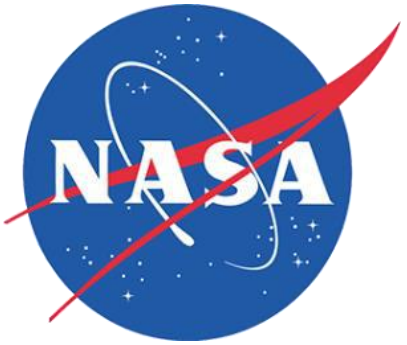
Advancing Application of Machine Learning Tools for NASA's Earth Observation Data

Jan. 21-23, 2020 | Washington, D.C.
Workshop Report



Workshop Committee

- Hamed Alemohammad, Radiant Earth Foundation
- Manil Maskey, NASA/ESDS
- Lyndon Estes, Clark University
- Pierre Gentine, Columbia University
- Dalton Lunga, Oak Ridge National Laboratory
- Zhuang-Fang (Nana) Yi, Development Seed

Sponsored by grant # 19-TWSC19-0010 from NASA Earth Science Data Systems

Table of Contents

Executive Summary

Radiant Earth Foundation, with sponsorship from NASA's Earth Sciences Data Systems (ESDS) Program, hosted the "Advancing Application of Machine Learning Tools for NASA's Earth Observation Data" workshop Jan. 21-23, 2020, in Washington, D.C. During the workshop, experts from academia, industry, government, and nonprofit organizations discussed opportunities for advancing Earth sciences by way of using machine learning techniques to analyze NASA Earth observation data.

Workshop participants were separated into three working groups, each of which generated findings and recommendations according to a theme of the group:

- Working Group 1: Training Data Generation and Accounting for Errors/Uncertainties

  Training data is the main component of supervised machine learning techniques and is increasingly becoming the main bottleneck to advance applications of machine learning techniques in Earth science. Several concerted efforts have started in recent years to catalog and publish benchmark datasets to support new model development and running data science challenges. However, more investment must be made to increase diversity and representativeness of open training datasets across all science disciplines.

  While development and sharing of benchmark training datasets is essential, research and investment in techniques that require less training data (such as semi-supervised learning and active learning) is required. Use of synthetic and simulation datasets to build machine learning models that mimic the physics of Earth system was recommended by the group. Finally, the group also recommended several best practices to account for errors and uncertainties in training datasets as it pertains to Earth science problems.

- Working Group 2: Modeling Approaches and Best Practices for Building the Best and Computationally Optimum Model

  Machine learning models are built on data, and by design they do not incorporate any physical law (such as mass and energy balance) and do not extrapolate well beyond the range of the training data. Different techniques such as regularization and embedding the physical constraints into the model architecture have been successfully implemented to address this issue. However, more research and investment in building domain-aware models and using techniques such as reinforcement learning is a high priority.

  Earth science data include a wealth of ground observations that do not fit into the common gridded data frameworks (i.e. images) and satellites and models provide

multi-band and multi-modal raster images that are not necessarily three bands which is the common data type in computer vision problems. Therefore, new model architectures that can benefit from the characteristics of Earth science data need to be developed and adopted.

Similar to training data, sharing baseline pre-trained machine learning models is essential to ensure research results are reproducible and to facilitate development of new models and applications.

- Working Group 3: Best Practices for Sharing and Publishing Machine Learning Applications (Model, Training Data, and Results)

  Training data and models should be well documented and shared following the FAIR (findability, accessibility, interoperability, and reusability) data management principles. It is expected that journals enforce more strict policies for sharing training data and models, and funding agencies provide support for the required efforts through grants data management plans.

  Training data and model catalogs should have sufficient metadata in a standard format. The SpatioTemporal Asset Catalog (STAC) specification provides a framework for cataloging various geospatial data types but it has shortcomings for storing non-raster data and it needs further investment. A similar specification to STAC needs to be adopted by the community for cataloging and publishing machine learning models.

  A cultural shift is needed to incentivize and reward researchers, developers and practitioners to share their scripts used to generate training data and build models. There should also be more collaborative efforts, supported by funding agencies, to develop and maintain open-source scientific software packages that facilitate research studies.

Acknowledgement

Main Sponsor

- Kevin Murphy, Earth Science Data Systems Program, NASA/HQ
- Manil Maskey, Earth Science Data Systems Program, NASA/HQ

Co-Sponsor

- Lyndon Estes, Clark University, with support from Omidyar Network and PlaceFund
- IEEE Geoscience and Remote Sensing Society (Technical Sponsor)

Technical Advisor

- Rahul Ramachandran, Project Manager, Senior Research Scientist, NASA Marshall Space Flight Center

1. Introduction

During the last decade, machine learning (ML) techniques have improved significantly due to an abundance of data and advancements in high-performance computing. ML-based applications are now being deployed across diverse markets to provide new, faster, and/or more accurate solutions, or to augment human intelligence and increase human efficiency. Increasingly, the science community also is using these techniques to better harness the ever-increasing volume of Earth observation (EO) data for geospatial analysis—i.e., modeling the Earth system and its changing environment.

In order to increase the adoption of ML techniques for geospatial analysis by researchers, practitioners and application developers, several challenges must be addressed. A key component of many ML techniques, for example, is training data (TD), which are used to develop and train models. Such models "learn" patterns in the TD by identifying relationships between input data and observed outputs, allowing them to predict future scenarios. Therefore, it is essential to have comprehensive and diverse TD for building models that can be applied to the broad spectrum of possible input and output scenarios. This is essential for geospatial analysis, as satellites capture EO data with global coverage that can be used to address problems spanning numerous and diverse geographies and time frames.

NASA's Earth Science Division is implementing a cloud-based solution to address the challenges posed by future high-data-volume missions. The collocation of large data repositories and cloud computing democratizes science by allowing anyone to pursue new discoveries without investing in substantial hardware and with only minimal need for large data transfers. Furthermore, cloud computing is expected to make it easier for researchers to exploit new ML techniques that utilize data to solve challenging problems in Earth science. The goal of this new cloud-based approach is to build ML-ready datasets for rapid prototyping and to create common frameworks that enable decentralized development and sharing. This requires shifting from the classical paradigm of data curation and model development to a cloud-centric and data-driven approach. This workshop sought to address the scientific and technological challenges underlying this paradigm shift by presenting successful use case implementations, identifying existing gaps, and developing best practices that will help the community benefit from advancements in ML techniques.

NASA's Earth Science Data Systems (ESDS) Program selected Radiant Earth Foundation—a nonprofit organization whose mission is expanding the application of Earth observations to address global development and environmental challenges—to host this international expert workshop and compile a report that's based on it.

1.1 Definitions

1.1.1 Machine Learning

Machine learning (ML) is the science of getting computers (specifically, algorithms or models) to learn from data and act accordingly in future situations. Generally, there are three types of learning: supervised, unsupervised, and reinforcement. In supervised learning, data contains inputs and outputs (i.e., labels), and models try to learn patterns between them. Regression and classification are the two main types of supervised learning. In unsupervised learning, data doesn't have any labels; models therefore try to find and group patterns based on similarities. Clustering is the most common unsupervised learning problem. Finally is reinforcement learning, wherein the model continuously learns from data based on a reward system. In contrast to supervised learning, where the model can learn patterns just by analyzing the data, models in reinforcement learning are rewarded based on the quality of their predictions and continue to improve until a reward threshold is reached.

1.1.2 Analysis Ready Data

With rapid expansion of EO data, more datasets are becoming available with varying resolutions, processing levels, accuracy, and formats, among other properties. Against that backdrop of data diversity exists the concept of Analysis Ready Data (ARD), which seeks to reduce the preprocessing requirements for users. Specifically, the term "ARD" refers to products that are stored in ARD format or a reproducible pipeline to generate data when it's needed. The processing steps that go into an ARD product depend on the source data and applications for which the data is being prepared. These can include atmospheric correction, masking unusable data (e.g., clouds in visible band imagery), harmonization of different instruments, reprojection, and regridding, among others.

Currently, the U.S. Geological Survey (USGS) is developing ARD for its Landsat mission [1]–[3]. The Committee on Earth Observation Satellites (CEOS) likewise has worked with the community to develop a framework for "ARD for Land" products.[1] Other efforts and partnerships among commercial providers also are progressing to define community standards for ARD.[2] This is an emerging topic that eventually will enable many large-scale ML applications.

2. Workshop Overview

The two-and-a-half-day workshop hosted a diverse group of researchers, practitioners, developers, and stakeholders from academia, industry, government, and nonprofit

[1] http://ceos.org/ard/

[2] https://medium.com/radiant-earth-insights/the-first-satellite-data-interoperability-workshop-is-happening-next-week-fae9539f81f9

organizations. The schedule included two half-day oral presentations by participants, and three half-day working group discussions.

The workshop scientific committee set up three working groups:

- Working Group 1: Training Data Generation and Accounting for Errors/Uncertainties
- Working Group 2: Modeling Approaches and Best Practices for Building the Best and Computationally Optimum Model
- Working Group 3: Best Practices for Sharing and Publishing ML Applications (Model, Training Data, and Results)

To better organize the discussions, participants were assigned to a working group in advance of the workshop and received a set of questions to start the conversation. A member of the scientific committee moderated each working group while rapporteurs captured all the discussions. During three planned report-out sessions, moderators presented their working group's findings to all workshop participants and received feedback along with potential topics to discuss further in their respective groups.

The workshop schedule and list of presenters are available on the workshop website[3] and recorded presentations on YouTube.[4] The full list of workshop attendees also is provided in Appendix B: Workshop Participants.

3. State of Research and Gaps in Training Data and Models for Earth Science ML

Results from a recent study conducted by the University of Alabama and NASA's Marshall Space Flight Center shows significant trends in using ML across different Earth science disciplines between 2008 and 2018, including a 40% increase in atmospheric science papers using supervised ML published by the American Meteorological Society (AMS) in its journals, a 90% increase in similar papers published by the American Geophysical Union (AGU) in its journals, and a 10-fold increase in geoscience papers published by the Institute of Electrical and Electronics Engineers (IEEE). Data from *in situ* instruments, models, and satellites are the most used for building these ML models. Other data that are used include reanalysis, physical samples, and airborne data. There is also diversity across sub-domains in the dominant source of data for ML applications [4]. The following summarizes the latest advancements and gaps in ML applications using EO based on participants' presentations and input.

[3] https://www.radiant.earth/events/nasa-ml-2020/

[4] http://bit.ly/ML4EO20

3.1 State of Research

3.1.1 Training Datasets

Several organizational and academic efforts are underway to publish and share open source benchmark TD for EO in support of ML applications, including SpaceNet,[5] Radiant MLHub,[6] BigEarthNet [5], Dynamic World, and WeatherBench.[7]

Although size of TD is important for its impact on models' performance, another key factor is its representativeness with respect to the distribution of phenomenon of interest. Moreover, different model architectures cannot be absolutely ranked for their performance because their accuracy is highly dependent on the number of samples[8].

Using simulation data as TD is key to assess whether ML models can learn the physics of a geoscience phenomenon. Several examples were presented across science disciplines, including ocean science, climate modeling, and soil moisture retrieval [6]–[9].

3.1.2 Algorithms

Active learning techniques that involve humans in the loop provide a suitable framework for building accurate ML models by directing label generation where the model has low accuracy [10]. This technique can also be used to fine-tune existing models for new spatial or temporal domains [11]–[12].

Manifold embedding to partition massive amounts of data into homogeneous distributions and then fitting simple models helps optimize resource utilization when working with massive geospatial data [13].

To take advantage of multi-resolution bands in satellite imagery or model outputs, a customized Convolutional Neural Network (CNN) can be built that receives multi-resolution data as input [14].

Many successful ML applications are developed and in some cases commercialized to provide crop type classification using satellite data in-season and post-harvest [15]–[19].

3.1.3 Tools and Analytic Frameworks

Analytic frameworks (cloud-based or local) have progressed significantly in recent years and continue to add new functionalities and features to enable development of scalable models.

---

[5] https://spacenet.ai/
[6] https://mlhub.earth/
[7] https://github.com/pangeo-data/WeatherBench
[8] https://medium.com/the-downling/robustness-of-limited-training-data-part-3-9df24c58c2

Pangeo is an open-source architecture that is built on top of other open source tools such as Jupyter, xarray, Dask-Kubernetes, and Binder. It provides ARD with parallel computing and an interactive interface. Future versions will have enhanced integration with ML packages [20].

Open Data Cube (ODC) is an open source data management and analysis platform that provides a flexible framework for hosting multi-spectral data as ARD and accessing it for analysis at scale. Similar to Pangeo, it has a Jupyter interface.

Google Earth Engine is widely used in geoscience research and recently released an integration with TensorFlow and Google Colab for deploying ML models.

RasterVision,[9] Solaris,[10] and eo-learn[11] are examples of open source Python packages for conducting end-to-end ML on satellite imagery.

There are several open source labeling tools to help the community generate high-quality labels from satellite imagery, including: Image Labeler,[12] Geo-Wiki,[13] Picture Pile, FotoQuest Go, LACO-Wiki Validation Platform, and Groundwork[14].

Several competitions have been held in recent years to build and open source ML models for various Earth science applications (clouds organization[15], understanding the Amazon[16], tree detection/classification[17], crop type classification[18], roads and buildings detection[19], and sea lion counts[20]). Because these are very effective for showcasing the state of the art for the corresponding benchmark data, more efforts in this regard are required.

3.2 Gaps

3.2.1 Training Datasets

Lack of open benchmark training datasets across all science disciplines is the main bottleneck in advancing ML application in Earth science and it's also causing

[9] https://rastervision.io/
[10] https://github.com/CosmiQ/solaris
[11] https://github.com/sentinel-hub/eo-learn
[12] https://labeler.nasa-impact.net/
[13] https://www.geo-wiki.org/
[14] https://groundwork.azavea.com/
[15] https://www.kaggle.com/c/understanding_cloud_organization
[16] https://www.kaggle.com/c/planet-understanding-the-amazon-from-space
[17] https://www.ecodse.org/
[18] https://zindi.africa/competitions/iclr-workshop-challenge-2-radiant-earth-computer-vision-for-crop-recognition
[19] https://spacenet.ai/challenges/
[20] https://www.kaggle.com/c/noaa-fisheries-steller-sea-lion-population-count

reproducibility issues and making it difficult to compare results across studies. Moreover, coordinated efforts to collect and publish ground reference data for applications that require those data as labels (agricultural and oceanographic applications, in particular) are needed.

3.2.2 Algorithms

There is limited availability of baseline models for Earth science data that allow for integration of multi-resolution, multi-modal (e.g., EO, SAR, model), multi-scale, and high-dimensional data.

Geoscience models must generalize across space and time; however, for supervised learning one needs large training datasets to build generalizable models. Potential innovative solutions in which to invest include techniques that:

- Require less data, such as transfer learning, representation learning, semi-supervised learning, and unsupervised learning [21]–[25]; and/or
- Use alternative data sources, such as weak and crowdsourced labels [26].

Because ML models generally do not preserve physical laws (e.g., mass and energy balance), innovative ways to incorporate physics in ML models are needed [8]. Moreover, there are complex spatio-temporal relationships—including non-trivial and lagged long-distance relationships (i.e., teleconnections) between geophysical variables—that must be adequately modeled in ML architectures.

The majority of ML models do not generalize to data outside of their training dataset; therefore, they cannot be used for extrapolating (e.g., forecasting extreme events or climate change). This is a major barrier for their application in Earth sciences and requires further research—for example, using reinforcement learning.

Unstructured data are key in some disciplines, such as oceanic and atmospheric sciences. However, the majority of ML models are built for structured (e.g., gridded) data. A potential solution that requires more investment is using graph neural networks [27]. Meanwhile, sparsity of these data in space and time makes it more challenging to use them for training and validation. It's necessary to quantify the uncertainty that is caused by their heterogeneous sampling, and to propagate that into the model predictions.

Improved characterization of models' confidence and accuracy is essential to their wide adoption. Bayesian/probabilistic inference should be integrated into models because such inference allows for explicit representation and propagation of uncertainties.

Lack of interpretability in geoscience models reduces the trust in model predictions. More research in developing interpretability metrics is essential.

3.2.3 Tools and Analytic Frameworks

There are limited open source software and ML frameworks for deploying Earth science ML models, and more investment is required to expand existing resources and build new cost-effective, scalable, and maintainable frameworks.

Data preparation and preprocessing is still time-consuming while new cloud-based ARD are gradually filling this gap.

4. Working Group Discussions and Recommendations

4.1 Working Group 1: Training Data Generation and Accounting for Errors/ Uncertainties

TD is the building block of supervised ML models which constitutes majority of ML applications in Earth science. ML models iteratively learn from TD, and any uncertainty or error in TD will be propagated to the model and its outputs. Therefore, it's essential to ensure quality of training datasets by quantifying and documenting their uncertainties and characterizing their representativeness. To that end, Working Group 1 was asked to review existing practices and provide recommendations for future applications. In particular, the group started by reviewing the results of a January 2019 workshop at Clark University, the focus of which was errors in TD and approaches to account for them [28].

4.1.1 Findings

Errors in TD can be design- or collection-related. In design-related cases, the data lacks spatial or temporal representativeness due to several factors, including out-of-date data, rare classes not captured, labels collected by users who are biased, or *in situ* point data that is not scalable for labeling gridded data (e.g., pixels in the image or model output). In collection-related cases, there can be interpretation errors due to vagueness of class definitions for categorial data or observation errors for continuous data.

"How much data is enough?" and "Is the data spatially and temporally representative?" are common questions asked by researchers who are interested in developing or using an ML model. However, there is no universal answer to these questions. Many factors can impact how much TD is needed to build an ML model, including: whether it's a classification (e.g., segmentation, object detection) or regression problem, architecture of the model being used, preferred accuracy level, generalizability of the model, and sampling technique used to generate the TD.

In order to best utilize annotations, *in situ* measurements, or other outputs as TD, it is necessary to capture basic metadata and report it in the TD catalog. This may encompass dates of measurement or imagery used for annotation, spatial coordinates, how

measurements were taken, how labels were generated, and uncertainty of the measurement device or annotator.

Class imbalance (i.e., one class or a range of values is under-represented or over-represented in TD) is a common problem. In categorical data, certain class(es) may not be well represented, depending on the geographical domain from which the data is generated. For continuous data, extreme events are usually rare in TD, which makes it hard to train the model for such scenarios and achieve reasonable accuracy.

Advancing applications of ML techniques on EO is contingent on "benchmark training data" that the community can use to assess models' performance. These benchmarks are application-specific and would require due diligence to generate and share with the community.

NASA provides a wealth of data across all Earth science disciplines. These data provide a unique opportunity to build new ML models—potentially by fusing multiple data sources—to detect new patterns and improve our understanding of the Earth system.

4.1.2 Recommendations

Generating TD and accessing TD from previous studies is a major bottleneck in building new ML models and benchmarking existing ones in Earth science. The group compiled the following recommendations for new areas of research and investment to address TD issues.

TD errors: The group recommended the guidelines and use cases presented in [28] as best practices to learn more about ways to account for errors in TD.

TD sparsity: More investment is needed in order to advance techniques that require less training data (e.g., active learning and semi-supervised learning). Innovative solutions that exploit alternative datasets, transfer learning, and/or fusion of multiple datasets (either by taking the most consistent one over time or weighting them by quality [29]) should be supported across all disciplines—with the caveat that some models are not transferrable and researchers must explicitly define the applicable spatial and temporal domain of the TD and model.

Size of TD: Researchers should consider the physics of the problem and the data range/distribution before deciding on their TD size. Comparing model performance for different TD sizes and for different model architectures (from simple to complex) should be routine in order to justify the size of TD. While this analysis will be informative for the problem of interest, it may not be easily generalized to other problems or to the same problem in a geographically diverse region [30].

Class Imbalance: To generate new TD, a representative sampling scheme should be developed in space and time (for annotation or otherwise generating labels). For existing

TD, mitigation strategies such as penalizing certain classes in the cost function or under-/over-sampling during training can be implemented. Moreover, architectures such as random forests can handle class imbalance better than neural networks.

Benchmark TD: Significant investment in generating and sharing high-quality open-access benchmark data is requested from NASA and other federal funding agencies to advance ML applications in geosciences. These data should have sufficient diversity and sample size with respect to space, time, and scale to allow exploring different model architectures and their performance. Ideally, generating the TD should be grounded in a peer-review process that includes community consensus, adaptation, documentation, domain specificity, transparency, and validation of quality and accuracy in order to build trust in the dataset.

4.2 Working Group 2: Modeling Approaches and Best Practices for Building the Best and Computationally Optimum Model

Unlike common ML problems that are defined on deterministic systems, the Earth system is stochastic; therefore, scientists must incorporate this characteristic in modeling frameworks and develop uncertainty-aware models. In addition, EO and many model-based datasets have more than three bands of observation, which is the typical case in computer vision problems. These data are usually collected over time at fixed temporal frequencies, which makes them unique in nature. These properties require domain-specific modeling practices, which was the focus of Working Group 2.

4.2.1 Findings

A key challenge pertaining to domain-aware models for ML in EO applications, the group found, is effectively incorporating prior knowledge, such as physical principles, constraints, and computational simulations. Among the biggest challenges with building domain-aware models for supervised learning with hard and soft constraints on TD spaces are: subjectivity in labeling data, lack of common standards in tools for TD collection, and tracking temporal changes (labels are snapshots in time). Incorporating regularization techniques—such as data augmentation, label smoothing, incremental partial training, time-dependent labels, and active learning methods that include a human in the loop—can help create generalizable models and dramatically reduce data requirements.

Building domain-aware models using unsupervised learning suffers from the development of features that are representative, informative, interpretable, and generalizable; the incorporation of prior knowledge, such as physical constraints and computational simulations, into features; and the evaluation of features' effectiveness. The group recognized that dimensionality reduction methods, generative models, and manifold and metric learning are among the state-of-the-art techniques available to

address these challenges and improve unsupervised techniques' interpretability, robustness, and generalization.

Challenges in finding effective strategies for incorporating multi-band data into ML models include interpretation of new bands and lack of baseline remote sensing pre-trained models. Moreover, these pre-trained models are application-specific and must be built by addressing the requirements of those applications. Another challenge is the curse of dimensionality and the Hughes effect. In 1968, Gordon Hughes developed a numerical relationship between measurement complexity and mean accuracy of a classifier [31]. His work shows that increasing measurement complexity increases model accuracy up to a certain point, after which further increasing measurement complexity reduces model accuracy (ref. Fig 3. in [31]). This calls for an assessment of optimal model architectures for a given application and TD.

Geospatial data have special characteristics such as projection, spatial grid, and temporal revisit that are rarely the same across two datasets. In order to best utilize these data and increase their interoperability, a set of community standards should be compiled by implementing a wide range of use cases. Moreover, we must define application-specific reference frameworks (including the TD and performance metric) in order to effectively increase interoperability of heterogenous input data. Such frameworks can then guide development of model architectures to incorporate these data.

Finally, the group reviewed the potential use of ML models and existing TD and modeling frameworks to inform new data collection strategies. According to the group, having a human in the loop to compare human label and model confidence estimates can inform what new bands of information are needed in order to enhance separability in feature space or develop guidelines for better data collection for the specific application (e.g., maybe no cloud data).

4.2.2 Recommendations

ML models are actively changing with new techniques being developed continuously. Therefore, the group provided priorities in three categories to advance ML models in Earth science and increase their adoption for addressing challenging problems.

Domain-aware models: While several successful examples of domain-aware models were presented and reviewed, more research and development is needed around incorporating regularization techniques, implementing techniques for constant learning from the environment (e.g., reinforcement learning), utilizing simulation and synthetic data, and enhancing domain-specific validation metrics. Further research into using Bayesian-based simulation frameworks also is required. These models have the potential to improve interoperability, accuracy, and uncertainty propagation while also simplifying models' TD requirements. Finally, more research is needed into advancing unsupervised

models to automatically learn features by satisfying a given set of constraints along with learning features for processes described by large heterogeneous datasets. These can help achieve increased interpretability and improve models' robustness.

Multi-band/multi-source ML models: This area should be a top priority for researchers and NASA programs in order to accelerate adoption of ML techniques across different science disciplines. In particular, areas that need further support include: developing benchmark pre-trained models using multi-band/multi-source data, establishing a repository for sharing these models to be used as a backbone for new developments, incorporating interpretability metrics into these models to understand the value of each input, publishing review papers to benchmark different model architectures in each domain, and providing guidance for future developments.

Productionizing models: Several steps should to be taken during model development to enable scaling of prototypes for production later on, including: building models for scale from the beginning (in particular, model compression/pruning—e.g., using TensoRT[21]); conducting out-of-sample test runs on baseline models before moving to production, including spatially and temporally varying samples in the training and test sets to better assess performance of the model in production; striking a balance between model versioning and robustness of models; and back testing models to measure and correct for catastrophic forgetting.

4.3 Working Group 3: Best Practices for Sharing and Publishing ML Applications (Model, Training Data, and Results)

To ensure that ML applications and research studies are reproducible, it is necessary to define best practices for documenting, storing, and sharing models, TD, and their results/outputs. Moreover, proper cataloging of TD will allow for development and evaluation/benchmarking of models, as well as for running data science challenges. Working Group 3 was asked to discuss existing practices around this topic and to provide recommendations for the community. The group emphasized the importance of holding TD and ML models to the same documentation and reproducibility standards as other datasets and models while also making necessary adaptations to general data management standards in order to account for the unique characteristics of ML techniques.

4.3.1 Findings

The group agreed that TD and ML models should adhere to FAIR (findability, accessibility, interoperability, and reusability) data management principles,[22] which recommend generating rich metadata interpretable by humans and machines [32]. Moreover, it said,

[21] https://developer.nvidia.com/tensorrt
[22] https://www.go-fair.org/fair-principles/

TD and ML models should be version-controlled to ensure usability, accessibility, and permanency, and should be deposited in an open and trusted repository.

Although many journals currently require authors to share their data, the process and specifications are highly variable. A stricter enforcement by journals to ensure that data is properly shared, accessible, and permanently available is required. Additionally, open data sharing could lead to more co-author and citation credits, which could motivate the community to adhere to best practices for data sharing. The group agreed that different incentives should be developed to foster data sharing by the community according to best practices, and that the community should reframe TD as a deliverable that should be documented and published at the end of a project.

4.3.1.1 Training Datasets

The group underscored that TD documentation should include general spatiotemporal information and TD-specific attributes such as the original data used to generate labels, data processing, and type(s) of applicable ML model/algorithm. The TD's metadata should include these same attributes, but in a machine-readable format to facilitate discovery and query of these data.

The group discussed the SpatioTemporal Asset Catalog (STAC) as an example of a detailed and queryable metadata schema that could be utilized for TD documentation and dissemination.[23] The STAC label extension would be a particularly useful specification for documenting TD. Storing TD catalogs using STAC specification could create a centralized system with a human interface (STAC browser) and machine interface (STAC API) where one could discover imagery, TD, and corresponding metadata.

With increasingly more data becoming available on public clouds, and the paradigm shift in using cloud-based data, it is necessary to store data in formats that facilitate easy access to and usage of TD. NetCDF format,[24] which has been historically used to store many spatiotemporal geoscience datasets, is an option that can be pushed onto the cloud with a converter for cloud optimization as an interim format until Unidata releases an integrated NetCDF and Zarr format.[25] Zarr is a more ML-ready format but is relatively new and will need to go through NASA and other agency standard offices in order to be accepted as an official storage format. The group identified a need for a hierarchal approach to storing mission data with "archival" and "reproducibility" levels. A tiered

[23] https://github.com/radiantearth/stac-spec

[24] https://www.unidata.ucar.edu/software/netcdf/

[25] https://zarr.readthedocs.io/

system of data storage would allow the community to have a long-term record without sacrificing accessibility.

The group noted that there are varying levels of sophistication within ML's user base. As such, it's important to recognize that not everyone accesses data in the same way or at the same point in the pipeline. While a centralized repository of all would be ideal, it is therefore not feasible in the short term. Instead, the group proposed that existing repositories like NASA's CMR add TD metadata and query parameters for TD and ML models to ensure that these data are searchable.

4.3.1.2 Scripts and ML Models

Some TD are not stored as data, but rather are generated on-demand using a script or scripts that train models without saving the intermediate data. This is more common when using outputs of physical models or generating synthetic data for TD. For the sake of reproducibility, the group agreed that such scripts should also be documented and shared.

Sharing ML models themselves could similarly increase reproducibility. Models should be shared in a universal format (e.g., ONNX[26]) or in other formats that can be easily converted to other universal/standard formats. Model metadata should be recorded using the standard terminology of the model's domain. If test cases or benchmark datasets are included with shared models, they should be documented appropriately.

The code used in both TD generation scripts and ML models should be maintained so it does not become obsolete after publication. The group proposed that researchers should be responsible for maintaining code for three years—the typical grant funding cycle. However, an alternative data-driven approach would measure the code's usage and determine how long it should be maintained based on a popularity metric. The community could reduce the need for maintenance if everyone continuously integrated code and extended features from the robust community archive. The adoption of "community code" would require a huge cultural shift and raises questions about how to report one's contribution to integrated code on resumes, etc. A successful example of such community code is the Pangeo[27] project.

4.3.1.3 Data and Knowledge Gaps

The group identified several geographical, scientific domains, and methodological gaps in existing TD catalogs and ML applications. It recognized that applications using labels from satellite image annotation currently are overrepresented in the community due to their similarity with common computer vision problems. However, there are other sources of

[26] https://onnx.ai/

[27] https://pangeo.io/

EO data (e.g., non-optical imagery) and *in situ* or simulation data that can be utilized. Non-traditional data sources such as NOAA storm events,[28] Data.gov, or ancillary data like plane routes are not stored in the same repositories as EO and provide valuable information for creating training data.

4.3.2 Recommendations

Understanding the community effort and cultural shift that is required to encourage researchers, practitioners and developers to document and share their TD and models, the group provided specific recommendations utilizing existing resources standards.

Cataloging TD: TD catalogs should follow FAIR principles for data management. In particular, they should have: permanent identifier and storage, ReadMe/documentation, data dictionary, version, license, citation, and machine-readable metadata. Scripts/codes that generate TD as intermediary products should be preserved by containerizing the code (ideally using lightweight Docker containers) or using binders. TD generation scripts should also have appropriate metadata and documentation about their attributes, particularly for the ML algorithms with which they are used.

The group suggested two options to address STAC's shortcomings for storing non-regularly gridded data: 1) removing the requirement for regular latitude and longitude grids in STAC, and 2) creating a more generalized version of a data catalog specification (of which STAC would be a subset) that could be used with a more diverse set of geoscience.

Research proposals should allocate financial and human resources in their data management plan (DMP) to ensure that cataloging and documentation of code/data is implemented throughout the project.

Cataloging models: ML models should be shared with specific metadata to ensure reproducibility. At minimum they should include:

- Any metadata required by the format that the model object is stored in;
- List of input(s) and output(s) with a harmonized vocabulary (domain-specific), along with the format, structure/shape, unit, and any normalization of the input data;
- Spatial and temporal extent that the model is applicable;
- URL to documentation;
- URL/identifier of the TD used to train the model;
- License;
- Provider; and

[28] https://www.ncdc.noaa.gov/stormevents/

- Accuracy (optional): If model has been tested against a benchmark dataset, it is informative to report the accuracy in the metadata.

The group reviewed the STAC collection specification[29] and recommended including all of the fields for ML model catalogs. The group also called for a coordinated community effort to develop a standard specification for cataloging ML models.

Data Format: The group agreed that TD: 1) must use a well-documented format, and 2) should be stored in a cloud-optimized format.

Incentives: Understanding that data and software development and maintenance is an extensive effort, researchers should be incentivized to focus on them. Incentives may include employer recognition (especially at academic institutions), legal support to ensure intellectual property (IP) and proper use of their work by others, as well as proper citation mechanisms. Data agreements that stipulate open sharing of ML applications could provide legal motivation for data sharing, particularly for companies and NGOs.

Tooling: Development of a geospatial community tool that would run/containerize a code based on a specification file is required. Because geospatial pipelines have unique requirements to the characteristics of the data, developing a tool that can read from a specification file and generate the processing pipeline in a modular format would be of significant value for reproducibility of scientific results.

Existing data: Develop tools to exploit the wealth of *in situ* data in oceanic and atmospheric science applications that are usually stored in a Lagrangian coordinate system and therefore not easily convertible to a Euclidean system of EO and Earth science models. Encourage utilization of Cal/Val data in ML applications by investing in methodologies to make this data more ML-ready or build community tools that would help users access these data along with EO more easily. Finally, increase awareness about the potential of using non-traditional geospatial data sources as TD with EO.

Data Management Plans: ML applications need data, and their applications go beyond the scope of one government agency. Therefore, coordination across government agencies to standardize DMPs and data storage practices would significantly help researchers and grantees.

[29] https://github.com/radiantearth/stac-spec/blob/master/collection-spec/collection-spec.md

5. References

[1] J. L. Dwyer, D. P. Roy, B. Sauer, C. B. Jenkerson, H. K. Zhang, and L. Lymburner, "Analysis ready data: Enabling analysis of the landsat archive," *Remote Sens.*, vol. 10, no. 9, pp. 1–19, 2018.

[2] C. Anderson *et al.*, "The U. S. Geological Survey's Approach to Analysis Ready Data," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 5541–5544.

[3] P. Potapov *et al.*, "Landsat Analysis Ready Data for Global Land Cover and Land Cover Change Mapping," *Remote Sens.*, vol. 12, no. 3, p. 426, Jan. 2020.

[4] K. Virts *et al.*, "A Quantitative Analysis on the Use of Supervised Machine Learning in Earth Science," in *IGARSS 2020- 2020 IEEE International Geoscience and Remote Sensing Symposium*, 2020.

[5] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigearthnet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 5901–5904.

[6] P. Gentine, M. Pritchard, S. Rasp, G. Reinaudi, and G. Yacalis, "Could Machine Learning Break the Convection Parameterization Deadlock?," *Geophys. Res. Lett.*, pp. 1–10, Jun. 2018.

[7] J. Kolassa *et al.*, "Estimating surface soil moisture from SMAP observations using a Neural Network technique," *Remote Sens. Environ.*, vol. 204, pp. 43–59, Jan. 2018.

[8] T. Beucler, S. Rasp, M. Pritchard, and P. Gentine, "Achieving Conservation of Energy in Neural Network Emulators for Climate Modeling," in *ICML 2019 Workshop. Climate Change: How Can AI Help?*, 2019.

[9] A. Sinha, "Temporal Variability in Ocean Mesoscale and Submesoscale Turbulence," Columbia University, 2019.

[10] S. R. Debats, L. D. Estes, D. R. Thompson, and K. K. Caylor, "Integrating active learning and crowdsourcing into large-scale supervised landcover mapping algorithms.," *PeerJ*, vol. 5, 2017.

[11] K. Malkin *et al.*, "Label Super-Resolution Networks," in *ICLR*, 2019, pp. 1–22.

[12] C. Robinson *et al.*, "Large Scale High-Resolution Land Cover Mapping With Multi-Resolution Data," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12718–12727.

[13] D. Lunga, J. Gerrand, L. Yang, C. Layton, and R. Stewart, "Apache Spark Accelerated Deep Learning Inference for Large Scale Satellite Image Analytics," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 271–283, 2020.

[14] G. Sumbul and B. Demir, “A Novel Multi-Attention Driven System for Multi-Label Remote Sensing Image Classification,” in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 5726–5729.

[15] V. Zalles *et al.*, “Near doubling of Brazil’s intensive row crop area since 2000,” *Proc. Natl. Acad. Sci.*, vol. 116, no. 2, pp. 428–435, Jan. 2019.

[16] H. R. Kerner, C. Nakalembe, and I. Becker-Reshef, “Field-level crop type classification with k-nearest neighbors: A baseline for a new Kenya smallholder dataset.,” in *Proceedings of the International Conference on Learning Representations (ICLR) Workshops.*, 2020.

[17] L. Zhong, L. Hu, and H. Zhou, “Deep learning based multi-temporal crop classification,” *Remote Sens. Environ.*, vol. 221, no. December 2018, pp. 430–443, Feb. 2019.

[18] M. K. Gumma *et al.*, “Agricultural cropland extent and areas of South Asia derived using Landsat satellite 30-m time-series big-data using random forest machine learning algorithms on the Google Earth Engine cloud,” *GIScience Remote Sens.*, vol. 00, no. 00, pp. 1–21, Nov. 2019.

[19] A. J. Oliphant *et al.*, “Mapping cropland extent of Southeast and Northeast Asia using multi-year time-series Landsat 30-m data using a random forest classifier on the Google Earth Engine Cloud,” *Int. J. Appl. Earth Obs. Geoinf.*, vol. 81, no. November 2018, pp. 110–124, Sep. 2019.

[20] T. E. Odaka *et al.*, “The Pangeo Ecosystem: Interactive Computing Tools for the Geosciences: Benchmarking on HPC,” 2020, pp. 190–204.

[21] S. Wang, W. Chen, S. M. Xie, G. Azzari, and D. B. Lobell, “Weakly Supervised Deep Learning for Segmentation of Remote Sensing Imagery,” *Remote Sens.*, vol. 12, no. 2, p. 207, Jan. 2020.

[22] S. Wang, G. Azzari, and D. B. Lobell, “Crop type mapping without field-level labels: Random forest transfer and unsupervised clustering techniques,” *Remote Sens. Environ.*, vol. 222, no. November 2018, pp. 303–317, Mar. 2019.

[23] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, and S. Ermon, “Tile2Vec: Unsupervised Representation Learning for Spatially Distributed Data,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 3967–3974.

[24] S. Wang *et al.*, “Mapping Crop Types in India with Crowdsourced Data and Deep Learning,” in *American Geophysical Union, Fall Meeting 2019*, 2019.

[25] D. Lunga, H. L. Yang, A. Reith, J. Weaver, J. Yuan, and B. Bhaduri, “Domain-Adapted Convolutional Networks for Satellite Image Classification: A Large-Scale Interactive Learning Workflow,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 11, no. 3, pp. 962–977, Mar. 2018.

[26] S. Wang *et al.*, “Mapping Crop Types in India with Crowdsourced Data and Deep

Learning," in *American Geophysical Union, Fall Meeting 2019*, 2019.

[27] J. Zhou *et al.*, "Graph Neural Networks: A Review of Methods and Applications," Dec. 2018.

[28] A. Elmes *et al.*, "Accounting for Training Data Error in Machine Learning Applied to Earth Observations," *Remote Sens.*, vol. 12, no. 6, p. 1034, Mar. 2020.

[29] S. H. Alemohammad *et al.*, "Water, Energy, and Carbon with Artificial Neural Networks (WECANN): a statistically based estimate of global surface turbulent fluxes and gross primary productivity using solar-induced fluorescence," *Biogeosciences*, vol. 14, no. 18, pp. 4101–4124, Sep. 2017.

[30] G. Sumbul, R. G. Cinbis, and S. Aksoy, "Fine-Grained Object Recognition and Zero-Shot Learning in Remote Sensing Imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 770–779, Feb. 2018.

[31] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. 14, no. 1, pp. 55–63, Jan. 1968.

[32] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci. Data*, vol. 3, no. 1, p. 160018, Dec. 2016.

Appendix A: Acronym List

Acronym	Description
ACCESS	Advancing Collaborative Connections for Earth System Science
AGU	American Geophysical Union
AMS	American Meteorological Society
ARD	Analysis Ready Data
CEOS	Committee on Earth Observation Satellites
DL	Deep Learning
DMP	Data Management Plan
DOI	Digital Object Identifier
EO	Earth Observations
ESDS	NASA Earth Science Data Systems
FAIR	Findable, Accessible, Interoperable, and Reproducible
HQ	NASA Headquarters
IEEE	The Institute of Electrical and Electronics Engineers
IEEE GRSS	IEEE Geoscience and Remote Sensing Society
IP	Intellectual Property
ML	Machine Learning
NASA	National Aeronautics and Space Administration
NOAA	National Oceanic and Atmospheric Association
ONNX	Open Neural Network Exchange
R&D	Research and Development
SAR	Synthetic Aperture Radar
STAC	SpatioTemporal Asset Catalog
TD	Training Data
USGS	U.S. Geological Survey

Appendix B: Workshop Participants

Name	Organization	Role
Alando Ballantyne	Radiant Earth Foundation	Participant
Alex Leith	Geoscience Australia	Presenter
Andreas Gros	Facebook	Presenter
Anne Miglarese	Radiant Earth Foundation	Presenter
Anu Swatantran	Corteva	Presenter
Benjamin Goldenberg	Planet	Presenter
Brookie Guzder-Williams	World Resources Institute	Presenter
Caleb Robinson	Microsoft AI for Earth, Georgia Institute of Technology	Presenter
Christopher Brown	Google	Presenter
Christopher Lynnes	NASA	Presenter
Dalton Lunga	Oak Ridge National Lab	Presenter/Moderator/Scientific Committee
Daniel Hogan	CosmiQ Works	Presenter
Elizabeth M. Fancher	NASA	Participant
Gencer Sümbül	TU Berlin	Presenter
George Percivall	Open Geospatial Consortium	Participant
Gracie Pearsall	Radiant Earth Foundation	Participant/Note-taker
Hamed Alemohammad	Radiant Earth Foundation	Presenter/Moderator/Scientific Committee
Hannah Kerner	University of Maryland College Park	Presenter

Ilkay Altintas	San Diego Supercomputer Center, UC San Diego	Presenter
Joe Flasher	Amazon Web Services	Presenter
Joe Hamman	National Center for Atmospheric Research (NCAR)	Presenter
Justin Hnilo	U.S. Department of Energy	Participant
Kethelyn Papp	The George Washington University	Participant/Note-taker
Kevin Booth	Radiant Earth Foundation	Participant
Kevin Murphy	NASA	Presenter/Sponsor
Konrad Wessels	George Mason University	Presenter
Lewis Fishgold	Azavea	Presenter
Louisa Nakanuku-Diggs	Radiant Earth Foundation	Participant/Logistics
Lyndon Estes	Clark University	Presenter/Moderator/Scientific Committee/Sponsor
Manil Maskey	NASA	Presenter/Scientific Committee/Sponsor
May Casterline	NVIDIA	Presenter
Mike Humber	NASA Harvest, University of Maryland College Park	Participant
Murali Krishna Gumma	ICRISAT	Presenter
Nadya Vinogradova-Shiffer	NASA	Participant
Olha Danylo	International Institute for Applied Systems Analysis	Presenter
Patrick Grimont	ESA	Presenter
Peter Doucette	USGS	Presenter
Rahul Ramachandran	NASA	Participant/Technical Advisor
Ramakrishna Nemani	NASA	Participant

Richard Choularton	Tetra Tech	Presenter
Ryan Abernathey	Columbia University	Presenter
Sherrie Wang	Stanford University	Presenter
Soukayna Mouatadid	University of Toronto	Participant
Steven Brumby	National Geographic	Presenter
Subit Chakrabarti	Indigo, IEEE GRSS	Presenter/Note-taker/Sponsor
Sujata Emani	U.S. Department of Energy	Participant
Tasso Azevedo	MapBiomass	Presenter
Teresa Tinta	University of Maryland College Park	Participant/Note-taker
Tom Snitch	Bowling Green State University	Participant
Yigit Aytac	NASA	Participant
Yonah Bromberg-Gaber	Radiant Earth Foundation	Participant/Logistics
Zhuangfang NaNa Yi	Development Seed	Presenter/Scientific Committee