Hurricane Dorian (9/2/2019): "A Rogues' Gallery of the Five Category 5 Storms of 2019," 9 Jan. 2020, https://blogs.scientificamerican.com/eye-of-the-storm/a-rogues-gallery-of-the-five-category-5-storms-of-2019/. Accessed 17 Mar. 2021.

Influence of sea-surface temperatures on wind speed in the North Atlantic basin using the International Comprehensive Ocean-Atmosphere Data Set

James B. Carter

May 17, 2021

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Abstract

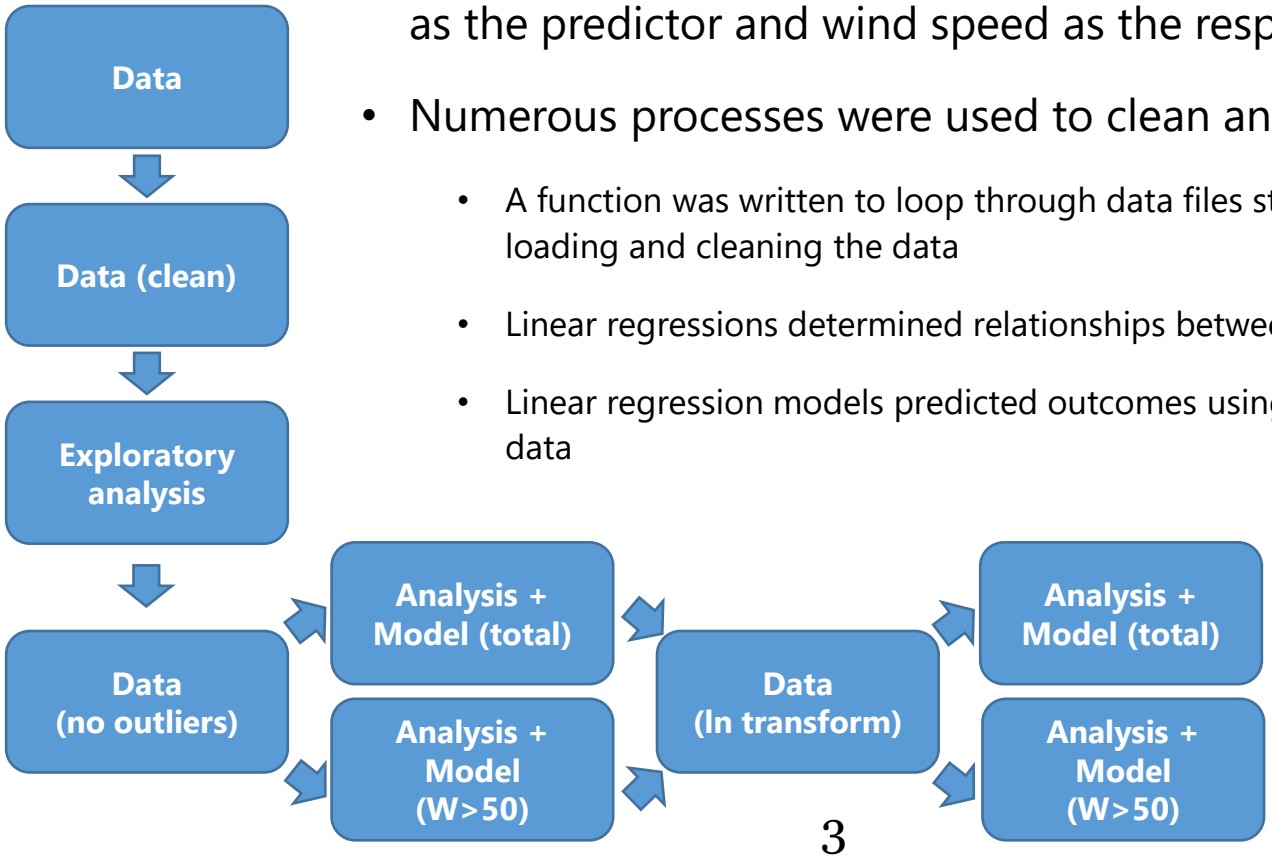
- The purpose of this project is to determine the influence of sea-surface temperature (SST) on wind speed in the North Atlantic basin
  - The period being considered is January 2000 - March 2021
  - The dataset used is International Comprehensive Ocean-Atmosphere Data Set (ICOADS)
- Linear regressions were used to determine the influence of sea-surface temperature on wind speed
  - Data was loaded on a monthly basis, cleaned, and placed in a dataframe using a function
  - The model used was a linear regression of wind speed as a function of sea-surface temperature
- The analyses indicated significant and non-significant relationships between SST and wind speed depending on the subset of data analyzed
- The ICOADS appeared to contain values that were suspect and therefore influenced interpretations

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Introduction

- The project uses data cleaning techniques and exploratory analyses to clean the data and remove outliers. It uses ordinary least squares (OLS) linear regressions to analyze the data and linear regression models to create predictive models using SST as the predictor and wind speed as the response
- Numerous processes were used to clean and analyze the data
  - A function was written to loop through data files streamlining the process of loading and cleaning the data
  - Linear regressions determined relationships between SST and wind speed
  - Linear regression models predicted outcomes using a train/test set of the data

Workflow



THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

The Problem (Context)

- Climate change leads to increasing ocean temperatures and as a result climatologists predict increases in the frequency and severity of weather events
    - Understanding the influence of SST on wind speed is necessary to predict tropical cyclones (Michaels, P. J. et al., 2006)
- Is there a relationship between SST and wind speed in the North Atlantic basin via the ICOADS between January 2000 and March 2021?
    - If so, can that relationship create a predictive model?
- The ICOADS is a large data set consisting of all world-wide ocean-atmosphere data collected with over 200 variables recorded. This project must filter out only time, latitude, longitude, SST and wind speed data for the North Atlantic basin
    - Must filter out significant outliers as some SST values fall well outside normal ranges for ocean temperatures

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Purpose of the Study

- Desired outcome of this project is to:
  - Identify a relationship between SST and wind speed in the North Atlantic basin using the ICOADS
  - Create a model capable of predicting wind speed values using SST as the predictor variable
- Michaels et al. (2006) indicated a significant relationship between SST and wind speed using the Hurricane Databases (HURDAT) data from 1982-2005.

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Project Description

- This project will determine if there is a significant relationship between SST and wind speed in the North Atlantic basin and create a model to predict wind speed from SST
  - An initial piece of the ICOADS will be loaded to establish a protocol for loading and cleaning the whole January 2000 – March 2021 dataset
  - A function will be written and used to load, clean, and append the dataset into a single dataframe
  - An exploratory linear regression will be performed on the total data
  - Outliers will then be removed
  - A linear regression analysis and model using train/test will be performed on the total data and wind speed greater than 50 m/s (W>50) data
  - The wind speed data will be natural log transformed to normalize the right skewness
  - A linear regression analysis and model using train/test will be performed on the total natural log transformed data and wind speed greater than 50 m/s (W>50) natural log transformed data

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Project Description: Hypothesis

Null hypothesis:

There is no relationship between sea-surface temperature and wind speed

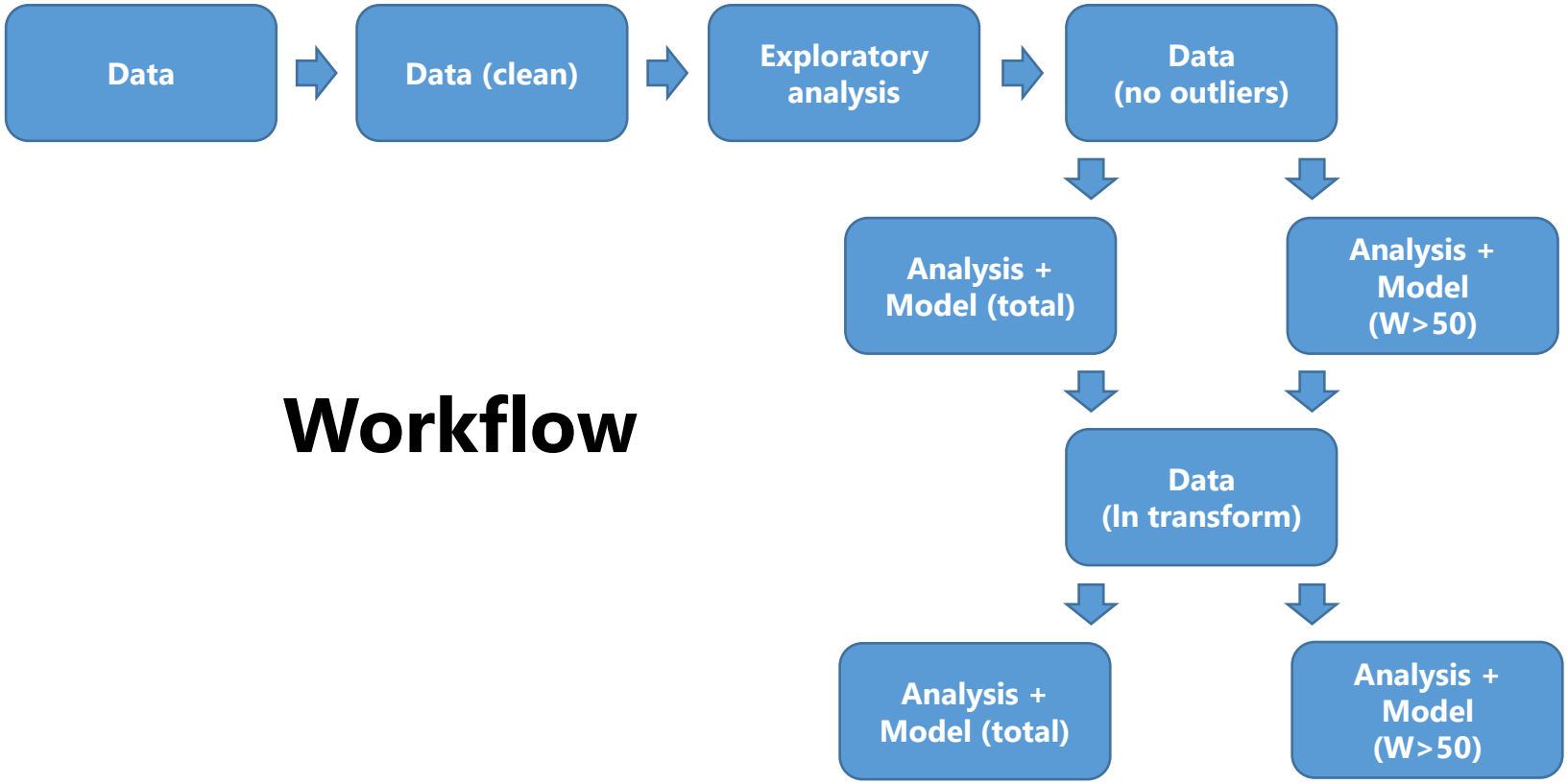
Alternate hypothesis:

There is a relationship between sea-surface temperature and wind speed

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Project Description: Workflow

Workflow



THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Project Description: Workflow

- Due to the size of the ICOADS dataset it was necessary to load one month of data to establish a protocol for cleaning the dataset
- A function was written to load, clean, and append the dataset into a single dataframe as follows:
    - Drop all columns except time, latitude, longitude, wind speed indicator, wind speed, and sea-surface temperature (SST)
    - Drop all rows with NaN values in wind speed indicator, wind speed, and SST
    - Drop all rows with 2, 5, or 6 in the wind speed indicator column
        - 2, 5, and 6 indicate estimation from unknown units or Beaufort force (categorical values)
    - Drop all rows with latitude and longitude values outside of the North Atlantic basin
    - Convert wind speed values to meters per second from knots when the wind speed indicator value is equal to 3 or 4
    - Convert wind speed to a float datatype
    - Drop wind speed indicator column because it is no longer needed
    - Append to created dataframe

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Project Description: Workflow

```python
# function to load/clean data
def read_icoads(df, icoads):
    ds = xr.open_mfdataset(icoads, parallel = True)
    data = ds.to_dataframe()
    # dropping all unnecessary columns
    data.drop(data.columns.difference(['time', 'lat', 'lon', 'WI', 'W', 'SST']), axis = 1, inplace = True)
    # dropping all rows with necessary missing data
    data.dropna(subset = ['WI', 'W', 'SST'], inplace = True)
    # dropping rows with WI = 2, 5, 6 which are inaccurate
    data.drop(data[(data['WI'] == 2)].index, inplace = True) # 2 indicates estimated from unknown units
    data.drop(data[(data['WI'] == 5)].index, inplace = True) # 5 indicates Beaufort force which is categorical
    data.drop(data[(data['WI'] == 6)].index, inplace = True) # 6 indicates estimated from unknown units and method
    # dropping rows with lat/lon outside the North Atlantic
    data.drop(data[(data['lat'] < 0.936028)].index, inplace = True)
    data.drop(data[(data['lat'] > 68.638722)].index, inplace = True)
    data.drop(data[(data['lon'] < 12.005944)].index, inplace = True)
    data.drop(data[(data['lon'] > 98.053917)].index, inplace = True)
    # convert W (wind speed) values to m/s following conversions for WI (wind indicator) categories
    # WI = 0, 1, 8 in m/s
    # WI = 3, 4 in knots
    # 1 kn = 0.5144 m/s
    data['W'] = np.where(data['WI'] == 3, data['W'] * 0.5144, data['W']) # convert WI = 3 from kn to m/s
    data['W'] = np.where(data['WI'] == 4, data['W'] * 0.5144, data['W']) # convert WI = 4 from kn to m/s
    data['W'] = data['W'].astype(float)
    # all of wind speed ('W') is now in m/s
    # deleting Wind Speed measure method - no longer useful
    del data['WI'] # deleting WI column
    df = df.append(data, ignore_index = True)
    return df
```


THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Project Description: Workflow

- After the function loaded, cleaned, and appended the dataset into a single dataframe all duplicate rows were dropped
- The time, latitude, and longitude columns were dropped
    - These values were no longer needed for the analysis

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Project Description: Workflow

- Ordinary least squares (OLS) linear regressions
  - Dependent variable: Wind speed
  - Independent variable: SST
  - OLS linear regressions performed:
    - Total data
    - Total data, no outliers
    - Wind speed > 50 m/s, no outliers
    - Total data, no outliers, log transformed wind speed values
    - Wind speed > 50 m/s, no outliers, log transformed wind speed values

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Project Description: Workflow

```
# linear regression
X = sm.add_constant(df['SST'])
y = df['W']
```

```
#Linear regression
model = sm.OLS(y, X)
results = model.fit()
print(results.summary())
```

OLS Regression Results

Dep. Variable:	W	R-squared:	0.048
Model:	OLS	Adj. R-squared:	0.048
Method:	Least Squares	F-statistic:	1.734e+05
Date:	Mon, 17 May 2021	Prob (F-statistic):	0.00
Time:	05:53:18	Log-Likelihood:	-9.0617e+06
No. Observations:	3437087	AIC:	1.812e+07
Df Residuals:	3437085	BIC:	1.812e+07
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.2737	0.004	1484.867	0.000	6.265	6.282
SST	-0.0786	0.000	-416.427	0.000	-0.079	-0.078

Omnibus:	2152391.950	Durbin-Watson:	1.378
Prob(Omnibus):	0.000	Jarque-Bera (JB):	109101727.685
Skew:	2.360	Prob(JB):	0.00
Kurtosis:	30.194	Cond. No.	52.0

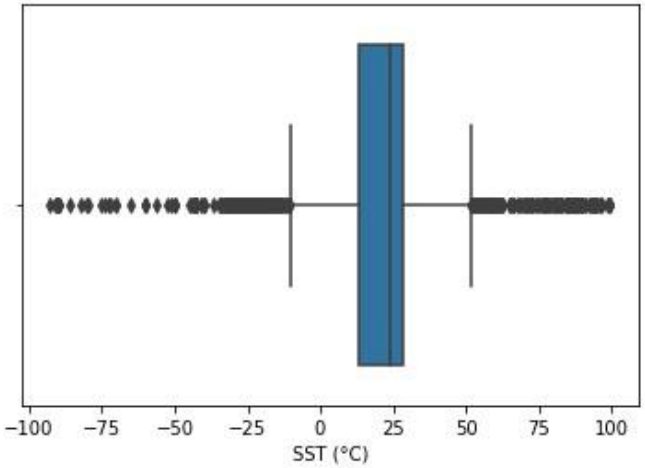
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

- OLS linear regression of total data

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Project Description: Workflow

```
# finding outliers - boxplot
sns.boxplot(x = df['SST'])
plt.xlabel('SST ('+deg+'C)')
plt.show()
```



- Outliers for SST were determined by using a boxplot
- The interquartile range (IQR) was used to remove outlier values

```
# finding outliers - IQR
Q1 = df['SST'].quantile(0.25)
Q3 = df['SST'].quantile(0.75)
IQR = Q3 - Q1
print(IQR)
```

```
15.600000381469727
```

```
# df with outliers removed
df_o = df[((df['SST'] > (Q1 - 1.5 * IQR)) & (df['SST'] < (Q3 + 1.5 * IQR)))]
df_o
```


THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

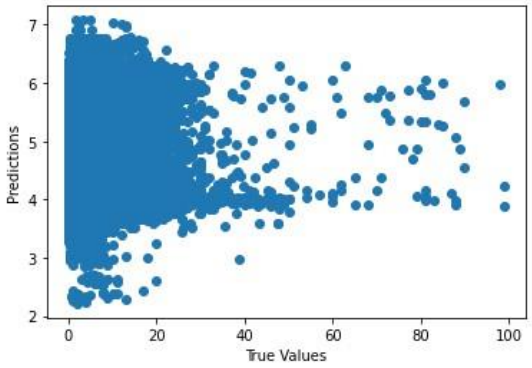
Project Description: Workflow

- Ordinary least squares (OLS) models
  - Dependent variable: Wind speed
  - Independent variable: SST
  - OLS linear models created:
    - Total data, no outliers
    - Wind speed > 50 m/s, no outliers
    - Total data, no outliers, log transformed wind speed values
    - Wind speed > 50 m/s, no outliers, log transformed wind speed values

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Project Description: Workflow

```
# train/test linear model without outliers
X = df_o[['SST']]
y = df_o['W']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2)

lm = linear_model.LinearRegression()

model = lm.fit(X_train, y_train)
pred = lm.predict(X_test)

plt.scatter(y_test, pred)
plt.xlabel('True Values')
plt.ylabel('Predictions')
plt.show()
```



- OLS model using train/test set from total data (no outliers)

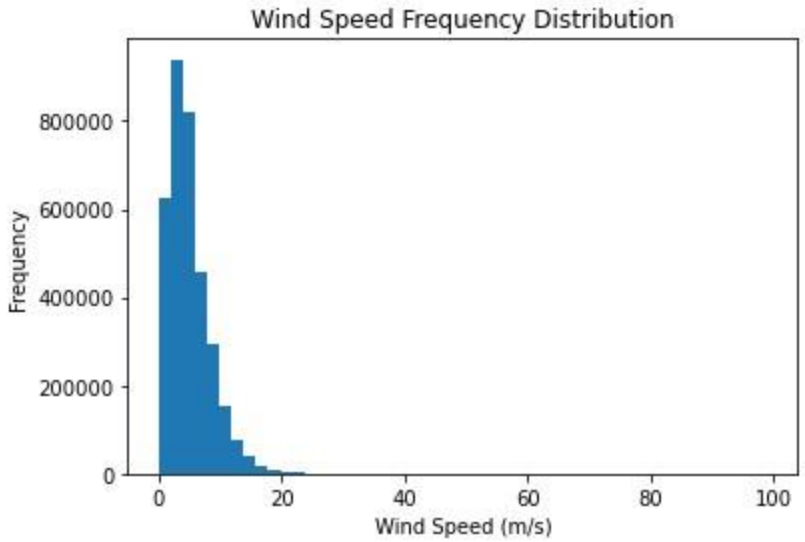
```
# MAE, MSE, RMSE without outliers
mae = metrics.mean_absolute_error(y_test, pred)
mse = metrics.mean_squared_error(y_test, pred)
rmse = sqrt(metrics.mean_squared_error(y_test, pred))

print("MAE: {}".format(mae))
print("MSE: {}".format(mse))
print("RMSE: {}".format(rmse))
```

```
MAE: 2.508607470264717
MSE: 11.274650157949564
RMSE: 3.357774584147894
```


THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Project Description: Workflow

```
# frequency distribution of W
plt.hist(df['W'], bins = 50)
plt.title('Wind Speed Frequency Distribution')
plt.xlabel('Wind Speed (m/s)')
plt.ylabel('Frequency')
plt.show()
```



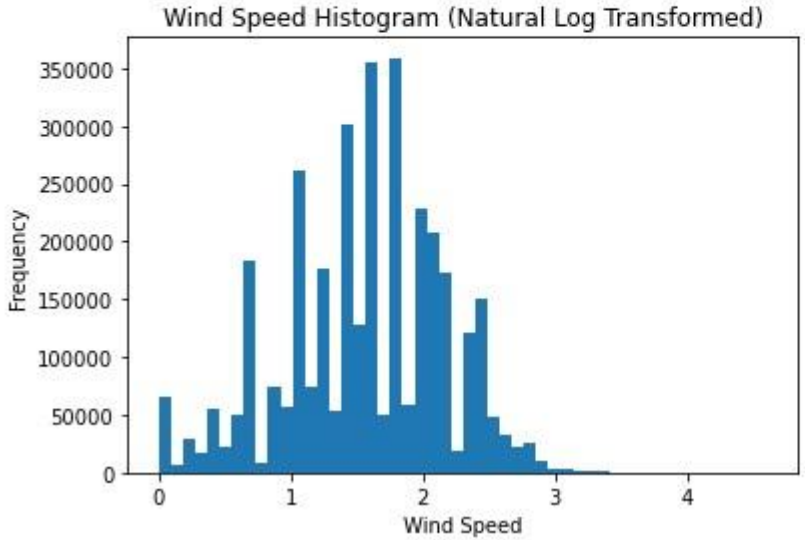
- Frequency distribution of wind speed shows right skewness

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Project Description: Workflow

```
# creating log transformed dataframe
df_log = pd.concat([df_o['SST'], np.log(df_o['W'].add(1))], axis = 1, keys = ['SST', 'W'])
df_log
```

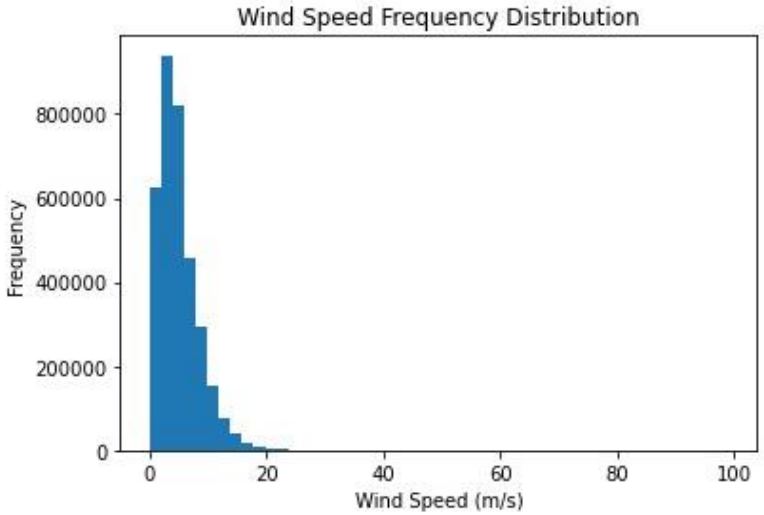
```
# frequency distribution of W from log transformed dataframe
plt.hist(df_log['W'], bins = 50)
plt.title('Wind Speed Histogram (Natural Log Transformed)')
plt.xlabel('Wind Speed')
plt.ylabel('Frequency')
plt.show()
```

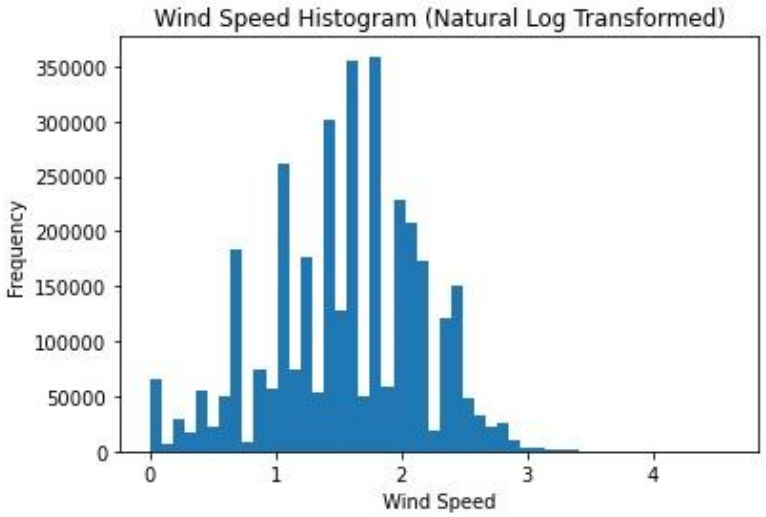


- Wind speed was natural log transformed to normalize wind speed values due to the right skewness
- The natural log transformation was performed by taking the natural log of the wind speed value plus one
  - ln (wind speed + 1)

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Project Description: Workflow





- Frequency distribution histograms of wind speed before and after the natural log transformation

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Project Description: Methods

- Function to load/clean ICOADS
  - A function was written to load, clean, and append the ICOADS into a single dataframe on a monthly basis because the ICOADS from January 2000 – March 2021 is too large to load into a dataframe and clean manually in Python
- Linear regressions and models
  - Ordinary least squares was selected because the purpose of this project is to predict wind speed as a function of SST
- Plots
  - Histograms were used for frequency distributions of SST and wind speed
  - Scatterplots were used to plot data and visualize the relationship between SST and wind speed
  - A boxplot of SST was used to determine outliers for removal
- Outliers
  - IQR was used to remove outliers for SST to prevent bias in removing SST values outside normal ocean temperatures
- Value normalization
  - Wind speed was natural log transformed to normalize wind speed values due to the right skewness

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Project Description: Analysis

- Ordinary least squares (OLS) linear regressions
  - Total data
    - Significant relationship ($R^2 = 0.0048$, $P < 0.001$), negative
  - Total data, no outliers
    - Significant relationship ($R^2 = 0.0048$, $P < 0.001$), negative
  - Wind speed > 50 m/s, no outliers
    - Nonsignificant relationship ($R^2 = 0.0000$, $P = 0.792$)
  - Total data, no outliers, log transformed wind speed values
    - Significant relationship ($R^2 = 0.0045$, $P < 0.001$), negative
  - Wind speed > 50 m/s, no outliers, log transformed wind speed values
    - Nonsignificant relationship ($R^2 = 0.000$ $P = 0.957$)

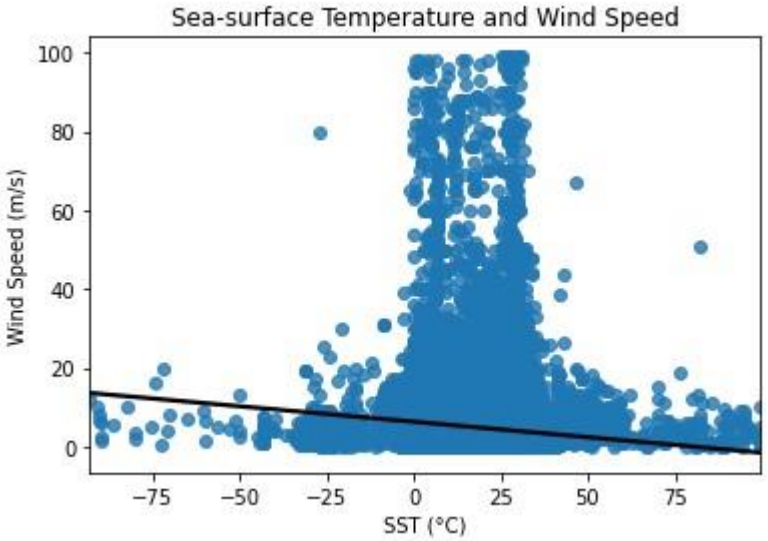
THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

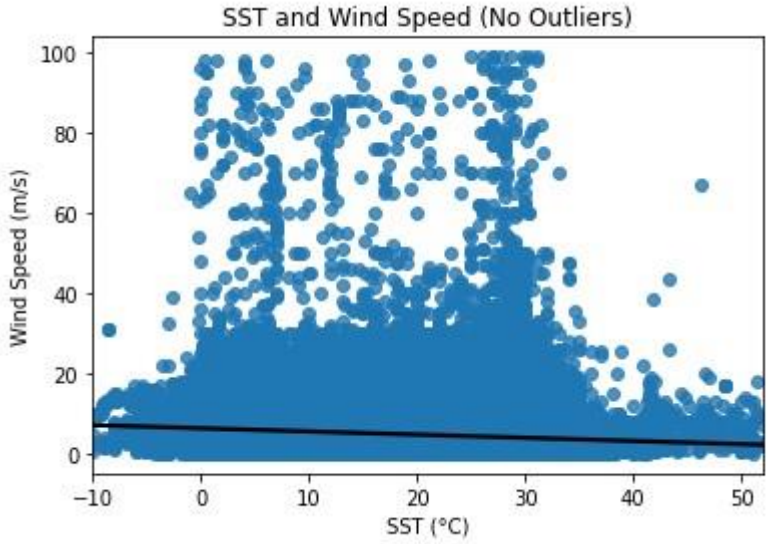
Project Description: Analysis

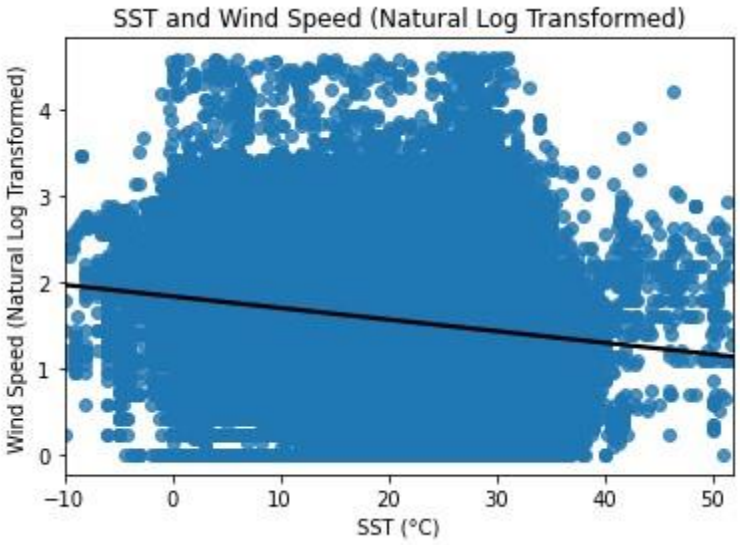
Data	n	R2	P-value	Coeff	MAE
Total	3437087	0.048	< 0.001	-0.0786	n/a
Total, no outliers	3436097	0.048	< 0.001	-0.0793	2.51
W > 50, no outliers	332	0.000	0.792	0.0199	12.94
Total, no outliers, log trans	3436097	0.045	< 0.001	-0.0134	0.478
W > 50, no outliers, log trans	332	0.000	0.957	0.000	0.187

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Project Description: Analysis



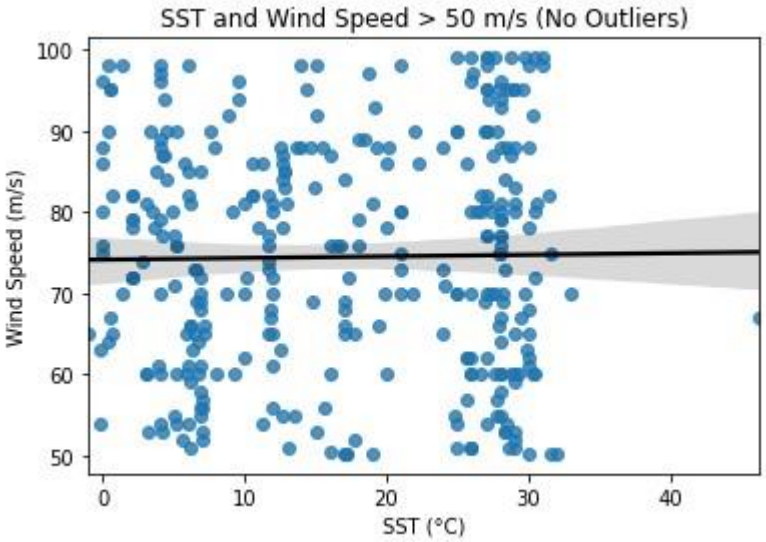


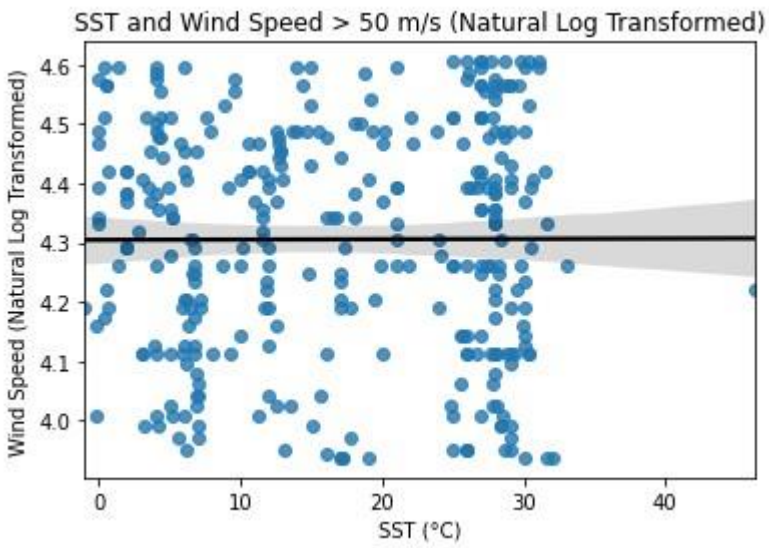


- Regression line shows a negative relationship in all three figures of the total data

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Project Description: Analysis





- Regression line shows no relationship in both figures of the wind speed greater than 50 m/s data

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Project Description: Analysis

- Ordinary least squares (OLS) models
  - Total data, no outliers
    - Minimal error, MAE = 2.51
  - Wind speed > 50 m/s, no outliers
    - Large error, MAE = 12.94
  - Total data, no outliers, log transformed wind speed values
    - Minimal error, MAE = 0.478
  - Wind speed > 50 m/s, no outliers, log transformed wind speed values
    - Large error, MAE = 0.187

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Project Description: Analysis

Data	n	R2	P-value	Coeff	MAE
Total	3437087	0.048	< 0.001	-0.0786	n/a
Total, no outliers	3436097	0.048	< 0.001	-0.0793	2.51
W > 50, no outliers	332	0.000	0.792	0.0199	12.94
Total, no outliers, log trans	3436097	0.045	< 0.001	-0.0134	0.478
W > 50, no outliers, log trans	332	0.000	0.957	0.000	0.187

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Project Description: Results

- Significant relationship between SST and wind speed of total dataset
  - Relationships were all negative
  - All analyses using total data were significant
    - These significant relationships were likely due to the high number of observations in the dataset
- Nonsignificant relationship between SST and wind speed of wind speed greater than 50 m/s dataset
- Models using total data predicted wind speed using SST with minimal errors
- Models using wind speed greater than 50 m/s data predicted wind speed using SST with large errors
- Contrasts with Michaels et al. (2006) where SST had a positive influence on wind speed

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Conclusion

- Analyses and models for total data indicate significant negative relationship between SST and wind speed in North Atlantic basin
    - Contrasts with Michaels et al. (2006) which indicated a significant positive relationship
    - All analyses performed on wind speed greater than 50 m/s were nonsignificant
    - Models using total data were more accurate at predicting wind speed from SST than using the wind speed greater than 50 m/s data
- Learning from the study
    - The ICOADS has numerous datapoints outside expected values for variables
    - Loading and cleaning the ICOADS in preparation for the analyses and models took a significant amount of time – it required 10 hours of processing to load and clean all the data
    - The significant relationship indicated for the total data was likely due to the high n of the dataset
- Future directions
    - Determine why ICOADS has high wind speed values coupled with low SST values
    - Use ICOADS to replicate statistical analyses performed by Michaels et al. (2006)
        - Identifying the data associated with tropical cyclones from the previous seven days of maximum wind speed

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

Bibliography

- Deisenroth, Marc P., Faisal A. A., Ong, Cheng S. 2020. *Mathematics for Machine Learning*. Cambridge University Press.
- Michaels, Patrick J., Knappenberger, Paul C., Davis, Robert E. Sea-surface temperatures and tropical cyclones in the Atlantic basin. 10May 2006. Geophysical Research Letter, Vol. 33, Issue 9. https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2006GL025757. Accessed 02 005 2021.
- National Oceanic and Atmospheric Administration, ICOADS Release 3.1 (R3.1), September 2016,https://rda.ucar.edu/datasets/ds548.0/docs/R3.0-imma1_short.pdf. Accessed 27 004 2021.
- Research Data Archive/Computational and Information Systems Laboratory/National Center for Atmospheric Research/University Corporation for Atmospheric Research, Physical Sciences Laboratory/Earth System Research Laboratory/OAR/NOAA/U.S. Department of Commerce, Cooperative Institute for Research in Environmental Sciences/University of Colorado, National Oceanography Centre/University of Southampton, Met Office/Ministry of Defence/United Kingdom, Deutscher Wetterdienst (German Meteorological Service)/Germany, Department of Atmospheric Science/University of Washington, Center for Ocean-Atmospheric Prediction Studies/Florida State University, and National Centers for Environmental Information/NESDIS/NOAA/U.S. Department of Commerce(2016): International Comprehensive Ocean-Atmosphere Data Set (ICOADS) Release 3, Individual Observations. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. Dataset. https://doi.org/10.5065/D6ZS2TR3. Accessed 27 004 2021.
- Shafer, Douglas S., Zhang, Zhiyi. 2012. *Introductory Statistics*. Saylor Foundation.

THAYER SCHOOL OF ENGINEERING AT DARTMOUTH

THAYER SCHOOL OF
ENGINEERING
AT DARTMOUTH