Research Article Vol. 59, No. 21 / 20 July 2020 / Applied Optics 6379

applied optics

Machine learning informed predictor importance measures of environmental parameters in maritime optical turbulence

Christopher Jellen,[1,*] John Burkhardt,[1] Cody Brownell,[1] and Charles Nelson[2]

[1]Mechanical Engineering Department, United States Naval Academy, 1 Wilson Rd., Annapolis, Maryland 21402, USA
[2]Electrical and Computer Engineering Department, United States Naval Academy, 1 Wilson Rd., Annapolis, Maryland 21402, USA
*Corresponding author: cdjellen@gmail.com

Received 15 May 2020; revised 19 June 2020; accepted 20 June 2020; posted 22 June 2020 (Doc. ID 397325); published 17 July 2020

Prediction of the index of refraction structure constant C_n^2 in the low-altitude maritime environment is challenging. To improve predictive models, deeper understanding of the relationships between environmental parameters and optical turbulence is required. To that end, a robust data set of C_n^2 as well as numerous meteorological parameters were collected over a period of approximately 15 months along the Chesapeake Bay adjacent to the Severn River in Annapolis, Maryland. The goal was to derive new insights into the physical relationships affecting optical turbulence in the near-maritime environment. Using data-driven machine learning feature selection approaches, the relative importance of 12 distinct, measurable environmental parameters was analyzed and evaluated. Random forest nodal purity analysis was the primary machine learning approach to relative importance determination. The relative feature importance results indicated that air temperature and pressure were important parameters in predicting C_n^2 in the maritime environment. In addition, the relative importance findings suggest that the air–water temperature difference, temporal hour weight, and time of year, as measured through seasonality, have strong associations with C_n^2 and could be included to improve model prediction accuracy.

https://doi.org/10.1364/AO.397325

1. INTRODUCTION

Atmospheric optical turbulence has a deleterious effect on the operational capabilities of naval laser weapons, power transmission, and communication systems. Optical turbulence may result in significant beam spread beyond the diffractive limit and in optical scintillation at the target. For naval weapon and power transmission systems, this increases the time-on-target required to achieve desired effects, and for communication systems, this can result in lost data transmission. The impact of atmospheric factors on laser beam propagation motivates the need for a deeper understanding of the relationships between environmental parameters and optical turbulence.

Optical turbulence is the result of fluctuations in the index of refraction along the path of an electromagnetic beam, caused primarily by the turbulent mixing of temperature variations in the atmosphere. As the beam passes through regions of varying temperatures, and therefore varying refractive indices, its wave front becomes perturbed, causing additional beam spread, increased variance, and a reduction in irradiance on target. Because the turbulent structure of the atmosphere depends on both large-scale forcing (e.g., solar radiation, pressure) and on local conditions (e.g., surface roughness, albedo), laser propagation will also depend indirectly on a wide range of parameters that influence the spatial structure of the propagation medium. Metrics at the target, such as irradiance, beam spreading, and scintillation are due to the path-integrated effects of these atmospheric parameters.

To form semi-empirical models for laser propagation based on environmental variables, theories of optical propagation through random media are combined with physics-based models of atmospheric turbulence and heat transfer. These models are highly effective in high-altitude environments where optical turbulence is lower and local conditions are less significant, but predictions become inaccurate close to the ground [1–6]. Similar physics-based models have been employed to predict optical turbulence in low-altitude terrestrial environments, which may rely on strong diurnal variations [4–6]. The relationships between environmental parameters and optical turbulence is less well-defined in low-altitude maritime environments [7–11], where a wide variety of environmental parameters that are often neglected by overland as well as high-altitude turbulence models can have a material impact. One Naval Postgraduate School study suggested that other environmental factors such as the properties of the ocean below the propagation path make current models inaccurate for maritime environments [7,8]. Further studies by the Naval Research Laboratory's

Chesapeake Bay Detachment suggest that the temperature difference between air and water, as well as the wind speed along the propagation path, may affect optical turbulence [10].

An improved understanding of which environmental parameters are most important for optical turbulence prediction could enhance the accuracy of current models in the low-latitude maritime environment. This understanding could be further leveraged to create a framework for generating data-driven predictions for optical turbulence in a wide variety of microclimates. Rather than using a universal model, a machine learning based model framework could be trained on local data to generate improved understanding of low-altitude near-maritime optical turbulence. This investigation focuses on analyzing these models to gain insight into the relative importance of measured environmental parameters in predicting C_n^2. To our knowledge, the data-driven modeling techniques developed in machine learning have yet to be applied to exploration of the physical interactions that generate optical turbulence in the near-maritime environment.

Using a newly acquired 15 month long data set, this paper applies a data-driven machine learning approach to determine the relative importance of 12 environmental variables for the prediction of C_n^2, a common measure of optical turbulence in the environment. In summary, the theoretical foundation of C_n^2 as a measure of optical turbulence is presented in Section 2. Section 3 describes the acquisition and preparation of the data set. Section 4 outlines the machine learning techniques used to determine the relative importance of the environmental variables considered. Section 5 describes the results of the analysis, while Section 6 provides concluding remarks.

2. THEORY

The primary driver of electromagnetic wavefront distortion in the atmosphere is random fluctuations in the index of refraction of the medium. At optical frequencies, this is driven primarily by the absolute temperature difference between eddies, but also experiences contributions due to the pressure and humidity of the atmosphere along the beam path [1,2,4]. These effects can be extreme due to shear stress and temperature gradients, which are largest at the surface. Due to boundary layer effects in the near-maritime environment, the refractive index of air can change rapidly over small distances and time intervals.

In order to predict the effect of atmospheric turbulence on beam propagation, the atmosphere can be considered locally homogenous [1]. With this assumption, changes in the refractive index of the atmosphere can be expressed using Eq. (1) [1,2]:

$$n(\hat{r}, t) = n_0 + n_1(\hat{r}, t). \tag{1}$$

In Eq. (1), the local refractive index of the atmosphere n is defined to be the mean index of refraction for the non-turbulent atmosphere n_0 with the added turbulence effects given by n_1, which is a function of both the position along the propagation path $\hat{r}$ and time t. This is the Reynolds decomposition for turbulent flows applied to the refractive index. Normalizing the mean index of refraction n_0 to one and applying assumptions of temporal homogeneity discussed in [1,2], the time dependence in the local refractive index is removed. The resulting Eq. (2) approximates the local refractive index of the atmosphere as a function of spatial coordinate $\hat{r}$ along the propagation path [1,2]:

$$n(\hat{r}) = 1 + n_1(\hat{r}). \tag{2}$$

In Eq. (2), $n(\hat{r})$ has been normalized by its mean value. The behavior of the refractive index is tied to the physical properties and interactions of the local atmosphere. Prior studies suggest that these fluctuations in the local index of refraction are related to corresponding temperature and pressure fluctuations [2]. For optical and infrared wavelengths, this relationship is expressed in Eq. (3) [2]:

$$n(\hat{r}) \cong 1 + 79 \times 10^{-6} \frac{P(\hat{r})}{T(\hat{r})}. \tag{3}$$

In Eq. (3), the structure of the local index of refraction n is predicted from the pressure P in $[mbar]$ and the temperature T in $[K]$ along the path $\hat{r}$. Prior studies suggest that local fluctuations in pressure are generally negligible [2], and as a result, the random fluctuations in the index of refraction occur primarily due to temperature effects. Different regions in the local atmosphere along a propagation path vary in temperature [1,2,5–8]. As with a velocity field, the temperature distribution within a turbulent medium is composed of regions of local coherence, described as eddies. These eddies span a range of sizes within the flow, and are continuously generated by large-scale thermal forcing and dissipated through diffusion. One way to describe the fine-scale characteristics of the temperature distribution is through the structure function for temperature, applied between an outer eddy scale L_0 and an inner scale of l_0. Because temperatures of neighboring points are correlated, the structure function calculates a difference in temperature at two points as a function of the proximity between the points, as in Eq. (4). Using dimensional arguments and assuming the flow to be fully turbulent (locally homogenous and isotropic, large separation of scales), the structure function for temperature within the energy-containing scales is further assumed to depend only on the distance between points, separated by R, and on a constant, C_T^2, which characterizes the magnitude of the differences:

$$D_T(R) = (T_1 - T_2)^2 = \begin{cases} C_T^2 l_0^{-\frac{4}{3}} R^2 & 0 \le R \ll l_0 \\ C_T^2 R^{2/3} & l_0 \le R \ll L_0 \end{cases}. \tag{4}$$

In Eq. (4), C_T^2 is the temperature structure constant, and T_1 and T_2 denote the temperature at the two points separated by R. The brackets in Eq. (4) represent the ensemble average of the squared difference in temperatures T_1 and T_2. The concept of the structure function can be extended to describe local refractive index fluctuations on a path, also with outer scale L_0 and inner scale l_0 in Eq. (5) [1,2]. The structure function describing the refractive index within statistically homogenous isotropic turbulence is given by

$$D_n(R) = (n_1 - n_2)^2 = \begin{cases} C_n^2 l_0^{-\frac{4}{3}} R^2 & 0 \le R \ll l_0 \\ C_n^2 R^{2/3} & l_0 \le R \ll L_0 \end{cases}. \tag{5}$$

In Eq. (5), C_n^2 is the index of refraction structure constant, with units of $[m^{-\frac{2}{3}}]$. Using the relationship in Eq. (3), and the

assumptions outlined in the formulation of Eq. (4), the constant C_n^2 can be calculated as in Eq. (6) [1,2]:

$$C_n^2 = \left(79 \times 10^{-6} \left[\frac{K}{mbar}\right] \frac{P}{T^2}\right)^2 C_T^2. \tag{6}$$

In Eq. (6) from [2], temperature T is measured in $[K]$, pressure P is in $[mbar]$, and C_T^2 has units of $[K^2 m^{-\frac{2}{3}}]$. The index of refraction structure constant C_n^2 is a measure of the intensity of the fluctuations in the refractive index of the atmosphere. The temperature structure function in Eq. (4) can be approximated using the mean square temperature difference between measurements taken by two thermometers separated by a known distance. The assumptions of path invariance allow for approximation of the temperature structure parameter, using point measurements of environmental parameters. And while these assumptions simplify the process of predicting C_n^2, they could contribute to the overestimation of the intensity of optical turbulence in low-altitude near-maritime environments [7,8]. In these environments, a wide range of environmental parameters may be required to accurately predict the intensity of optical turbulence.

In addition to point-wise measurements of the medium, the refractive index structure parameter can also be determined, in a path-averaged sense, from optical measurements at a target. To accomplish this, measures of the variance in beam irradiance are normalized to generate a scintillation index σ_I^2, defined in Eq. (7) [1,2,4,5,7,12]:

$$\sigma_I^2 = \frac{\langle I^2 \rangle - \langle I \rangle^2}{\langle I \rangle^2}. \tag{7}$$

The scintillation index is thus a path-dependent measurement of the normalized variance in irradiance along a propagation path. In Eq. (7), I is the irradiance of the beam at the target at a given time. In this context, the brackets in Eq. (7) denote time averaging, which assumes ergodicity [1,2]. The scintillation index for a plane wave is used to calculate C_n^2 using the relationship in Eq. (8) [1,2]:

$$\sigma_I^2 = 1.23 C_n^2 k^{\frac{7}{6}} L^{\frac{11}{6}}. \tag{8}$$

In Eq. (8), the path length L and wave number k of a beam are related to the observed scintillation index through the structure constant C_n^2. Experimentally, this measurement can be done with a scintillometer, a device that records variance in irradiance of a known beam at a receiver positioned a distance L from the transmitter.

Estimating C_n^2 directly can be achieved through use of Eq. (6). This process relies on knowledge of local temperature within each of the eddies along the path. Approximations for the temperature structure function can be made through temperature measurement of large- and small-scale eddies along the propagation path. This process is rarely done experimentally, as small-scale eddies are typically only 1–2 mm across in a near-maritime environment [1,2]. Rather than measuring the absolute temperature difference between different eddies directly, bulk atmospheric similarity theories employ the use of mean flow characteristics along the propagation path. Monin–Obukhov similarity (MOS) theory is used to describe flow behavior for use in C_n^2 prediction [1–3,8]. MOS theory calculates average properties along a path by assuming parameters such as temperature, humidity, and wind speed are path-invariant. Using MOS theory, C_n^2 is predicted using a combination of temperature and humidity effects, where temperature and humidity are modeled with structure parameters from bulk measurements assumed to be invariant along the beam propagation path [1–3,7].

These assumptions of path invariance allow for approximation of the temperature and humidity structure parameters using point measurements of environmental parameters. And while these assumptions simplify the process of predicting C_n^2, they could contribute to the overestimation of the intensity of optical turbulence in low-altitude maritime environments [7,8,10]. In these environments, a wider range of environmental parameters than those used in MOS theory may be required to accurately predict the intensity of optical turbulence.

3. DATA SET

A. Acquisition

In order to investigate optical turbulence in the low-altitude maritime environment, an 890 m propagation path was established over the Severn River at the United States Naval Academy in Annapolis, Maryland. A scintillometer was used to establish a standardized measure of optical turbulence, C_n^2, with which to compare model predictions. The scintillometer link was approximately 2 m above the surface of the water with more than 95% of the of the propagation path over water. This propagation environment is characterized as "near-maritime," as the local atmosphere is affected by the land masses on either side of the path [9,11,13,14]. As the link was not over open ocean, tidal changes were modest and the path's elevation over water was within 1 m of its starting elevation for the duration of the data collection period. A diagram of the scintillometer link location is given in Fig. 1.

In addition to collecting measured readings of optical turbulence through C_n^2, an experimental system to collect measurements of a wide variety of local environmental parameters was also deployed. These environmental parameters included both atmospheric and oceanographic features, taken in reasonably close proximity to the scintillometer link. Most data sources were located with the link, but additional measurements were taken from stations 2 km and 10 km away. Accurate local environmental parameter measures were essential to inform feature importance insights.

The United States Naval Academy's Severn River Watershed Observatory is co-located with the receiver of the scintillometer

Fig. 1. Diagram of location of scintillometer beam propagation path with receiver (left) and transmitter (right) [14].

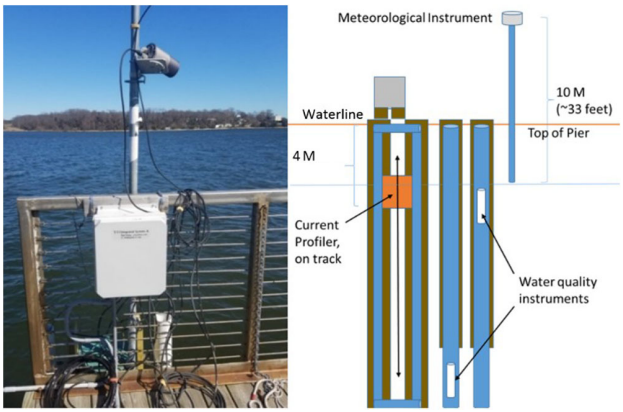


Fig. 2. Diagram of local weather station equipment configuration.

link [15]. This data source is maintained by the United States Naval Academy Oceanography Department, and includes a variety of instruments to measure both meteorological and oceanographic parameters. A Vaisala-WXT536 weather station is fixed approximately 10 m above the Severn River, which reports a variety of atmospheric measurements at 2 or 4 min intervals [15]. The wind speed, air temperature, relative humidity, atmospheric pressure, and rain measurements from this station formed the core of the environmental parameter data set [15,16]. The local water temperature was also measured using a Sontek SL500 side-look ADCP current profiler located approximately 2 m below the water surface. This instrument recorded the water temperature every 10 min [15]. The water temperature 2 m below the surface may be different from on-surface water temperature. In the Severn River environment, there is generally significant mixing within the surface water that mitigates the difference between bulk and surface temperatures. Both the weather station and current profiler recorded measurements from 2 May 2019 until 19 November 2019. A diagram of the local weather station setup, including measurement heights, is provided in Fig. 2.

To supplement the meteorological measurements taken at the Severn River Watershed Observatory, data from the National Oceanographic and Atmospheric Association's (NOAA) WBAN 13752 weather station was used. The NOAA WBAN 13752 weather station is located approximately 2 km from the scintillometer link [14]. This station recorded wind speed, air temperature, relative humidity, atmospheric pressure, and rain measurements to complement those measurements taken at the Watershed Observatory weather station [17]. Additionally, the NOAA WBAN 13752 weather station measured the local visibility in nautical miles. These data were available from 4 January 2019 until 31 March 2020 in quality-controlled hourly intervals [17]. The NOAA quality controls involved averaging measurements taken over the course of each hour to report an overall measurement for a given time stamp [17]. The reported atmospheric parameters with units of measure, time interval, and measurement instrument are given alongside the weather station location in Table 1.

In addition to the water temperature data collected locally, measurements taken from a National Data Buoy Center (NDBC) station were used. The NDBC station at Thomas Point Light in the vicinity of the scintillometer link measured the water temperature approximately 1 m below the water surface [18]. The reported oceanographic parameters with units of measure, time interval, and measurement instrument are given alongside the data buoy location in Table 2.

Similar to the NOAA weather station, the observation station in Table 2 reported quality controlled hourly readings [18]. Multiple measurements were taken in each time interval to produce an hourly observed local water temperature. These hourly data were available from 4 January 2019 until 31 March 2020 [18].

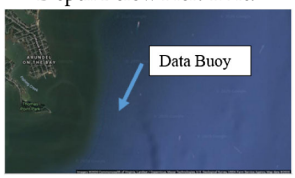
Two additional temporal parameters were added to the data set before final standardization and analysis. In an attempt to capture the potential relationship between optical turbulence and sunlight, the temporal hour weight was calculated for each observation of C_n^2. Temporal hour weight is obtained by dividing the difference between the time of each observation and local sunrise time by 1/12 the time between local sunrise and sunset, as in Eq. (9) [1]:

$$W_{\text{th}} = \frac{12(t_{\text{observed}} - t_{\text{sunrise}})}{t_{\text{sunset}} - t_{\text{sunrise}}}. \tag{9}$$

Table 1. NOAA Weather Station Data Summary

Parameter	Measurement Interval	Units	Source:
Dew point temperature	Observations reported hourly using intermittent averaging	[°F, °C]	NOAA Weather Station WBAN 13752 Date Range: January 1st 2019 to 31st March 2020 Latitude: 38.9913 °N Longitude: 76.4891 °W. Elevation from Ground: 6 ft.
Atmospheric pressure	Observations reported hourly using intermittent averaging	[bar]	
Hourly pressure change	Observations reported hourly based on the observed change in reported atmospheric pressure	[bar]	
Hourly rainfall	Observations reported hourly, total rainfall for the duration reported	$[\frac{cm}{hr}]$	
Relative humidity	Observations reported hourly intermittent averaging	[%]	
Visibility	Observations reported hourly	[nm]	
Wind direction	Observations reported hourly, none listed when local relative wind has no clear and consistent direction	$[\frac{m}{s}]$	
Average wind speed	Observations reported hourly using extreme outlier rejection and intermittent averasing	$[\frac{m}{s}]$	
Gust wind speed	Observations reported hourly, highest wind speed with multiple observations reported	$[\frac{m}{s}]$	

Table 2. NDBC Data Buoy Summary

Parameter	Measurement Interval	Units	Source:
Water temperature	Observations reported at 15 minute intervals	[°C]	NDBC Station TPLM2 Date Range: January 1st 2019 to 31st March 2020 Latitude: 38.899 °N. Longitude: 76.436 °W. Depth Below Pier: 12 ft. 

The recorded time of each observation was determined using the scintillometer internal clock, while local sunrise and sunset were recorded by the USNA Oceanography Department sunrise data set [1,19]. Additionally, seasonality was measured as the fraction of the year elapsed at the time of each C_n^2 scintillometer observation using the native scintillometer clock using Eq. (10):

$$S = \frac{d_{\text{observed}}}{d_{\text{end}} - d_{\text{start}}}. \tag{10}$$

In Eq. (10), S represents the fraction of the total observation time between 4 January 2019 (d_{start}) and 30 March 2020 (d_{end}) that has elapsed. And d_{observed} is the number of days from 4 January 2019.

In addition to the hourly-averaged data set in Table 1, a secondary data set using only higher-frequency local environmental parameter measurements and local water temperature was collected. Between 10 December 2019 and 31 March 2020, a Davis VANTAGE PRO2 weather station was established approximately 4 m from the receiver of the scintillometer link. This weather station measured atmospheric parameters at an elevation of approximately 1–2 m above the surface of the water and near ground height next to the scintillometer (see Fig. 1), from the edge of the USNA Severn River Basin pier. Local wind speed, air temperature, relative humidity, atmospheric pressure, and rain measurements were collected. Additionally, the solar radiation, ultraviolet index, and air density were measured. The reported atmospheric parameters with units of measure, time interval, and measurement instrument are given alongside the weather station location in Table 3.

These readings were reported at 10 min intervals between 10 December 2019 and 31 March 2020. Of the 15,600 possible 10 min readings in this period, 10,253 were collected. Many short dropouts, often less than 2 h, occurred due to low battery power. Occasional longer dropouts occurred due to system maintenance. The temporal hour weight and seasonality parameters generated using the internal scintillometer clock were also added to this data set using the same 10 min intervals.

B. Preparation

Before progressing to the final compilation and standardization of the time series data, the scintillometer measurements were cleaned to eliminate observations with missing values or non-physical measurements. For example, due to temporary obstructions such as water traffic or dense fog, some scintillometer readings were reported as missing. Additional scintillometer readings that were above 1×10^{-12} $[m^{-\frac{2}{3}}]$ were removed from the data set as non-physical. Local weather data collected over 15 min intervals were averaged similarly in order to ensure consistent measurement of environmental parameters. Missing or non-physical measurements reported by the Severn River Watershed Observatory were eliminated using hourly averaging of the reported data, in a manner similar to the NOAA and NDBC stations [17,18]. Both the NOAA and NDBC parameter measurements were obtained from their respective quality-controlled data bases, where the readings were reported as hourly averages, with outliers and missing data handled by the NOAA or NDBC procedures before publication [17,18]. When measurements from the local observatory were not available, the corresponding NOAA or NDBC measurements were used to complete the data set. Rather than using the water temperature as measured directly, the air–water temperature difference was obtained by subtracting the measured water temperature from the measured air temperature for each observation.

These atmospheric, oceanographic, and temporal parameters had varying measurement intervals. In order to perform analysis

Table 3. Local Davis Weather Station Data Summary

Parameter	Measurement Interval	Units	Source:
Air temperature	Observations reported at 2-4 minute intervals	[°F, °C]	Davis VANTAGE PR02 Weather Station Date Range: December 10th 2019 to 31st March 2020 Latitude: 38.9832 °N. Longitude: -76.4792 °W. Elevation from Ground: 2 ft. 
Relative humidity	Observations reported at 2-4 minute intervals	[bar]	
Dew point temperature	Observations reported at 2-4 minute intervals	[bar]	
Wind speed	Observations reported at 2-4 minute intervals using interval averaging	$[\frac{m}{s}]$	
Atmospheric pressure	Observations reported hourly intermittent averaging	[hPa]	
Rainfall	Observations reported hourly, total rainfall for the duration reported	[mm]	
Solar radiation	Observations reported at 2-4 minute intervals using interval averaging	$[\frac{W}{m^2}]$	
UV index	Observations reported at 2-4 minute intervals. Derived from pyranometer measurements	$[\frac{m}{s}]$	
Air density	Observations reported at 2-4 minute intervals	$[\frac{g}{m^3}]$	

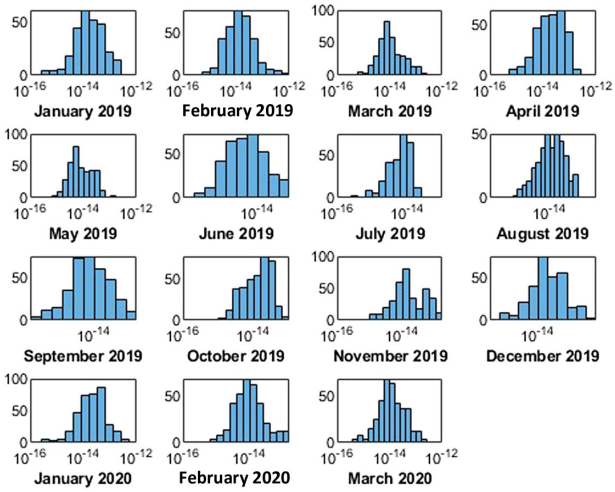


Fig. 3. Monthly histograms of counts of observed C_n^2 $[m^{-\frac{2}{3}}]$ from January 2019 to March 2020.

on the data set, the time of each observation had to be standardized, so that for each index in the time series, a scintillometer reading had corresponding environmental parameter readings. An hourly time series of readings from 4 January 2019 until 31 March 2020 was generated using time synchronization. In order to bring the timing of all of the data sources into reasonable alignment, hourly data averages were computed. This allowed both a wide range of environmental parameters, as well as use of quality-controlled NDBC and NOAA data. Between 4 January 2019 and 31 March 2020, there were 8676 possible hourly observations. The final data set contained 7501 complete observations during this time period. There were some short outages in data recording due to scintillometer misalignment, snowfall, loss of power due to extended overcast days, as well as local boat traffic blockages across the propagation path in the Severn River. A sample of the 12 environmental parameters used in our analysis from the data set is provided in Table 4.

After cleaning, standardizing, and compiling, the environmental parameter readings and C_n^2 measurements from 4 January 2019 to 31 March 2020 were analyzed. The central tendency and distribution of C_n^2 measurements were analyzed, and we experienced a range of values from 2.23×10^{-16} $[m^{-\frac{2}{3}}]$ to 9.96×10^{-14} $[m^{-\frac{2}{3}}]$. For applications in laser communication and weapon systems, C_n^2 values from 1×10^{-16} $[m^{-\frac{2}{3}}]$ to 5×10^{-13} $[m^{-\frac{2}{3}}]$ are generally expected, and this is in line with what we experienced. The field measurements of C_n^2 along the propagation path were partitioned by month and the resulting observation sets plotted as histograms. Both the range and distribution of C_n^2 readings were analyzed and compared to existing literature [5–11]. Monthly histograms of readings from January 2019 to March 2020 are given in Fig. 3.

Both the range of C_n^2 values observed and their distribution are similar to the distributions presented in literature [5–11]. For each month in which data were collected, the minimum C_n^2 value in the data set was around $2 \times 10^{-16}[m^{-\frac{2}{3}}]$, while the maximal value was on the order of $1 \times 10^{-13}[m^{-\frac{2}{3}}]$. The distributions of C_n^2 showed a unimodal distribution with fat

Table 4. Sample of Data Sets for 12 Selected Environmental Parameters

Time	C_n^2 $[m^{-\frac{2}{3}}]$	Air Temperature [C]	Air-Water Temperature Difference [C]	Dew Point Temperature [F]	Pressure [bar]	Hourly Pressure Change [bar]	Relative Humidity [%]	Wind Direction [deg T]	Gust Wind Speed [m/s]	Average Wind Speed [m/s]	Visibility [nm]	Seasonality	Temporal Hour Weight
1/2/2019 01:00	4.79E-16	6.60	−3.00	46	29.78	0.08	93	140	18	3	0.75	0	−0.68
1/2/2019 02:00	4.77E-16	6.60	−2.00	48	29.78	0.08	89	150	18	3	1	0	−0.57
1/2/2019 03:00	5.29E-16	6.60	−1.70	46	29.78	0.08	88	140	18	3	0.5	0	−0.46
1/2/2019 04:00	3.32E-16	6.60	−1.10	47	29.78	0.08	89	140	18	3	0.25	0	−0.36
1/2/2019 05:00	3.25E-16	6.60	−0.90	45	29.78	0.08	86	130	18	6	0.5	0	−0.25
1/2/2019 06:00	1.59E-15	6.60	−0.70	45	29.75	0.08	86	130	18	9	0.25	0	−0.15
1/2/2019 07:00	5.01E-15	6.60	0.00	46	29.75	0.08	90	130	18	5	0.5	0	−0.04
1/2/2019 08:00	3.18E-15	6.60	0.30	46	29.75	0.08	86	130	18	6	0.5	0	0.06

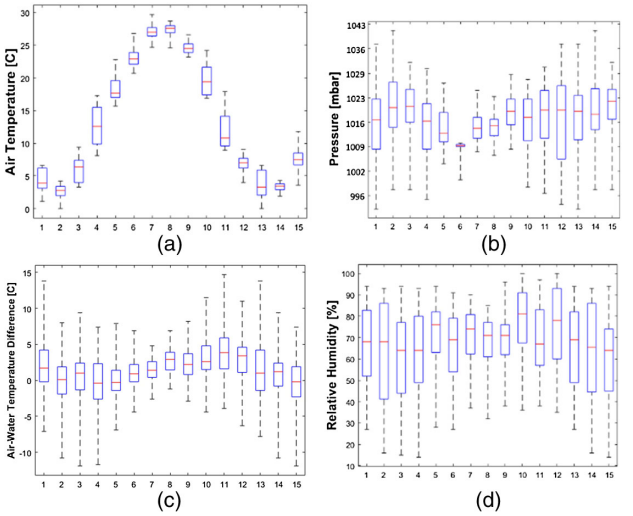


Fig. 4. Aggregate box and whisker plots of readings from January 2019 to March 2020 for (a) temperate, (b) pressure, (c) air–water temperature difference, and (d) relative humidity.

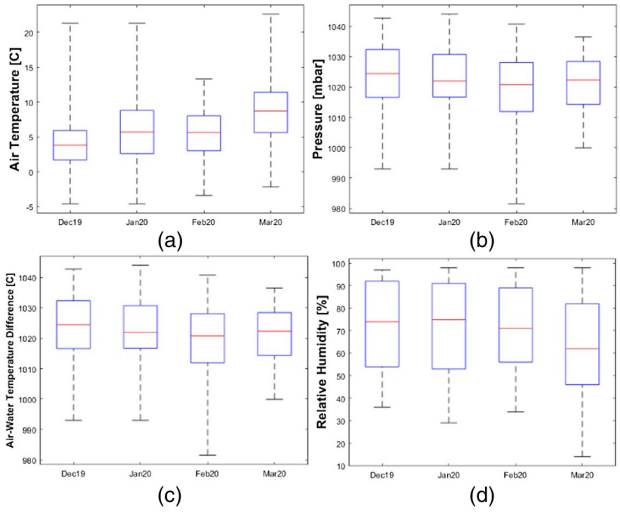


Fig. 5. Aggregate bar and whisker plots of local weather station readings from December 2019 to March 2020 for (a) temperate, (b) pressure, (c) air–water temperature difference, and (d) relative humidity.

tails similar to those presented in [9–11,13]. Both the range and distribution of C_n^2 measurements taken along the Severn River link were in line with the results of other near-maritime links [7,8,10].

The ranges and distributions of the measured hourly-averaged environmental parameters were also examined to ensure they fell within the bounds of existing literature. The recorded observations for the full 15 month period are given as bar and whisker plots in Fig. 4.

As with C_n^2, both the range and distribution of environmental parameter observations are similar to the distributions presented in literature [3,7]. After validation of the hourly-averaged data set, the high-frequency local data were analyzed in a similar way. The ranges and distributions of the 10 min interval readings in each calendar month are presented as bar and whisker plots in Fig. 5.

In addition to analyzing the measurements collected in each of the data sets, the potential relationship between the measured air–water temperature difference and C_n^2 was visualized. Between January 2019 and March 2020, the air–water temperature difference was calculated and compared with the corresponding measurement for C_n^2. Prior investigation suggests that some relationship may exist between these parameters in the near-maritime environment [6,7,10]. The joint distributions of air–water temperature differences and $\mathrm{Log}_{10} C_n^2$ are given in Fig. 6.

The visualization in Fig. 6 confirms lower observed C_n^2 when the air–water temperature difference was near 0°C, with higher C_n^2 as the magnitude of temperature difference increases. Figure 6 shows a relative minimum C_n^2 near 0°C air–water temperature difference and increased C_n^2 for greater differences between air and water temperature. This is in agreement with the results in [7,8,10]. More data were collected with sub-zero air–water temperature difference, which is reflected in the mean air–water temperature difference lying below 0°C.

After exploratory data analysis and validation of scintillometer readings, machine learning techniques were applied

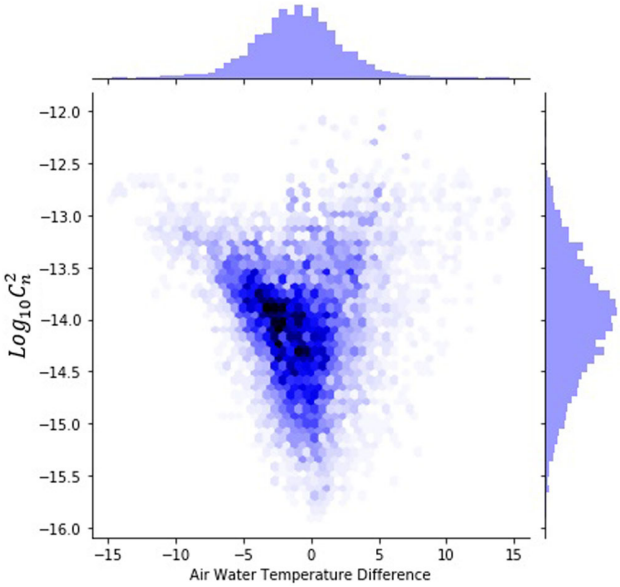


Fig. 6. Joint distribution of C_n^2 versus air–water temperature difference from January 2019 to March 2020.

to generate new understanding of the potential physical relationships between environmental parameters and optical turbulence. Both the hourly 15 month long data set and the 10 min averaged local data set were used to train, develop, test, and investigate the potential physical relationships through feature importance [20–24]. Machine learning techniques were applied to generate new understanding of the potential physical relationships between environmental parameters and optical turbulence using this 15 month long data set. Ultimately, regression trees were the key novel approach that we found to be effective for both modeling and feature importance investigation. The theoretical foundation and a brief example of our regression tree approach are highlighted in the following section

[20–22,24,25]. In addition to basic regression trees, we explored additional machine learning models that utilize regression trees as a base component, such as random forests. [20,21,25].

4. VARIABLE IMPORTANCE MEASURES

The first step in training data-driven machine learning models was to divide the set of all observations previously discussed into a training subset, a development subset, and a test subset. The training subset contained 60% of the 10,253 available observations, with 20% set aside for the development set, and the final 20% set aside for the test set. The models were trained by fitting to the data in the training subset, with no forced physical relationships incorporated in the models. These models were then tuned and re-trained by comparing their predictions for C_n^2 with the measured values of C_n^2 set aside in the development subset. Specific model parameters such as the number of splits in each tree and the number of observations in each terminal node were fixed before training. The appropriate values for these model parameters were determined by training over a wide range of possible values, and selecting the values used in the model with the lowest prediction error [20,21,25].

A. Regression Trees

A regression tree is a nonparametric model that seeks to divide the data set into a number of distinct, non-overlapping regions in order to generate a prediction for a response variable given an input predictor observation.

In this paper, regression trees will seek to sub-divide $\mathbb{R}^{12}$, such that given some new observation $\boldsymbol{x}^{[i]} = \langle x_1, x_2, x_3, \ldots, x_f, \ldots, x_{12}\rangle^{[i]}$, there exists a distinct prediction $\hat{y}^{[i]}$. In order to train a regression tree model, a sample of the global training data set is selected. This data set becomes the first node in the regression tree. In the context of regression trees, a *node* is a collection of observations at some level in the tree. To begin, all observations are in the same node at the top of the regression tree. In order to generate lower levels in the tree, the initial parent node is split into two child nodes [20,21,25]. We pick one variable and a value, and separate the observations in the top node into two groups based on whether the observation is higher or lower than that value. In order to identify the location of a split within an environmental parameter, a splitting criterion is used [20,21,25]. The standard splitting criterion is the sum of square error (SSE) objective function in Eq. (11):

$$SS_{\text{error}} = \sum_{i=1}^{n} \left(y^{[i]} - \hat{y}^{[i]}\right)^2. \tag{11}$$

In Eq. (11), n is the number of observations in the node, $\hat{y}$ is the predicted response, and y is the observed response. The sum of squared differences between $\hat{y}$ and each observed y in a set of two candidate nodes is compared. The lower SSE determines which nodal split is most appropriate [20,21,25]. While a split could, in theory, happen anywhere between two observations, many potential splits result in the same set of child nodes. As there are at most 10,253 training observations and 12 environmental parameters in the data set, the total number of unique splits at each decision point is at most 123,036. All unique splits within each of the 12 predictor variables are compared. The potential split that minimizes the SSE for a given node becomes the final split that divides the node into two smaller, non-overlapping child nodes whose union is the parent node. The order of parameters in the data set does not impact which of the possible splits is chosen, as all unique splits are compared against one another. This process is called recursive binary splitting. When evaluating model performance, the mean value of SSE in each node was used. This metric is the mean square error (MSE). In order to train a regression tree, a training set of 60% of the 10,253 observations was randomly selected without replacement. The regression tree generated on a data set using recursive binary splitting is determined by the training data set and the metric. If the same training set is used to generate two regression trees using the recursive binary splitting approach, the resulting trees will be identical. To illustrate this process, the training set generated on the 10 min average data set is investigated.

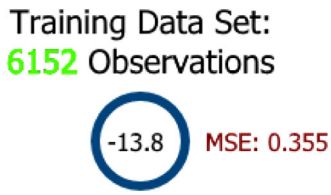


Fig. 7. Training data set as a top-level node with predicted (mean) $\text{Log}_{10}C_n^2$ given in the center of the node.

Using the recursive binary splitting approach, a basic regression tree was trained that sought to divide the training data set into two child nodes with a lower residual sum of squares. The model used the mean of all observation in each node as the predicted C_n^2 for the node. The tree, with both terminal node's prediction for $\text{Log}_{10}C_n^2$ and the corresponding MSE, is presented in Fig. 7.

The split selected for this data set was on measured air density, with observations below 1.28 $[\frac{g}{m^3}]$ separated from those whose measured air density was above 1.28 $[\frac{g}{m^3}]$. The selected split and MSE for the resulting child nodes are given in Fig. 8.

Next, the optimal split was determined within each of the two child nodes, to further divide the measured C_n^2 and corresponding environmental parameter measurements. The next level of splits was on the measured dew point temperature, with 7.15 [°C] and 7.25 [°C] selected as the splits for the left and right nodes, respectively. This process is highlighted in Fig. 9.

Further splits within each of these child nodes occur using the recursive binary splitting algorithm. The visual description of this process is given in Fig. 10.

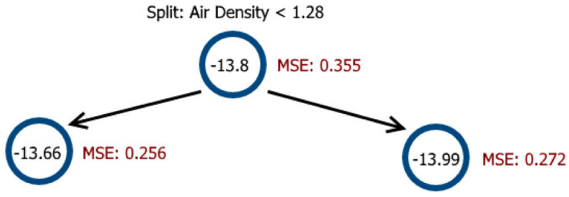


Fig. 8. Resulting child nodes with predicted (mean) $\text{Log}_{10}C_n^2$ in the center of each node.

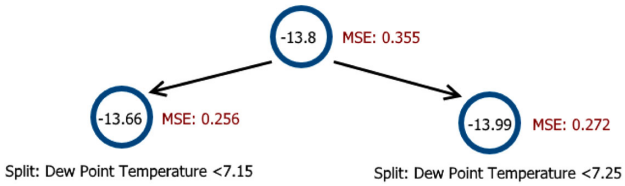


Fig. 9. Second-level resulting child nodes with predicted (mean) $\mathrm{Log}_{10}C_n^2$.

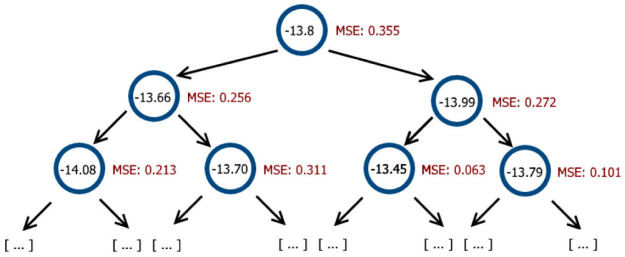


Fig. 10. Recursive binary splitting process for training deep regression trees.

B. Random Forest Algorithm

While a regression tree can act as a full model, generating predictions for any input test observation, it is more commonly used as the building block for a larger ensemble machine learning model. One key challenge of applying individual regression trees as models is the high variance between regression trees trained on different random samples of the same test population. By generating an ensemble of individual regression trees, the overall variance in the model is reduced [20,21,25]. Further, as each tree is unbiased and has the potential to capture the physical interactions hidden in the data set, regression trees are valuable as components of a larger model, called an ensemble model [20,21,23].

Ensemble models generate predictions by leveraging the predictions of their constituent components. In the case of regression tree ensembles, a number of distinct regression trees are trained and incorporated into a new model using a weighting scheme [20,21,25]. For the random forest model, equal weighting between unique regression trees was selected in order to reduce the risk of over-fitting [20,21,25]. The ensemble model prediction is thus informed by the predictions of more than one regression tree, in an effort to maintain low model bias but reduce the high variance of a single tree. The process used to train a random forest in this study considered 500 regression trees, with nine of 12 predictors randomly sampled as splitting candidates for each partition. Although the group of splitting candidates was randomly selected, predictors with a greater impact on MSE reduction were split on more often. At each decision point, a new subset of nine of the 12 overall predictors was randomly generated. As the subset changes for each node in the tree, predictors with stronger influence are expected to generate the nodal split more often. Limiting the subset of parameters available for splitting increases the diversity between trees, which has been suggested to reduce over-fitting to the data. The process is outlined in the following algorithm [20].

1. Generate a random sample $\boldsymbol{X}^{\text{sample}}$ with replacement of size M from the data set.
2. Use this sample $\boldsymbol{X}^{\text{sample}}$ as the top-level parent node.
3. Randomly select nine predictors as splitting candidates, and apply the recursive binary splitting approach highlighted in [20,21,25].
4. Determine the best split within the candidate predictors.
5. Take this best split as the partition, and generate two child nodes.
6. Randomly generate new subsets of nine predictors for the two resulting child nodes.
7. Continue applying recursive binary splitting until the stop condition of a certain minimum number of observations is reached.
8. Repeat steps (2)–(7) until 500 regression trees, $\boldsymbol{T}_1, \ldots, \boldsymbol{T}_{500}$ are fully trained.
9. Generate the random forest prediction, $\hat{f}(\boldsymbol{x})$ as

$$\hat{f}(\boldsymbol{x}) = \frac{1}{500}\sum_{i=1}^{500} T_i(\boldsymbol{x}).$$

As each regression tree is likely to be unique, given that both the training data and parameters available for splitting at each node are randomized, the random forest model can provide a low-biased estimate of observed optical turbulence y^k based on the corresponding environmental parameter $\boldsymbol{x}^k$.

C. Random Forests for Predictor Importance

While regression trees provide a foundation for nonlinear machine learning regression models, they can also be used to investigate the relationships between environmental parameters, x_f, and the response variable C_n^2 [20–25]. Examining which parameter $x_f \in \boldsymbol{X}_i = \langle x_1, x_2, \ldots, x_{12}\rangle$ in feature space is used in nodal partitioning can offer basic insight into variable importance. Parameters that are more important are expected to have greater reductions in the sum of squared deviation when selected for nodal partitioning.

This process can be used to generate variable importance information by analyzing the impact of various parameters on the MSE of the random forest [20–25]. Permutation based MSE reduction is one common technique for generating variable importance information.

In order to induce diversity in the set of regression trees, each tree was limited to find the most effective split using a randomly selected subset of nine of the 12 environmental parameters. These trees can be combined to form a simple random forest model, where the prediction of all three trees is averaged to generate a final prediction for a test observation. By using more than just one regression tree, model prediction and feature importance insights are more stable and have lower error [20,21,25].

As the random forest incorporates regression trees to inform its final predictions, the constituent trees can be investigated to develop relative feature importance metrics. Each model consists of N regression trees $T_1, \ldots, T_N$. Within each of

these trees, there are $M-1$ internal nodes, which excludes the top-level node of each tree [20,25]. For each internal node $t \in \{1, \ldots, M-1\}$, the nodal split that minimizes SSE occurs within one of the 12 environmental parameters x_j, $j \in \{1, \ldots, 12\}$. This split results in two child nodes for the internal node t. The squared relative importance i^2 of environmental parameter x_j in tree T is the sum of nodal MSE reductions resulting from all splits on x_j on the $M-1$ internal nodes, divided by the number of such splits [20].

As random forests use the average of all constituent trees to make a final C_n^2 prediction, the average squared relative importance I^2 of each environmental parameter can be computed using Eq. (12):

$$I^2 = \frac{1}{N}\sum_{k=1}^{N} i_k^2. \tag{12}$$

After I^2 is computed for each of the 12 environmental parameters, the squared relative importance values $I_1^2, \ldots, I_{12}^2$ for each parameter can be normalized by taking the sum of all I^2 terms and dividing by 12.

5. RESULTS

A full random forest model was trained on the set of 6152 training observations using the environmental parameters in Table 1. In order to gain insight into the relative importance of each environmental parameter in the feature space, the average square relative importance I^2 was computed [20,21,25]. In order to measure the accuracy of the random forest model in generating predictions for $\mathrm{Log}_{10} C_n^2$ on the test set, the base 10 logarithm of an observation of C_n^2 from the test set was compared to the model prediction using the corresponding 12 environmental parameters. A scatterplot comparing the predicted $\mathrm{Log}_{10} C_n^2$ and the corresponding measured $\mathrm{Log}_{10} C_n^2$ is given in Fig. 11.

In Fig. 11, the red trend line shows perfect prediction The MSE of predictions from this random forest were computed on the test set, so that the constituent regression trees could not train and test on the same observations. In Fig. 11, the MSE in

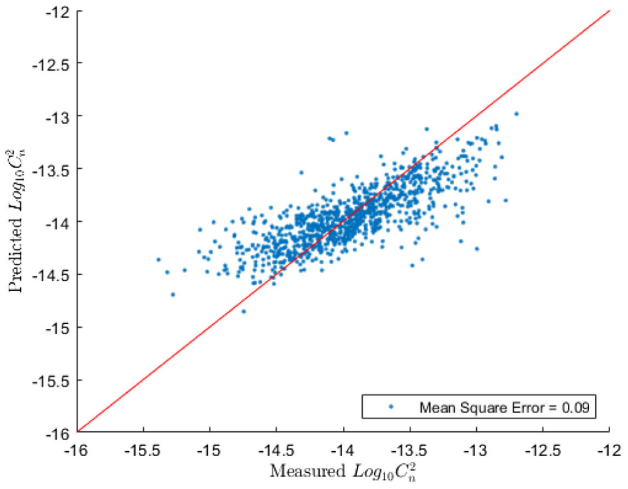


Fig. 11. Random forest model prediction and measured $\mathrm{Log}_{10} C_n^2$ for hourly-averaged data in the test set.

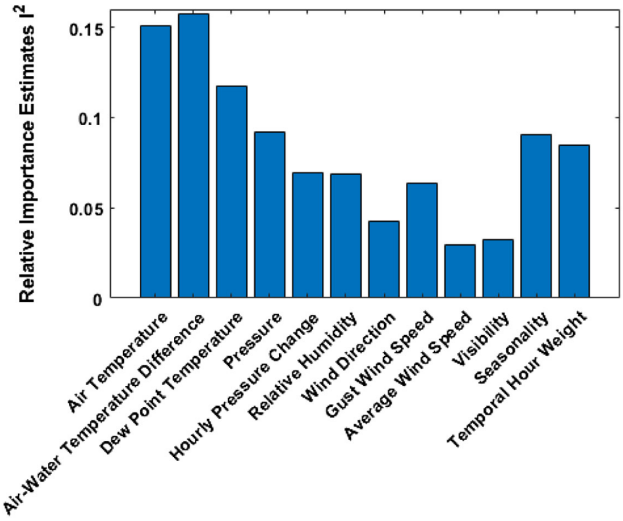


Fig. 12. Environmental parameter relative predictor importance I^2 generated by mean square error descent for hourly-averaged data.

prediction for the $\log_{10} C_n^2$ was 0.09. Notably, the model prediction was less accurate for lower and higher observed values of C_n^2, especially those below $5 \times 10^{-15} [m^{-\frac{2}{3}}]$. For C_n^2 values near the mean of the distribution of field measurements, prediction error measured by the model MSE was low. This is notable, especially when considering the difficulties posed by the near-maritime environment. In order to understand which of the 12 environmental parameters were most important in generating these predictions, the random forest model was analyzed to derive predictor importance insights. The potential change in MSE of each split was used to compute the average squared relative predictor importance I^2. Parameters with high relative predictor importance had splits with the highest average reduction in model MSE. A comparison of the estimated relative square predictor importance for each environmental parameter is given in Fig. 12.

As expected from similarity theory, atmospheric temperature and pressure were important parameters for predicting C_n^2 [1–4]. Additionally, seasonality, temporal hour weight, and most notably, the air–water temperature difference were found to be important. The potential dependence on air–water temperature difference is notable, as this relationship has been suggested by existing literature [7–11]. The relative importance of seasonality and visibility is also notable, as the seasonal dependence of C_n^2 in the near-maritime environment is not well understood. The importance of this parameter indicates that the MSE in C_n^2 in the random forest prediction was lower when the model could split based on the time of year a given observation was taken.

The potential relationship between temporal hour weight and C_n^2 was also hypothesized in prior literature [1,2]. This relationship is strong over land, where a diurnal dependence has been noted. The results in Fig. 11 indicate that, while air temperature and pressure are important in predicting C_n^2, the air–water temperature difference also contains substantial predictive power. The results also seem to indicate a stronger temporal dependence on the near-maritime environment when compared to open ocean environments.

6. CONCLUSION

Comparing the relative efficacy of a wide range of measured environmental parameters in predicting C_n^2 in a low-altitude maritime environment indicates that parameters other than absolute air temperature can be useful in minimizing model error. In concordance with current similarity and bulk atmospheric theory, air temperature and local pressure are found to have had the greatest predictive power. The analysis in Fig. 11 confirms that both of these parameters contribute to a reduction in model MSE. The results also suggest, however, that the air–water temperature difference, temporal hour weight, and seasonality are important parameters to consider when modeling C_n^2 in a low-altitude maritime environment. The analysis further suggests a weak relationship between the measured wind speed and C_n^2 in this domain; however, this effect may be captured elsewhere in the models. While relative humidity was an important parameter for minimizing model MSE, the relationship was not as strong as expected for an over-water link. This finding suggests that the maritime environment of the Severn River is distinct from longer over-water links as well as terrestrial links. The relative importance of water temperature, wind speed, and seasonality suggest that a wider range of environmental parameters, including the air–water temperature difference, temporal hour weight, and seasonality, should be considered in models for C_n^2 in the near-maritime environment. Further investigation into the relationships between environmental parameters and C_n^2 in maritime environments, as opposed to terrestrial or open-ocean environments, will improve the predictive power of optical turbulence models.

The statistical techniques used to derive measures of relative feature importance can also be used to train new models for predicting C_n^2. The demonstrated relative importance of air temperature, pressure, and air–water temperature difference suggests that machine learning could be used to more accurately predict C_n^2. Future work will further investigate the application of machine learning models in an effort to generate improved predictions of C_n^2 in the near-maritime environment.

Funding. Directed Energy Joint Technology Office; Office of Naval Research; Office of Academic Research, U.S. Naval Academy.

Acknowledgment. This work could not have been accomplished without instruction and review from Professors William Traves, Douglas Vanderwerken, Daren Creutz, Daniel Roche, and Emily Retzlaff, and CAPT Chris Ruth, USN. The authors would also like to thank the United States Naval Academy Mechanical Engineering, Electrical Engineering, Oceanography, and Mathematics Departments for their assistance in data collection, model critique, and editorial review. Finally, we would like to thank the Trident Committee and Professor Maria Schroeder.

Disclosures. The authors declare no conflicts of interest.

REFERENCES

1. R. Barrios, F. Dios, and D. Narottam, *Wireless Optical Communications Through the Turbulent Atmosphere: A Review* (InTech, 2012), pp. 3–40.
2. L. C. Andrews and R. L. Phillips, *Laser Beam Propagation Through Random Media* (SPIE, 2005).
3. T. Foken, "50 years of the Monin–Obukhov similarity theory," Boundary Layer Meteorol. **119**, 431–447 (2006).
4. G. P. Perram, S. J. Cusumano, R. L. Hengehold, and S. T. Fiorino, *An Introduction to Laser Weapon Systems* (Directed Energy Professional Society, 2010).
5. A. D. Tunick, "The refractive index structure parameter/atmospheric optical turbulence model," CN2 ARL-TR-1615 (Army Research Lab, 1998).
6. A. Tunick, N. Tikhonov, M. Vorontsov, and G. Carhart, "Characterization of optical turbulence (C_n^2) data measured at the ARL A_LOT facility," ARL-MR-625 (US Army Research Laboratory, 2005).
7. P. A. Frederickson, K. L. Davidson, C. R. Zeisse, and C. S. Bendall, "Estimating the refractive index structure parameter (C_n^2) over the ocean using bulk methods," J. Appl. Meteorol. **39**, 1770–1783 (2000).
8. P. A. Frederickson, S. Hammel, and D. Tsintikidis, "Measurements and modeling of optical turbulence in a maritime environment," Proc. SPIE **6303**, 630307 (2006).
9. S. Avramov-Zamurovic, C. Nelson, and M. Hyde, "Scintillation experiments with non-uniformly and uniformly correlated spatially partially coherent laser beams propagating underwater," J. Mod. Opt. **66**, 1998–2007 (2019).
10. R. Mahon, C. I. Moore, H. R. Burris, W. S. Rabinovich, M. Stell, and L. M. Thomas, "Analysis of long-term measurements of laser propagation over the Chesapeake Bay," Appl. Opt. **48**, 2388–2400 (2009).
11. M. Oakley, C. Nelson, C. J. Brownell, and D. Nelson, "Experimental analysis of atmospheric optical turbulence and laser beam scintillation in a near-maritime environment," in *13th Directed Energy Systems Symposium*, Virginia, USA, September, 2018.
12. J. M. Wilczak, "Large-scale eddies in the unstably stratified atmospheric surface layer. Part I: Velocity and temperature structure," J. Atmos. Sci. **41**, 3537–3550 (1984).
13. W. Bourque, C. Nelson, and D. Nelson, "Evidence and implications of differences in atmospheric optical turbulence behavior on opposite coastal environments," J. Directed Energy **6**, 187–197 (2017).
14. Google, Map of the United States Naval Academy and Surrounding Areas (2020).
15. "United States Naval Academy Oceanography Department EDMAPS online data repository," 2020, from https://www.usna.edu/Users/oceano/adavies/Research/index.php.
16. O. Vaisala, *Vaisala Weather Transmitter WXT530 Series User Guide* (2017).
17. NOAA, National Centers for Environmental Information Station Metadata, 2020, https://www.ncdc.noaa.gov/data-access/land-based-station-data/station-metadata.
18. NOAA, "National Data Buoy Center US Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service, and National Data Buoy Center," 1996, https://www.ndbc.noaa.gov/measdes.shtm.
19. NOAA, "ESRL Improved Sunrise/Sunset Calculator," 2020, https://www.esrl.noaa.gov/gmd/grad/solcalc/sunrise.html.
20. T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction* (Springer, 2009).
21. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning* (Springer, 2013), Vol. **112**, pp. 3–7.
22. C. Shalizi, "Advanced data analysis from an elementary point of view," 2013, https://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/ADAfaEPoV.pdf.
23. C. Strobl, A. L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: illustrations, sources and a solution," BMC Bioinf. **8**, 25 (2007).
24. P. Orbanz and Y. W. Teh, "Bayesian nonparametric models," in *Encyclopedia of Machine Learning* (2010), Vol. **1**.
25. L. Breiman, "Random forests," Mach. Learn. **45**, 5–32 (2001).