Contract No:

This document was prepared in conjunction with work accomplished under Contract No. 89303321CEM000080 with the U.S. Department of Energy (DOE) Office of Environmental Management (EM).

Disclaimer:

This work was prepared under an agreement with and funded by the U.S. Government. Neither the U.S. Government or its employees, nor any of its contractors, subcontractors or their employees, makes any express or implied:

1) warranty or assumes any legal liability for the accuracy, completeness, or for the use or results of such use of any information, product, or process disclosed; or
2) representation that such use or results of such use would not infringe privately owned rights; or
3) endorsement or recommendation of any specifically identified commercial product, process, or service.

Any views and opinions of authors expressed in this work do not necessarily state or reflect those of the United States Government, or its contractors, or subcontractors.

Title of Project
The Application of Machine Learning Techniques to Meteorological Forecasting

Project Start and End Dates
Project Start Date: October 1, 2020
Project End Date: September 30, 2021

Project Highlight
No more than one or two sentences to highlight the impact for the nation and overview the technology in "layman's terms." Please do not use acronyms. Should include one image that reflects the key achievement or importance of the project.

A labeled data set comprising the two predictands (fog and the sea breeze) and related meteorological predictors has been developed, and several machine learning algorithms have been trained to find relationships between the predictors and the subsequent occurrence of the predictand events.

Project Team
Principal Investigator: David Werth
Team Members: Thomas Danielson, Elizabeth LaBone, Stephanie Gamble, Eric Hoar, Stephen Noble, Brian Viner
External Collaborators (all external collaborators and their respective organizations that participated in this project): Ajay Kumar Gogineni, Brian Mayer, Naren Ramakrishnan; Sanghani Center for Artificial Intelligence and Data Analytics

Abstract
Abstract in 150 words or less that describes the impact of the research in plain English.

Fog and inland-penetrating sea-breezes occur often at SRS and have a strong impact on site operations. Site personnel therefore require accurate forecasts of these events, but both are difficult to forecast using traditional techniques. Our goal is to apply machine learning (ML) techniques to the problem of forecasting fog and the sea breeze at the Savannah River Site. We apply several such techniques - decision trees, regression, and a series of classification/regression techniques – and train them using the large datasets collected by our group at SRS and from external organizations that maintain databases of regional meteorological variables.

The proposed methods have been developed and have shown skill in predicting fog when compared to existing forecasting tools. We will further advance the use of these algorithms for weather forecasting, taking them towards becoming wide-spread forecasting tools.

Objectives

- Bulleted list of specific project objectives
- Collect fog and sea breeze data from onsite sources, and label all data with its respective category
- Develop Random Forest algorithm
- Develop Ordinal Regression algorithm
- Develop Neural Network algorithm

REVIEWS AND APPROVALS

1. Authors:

_____

_____

_____

Name and Signature Date

2. Technical Review:

_____

Name and Signature Date

3. PI's Manager Signature:

_____

Name and Signature Date

4. Intellectual Property Review:

This report has been reviewed by SRNL Legal Counsel for intellectual property considerations and is approved to be publicly published in its current form.

SRNL Legal Signature

_____

Name and Signature

Introduction

Present the background for the work and explain what work was done in this project and its significance in terms of advancing the state of science. Should be 3 paragraphs maximum. *If an image, chart, figure is included, mention the figure and number but include the image, chart or figure at the end of this document (see below).*

Our objective is to apply multiple machine learning (ML) algorithms to the problem of forecasting fog and the sea breeze at the Savannah River Site. This task comprises collecting and organizing data from onsite and offsite sources, labeling periods during which these events occurred, training the algorithms to recognize the conditions that preceded their occurrence, then testing them using new data to verify that the forecasts are accurate. For both applications, the ML techniques are to identify patterns and correlations from this labeled data, allowing classification of future predictors to the proper category (e.g., fog or no fog will occur tomorrow). The methods we have selected have been proven useful for a variety of applications, and by developing multiple methods, we ultimately plan to assess how an ensemble of predictions can be used to estimate the uncertainty of the forecasts.

For fog, we put together a large, labeled dataset of predictors and associated visibility readings, which serve as a proxy for fog formation. The selected algorithms were each applied to the dataset and demonstrated skill at predicting fog on the subsequent day. Currently, the existing Model Output Statistics (MOS) model yields a visibility forecast, and serves as the baseline forecast upon which the ML must improve. Several of the methods have significantly improved on this forecast, highlighting their benefit.

Fog and the sea breeze are affected by small-scale variations in temperature, moisture, etc., that are not reproduced well by existing computer weather models, so this project has the potential to improve the forecasting of these complex weather phenomena. The ML algorithms are executed using existing software, and the developed modeling system could be i) applied at any location with sufficient training datasets, or ii) applied to predict other weather phenomena with a clearly defined predictand and data to do the training.

Approach

Explain the approach used to conduct the research. Should be 1 - 2 paragraphs. This narrative should not outline in detail the steps that will be taken but should describe the overarching approach for achieving the objectives of the project. *If an image, chart, figure is included, mention the figure and number but include the image, chart or figure at the end of this document (see below).*

The predictors comprised both observations and predictions from existing weather models. Data from our site towers (including visibility data, a proxy for fog formation), radar maps (to identify sea breezes) and archived forecasts from two weather models were collected. These were composited into a single dataset for each day by averaging or selecting minima or maxima (e.g., the observed morning low temperature and the forecasted average dewpoint for the next day from the existing weather forecast model), creating a set of inputs and desired outputs (e.g., the next day's minimum visibility onsite between 0600UTC and 1200UTC) for the machine learning algorithms to train on. To identify the inputs that best serve as precursors to fog, the team developed a Pearson correlation matrix (Fig. 1), which

indicates that several model and observed variables correlate with the site visibility values on the subsequent day.

Several methods – the random forest, logistic regression, and a series of classification/regression models - were applied to the data assuming both a continuous predictand (the measured visibility value) or a binary predictand (setting a visibility threshold for 'fog' or 'no fog'), and also using various sampling methods and variable selection methods. Equitable threat scores (ETS, for which higher values indicate more accurate forecasts) and other similar metrics are used to validate the forecasts.

Accomplishments

Brief description of accomplishments to date in bullet form. Whenever possible, accomplishments should be stated quantitatively, as in the examples shown below, and indicate the contribution to meeting the objectives, as well as the magnitude of the improvement over past work.

- When using a random forest with a binary predictand, the best classification model has an accuracy (#correct/#attempts) of 53%. Predicting 'no fog' is more accurate than fog predictions - the model has a 96% precision (True Positives / (True Positives + False Positives), where 'positive' means no fog) when predicting no fog, but only 6% precision when predicting fog (where 'positive' now means fog), because of a high number of false positives. False positives are preferred for this particular forecasting system, though continuing efforts seek to improve the performance metrics of the model.

- When we use a random forest algorithm to forecast the actual values of visibility, the model results indicate an ability to accurately train on known data with limited ability to predict future data due to limited fog data available (Fig. 2). The plot shows the model is able to accurately predict the visibility for data points utilized in the training dataset but is limited in its predictive capability for the testing dataset.

- The number of foggy days is much lower than the number of clear days, which can frustrate attempts to train an ML algorithm to predict fog. To compensate for this, the logistic regressions were trained using sampling methods to balance the number of days with and without fog, either by i) under sampling, ii) over sampling, or iii) creating synthetic data. These performed better than a logistic regression trained with the unmodified data. The highest ETS values for two data sets were 0.29 for a data set using data from the current MOS forecast model as predictors. This is a large improvement on the MOS ETS of 0.18.

- As a precursor to the neural net techniques, we applied a series of other classification/regression techniques that did very well (exceeding the 0.18 MOS ETS) for predicting continuous visibility or when a binary predictand was used (Table 1).

Future Directions

Describe what will be done next and future anticipated accomplishments. Bulleted lists are recommended. *If an image, chart, figure is included, mention the figure and number but include the image, chart or figure at the end of this document (see below).*

- The team is working toward fine tuning the datasets and the machine learning models. A key challenge in the datasets is the presence of missing values, class imbalance (many more non-fog days than foggy days), and a variety of environmental conditions represented by the various SRS towers at different times of the year. Currently, more sophisticated approaches are being sought that train separate random forest models both spatially and temporally, which should improve the predictive power.

- Applying the neural network models.

- Further assessment of inputs to determine which ones have the most impact on fog prediction and creating new features as appropriate.

- We also have access to the 2D surface pressure information and are exploring the use of this, by considering it as an image, to predict fog.

FY 2021 Peer-reviewed/Non-peer reviewed Publications

List all peer-reviewed/non-peer-reviewed manuscripts (published, accepted for publication or submitted/under review) written from this work during the year. Please specify if SRNL is the primary research organization for this manuscript (first author and/or corresponding author). Do not include internal reports or this annual summary report. ACS Format

Intellectual Property

List all invention disclosures, copyright disclosures, patent applications, and patents granted.

Total Number of Post-Doctoral Researchers

List total number of post-doctoral researchers involved even through subcontractors. Indicate their name(s) and if they performed research on or off site.

Eric Hoar
Stephanie Gamble

Total Number of Student Researchers

List total number of students supported by the project, including those involved through university subcontracts. Indicate their name(s), if they were undergraduate or graduate students, and if they performed research on or off site.

Ajay Kumar Gogineni, Sanghani Center for Artificial Intelligence and Data Analytics

Include all images, charts and figures with captions, as shown below.

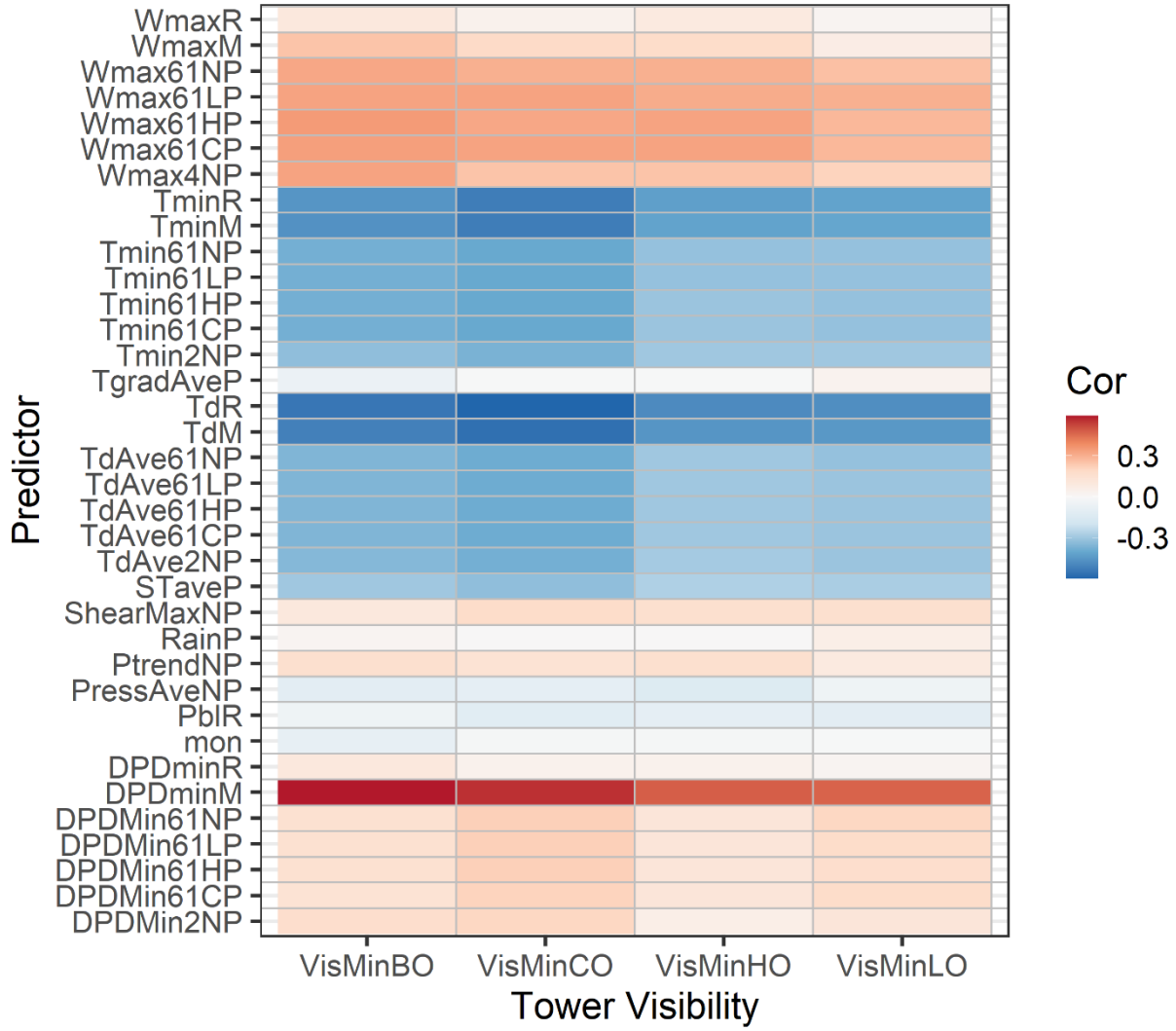

Figure 1 Correlation matrix showing the Pearson correlation coefficients between the observed predictor variables (left column) and the next-day observed morning minimum visibility at the four site towers (bottom row - B, C, H, and L).

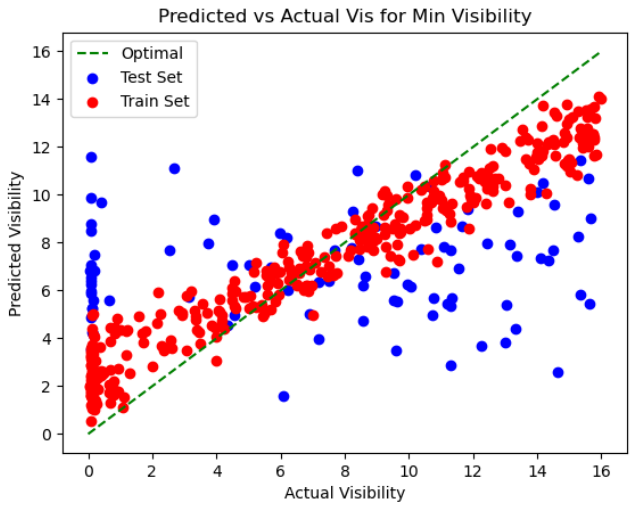

Figure 2: Plot of predicted visibility vs actual visibility in kilometers illustrating the 1-day predictive capability of the Random Forest algorithm.

	Binary classification	Continuous visibility values
Elastic Net	0.16	0.47
Cat Boost	0.467	0.35
Light GBM	0.467	0.38

Table 1 Equitable threat scores (ETS) for various classification/regression methods when forecasting next-day fog.