Diffuse Attenuation Coefficient (K_d) from *ICESat-2* ATLAS Spaceborne Lidar Using Random-Forest Regression

Forrest Corcoran and Christopher E. Parrish

Abstract

This study investigates a new method for measuring water turbidity—specifically, the diffuse attenuation coefficient of downwelling irradiance K_d—using data from a spaceborne, green-wavelength lidar aboard the National Aeronautics and Space Administration's ICESat-2 *satellite. The method enables us to fill nearshore data voids in existing K_d data sets and provides a more direct measurement approach than methods based on passive multispectral satellite imagery. Furthermore, in contrast to other lidar-based methods, it does not rely on extensive signal processing or the availability of the system impulse response function, and it is designed to be applied globally rather than at a specific geographic location. The model was tested using K_d measurements from the National Oceanic and Atmospheric Administration's Visible Infrared Imaging Radiometer Suite sensor at 94 coastal sites spanning the globe, with K_d values ranging from 0.05 to 3.6 m^{-1}. The results demonstrate the efficacy of the approach and serve as a benchmark for future machine-learning regression studies of turbidity using* ICESat-2.

Introduction

Measurement and long-term monitoring of water clarity is an important undertaking in oceanography, marine eco-forecasting, pollution and runoff modeling, and coral-reef ecosystem health assessment (National Academies of Sciences, Engineering, and Medicine 2018). Turbidity, which refers to the capacity of a body of water to attenuate light, has been used across numerous disciplines, including in classifying water types (Jerlov 1976; Sarangi *et al.* 2002), determining the vertical distribution of algae species (Saulquin *et al.* 2013), protecting submersed aquatic vegetation and coastal estuaries (Gallegos 2001; Doxaran *et al.* 2006), and modeling colored dissolved organic matter in shallow estuaries (Branco and Kremer 2005). Spatially and temporally varying measurements of turbidity are also frequently used in airborne bathymetric lidar project planning (Richter *et al.* 2017; Saylam *et al.* 2017; Forfinski-Sarkozi and Parrish 2019). One of the most common metrics used to quantify turbidity is the diffuse attenuation coefficient of downwelling irradiance K_d, an apparent optical property (AOP) defined by Equation 1 (Mobley *et al.* 2020) and typically specified in units of m^{-1}:

$$K_d(\lambda) = -\frac{1}{E(\lambda)}\frac{\partial E(\lambda)}{\partial z} \tag{1}$$

where z is depth, E is downwelling irradiance, and λ is wavelength.

It is important to recognize that K_d, an AOP, is different from the beam attenuation coefficient, defined as $c(\lambda) = a(\lambda) + b(\lambda)$, where a is the absorption coefficient and b is the scattering coefficient. The beam attenuation coefficient is an inherent optical property, whereas K_d is an AOP, with the distinction between the two being that AOPs depend on both the medium (i.e., the inherent optical properties) and the light field in which they are measured. According to Guenther (2007), for coastal waters K_d is generally smaller than c by a factor of 2 to 6 for green light.

Although the validity of the use of K_d in the Beer–Lambert law (Equation 2, which is a particular solution of the differential equation in Equation 1) has been the subject of discussion in the literature (Gordon, 1989), the Beer–Lambert law is generally assumed to hold for most water types, providing estimates of E as a function of depth:

$$E(z) = E_0^{-K_d z} \tag{2}$$

Traditionally, K_d has been obtained from in situ techniques such as Secchi depth measurements (Guenther 1985; Z. Lee *et al.*, 2015) and submarine photometry (Koenings and Edmundson 1991); however, advances in satellite imaging and the availability of remotely sensed data have allowed for daily, near-global measurements of K_d (Z.-P. Lee *et al.* 2005). Currently, data from the European Space Agency's Medium Resolution Imaging Spectrometer, the National Aeronautics and Space Administration's (NASA's) Moderate Resolution Imaging Spectroradiometer, and the National Oceanic and Atmospheric Administration's Visible Infrared Imaging Radiometer Suite (VIIRS) are used to generate K_{d490} (K_d at a wavelength of 490 nm) maps of the Earth's oceans (M. Wang *et al.* 2017). The VIIRS instrument aboard the Suomi National Polar-orbiting Partnership and *Joint Polar Satellite System-1* and -2 is a passive radiometer used to detect visible and infrared electromagnetic spectra with the objective of measuring global ocean color (M. Wang *et al.* 2017).

While this category of passive remote-sensing techniques provides an effective method for measuring K_d over large spatial extents at daily intervals, it relies strictly on observations of water-leaving irradiance and does not directly measure the absorption and attenuation of light at depth. In this study, we propose an active remote-sensing method for measuring K_{d532} (K_d at a wavelength of 532 nm) using NASA's Advanced Topographic Laser Altimeter System (ATLAS) aboard the *Ice, Cloud and Land Elevation Satellite-2* (*ICESat-2*). A key goal is to produce output

Forrest Corcoran and Christopher E. Parrish are with the School of Civil and Construction Engineering, Oregon State University, Corvallis (corcoraf@oregonstate.edu).

Contributed by Qunming Wang, March 25, 2021 (sent for review May 4, 2021; reviewed by Yue Ma, Yanli Zhang, Yao Li).

Photogrammetric Engineering & Remote Sensing
Vol. 87, No. 11, November 2021, pp. 831–840.
0099-1112/21/831–840

doi: 10.14358/PERS.21-00013R2

that is compatible with the imagery-based K_{d490} data sets and that can be used to fill in the data gaps and sparse areas that often exist in the imagery-based K_d products near shorelines with high-resolution, active-sensing data. Additionally, the ability of the ATLAS lidar to penetrate the water column (Jasinski *et al.* 2016) allows for more direct measurement of turbidity at depth. Despite the uncertainty inherent in the VIIRS K_d data, we consider VIIRS to be a viable source of reference data in this study, due to the fact that VIIRS K_{d490} has been well characterized in the literature (Z.-P. Lee *et al.* 2005; M. Wang *et al.* 2017) and is already being used for a number of science objectives (Qi *et al.* 2015; Shi and Wang 2015; M. Wang and Wilson 2017; Liu *et al.* 2017; van Hooidonk 2020).

The ATLAS instrument, a 10-kHz photon-counting lidar system operating at a wavelength of 532 nm, is the sole instrument aboard the *ICESat-2* satellite. At this wavelength, ATLAS is able to penetrate bodies of water up to a depth of approximately 40 m in areas of low turbidity (Parrish *et al.* 2019). It measures the time of flight of discrete photons reflected by the Earth and the Earth's atmosphere. A diffractive optical element within the ATLAS system splits each laser pulse into six beams, grouped into three pairs and oriented roughly perpendicular to the satellite flight direction. The beam pairs are separated by approximately 3.3 km across the track, with each pair made up of a strong and a weak beam. The strong and weak beams have an energy ratio of approximately 4:1 and are separated by 90 m in the across-track direction and approximately 2.5 km in the along-track direction. Figure 1 shows the footprint pattern of the ATLAS beams (Neumann *et al.*, 2020a).

Several studies have already demonstrated the ability to extract K_d measurements from spaceborne lidar systems (Lu *et al.* 2014, 2019, 2020). However, these studies focus on specific sites of limited spatial extent, making it difficult to generalize their findings to a global scale. Additionally, they rely on deconvolving the received lidar signal using an estimate of the system impulse response function. The ATL13 inland water product also includes a subsurface attenuation coefficient, defined as the sum of the absorption and scattering coefficients and computed as described by Jasinski *et al.* (2020). In contrast, the techniques used in the present study require minimal signal preprocessing and instead favor an ensemble machine-learning approach to derive K_{d532} from patterns in the shapes of ATLAS pseudo-waveforms (vertical histograms representing the number of photons within discrete elevation intervals, used to approximate a full waveform response from the photon-counting point cloud). This technique is advantageous because it does not require knowledge of the system impulse response, deconvolution of the pseudo-waveform, or any curve fitting. Instead, it requires only the computation of a few simple statistical features, which the trained model uses to make predictions. To demonstrate the validity of this technique, we extracted pseudo-waveforms from 543 ground tracks, collected from 94 sites across the world, and performed a random-forest regression between the ATLAS pseudo-waveforms and VIIRS K_{d532} measurements (derived from VIIRS K_{d490}) observed at the same approximate locations and times. The R^2 of the regression was 0.67 ± 0.12, with a mean squared error of 0.34 ± 0.14 m^{-2}, a mean absolute error of 0.21 ± 0.4 m^{-1}, and a mean relative difference of 1.07 ± 0.25, over the range of 0.05 to 3.6 m^{-1}, indicating that ATLAS pseudo-waveforms can be used to complement VIIRS K_d data and fill in data voids, especially in nearshore regions.

This study is the first to report accuracy metrics, aside from the mean relative difference (Lu *et al.* 2014, 2019, 2020), for K_d retrieval from a spaceborne lidar system. Therefore, these metrics stand as a benchmark for future studies. Importantly, our assessment of the model's accuracy does not rely solely on a single training–test split but rather on the average score of a randomized cross-validation approach, making our evaluation robust to any biases introduced by the training–test split and the inherent randomness associated with the convergence of random-forest regression. Figure 2

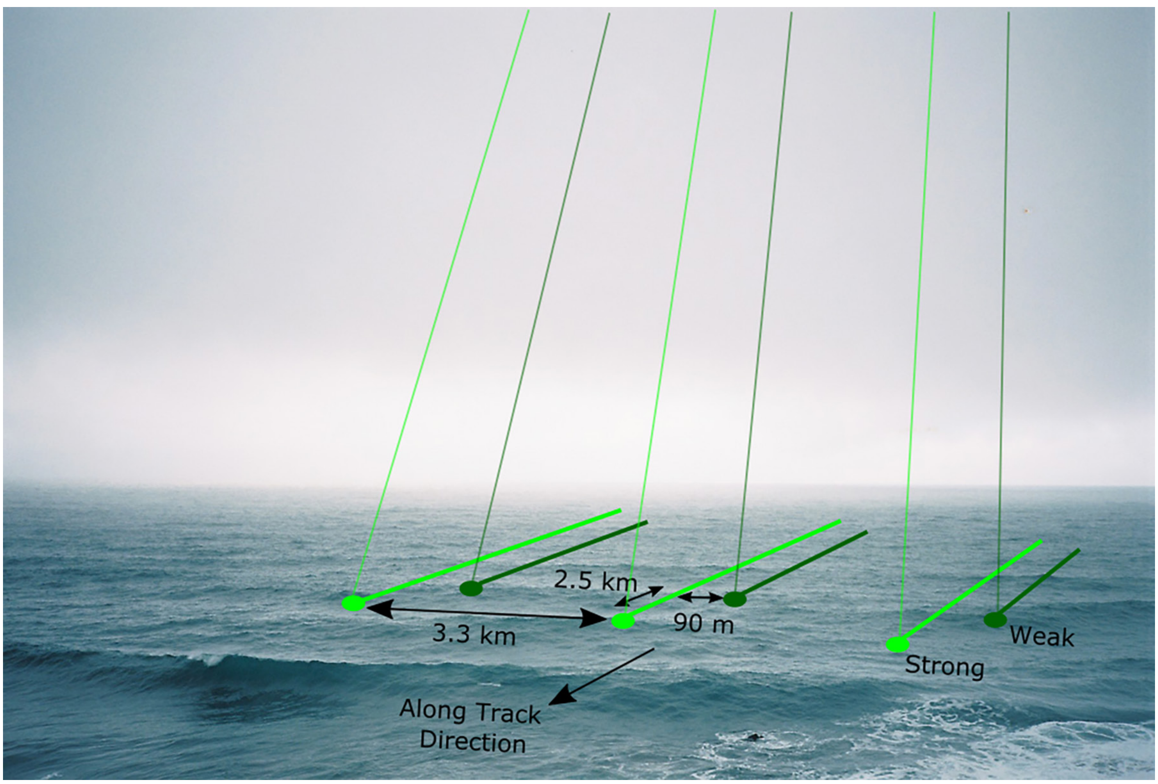

Figure 1. Schematic of ground-track pattern made by ICESat-2's Advanced Topographic Laser Altimeter System lidar. Beams are seperated into three pairs of strong and weak. The pairs are separated by 3.3 km, and within each pair, the strong and weak beams are separated by 2.5 km in the along-track direction and 90 m in the across-track direction.

shows an outline of the general workflow used to develop the random-forest regression model for K_{d532}.

Methods

Bathymetric or topobathymetric lidar has long been used for hydrographic surveys in both inland and nearshore coastal waters (Muirhead and Cracknell 1986). The majority of this research has been conducted using airborne, full-waveform lidar systems which calculate depths (or seafloor elevations, relative to a defined vertical datum) from digitized return waveforms (Walker *et al.* 1999; Klemas 2011; J. H. Lee *et al.* 2013; Rogers *et al.* 2015, 2016; C. Wang *et al.* 2015; Richter *et al.* 2017; Saylam *et al.* 2017). These waveforms typically display two peaks—an upper peak corresponding to the water surface and a lower peak corresponding to the bathymetric bottom—along with an exponentially decaying signal contribution between the two peaks, typically referred to as "volume backscatter" and corresponding to returns from water-column constituents (Guenther 2004). The intensity of the waveform diminishes between these peaks as a result of the water turbidity and can be modeled with an exponential decay function. The decay coefficient of this exponential function represents K_d, as can be readily seen from Equation 2.

By contrast, *ICESat-2*'s ATLAS instrument is a photon-counting lidar system that measures the flight time of discrete photons. As a result, ATLAS does not generate full waveforms but instead produces two-dimensional photon-cloud profiles. In this work, we generated "pseudo-waveforms" that serve the same purpose as waveforms (i.e., to indicate the "energy" or strength of return in vertically binned depth ranges). This was done using a moving window with an along-track length of 20 m and a height of 1 dm. At each decimeter depth interval, we counted the number of photons within the 20-m along-track distance. The count at each interval is proportional to the amplitude of the pseudo-waveform at that depth. By stacking these discrete depth-interval bins vertically, we created pseudo-waveforms from each of the point-cloud profiles in our data set. Finally, we cropped these pseudo-waveforms between −10 and 10 m (an empirically determined range) along the vertical axis to standardize the boundaries of the waveform above and below the water surface. This was a necessary step because it removed unwanted peaks in the waveform above the water surface corresponding to atmospheric phenomena, as well as unwanted peaks below the water surface due to instrument effects or signal noise. All the data used in the study were far enough offshore that bathymetric returns were not considered a factor within the elevation window from −10 to 10 m. Figure 3 shows examples of pseudo-waveforms extracted from the ATLAS data set used in this study.

An important question that arises with respect to Figure 3 is: Why not just directly estimate K_d by fitting an exponential decay curve to the pseudo-waveform? While this is, in theory, possible (at least if field-of-view loss and other system variables

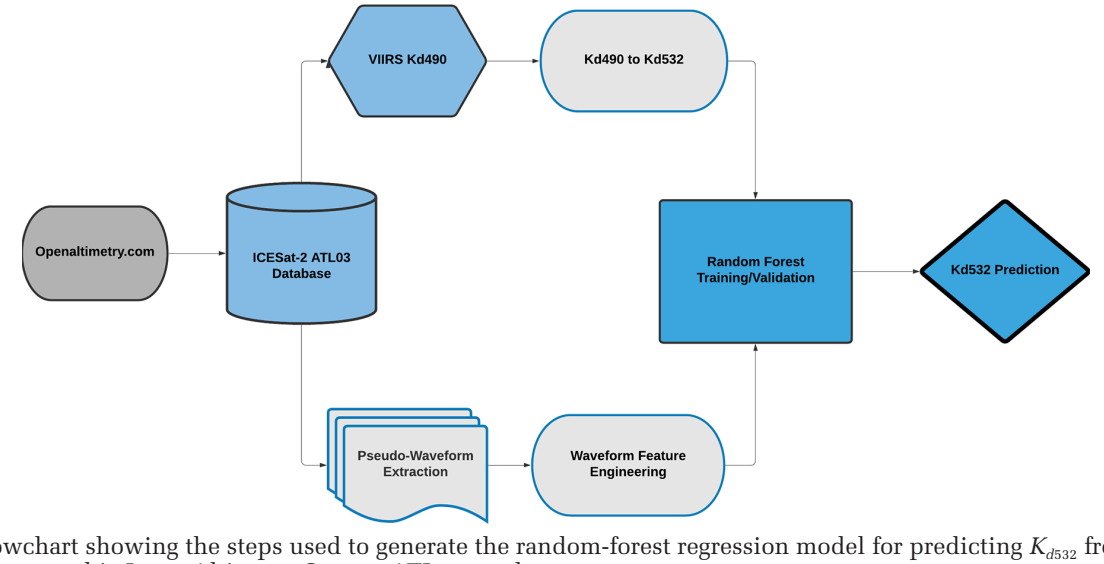

Figure 2. Flowchart showing the steps used to generate the random-forest regression model for predicting K_{d532} from the Advanced Topographic Laser Altimeter System ATL03 product.

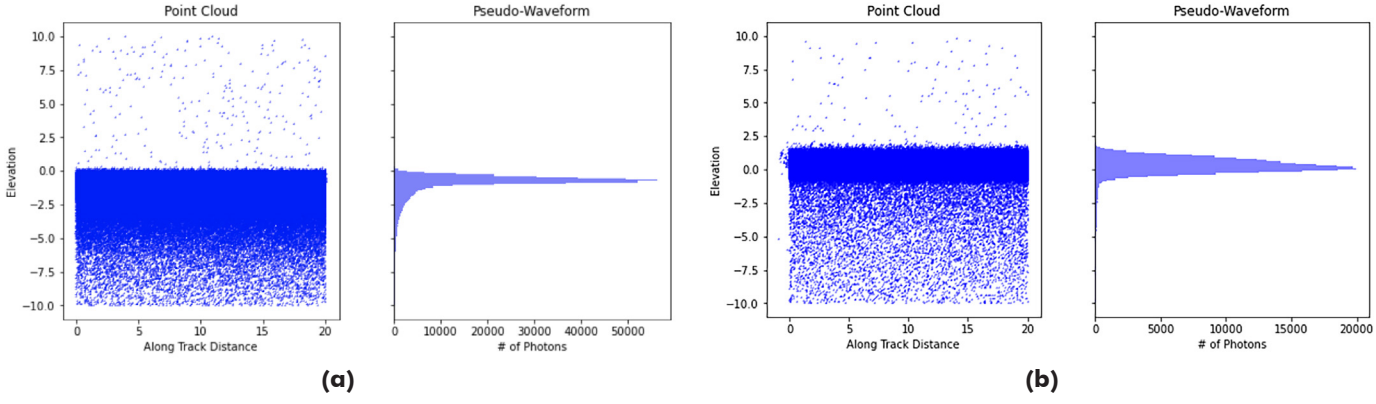

Figure 3. (a) Comparison of the point cloud and corresponding pseudo-waveform measured by the GT1R beam off the coast of Dwarka, India. (b) Comparison of the point cloud and corresponding pseudo-waveform measured by the GT1R beam off the coast of Portland, OR, USA.

are accounted for; Guenther 2007), the situation is complicated by the fact that the decay is a function not solely of the volume backscatter but also of the system impulse response function and other system-specific parameters. It should be noted that the sample pseudo-waveforms shown in Figure 3 are fairly "clean" examples; artifacts such as ringing or after-pulsing after the strong water-surface return are often present. In addition, the ATLAS sensor is susceptible to solar-induced background noise, particularly during the daytime, which can affect the signal decay (Neuenschwander and Macgruder 2019; Malambo and Popescu 2020; McGarry *et al.* 2021).

Based on these considerations, there are two fundamentally different approaches to computing K_d from ATLAS pseudo-waveforms. The first is to apply deconvolution, noise removal, and/or other signal processing as preprocessing steps before curve fitting and then, if needed, to apply an additional step of converting from the lidar attenuation coefficient to K_d (Feygels *et al.* 2003; Churnside 2013; Carr and Tuell 2014; Zhang *et al.* 2021). This general approach has been tested by others (Lu *et al.* 2019, 2020). The second approach is to avoid additional preprocessing and simply use the pseudo-waveforms (and/or derived features) "as is" in machine-learning algorithms, which should be able to learn the associations, even in the presence of noise or artifacts. While neither of these two fundamentally different approaches is inherently right or wrong, and both have associated trade-offs, based on experimentation with both we prefer the latter. Its advantages include the fact that it is simpler, avoids extensive preprocessing (which may introduce complications or errors if the preprocessing algorithms are not tuned correctly), is more robust to solar-induced background noise, and does not require knowledge of the system impulse response function, which may not be available and may change over time. Additionally, because of the step in our procedure of training the model with K_d data, as long as the relationship between the lidar attenuation coefficient and K_d can be modeled, a separate conversion from the lidar attenuation coefficient to K_d is unnecessary, as it is inherently accounted for in the training procedure.

Feature Engineering

In order to describe the shapes of the pseudo-waveforms, we treated the pseudo-waveforms as statistical distributions and calculated several features of the data: the mean, median, standard deviation, median absolute deviation, skewness, and kurtosis. In addition to these statistical features, we calculated several nonstatistical indices to describe the pseudo-waveforms: the number of peaks, the ratio of the areas under the curve between 0 and −1 m elevation and between −1 and −10 m, and the maximum slope.

Table 1 shows the equations and algorithms used to calculate the features of the pseudo-waveforms. The statistical features are based on statistical moments, which are calculated using Equations 3 and 4 (Parrish *et al.* 2014), where $\bar{n}$ is the distribution mean and m_i is the i^{th} moment of the distribution:

$$\bar{n} = \frac{\sum_{n=0}^{N-1} n \cdot y[n]}{\sum_{n=0}^{N-1} y[n]} \tag{3}$$

$$m_i = \frac{\sum_{n=0}^{N-1} (n-\bar{n})^i \cdot y[n]}{\sum_{n=0}^{N-1} y[n]} \tag{4}$$

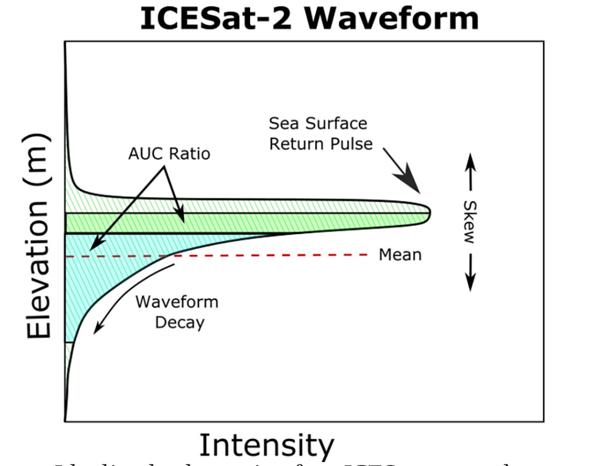

Figure 4. Idealized schematic of an *ICESat-2* pseudo-waveform showing some of the features of interest used in the process of developing the regression model.

One important consideration in calculating these pseudo-waveform statistics and features was the distinction between strong and weak beams in *ICESat-2* ground tracks. To determine whether the weak- and strong-beam data could be combined in the modeling process, we segmented the data by beam type and conducted *t*-tests on the distributions of the features from the strong versus weak beams, with a significance level of 95%. The *p*-values from this analysis (Table 2) indicated that the majority of the waveform features from the strong and weak beams could be considered parts of the same populations with greater than 95% confidence, although low *p*-values were noted for the quartiles, median absolute deviation, and number of peaks. Since these features are particularly sensitive to the signal-to-noise ratio, it makes sense that they would differ between the strong and weak beams. Given that the majority of features, including the three with the highest predictive power (kurtosis, standard deviation, mean), met the 95% confidence criteria, we decided to combine the data from the weak and strong beams. This decision had the added benefit of allowing us to generate one model, as opposed to a weak-beam model and a strong-beam model.

Regression Analysis

Random-forest (RF) regression is a supervised machine-learning technique used to approximate a function between a set of independent variables (e.g., features) and a continuous dependent variable (e.g., ground truth). It is an ensemble method that builds a large number of decision-tree (DT) predictors that depend on randomly sampled, independent, identically distributed vectors within the feature space (Breiman 2001). Each DT predictor in the forest learns a different mapping from the feature space to the dependent variable, using the Binary Recursive Partitioning algorithm (Cutler *et al.* 2011). The final prediction of the RF regression model is the average value of the DT predictors; as the number of DT predictors increases, the error of the model converges *almost surely* (Breiman 2001).

In order to determine the accuracy of the model, the data are first split into training and test subsets. The RF regression model is built using the training data and then applied to the test data. Accuracy metrics can then be calculated by comparing the known values y of the dependent variable from the test set with the model-predicted values $\hat{y}$. Due to the inherent randomness of the algorithm, the model generated in one round of training is often not an exact replica of a model generated by a subsequent round of training. Additionally, different partitions of the data set into training and test subsets introduce different biases into the models

Table 1. Features of pseudo-waveforms.

Feature	Description	Equation, Pseudo-Code, or Reference for Algorithm
AUC ratio	The ratio of the AUC between 0 and −1 m to the AUC between −1 and −10 m	$\text{AUC ratio} = \frac{\text{AUC}_{0\text{-}(-1)}}{\text{AUC}_{(-1)\text{-}(-10)}}$
A/B ratio	The ratio of the AUCs above and below 0 m	$\text{A/B ratio} = \frac{\text{AUC}_{\text{Above}}}{\text{AUC}_{\text{Below}}}$
Number of peaks	The number of peaks in the waveform with a prominence greater than 16 (i.e., the saturation point of the ATLAS sensor)	scipy.signal.find_peaks (# photons, prominence = 16) (Virtanen *et al.* 2020)
5th percentile	The noise in the waveform as measured by the value separating the smallest 5% of photon counts from the other 95%	$P_5 = 0.05\ N$
Mean	The center of the waveform as measured by the statistical mean of the distribution	$\mu = \frac{\sum_{i=1}^{N}(i \cdot x_i)}{\sum_{i=1}^{N} x_i}$
Median	The center of the waveform as measured by the value that separates upper and lower portions of the waveform equally	$\text{Median} = \frac{m_4}{m_2^{3/2}}$
Mode	The center of the waveform as measured by the location of the largest photon-counting bin	Mode = Elev. [max (# of photons)$_{\text{index}}$]
Standard deviation	The spread of the waveform as measured by the square root of the variance (i.e., the second moment of the distribution)	$\sigma = \sqrt{m_2}$
Skewness	The direction and magnitude of the waveform tail as measured by the population skewness	$\text{Skewness} = \frac{m_3}{m_2^{3/2}}$
Kurtosis	The two-sided magnitude of the waveform tail as measured by the population kurtosis	$\beta_2 = \frac{m_4}{m_2^2}$
Amplitude	The peak size of the waveform as measured by half the difference between the minimum and maximum	$\text{Amp.} = \frac{1}{2}(\max(\text{photons}) - \min(\text{photons}))$
Maximum slope	The steepness of the waveform as measured by the largest rate of change between consecutive photon bins	$\text{Max slope} = \max_{i=0,1,\ldots,n-1}(x_i - x_{i+1})$
Median absolute deviation	The spread of the waveform as measured by the median distance from the mean	$\text{MAD} = \text{median}\,\lvert x_i - \mu \rvert$
Pearson 1st coefficient	The direction and magnitude of the waveform tail	$\text{Pearson 1} = \frac{\mu - \text{mode}}{\sigma}$
Pearson 2nd coefficient	The direction and magnitude of the waveform tail	$\text{Pearson 2} = \frac{\mu - \text{median}}{\sigma}$
Q_1	The noise present in the waveform tail as measured by the first quartile	$Q_1 = 0.25N$
Q_2	The shape of the waveform as measured by the second quartile	$Q_2 = 0.50N$
Q_3	The shape of the waveform as measured by the third quartile	$Q_3 = 0.75N$

ATLAS = Advanced Topographic Laser Altimeter System; AUC = area under the curve.

and subsequent accuracy metrics. Because of these two facts, it is common practice to train, test, and retrain a model many times to create a distribution of accuracy metrics that can be used to determine the overall accuracy.

We chose RF regression for this study because of its high level of interpretability compared with other machine-learning methods. In particular, we were interested in understanding the predictive power of the features of the waveform in order to better understand how turbidity affects the distribution of photon returns. We implemented the RF regression model using the RandomForestRegressor method from the Scikit-Learn version 0.23.2 (Pedregosa *et al.* 2011) package in Python.

Estimation Process

We collected ATL03 ground-track data from 94 coastal sites around the world using the OpenAltimetry.org web-based user interface for *ICESat-2* data (Neumann *et al.* 2019, 2020b). Sites were selected based on the availability of corresponding VIIRS K_{d490} data, with an emphasis on acquiring a wide range of VIIRS K_{d490} values and good geographic distribution. We focused exclusively on nearshore locations, as these areas are

Table 2. *p*-values calculated from *t*-tests.

Feature	*p*
AUC ratio	0.385
A/B ratio	0.391
Number of peaks	0.017
5th percentile	0.008
Mean	0.538
Median	0.540
Mode	0.540
Standard deviation	0.360
Skewness	0.992
Kurtosis	0.558
Amplitude	0.694
Maximum slope	0.312
Median absolute deviation	0.009
Pearson 1st coefficient	0.595
Pearson 2nd coefficient	0.935
Q_1	0.007
Q_2	0.001
Q_3	0.008

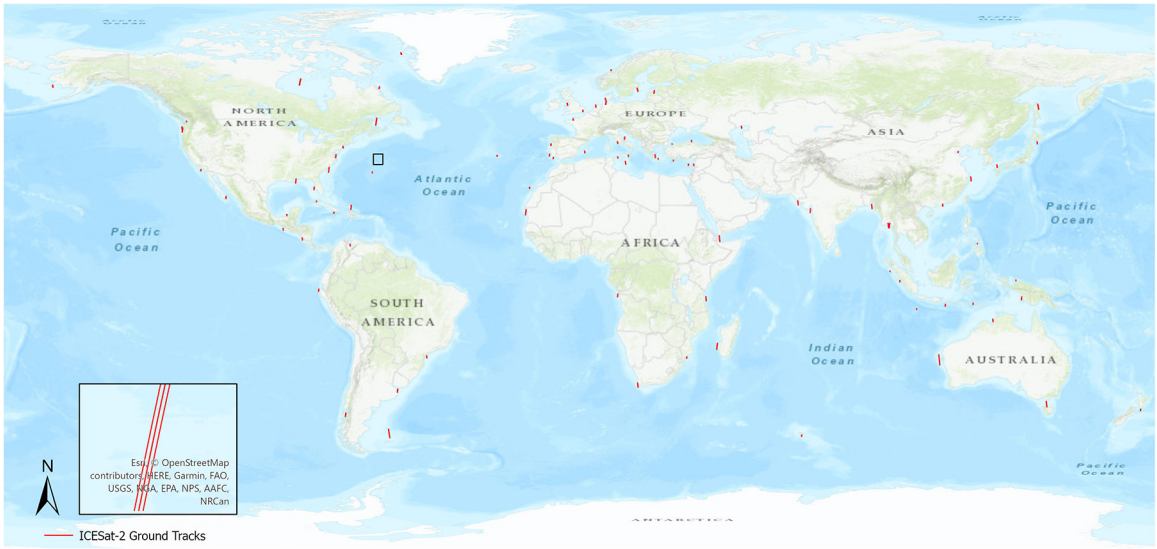

Figure 5. Locations of the Advanced Topographic Laser Altimeter System ground tracks collected for this study.

of the greatest interest in ecological and engineering projects that rely on turbidity estimates. Additionally, due to the coarse spatial resolution of passive spaceborne sensors used to map K_d, such as VIIRS, these coastal areas represent a significant gap in turbidity data. Figure 5 shows the locations of the ATLAS ground tracks we acquired for this study.

For each ground track in our ATLAS data, we acquired three VIIRS K_{d490} values, corresponding to the midpoint and both endpoints, and used the average of these three values as the ground-truth value for that ground track. Coordinates with missing K_{d490} values, whether due to atmospheric conditions, sun glint, or satellite orbital patterns, were ignored. In collecting the VIIRS K_{d490} values, we selected values that minimized the time difference between the corresponding ATLAS and VIIRS measurements. No two corresponding measurements were taken more than 24 hr apart. For comparison, the range of K_{d490} values in this data set was 0.02 to 5.2 m^{-1}. This translates to a K_{d532} range of 0.05 to 3.6 m^{-1}.

We then converted the photon heights for each ground track from ellipsoidal to orthometric heights using the Earth Gravitational Model 2008 geoid model and applied the moving-window binning method with an along-track distance of 20 m and height of 1 dm to generate pseudo-waveforms. (The reasons for converting from ellipsoid height to orthometric height were to remove the water-surface tilt that is common when using ellipsoid heights, due to the geoid gradient, and to set the water-surface height near zero; Babbel *et al.* 2021). Finally, we truncated each pseudo-waveform between 10 m and −10 m and calculated the set of features described in Table 1.

Next we converted the K_{d490} values from VIIRS to K_{d532} using the following empirical relationship (Lu *et al.* 2016):

$$K_{d532} = 0.68(K_{d490} - 0.022) + 0.054 \quad (5)$$

This conversion allowed for a more direct comparison of the ATLAS-derived K_d and the VIIRS K_d, as well as a calculation of K_d in the wavelength native to ATLAS. This is particularly advantageous for future bathymetric studies using ATLAS, because it can be used to estimate the depth of the lidar penetration in the water column, and thus the maximum depth at which bathymetry can be retrieved, for a given area.

Using these features and the K_{d532} values calculated from VIIRS, we conducted a preliminary regression using a single DT predictor on all the data with all the features included.

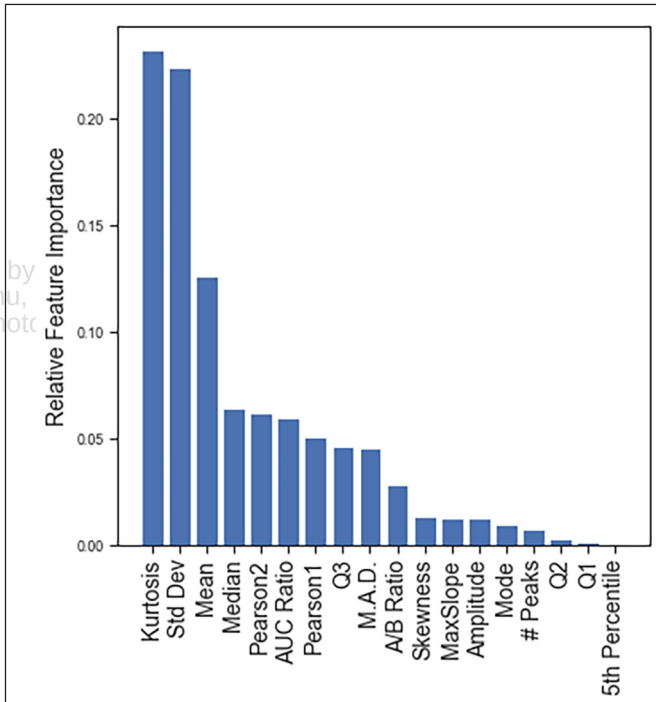

Figure 6. Relative importances of the pseudo-waveform features used in the K_{d532} model. The sum of the importances is 1.0.

Based on the results of this preliminary test, we determined the relative importance of each feature in partitioning the data (Figure 6).

Next we generated a correlation matrix of all the features (Figure 7). Using the correlation matrix and relative feature importances, we systematically trimmed features from the data set by comparing pairs of features with a correlation greater than 0.75 or less than −0.75 and removing the feature with the lesser importance from the data set.

The resulting features are shown in Figure 8. The main purpose of this feature reduction procedure was to reduce correlation in the final training data set and avoid overemphasizing particular attributes of the waveforms in the modeling process. However, reducing the number of features in the data

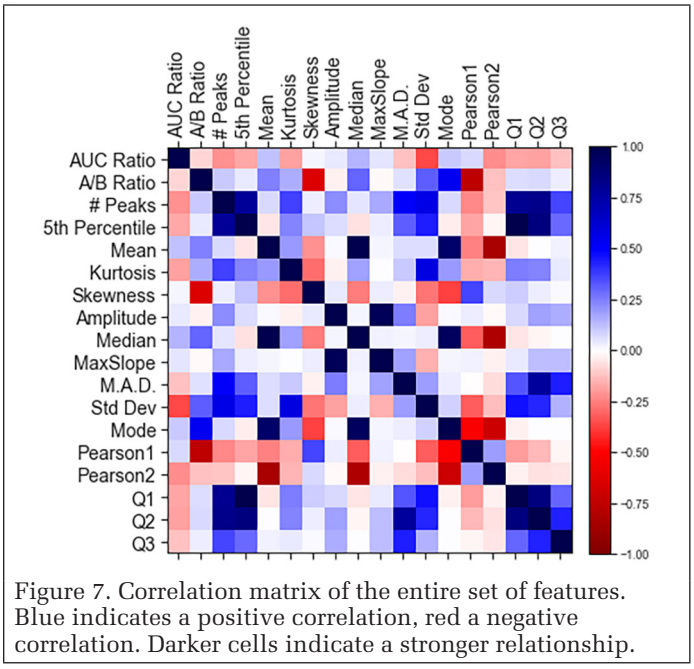

Figure 7. Correlation matrix of the entire set of features. Blue indicates a positive correlation, red a negative correlation. Darker cells indicate a stronger relationship.

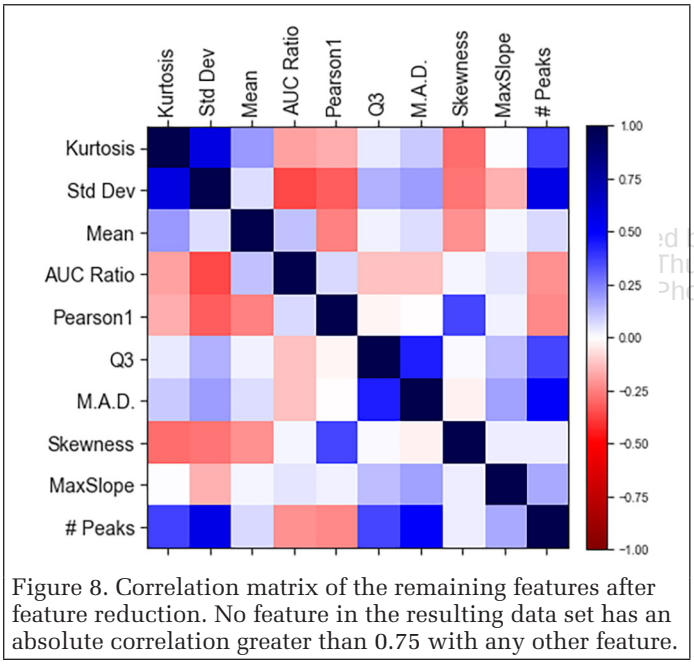

Figure 8. Correlation matrix of the remaining features after feature reduction. No feature in the resulting data set has an absolute correlation greater than 0.75 with any other feature.

also decreased the computation time, allowing us to train models with significantly more DT predictors. Additionally, the reduced computation time allowed us to retrain the model many times on different subsets of the data, thereby mitigating the bias introduced by the training–test split and providing more reliable accuracy metrics.

Using these features, we trained 5000 RF regression models, each with 100 DT predictors. For each model, the data set was randomly split into 80% training and 20% test data. After each round of training we evaluated the model using the test data and calculated the coefficient of determination (R^2), mean squared error (MSE), mean absolute error (MAE), and mean relative difference (MRD):

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} \tag{6}$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{7}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{8}$$

$$\mathrm{MRD} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|} \tag{9}$$

where $\hat{y}$ is the model-predicted K_{d532}, y is the VIIRS K_{d532}, and $\bar{y}$ is the average VIIRS K_{d532}.

Equation 6 (R^2) is used to determine the amount of variance in the VIIRS K_{d532} explained by the model. MSE and MAE (Equations 7 and 8) are measures of the error between the K_{d532} from VIIRS and the model K_{d532} in units of m^{-1}. MRD (Equation 9) is a unitless measure of the error between the VIIRS and model K_{d532}. After the 5000th training round, we computed the averages of these four metrics, which serve as the overall modeling accuracy metrics for the final model.

Finally, we trained a model using the entire data set, which can be used to predict K_{d532} from future ATLAS data.

Results

The results of the model evaluation show that on average, the RF regression model is able to explain 67% ± 12% of the variance in K_{d532}. The average MSE of the model is 0.16 ± 0.06 m^{-2}, with an average MAE of 0.21 ± 0.03 m^{-1}. The standard uncertainty ($\pm\sigma$) of each of these metrics provides an indication of the ability of the model to generalize to new data. This indicates that the predictions of K_{d532} made on new data by the final model can be expected to differ from VIIRS K_{d532} by 0.21 m^{-1}. Additionally, the average MRD is 107% ± 25%. While this value is substantially larger than the MRD observed by Lu *et al.* (2020), who reported an MRD of 10%, their geographic and temporal scope was limited to three ATLAS ground tracks collected over a period of 1 month around the Antarctic coast. Additionally, the K_{d532} values they used covered a relatively narrow, low-turbidity range (0.05 to 0.2 m^{-1}), whereas the K_{d532} values in the present study cover a much greater range extending into substantially more turbid water: 0.05 to 3.6 m^{-1}. Figure 9 shows the distributions of the metrics over 5000 training rounds.

Figure 10 shows the results of the final model after training on all the data, with ground tracks sorted by VIIRS K_{d532} on the horizontal axis. A visual comparison shows strong agreement between the observed VIIRS K_{d532} and the model-predicted values. Furthermore, the final model fits the training data with an R^2 of 0.91, indicating that the features selected for the modeling process are able to capture 91% of the variability in the VIIRS K_{d532}. It is important to note that this R^2 indicates only the level of agreement between the model-predicted K_{d532} and the VIIRS K_{d532} within the training data set, not the accuracy of the final model when applied to previously unseen pseudo-waveform data (which is given instead by the metrics in Figure 9). Though subtle, this difference is extremely important to note for future work building on the results of this study.

Figure 11 shows the residuals of the modeled K_{d532}. The maximum and minimum residuals are 1.63 and –1.17 m^{-1}, with a mean residual of –0.01 m^{-1}. The largest residuals are observed between VIIRS K_{d532} of 0.5 and 1.5 m^{-1}. The algorithm does an especially good job fitting to values lower than 0.5 m^{-1}, likely because the bulk of the training data is clustered between 0.0 and 0.5 m^{-1}. We do not consider the abundance of low K_{d532} values in our training data to be an

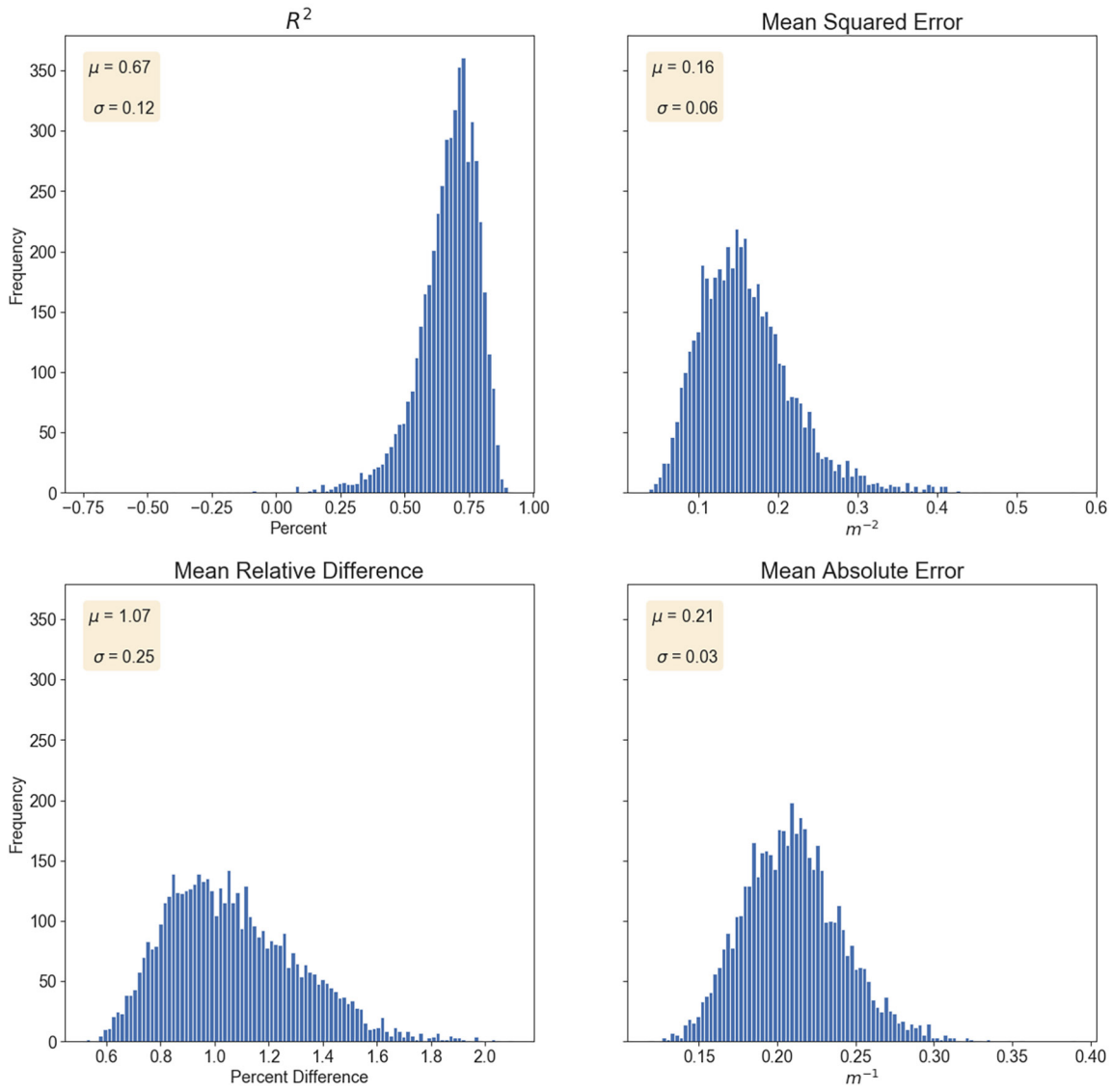

Figure 9. Distributions of R^2, mean squared error, mean relative difference, and mean absolute error scores across the 5000 model training runs.

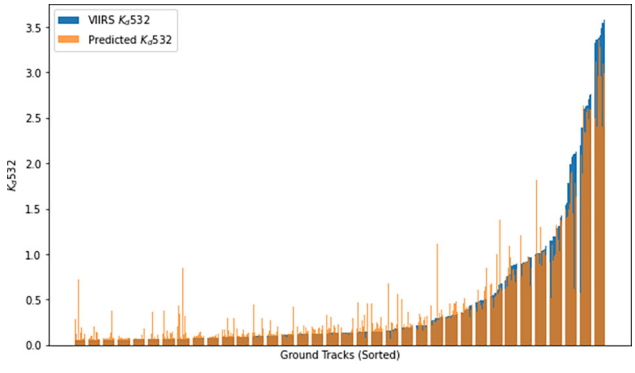

Figure 10. Comparison of the Visible Infrared Imaging Radiometer Suite (VIIRS) K_{d532} (blue) and the model-predicted K_{d532} (yellow). Ground tracks are shown in ascending sorted order, with the lowest VIIRS K_{d532} on the left and the highest on the right.

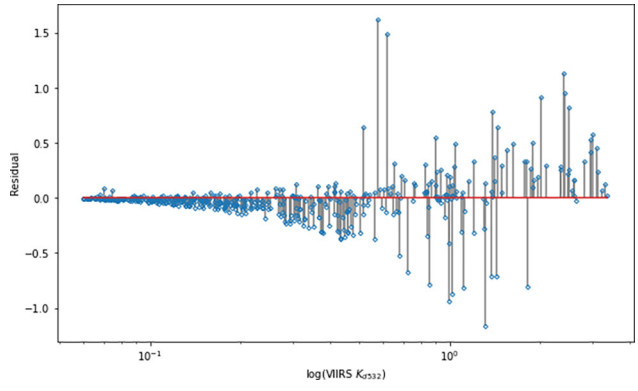

Figure 11. Residuals of the modeled K_{d532}.

overrepresentation, but instead an accurate reflection of the distribution of K_{d532} on a global scale. It is also worth mentioning that many imagery-based K_{d490} retrieval algorithms have been shown to perform poorly in higher-turbidity waters, and it is possible that the larger residuals shown in Figure 11 indicate that this problem persists into the current iteration of the retrieval algorithm (M. Wang *et al.* 2009; Zhao *et al.* 2013).

Conclusion

Measuring turbidity is an important task in many fields of coastal and oceanographic study. Currently, large-scale efforts to measure turbidity on a global level rely solely on satellite imagery. While these techniques have been shown to be effective, they are unable to measure K_d at depth, and rely instead on measurements of water-leaving irradiance. In this study we demonstrated a machine-learning-based approach to extracting K_d from the ATLAS instrument aboard NASA's *ICESat-2*. Using 543 ground tracks, collected from 94 sites across the world, we generated a regression model with an R^2 of 0.67 ± 0.12, an MSE of 0.16 ± 0.06 m^{-2}, an MAE of 0.21 ± 0.03 m^{-1}, and an MRD of 1.07 ± 0.25. While other studies comparing ATLAS-derived K_d and values derived from satellite imagery have reported higher accuracies, our work included data from around the world, rather than a small geographic extent, as well as a wide range of K_d values, extending into higher-turbidity waters. The methods developed here have the additional advantages of bypassing the need for knowledge of the ATLAS system impulse response, simplifying the signal preprocessing procedure, being applicable over much wider ranges of K_d, and providing data comparable to the imagery-based K_d data sets, such that they can be merged and used to fill nearshore gaps in the imagery-based products. Additionally, in evaluating the level of agreement between the *ICESat-2*-derived K_d obtained using the methods of this work and the VIIRS K_d data, it is important to note that the two are generated using fundamentally different types of sensors (active versus passive) and processing workflows, ensuring their independence. Furthermore, this is the first study to rigorously document the achievable accuracies, and thus it can serve as a benchmark for future studies on extraction of K_d from satellite-based lidar.

Another contribution of this study is the development of the suite of pseudo-waveform features, which may be investigated in follow-on work to determine their ability to predict a range of seafloor characteristics (e.g., substrate and cover type) in shallow-water areas. A serialized copy of the final model generated by this study is available at https://github.com/fpcorcoran/ATLAS_Kd532, along with detailed documentation for using the serialized model in a Python environment. While this study demonstrates that K_d can be extracted from *ICESat-2*'s ATLAS instrument across many coastal environments and levels of turbidity using machine learning, we recommend that future studies explore the efficacy of other machine-learning algorithms, such as neural networks, in extracting K_d from *ICESat-2*'s ATLAS instrument.

Acknowledgments

Funding for this research was provided by NASA ROSES Grant 80NSSC20K0964, "*ICESat-2* Bathymetric Studies, Product Development and Data Validation," and subaward UTA20-000752 from Applied Research Laboratories, The University of Texas at Austin, to Oregon State University. We gratefully acknowledge the support of Lori Magruder, University of Texas Principal Investigator and *ICESat-2* Science Team Lead.. We would also like to express our gratitude to the research teams at OSU and UT Austin for ongoing collaboration on *ICESat-2* bathymetric mapping research and to David Harding of NASA's Goddard Space Flight Center for technical input. Additionally, we are grateful for the helpful comments of the three anonymous reviewers.

References

Babbel, B. J., C. E Parrish and L. A. Magruder. 2021. *ICESat-2* elevation retrievals in support of satellite-derived bathymetry for global science applications. *Geophysical Research Letters* 48(5): e2020GL090629.

Branco, A. B. and J. N. Kremer. 2005. The relative importance of chlorophyll and colored dissolved organic matter (CDOM) to the prediction of the diffuse attenuation coefficient in shallow estuaries. *Estuaries* 28(5):643–652.

Breiman, L. 2001. Random forests. *Machine Learning* 45:5–32.

Butler, W. L. 1962. Absorption of light by turbid materials. *Journal of the Optical Society of America* 52(3):292–299.

Carr, D. and G. Tuell. 2014. Estimating field-of-view loss in bathymetric lidar: Application to large-scale simulations. *Applied Optics* 53(21):4716–4721.

Churnside, J. H. 2013. Review of profiling oceanographic lidar. *Optical Engineering* 53(5):051405.

Cutler, A., D. R. Cutler and J. R. Stevens. 2011. Random forests. In *Ensemble Machine Learning: Methods and Applications*, edited by C. Zhang and Y. Ma, 157–176. Boston, Mass.: Springer.

Doxaran, D., N. Cherukuru and S. J. Lavender. 2006. Apparent and inherent optical properties of turbid estuarine waters: Measurements, empirical quantification relationships, and modeling. *Applied Optics* 45(10):2310–2324.

Feygels, V. I., C. W. Wright, Y. I. Kopilevich and A. I. Surkov. 2003. Narrow-field-of-view bathymetrical lidar: theory and field test. Pages 1–11 in *Ocean Remote Sensing and Imaging II, Proceedings Vol. 5155: Optical Science and Technology, SPIE's 48th Annual Meeting*, held in San Diego, CA, 3–8 August 2003. Edited by R. J. Frouin, G. D. Gilbert and D. Pan. City, St.: International Society for Optics and Photonics.

Forfinski-Sarkozi, N. A. and C. E. Parrish. 2019. Active-passive spaceborne data fusion for mapping nearshore bathymetry. *Photogrammetric Engineering and Remote Sensing* 85(4):281–295.

Gallegos, C. L. 2001. Calculating optical water quality targets to restore and protect submersed aquatic vegetation: Overcoming problems in partitioning the diffuse attenuation coefficient for photosynthetically active radiation. *Estuaries* 24(3):381–397.

Gordon, H. R. 1989. Can the Lambert-Beer law be applied to the diffuse attenuation coefficient of ocean water? *Limnology and Oceanography* 34(8):1389–1409.

Guenther, G. C. 1985. *Airborne Laser Hydrography: System Design and Performance Factors.* NOAA Professional Paper Series, *National Ocean Service 1.* Rockville, MD. NOAA Professional Paper Series.

Jasinski, M. F., J. D. Stoll, W. B. Cook, M. Ondrusek, E. Stengel and K. Brunt. 2016. Inland and near-shore water profiles derived from the high-altitude Multiple Altimeter Beam Experimental Lidar (MABEL). *Journal of Coastal Research* 76(10076):44–55.

Jerlov, N. G. 1976. *Marine Optics*, 2nd ed. Amsterdam: Elsevier Scientific Publishing Co.

Klemas, V. 2011. Beach profiling and LIDAR bathymetry: An overview with case studies. *Journal of Coastal Research* 27(6):1019–1028.

Koenings, J. P. and J. A. Edmundson. 1991. Secchi disk and photometer estimates of light regimes in Alaskan lakes: Effects of yellow color and turbidity. *Limnology and Oceanography* 36(1):91–105.

Lee, J. H., J. H. Churnside, R. D. Marchbanks, P. L. Donaghay and J. M. Sullivan. 2013. Oceanographic lidar profiles compared with estimates from in situ optical measurements. *Applied Optics* 52(4):786–794.

Lee, Z., S. Shang, C. Hu, K. Du, A. Weidemann, W. Hou, J. Lin and G. Lin. 2015. Secchi disk depth: A new theory and mechanistic model for underwater visibility. *Remote Sensing of Environment* 169:139–149.

Lee, Z.-P., K.-P. Du and R. Arnone. 2005. A model for the diffuse attenuation coefficient of downwelling irradiance. *Journal of Geophysical Research: Oceans* 110(C2):C02016.

Li, J., Y. Hu, J. Huang, K. Stamnes, Y. Yi and S. Stamnes. 2011. A new method for retrieval of the extinction coefficient of water clouds by using the tail of the CALIOP signal. *Atmospheric Chemistry and Physics* 11(6):2903–2916.

Li, Y., H. Gao, M. F. Jasinski, S. Zhang and J. D. Stoll. 2019. Deriving high-resolution reservoir bathymetry from *ICESat-2* prototype photon-counting lidar and Landsat imagery. *IEEE Transactions on Geoscience and Remote Sensing* 57(10):7883–7893.

Liu, J., W. J. Emery, X. Wu, M. Li, C. Li and L. Zhang. 2017. Computing coastal ocean surface currents from MODIS and VIIRS satellite imagery. *Remote Sensing* 9(10):1083.

Lu, X., Y. Hu, J. Pelon, C. Trepte, K. Liu, S. Rodier, S. Zeng, P. Lucker, R. Verhappen, J. Wilson, C. Audouy, C. Ferrier, S. Haouchine, B. Hunt and B. Getzewich. 2016. Retrieval of ocean subsurface particulate backscattering coefficient from space-borne CALIOP lidar measurements. *Optics Express* 24(25):29001–29008.

Lu, X., Y. Hu, C. Trepte, S. Zeng and J. H. Churnside. 2014. Ocean subsurface studies with the CALIPSO spaceborne lidar. *Journal of Geophysical Research: Oceans* 119(7):4305–4317.

Lu, X., Y. Hu and Y. Yang. 2019. Ocean subsurface study from *ICESat-2* mission. Pages 910–918 in *2019 Photonics & Electromagnetics Research Symposium–Fall (PIERS-Fall)*, held in Xiamen, China, 17–20 December 2019. Edited by J. Editor. Piscataway, N.J.: IEEE.

Lu, X., Y. Hu, Y. Yang, P. Bontempi, A. Omar and R. Baize. 2020. Antarctic spring ice-edge blooms observed from space by *ICESat-2*. *Remote Sensing of Environment* 245:111827.

Malambo, L. and S. Popescu. 2020. PhotonLabeler: An inter-disciplinary platform for visual interpretation and labeling of *ICESat-2* geolocated photon data. *Remote Sensing* 12(19):3168.

McGarry, J. F., C. C. Carabajal, J. L. Saba, A. R. Reese, S. T. Holland, S. P. Palm, J.-P.A. Swinski, J. E. Golder and P. M. Liiva. 2021. *ICESat-2/ATLAS* onboard flight science receiver algorithms: Purpose, process, and performance. *Earth and Space Science* 8(4):e2020EA001235.

Mobley, C., E. Boss and C. Roesler. 2020. Ocean Optics Web Book. <https://www.oceanopticsbook.info> 25 September 2021.

Muirhead, K. and A. P. Cracknell. 1986. Airborne lidar bathymetry. *International Journal of Remote Sensing* 7(5):597–614.

National Academies of Sciences, Engineering, and Medicine. 2018. *Thriving on Our Changing Planet: A Decadal Strategy for Earth Observation from Space: An Overview for Decision Makers and the Public*. Washington, DC: National Academies Press.

Neuenschwander, A. L. and L. A. Macgruder. 2019. Canopy and terrain height retrievals with *ICESat-2*: A first look. *Remote Sensing* 11(14):1721.

Neumann, T. A., A. J. Martino, T. Markus, S. Bae, M. R. Bock, A. C. Brenner, K. M. Brunt, J. Cavanaugh, S. T. Fernandes, D. W. Hancock, K. Harbeck, J. Lee, N. T. Kurtz, P. J. Luers, S. B. Luthcke, L. Magruder, T. A. Pennington, L. Ramos-Izquierdo, T. Rebold, J. Skoog, T. C. Thomas. 2019. The Ice, Cloud, and Land Elevation Satellite – 2 mission: A global geolocated photon product derived from the Advanced Topographic Laser Altimeter System. *Remote Sensing of Environment* 233:111325.

Neumann, T., A. Brenner, D. Hancock, J. Robbins, J. Saba, K. Harbeck, A. Gibbons, J. Lee, S. Luthcke and T. Rebold. 2020. *Algorithm Theoretical Basis Document (ATBD) for Global Geolocated Photons (ATL03), Release 003*. Greenbelt, Md., Goddard Space Flight Center.

Parrish, C. E., L. A. Magruder, A. L. Neuenschwander, N. Forfinski-Sarkozi, M. Alonzo and M. Jasinski. 2019. Validation of *ICESat-2* ATLAS bathymetry and analysis of ATLAS's bathymetric mapping performance. *Remote Sensing* 11(14):1634.

Parrish, C. E., J. N. Rogers and B. R. Calder. 2014. Assessment of waveform features for lidar uncertainty modeling in a coastal salt marsh environment. *IEEE Geoscience and Remote Sensing* Letters 11(2):569–573.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos and D. Cournapeau, D. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Qi, L., C. Hu, J. Cannizzaro, A. A. Corcoran, D. English and C. Le. 2015. VIIRS observations of a Karenia brevis bloom in the northeastern Gulf of Mexico in the absence of a fluorescence band. *IEEE Geoscience and Remote Sensing Letters* 12(11):2213–2217.

Richter, K., H.-G. Maas, P. Westfeld and R. Weiß. 2017. An approach to determining turbidity and correcting for signal attenuation in airborne lidar bathymetry. *PFG—Journal of Photogrammetry, Remote Sensing and Geoinformation Science* 85(1):31–40.

Rogers, J. N., C. E. Parrish, L. G. Ward and D. M. Burdick. 2015. Evaluation of field-measured vertical obscuration and full waveform lidar to assess salt marsh vegetation biophysical parameters. *Remote Sensing of Environment* 156:264–275.

Rogers, J. N., C. E. Parrish, L. G. Ward and D. M. Burdick. 2016. Assessment of elevation uncertainty in salt marsh environments using discrete-return and full-waveform lidar. *Journal of Coastal Research* 76(10076):107–122.

Sarangi, R. K., P. Chauhan and S. R. Nayak. 2002. Vertical diffuse attenuation coefficient (K_d) based optical classification of IRS-P3 MOS-B satellite ocean colour data. *Journal of Earth System Science* 111(3):237–245.

Saulquin, B., A. Hamdi, F. Gohin, J. Populus, A. Mangin and O. F. d'Andon. 2013. Estimation of the diffuse attenuation coefficient KdPAR using MERIS and application to seabed habitat mapping. *Remote Sensing of Environment* 128:224–233.

Saylam, K., R. A. Brown and J. R. Hupp. 2017. Assessment of depth and turbidity with airborne Lidar bathymetry and multiband satellite imagery in shallow water bodies of the Alaskan North Slope. *International Journal of Applied Earth Observation and Geoinformation* 58:191–200.

Shi, W. and M. Wang. 2015. Decadal changes of water properties in the Aral Sea observed by MODIS-Aqua. *Journal of Geophysical Research: Oceans* 120(7):4687–4708.

van Hooidonk, R. J. 2020. Decision Support Tool to Promote Long-Term Survival of Acropora cervicornis on the Florida Reef Tract, NOAA Technical Report, OAR-AOML-53. City, St.: Publisher.

Virtanen, P., R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregtand SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods* 17(3):261–272.

Wang, C., Q. Li, Y. Liu, G. Wu, P. Liu and X. Ding. 2015. A comparison of waveform processing algorithms for single-wavelength LiDAR bathymetry. *ISPRS Journal of Photogrammetry and Remote Sensing* 101:22–35.

Wang, M., X. Liu, L. Jiang and S. Son. 2017. The VIIRS Ocean Color Product Algorithm Theoretical Basis Document, Version 1.0. College Park, MD. NOAA NESDIS Center for Satellite Applications and Research.

Wang, M., S. Son and L. W. Harding Jr. 2009. Retrieval of diffuse attenuation coefficient in the Chesapeake Bay and turbid ocean regions for satellite ocean color applications. *Journal of Geophysical Research: Oceans* 114(C10):C10011.

Wang, M. and C. Wilson. 2017. Applications of satellite ocean color products. Pages 2794–2797 in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS),* held in Fort Worth, Texas, 23–28 July 2017.

Zhang, W., N. Xu, Y. Ma, B. Yang, Z. Zhang, X. H. Wang and S. Li. 2021. A maximum bathymetric depth model to simulate satellite photon-counting lidar performance. *ISPRS Journal of Photogrammetry and Remote Sensing* 174:182–197.

Zhao, J., B. Barnes, N. Melo, D. English, B. Lapointe, F. Muller-Karger, B. Schaeffer and C. Hu. 2013. Assessment of satellite-derived diffuse attenuation coefficients and euphotic depths in south Florida coastal waters. *Remote Sensing of Environment* 131:38–50.