

MULTI-PATH FUSION: A HIERARCHICAL MACHINE LEARNING APPROACH FOR COMBINING DIVERSE DATA SETS FOR A FOREST MONITORING NEW OBSERVING SYSTEM

James MacKinnon¹, David Harding¹, Mark Moussa¹, Matt Brandt¹, Paul Montesano^{4,1}, Mark Carroll¹, Randolph Wynne², Valerie Thomas², Fred Huemmrich³, Jon Ranson¹

¹NASA Goddard Space Flight Center ²Virginia Tech

³University of Maryland Baltimore County ⁴ADNET Systems, Inc.

ABSTRACT

New Observing Systems (NOS) will be NASA's next generation approach for Earth remote sensing, utilizing many diverse observing capabilities to produce optimized measurements integrated from multiple vantage points and in multiple dimensions. NOS will require strong data fusion foundations to be able to intelligently combine, and retrieve information from, data coming from assets differing in characteristics like instrument type, spectral domain, and spatial and temporal resolution. We are developing an end-to-end data fusion framework employing advanced Artificial Intelligence (AI) Machine Learning (ML) techniques with the primary purpose to drive the design and operation of multi-sensor NOS for Earth sciences and beyond. This work requires building ML-enabled analytic tools and advanced environments to take advantage of high-performance computing systems for the creation of a NOS workflow that utilizes large amounts of diverse airborne and satellite observations along with ancillary information including climate and drought time series and soil properties. We are demonstrating the framework using a forest productivity and degradation use case, but the framework is designed to be applicable to a wide variety of NOS scientific objectives.

Index Terms— Data Fusion, Machine Learning, Artificial Intelligence, Remote Sensing, Forest Productivity

1. INTRODUCTION

The increasing availability of Earth observations from a diverse array of sensors on airborne and satellite platforms, acquiring data from multiple vantage points and in multiple spatial, spectral, and temporal dimensions, has led NASA to develop a New Observing System (NOS) approach for coordinated operation of multiple remote sensing assets [1]. This is necessary to meet the recommendations to NASA by the National Academy of Sciences in “Thriving on Our Changing Planet: A Decadal Strategy for Earth Observations from Space” decadal survey [2]. Optimizing the design and operation of NOS systems, to maximize their scientific capabilities, is a complex undertaking. We are developing a

framework for NOS optimization based on Artificial Intelligence (AI) Machine Learning (ML) techniques that implements a multi-path fusion system to identify the data types, their characteristics and methods of acquisition needed to achieve a NOS scientific objective. To demonstrate the usefulness of our approach for the design and operation of a NOS, we are applying our framework to a forest productivity and degradation use case (Figure 1). This was chosen because of the crucial role forests play in the sustainable provision of many critical services for ecosystem and human well-being. However, we have developed the system to also be applicable to other use cases requiring fusion of disparate data types. For the forest use case, with our technique we can determine which set of climate parameters and forest physiologic, morphologic, geographic and soil parameters are the key predictors of forest productivity and naturally occurring degradation.

We are demonstrating this capability at NSF-funded National Ecological Observatory Network (NEON) eddy covariance flux tower sites, distributed across the United States. At these sites, airborne high-resolution hyperspectral imaging and lidar mapping are done every year (or sometimes every other year) during the peak growing season [3]. Our ML truth labels are based on lidar-derived canopy height change per year between flight years. Models are being developed to predict height increase, representing productivity, and height decrease, representing degradation due to disturbance and mortality. Separate models are being developed because the predictive drivers of growth and degradation are likely to differ. The training datasets we have assembled are documented in Table 1. The dataset is built from a large variety of sources and includes things like hyperspectral, lidar, and meteorological data, all co-registered on a 1 meter grid. The size of this dataset is roughly 100 terabytes and encapsulates enough variability to allow for training of models covering North America.

Our near-term objective is to use a terrestrially based, maximum-accuracy model, deployed on a high-end

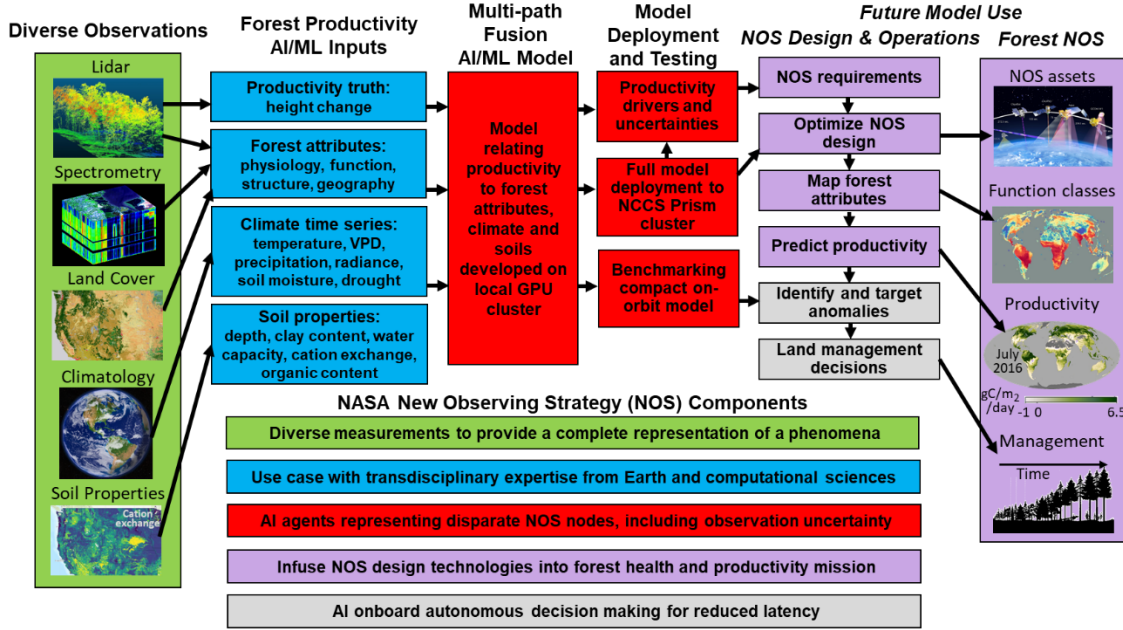


Figure 1: Our multi-path fusion framework for the forest use case, with elements colored according to components that make up the NOS approach for NASA’s future Earth observation systems. In our current work we have assembled a comprehensive set of data, from multiple sources, and are conducting model development, deployment, and testing. In future work, the framework will be used for the design and operation of a forest NOS, and ultimately for a wide range of NOS objectives.

computing (HEC) platform, for NOS design purposes. Emulating satellite observations derived from the high-resolution airborne remote sensing, to represent current, planned, and potential future orbital assets, we will conduct sensitivity studies to guide sensor and mission requirements traceability for a forest monitoring and prediction NOS. Ultimately, the model would be used during NOS operations to map global forest functional diversity, document

productivity and degradation, predict forest change under differing climate scenarios, and identify anomalies to provide actionable information for forest management decisions.

2. DATA PROCESSING PIPELINE

Our approach uses a combination of bespoke python modules in combination with open-source tools to create a framework

Table 1: Truth labels, training data and exclusion criteria for the forest productivity and degradation use case.

Products	Parameters	Sources
Truth labels	Canopy height increase between flight years for productivity models Canopy height decrease between flight years for degradation models	NEON Airborne Observing Platform lidar mapping
Spectrometry training data	VNIR-SWIR spectral curves and their derivative Broad band albedo 20 derived vegetation composition and function indices Solar incidence and azimuth angles	NEON Airborne Observing Platform hyperspectral imaging
Lidar training data	Topography elevation, slope and azimuth Forest canopy roughness Derived canopy waveforms and height percentiles Canopy illumination patch classification Illumination patch frequency, size and connectivity	NEON Airborne Observing Platform lidar mapping
Climate training data	Daily time series of air and surface temperatures, vapor pressure deficit, precipitation, evaporation, solar irradiance and soil moisture	NASA Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) [4]
Drought training data	Five-day time series of Palmer Drought Severity and Z indices from extremely wet to extreme drought	NOAA National Integrated Drought Information System (NIDIS) [5]
Soil training data	Depth, clay content, cation exchange capacity, water storage and organic matter	USDA Gridded National Soil Survey Geographic Database (gNATSGO) [6]
Data rejection criteria	Probability of forest land use < 70% Cloud cover > 10% Cloud shadow on ground	USFS Landscape Change Monitoring System (LCMS) [7] NEON Airborne Observing Platform

for rapidly prototyping various combinations of machine learning layer types and input data.

At the highest level, the training of our models is coordinated by the MLFlow tool. This is a MLOps platform that helps organize experiments, track performance metrics, and overall enables reproducibility and tracking of machine learning model development during training. [8] This tool integrates directly with our custom modules due to its simple interoperability with TensorFlow, which is the machine learning framework we used for this effort. Using this tool prevented us from needing to write a platform like this from scratch and allowed us to spend more time on model development.

The second open-source tool we leveraged was Optuna, a hyperparameter optimization framework which has a high level of integration with MLFlow. [9] Hyperparameter search is an essential part of producing high performance ML models, and this tool simplifies and automates the majority of this process. This process uses a guided random style search strategy to find the optimal hyperparameter configuration for training. Hyperparameters include things like batch size, optimization algorithm, loss functions, and most importantly in our case, the model architecture. Since we have to search a diverse set of possible inputs, and several candidate NN layer types, having an efficient way to explore this solution space is paramount.

Our framework consists of two primary modules, the first of which is a sophisticated data generator that is tasked with producing batches of data for the training process. This is a highly parametrized class to handle the large numbers of different input sources from our training dataset. When a training run is initialized, this data generator is provided with a configuration file telling it what input sources to draw from, how much data to grab, from what time frame to select from, and from what geographical extent or context. This module is able to handle differing dimensional slices from the data depending on whether that data will be fed into a simple densely connected layer or something more complex like a convolutional layer that requires inputs to be in a 2D (e.g., spatial) or 3D (e.g., spatial plus spectral) arrangement.

The data generator is essential for sampling from our dataset in an efficient way and for providing the correct arrangement of the data slices to be valid for a particular neural network architecture. This is essential for keeping our NN fed during the training process, but we also must tackle the problem of constructing this NN before training can begin. We accomplish this with our second module: the runtime NN constructor. This module takes the same configuration file as the data generator and uses it to dynamically build up the NN architecture so that it will properly consume the data from the generator. This process involves building the NN layer by layer from the specification laid out in the config file. By

leveraging a dynamic approach, we can minimize the amount of code we have to write by automating the NN construction process. Ultimately, this process builds up a hierarchical NN capable of ingesting diverse sources of data.

3. DATA FUSION AND MODEL ARCHITECTURE

The centerpiece of this method is our NN model, a hierarchical complex of deep neural networks that can handle a dataset comprised of a diverse number of input sources (Figure 2). The NN allows us to combine vastly different types of data together by training sub-component neural nets that can extract features from data, transform them into a latent space which then can be directly analyzed with downstream neural networks. These downstream neural networks can perform a number of tasks like regression or classification.

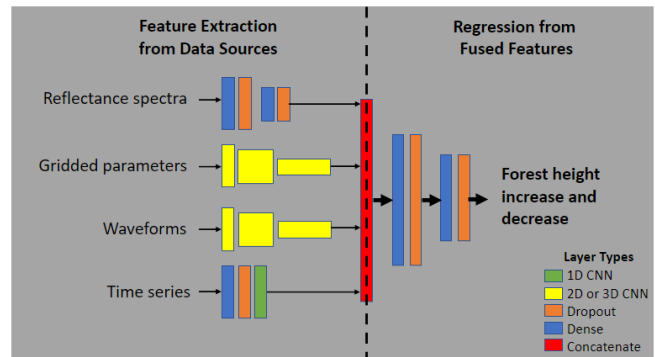


Figure 2: This shows the flow of data through our multi-path ML model. A diverse set of inputs are matched to appropriate ML layer types to extract features, which are then concatenated together and passed to a regression stage that predicts the desired science product. In our case, forest height increase, representing productivity, and decrease, representing degradation.

Each NN is unique to the data source, to be able to best take advantage of NN layers best suited for the data type. This means for spatial data or raster data we can utilize 2D or 3D Convolutional neural networks to extract features relating to spatial (or spectral) locality, additionally for temporal data we can use Long Short-Term Memories, Temporal Convolutional Networks, or Transformers, and we can use more traditional fully connected networks or 1D convolutions to pull features out of vector data. The outputs of these models can then be fused together with simple concatenation. This method allows us to arbitrarily fuse data sources in a way that improves our overall system performance over traditional approaches. This sophisticated architecture requires an equally sophisticated and well-vetted dataset for truth labels and training parameters. Because the requisite data sets for our forest use case had not previously been compiled, a substantial component of our work was data wrangling, including the definition of data product

requirements, accessing multiple data servers and co-registration in data stacks.

4. CONCLUSIONS

NASA's next generation Earth science missions will utilize the NOS approach to fulfill the demanding challenges for advancing understanding of Earth systems laid out by studies like the "Thriving on Our Changing Planet: A Decadal Strategy for Earth Observations from Space" decadal survey. Large numbers of observations from a diverse set of platforms and instruments, with different field of views and temporal, spatial and spectral resolutions, will require a level of data fusion that does not currently exist. This effort lays out a framework for dealing with these issues through a novel combination of existing tools, machine learning approaches, and data processing techniques. Using this framework, we show a state-of-the-art result using forest productivity and degradation as a case study.

5. ACKNOWLEDGEMENTS

We would like to thank the NASA Earth Science and Technology Office for funding this work, proposal # NNH21ZDA001N-AIST, under the Advanced Information Systems Technology program.

Resources supporting data preparation and manipulation were provided by the NASA High-End Computing (HEC) Program through the NASA Computational Information Science and Technology Office at Goddard Space Flight Center.

6. REFERENCES

- [1] M. Little, M. Seabloom, B. Quam, J. LeMoigne-Stewart, D. Crichton, and M. Cole, "Sensor Webs of Earth Science Instruments to Improve Our Understanding of Natural Phenomena and Physical Processes", 2018. IEEE International Geoscience and Remote Sensing Symposium, pp. 264-266. 2018. DOI: 10.1109/IGARSS.2018.8517513.
- [2] National Academies of Sciences, Engineering, and Medicine. "Thriving on Our Changing Planet: A Decadal Strategy for Earth Observation from Space". Washington, DC: The National Academies Press. 2018. <https://doi.org/10.17226/24938>.
- [3] M. Musinsky, T. Goulden, G. Wirth, N. Leisso, K. Krause, M. Haynes, and C. Chapman, "Spanning scales: The airborne spatial and temporal sampling design of the National Ecological Observatory Network", *Methods in Ecology and Evolution*, 13, pp. 1866-1884. 2022. DOI:10.1111/2041-210X.13942.
- [4] R. Gelaro, W. McCarty, M.J. Suárez, et al., "Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA-2)", *Journal of Atmospheres*, 30(14), pp. 5419-5454, 2017. DOI:10.1175/JCLI-D-16-0758.1.
- [5] J. Rhee and G. J. Carbone, "A Comparison of Weekly Monitoring Methods of the Palmer Drought Index", *Journal of Climate*, 20(24), pp. 6033-6044, 2007. DOI:10.1175/2007JCLI1693.1.
- [6] National Soil Survey Center, "Gridded Soil Survey Geographic (gSSURGO) Database User Guide", U.S. Department of Agriculture. 2016. https://geodata.lib.ncsu.edu/fedgov/nrcs/gSSURGO_UserGuide.pdf.
- [7] USDA Forest Service. "USFS Landscape Change Monitoring System Conterminous United States version 2021-7". Salt Lake City, Utah. 2022. <https://data.fs.usda.gov/geodata/rastergateway/LCMS/index.php>.
- [8] A. Chen., et al. "Developments in MLflow: A System to Accelerate the Machine Learning Lifecycle". DEEM'20: Proceedings of the Fourth International Workshop on Data Management for End-to-End Machine Learning, pp. 1-4. 2020. <https://doi.org/10.1145/3399579.3399867>.
- [9] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama. . "Optuna: A Next-generation Hyperparameter Optimization Framework." KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2623-2631. 2019. <https://doi.org/10.1145/3292500.3330701>.