Fluvius Executive Report

Prepared for: Microsoft AI for Earth

By: Analytics Lab at CSP

Vincent A. Landau
Data Scientist

Luke J. Zachmann
Senior Scientist

Tony Chang
Data Scientist and Principal Investigator

ANALYTICS LAB
CONSERVATION SCIENCE PARTNERS

Fluvius Executive Report

by

Vincent A. Landau
Data Scientist

Luke J. Zachmann
Senior Scientist

Tony Chang
Data Scientist and Principal Investigator

Introduction

Since the 1970s, forests in the Amazon region have been increasingly converted to pasture lands and human settlements, and have undergone increased natural resource extraction. As of 2006, nearly 95% of deforestation was within 5.5 km of roadways or within 1 km of navigable rivers, resulting in significant impacts on freshwater ecosystems (Souza-Filho et al. 2016).

One measurable impact on the riparian system is greater discharge and increased sediment flux in the channels of the main rivers, and these changes can be used as a proxy for ecosystem degradation (Coe et al. 2011, Costa et al. 2003, Latrubesse et al. 2009). One means of understanding the impact of deforestation or mitigation through reforestation is by monitoring the sediment flux over time. Understanding the dynamics of hydro-sedimentological processes in river basins provides critical data for decision-making and supports management planning for the rational use of natural resources. However, due to the logistical difficulties in measuring sediment flux in a distributed way through time and space, there is a significant opportunity to increase monitoring frequency of water bodies through satellite remote sensing.

New artificial intelligence (AI) techniques such as convolutional neural networks (CNN) have shown promising performance in the application of object detection and continuous forest structure estimation when combined with remote sensing imagery (Chang et al. 2019). CNNs can draw on the spatial contextual information embedded in images, and often outperform traditional remote sensing approaches that are limited to spectral analysis of individual pixels alone.

In this report, we describe a joint effort by Microsoft, Instituto Tecnológico Vale (ITV), and Conservation Science Partners (CSP) to prototype and deploy a new machine learning model that is capable of estimating suspended sediment concentration (SSC) based on the remotely sensed Harmonized Landsat Sentinel image collection. The goal was to develop a tool that can be used in "near real-time" to monitor the impact of forest land use/cover change in the Itacaiúnas River watershed.

Ideally, a monitoring system that achieves near real-time predictions would have a high frequency of cloud-free observations in both the rainy and dry seasons. preliminary invsetigation showed that number of satellite images that have less than 60% percent cloud decreased significantly during the rainy season, meaning that there will be far fewer cloud-free pixels in the rainy season from which to make predictions of SSC compared to the dry season. Developing a method to identify cloud-free "water" pixels to use for modeling, particularly in the rainy season, is essential for enabling the generation of time series predictions of SSC with the highest cadence possible.

This report will cover all core project milestones and objectives. Briefly, these included the following items:

- Co-development of a continuous pipeline for core datasets, preprocessing, and storage on a cloud based platform.
- Prototyping of various experimental deep learning models to analyze remotely sensed image "chips" of rainforest watersheds and estimate suspended sediment concentration.
- Diagnose the deep learning model on field based measurements and compare to performance of current state-of-the-art spectral based models.
- Deploy the deep learning model API on the Itacaiúnas River watershed over space and time, and provide metadata and documentation write-up.
- Co-production of a peer-reviewed scientific manuscript.
- Web-application creation and deployment.

Methods

The goal of the Fluvius model was to enable predictions of suspended sediment concentration (SSC) from satellite imagery. Our approach entailed extracting image "chips" for all dates and locations for which we had SSC observations from gauging stations, then using the spectral characteristics of those chips to predict SSC. A properly trained and validated model can make predictions of SSC not only for sites and times

at which we have ground-based measures, but also for sites and times for which such measures are lacking entirely. Our workflow primarily consisted of two components: data and modeling. Each of these are detailed in respective sections below.

Data

Gathering and formatting training, validation, and test data for model fitting, optimization, and performance evaluation is the first step in any machine learning modeling effort. The source data for this modeling effort comes from in situ observations of SSC from rivers and streams in both Brazil and the United States. We compiled data from each of the four data sources described in Table 1 and converted SSC measurements to units of milligrams per liter (mg/L). Each SSC observation has associated metadata and coordinates describing the time and site at which the measurement was taken.

Table 1: Descriptions of data sources, number of sampling sites, and total number of suspended sediment concentration (SSC) observations. The total number of observations available *after* QA/QC are indicated parenthetically. Note that the process of selecting reliable observations meeting all analysis criteria removes a significant fraction of the total available samples.

Data Source	Description	Number of Sites	Number of Observations
USGS	High-accuracy ($R^2 > 0.9$) modeled daily SSC from the United States Geological Survey (USGS) monitoring stations in the U.S.	35 (14)	54619 (846)
USGSI	in situ measurements of SSC at monitoring stations in the U.S.	43 (2)	702 (8)
ITV	Measurements from Instituto Tecnológico Vale monitoring stations	16 (11)	292 (45)
ANA	Measurements from Brazilian National Water Agency (ANA) monitoring stations	22 (18)	396 (47)

After compiling and formatting the tabular data, we used geolocation and timestamp information to associate each observation with a satellite image to use for model training, wherein the spectral characteristics of the image chip were features in the model, and the associated SSC measurement (log-transformed) was the response. We used data from the Sentinel-2 L2A image archive (Drusch et al. 2012). In order for a Sentinel-2 L2A image tile to be matched with an SSC observation, two initial criteria had to be met: the sensing date had to be within eight days of the SSC observation date (except for USGS stations, which had daily measurements; see Table 1), and the image tile had to have total cloud cover no higher than 80%. If these criteria were met, then a 1 km^2 square chip centered on the water monitoring station was extracted from the Sentinel-2 L2A tile for further QA screening and pre-processing.

To reduce noise in our input data, we removed clouds from the image chip and retained only pixels that corresponded to water. To remove non-water pixels, we retained only pixels that were classified as water in the Impact Observatory land use/land cover data set (Karra et al. 2021) *and* that had a Modified Normalized Difference Water Index (MNDWI) value greater than 0. Following the isolation of water pixels, we calculated the mean reflectance values in water pixels within the image chip for each Sentinel-2 L2A band. These aggregated reflectance values then served as features in our model (we additionally added an indicator variable denoting whether the site was in Brazil as a feature; Table 2). Observations for which a matching Sentinel-2 L2A chip could not be obtained (either due to cloud cover or a lack of pixels classified as water following cloud masking) were dropped.

To further improve the quality of input data, we manually screened the image chips and associated water masks for each image chip that was matched to an SSC observation. We dropped observations for which the associated image chip had poor quality (i.e. inaccurate cloud or water masks, excessive cloud shadow,

Table 2: Features derived from Sentinel-2 L2A imagery and observation metadata used in our models.

Feature	Description	Spatial resolution (m)
AOT	Aerosol optical thickness	20
B02	Blue	10
B03	Green	10
B04	Red	10
B05	Red edge 1 (705nm)	20
B06	Red edge 2 (740nm)	20
B07	Red edge 3 (783nm)	20
B08	Near infrared (842nm)	10
B8A	Narrow near infrared (865nm)	20
B11	Short wave infrared 1 (1610nm)	20
B12	Short wave infrared 2 (2190nm)	20
I_{Brazil}	An indicator denoting whether the site is in Brazil ($I_{Brazil} = 1$ for Brazil sites, and 0 otherwise)	-

etc.). Additionally, we only retained observations for which there were at least 20 water pixels in the associated image. This served to reduce noise in the training data by mitigating the impact of outlier pixels on aggregated reflectance values. The combination of all of the QA measures described above significantly decreased the number of samples available for modeling (see sample sizes indicated parenthetically in Table 1). However, such measures are necessary to ensure the model was trained using only the best available information.

We wanted the model to be able to make accurate predictions of SSC at sites it never saw during training, so we partitioned the data by allocating sites into train and test sets. Specifically, we withheld half of all ITV and ANA sites entirely from training, and used 100% of all USGS observations in training. The USGS observations were critical in training the model, but were not from the focal geography (the Amazon). We wanted to know how well the model performed in its intended geography. Thus, we had to ensure we had out-of-sample testing observations from the Amazon. After the initial train / test split, the training set was further subdivided into train and validation sets over $k = 5$ folds. We visually inspected the distributions of all features and response values in each fold to ensure each fold was appropriate / representative prior to conducting the grid search, which we describe below.

Model

We trained a multi-layer perceptron model (MLP), a type of neural network, using the feature data we extracted for the SSC observations using the PyTorch (Paszke et al. 2019) library in Python. We used stochastic gradient descent with momentum as our optimization algorithm, and each model was trained for 15,000 epochs. We employed a learning rate scheduler with a gamma of 0.1 with milestones (at which the current learning rate is multiplied by the gamma value) at epochs 7,500, 10,000, 12,500, and 14,000. All features were normalized prior to model fitting, and we used the log-transformed SSC as the response. We conducted a hyperparameter grid search in which we varied model features, learning rate, batch size, weight decay, activation function, and model architecture (i.e. number and size of the hidden layers in the neural network) (Table 3).

Due to data limitations, we used a k-fold cross validation approach, in which the training partition described above was split into five folds, and fit the model five times. For each model fit, a different fold was used as the validation partition, and the final performance of each model was calculated as the mean loss across all five model fits. We calculated two forms of loss for evaluating model performance. For the first form, we used the root mean squared error (RMSE) across all observations. For the second, we calculated

Table 3: Hyperparameters (and values) that were included in the grid search.

Hyperparameter	Description	Values
Learning rate	The learning rate used during optimization	$1 \times 10^{-4},\ 5 \times 10^{-4}$
Batch size	The number of samples used per batch during training	$16, 32$
Features	The set of features used in the model (see Table 2 for descriptions)	[B02, B03, B04, B05, B06, B07, B08, B8A, I_{Brazil}], [AOT, B02, B03, B04, B05, B06, B07, B08, B8A, I_{Brazil}], [B02, B03, B04, B05, B06, B07, B08, B8A, B11, B12, I_{Brazil}], [AOT, B02, B03, B04, B05, B06, B07, B08, B8A, B11, B12, I_{Brazil}]
Activation function	The function to used in activation layers between hidden layers in the MLP	`nn.SELU()`,`nn.PReLU()`
Weight decay	The weight decay used with SGD optimization	$0, 0.01$
Architecture	The architecture of the model, represented by a list of length equal to the number of hidden layers, with elements corresponding to the number of neurons in each layer.	$[5, 3, 2], [4, 4, 2], [4, 2, 2], [6, 5], [6, 3]$

RMSE per site, then averaged the site-level RMSE values. These two loss metrics were not equivalent because each site had a different number of observations. We found that using the mean of these two forms of loss as our final loss metric resulted in a good balance between having a low error rate for as many observations as possible, while also ensuring that the model performs well across a variety of different sites. To evaluate the final model, we calculated the R^2 of the observed vs predicted SSC on the native scale (mg/L).

Following the identification of the top model, we extracted image chips for the entire Sentinel-2 L2A archive at each Brazilian water monitoring station to deploy our model and generate time series predictions of SSC. Each image chip was generated and processed using the same methods for preparing training data described above, but chips were not manually screened.

Code to reproduce our entire analysis is available on GitHub, and documentation will be made available on a standalone web page.

Results and Discussion

We fit a total of 320 different models for our hyperparameter grid search. The top model (Table 4) had an R^2 of 0.72 for the withheld test set. Importantly, this test set consisted entirely of sites for which no observations were "seen" by the model. A plot of observations vs. predictions for the withheld test set is shown in Figure 1. We applied the top model to the full times series of Sentinel-2 L2A images at each ITV motitoring station to generate time series plots of SSC (Figure 2).

As is evident in Figure 2 (orange points), observations of high SSC were rare in the training and test data. As a result, our model was primarily tuned to predicting low SSC. Focusing future monitoring on capturing high SSC events and retraining models will likely serve to improve predictive performance across the full range of possible SSC values. One challenge to this is that high SSC events may be correlated with high cloud cover (i.e. during and following large rain events, and more broadly in the rainy season), meaning finding matching cloud-free Sentinel-2 L2A images may not be possible.

Table 4: Hyperparameter values for the top model. See Table 2 for more information on the features in the model.

Hyperparameter	Value
Learning rate	5×10^{-4}
Batch size	16
Features	[B02, B03, B04, B05, B06, B07, B08, B8A, I_{Brazil}]
Activation function	`nn.PReLU()`
Weight decay	0.01
Architecture	$[4, 4, 2]$

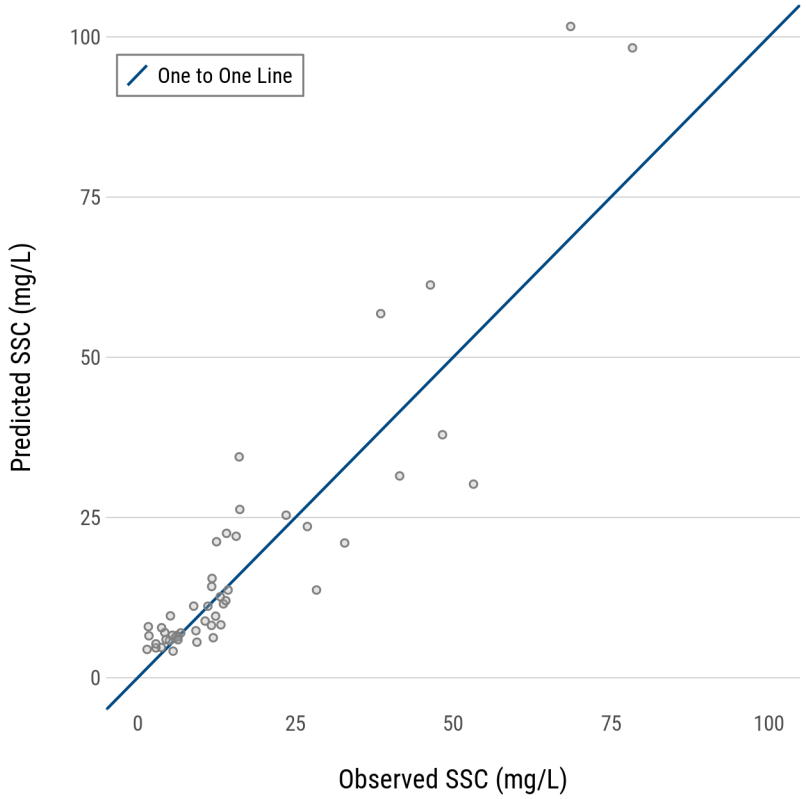

Figure 1: The one-to-one plot of observed versus predicted SSC for the withheld test partition.

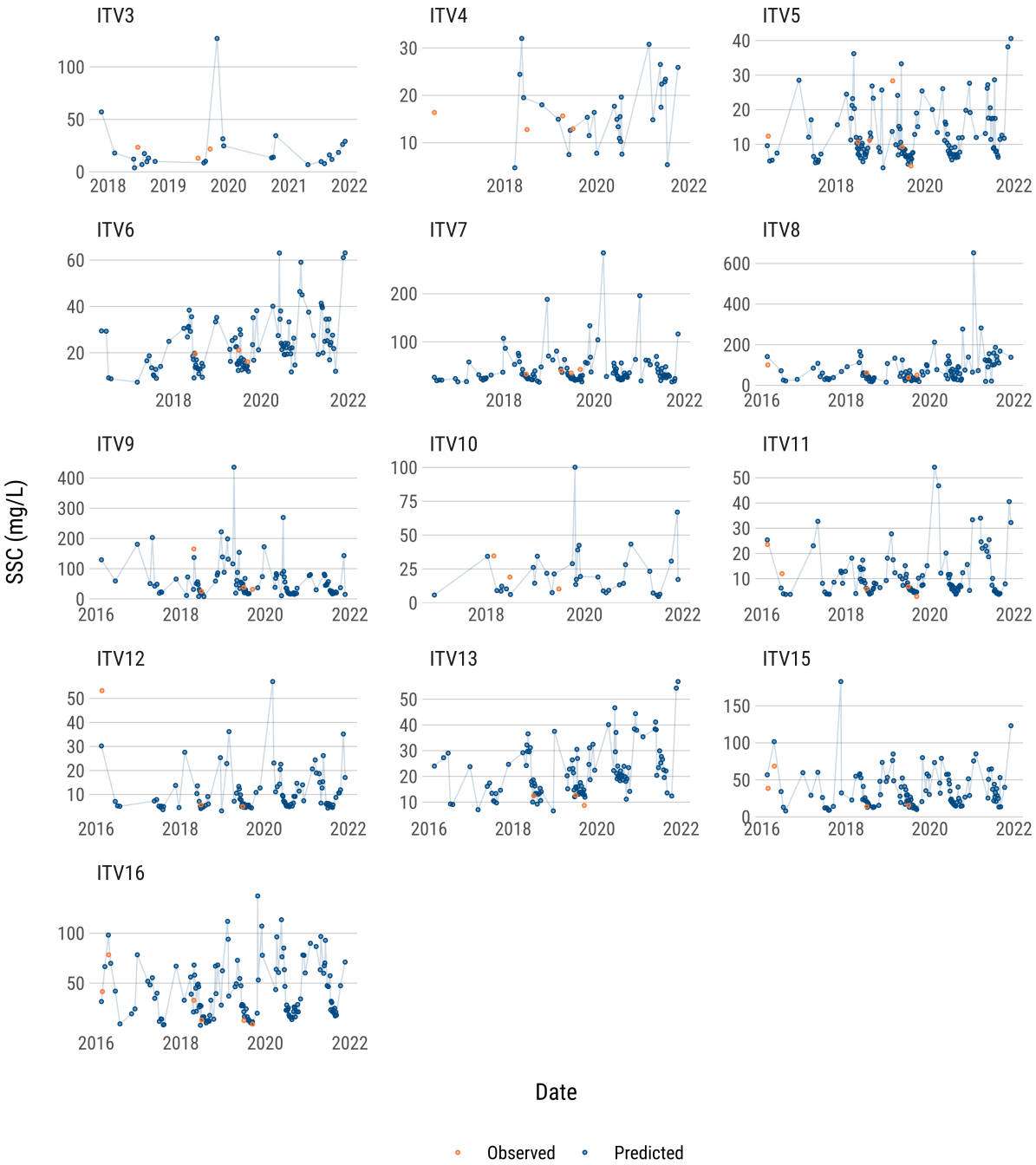

Figure 2: A plot showing time series predictions of SSC based on the top model (blue points), with in situ field observations of SSC overlaid (orange points) for ITV monitoring sites.

References

Chang, T., B. P. Rasmussen, B. G. Dickson, and L. J. Zachmann. 2019. Chimera: A multi-task recurrent convolutional neural network for forest classification and structural estimation. Remote Sensing, 11:768.

Coe, M., E. Latrubesse, M. Ferreira, and M. Amsler. 2011. The effects of deforestation and climate variability on the streamflow of the araguaia river, brazil. Biogeochemistry, 105:119–131.

Costa, A. T., H. A. Nalini, J. C. de Lena, K. Friese, and M. Mages. 2003. Surface water quality and sediment geochemistry in the gualaxo do norte basin, eastern quadrilátero ferrífero, minas gerais, brazil. Environmental geology, 45:226–235.

Drusch, M., U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, et al. 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sensing of Environment, 120:25–36.

Karra, K., C. Kontgis, Z. Statman-Weil, J. C. Mazzariello, M. Mathis, and S. P. Brumby. 2021. Global land use/land cover with sentinel 2 and deep learning. *in* 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Pages 4704–4707. IEEE.

Latrubesse, E. M., M. L. Amsler, R. P. de Morais, and S. Aquino. 2009. The geomorphologic response of a large pristine alluvial river to tremendous deforestation in the south american tropics: The case of the araguaia river. Geomorphology, 113:239–252.

Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32.

Souza-Filho, P. W. M., E. B. de Souza, R. O. S. Júnior, W. R. Nascimento Jr, B. R. V. de Mendonça, J. T. F. Guimarães, R. Dall'Agnol, and J. O. Siqueira. 2016. Four decades of land-cover, land-use and hydroclimatology changes in the itacaiúnas river watershed, southeastern amazon. Journal of environmental management, 167:175–184.