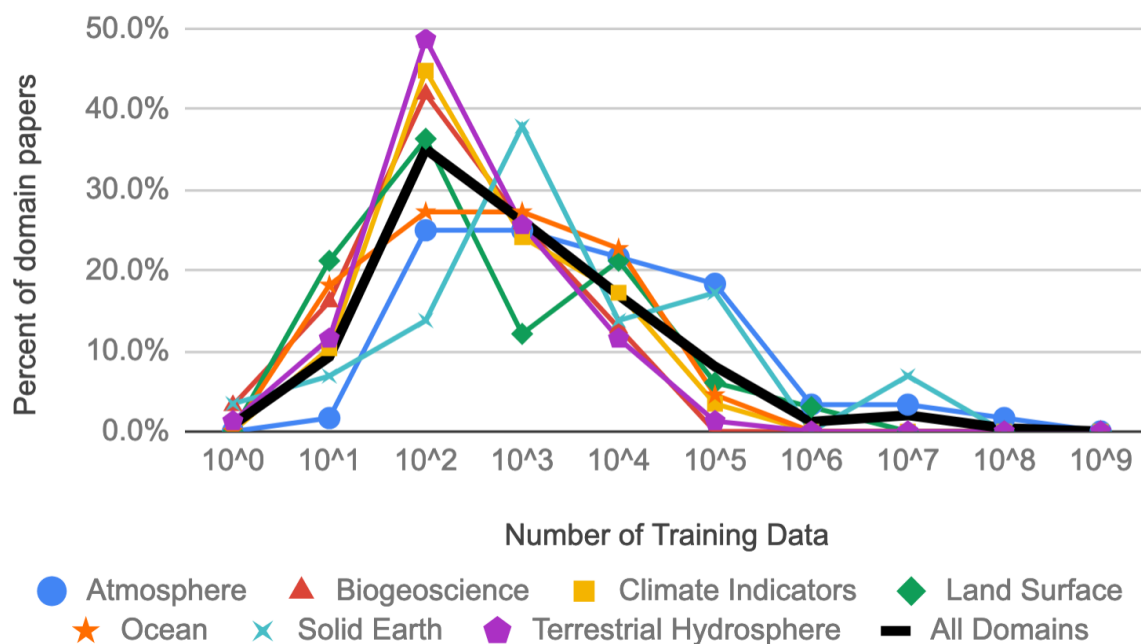


Fig. 4 Number of labeled data used to train supervised ML algorithms.

What kinds of training data are used?

In situ data (e.g., meteorological station data, streamflow gauges, buoys, seismic waveforms) is **most commonly used as training data for most domains**, while model output is used frequently in climate studies. Land surface and ocean researchers are most likely to use satellite data.

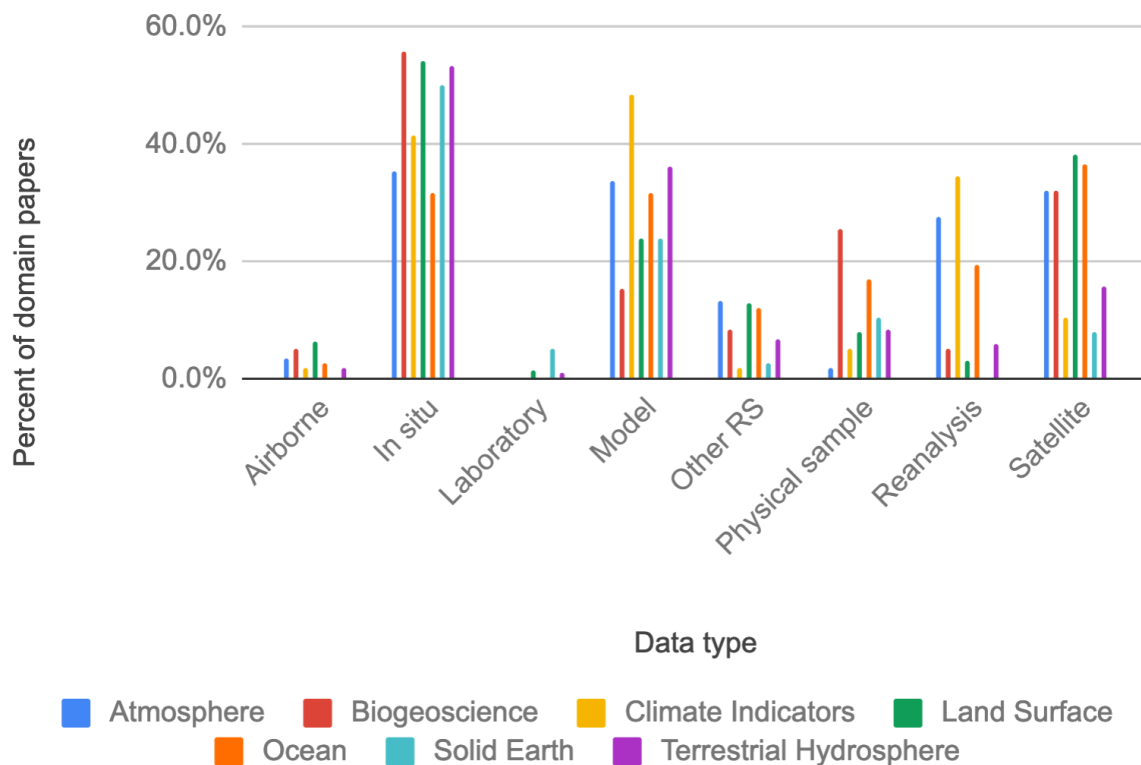


Fig. 5 Type of data used to train supervised ML algorithms.

How are the training data labeled?

Overall, ~90% of papers use automatically labeled training data. Solid Earth studies are the outlier, in which semi-automatic and manual labeling methods are used in about 40% of papers.

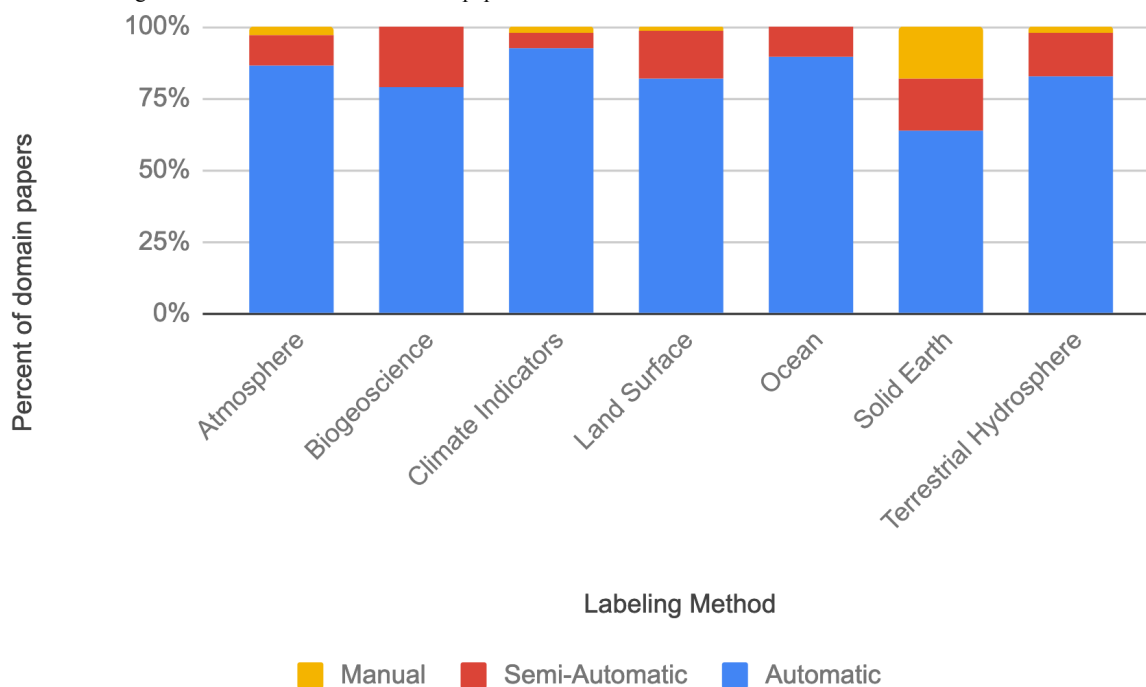


Fig. 6 Labeling method of data used to train supervised ML algorithms.

Acronyms

BA = bootstrapped aggregation (bagging)

BN = Bayesian network

CART = classification and regression tree

CNN = convolutional neural network

FCNN = fully connected neural network

kNN = k-nearest neighbor

RF = random forest

RNN = recurrent neural network

SOM = self-organizing map

SVM = support vector machines

CONCLUSIONS

A 10-year literature review shows that **supervised machine learning use in Earth science research is growing rapidly**, as illustrated by the ~90% increase in AGU papers and >10-fold increase in IEEE papers using supervised ML approaches. Adoption is particularly notable in the biogeoscience and land surface communities.

A two-year deep dive reveals a **lack of training data** in Earth science studies: while sample size varies across multiple orders of magnitude, the plurality of supervised ML algorithms were trained using only hundreds of labeled samples. Biogeoscience and terrestrial hydrosphere studies used the smallest training data samples on average.

There is a need for **open sharing** of available training data and the **development of larger training datasets** to support the application of supervised ML methods to Earth science problems and to enable the use of DL techniques that require even larger volumes of labeled training data.

Certain types of data in the archives, including satellite and airborne data, are **under utilized**. Systematic curation efforts are needed in order to find, process, clean, and construct labels using these unexploited sources of information.

Acknowledgment

This research is supported by NASA Grant NNM11AA01A as part of the Interagency Implementation and Advanced Concepts Team (IMPACT) program.

References

Ball, J. E., Anderson, D. T., and Chan, C. S. (2017). Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges from the community. *Journal of Applied Remote Sensing*, 11(4), <https://doi.org/10.1117/1.JRS.11.042609>

Reichstein, M., Camps-Valls, G., Stevens, B., et al. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, 566, <https://doi.org/10.1038/s41586-019-0912-1>

AUTHOR INFORMATION

Katrina Virts¹, Ashlyn Shirey¹, George Priftis¹, Muthukumaran Ramasubramanian¹, Hassan Muhammad¹, Ashish Acharya¹,
Rahul Ramachandran²

ABSTRACT

Review articles on the application of machine learning (ML) techniques to Earth science problems commonly cite the lack of labeled training data as a challenge to supervised learning. Our objective is to understand the landscape of supervised ML within the Earth sciences. To that end, we have analyzed 10 years of articles published in AGU, AMS, IEEE, and SPIE journals to understand trends in the application of supervised ML techniques to Earth science problems. We have further performed intensive analysis of AGU papers from 2018-2019, extracting information such as the Earth science sub-domain, ML algorithm(s) applied, number and type of training data, and methods used to label the training data. We find that ML use in the Earth sciences is growing rapidly, but the plurality of ML algorithms were trained using only hundreds of labeled samples. Our results underscore the lack of training data across the Earth science research community and the need for open sharing of the available data and the development of large training datasets.