

A QUANTITATIVE ANALYSIS ON THE USE OF SUPERVISED MACHINE LEARNING IN EARTH SCIENCE

Katrina Virts, Ashlyn Shirey, George Priftis, Muthukumaran Ramasubramanian,
Ankur Kumar, Hassan Muhammad, Ashish Acharya

University of Alabama in Huntsville

Rahul Ramachandran
NASA/MSFC

AMS Annual Meeting 2021



Overview

Context

Recent review papers^[1, 2] have investigated the challenges in applying machine learning (ML) techniques to Earth science problems. These challenges listed range from interpretability of the results to computational demand to **data issues**.

Goal

In this study, we seek to understand ***the landscape of supervised ML in Earth science research***, with a focus on the ***number and type of labeled data*** used to train the ML and deep learning (DL) algorithms.

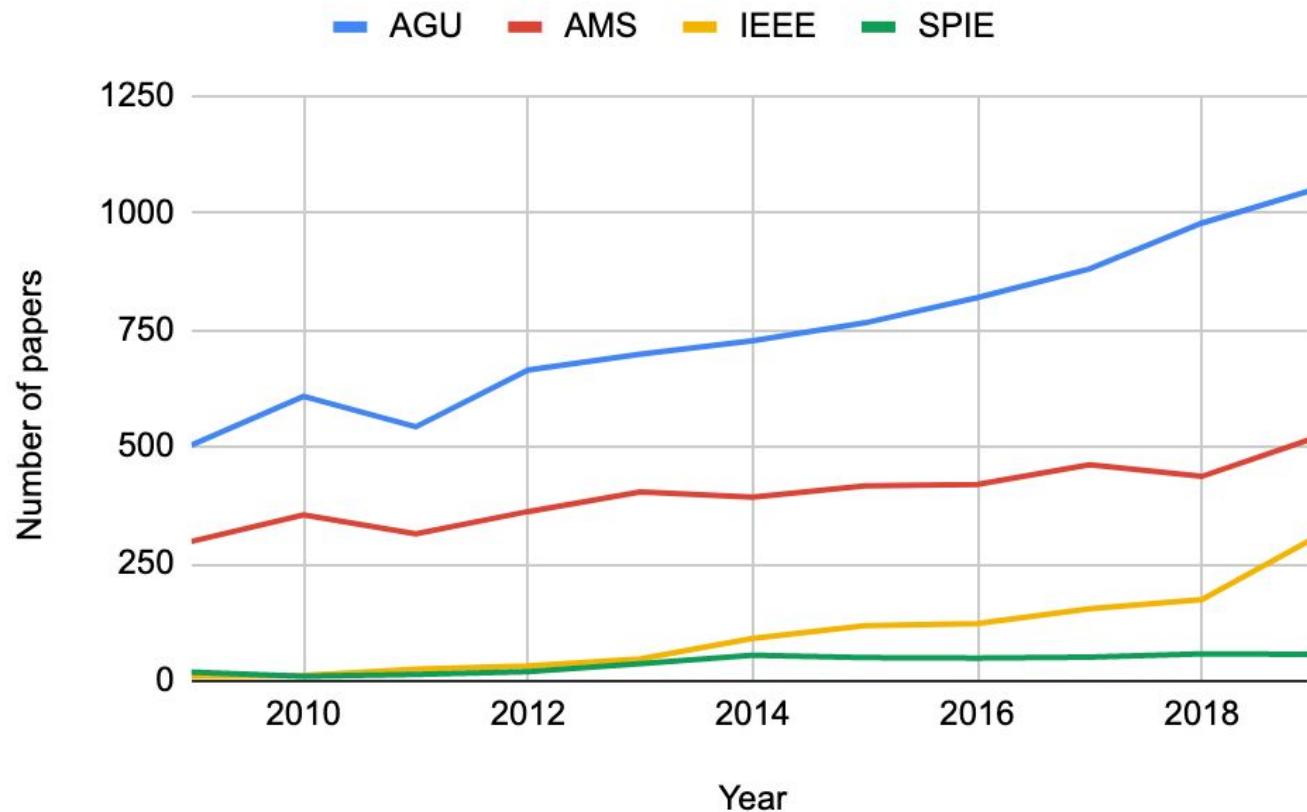
Approach

We have ***conducted a literature survey to better understand these challenges***. We analyze the 10-year trend in ***AGU, AMS, IEEE and SPIE*** Earth science journal papers that use supervised ML techniques, and conduct deep-dive analysis into ~2,000 AGU papers from 2018-2019 to investigate the use of ML algorithms and the ***number and type of training data***.

10-Year Overview

10-Year Trend

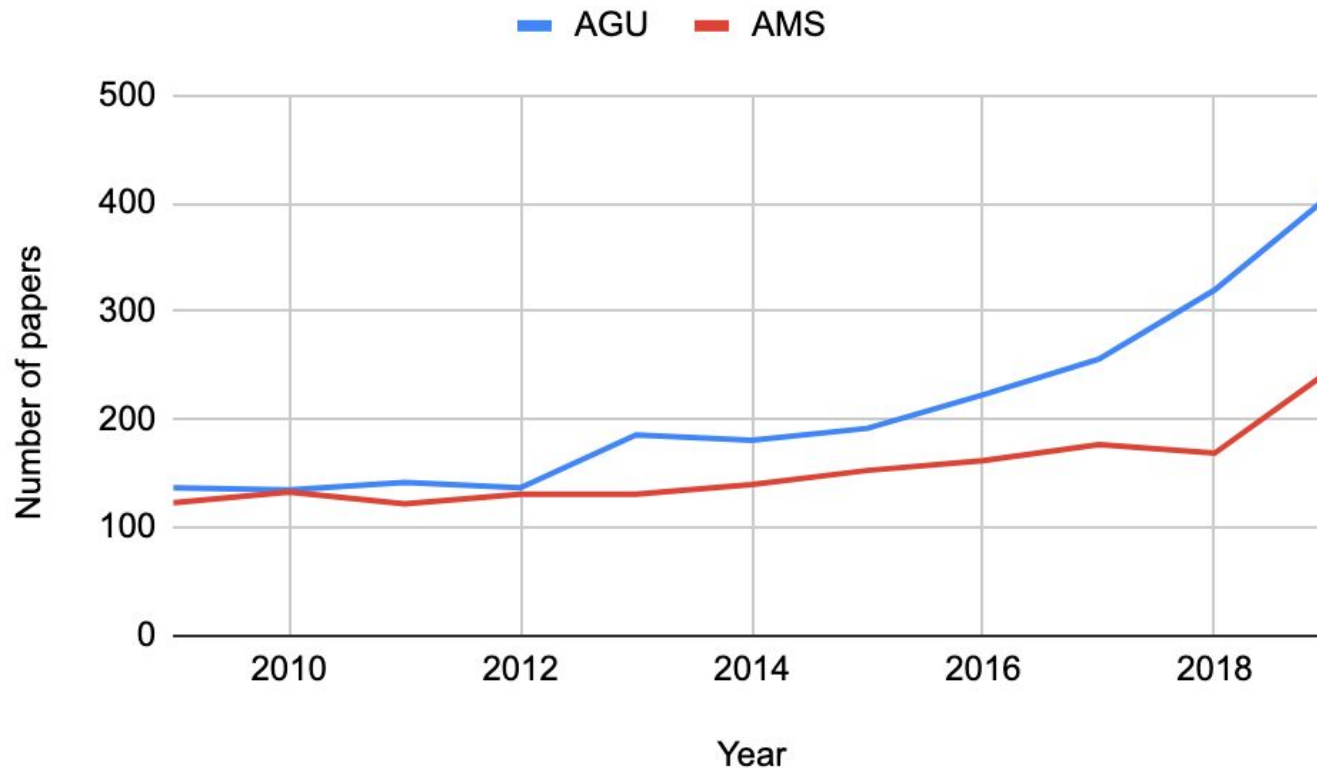
2009-2019 Trend in Earth Science Papers using Supervised Machine Learning Techniques



The use of supervised machine learning techniques in Earth science research has **increased significantly in the last decade**. The number of atmospheric science papers (i.e., from AMS journals) using ML approaches has increased by over 60%. Across all of Earth science even larger changes have occurred, including a >90% increase in AGU papers and a >10-fold increase in IEEE papers using ML.

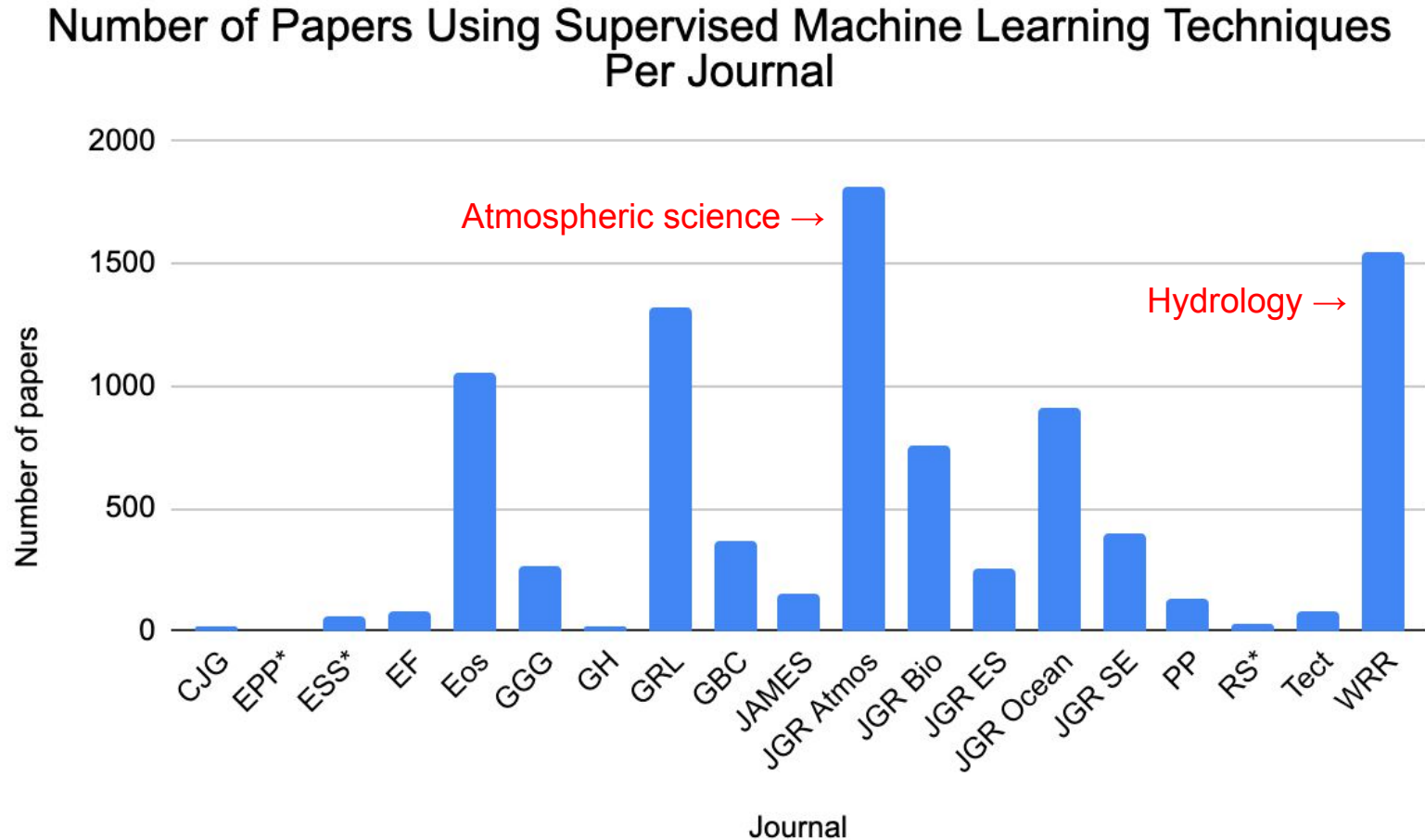
10-Year Trend

2009-2019 Trend in Earth Science Papers Using Supervised Machine Learning Techniques (Excluding Linear Regression)



Linear regression is the most established of the supervised ML techniques included in this review; it is used in ~75% of the surveyed papers. This figure shows the **10-year trend for ML algorithms other than linear regression**. Initially, the number of AGU and AMS papers is similar, but the increasing separation between them indicates the **more rapid adoption of newer ML methods by the broader Earth science community**. This motivates us to look deeper at ML usage within various Earth science sectors.

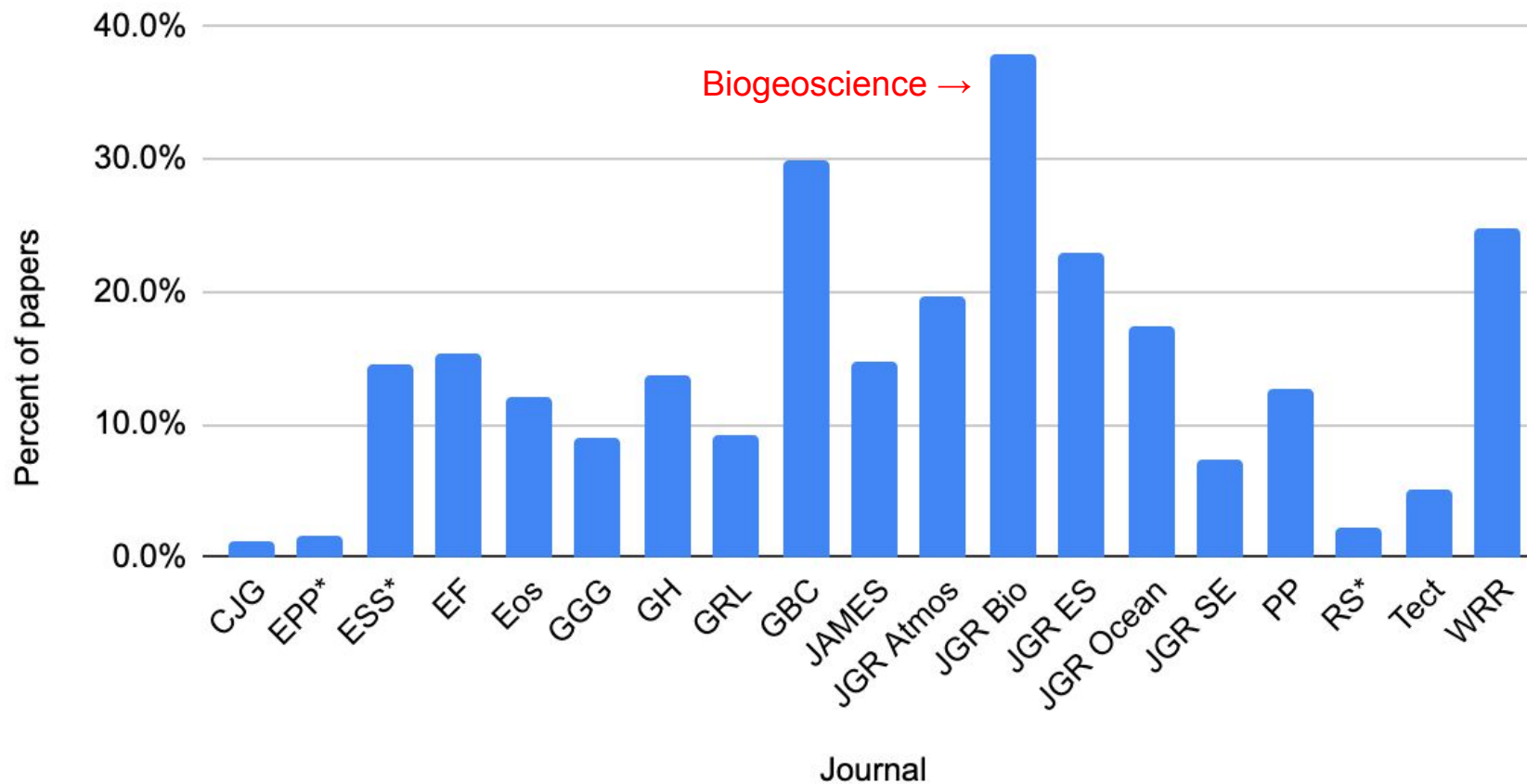
ML Use in Earth Science Sub-Domains



This chart shows the number of papers from each journal that use supervised ML techniques. By this metric, atmospheric science, hydrology, and ocean science have most frequently used supervised ML. But this chart does not tell the whole story.

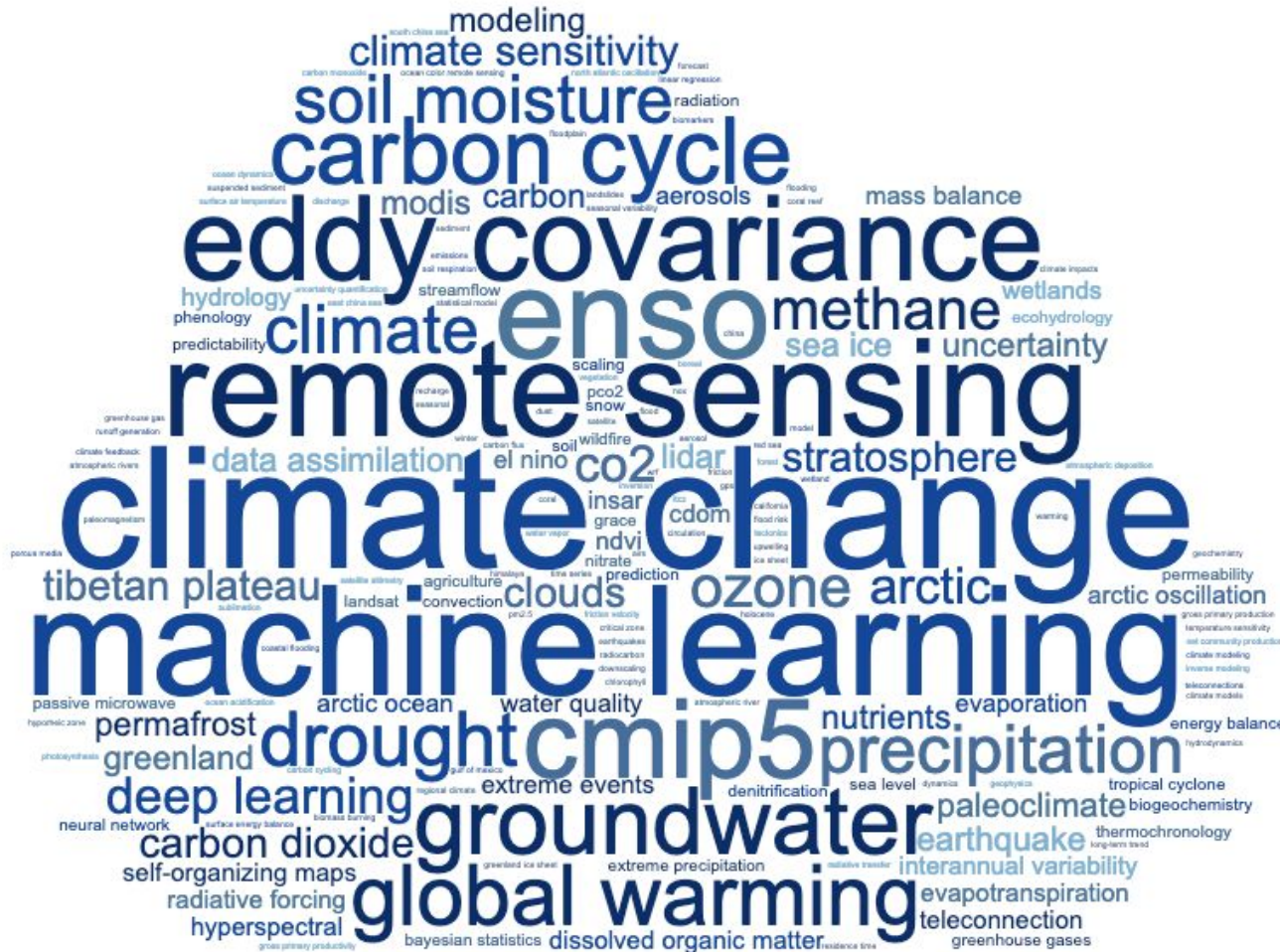
ML Use in Earth Science Sub-Domains

Percent of Papers Using Supervised Machine Learning Techniques Per Journal



In this chart, the number of supervised ML papers from the previous chart is divided by the total number of papers published in each journal during the 10-year window covered by the literature survey. This gives a view of the **prevalence of the use of supervised ML in each research domain**. The biogeoscience and land surface research communities lead in this area: **over 20% of papers published in Global Biogeochemical Cycles, JGR Biogeosciences, JGR Earth Surface, and Water Resources Research use supervised ML techniques**, including over 35% of the papers in JGR **Biogeosciences**.

Keywords in ML Papers

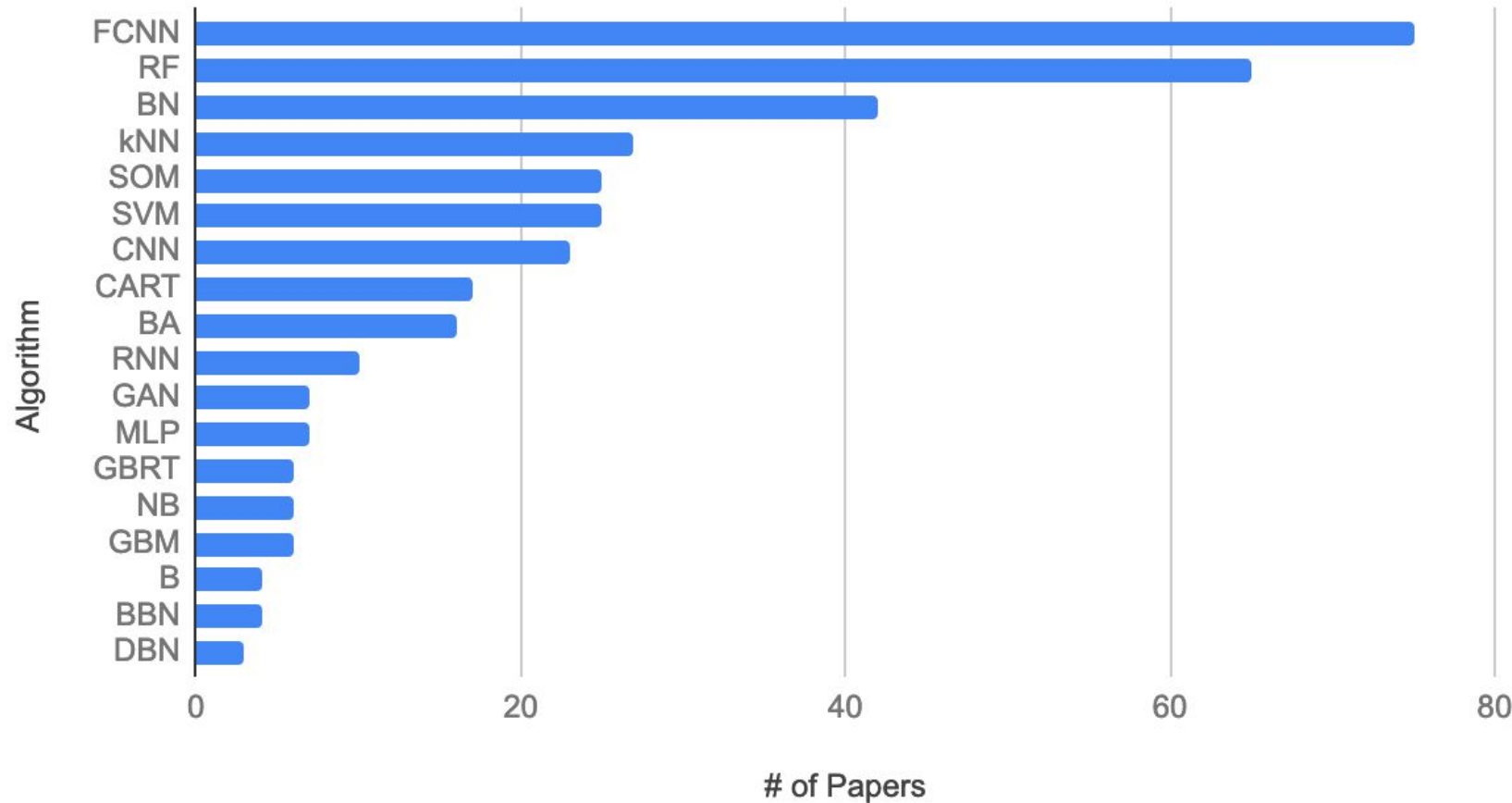


AGU authors choose keywords for their paper to represent its major topics and make it more discoverable. For papers using supervised ML techniques, **climate change was the most frequent chosen keyword for each year** from 2009-2019, illustrating the **application of these emerging techniques to this significant research area**. ~20% of ML papers from the two-year deep dive included climate change as a keyword.

2-Year Deep Dive

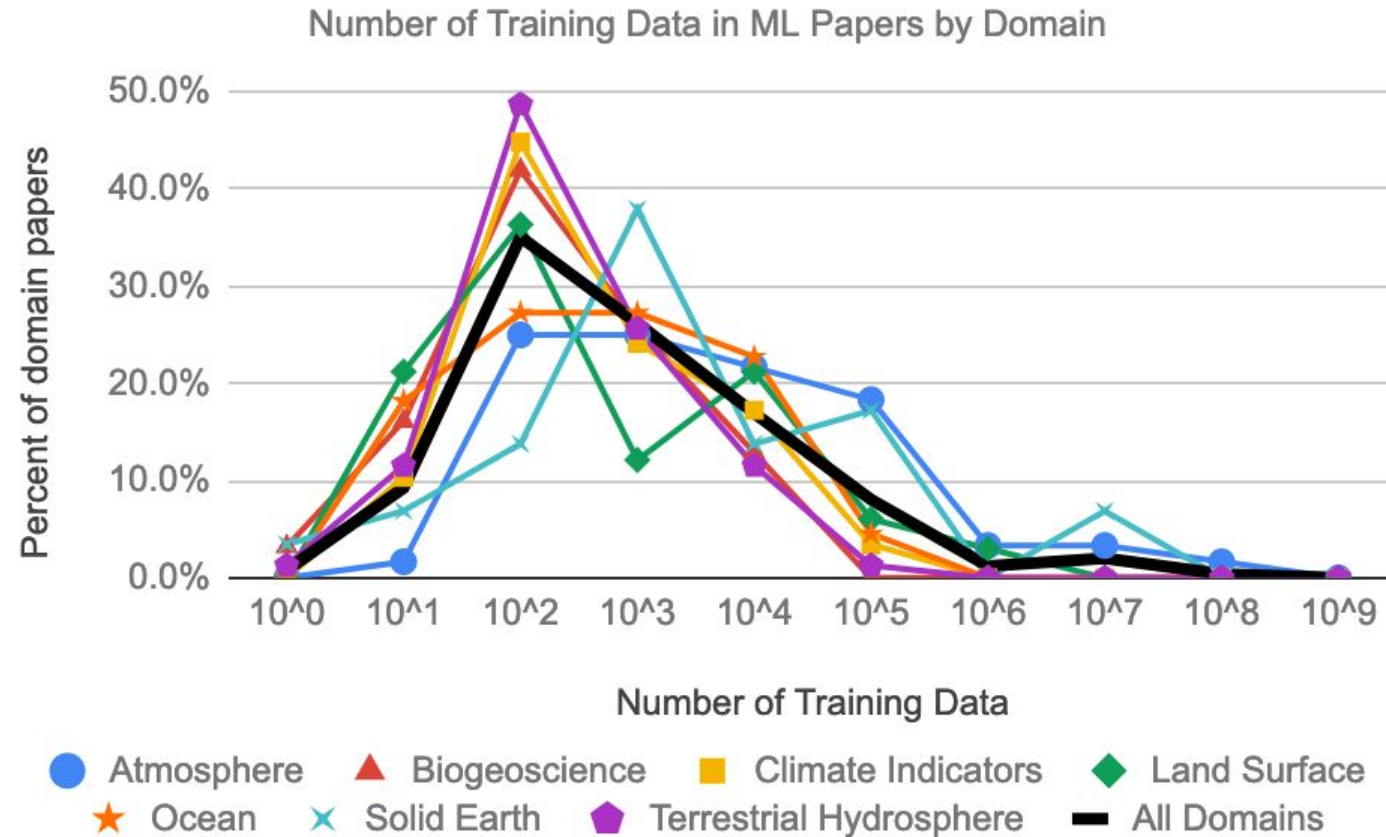
ML Algorithms Used

Supervised ML Algorithms Used in Earth Science



Ensemble, instance-based, neural networks, Bayesian, and deep learning tools are most applicable for addressing Earth science problems. **Fully connected neural networks** (FCNN) are the most frequently used techniques, in 18% of the analyzed papers and particularly in atmospheric science studies. The ensemble algorithm **random forest** (RF) is used in 15% of all supervised ML papers and is one of the most frequent algorithms in each Earth science sub-domain. Others frequently used include Bayesian networks (BN), self-organizing maps (SOM), k-nearest neighbor (kNN), and support vector machines (SVM). Of **deep learning** (DL) algorithms, **convolutional neural networks** (CNN) are used most frequently, in ~5% of the analyzed papers.

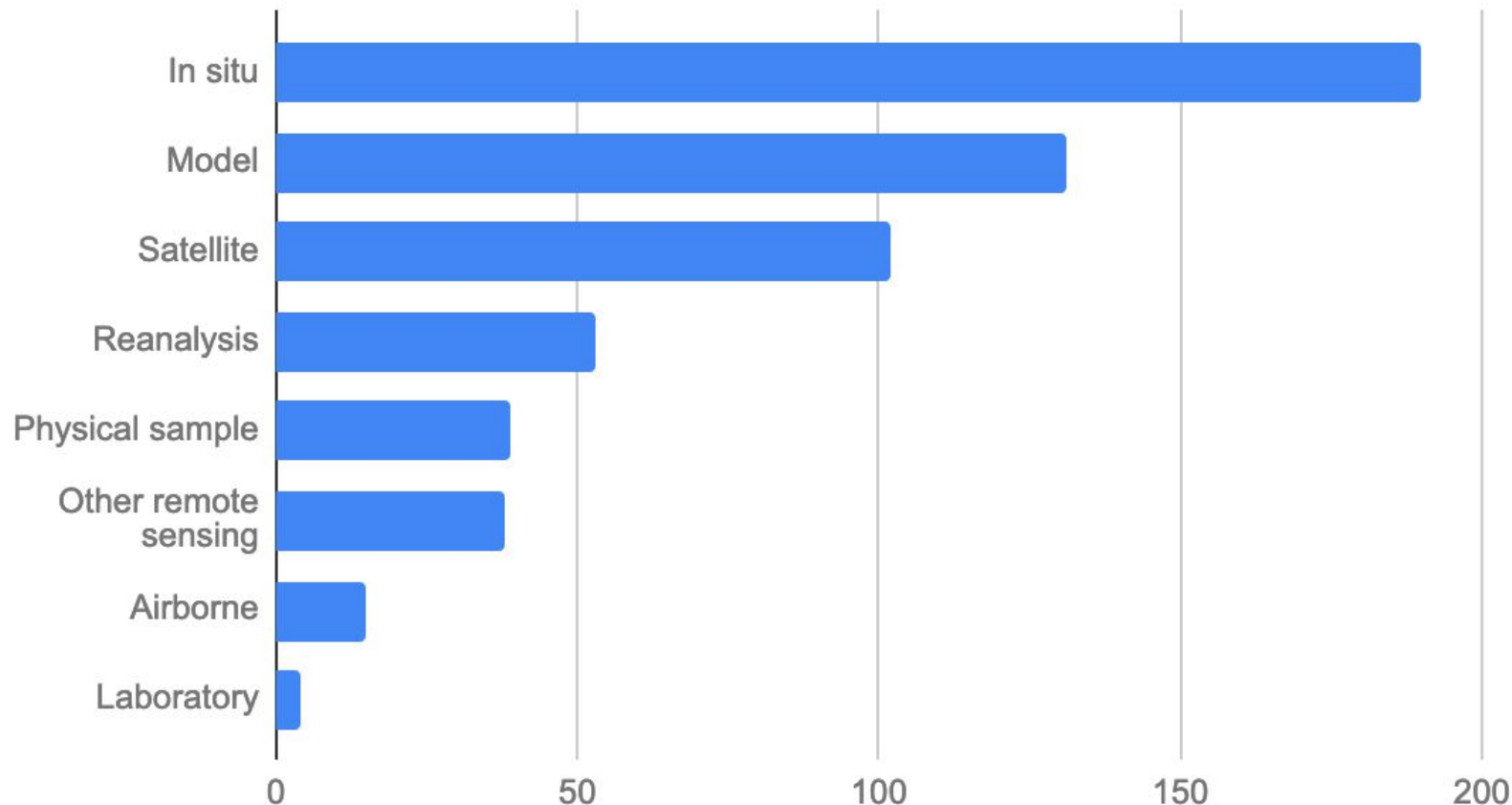
Number of Training Data



The availability of labeled training data in Earth science is reflected in the number of training data used in supervised analysis. In addition to the 10-year analysis above, we have **manually extracted from AGU papers from the years 2018-2019** other relevant information including the **ML algorithm** that was applied, the **number of labeled training data**, and the **data type** (model output, satellite, in situ, reanalysis, etc.). In the papers we surveyed, most **ML algorithms were trained using hundreds of labeled samples**. However, for some applications using model output or large, established datasets, the number of training data ranged several orders of magnitude greater. The lack of training data is particularly **acute in biogeoscience and terrestrial hydrosphere (hydrology) studies**.

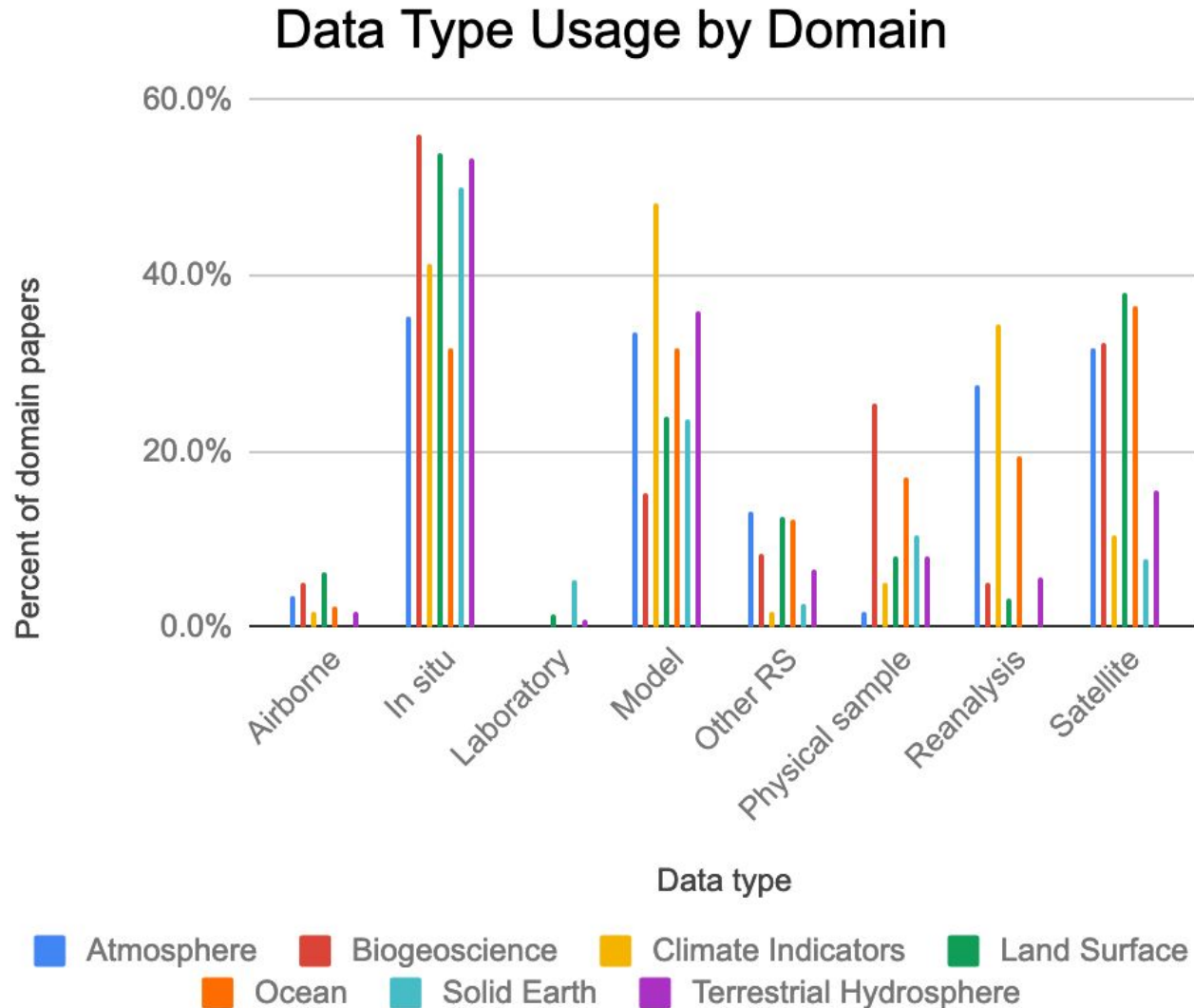
Type of Training Data

Type of Data Used in Earth Science Supervised Machine Learning Papers



Closely related to the number of training data is the data type. Across Earth science, **in situ data** (e.g., meteorological station data, streamflow gauges, buoys, seismic waveforms, etc.) **are most commonly used to train supervised ML algorithms**. Although large archives of **satellite data** are available, these data are used in ML applications **less often than in situ data and model output**. Reasons for this may include the lack of spatiotemporal continuity in low-Earth orbit satellite data and the amount of necessary preprocessing. The least-used data types in supervised ML analysis are airborne and laboratory data, suggesting that those **airborne datasets** large enough for ML purposes **may be under-utilized**.

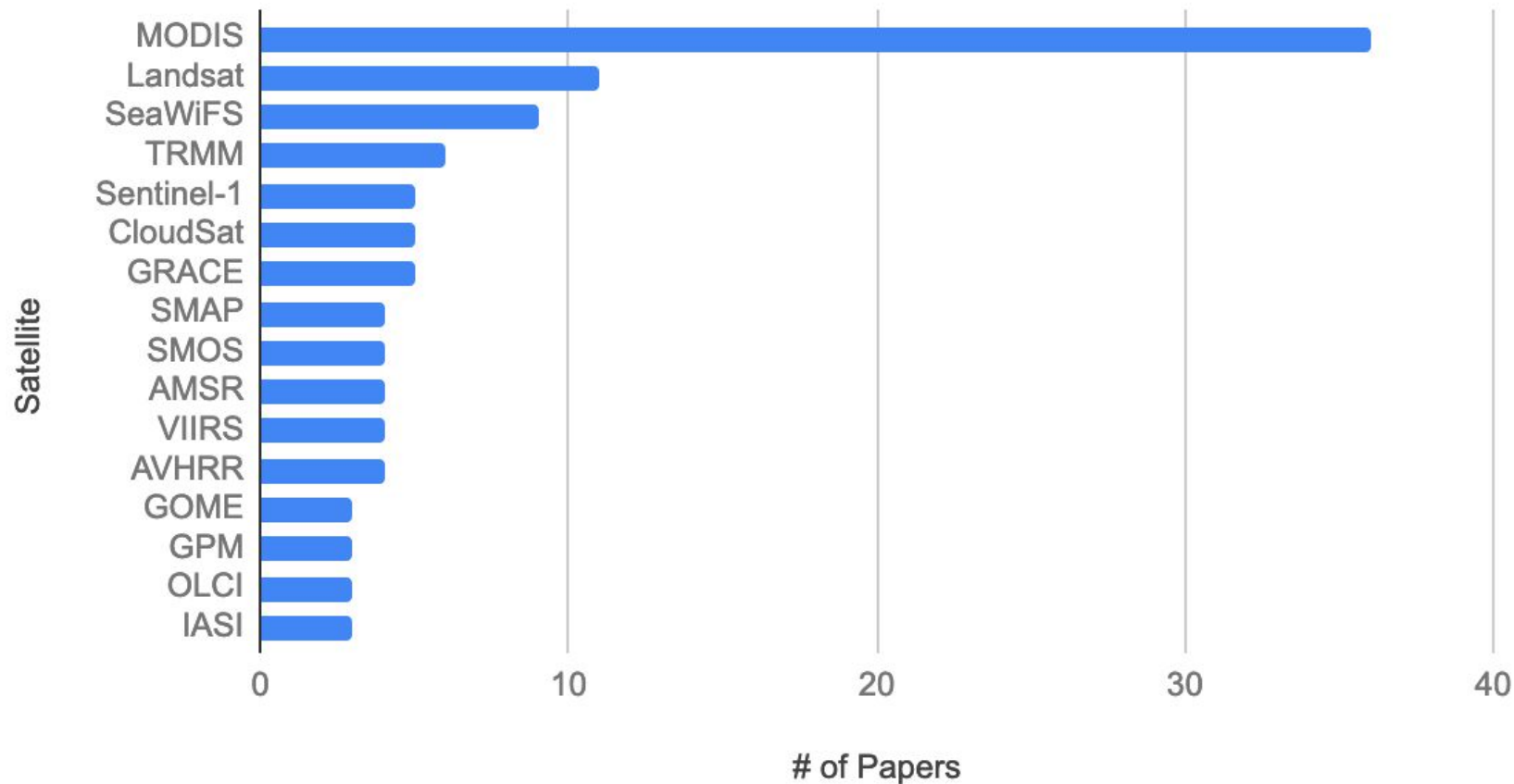
Type of Training Data by Domain



For most sub-domains, **the majority of ML algorithms are trained using in situ data**. Climate and hydrology studies depend heavily on **model output**, while atmospheric studies frequently use **model output, satellite, and reanalysis data** for training. Ocean studies use satellite, model, and in situ data approximately equally. The use of data derived from physical samples is more frequent in biogeoscience research than any other sub-domain, while solid Earth studies are the most likely of any sub-domain to use laboratory data.

Satellites Used for Training

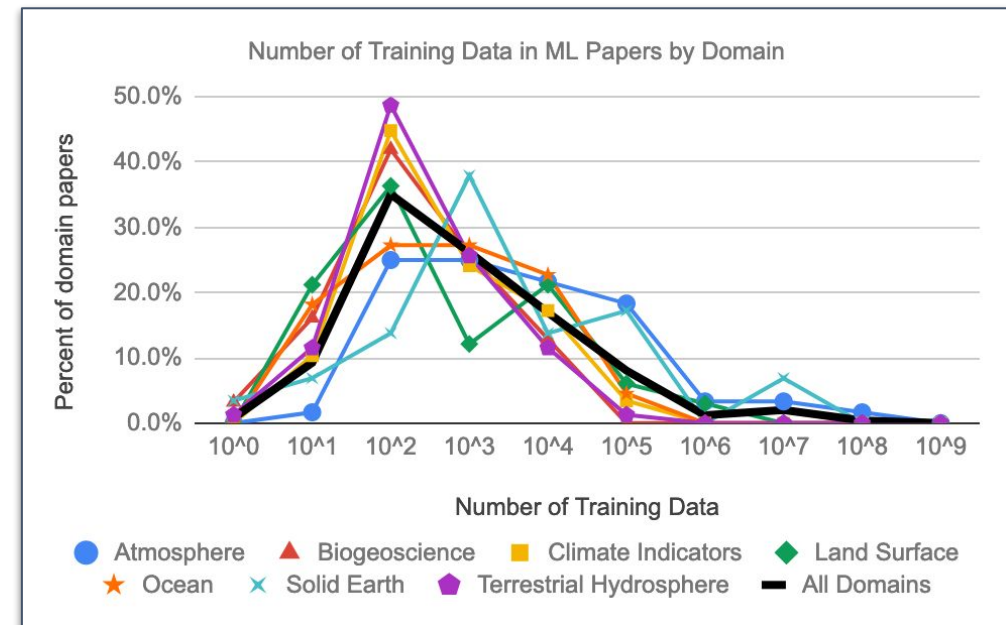
Satellites Used in Earth Science Supervised Machine Learning Papers



In studies using satellite data to train supervised machine learning algorithms, **MODIS was most frequently used** (in ~35% of such studies), followed by **Landsat** (~11%) along with other long-duration missions such as SeaWiFS, TRMM, CloudSat, and GRACE.

Conclusions

- A 10-year literature review shows that **machine learning** use in Earth science research is **growing rapidly**, especially in the biogeoscience and land surface communities
- A 2-year deep dive reveals a **lack of training data** in Earth science studies: machine learning algorithms are typically trained using only hundreds of labeled samples
- There is a need for **open sharing** of available training data and the **focused development of larger training datasets** to support the application of supervised machine learning and deep learning methods
- **Certain types of data** in the archives are **under utilized**, including satellite and airborne data



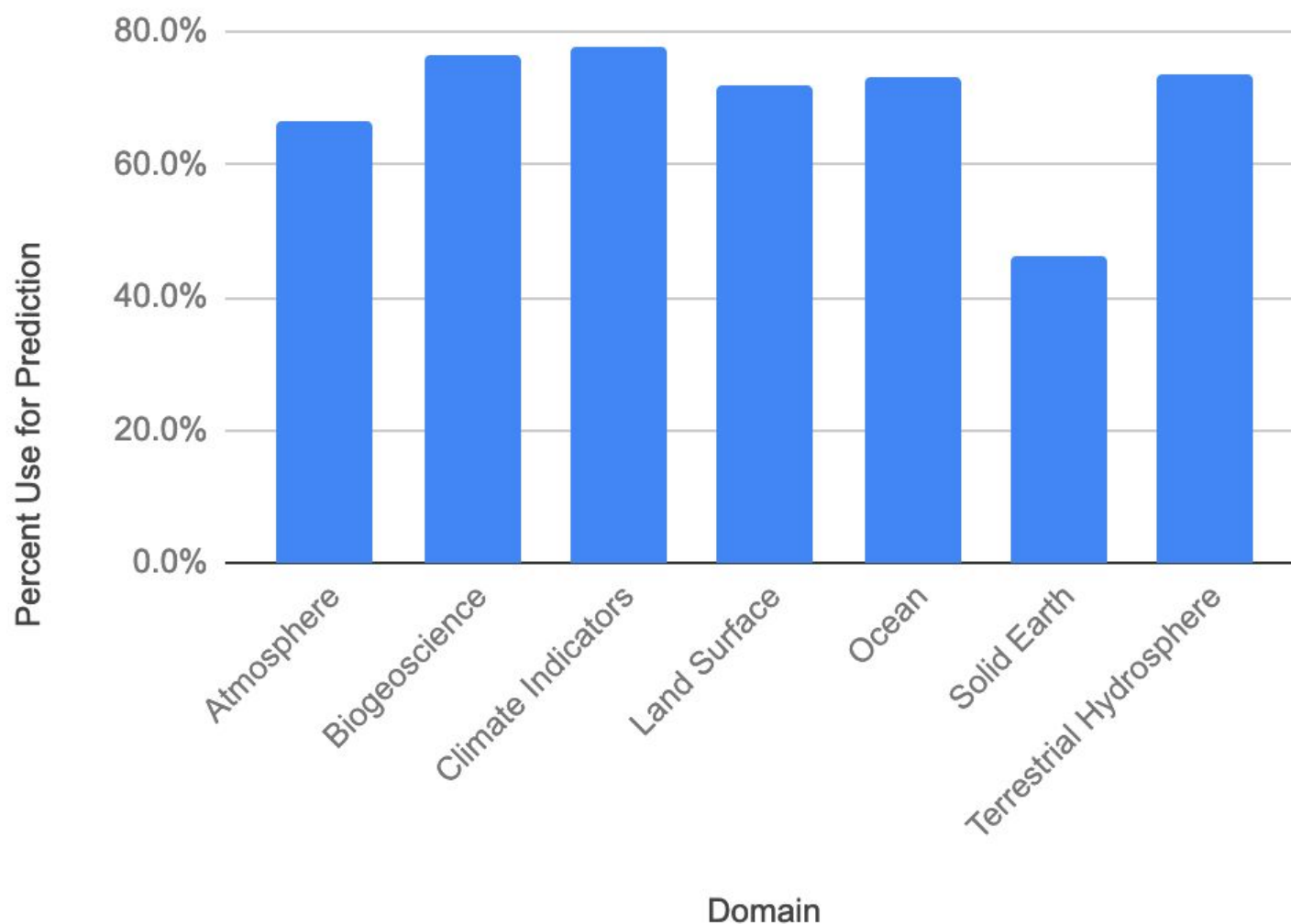
Questions?

Please contact:

Katrina Virts

katrina.virts@uah.edu

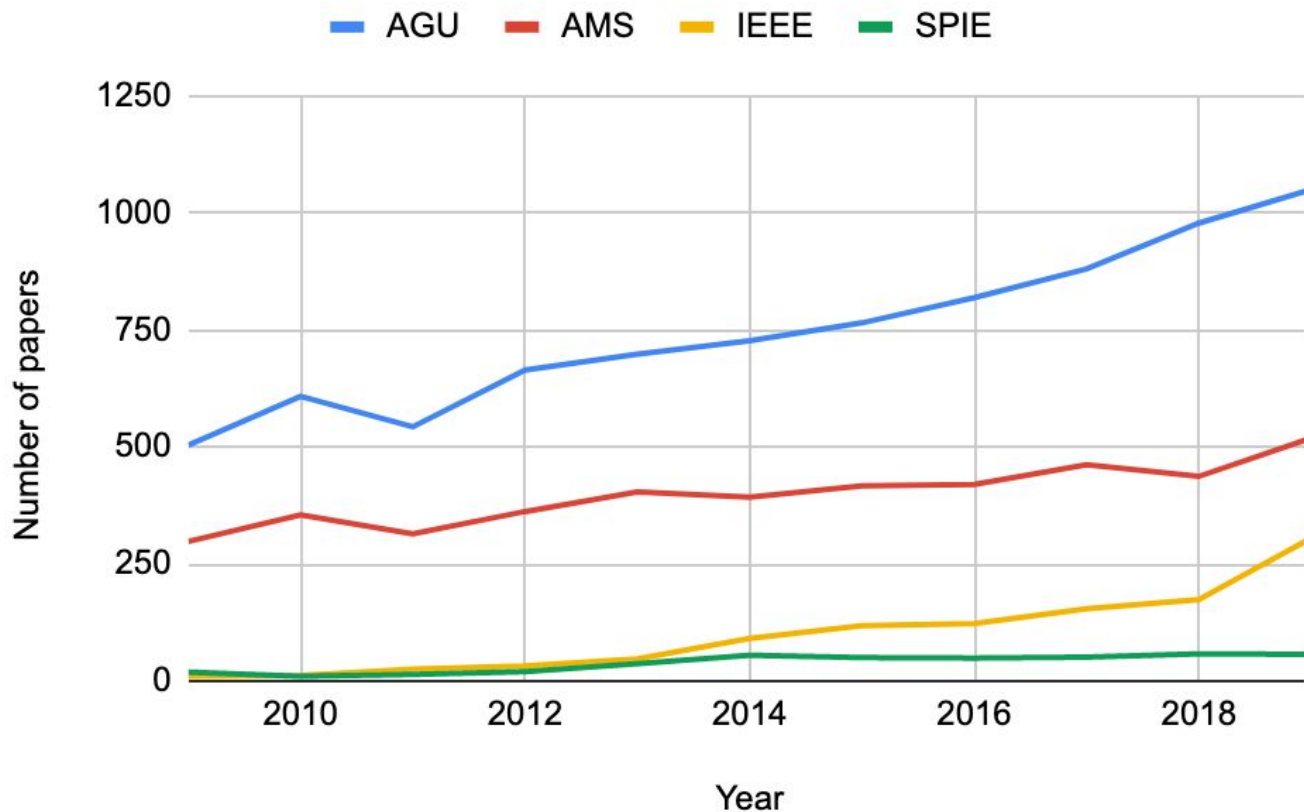
Two-Year Deep Dive



Across Earth Science, **ML techniques are predominantly used for prediction** (~70%). In solid Earth studies, however, ML techniques are applied more frequently for classification studies, which frequently involve earthquake early detection methods.

Trend Analysis

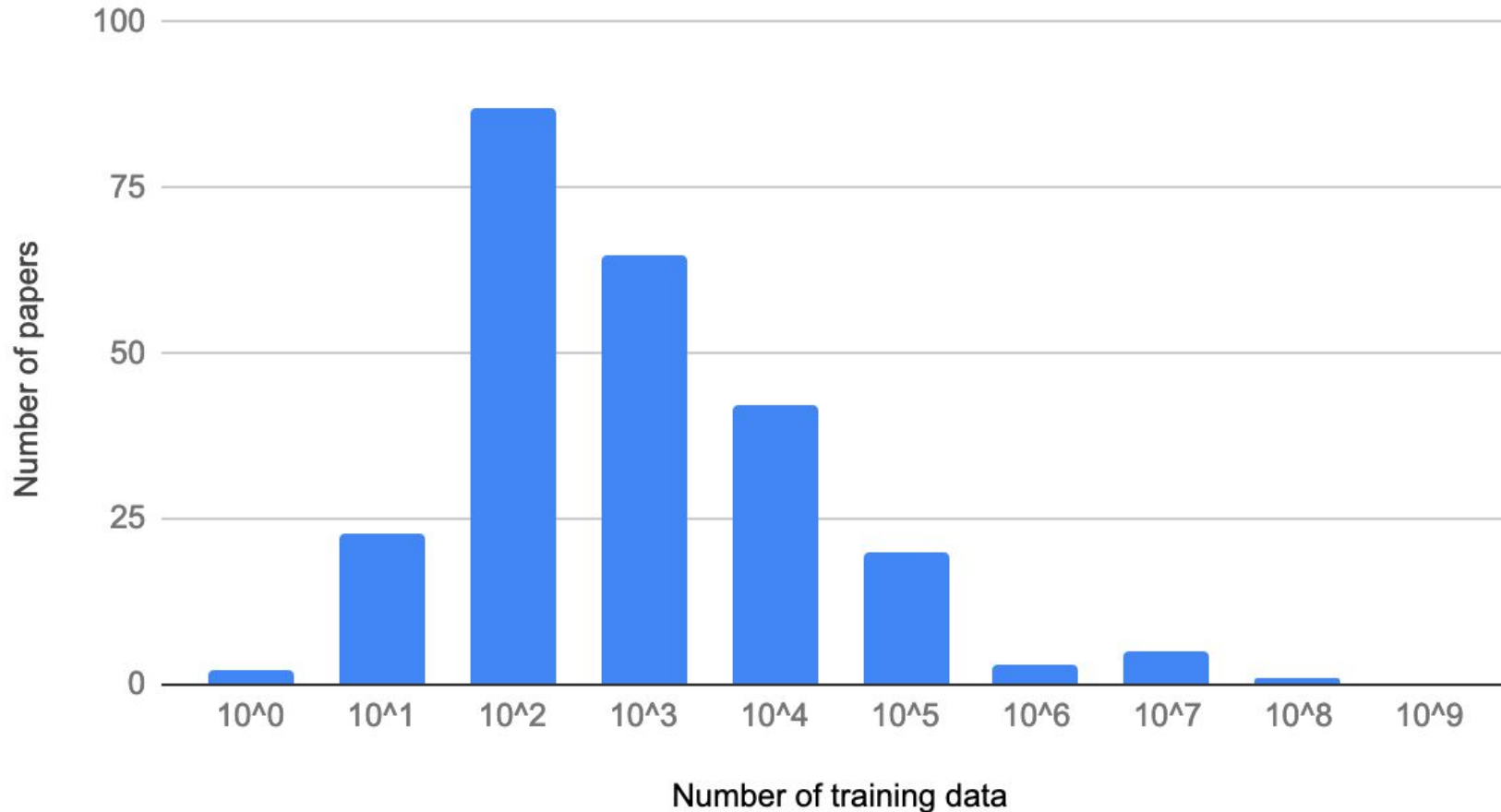
2009-2019 Trend in Earth Science Papers using Supervised Machine Learning Techniques



The use of supervised machine learning techniques in Earth science research has **increased significantly in the last decade**. The number of atmospheric science papers (i.e., from AMS journals) using ML approaches has increased by over 60%. Across all of Earth science even larger changes have occurred, including a >90% increase in AGU papers and a >10-fold increase in IEEE papers using ML.

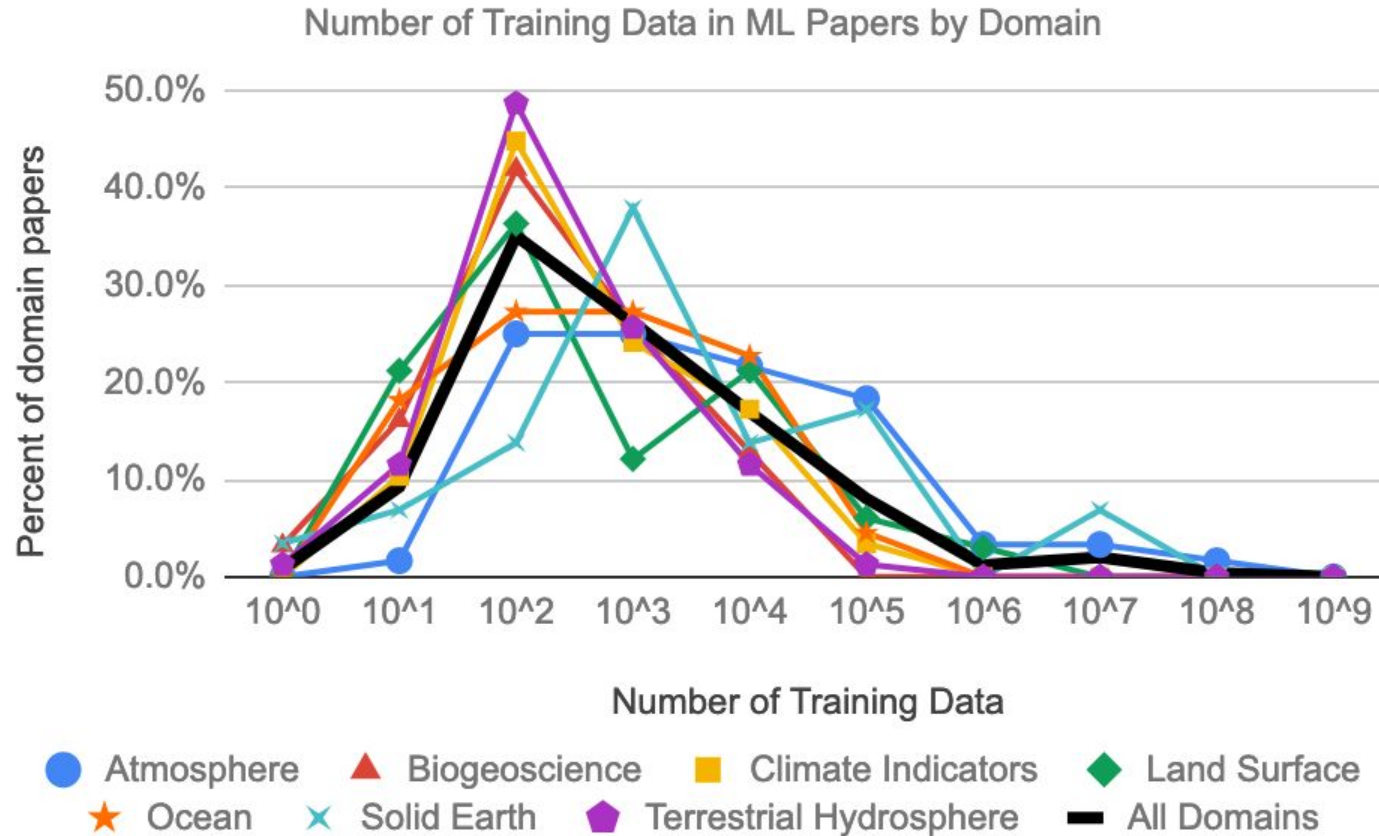
Two-Year Deep Dive

Number of Training Data in ML Earth Science Papers



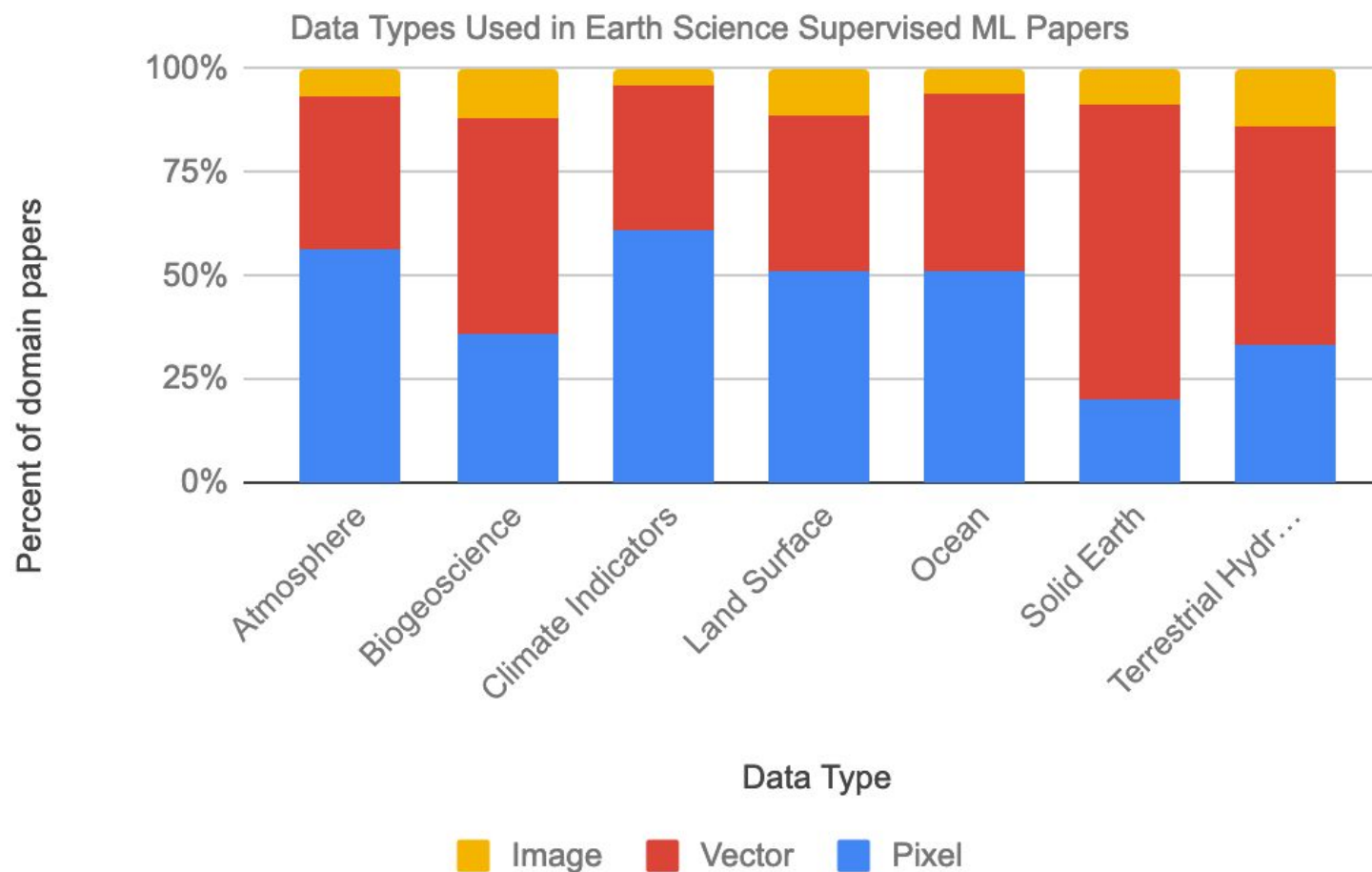
The availability of labeled training data in Earth science is reflected in the number of training data used in supervised analysis. In addition to the 10-year analysis above, we have **manually extracted from AGU papers from the years 2018-2019** other relevant information including the **ML algorithm** that was applied, the **number of labeled training data**, and **the data type** (model output, satellite, in situ, reanalysis, etc.). In the papers we surveyed, most **ML algorithms were trained using hundreds of labeled samples**. However, for some applications using model output or large, established datasets, the number of training data ranged several orders of magnitude greater. More than **80% of studies used automatically labeled training data** rather than manually labeled.

Two-Year Deep Dive



The lack of training data is particularly **acute in biogeoscience and terrestrial hydrosphere (hydrology) studies**. In contrast, training data sizes in atmospheric science and solid Earth studies are more than an order of magnitude larger.

Data Type



Vector and pixel format training data are used about equally in Earth Science supervised ML studies (48% and 43%, respectively). Among sub-domains:

- **Vector** data used most frequently in **solid Earth** studies (71%). E.g., seismic waveforms.
- **Pixel** data used most frequently in **climate indicators** studies (61%). E.g., climate model simulations.
- **Image** data used most frequently in **terrestrial hydrosphere** studies (14%). E.g., satellite-based water fraction imagery.

Summary slide (to be shown during discussion)

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

20th Conference on Artificial Intelligence for Environmental Science

1.5

Katrina Virts, University of Alabama in Huntsville

UAH - Ashlyn Shirey, George Priftis, Muthukumaran Ramasubramanian, Ankur Kumar,

Ashish Acharya; NASA/MSFC - Rahul Ramachandran

Monday, 11 January 2021, 9:50-9:55 AM

- We have conducted a **10-year literature review** of Earth science articles in AGU, AMS, IEEE, and SPIE journals and **deep dive analysis of ~2,000 AGU papers** from 2018-2019 that use supervised machine learning techniques
- **Machine learning** use in Earth science research has **increased significantly in the last decade**, especially in the biogeoscience and land surface communities
- There is a **lack of training data** in Earth science studies: supervised machine learning algorithms are typically trained using only **hundreds of labeled samples**
- There is a need for **open sharing** of available training data and the **focused development of larger training datasets** to support the application of supervised machine learning and deep learning methods
- **Certain types of data** in the archives are **under-utilized**, including satellite and airborne data

