C204 Project Final

The goal of the project is to be apply machine learning techniques to predict final OH concentration in time-resolved measurements. This measurement is needed to calculate slow phase OH formation.

1. Introduction

Reactive oxygen species (ROS), such as the hydroxyl (OH) radical, play an important role in aerosol and cloud water chemistry[1]. The OH radical has both direct and indirect impacts on the climate by contributing to the processing of aerosols (e.g. chemical composition and size distributions) which in turn can alter scattering properties and plays a role in brown carbon formation[2]. There are several contributing formation processes of the hydroxyl radical, both aqueous and gas-phase, but the majority tend to be slow[3]. A significant burst of aqueous hydroxyl radical is rapidly generated when particles are grown into cloud droplets with concentrations ranging from 0.1 to 3.5 μM[4]. The magnitude of the burst is dependent on particle concentration, composition and the degree of cloud processing the particles undergo. The burst can be generalized as the hydroxyl radical concentration within moments of the particle growing into a cloud droplet.

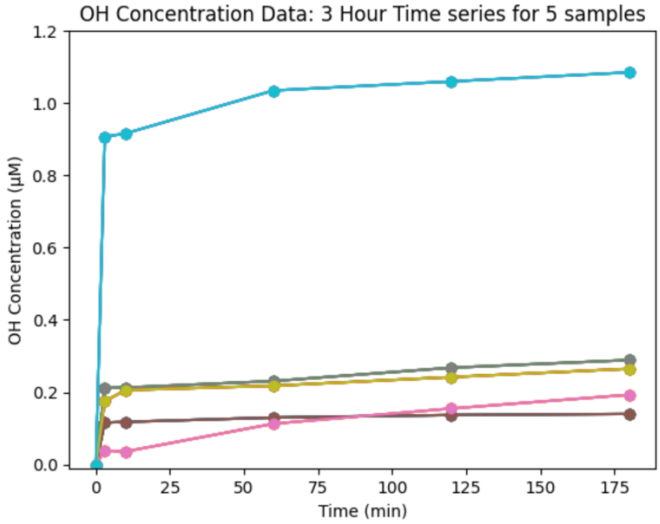

Figure 1. Plot of OH concentration as a function of time for 5 samples during field deployment in La Jolla, CA. Measurements were taken at 3, 10, 60, 120, and 180 minutes.

To measure OH, we collect particles directly into a terephthalate probe that scavenges the OH formed, and measure this complex via fluorescence spectroscopy to derive the concentration. This measurement is not able to isolate the contribution from just the aqueous burst and includes slow-phase OH processes. To characterize the OH concentration over time and account for slow-phase formation, we take fluorescence measurements at 3 minutes (burst measurement), 10 minutes, 60 minutes, and so on, for as long as half the sample collection period. For this dataset the final time data point is 180 minutes. The typical trend this data takes on is a rapid increase of OH followed by a slow increase, indicating that the main contribution of OH for the particles collected is due to this aqueous chemistry (Fig. 1). To determine the OH concentration contributed by the slow-phase, we subtract the final concentration measurement from the 10 minute

measurement; as all activity due to the burst would have been generated by now. We then take the OH concentration attributed to the slow-phase and subtract this from the 3 minute initial OH concentration to get the OH concentration contributed by the burst. This process is arduous, and a new sample cannot be started until the final concentration in the time series is measured. Thus, it would be useful to train the time series data to predict the contribution of the slow-phase from the first hour's data; specifically we aim to predict the final data point in the time series. This will aid in isolating the OH concentration that is solely formed via the burst pathway. I will apply supervised regression techniques in order to make the prediction. First, a linear regression model with a polynomial fit will be applied on the time series data to predict the 180 minute OH measurement. Next, I will use an autoregression model to see if this method will do a better job at predicting the final 180 minute measurement using the past values in the time series.

2. Data

The dataset being used for this project is from field measurements taken during the Eastern Pacific Cloud Aerosol Precipitation Experiment (EPCAPE) in La Jolla, California. The dataset is comprised of 71 samples collected at the site and features include the sample collection period (minutes), calculated OH concentration for the measured time points (μM), averaged particle concentration ($\mu g/m^3$), atmospheric pressure (kPA), mean surface temperature (°C), relative humidity (%RH), vapor pressure (kPa), wind speed (m/s), and wind direction vector mean (degree). Particle composition data acquired by the Aerosol Chemical Speciation Monitor (ACSM) are also included in the dataset to see how composition might influence the OH time series. ACSM and Meteorological (MET) data can be downloaded on the ARM data browser: ARM Data. The raw OH data was preprocessed, converting fluorescence signals for each time point into concentration (μM). Most of the OH concentrations in the data set fall into the 0.1 – 0.3 μM range, however, there are a few series that see OH concentrations well above 1 μM, appearing to be outliers. These samples were not removed as this is well within the range we expect for OH formation from previous field data. The data from the ARM data browser was averaged over the sampling period and concatenated with EPCAPE OH field data.

3. Modeling

Due to the constraints of the sample size for the data set (n = 71), I am limited in which machine learning model I can apply for the time prediction series. I first used a linear regression model with a polynomial fit using the first four points in the time series data to predict the final measurement. When choosing the degree of the polynomial fit we must consider several factors. First, we look at the relationship between the independent (time) and dependent (OH concentration) variables. For the case of this dataset, the relationship appears to be more linear, thus a lower order polynomial will be selected. If we choose a high order polynomial the model fit for the data will be subject to overfitting. Once the best fit is determined, we will look at features from the data set would improve the fit of the data. I applied a covariance matrix on the remaining features from the ACSM and MET as well as the time at 180 minutes to see if there was covariance between these variables. Scikit-learn packages for linear regression, polynomial fit, mean square error and mean absolute error were imported into the workbook[5]. These packages evaluate the performance and the error within the model. I will first run the entire time series (time 0 to time 120) using the regression model to get an initial fit. I will then run the model using

the first four data points (time 0 to time 60) as the goal of the project is to be able to acquire enough data without compromising the fit. When splitting the data for testing and training, I will use a test size of 30% such that the model will have more samples to evaluate for the performance.
The next regression model technique I tried to implement was an autoregression model to again, see if the past OH concentrations could inform the final OH concentration of the series, in short we want to predict the temporal relationship. For this model we import the stats statsmodel autoregression package into our file. To set up the data frame such that previous measurements inform the target measurement, we implement the concept of lagging data. This is where we shift variables over the time series to inform the next time points prediction. I will apply the same split for the training and test data set as the linear regression model , as well as the same error metrics.

```
from statsmodels.tsa.api import AutoReg
# create the data shift for time series prediction
data_shift = pd.DataFrame({
    'time_0_t-1': df['time_0'].shift(1),
    'time_3_t-1': df['time_3'].shift(1),
    'time_10_t-1': df['time_10'].shift(1),
    'time_60_t-1': df['time_60'].shift(1),
    'time_120_t-1': df['time_120'].shift(1),
    'time_180_t': df['time_180']
})

# Reset index to default integer index
data_shift = data_shift.reset_index(drop=True)
# Set a valid DateTime index for the lagged_data DataFrame
data_shift = data_shift.dropna()
```

4. Results

4.1. Linear Regression Model and Covariance Matrix

Below are the model testing data results for the linear regression and autoregression applied to the OH data set. Figure 2 shows the results of the linear regression for the actual 180 minute OH concentration versus the predicted 180 minute OH concentration using an order of one. The figure on the left (purple) includes all time data points in the set. The right hand side of the figure (blue) includes the first three time data points. Figure 3 displays the entire OH time series for the linear regression model used. Data points to the left of the red dotted line indicate the X_test values for OH concentration, and to the right indicate the data points generated by the model prediction. The top figure includes all time data points in the set and the bottom figure considers only the first three points.

Table 1 displays the results for ranking the features in the covariance matrix for variables from the ACSM and DOE MET data set, as well as the actual OH concentration at time point 180 minutes.

4.2. Autoregression Model

Figure 4 shows the results of the OH concentration time series using the autoregression model. The red dotted line indicates the same as Figure 3 with the left hand side

displaying the X_test data, and the right hand side displaying the predicted values at 180 minutes.

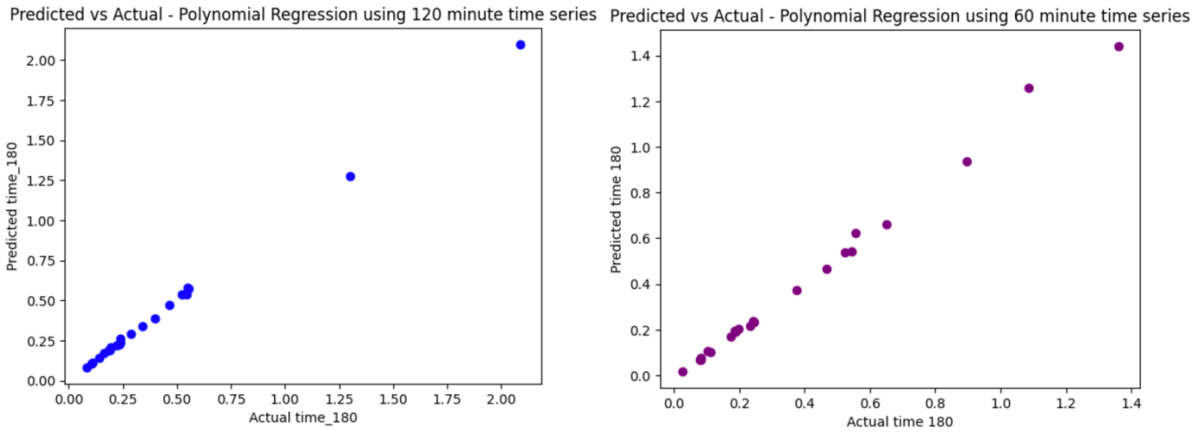

Figure 2 . Plot of real OH concentration at time point 180 minutes, model predicted OH concentration at time point 180 minutes using the 120 minute time series (blue) and the 60 minute time series (purple).

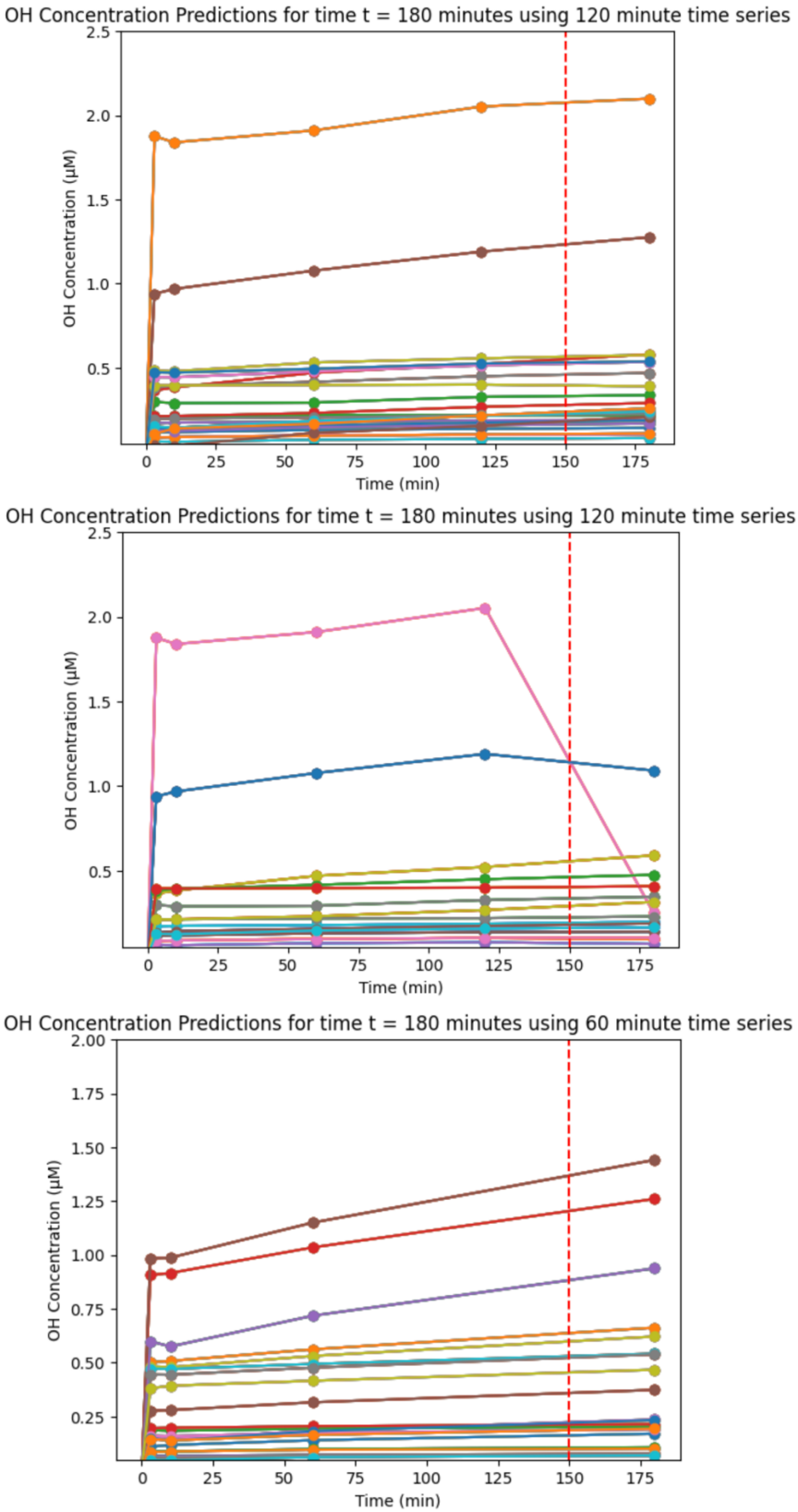

Figure 3. Cumulative OH concentration over the time series of the slow phase measurements using the linear regression model. Data points to the left of the vertical line (red) indicate points used in the X_test

set; points to the left represent the model predicted 180 minute OH concentration. (top) includes all measurements up to 120 minutes for polynomial order 1, (middle) includes all measurements up to 120 minutes for polynomial or 2, and (bottom) includes all measurements up to 60 minutes for polynomial order 1.

Table 1. Covariance matrix feature ranking for the ACSM and MET data features.

Features	Rank
wdir_vec_mean	4934.86
rh_mean	129.17
mass_conc	108.42
temp_mean	69.53
wspd_arith_mean	33.87
vapor_pressure_mean	11.46
time_180	3.74
fraction_total_organic	2.66
total_sulfate	2.60
total_nitrate	1.22
atmos_pressure	1.16
total_chloride	0.39
total_ammonium	0.19

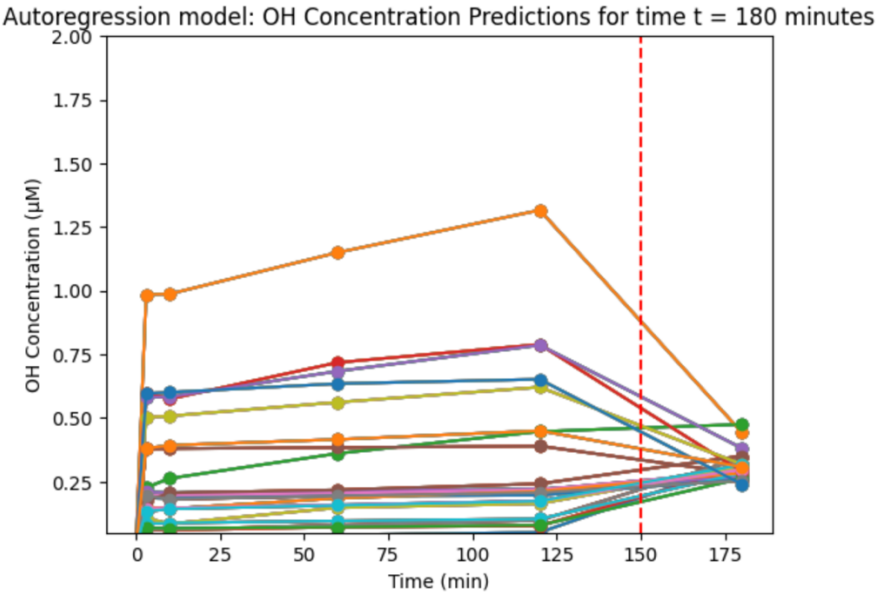

Figure 4. Cumulative OH concentration over the time series of the slow phase measurements using the autoregression model. Data points to the left of the vertical line (red) indicate points used in the X_test set; points to the left represent the model predicted 180 minute OH concentration.

5. Discussion

5.1. Linear Regression Model and Covariance

The linear regression model had the most success predicting OH concentration at the final time point, 180 minutes, in comparison to when a polynomial fit was added. The data set was run trying 3 polynomial orders (1-3), with the 1st polynomial order being most successful. When using the first 4 data points (time intervals at 0, 3, 10 and 60 minutes) the model was able to predict the time series and variable in question with a mean square error (MSE) of 0.2×10^{-3} and a mean absolute error (MAE) of 2.2×10^{-2} (See Fig. 2). These error metrics indicated the model was able to predict the target values with accuracy and can be used for future measurements. When the polynomial order was

increased, the model was still able to closely predict the OH concentration, however there is evidence that some results began to be overfit. Perhaps the model may have also been biased as most samples had concentrations in the range of 0.1 to 0.3 μM, so higher activity samples may be skewed to these ranges (See Fig. 3). A covariance matrix was also performed to see if additional features would be worth adding to improve the fit of the model. There was little covariance with the target feature (time 180) across all features. The feature that ranked the highest was wind direction vector, however it does not appear to really have importance on the actual OH concentration. I attempted to use the fraction of organics but no significant improvement was seen. It is unclear whether adding more features would truly help the prediction as the simple model runs the best fit; thus it is able to predict the final OH concentration with great confidence.

5.1. Autoregression Model

The autoregression model seems to underpredict high activity samples (> 0.5 μM) and overpredict low activity samples (< 0.25 μM) for the final OH concentration (See Fig. 4). Several issues occurred when processing the data. Similar to the polynomial fit on the linear regression, the autoregression seems to be ignoring the OH concentrations of in the series X_test and making its own prediction. Although the model does run there are several warnings that display citing forecasting and the time index being a problem. When a forecasting window was implemented in the code an error appeared when the window went above 71, the sample size. This may indicate that there is indeed an issue with the data frame, and a time index may not have been assigned. This may also be a sample number issue, with the data set not being large enough. The autoregression model might work with restructuring but at the moment it is not able to capture the underlying patterns that should predict the final OH concentration.

6. Conclusion

From this work the linear regression model has the best fit and would be the most useful for forecasting the OH concentration time series. Although the addition of a polynomial fit with an order of two, results in a decent prediction for the OH time series, it does not do as good a job compared to the linear regression model. It is not worthwhile to then use this fit unless the shape of the data would change drastically. Additional features, such as particle mass concentration or fraction of organic species, do not significantly improve the fit for our model. This was unexpected as OH formation thought to be dependent on particle mass concentration, organic species, and metals (not in the data set). When attempting the autoregression fit, a deep knowledge of how the data set is should be structured is needed. There were several warnings and issues that do not seem to be resolved, but with more work it can be fine-tuned. However, the linear regression appears to do an excellent job with predicting the final OH outcome, so the autoregression model may not be warranted.

To make the model more robust further work could be achieved. Such work includes, adding more samples to the data set, analysis of slow phase dependent on sampling location, and the addition of trace metal analysis.

7. References

1. Fuller, S.J., et al., Comparison of on-line and off-line methods to quantify reactive oxygen species (ROS) in atmospheric aerosols. Atmos. Environ., 92, 97-103 (2014).

2. Ervens, B., et al., Key parameters controlling OH initiaed formation of secondary organic aerosols in the aqueous phase (aqSOA). J. Geophys. Res. Atmospheres, 119.7, 3997-4016 (2014).

3. Zhang, Z-H., et al. , Are reactive oxygen species a suitable metric to predict toxicity of carbonaceous aerosol particles? Atmos. Chem. Phys., 22.3, 1793-1809 (2022).

4. Paulson, S.E., et al. A light-driven burst of hydroxyl radicals dominates oxidation chemistry in newly activated cloud droplets. Sci. Adv. 5.5 (2019).

5. Pedregosa, F., et al., Scikit-learn: Machine Learning in Python. JMLR., 12(85):2825-2830 (2011).