SMU Data Science Review

Volume 2 | Number 3 Article 9

2019

Machine Learning and Deep Learning Applications for International Ocean Discovery Program Geoscience Research

Brandon De La Houssaye
Southern Methodist University, bdelahoussaye@mail.smu.edu

Peter Flaming
Southern Methodist University, pflaming@mail.smu.edu

Quinton Nixon
Southern Methodist University, qnixon@mail.smu.edu

Gary Acton
Texas A&M University, acton@iodp.tamu.edu

Machine Learning and Deep Learning Applications for International Ocean Discovery Program Geoscience Research

Peter Flaming[1], Brandon De La Houssaye[1], Quinton Nixon[1], and Gary Acton[2]

[1] Master of Science in Data Science, Southern Methodist University, Dallas TX 75275 USA {pflaming,bdelahoussaye,qnixon}@smu.edu
[2] International Ocean Discovery Program, Texas A&M University, College Station, Texas 77845 USA acton@iodp.tamu.edu

Abstract. In this paper, we present the use of a machine learning-based architecture for use in regression tasks with image data collected by the International Ocean Discovery Program (IODP) – in part funded by the U.S. National Science Foundation (NSF) – for scientific use during seagoing operations aboard the JOIDES Resolution scientific drillship as well as onshore at the Texas A&M University IODP Headquarters. Our data science-driven approach integrates modern programming techniques in Python, computer vision, machine learning, and deep learning applications with the traditional geoscience linear regression architecture such that the modeling of high-resolution spectral data with the vast amounts of petrophysical and geochemical data can be trained end-to-end with open source applications to predict independent global proxies O^{18}/O^{16} isotope ratio (δ^{18}O) for geologic age of ocean sediments through time. First, we show that computer vision applications like OpenCV can be employed to extract, transform and load spectral data into a continuous data array in an automated function that webscrapes the online IODP database. Next, we present generalizable machine learning regression modeling of IODP core data driven by the Support Vector Machine (SVM) Regression (SVR) algorithm and hyperparameter tuning with a K-fold cross validation CV grid search technique using popular Scikit-learn packages and functional programming. Finally, we demonstrate a K-fold cross validated deep learning prediction of the δ18O variations with a deep neural network using Keras and TensorFlow 2.0 with generalizability for the vast amounts of ocean sediment data maintained by IODP. We find that the machine learning and deep learning approaches both extrapolate quite well outside of the training data, and the use of these methods pave the way for future data science applications in scientific exploration and discovery at IODP.

Keywords: Data Science · Geoscience · δ^{18}O · IODP · JOIDES Resolution · Texas A&M University · NSF · Open Source · Python · OpenCV · Scikit-learn · Keras · TensorFlow 2.0 · Machine Learning · Deep Learning · Regression · DNN · Scientific Exploration · Discovery.

1 Introduction

There is significant work undertaken by a broad array of scientists, engineers, industrialists, and officials seeking to understand Earth's dynamic history. A widely studied topic today is Earth's climate, and it's particularly a hot one. Scientists from all around the world and especially the geoscientists at the International Ocean Discovery Program (IODP) Headquarters located at Texas A&M University have a renowned ability to build scientific models that explain the past, present and future of Earth processes by analyzing the vast amounts of data (petrophysical, geochemical, etc.) collected during IODP's past fifty years of operation. However, the sheer volume of information that comes from each of the scientific expeditions conducted varies across the years as technological advances made new types of data available (high-resolution images, etc.) to be processed into a clean and normalized format where the geologists can perform their scientific studies.

As it relates to this paper, of particular interest are those data consisting of color images of sediment samples taken from the sediment cores recovered from beneath the ocean floor. Historically at IODP, the image data collected from the sediment core samples were only able to be integrated by select individuals who could write custom programs to collect the spectral data and personally model the relationships to the past and present climate. New camera technologies employed during expeditions aboard the JOIDES Resolution can capture data in near real-time and in high-resolution, presenting the need for a modernized approach to extract, transform, and load the data before it can be used to analyze the color variation relationships contained within the images with other features that model the past and future of the Earth's climate with modern statistical techniques used by data scientists with machine learning and deep learning models.

The climate models today would benefit greatly with the astronomical tuning of past temperature variations using relationships found in the color variation datasets derived from high-resolution images of ancient sediments. Modeling the variations with many complex phenomena elucidate the underlying dynamics of a particular system and can provide new features for predictions over a very wide range of conditions. However, in order for machine learning approaches to be widely accepted by and applied to geoscience, there is a need for interpretable and generalizable models that can extract meaningful information from complex datasets and extrapolate outside of the training data. As more cores are gathered, and images are captured the scientific community seeks to gain better insights into the whole of what the core image can tell us by using more precise and new types information gathering tools, there is a real need to transform this additional information into a format where scientific understanding can occur.

In order to extract the color spectrum at scale, we set out to first understand the core sample photographs. We have to first collect the photographs and place them in a format whereby the pixels can be analyzed, recorded, and classified accordingly. The means for extracting color spectrum layers is not difficult in that a variety of computing packages are readily available that perform such tasks.

However, the key in performing this step with accuracy is to remove outlier values from color spectrum observations at the pixel level. Under the tutelage of a subject matter expert, the team worked to extract the target color spectrum and remove edges, cracks, and other related matter from the extracted pixels so that what remained was data of interest. Additionally, the subject matter expert provided a test dataset with known pixel data to the team so that a prediction model could be compared to the true data established from the known extracted pixel information provided.

In building the model, the key focus is one of interpretability and scalability. Scientists can – and have – gone through the sample photos and translated the colors into a flat (data) file based on color spectrum using spectrophotometry. This means that for the core sample, at a given location in the photo, on a grey scale from 0 to 255, the color shade is provided. The results are then mapped and plotted, and the resulting data can be statistically analyzed. So, the question at hand is why would a team of data scientists be needed? The answer is that it has been done 'by hand' over a painstakingly long and expensive process. What data science can do is develop a model that for a given image, the resulting classification of the image (against a grey scale) becomes automated. The explainable model, once it has been trained to achieve sought accuracy metrics, is scalable. The model can be employed in real time to additional core photographs contained within a relational database schema. The end result is a data tidal wave of new features to use within their models for the population of geoscientists patiently waiting to run their experiments.

In a practical sense, this paper can be described as containing a 'proof of concept' model that focuses on transforming one element of the information to be gleaned from the sample (the high-resolution photographs) into a data-driven usable format where existing statistical models can be applied in order to better understand the Earth's history. The end result is an explainable machine learning-based model that extracts spectral data, transforms that data into new color features, and loads the multi-dimensional images into one-dimensional arrays that can be used with other scientific data to predict other geoscience-related data. The color features engineered are used as independent variables with known features used for linear regression models created by IODP scientists to predict many physical properties of the geoscience data related to the drill site where the core samples were collected.

The remainder of this paper includes an overview of the data gathered and utilized; the methods and experiments performed; the results (of the model developed); an analysis of those results (i.e., accuracy indications); ethical considerations; and overall conclusions. Finally, this paper contains a summary of potential future work as well as the associated references and appendices.

2 IODP Expedition 306 Research Background

IODP Expedition 306, North Atlantic Climate 2 (NAC2) [8] consisted of three drilling locations (Sites U1312, U1313, U1314) designed to generate a late Neo-

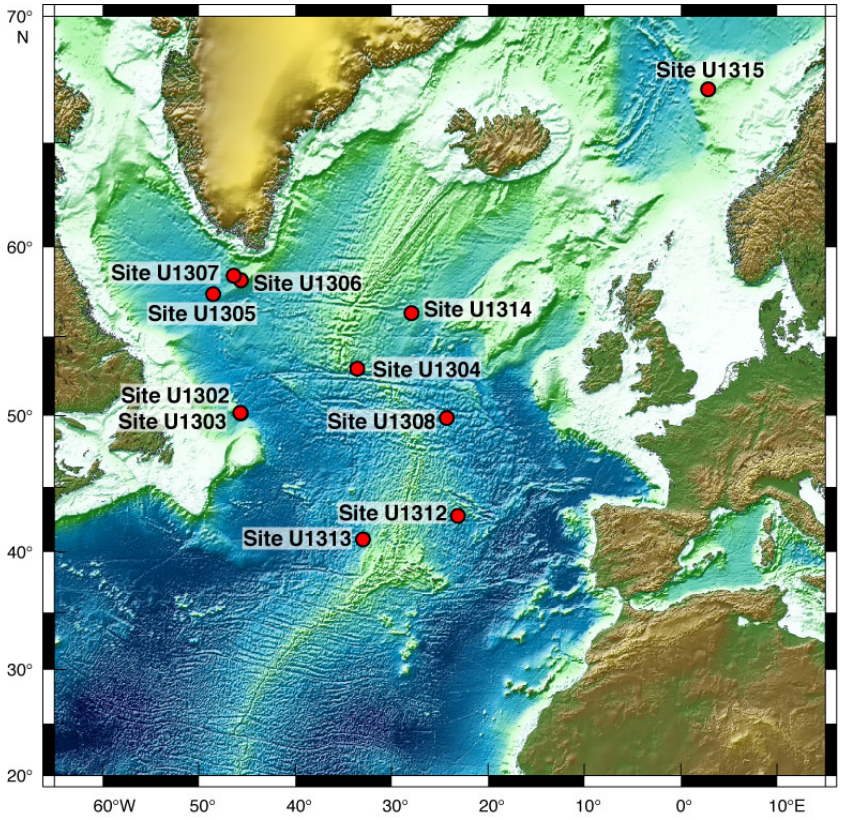

Fig. 1: IODP Expedition 303/306 Drill sites.

gene–Quaternary chronostratigraphic template for North Atlantic climate proxies with independent global stratigraphic signals known as oxygen isotopes δ18O and relative paleointensity (RPI) which allows for their correlation at a sub-Milankovitch scale and their export to other parts of the globe using a paleointensity assisted chronology (PAC)[1].

The North Atlantic Climate 2 paleoceanography study [8] recovered complete sedimentary stratigraphic sections at three sites and the excellent sections from Sites U1313 and U1314 provided the proxy requirements, including adequate sedimentation rates, to study millennial-scale environmental variability in terms of ice sheet-ocean interactions, deep circulation changes, and sea-surface conditions. Site U1313 provides a unique and complete Pliocene–Pleistocene sediment section with remarkably constant sedimentation rates of 4–10 cm/k.y.[4], and the recovered sections extend back to 3.2 and 5 million years (Ma), respectively.

Site U1313 will allow for an optimal reconstruction of the phasing of temperature records and their relationship to ice sheet instability and changes in deep water circulation throughout the last 5 million years [9] at a higher resolution

scale. The dataset provided represents the persevered geologic history discovered by IODP during drilling operations at Sites throughout the North Atlantic Ocean.

2.1 Site U1313 Summary

Site U1313 and the core section images associated with them are the focus of this analysis due to the benchmark it sets for the long-term (millions of years) surface and deep ocean climate records from the subpolar North Atlantic [5] [7]. Four holes (A, B, C, and D) were drilled at Site U1313, reaching a maximum depth of 308 meters below sea floor (mbsf) provides the means of constructing an ostensibly complete composite section that dates back into the late Miocene ˜6 Ma with mean sedimentation rates of ˜5 cm/k.y. for the Pliocene–Quaternary and perhaps as high as ˜13–14 cm/k.y. for the upper Miocene [7].

Stratigraphic correlation among holes was done unambiguously down to 168 mbsf, based largely on sediment color reflectance data recorded from shipboard measurements that mimic the marine oxygen isotope curve [9]. Magnetic stratigraphy was interpretable through most of the Gauss Chron data down to ˜150 mbsf. The deeper depths from 150-308 mbsf consist of low magnetization intensities and were not useful in the stratigraphic correlation. Diatom flora are present within the upper ˜70 mbsf (Pliocene–Pleistocene interval) and calcareous nannofossils were abundant and preserved enough to complete a succession of nannofossil datums for the upper most 158 mbsf, dating back to ˜3 Ma [3] [4].

Studied Pliocene–Pleistocene planktonic foraminifer events at this site [2] and assigned ages using the correlation of sediment color/reflectance data to an oxygen isotope reference curve. Those ages have been compared with the same datums in the Mediterranean where calibration was previously achieved through astrochronology. An oxygen isotope stratigraphy has been generated from planktonic and benthic foraminifers for part of the record, specifically the MIS 10–16 interval [10], providing both chronological control and information about surface and deep-water conditions.

Planktonic and benthic oxygen isotope data, lithic fragment, and physical grain size data at Site U1313 indicated that major Heinrich-type ice-rafting events were associated with Terminations V and VII and with MIS 10 and 12 [10]. Lower benthic δ13C values indicate weakened meridional overturning circulation (MOC) and the presence of Antarctic Bottom Water during all glacial periods in this interval (MIS 10–16) with strong MOC during all interglacials [6]. Biomarker analyses, performed shipboard, indicate that alkenone-derived SSTs show variability from ˜13 to 19 Celsius during the Pleistocene [9] with a few data points from the late Pliocene interval display SST values of ˜17 to 22 Celsius.

3 IODP Core Dataset and Image Preprocessing

3.1 Expedition 306 Site U1313 Dataset

The core dataset containing all of the petrophysical and geochemical measurements made and also the independent global proxies previously correlated to Site

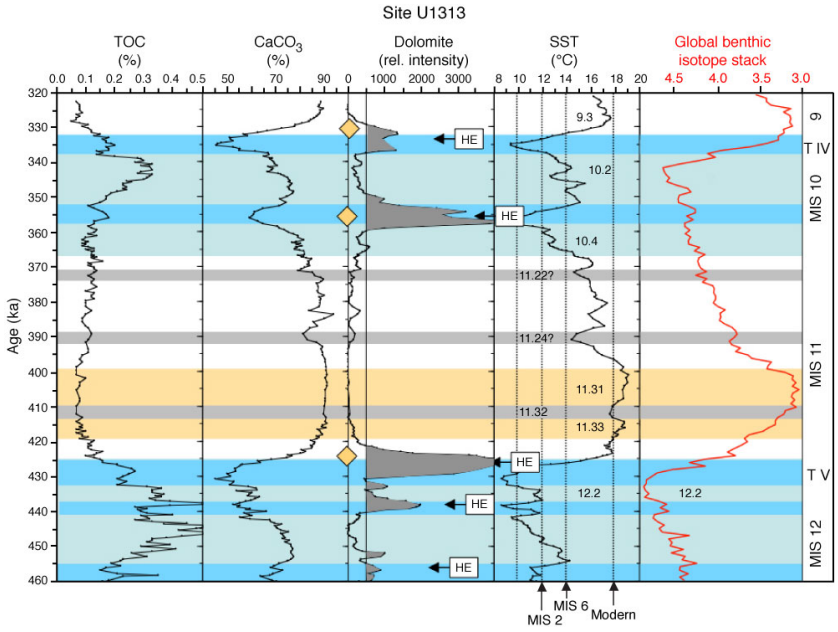

Fig. 2: Three-point moving averages of total organic carbon (TOC), calcium carbonate (CaCO3), relative (rel.) XRD intensities for dolomite, and alkenone-based sea-surface temperatures (SST). Global benthic isotope stack from Lisiecki and Raymo (2005). Marine isotope stage (MIS) and substage numbering from Bassinot et al. (1994). T = Termination. Orange rhombs indicate occurrence of Heinrich-like events (HE) at Site U1308 (from Hodell et al., 2008). Modern SST (from Locarnini et al., 2006) and estimated SSTs for MIS 2 and 6 are indicated. Figure from Stein et al. (2009). Source: IODP Expedition 306 Proceedings

U1313 needed to build the model was provided by Texas A&M IODP Headquarters and is freely available as an open source download from the IODP database. The Site U1313 core section images needed for the color extraction techniques are also available as high-resolution downloads from the same IODP database as the core dataset.

The core dataset is comprised of six continuous features all standardized (by subtracting their mean and dividing by their standard deviation) accounting for the last 4,260 ka of time with just 1,903 instances of data measured by the participating scientists and technicians from IODP Expedition 306 at Site U1313 in the North Atlantic Sea from 2 March – 25 April 2005 [8]. Dataset features were derived from shipboard measurements made by IODP participants during their time at sea with Expedition 306 in the North Atlantic Sea at Site U1313. Below in Table 1 are descriptions of the IODP features provided:

Table 1: IODP Expedition 306 Site U1313 Core Data.

Geologic Feature	Geologic Description of Features Provided by IODP Research
Age (ka)	Geologic Age of sediments derived from δ18O proxies
d18O (ng/g)	δ18 Oxygen isotope values derived from δ18O
SST (C)	Sea Surface Temperatures derived from alkenone variations
Alkane (ng/g)	Hydrocarbon concentrations from Calcium Carbonate
Alkanol (ng/g)	C26 hydrocarbon concentrations from Calcium Carbonate
Sediment (cm/ka)	Sedimentation rate derived from biomarker accumulations

3.2 Core Image Data

High resolution cameras are currently in-use as scanning devices known as a Section Half Image Logger (SHIL) on every seagoing expedition today and these SHIL images produce the spectrophotometry data for every core section recovered during operations at sea. The spectral data produced by the SHIL is fairly new with respect to the 50-year history of IODP, and the vast amount of spectral data waiting to be explored is contained within .JPEG and .PNG image files throughout the IODP Database.

The SHIL images represent the cross section of the core sample taken from deep beneath the ocean floor surface. As drilling occurs, the sample is pulled up as a core approximately 9 meters in length by ˜6.5 centimeters in diameter and cut into sections approximately 1.5 meters long that are analyzed as a whole and as well as in half to create two samples (one working and one archive) with many petrophysical and geochemical instruments. The archive section half is the preserved sample that will be photographed, and those images are used for our project. Each image is stored and tagged with a unique sample identification number that points to the Expedition-Site-Hole-Section of each image captured.

The Expedition 306 Site U1313 section half images are organized according to their core and section for each drill site (the drilling location on Earth) and represent a unique sequence of sediment deposition that is a record of many geologic processes occurring in and around that region.

The core section halves are photographed with a high-resolution camera capturing a sub-millimeter resolution of the geologic sediments and rocks contained within the core. Below Fig. 3 shows a batch of the high-resolution photos for all sections collected from a single core.

Most of the images are stored like that shown in Fig. 3 for the older IODP Expeditions and the use of machine learning to automate the tedious task of pixel value extraction with object detection and spectrophotometry color variation extraction offers a great service to the scientific participants who wish to use these data for modeling and don't have the time and or means to extract those data for themselves. As seen in Fig. 4 is how IODP offers a closer look at a single section available with (consumer image) and without (cropped image) a depth scale: Figure 4: Consumer and Cropped Image side by side

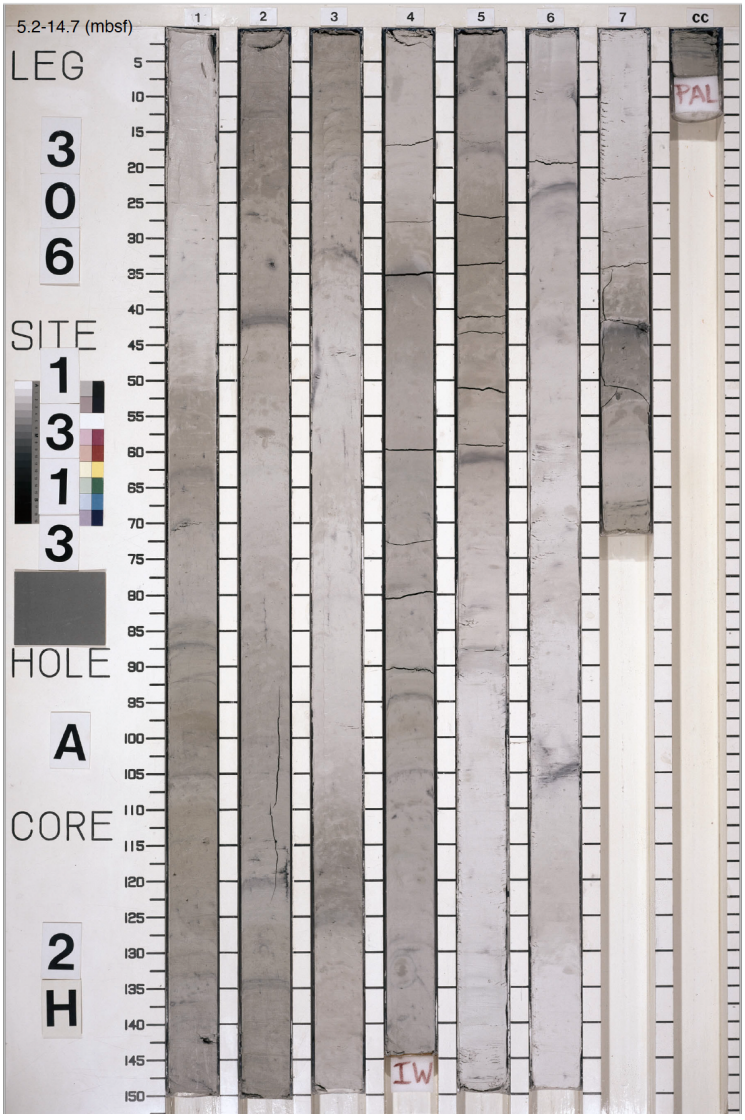

Fig. 3: IODP Expedition 306-U1313A-02 Core Photo. Source: IODP Expedition 306 Proceedings

Easily seen with the human eye, there are variations of color in the extracted sediment. These changes in the colors are what the team sought to better understand and analyze for Expedition 306 Site U1313 and all of its sections recovered in the North Atlantic Sea. From the consumer image in Fig. 4 above, a ruler on the left-hand side of the core provides measures of depth in centimeters of sediments downhole which is of particular use in scaling the extracted three-

dimensional array of color pixels for each image. Cores and their sections are organized sequentially in the order they were extracted from a hole and their placement within the core, respectively.

The scale of age with respect to depth was produced by the scientific participants and independent features like the δ18 Oxygen are used as age proxies that help to correlate global patterns of marine sediments and geochemistry to the right geologic age. The use of proxies help many scientists to understand the complex history of the Earth by analyzing the physical and chemical world with well-known attributes like the δ18 Oxygen and SST trends, and the use of color variations of sediments aim to astronomically tune the geologic time preserved at Site U1313.

Fig. 4: IODP Expedition 306-U1313A-01 Core Section Photo. Source: IODP Expedition 306 Proceedings

3.3 Data Engineering and Image Preprocessing

Five spectrophotometry color variations are extracted, transformed and loaded (ETL) into the dataset from the core images queried from the IODP database and are the focus of the modeling process due to their ability to act as real-time proxies of high-resolution geologic age at the hundreds of years scale.

All of the core data has an adjusted depth scale (amcd) in meters across the holes for all U1313 drilling Sites (A, B, C, and D), and in preparation to the statistical analysis and correlation of the IODP provided features with the image extracted color variations a necessary scaling of the image pixels was needed to align the top and bottom of core samples with the pixel values extracted.

Table 2: Extracted Core Image Features.

Color Feature	Pixel Value Range
Red	0-255
Blue	0-255
Green	0-255
Grey:	0-255
L*	0-100

The sediment and rock can be damaged and moved as cores are extracted from deep beneath the sea floor, and only unaltered sediments are of use in the correlation to the features aligned to the (amcd) depth, so the images are reduced to only one value down the center of the image represented by the mean pixel values aligned to the (amcd) depth provided by IODP in order to make comparison across all holes of Site U1313. Overall, the extracted color data is as follows in Below:

The depth corrected image data are extracted programmatically using modern computer vision processing techniques within an iterative loop to produce flattened vectors of section images contained within one matrix that can easily be trained with all other features all sequentially aligned with respect to the (amcd) depth. The spectrophotometry data extracted can be represented as D = X,Y , where X = the 16,000 x 200 x 3 core images where each pixel records the values for Red, Green, and Blue, and Y = 1,903 x 5 matrix of filtered RGB, Grey, and L* mean color values extracted along the amcd depth scale.

With the dataset preprocessed and aligned with the globally independent age proxy of provided benthic $\delta 18$ Oxygen isotope variations, the use of color variations derived from IODP image data can be used to build predictive models that correlate with well-known and accepted age proxies.

3.4 Image Extraction Automation

The first step was to build a machine learning architecture that could automate the extraction, transformation and loading (ETL) of the core image data into a

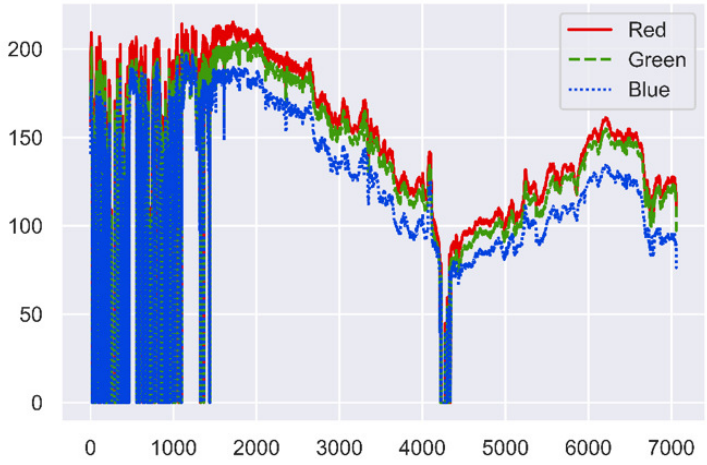

Fig. 5: Extracted Raw Color Variations

scaled dataset for predictive modeling of globally independent variables like the benthic δ18 Oxygen isotope variations produce by δ^{18}O.

Next the use of modern computer vision modeling with OpenCV was used to load the images into Python, where we built functions to extract color variations from each of these cropped images with an iterative loop that loads and transforms each image into the desired RGB, Grey, and LAB (L*a*b*) outputs. The function is extracting only the central portions of the images defined by the middle 200-pixel columns of each image and down every 16,000 rows resulting in a matrix of shape 16,000 x 200 x 3 that are ready to preprocess and align with the δ^{18}O age proxy.

The image matrices were much higher resolution than the δ^{18}O proxy data with approximately 16,000 rows in comparison to the δ^{18}O dataset with 1,903 rows. The preprocessing of the pixels was done with a function that organized each image sequentially down hole, used OpenCV to convert and extract the red, green, blue, grayscale, and L* color spectrum from each pixel, cropped the image to the middle 200 pixels that were then sorted by L* and 35% of darkest and 15% of lightest (as measured against the 0 to 100) L* value scale were removed to account for cracks, reflections, distortions, and other related anomalies in the sample, then the remaining pixels (approximately 100) were averaged (across each row) to establish a single-color spectrum of data for each extracted color given along the depth of each image 16,000 pixels for every 150 cm of sediment imaged. The extracted color variations for each image are saved as a Python matrix variable with a shape of 16,000 x 5, Fig. 6 shows a plot of the data from the resulting 16,000 x 5 matrix.

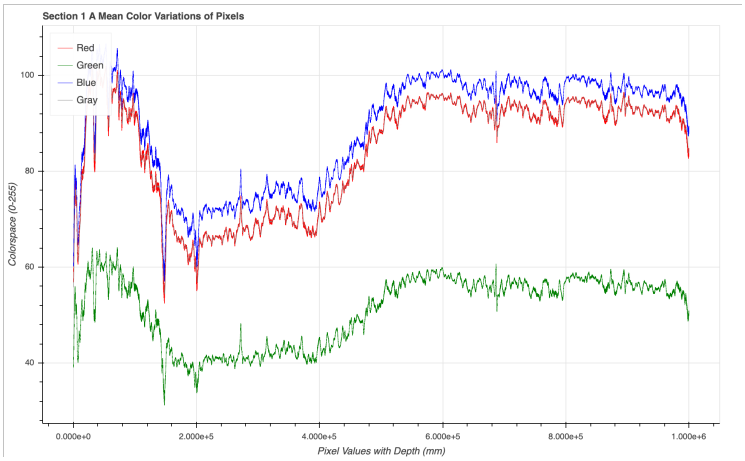

Fig. 6: The Extracted Color Variations of U1313A-01H-1

Next, we take the extracted color variations and scale it to the 1,903 age values provided by the benthic δ18 Oxygen isotope variations produce by (δ^{18}O). This is done with aligning the amcd (adjusted depths) of the extracted color variations with the age values of δ^{18}O proxy data provided by IODP as seen Fig. 7 focusing on expedition 306 Site U1313, Hole A, Cores 1-20 (equaling 137 different sections) photos representing the sediments recovered 200 mbsf.

The final dataset consists of normalized data, where we subtract the mean and divide by the standard deviation for every feature except age (in order to account for difference in measurement scaling of each feature), uniform amcd depths across all data (extracted and known) in order to align on the sequential age ranging from (0 – 4,260 ka). This final dataset is cleaned and uniformly spaced by geologic age and ready to model for linear regressive prediction purposes.

3.5 Exploratory Data Analysis

The first step in experimenting was to the linear relationships between all features provided by IODP, and most importantly the exploring of L* and δ^{18}O results for correlation. Fig. 8 is an example of the exploratory data analysis (EDA) used to visualize statistical relationships between features.

The EDA clearly show the strength of these linear relationships between the variables for use as geologic age proxies. To better measure how each feature is related to one another the team explored the influence of δ^{18}O variations on L* and SST through time, and a visualization gives valuable insight into the responses of SST and L* as the δ^{18}O values change through time as seen illustrated in Fig. 9.

The covariance table below in Fig. 10 clearly shows that the color spectrums have high degrees of inter-relatedness which is expected. Now that the dataset

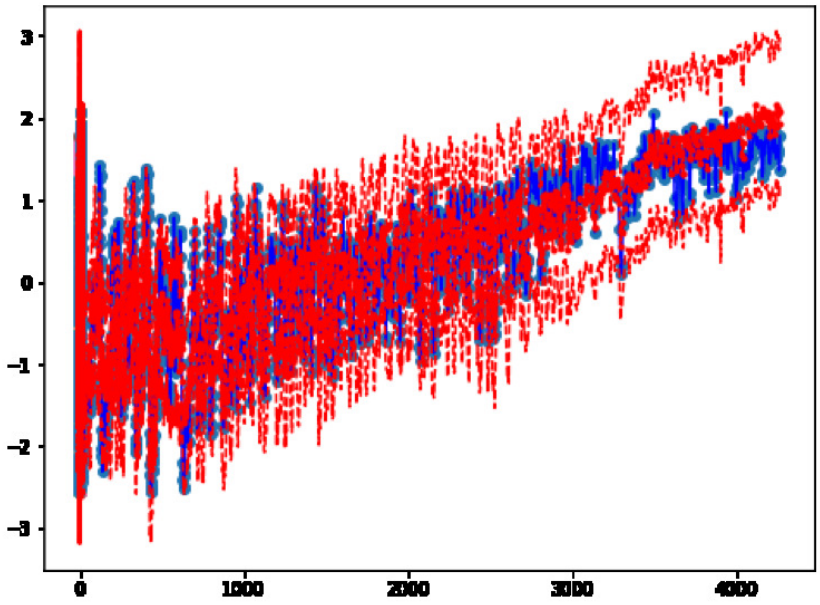

Fig. 7: The Scaled Color Variations of U1313A-01H-1

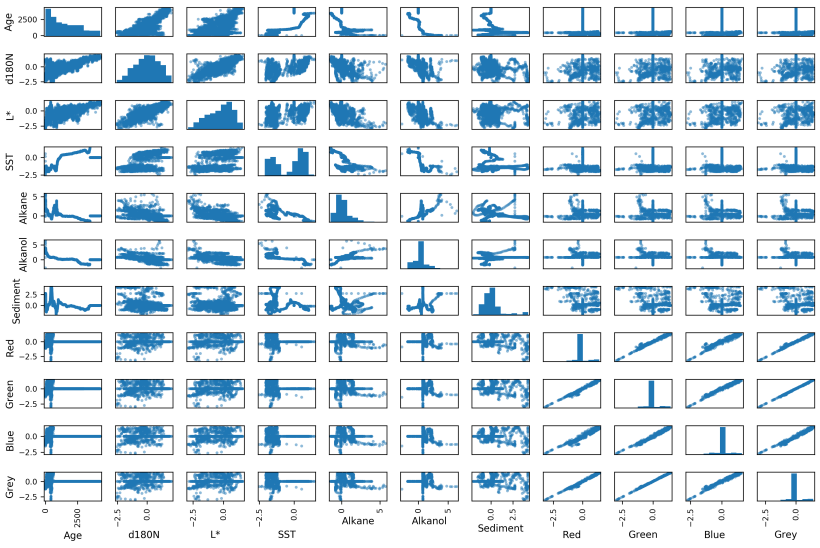

Fig. 8: Scatter Matrix of features

is preprocessed and the feature relationships are explored we are ready to run many regressor models that predict δ^{18}O values from various combinations of

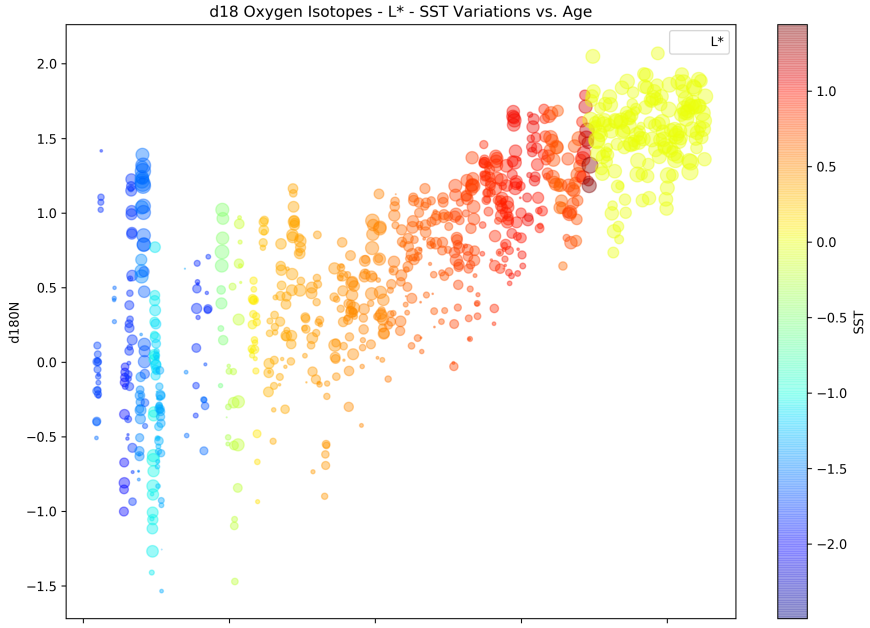

Fig. 9: δ^{18}O Variations with L* and SST Rates vs. Age

features with k-fold cross validation techniques to ensure the best fit models will be generalizable to the other holes of U1313 and also to other sites and expeditions of the past and the future.

4 IODP Modelling Developments

4.1 Automation of Image Data Acquisition

The team had automated the ETL process for IODP by first building a function that downloads the cropped images and saves them to a local directory from the online IODP database repository. The use of open source software like, the Selenium Firefox WebDriver module in Python was wrapped within a function that navigates the IODP database and selects the browser execution scripts needed to automatically click and download each image into the local machine directory. The image data automation technique created for IODP is highly desirable for future use in IODP research due to the automatic ETL of color variations into .CSV files that are used by the staff and researchers on a daily basis.

4.2 End-to-End Usability for IODP Research

The use of open source software like Python and TensorFlow allow any user to use these modeling techniques for research with IODP data whether it is

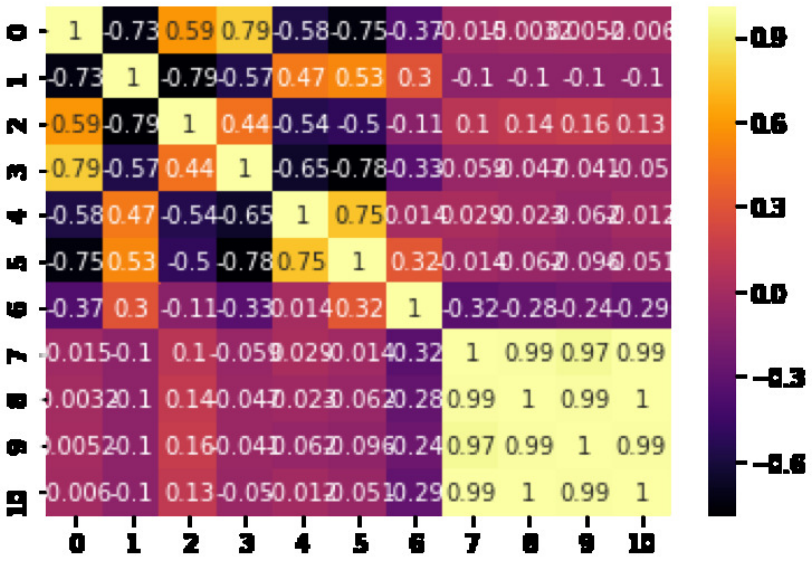

Fig. 10: Correlation Matrix of the Final Dataset

continuous or image data. The reproducibility of results is ensured and provided with heavily commented code and theoretical notes for models and algorithms providing all levels of the IODP staff with a functional application of machine learning and deep learning to learn from and implement into their current and future projects immediately.

4.3 Visualization of Machine Learning Results

There are numerous ways to visualize how the machine is learning training data. We produce multiple examples of visualizations that represent how the machine is learning along with comments and discussions on what insights can be gained from those visualizations from the per-epoch level all the way up to the more generalized fit.

4.4 Interpretability of Results

Interpretability of the models and their results were very important and thought of during model selection and throughout the building of deep neural networks. We chose to use the MAE metric due to the results having the same metric as the data it is evaluating, along with the easily interpreted and r2 metric that is heavily used throughout scientific research at IODP.

4.5 Generalizability

The best practice in modeling data to be generalizable is to use K-fold cross-validation. We implemented K-fold splitting on the data into K partitions, then instantiating K identical models, and training each one on K-1 partitions while evaluating on the remaining partition. The validation score for the model used would then be the average of the K validation scores obtained. We also give insight on how to handle small and large datasets when building models that will generalize out of the scope of the training data.

5 Results

5.1 Machine Learning Approach

The team first performed a basic linear regression on δ^{18}O and L* data at the guidance of the subject matter expert from IODP. The basic linear model describes the benthic oxygen isotope ratios δ^{18}O modeled for each of the variations in sediment lightness (L*) by only modeling the one independent feature against δ^{18}O. The basic linear model used is described by the following form:

$$y = mx + b \tag{1}$$

The linear relationship and line of best fit are plotted in Fig. 11 below, along with the accuracy measures for the basic linear model are shown in Table 3. The linear relationship between L* and δ^{18}O shown in Fig. 11 showed the promising use of the linear regressors predictability of δ^{18}O from color variations.

Table 3: Linear Regression Model Results for δ^{18}O Prediction.

Metric	Result
MAE =	0.47475
MSE =	0.42316
RMSE =	0.65051
r^2 =	0.57534

The predictability of the basic model is further illustrated by visualizing the residual pairs being plotted together side-by-side as shown in Fig. 12 below:

5.2 Hyperparameter Tuning

A common data science technique of hyperparameter tuning with grid searching cross validation (CV) is used to find the best generalized model fit controlled by under the hood decision functions that are user defined violation boundaries that must be kept. The use of the support vector machine (SVM) algorithm to

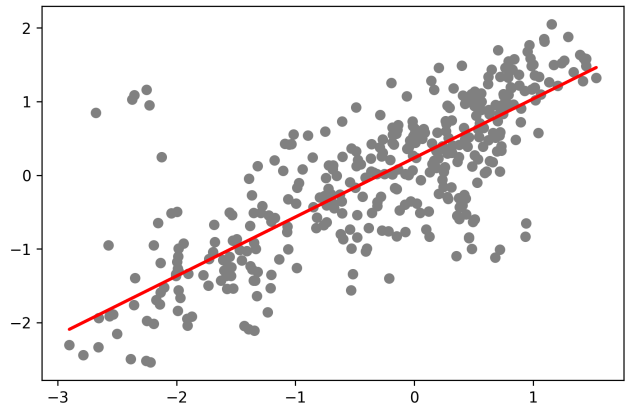

Fig. 11: Linear Regression Line of Best Fit Plot for δ^{18}O and L*

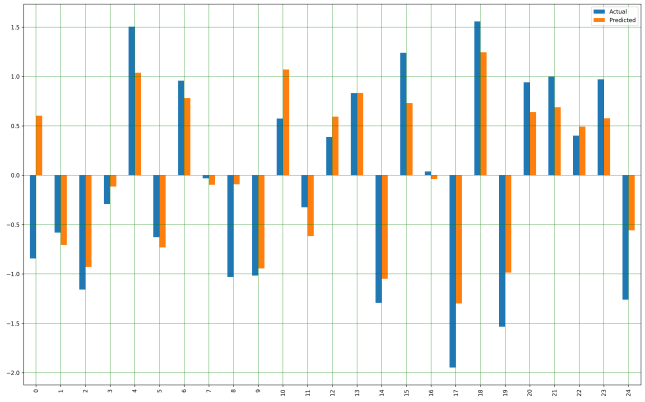

Fig. 12: Residuals Pair Plot for Actual and Predicted Values

grid search the best fitting CV model is used with a support vector machine regression (SVR) algorithm utilizing a grid of parameters and kernel tricks to find the best fitting regressor for predicting δ^{18}O values through time.

The SVR algorithm tries to fit as many instances as possible on the line of best fit while limiting the number of violations controlled by the hyperparameters epsilon, gamma, and a decision function C with kernel trick transformations that take the dot products of the weight vectors and project them into a higher dimension allowing for algorithm to minimize their distances from the best fitting line through the dependent feature space. SVM Regression Model we found to perform the best used a linear kernel, which was expected due to the nature of the data relationships.

The SVM Regression Model in Basic form:

$$y = w^T x + b \tag{2}$$

The SVR Model we used with a linear kernel is modeled as follows:

$$f(\mathbf{x}) = \phi(\mathbf{x})^T \mathbf{w} + b = \sum_{j=1}^{m} \alpha_j y_j K(\mathbf{x}, \mathbf{x}_j) + b$$

The parameter b can be found from any support vectors $\mathbf{x}_i$:

$$b = y_i - \phi(\mathbf{x}_i)^T \mathbf{w} = y_i - \sum_{j=1}^{m} \alpha_j y_j (\phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)) = y_i - \sum_{j=1}^{m} \alpha_j y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

SVM Regression: Utilizing a linear kernel.
Assume $\mathbf{x} = [x_1, \cdots, x_n]^T$, $\mathbf{z} = [z_1, \cdots, z_n]^T$,

$$K(\mathbf{x}, \mathbf{z}) = \mathbf{x}^T \mathbf{z} = \sum_{i=1}^{n} x_1 z_1$$

First the SVR model was manually tuned with out-of-box parameters provided with the Scikit-learn SVM module and then a grid search function was written in Python to find the best model hyperparameter combinations of kernel, C, and gamma values for the dataset. The out-of-box, and custom grid search CV models produced the following coefficient of determination (r^2) accuracy scores Tab. 4:

Table 4: SVR Regression Model Results for δ^{18}O Prediction.

Modeling Technique	Accuracy
Out-of-Box SVR r^2	0.78445
Hypertuned SVR r^2	0.89158
Hypertuned 4-fold CV SVR r^2	0.95071

These machine learning techniques used have increased the predictability of the δ^{18}O variations drastically with the use of more geoscience related features provided by IODP and also by the extracted color variations of sediment images from Site U1313. To further modernize predictive modeling techniques of continuous features at IODP, we set provide a step-by-step deep learning model of the same data that learns the δ^{18}O variations of Site U1313 which can be used to model other sites and predict oxygen isotope variates throughout the past present and future seagoing expeditions.

5.3 Deep Learning Approach

The deep learning architecture is based on the Keras deep neural network (DNN) API using TensorFlow 2.0 and popular Python modules like NumPy, Scikit-learn, Pandas, and Matplotlib to build, train, test and visualize our DNN.

5.4 Building our network

Because so few samples are available, we will be using a very small network built into a callable function with two hidden layers, each with 64 units as seen in Fig. 13. In general, the less training data you have, the worse overfitting will be, and using a small neural network is one way to mitigate overfitting.

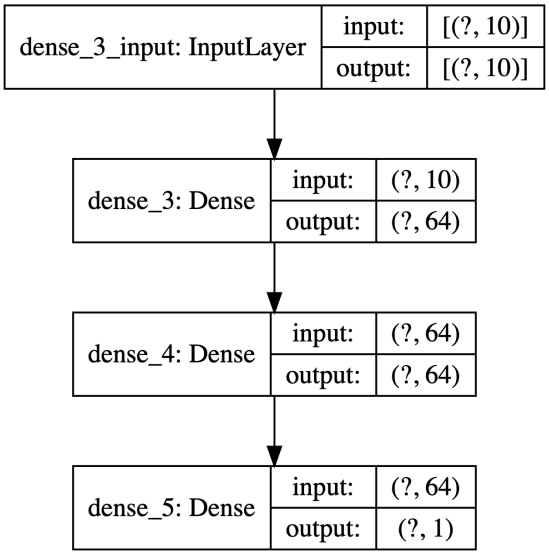

Fig. 13: DNN Regressor Model Architecture

Our network ends with a single vector of continuous output, and no activation (i.e. it will be linear layer). This is a typical setup for scalar regression (i.e. regression where we are trying to predict a single continuous value). Applying an activation function would constrain the range that the output can take; for instance, if we applied a 'sigmoid' activation function to our last layer, the network could only learn to predict values between 0 and 1. Here, because the last layer is purely linear, the network is free to learn to predict values in any range.

Note that we are compiling the network with the MSE loss function – Mean Squared Error, the square of the difference between the predictions and the targets, a widely used loss function for regression problems.

We are also monitoring a new metric during training: MAE. This stands for Mean Absolute Error. It is simply the absolute value of the difference between the predictions and the targets. For instance, a MAE of 0.5 on this problem would mean that our predictions are off by 0.5 (ng/g) on average.

5.5 Validating our approach using K-fold validation

To evaluate our network while we keep adjusting its parameters (such as the number of epochs used for training), we could simply split the data into a training set and a validation set, as we were doing in our previous examples. However, because we have so few data points, the validation set would end up being very small (e.g. about 100 examples). A consequence is that our validation scores may change a lot depending on which data points we choose to use for validation and which we choose for training, i.e. the validation scores may have a high variance with regard to the validation split. This would prevent us from reliably evaluating our model.

The best practice in such situations is to use K-fold cross-validation. It consists of splitting the available data into K partitions (typically K=4 or 5), then instantiating K identical models, and training each one on K-1 partitions while evaluating on the remaining partition. The validation score for the model used would then be the average of the K validation scores obtained. The first experiment MAE results for K=4 folds are plotted in Fig. 14.

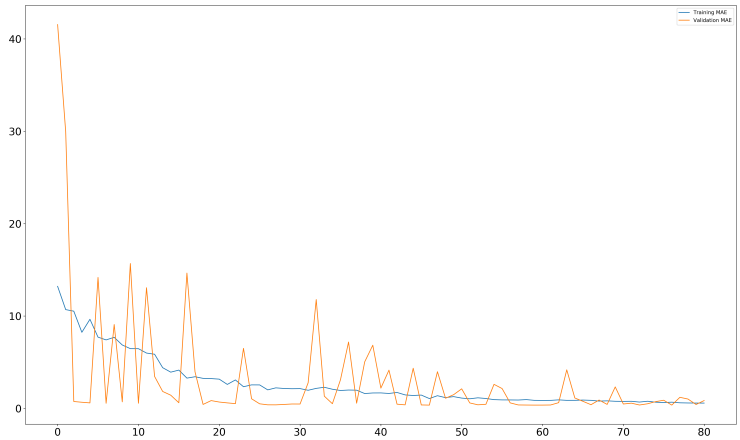

Fig. 14: 4-Fold Cross Validation MAE Results for 100 Epochs of Training

As you can notice, the different runs do indeed show rather different validation scores, ranging from 0.40 to 0.97. Their average (0.72) is a much more reliable metric than any single of these scores – that's the entire point of K-fold cross-validation. In this case, we are off by 0.719 x 0.498 (the product of the

MAE and standard deviation of δ^{18}O for the dataset) = 0.358 (ng/g) δ^{18}O on average, which is still significant considering that the δ^{18}O ratios range from 2.77 to 5.08. Let's try training the network for a bit longer: 500 epochs. To keep a record of how well the model did at each epoch, we modify our training loop to save the per-epoch validation score log, which allows us to compute the average of the per-epoch MAE scores for all folds and visualize the learning. Figure 14a shows the per-epoch learning over 500 epochs and Fig. 15 is a smoothed representation that filters out variation and allows the user to visualize the learning rate per-epoch and also when there may be points of overfitting.

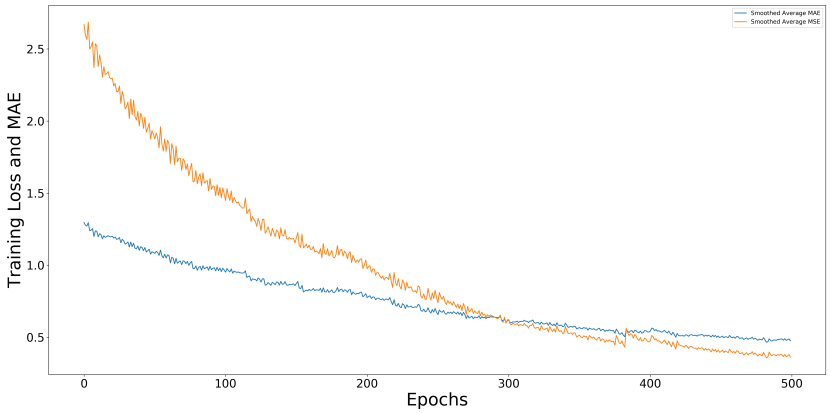

Fig. 15: Trained and Validated MAE and Smoothed MAE for 500 Epochs

6 Result Summary and Analysis

6.1 Data Engineering

The data science workflow and image preprocessing techniques used with the automated ETL of spectrophotometry data implementations are great tools for use by any scientist or staff at IODP. Our modeling approach has practical value for IODP researchers today and the potential use for any researcher who wishes to use IODP's open source data.

6.2 Basic Linear Regression

The basic linear model is unable to generalize well on the test dataset and the need for models that are capable of predicting globally independent features like the δ18 variations throughout geologic time within a region is highly sought after by many climate modeling scientists. We highly recommend the use of machine learning and deep learning to search for the best fitting line to make CV predictions and even regional forecasts.

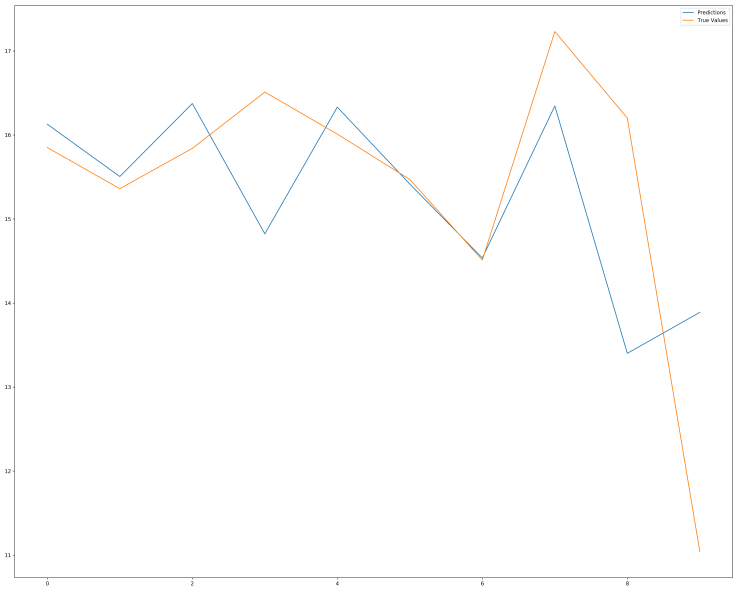

Fig. 16: Linear Regression Predicted and True Variations of δ^{18}O

6.3 Machine Learning Regression

Our machine learning model offers improvements in workflow with hypertuning parameter optimization along with automated ETL of image data to .CSV files for easy implementation in Python, and stratified K-fold CV modeling that increases learning performance and generalizes the models prediction ability.

The SVR model interpretability may be cumbersome due to the sophisticated statistics and linear transformations however, the results are in the same units as the independent feature being predicted which increases user interpretability of the models ability to accurately predict values of δ18 from features like L* and color variations from sediment images. The concern with overfitting is minimized with the stratified K-fold CV.

6.4 Deep Learning Regression

Our deep learning model offers improvements in workflow with hypertuning parameter optimization along with automated ETL of image data to .CSV files for easy implementation in Python, and stratified K-fold CV modeling that increases learning performance and generalizes the models prediction ability outside the scope of the training data.

The DNN regression model interpretability may seem more daunting and be just as cumbersome as the SVM model however, the results are in the same units as the independent feature being predicted which increases user interpretability of the models ability to accurately predict values of δ18 from features like L* and

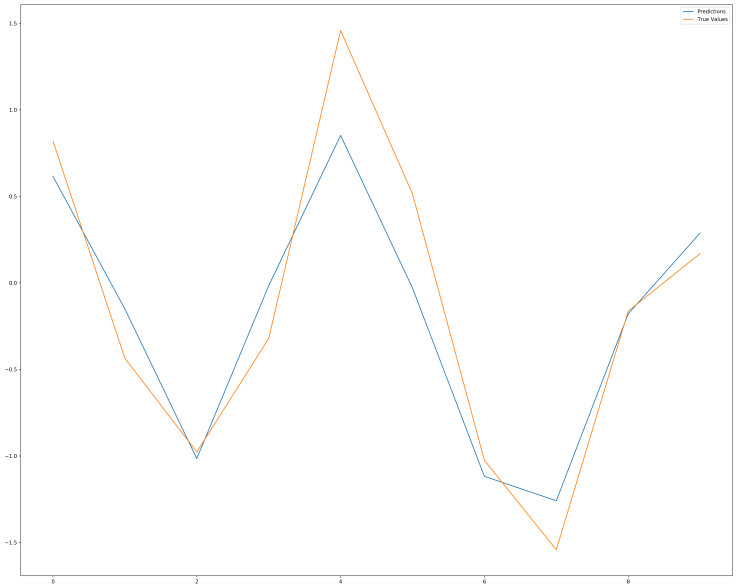

Fig. 17: SVM Regression Predicted and True Variations of δ^{18}O

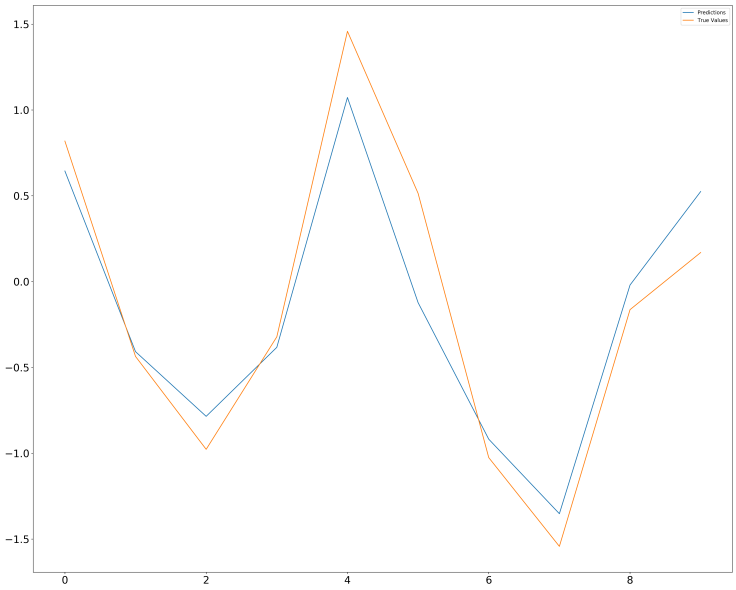

Fig. 18: Predicted and True Variations of δ^{18}O

color variations from sediment images. The concern with overfitting is minimized with the stratified K-fold CV.

Given the generalizable MAE results like those in Fig. 18, for the prediction of a globally independent geologic age proxy by both machine learning and deep learning models, the use of deep learning models should be preferred in cases when the models are used for continuous learning with . The reason for this recommendation is that for applications where the highest possible accuracy is not required but learning and preserving what has been learned is a higher priority, the DNN regression models can be used as pretrained layers for any model that wishes to use its weight vectors for training and predicting new features that are related to those trained in our model.

7 Ethics

The ethical considerations of applying machine learning to geoscience research should not be taken lightly. Millions of dollars and countless human and physical resources are expended to advance the objectives of the International Ocean Discovery Program. Many researchers have built their careers based on IODP work and many other professionals have launched theirs.

It's not uncommon for workers to be displaced when advance technology becomes available. While it's commonly thought that academics and "knowledge workers" are less susceptible to job displacement, that's not always true.

Utilizing tools (like our machine learning models) that save time and money does in fact reduce the demand for certain kinds of workers (junior researchers) and it is important to understand how applied machine learning will impact those workers going forward. One line of thinking will see this reduced demand as a negative. That line of thinking is incorrect. In fact, reducing the demand placed on junior researchers to undertake time intensive explorations and calculations enables them up to pursue harder problems that do not yet have clear answers. While some people fear what might happen as more and more machine learning models get deployed in the physical and chemical sciences, it is actually unethical not to pursue the use of these tools. The principal researchers who direct the budgets and the work of the IODP have a moral obligation to get the most scientific return on the research dollars invested and, presently, applied machine learning models based on open source tools, as we have demonstrated, enable researchers to do more work faster with comparable results.

This approach to research contributes to all scientific endeavors focused on better understanding the Earth. The faster applied machine learning is adopted into the sciences, the smaller the negative impact will be on the junior researchers who would be most vulnerable to having the demand for their services reduced. In fact, a junior researcher skilled in applied machine learning offers a tremendous value proposition to organizations and institutions.

8 Conclusions

The overall conclusions reached are that the data can be extracted from the photographs using certain basic assumptions around what pixels are valid for analysis and what pixels represent 'damaged' sediment and rock. From these extracted data, the conclusions are that color spectrums can be used to establish a correlation to δ^{18}O (oxygen levels in the ocean), with a modeling architecture that produces the following results with the ability to become even more accurate and precise with more data and training. Our DNN regression model predicted the δ^{18}O variations with a MAE: 0.26824036,RMSE: 0.13881794101904824, and r^2: 0.8537100907070401. It is important to conclude that δ^{18}O variations are well known proxies capable of inferring many oceanic features, and with the use of the modern data science modeling techniques discussed and provided in our capstone research many scientific researchers could employ the same methods to model their data and help many others astronomically tune the geological records influenced by global and region climatic temperatures.

It is important to note that in either case, using our machine learning model or deep learning model, an inexperienced researcher can obtain results comparable to those obtained by researchers with far more expertise following the step-by-step approach provided to IODP. Further, more experienced researchers using our models will be able to obtain new features that may generalize well with their data and increase their predictable results much faster. The Generalizability of the DNN regressor trained is able to extrapolate quite well with a small MAE = 0.27 on the roughly 100 test predictions Fig. 19 on the next page visulizes the models ability to accurately replicate the δ^{18}O variations down the core.

Machine learning and deep learning modeling of the core data with images are a unique use case of creating regressive models for IODP data. In the field of applied machine learning and deep learning, there are many more sophisticated models being built and trained daily with very similar data a objectives and we have high hopes that after reading our research findings the fields of data science and geoscience converge closer together with more open minded and competitive researchers who are inspired to explore new modeling techniques with IODP data.

9 Future Work

The models contained in this paper were built and trained using a limited number of sample images. As additional core sample images are gathered and analyzed, retraining the saved DNN model with the goal of increasing performance in terms of generalizable accuracy in predicting climate related features being discovered and measured by IODP researchers.

Additionally, as this paper sets out, the model we built is one geared towards extracting features in multi-dimensional images and then analyzing those features to establish new independent global proxies of age for geological purposes. As such, the possibility for future work is so broad that a complete list

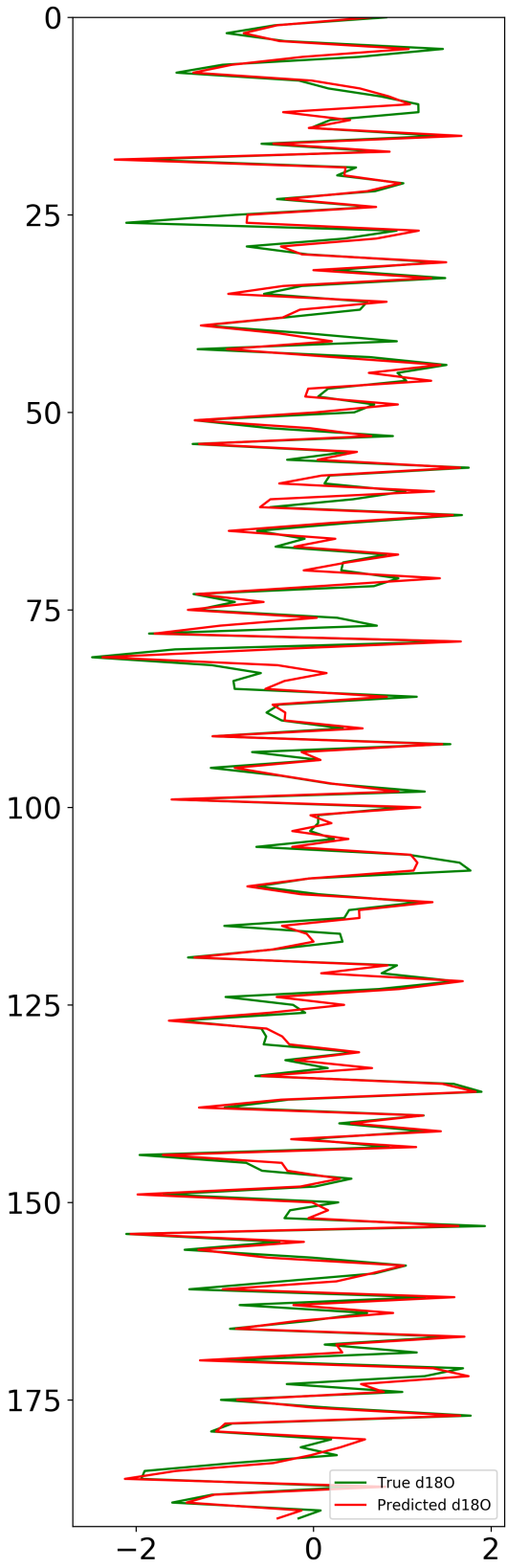

Fig. 19: DNN Predictions for δ^{18}O

of applications is outside of the scope of this paper, and we encourage the use of these techniques along with others related to explore new methods for global age proxies.

Taking the team's approach as an example, there will be multiple opportunities to continue deploying machine learning and deep learning applications for other issues in geoscience including models that perform facies classification and stratigraphic correlation from core images between the drilling sites of the samples; predicting many other characteristics of the oceans geochemistry from the color variations of the sediment core; and many other related climate or geoscience questions of interest.

10 Acknowledgements

The authors thank Dr. Gary D. Acton for supporting this study with his vast knowledge and experience in the modeling of geoscience data and the acquisition and interpretation of scientific data collected and published by IODP research, Dr. Daniel W. Engels for fruitful comments and inspiration during our capstone research project.

References

1. Channell, J.E.T., K.T.S.T.S.R.A.Z.C.M.M., the Expedition 303/306 Scientists: Proc. iodp, 303/306. Integrated Ocean Drilling Program Management International, Inc. (2006), doi:10.2204/iodp.proc.303306.2006
2. F.J. Sierro, I. Hernandez-Almeida, M.A.G., Flores, J.: Data report: Pliocenec-pleistocene planktonic foraminifer bioevents at iodp site u1313 (11 2008), doi:10.2204/iodp.proc.303306.205.2008
3. Hagino, K., Kulhanek, D.: Data report: calcareous nannofossils from upper pliocene and pleistocene, expedition 306 sites u1313 and u1314 1 (2009), 10.2204/iodp.proc.303306.206.2009
4. Kanamatsu, T., S.R., Alvarez Zarikian, C.: North atlantic climate ii addendum. IODP Sci. Prosp., 306 Add. (2005), doi:10.2204/iodp.sp.303306add.2005
5. Lamy, F., W.G., Alvarez Zarikian, C.: Expedition 383 scientific prospectus: Dynamics of the pacific antarctic circumpolar current (dynapacc). International Ocean Discovery Program (1986)
6. Naafs, B.D.A., H.J.A.G.H.G.M.G.A.P.R., Stein, R.: Strengthening of north american dust sources during the late pliocene (2.7 ma). Earth Planet. Sci. Lett. pp. 317–318:8–19 (2012), doi:10.1016/j.epsl.2011.11.026
7. Raymo: Timing of the descent into the last ice age deter- mined by the bipolar seesaw. Paleoceanography and Paleoclimatology (29), 489–507 (1989)
8. Scientists, E.: North atlantic climate 2. IODP Prel. Rept., 306 (2005), doi:10.2204/iodp.pr.306.2005
9. Stein, R., K.T.A.Z.C.H.S.C.J.A.E.O.M.A.G.: North atlantic paleoceanography: the last 5 million years. Eos, Trans. Am. Geophys. Union (87), 129 (2006), doi:10.1029/2006EO130002

10. Voelker, A.H.L., R.T.B.K.O.D.M.J.S.R.H.J., Grimalt, J.: Variations in mid-latitude north atlantic surface water properties during the mid-brunhes (mis 9–14) and their implications for the thermohaline circulation. Clim. Past (6), 531–552 (2010), doi:10.5194/cp-6-531-2010