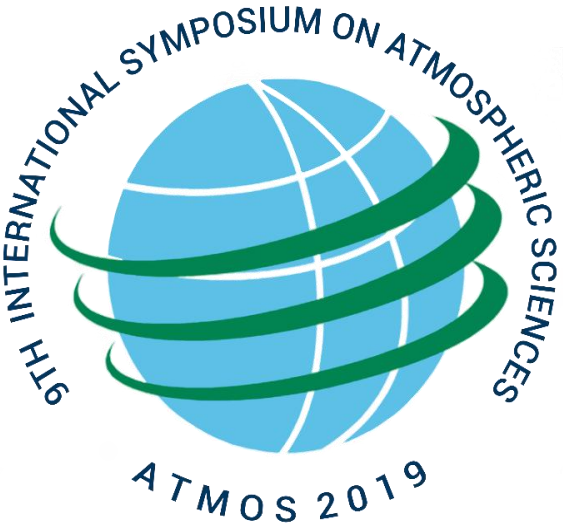

EDITORS

Dr. Kasım KOÇAK
Dr. Levent ŞAYLAN
Dr. Emine Ceren EYİGÜLER
Dr. Zerefşan KAYMAZ
M.Sc. Evren ÖZGÜR

e-ISBN: 978-975-561-482-3

e-Proceedings of the 9th International Symposium on Atmospheric Sciences

International Symposium on Atmospheric Sciences 23.10.2019 – 26.10.2019

ATMOS2019 supports and follows the ethical standards developed at the 2nd World Conference on Research Integrity, Singapore, July 22-24, 2010.

Air Pollution Forecasting for Ankara with Machine Learning Method

Zeynep Feriha Ünal[1], Umur Dinç[1], Cem Özen[1], Hüseyin Toros[1]

[1]Istanbul Technical University, Faculty of Aeronautics and Astronautics, Department of Meteorological Engineering, Istanbul, Turkey

unalz15@itu.edu.tr, dincum@itu.edu.tr, ozenc@itu.edu.tr, toros@itu.edu.tr

ABSTRACT

Environmental pollution problems have become more serious due to perceivable global climate change effects in the 21st century. Air pollution is one of the major pollution problems like water pollution, soil pollution. To protect people from fatal health diseases and extreme weather events as effects of air pollution, many countries have been trying to reduce their own emissions in keeping with agreements such as Paris Agreement, Kyoto Protocols. It is thought that remarking air pollution problems for Ankara in this study would be so essential for Turkey due to its critical daily air quality index values and also creating enlarging public awareness about air pollution. In this study, parameters like wind, pressure, temperature, humidity and precipitation are used for changeable air pollution parameter values like PM_{10}, SO_2 and NO_2. WRF model was operated for 2017 and 2018 years by using global NOAA's GDAS data with 9 km resolution. Air pollution measurement data from the Ministry of Environment and Urbanization for six points in Ankara and meteorological variables obtained from the WRF model were used to predict air quality for Ankara. To train about the relationship between air pollution and meteorological variables, the machine learning package H_2O was used in the R software. The educated model was tested with last one month data of 2018 and for the final step, the model performance rates and errors were obtained by RMSE (Root Mean Square Error), MAE (Mean Absolute Error) and correlation values.

***Keywords*:** Air pollution, machine learning, H_2O model, WRF model

INTRODUCTION

It is a fact that there are many different definitions of air pollution, but the simplest way to explain is basically the presence of substances in the atmosphere at the concentration level and duration that will harm living things or nature (Toros, 2000). Air pollutants are divided into two main groups: Primary pollutants and secondary pollutants. Primary air pollutants are emitted directly from the source into the atmosphere. Primary air pollutants include Particulate Substances (PM), Sulphur Oxides (SOx) Nitrogen Oxides (NOx), Hydrocarbons (HC) and Carbon Monoxide (CO). Secondary air pollutants are formed by chemical and/or physical reactions of primary air pollutants. These include PAN (Peroxy Acetyl Nitrate), Ozone (O_3), Sulphuric Acid (H_2SO_4). There are many methods to measure air pollution level in one region but generally the main parameters measured and they are analysed according to pollution criteria. The main parameters released from these sources in the category of air pollution are particulate matter ($PM_{2.5}$, PM_{10}, etc.), SOx (sulphur oxides), NOx (nitrogen oxides), CO (carbon monoxide) and finally O_3 (ozone). The causes of air pollution are mainly divided into two groups: natural and anthropogenic (sourced by human activities). Natural sources are volcanic eruptions, dust transport, forest fires (Choundary & Garg, 2015). The anthropogenic sources are divided into three sources: Fossil fuels used for heating in living areas as areal sources, transportation vehicles like planes, trains and cars as linear sources and factories, industrial activities and power plants as point sources. The first mention of the problem of air pollution in the world emerged as a result of the industrial revolution that started in the late 18th century and continued to exist in the early 19th century. The first event in which serious consequences of air pollution were observed and echoed in this case was the smog in London in 1952. Gasses from factories and fireplaces in homes created the smoke-mixed fog in the city, and the fog layer suspended in the air for several days, killing about 4,000 people in London. The effects of air pollution on ecological balance and living health are destructive and lethal, as can be seen from the example. Events such as acid rain, climate change, respiratory diseases, emergence of extreme weather conditions, decrease / increase in the number of species in the ecosystem could be considered among the effects of air pollution. Especially, human health is closely affected due to misconstruction in megacities, heavy traffic and population density. The impact of air pollution on humans is in respiratory diseases and allergic reactions in populous

cities such as Beijing (Incecik & İm, 2013). Furthermore, outdoor air pollution leads to 1.3 million premature deaths each year worldwide (WHO, 2008). For the last three decades, air pollution forecasting systems have been developed day by day. Air pollution forecasting is vital to create prepared and sensitive individuals for extreme events in the future. Common approaches for air pollution forecasting include simple empirical approaches, statistical approaches, and finally physics-based approaches. Each approach naturally has its own advantages and disadvantages. Today, it is known that the most frequently used approach is statistical approach. Air pollution forecasting with artificial intelligence models is closer to the statistical approaches than the basic approaches mentioned above. The concept of Artificial Intelligence is officially taking place since the times of World War II. The concept of artificial intelligence is mainly focused on the action of learning similar to human intelligence. This learning action is very vital for artificial intelligence to do desired job. Because the higher the learning performance, the more successful it can be done. Similarly, deep learning and machine learning are used for the concept of artificial intelligence. If artificial intelligence is considered as a cluster, machine learning should be considered as a subset of its machine learning set, while deep learning as its subset. This study has been conducted on the problem of air pollution in the capital Ankara, which has been in the literature since the 1960s. For the model learning, different datasets were compared and the most suitable meteorological and air pollution data were used. These data are divided into two for the learning and testing part of the models. For the air pollution forecasting, artificial intelligence models with different architectures were tried and the most suitable model was selected for the air pollution data set at 6 different stations and performance comparisons were made. Comparisons were made for PM_{10}, SO_2, NO_2 and CO parameters as individually in every station, not only on stations to stations. Performances of models were determined by calculation of RMSE (Mean Square Root), MAE (Mean Absolute Error) and correlation values.

STUDY AREA AND DATA

Study Area

The capital of Ankara, the country's second most populous city, has a population as 5,503,985 people with the year of 2018 (Wikipedia, 2019). The most preferred job of the people living in the city is sourced by trade and industrial facilities. Sincan, Yenimahalle and Polatlı regions are the most known industrial activity regions among the others. In terms of latitude and topographic location of Ankara, the continental climate is dominated in the middle and south of the city and the Black Sea climate prevail in the northern region. In Continental climate, cold and snowy winter, hot and dry summer seasons are observed and temperature differences between seasons and day and night are high. The city's surface area is approximately 25.632 km^2. Approximately 50% of the land is composed of agricultural areas, 28% is forested and heathland areas and the remaining percentages are meadows, pastures and non-agricultural lands. (Wikipedia, 2019). The northern part of Ankara is surrounded by The Northern Anatolian mountain range and there is a large plain in Ankara's south region. The metropolitan part of Ankara can be said to be the lowest point of a valley surrounded by mountains. The topographic feature and accelerating urbanization cause the city heat island effect increases from year to year especially in winter (Çiçek & Doğan, 2006). The topographic image of Ankara and its nearby regions is shown in Figure 1 below.

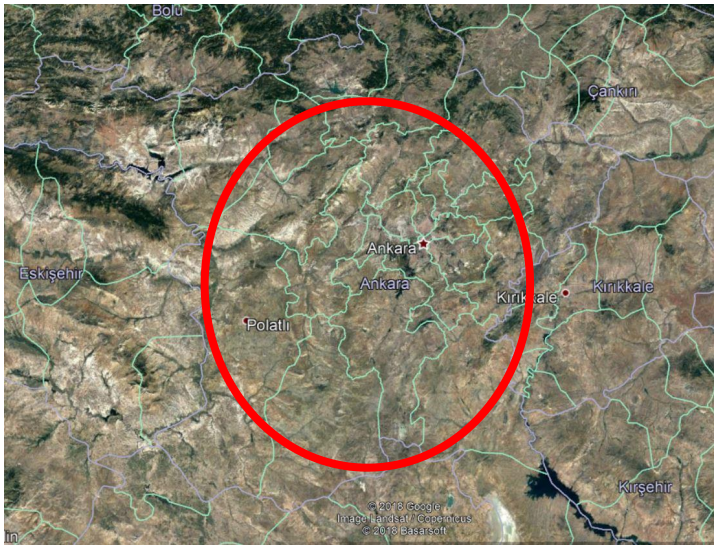

Figure 1. Topographic Map of Ankara Province (Google, 2019)

Meteorological Data

In this study, there were two different meteorological data sets to train machine learning model: Data set from Turkish State Meteorological Service and GDAS data from NOAA. Firstly, Meteorological Service's data set include wind speed, wind direction, temperature, pressure and humidity parameters. Only wind speed, wind direction and temperature parameters were found to be more complete than the other parameters. As a result of the observation, it was understood that the data which were considered to be complete contain occasional incorrect values. Therefore, in order to create a more consistent forecasting process by comparing, it was decided to use only GDAS data in the learning part. GDAS data is generated using GFS model data which is a model of the NCEP center operating under the United States' agency NOAA. This model's data has many resolutions which expressed in degrees. GDAS data indicates that the meteorological station measurement data and the GFS analysis data assimilation process. If simply stated, GDAS data is more current and consistent than GFS data.

In this study, GDAS data from January 1, 2017 to January 1, 2019 with a resolution of 0.25 degrees (~ 25 km), were taken and meteorological data were used as input for WRF model. Outputs from the WRF model are wind speed and wind direction (10 m), temperature (2 m), pressure (sea level pressure) and relative humidity (2 m) with 9 km resolution. Since meteorological data and pollutant data should be in a coherent way, outputs were made for the nearest grid point to the selected stations. Outputs from the WRF model were used for the training and test parts of the machine learning model for air pollution forecasting.

Pollutant Data

In this study, Ministry of Environment and Urbanization's six air quality stations located in Ankara are selected. Among the pollutant data which belongs the station data, PM_{10}, SO_2, NO_2 and CO data were found suitable for forecasting. The 2-year and hourly air pollution data from the Ministry of Environment and Urbanization's air quality stations data from January 1, 2017 to January 1, 2019 were used in the training and testing part of the machine learning model.

Table 1. Air Quality Monitoring Stations' Coordinates in Ankara

Stations	Latitude	Longitude
Ankara-Demetevler(Yenimahalle)	39.9678	32.7959
Ankara-Dikmen(Çankaya)	39.9180	32.8227
Ankara-Kayaş(Mamak)	39.9252	32.9266
Ankara-Keçiören(Keçiören)	39.9991	32.8555
Ankara-Sıhhiye(Çankaya)	39.9272	32.8594
Ankara-Ulus(Altındağ)	39.9400	32.8498

Selected stations' coordinates are shown in Table.1 above and also stations' locations in a map image is given in Figure.1 below.

METHOD AND MODEL CONFiGURATiON

As it is mentioned above, since the meteorological data were obtained from WRF model, it is needed to mention about its working scheme and also model configuration. These meteorological data obtained from WRF and pollutant data from Ministry of Environment and Urbanization are used for learning and testing part of H_2O package models. Later, the further information about the machine learning model package (H_2O package) from R program, the method and its configurations were explained and for the final step, RMSE, MAE and correlation formulas were explained which is used to test package models' performance rates and also model errors.

WRF Model

WRF model, began to be developed by the end of 1990 is a numerical weather prediction models. The model is used for atmospheric research as well as for operational weather forecasting. The scale can be operated from micro scale (ten meters) to synoptic scale (thousands of kilometers). It has multiple dynamic central codes, data assimilation system, parallel computing and open source software architecture. (NCAR, 2019).

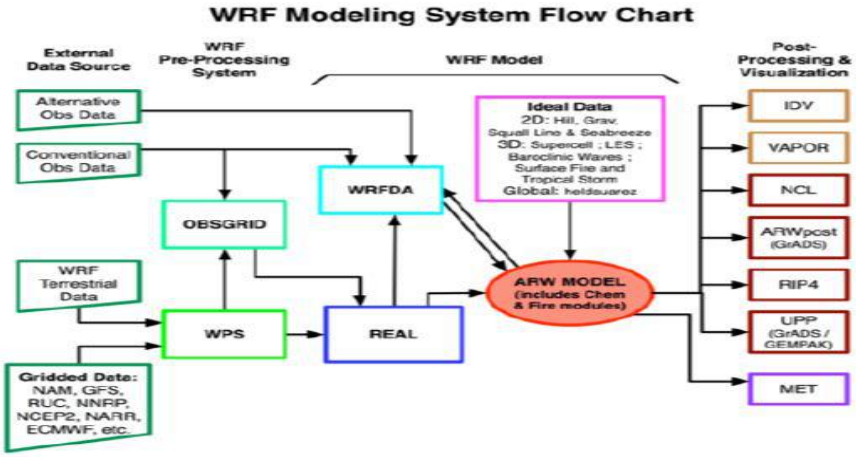

Figure 2. WRF Model System Flow Chart

WRF model running involves 4 major steps named as WRF Preprocessing System (WPS), Initialization (Real), Numerical Integration (WRF) and Visualization. In the study, 4.1.1 version of WRF model and 4.1 version of WPS were used to run GDAS data. As shown in Figure 3, 9 km single domain is used as the domain.

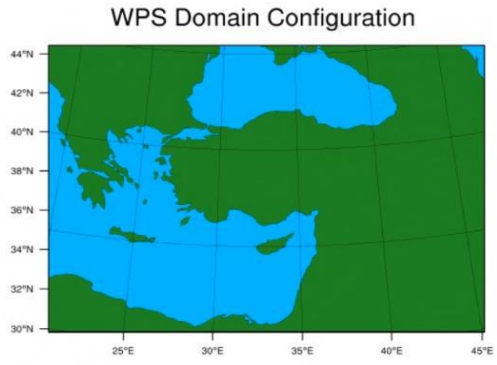

Figure 3. Model domain for this study

As shown in Table 2 below, the most appropriate settings for real case simulations are selected. The settings can be increased to make the simulations even more successful. The settings used in the current run are those recommended by developers and are the most used by users. Unmentioned settings are adjusting to be default.

Table 2. Physics Options used in this study

Physics Options	Value	Selected
Microphysics	8	Thompson
Longwave Radiation	4	RRTMG
Planetary Boundary layer	2	MYJ
Surface Layer	2	Janjic Eta
Land-Surface	2	NOAH

General Information about H_2O Package

The H_2O model is one of the packages included in the R program. The model is an open source, scalable machine learning platform. The platform includes an analytical environment based on detailed study and forecasting of big data ("Welcome to H_2O 3", 2019). The core code of H_2O is written in Java. It can be accessed via R and Python programs.

The package includes over 15 models with different algorithms. Common models in H_2O package are Deep Learning, Distributed Random Forest (DRF), Generalized Linear Model (GLM), Gradient Boosting Machine (GBM), Auto Model Selection (AutoML). In this study, these 5 model are operated and model results were obtained separately. In this study, 3.26 version of H2O package was used over R program. The same algorithm is used for the GBM, GLM, DRF, Deep Learning models except AutoML model. The AutoML model contain the same algorithm but also it has the settings such as learning times, selecting the best model and creating the ensemble model. If the parameters specific to the model are not specified in the script, they will have the default values automatically. Model developers stated that they recommend these values for people who are not in the expert category.

In the R script used, readxl package can be read easily from the data; The rWind package is used to use wind speed and direction data in East-West and North-South direction format; For the model estimation consistency analyzes (RMSE, MAE values), hydroGOF package was used. For GBM, GLM, DRF and Deep Learning models, the model sequence is as follows:

1. For each parameter, approximately 17,521 pollution data and the same number of meteorological data were read from Excel files for each station to be processed as input in the model.

2.The last 1000 data of the remaining data were converted into test data, the data other than the last 1000 data were used in the training process. (The idea behind that is the longer model learns, the better results can be given.)

3. For the results obtained by running models, correlation, RMSE and MAE values were measured and their performances were determined.

There are three common commands to all models are test (for data to be predicted), train (for data to be learned), seed (for generating random numbers). The purpose of generating random numbers in the Seed command is to determine how many models the ensemble models will generate in-house. The value used in this study is 7.

For the AutoML model, in addition to the 3 steps, there is a comparison process by obtaining separate model achievements as step 4. In the comparison process, after comparing all model performances, he creates an ensemble model by using all models and chooses the best model among them. Finally, these two models reach the final conclusion by comparing their performance. The commands used in step 4 are leaderboard_frame (to determine the best model), blending_frame (for the Stacked Ensembles model within the model), max_runtime_secs (to determine how long each model is run). Seed value 3 in AutoML; The max_runtime_secs value is 1800 (15 minutes training and testing process for each model). 10 different models were compared in AutoML.

Performance Testing

As it is mentioned above, hydroGOF package was used to determine RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) and correlation rates. Their equations are shown in Eqs. (1)-(2)-(3) below.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \quad (1)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \quad (2)$$

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,]\,[\,n\Sigma y^2 - (\Sigma y)^2\,]}} \quad (3)$$

RESULTS

As it can be seen from above sections for performance testing, RMSE, MAE are calculated as model errors but only correlation rates are considered for this study.

Demetevler

Table 3. Demetevler's Correlation Rates

Demetevler Correlation Rates	PM10	SO2	NO2
GBM	0.34	0.03	0.24
GLM	0.38	0.02	0.29
Deep Learning	0.09	-0.1	0.35
Random Forest	0.29	0.03	0.21
AutoML	0.41	0.11	0.36

Dikmen

Table 4. Dikmen's Correlation Rates

Dikmen Correlation Rates	PM10	SO2	NO2	CO
GBM	0.64	0.14	0.46	0.39
GLM	0.32	-0.03	0.32	0.24
Deep Learning	0.22	0.08	0.41	0.24
Random Forest	0.61	0.12	0.46	0.4
AutoML	0.69	0.38	0.54	0.45

Table 5. Kayaş's Correlation Rates

Kayaş Correlation Rates	PM10	SO2	NO2
GBM	0.48	0.01	0.33
GLM	0.4	0.05	0.25
Deep Learning	0.22	-0.07	0.21
Random Forest	0.47	0.01	0.34
AutoML	0.53	0.26	0.38

Keçiören

Table 6. Keçiören's Correlation Rates

Keçiören Correlation Rates	PM10	SO2	NO2
GBM	0.39	0.28	0.5
GLM	0.28	0.1	0.42
Deep Learning	0.25	0.27	0.38
Random Forest	0.37	0.19	0.43
AutoML	0.44	0.35	0.53

Sıhhiye

Table 7. Sıhhiye's Correlation Rates

Sıhhiye Correlation Rates	PM10	SO2	NO2	CO
GBM	0.4	0.13	0.55	0.4
GLM	0.42	0.25	0.4	0.24
Deep Learning	0.44	0.09	0.33	0.29
Random Forest	0.42	0.08	0.53	0.4
AutoML	0.49	0.34	0.69	0.51

Ulus

Table 8. Ulus' Correlation Rates

Ulus Correlation Rates	PM10	SO2	NO2	CO
GBM	0.17	0.23	-0.38	-0.03
GLM	0.13	0.17	-0.09	-0.01
Deep Learning	0.16	0.14	-0.26	-0.08
Random Forest	0.21	0.2	-0.4	-0.04
AutoML	0.265	0.26	0.7	0.22

When the five separate models in the H_2O package are compared separately for the parameters of each station, it is seen that the performance of the AutoML model for each parameter is higher than the performance of the other models. The reason for this is the AutoML model compares the other models and takes the ensemble or chooses the best model among them. The reasons for low correlation are thought to be due to the lack of data in the air pollution data set.

CONCLUSION

It is known that the air pollution problem mentioned in the previous sections is among the most important environmental problems. This important environmental problem is a major threat to living health and habitats. One of the measures that can be taken in this regard is to have information about the amount of air pollution in the near future. In line with the results obtained in the estimation phase of this study, it is planned to establish an estimation system for the near future pollution levels. In this estimation stage, the results were obtained by using various artificial intelligence models. Correlation, RMSE, MAE values were calculated for each parameter. The best correlation values were determined for the station parameters. The most successful model results among all models were found in AutoML model. Although there are many reasons for this, there are prominent reasons such as determining the flow of learning, comparing many models and making the best choice. The highest correlation was found for the NO_2 parameter at the Ulus station and the correlation value was 0.70. The reason why the correlation values of some parameters are not as high as desired is the lack of data in the data sets. Because, as mentioned earlier, the quality of the data in the learning process of the models is important in terms of their predictive performance. In the following studies, it is planned to measure the repeat performance by using the air pollution model results (WRF-Chem, CMAQ, etc.) in the missing air quality data such as using WRF model results instead of missing meteorological data. It is also contemplated to try different options for the configuration parameters suitable for the data. In another study, it was planned that comparing the performance of the model with other models would be better for the model performance analysis.

REFERENCES

Choundary, D. M. P., & Garg, V. (2015). Causes, Consequences and Control of Air Pollution. Control of Air Pollution, 1(August 2013), 9–11.

Çiçek, İ., Doğan, U. (2006). Detection of urban heat island in Ankara, Turkey. Nuovo Cimento Della Societa Italiana Di Fisica C, 29(4), 399–409. https://doi.org/10.1393/ncc/i2005-10028-2

Google (n.d.). Retrieved August 24, 2019, from https://www.google.com/maps/place/Ankara/@39.9032923,32.6226801,11z/

İncecik, S., İm, U. (2013). Hava Kirliliği Araştırmaları Dergisi. (Ocak 2013).

National Centers for Environmental Prediction/National Weather Service/NOAA/U.S. Department of Commerce (2015): NCEP GDAS/FNL 0.25 Degree Global Tropospheric Analyses and Forecast Grids. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. Dataset. https://doi.org/10.5065/D65Q4T4Z.

NCAR. (n.d). "Weather Research and Forecasting Model" Retrieved July 29, 2019 from https://www.mmm.ucar.edu/weather-research-and-forecasting-model

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, Z. Liu, J. Berner, W. Wang, J. G. Powers, M. G. Duda, D. M. Barker, and X.-Y. Huang, 2019: A Description of the Advanced Research WRF Version 4. NCAR Tech. Note NCAR/TN-556+STR, 145 pp. doi:10.5065/1dfh-6p97

Welcome to H2O 3 — H_2O 3.26.0.2 documentation. (2019). Retrieved August 3,2019, from http://docs.h2o.ai/h2o/latest-stable/h2o-docs/welcome.html#new-users

Wikipedia contributors. (2019, August 24). Ankara. In Wikipedia, The Free Encyclopedia. Retrieved August 1, 2019, from https://en.wikipedia.org/w/index.php?title=Ankara&oldid=912304374