Micro Climate Prediction Utilising Machine Learning Approaches

Anastasia Eleftheriou
Center for Technology Research & Innovation
Limassol, Cyprus
ael@cetri.net

Petros Karvelis
Computer Technology Institute & Press "Diophantus"
Patras, Greece
pkarvelis@gmail.com

Kostas Kouvaris
Center for Technology Research & Innovation
Limassol, Cyprus
k.kouvaris@cetri.net

Chrysostomos Stylios[1,2]
[1]Computer Technology Institute & Press Diophantus" Patras, Greece, stylios@cti.gr [2]Technological Educational Institute of Epirus, Arta, Greece stylios@teiep.gr

***Abstract*— The current study focuses on predicting the wind speed on short-term weather conditions for maritime vessels weather station. Several machine learning models were developed for different forecasting horizons and their efficiency for this study was evaluated across a number of metrics. A regression machine learning algorithm was chosen for sea trials on Lincoln vessels.**

Keywords*— *wind, forecasting, machine learning models

I. Introduction

The importance of maritime for the world economy is increasing. Europe is constantly in search for new solutions, products and services that will enable European industry to position itself as a key competitor and promote new business models, in order to create added value to the vessels.

The LINCOLN[1] project addresses this paradigm shift through a holistic perspective, where starting from the design phase and the new vessels concepts, it takes care of the final added market value. The LINCOLN project aims to propose added-value specialized vessels able to run requested services for Marine Aquaculture, Ocean Energy, Coastal Monitoring, Control, Surveillance and Rescue sectors in the most effective, efficient, economical valuable and eco-friendly way. In particular, innovative vessels are designed according to lean design tools (such as KbeML – Knowledge Based Engineering Modelling Language) and methodologies (such as SBCE – Set Based Concurrent Engineering), through an integrated IoT (Internet of Things) platform, able to provide knowledge and future services to the maritime sector actors. Specifically, the IoT platform consists of a physical part made of the following dedicated black boxes; i) the Universal Marine Gateway (UMG) black box for vessel prototypes and ii) the Marine Gateway (MG) black box for commercial versions. These black boxes host sensors and are connected to other vessels systems, such as the on-board weather station. The data gathered by the sensors are then collected and sent to a cloud system where they are analyzed and processed through specific algorithms. The generated information is published through a web interface to different users' categories, like designers, shipbuilders, suppliers and maintenance companies.

In order to predict changes in weather conditions, there are several meteorological parameters (such as air pressure and direction of wind), which can indicate the upcoming fronts. Specifically, the air pressure strongly influences the changes in wind speed, temperature and precipitation level. In moderate climates, which are observed in Europe, slow decrease of air pressure indicates that a low-pressure area is approaching. This brings clouds, precipitation, cold breeze in the summer time and temperature increase in the winter. In low pressure areas often high winds and warm air occur that results in creation of clouds and precipitation. It means that expected temperatures can be lower, more wind can be expected, without solar influence.

Wind direction and wind speed are factors with also big impact on weather, because they affect the surface water. They can cause higher or lower evaporation of water, creation of clouds and precipitation [1]. Also, wind direction and speed determine directions of fronts and how fast they are coming. For these reasons, this work focuses on the prediction of wind speed for LINCOLN vessels.

In the case of weather predictions for local weather conditions, the issue is more challenging than on the country level [2], [3]. Usually, the area for which the weather conditions need to be forecasted should be divided into squares. The smaller the squares, the better. In every square, some points should be where current parameters of the atmosphere like temperature, pressure, wind direction and speed, air humidity, amount of precipitation will be measured. The rest of the inputs like information about weather condition outside the area of local measurement, which are defined as boundary conditions, and information about fronts, clouds could be downloaded from outside meteorological stations, weather services, and bigger organizations, such as the Global Forecast System.

There are 4 main types of common weather forecasts: a) nowcast – predicts the current state of the weather, b) short-term – weather conditions are predicted for the next 72 hours maximum, c) medium-range – weather conditions are predicted for a period of 3-7 days and d) long-term – weather conditions are predicted for a period from one week to months, or even years.

Beside current conditions, the forecast system should also be supported by historical data about local weather conditions. The current study focuses on predicting the wind speed on short-term weather conditions.

II. Methods

A. *Using the causaLens Platform – Automated Machine Learning for Time-Series Predictions*

For the purpose of this work, the causaLens platform [4] was used to discover a machine learning model that predicts the wind speed.

This platform automates the process of discovering prediction models for time-series data. Given historical data

[1] LINCOLN: Lean Innovative Connected Vessels Project www.lincolnproject.eu. Horizon 2020 research and innovation program.

as input, the platform is capable of autonomously constructing an optimal prediction model that can subsequently be integrated in any workflow and deployed on any machine or device.

In a broad sense, a machine learning model consists of features extracted from the data, the algorithm and the parameters of the chosen algorithm. Theoretically, there are infinite set of models possible for a given dataset, therefore, researchers and data scientists rely on experience to choose a model. However, the causaLens platform automates this process and discovers models as fast as a collection of data scientists working together. The discovery process is complex and computationally intensive and it is not the subject of this paper.

For the current study, different models were developed for different forecasting horizons. In the following sections, we present results along with the technical specification for the resulting models developed for wind speed predictions. The hyper-parameters of each model have been optimized to provide predictive models with high generalisation performance. Lastly, a feature selection process has been incorporated so that to alleviate the problem of over-fitting. Given the vast number of features that can be considered in a time series model, robust feature selection methodology is essential to reduce the variance of the models, and thus avoid fitting.

The data was split into Training, Validation and Testing and time-series aware Cross-validation was performed to ensure robustness of the chosen models.

B. Portweather Prediction Model

In the LINCOLN project some of the major problems that we had to tackle concerning the weather prediction were the following:

- the ship had no internet connection to get or send data like local weather parameters from other sources like ports etc.
- the machine learning algorithm had to be implemented using limited memory.

For these two reasons we tested a regression machine learning algorithm and we named it Portweather, that has small memory footprint and was able to adaptively and online learn from data gathered from the weather station that was on board of the ship. This work is an expansion of our previous work presented in [2], [3] and for the same data.

Such a machine learning algorithm is the Linear Regression (LR) with a Stochastic Gradient Descent (SGD) update of its parameters. When it comes to LR we forecast wind speed y as a linear function of the following meteorological parameters Temperature (T) x_1, Dew Point Temperature (DPT) x_2, Humidity (H) x_3, Wind Direction (WD) x_4, Pressure (PR) x_5, Precipitation (PC) x_6 and Wind Speed (WS) x_7:

$$y(x^i) = a_0 + a_1 x_1 + \ldots + a_m x_m = \sum_{j=1}^{m} a_j x_j \,, \qquad (1)$$

where $x_0 = 1$ and the $\{a_i \mid i = 1,\ldots,\mathrm{m}\}$ are the parameters of the linear function.

When it comes to learning the parameters of the LR model a standard way would be to implement the Gradient Descent (GD) algorithm where the gradient defined as

$$\nabla J(\alpha) = {}^{1}\!/_{N}\left(y^T - aX^T\right)X \,, \qquad (2)$$

where N is the number of training samples x^i, $i = 1,\ldots,N$, $J(\alpha) = \frac{1}{2}\sum_{i=1}^{n}\left(y(x^i) - \mathrm{y}^i\right)^2$ and y^i, $i = 1,\ldots,N$ is the real recorded value of the wind speed from the training set.

Using the GD we can often have slow convergence because each iteration requires calculation of the gradient for every single training example mapping. Another way to learn the parameters of the LR is to use the Stochastic Gradient Descent method where we update the parameters each time by iterating through each training example making ideal for online learning and microcontroller use. This way we can get estimates despite the fact that we've done less work by defining the gradient as:

$$\nabla J(\alpha)_i = {}^{1}\!/_{N}\left(y^i - a^T X_i\right)X_i \,. \qquad (2)$$

III. DATA USED

A. Corfu Data Set

The data set that we have collected consisted of 15-minute records of seven parameters: Temperature (T), Dew Point Temperature (DPT), Humidity (H), Wind Direction (WD), Pressure (PR), Precipitation (PC) and Wind Speed (WS). They have been acquired by a local weather station that was installed by the Corfu Port Authority [5] and most of its operations are presented through a WebGIS application [6]. The gathering of the information ranges from January 1, 2017 until March 31, 2017.

B. Data pre-processing

Prior to the use of the two platforms, we had to ensure that we have high quality data and that our data are in the right format that is compatible. To do so, a multi-step data cleaning process has been developed.

C. Time series format

The data should be formatted in a tabular format, where rows correspond to different entries and columns corresponds to different features. Each row should contain i) an indication of the date and the time for the given entry, i.e., Year, Month, Day, Hour, Minute, ii) the values of the features that are used as input to the model, e.g., wind gust, pressure etc. and iii) the value of target variable used as an output of the model. The data should be ordered in a strict chronological order, so we make sure that there are not duplicated entries in our data.

D. Filling in missing values

Data gaps often appear in climatology time series data. In our data, extreme numbers were used to indicate that the data

for a given variable were missing. However, different numbers were used for each variable based on the data type. To deal with this, we translated all these entries from Not a Number (NaN) type that is compatible with and recognized as a format by Python environment.

Filling these missing values is essential, since having a continuous (without missing values) dataset is prerequisite in training our weather forecast models. To ensure data continuity and the generality of our methods, an efficient fill in the missing values methodology has been deployed; namely the Drift Method [7]. According to the drift method, missing values can be filled with estimates, using a simple forecasting method where the forecast values are equal to the last known values plus the average change over time (i.e., drift) in the historical data. Note that this method is equivalent to extrapolating into the future by drawing a line from the first observation to the last one.

E. Data normalisation

In theory, it is not necessary to normalize the numeric data that are to be used as predictors. In practice, however, normalizing input variables tends to lead to better predictors by facilitating a more efficient training process. A change in a parameter of the model will hence have a greater effect on the input values that is characterized by larger magnitude. To assist the development of high performance machine learning estimators, we perform data standardization for each feature via mean removal and variance scaling [8].

Following the transformation above, each feature follows a Gaussian distribution with zero mean and unit variance. Note that the individual transformation functions for all features are stored and subsequently used when the model is run for new data samples.

F. Training and test sets

Evaluating forecast accuracy based on how well the model fits past historical data is invalid. The accuracy of the forecasts can only be sufficiently determined by how well the learning model performs on new instances that have not been used in the training phase. In the current study, we keep 30% of the available data for testing. Because of the time dependency in the data, we preserve the order of the data during the split, by reserving the 30% most recent observations as test set, while keeping the first 70% of the data for training.

IV. RESULTS

A. Causalens Platform

We used the causaLens platform to automatically discover predictive models for different horizons i.e. between 1 hour and 24 hours. The platform evaluated hundreds of thousands models during the discovery process. The top performing model was selected on the validation dataset and not on the training dataset. This ensures that the performance of our models indicates how well the models have captured the true underlying structure of the problem. We present the results for 1 and 24hours forward prediction. The forecasting accuracy gradually gets reduced while forecasting time increases.

For each forecasting horizon and each target variable of interest, a separate model was discovered as the predictors and the parameters of the model are expected to vary.

In the case of climate predictions, statistical models tend to perform well on short term horizons. There is a point at which the performance of a physical model exceeds the performance of statistical models. This is beyond the scope of this research.

One-hour prediction

Given that the data was acquired approximately every 15 mins, one-hour prediction represents 4 steps forward prediction.

The winning model was Ridge Regression (see Table 1 for more details). The true out-of-sample performance is shown in Figure 1 (a). The model consists of ten features automatically derived from the original time-series. The top features or the key drivers were the wind speed itself (2 step lag and 1 step difference), the pressure (the actual value and the 2 step lag), the wind gust (median across a window size of 6 steps or 1.5 hours) and the 2 step lag of the temperature.

Every model was evaluated formally across a number of metrics. The Median Absolute Error is 0.66, the Mean Absolute Error equal to 0.97 and the Mean Squared Error equal to 1.99. All metrics are reported on the true out-of-sample performance.

Twenty-four-hour prediction

In the case of 24hours data was aggregated to one hour intervals.

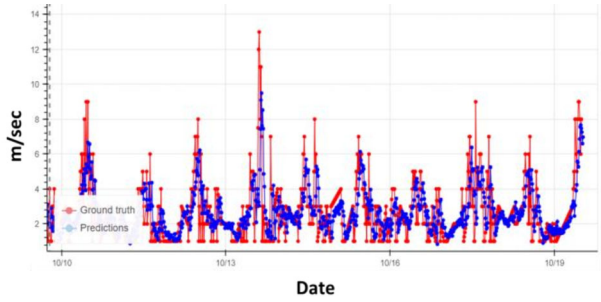
TABLE I. FORECASTING HORIZONS USING DIFFERENT ARCHITECTURES

Forecasting Horizon	Target Variable	Winning Model
1	Wind Speed	Ridge Regressor Alpha = 0.001 Window Size = 6 Type = Offline
24	Wind Speed	Elastic Net Regressor Alpha = 0.001 L1 ratio = 0.1983 Type = Offline

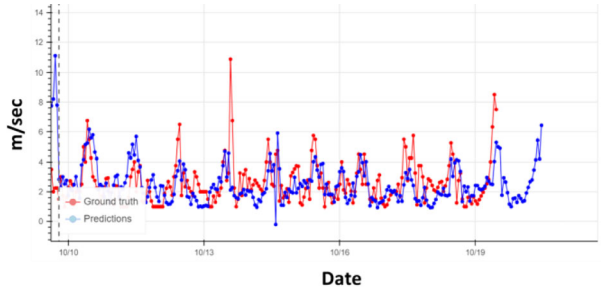
The winning model was an Elastic Net Regressor (see Table I for details). The true out-of-sample performance is shown in Figure 1 (b). The model consisting of ten features automatically derived from the original time-series. The top features or the key drivers were the wind speed itself (1 step difference and 1 step lag), the pressure (median value across a window size 6 and 1 step difference, 3 step lag), humidity (1 step and 3 step lags) the wind gust (2 step lag), wind direction (2 step lag) and the 1 step difference of the Heat Index.

Every model was evaluated formally across a number of metrics. The Median Absolute Error is 0.90, the Mean Absolute Error equal to 1.07 and the Mean Squared Error equal to 2.11. All metrics are reported on the true out-of-sample performance.

Figure 1 demonstrates that the resulting time series models that were derived from the platform achieve a high accuracy at predicting future wind speed in novel situations.



(a)



(b)

Fig. 1. Performance of over test data from causaLens platform. Wind speed predictions for 1 hour and 24 hours ahead (Top and bottom respectivelly).

B. Portweather prediction

The Portweather platform [6] could be further expanded by integrating the proposed prediction module, which is able to predict the value of the wind speed for the next 1 hour.

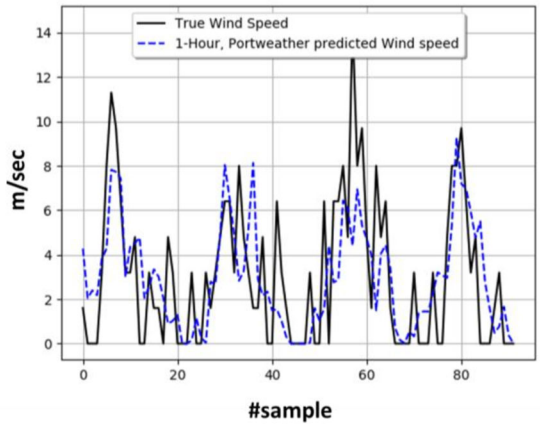


Fig. 2. Performance of the Portweather algorithm over test data for 1-hour prediction.

In order to determine the value of the time lag we did a correlation analysis for the output parameter using the training data. Figure 2 presents the accuracy of the Portweather algorithm, which is relatively high and comparable to the causaLens platform. We have implemented all the code using the Python [9] scripting language.

V. Conclusions

The aim of this study was to predict the wind speed on short-term weather conditions for an on-board vessel weather station. Several machine learning models were developed for different forecasting horizons and their efficiency for this study was presented across a number of metrics. A regression machine learning algorithm was chosen for sea trials on Lincoln vessels, due to the practical limitations of the application.

It is important to note that the models presented in this study did not make use of the testing data. The parameters of the model were not adjusted using this data and therefore the results can be interpreted as a proxy for "real-life" performance. When training data-driven models, we are interested in obtaining a model with the highest generalization performance. Good generalization means good predictive ability over previously-unseen instances. The generalization ability of a model is indicated by what is called as true error. The ultimate goal of a learning model is thus to minimize this true error and attain good generalization performance. This goal was met in the current study, by both approaches followed.

Acknowledgment

This study is supported by LINCOLN (Lean Innovative Connected Vessels) Project (www.lincolnproject.eu) Horizon 2020 research and innovation program (Grant Agreement: 727982). We would like to give special thanks to causaLens [4] for giving us access to their platform.

References

[1] http://www.fondriest.com/news/wind-speed-and-direction.htm. (accessed, July 2018)

[2] P. Karvelis, G. Georgoulas, S. Kolios and C. D. Stylios, Ensemble Learning for Forecasting Main Meteorological Parameters, 2017 IEEE International Conference on Systems, Man and Cybernetics (SMC), Banff, Canada, 2017.

[3] G. Georgoulas, P. Karvelis, S. Colios, C. Stylios, Examining nominal and ordinal classifiers for forecasting wind speed, in the 8th IEEE International Conference on Intelligent Systems IS'16.

[4] Causlanels, https://www.causalens.com/.

[5] Corfu Port Organization, https://www.corfuport.gr/

[6] S. Kolios, A. Vorobev, G. Vorobeva and C. Stylios, GIS and Environmental Monitoring: Application in the Marine, Atmospheric and Geomagnetic Fields, Springer ISBN 978-3-319-53084-0 (2017).

[7] R.J. Hyndman and G. Athanasopoulos. Forecasting: principles and practice. OTexts; 2014.

[8] G. Shrivastava, S. Karmakar, M.K. Kowar, P. Guhathakurta. Application of artificial neural networks in weather forecasting: a comprehensive literature review. International Journal of Computer Applications; 2012.

[9] Python, https://www.python.org/