frontiers in Big Data

ORIGINAL RESEARCH
published: 10 February 2022
doi: 10.3389/fdata.2022.781309

# Development of a Machine Learning Approach for Local-Scale Ozone Forecasting: Application to Kennewick, WA

*Kai Fan*[1,2,3], *Ranil Dhammapala*[4], *Kyle Harrington*[5], *Ryan Lamastro*[6], *Brian Lamb*[3] *and Yunha Lee*[1,2,3]*

[1] Center for Advanced Systems Understanding, Görlitz, Germany, [2] Helmholtz-Zentrum Dresden Rossendorf, Dresden, Germany, [3] Laboratory for Atmospheric Research, Department of Civil and Environmental Engineering, Washington State University, Pullman, WA, United States, [4] Washington State Department of Ecology, Olympia, WA, United States, [5] Max Delbrück Center for Molecular Medicine, Berlin, Germany, [6] Environmental Geochemical Science, School of Science and Engineering, State University of New York, New Paltz, NY, United States

Chemical transport models (CTMs) are widely used for air quality forecasts, but these models require large computational resources and often suffer from a systematic bias that leads to missed poor air pollution events. For example, a CTM-based operational forecasting system for air quality over the Pacific Northwest, called AIRPACT, uses over 100 processors for several hours to provide 48-h forecasts daily, but struggles to capture unhealthy $O_3$ episodes during the summer and early fall, especially over Kennewick, WA. This research developed machine learning (ML) based $O_3$ forecasts for Kennewick, WA to demonstrate an improved forecast capability. We used the 2017–2020 simulated meteorology and $O_3$ observation data from Kennewick as training datasets. The meteorology datasets are from the Weather Research and Forecasting (WRF) meteorological model forecasts produced daily by the University of Washington. Our ozone forecasting system consists of two ML models, ML1 and ML2, to improve predictability: ML1 uses the random forest (RF) classifier and multiple linear regression (MLR) models, and ML2 uses a two-phase RF regression model with best-fit weighting factors. To avoid overfitting, we evaluate the ML forecasting system with the 10-time, 10-fold, and walk-forward cross-validation analysis. Compared to AIRPACT, ML1 improved forecast skill for high-$O_3$ events and captured 5 out of 10 unhealthy $O_3$ events, while AIRPACT and ML2 missed all the unhealthy events. ML2 showed better forecast skill for less elevated-$O_3$ events. Based on this result, we set up our ML modeling framework to use ML1 for high-$O_3$ events and ML2 for less elevated $O_3$ events. Since May 2019, the ML modeling framework has been used to produce daily 72-h $O_3$ forecasts and has provided forecasts via the web for clean air agency and public use: http://ozonematters.com/. Compared to the testing period, the operational forecasting period has not had unhealthy $O_3$ events. Nevertheless, the ML modeling framework demonstrated a reliable forecasting capability at a selected location with much less computational resources. The ML system uses a single processor for minutes compared to the CTM-based forecasting system using more than 100 processors for hours.

**Keywords: machine learning, air quality forecasts, ozone, random forest, multiple linear regression**

OPEN ACCESS

***Edited by:***
*Rasa Zalakeviciute, University of the Americas, Ecuador*

***Reviewed by:***
*Zihan Lin, Michigan State University, United States*
*Yves Philippe Rybarczyk, Dalarna University, Sweden*

***Correspondence:**
*Yunha Lee*
*yunha.lee.00@gmail.com*

***Specialty section:***
*This article was submitted to Data-driven Climate Sciences, a section of the journal Frontiers in Big Data*

***Received:*** *22 September 2021*
***Accepted:*** *19 January 2022*
***Published:*** *10 February 2022*

***Citation:***
*Fan K, Dhammapala R, Harrington K, Lamastro R, Lamb B and Lee Y (2022) Development of a Machine Learning Approach for Local-Scale Ozone Forecasting: Application to Kennewick, WA. Front. Big Data 5:781309. doi: 10.3389/fdata.2022.781309*

## INTRODUCTION

Chemical transport models (CTMs) are widely used to simulate the temporal and spatial variation of air quality (Sportisse, 2007). Chemical transport models include various physical and chemical processes of the atmosphere as well as known sources and sinks. However, due to the lack of understanding of the important physical and chemical processes in the atmosphere (Seinfeld and Pandis, 2016), CTM simulations can suffer from significant uncertainties and errors, even though the accuracy of numerical models seems to improve over time. Most operational air quality forecast systems are based on CTM and thus can experience systematic biases and errors that result in failure to forecast poor air quality events. In addition, there is a high cost for those forecasts due to the demanding computational requirements and the need for well-trained personnel to operate complex models.

The Air Indicator Report for Public Awareness and Community Tracking (AIRPACT) was developed for air quality forecasting for the Pacific Northwest (PNW) of the United States. AIRPACT, operated by Washington State University, uses the Community Multiscale Air Quality Modeling System (CMAQ) model with Weather Research and Forecasting (WRF) meteorological inputs provided by the University of Washington. The AIRPACT domain covers Washington, Idaho and Oregon along with peripheral areas with 4-km horizontal grid cells and 37 vertical levels. AIRPACT uses the Carbon Bond version 5 (CB05) gas chemistry mechanism and AERO6 aerosol module. It provides 48-h forecasts produced daily, which are available via the web[1] for the public and local air quality agencies.

Within the AIRPACT domain, Kennewick, Washington (WA) is part of the Tri-cities metropolitan area with a total population of about 220,610 [the combined population of Kennewick (84,960), Pasco (77,100) and Richland (58,550) in 2020] (Washington State Office of Financial Management, 2020). The city is located 32 km north of Washington state's southern border with Oregon and is in a hot and dry portion of the state. Recent monitoring and a large field study have shown that $O_3$ mixing ratios can be unhealthy on days in the summer and early fall (Jobson and VanderSchelden, 2017). One EPA Air Quality System (AQS) monitoring site measures the $O_3$ mixing ratios at Kennewick, which identified several unhealthy air quality events in 2017 and 2018, while the daily forecasts struggle to identify unhealthy days in this area: e.g., excluding the wildfire affected days (more details will be discussed in Section $O_3$ Observations at Kennewick, WA), there were 10 days when the air quality was unhealthy for sensitive groups in 2017–2018, but AIPRACT missed all of them.

Machine learning (ML) models have been used to predict air quality in recent years (e.g., Feng et al., 2015; Freeman et al., 2018; Zamani Joharestani et al., 2019). The numerical air quality models require a huge computational power and many input data, such as the meteorological and emission data over the whole domain. Compared to numerical models, ML methods tend to be more computationally efficient, require less input data, and perform better for specific events. The ML models typically incorporate a variety of features, including observed pollutant levels and various meteorological variables as the basis for training and applying ML methods. For example, Feng et al. (2015) used trajectory-based geographic parameters, meteorological forecasts and associated pollutant predictors as input to an artificial neural network, to predict PM2.5 concentrations in Beijing, China. Freeman et al. (2018) used a recurrent neural network with long short-term memory to predict 72-h $O_3$ forecasting using hourly air quality and meteorological data. Zamani Joharestani et al. (2019) tested three machine learning approaches [i.e., random forest (RF), extreme gradient boosting, and deep learning] using 23 features to predict the PM2.5 concentrations in Tehran, Iran.

In this study, we developed a ML modeling framework to predict $O_3$ mixing ratios that is based on RF and multiple linear regression (MLR). Random forest is one of the most popular machine learning methods and has been used in air quality modeling and forecast studies. The RF method has been demonstrated to provide reliable forecasts for $O_3$ and PM2.5 with lower computational costs compared to physical models (Yu et al., 2016; Rybarczyk and Zalakeviciute, 2018; Zhan et al., 2018; Pernak et al., 2019). Random forest consists of an ensemble of decision trees; decision tree learning is a method for approximating discrete-valued functions (Kam, 1995; Mitchell, 1997; Breiman, 2001). The RF model can be used for classification and regression, but it was suggested that it could lead to the under-prediction of the high pollution events (Jiang and Riley, 2015; Pernak et al., 2019). Since this research aims to provide a reliable $O_3$ forecast, especially for the high pollution events, a MLR or second phase RF model is also used to improve the model performance for the high $O_3$ predictions to address the under-predictions of a simple RF model. Multiple linear regression is a regression method with one dependent variable and several independent variables. Previous studies that used MLR models to predict $O_3$ mixing ratios showed performance that matched more complex machine learning models (Chaloulakou et al., 1999; Sousa et al., 2007; Arganis et al., 2012; Moustris et al., 2012). Yuchi et al. (2019) used RF and MLR for indoor air quality forecasts, and RF provided better predictions for the data in the training dataset, while MLR provided better predictions for conditions that were not represented in the training dataset.

The goal of this study is to develop a reliable air quality forecast framework using machine learning approaches and to apply the system for Kennewick, WA with a focus on the predictability of unhealthy days related to $O_3$. Section Dataset and Modeling Framework presents the datasets and the ML forecast framework based on the two machine learning approaches. Section Results and Discussion presents the feature selection, evaluation of the model performance using 10-time, 10-fold cross-validation, and the ensemble

[1]http://lar.wsu.edu/airpact/

forecasts at Kennewick. Finally, Section Conclusions provides conclusions.

# DATASET AND MODELING FRAMEWORK

## Training Dataset of Kennewick

The training dataset for our ML models includes the previous day's observed $O_3$ mixing ratios from AQS data, time information (hour, weekday, month), and simulated meteorology from daily WRF forecasts from May to September during 2017–2020 at Kennewick, WA. Because heat and sunlight favor $O_3$ generation (Weaver et al., 2009), the observations for the training set only include from May to September. Weather Research and Forecasting meteorological output was obtained from the University of Washington (Mass et al., 2003), which is used in AIRPACT as an input to generate emissions and to drive the CMAQ forecast model. We use temperature (T), surface pressure (P), relative humidity (RH), wind speed, wind direction, and planetary boundary layer height (PBLH) in the training dataset. Time information is included in the training dataset due to the significant patterns of $O_3$ variations at diurnal, weekly, and monthly temporal scales.

## Machine Learning Modeling Framework

We have developed an air quality forecast modeling framework that consists of two independent ML models. The first machine learning model (ML1; **Figure 1A**) consists of RF classifier and MLR models. The *RandomForestClassifier* and *RFE* functions in the Python module *scikit-learn* were used (Pedregosa et al., 2011). In ML1, the WRF meteorology, time information, and previous day's $O_3$ mixing ratios were first used to train a RF classifier model to predict AQI categories. Because the AQI category is based on the 8-h averaged $O_3$, the training data for the RF classifier model used the previous day's 8-h averaged $O_3$ mixing ratios. Given that a highly polluted episode is generally a rare event, it makes the dataset unbalanced, and the unbalanced training data may produce a bias toward commonly observed $O_3$ events (Haixiang et al., 2017). To address this problem, the *balanced_subsample* option was used for the RF classifier. The *balanced_subsample* gives weights to the AQI category values based on their frequency in the bootstrap sample for each tree, so that high AQI values with low frequency in the training dataset are weighted proportionally more: this is an algorithm-level strategy commonly used in machine learning to reduce bias for the majority category for datasets with class imbalance. Without applying this strategy, our machine learning model fitting can be negatively impacted by a disparity in the frequencies of the observed classes (here, AQI classes). Separately, the observed AQI categories were added to the training dataset to train the MLR model (see the red dashed line shown in **Figure 1A**). When used for forecasting, the RF classifier model was first used to predict the AQI categories, which, in turn, provided input to the MLR model to predict the $O_3$ mixing ratios.

The second machine learning model (ML2; **Figure 1B**) was based on a two-phase RF regression model. Here, the *RandomForestRegressor* function in the Python module *scikit-learn* was used (Pedregosa et al., 2011). ML2 used the WRF meteorology, time information, and previous day's hourly $O_3$ mixing ratios to train an RF regression model to predict the concentrations. The entire historical dataset was used to train the first RF regression model (RF1 in **Figure 1B**). The training data were isolated when RF1 predicted $O_3$ mixing ratios that differ from the observations by more than 5 ppb, and then the isolated dataset was used to train the second RF regression model (RF2 in **Figure 1B**). The training dataset for RF2 was the subset of the entire training data, so RF2 required more decision trees (100 trees for RF1 and 200 trees for RF2). We found that using more decision trees in RF2 led to better performance without significantly increasing the computational cost. Jiang and Riley (2015) also used more decision trees in their second-phase model training (300 trees in the first phase and 500 trees in the second phase). This is why it is called a two-phase RF regression model. In ML2, the final $O_3$ mixing ratios are computed using Equation (1) with a set of weighting factors ($a_1$ and $a_2$).

$$Hourly\ prediction\ =\ a_1 * RF1\ + a_2 * RF2 \qquad (1)$$

The $a_1$ and $a_2$ are determined in the training process because the observed ozone data (truth) is available to the models. We divide the *RF*1 ozone predictions into three categories (low, mid, and high) and find the optimal weighting factors at each category using a linear regression equation in Equation (1). When

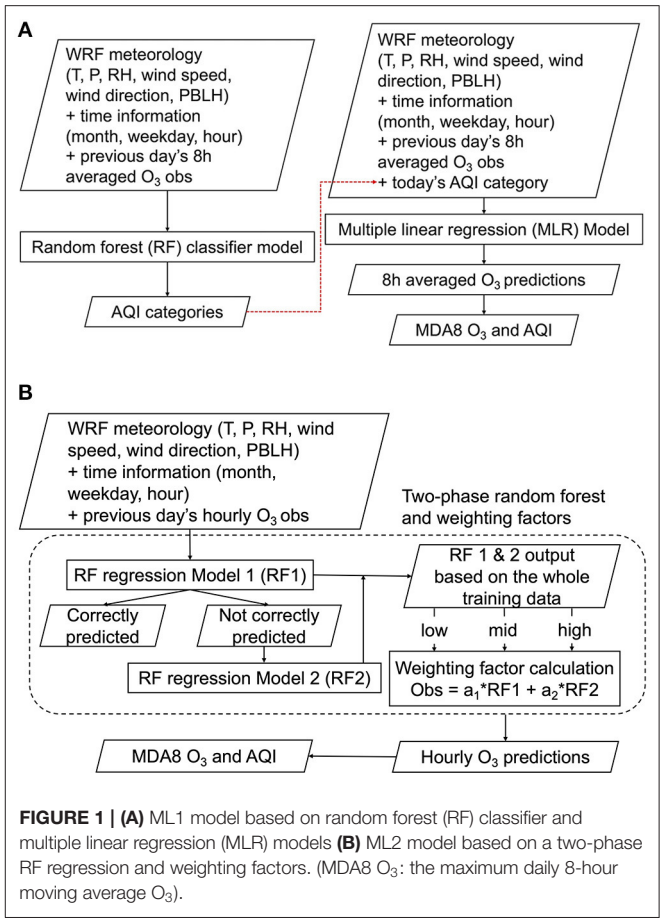


**FIGURE 1 | (A)** ML1 model based on random forest (RF) classifier and multiple linear regression (MLR) models **(B)** ML2 model based on a two-phase RF regression and weighting factors. (MDA8 $O_3$: the maximum daily 8-hour moving average $O_3$).

forecasting, *RF*1 and *RF*2 are computed first and then the *RF*1 prediction determines which weighting factors to use and the hourly $O_3$ prediction is computed using Equation (1).

## Computational Requirements

Our ML modeling framework requires much less computational power than the AIRPACT CMAQ system. The ML models use a single Intel E5 processor to train and evaluate the model. For the walk-forward cross-validation (more details will be discussed in the **Supplementary Materials**), ML1 takes about 8 min of CPU time to train the model and to predict daily $O_3$ at one location for the entire 2018–2020 ozone season (425 days in total), while ML2 takes about 27 min of CPU time to predict the same time period. These times are much less than AIRPACT that requires 360 h of CPU time (120 Intel E5 processors for 3 h) for a single 48-h forecast.

## Forecast Verifications for AQI Evaluation

Forecast verifications are used to evaluate the machine learning models: Heidke Skill Score (HSS), Hanssen-Kuiper Skill Score (KSS), and Critical Success Index (CSI). **Table 1** is a 2 × 2 contingency table that shows a simple unhealthy or good case: "unhealthy" refers to unhealthy air pollution events, and "good" refers to good air quality. Equations (2)–(4) show how HSS, KSS, and CSI are computed (Jolliffe and Stephenson, 2012), where a, b, c, and d refer to the numbers of hits, false alarms, misses and correct negatives, respectively; n refers to the total number of events.

$$HSS = \frac{a + d - a_r - d_r}{n - a_r - d_r} \tag{2}$$

where $a_r = \frac{(a+b)(a+c)}{n}$; $d_r = \frac{(b+d)(c+d)}{n}$

$$KSS = \frac{ad - bc}{(b + d)(a + c)} \tag{3}$$

$$CSI = \frac{a}{a + b + c} \tag{4}$$

Heidke Skill Score represents the accuracy of the model prediction compared with a reference forecast [$r$ in Equation (2)], which is from the random guess that is statistically independent of the observations (Wilks, 2011; Jolliffe and Stephenson, 2012). The range of the HSS is from $-\infty$ to 1. A negative value of HSS indicates a random guess is better, 0 indicates no skill, and 1 indicates a perfect score. Hanssen-Kuiper Skill Score measures the ability to separate different categories (Wilks, 2011; Jolliffe and Stephenson, 2012). The range is from −1 to 1 where 0 indicates no skill, and 1 indicates a perfect score. Critical Success Index is the number of hits divided by the total number of forecast and/or observed events (Wilks, 2011; Jolliffe and Stephenson, 2012), which shows the model performance for each category. The range of CSI is from 0 to 1.

The worst $O_3$ level at Kennewick was unhealthy for sensitive groups (AQI 3) during our study period (2017–2020), excluding the days when the air quality was affected by wildfire smoke. For a multi-category case such as in this study [AQI 1—Good, 2—Moderate, 3—Unhealthy for Sensitive Groups], we use the 3 × 3 contingency table in **Table 2** (Doswell and Keller, 1990). The skill scores are computed as follows (Jolliffe and Stephenson, 2012).

**TABLE 1 |** A 2 × 2 contingency table for forecast skill.

| Forecasts | Observations | | |
|---|---|---|---|
| | **Unhealthy** | **Good** | **Total** |
| Unhealthy | a = hits | b = false alarms | a + b |
| Good | c = misses | d = correct negatives | c + d |
| Total | a + c | b + d | a + b + c + d |

**TABLE 2 |** A 3 × 3 contingency table for forecast skill.

| Model AQI | Observed AQI | | |
|---|---|---|---|
| | **1** | **2** | **3** |
| 1 | $n_{11}$ | $n_{12}$ | $n_{13}$ |
| 2 | $n_{21}$ | $n_{22}$ | $n_{23}$ |
| 3 | $n_{31}$ | $n_{32}$ | $n_{33}$ |

$$HSS = \left(\sum_{i=1}^{3} p_{ii} - \sum_{i=1}^{3} p_i\hat{p}_i\right) / \left(\left(1 - \sum_{i=1}^{3} p_i\hat{p}_i\right)\right) \tag{5}$$

$$KSS = \left(\sum_{i=1}^{3} p_{ii} - \sum_{i=1}^{3} p_i\hat{p}_i\right) / \left(\left(1 - \sum_{i=1}^{3} p_i p_i\right)\right) \tag{6}$$

$$CSI_i = n_{ii} / \left(\sum_{i=1}^{3} n_i + \sum_{i=1}^{3} \hat{n}_i - n_{ii}\right) \tag{7}$$

The $p_{ii}$ is the sampling frequency when the observed and model predicted AQI is $i$, and $p_i$ and $\hat{p}_i$ are the observed and model predicted sample frequency when AQI = $i$. The $n_{ii}$ is the number of hits for $AQI_i$, and $n_i$ and $\hat{n}_i$ are the observed and model predicted event numbers when AQI = $i$.

# RESULTS AND DISCUSSION

## $O_3$ Observations at Kennewick, WA

This research covers the $O_3$ observations during the ozone seasons (May–September) from 2017 to 2020. The boxplot in **Figure 2** shows that the maximum daily 8-h moving average $O_3$ (MDA8 $O_3$) observations have decreased from 2017 through 2020. The 2017 and 2018 were fire years, which means they had several regional wildfire events, and there were fewer in 2019 and 2020. The COVID-19 pandemic in 2020 also reduced traffic and other air pollutant emissions.

The Washington State Department of Ecology explored the general relationship between $O_3$ level and temperature in the PNW and found that some MDA8 $O_3$ was beyond the normal level when the wildfire smoke was presented and there were 4 days identified in 2017–2018: no day identified in 2019–2020. The days affected by wildfire smoke in 2017–2020 have only about 0.75% occurrence rate, which is considered too rare to be predicted well by our ML models. Also, the wildfire smoke effect

is not easily predictable, so we exclude these 4 days affected by wildfire smoke from the dataset in this research to avoid the noise brought by the wildfire effects.

**Table 3** presents the general statistics of the MDA8 $O_3$ observations during the simulated period from May to September in 2017–2020. Here, we define a high-$O_3$ day as a day when the observed AQI category is worse than Moderate (i.e., AQI category > 2), which is considered an unhealthy $O_3$ event. There are six "high-$O_3$ days" for sensitive groups (i.e., AQI category 3) in 2017 and four in 2018. AIRPACT struggled to predict these high-$O_3$ days, and it missed all of the 10 "high-$O_3$ days". It is important to note that there were no unhealthy $O_3$ events in 2019 and 2020, and the forecasting performance of AIRPACT was better in 2019–2020 than in 2017–2018. It should also be noted that 2020 included potential emission reductions associated with COVID-19 reduced human activities. These emission changes were not incorporated into the AIRPACT emission system. However, ML models implicitly capture changes in emissions when relationships between meteorology and observed $O_3$ concentrations are updated during regular re-training (see section Ensemble Forecasts in 2019 and 2020).

## Machine Learning Model Evaluation at Kennewick WA

Cross-validation is commonly used for machine learning model evaluation by testing on subsets of the data (Raschka, 2015). Among various cross-validation methods available, we use both the 10-time, 10-fold, and walk-forward cross-validation techniques to evaluate our modeling framework. The result from the walk-forward cross-validation methods agrees with the 10-time 10-fold cross-validation, so we present the walk-forward cross-validation results in the **Supplementary Materials**.

For evaluation purposes, these forecasted hourly or 8-h averaged $O_3$ are computed into MDA8 $O_3$. We compare the evaluation results of this machine learning modeling framework against the AIRAPCT air quality forecasts for Kennewick, WA. This allows us to test how well this new machine learning-based forecasting system performs with respect to the existing CTM-based modeling framework.

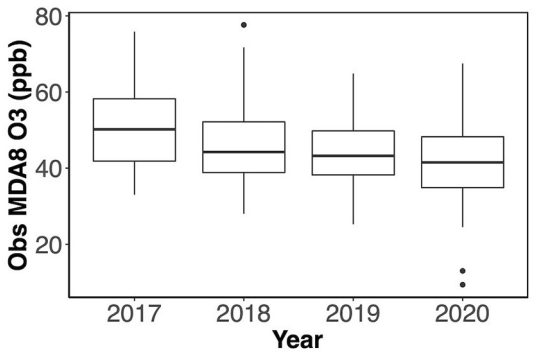


**FIGURE 2 |** Boxplot of observed (obs) MDA8 $O_3$ from May to September in 2017 – 2020.

## Feature Selection for Machine Learning Models

We initially provide 10 types of input data for the RF classifier and regression models and 11 types of input data for the MLR model; the additional data in the MLR model is the AQI classification. Since using too many features can cause an overfitting problem (Murphy, 2012), we used the following functions to do feature selection: *feature_importances_* in function *RandomForestClassifier/RandomForestRegressor* and *ranking_* in *RFE*. The selected features were preprocessed by *MaxAbsScaler* in the Python module *scikit-learn* and then used as input to train the model (Pedregosa et al., 2011).

For the RF classifier model used in ML1 and RF regression model in ML2, the feature selection function with the default setting computed the importance weights, and then the features with weights greater than the mean weight were selected. In this study, the mean weight is 0.1, so only features with weights >0.1 were selected: see the blue lines in **Figures 3A–C**. **Figure 3A** shows the weights of the features for the RF classifier model. The feature weights changed in each training process, but the ranking showed very little change. For instance, the previous day's $O_3$ observation and temperature were always the selected features, and the relative humidity, surface pressure and wind direction were selected in some cases.

The feature selection results of two-phase RF regression are shown in **Figures 3B,C**. Similar to the RF classifier model, the previous day's $O_3$ observation, temperature, and relative humidity were mostly above the 0.1 weight and thus were selected, but the ranking of the importance weights varied in the two phases. For the first phase RF regression model

**TABLE 3 |** Summary of historical air quality information from May to September in 2017–2020.

| Year | Simulated days | Mean | Median | 25th percentile | 75th percentile | # of days for each AQI | | | AQI > 2 |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 1 | 2 | 3 | |
| 2017 | 100 | 51 | 50 | 42 | 58 | 65 | 29 | 6 | 6.0% |
| 2018 | 148 | 46 | 44 | 39 | 52 | 119 | 25 | 4 | 2.7% |
| 2019 | 136 | 44 | 43 | 38 | 50 | 121 | 15 | 0 | 0 |
| 2020 | 142 | 42 | 42 | 35 | 48 | 132 | 10 | 0 | 0 |
| Total | 526 | 45 | 44 | 38 | 51 | 437 | 79 | 10 | 1.9% |

*The AQI categories are based on $O_3$ mixing ratios only.*

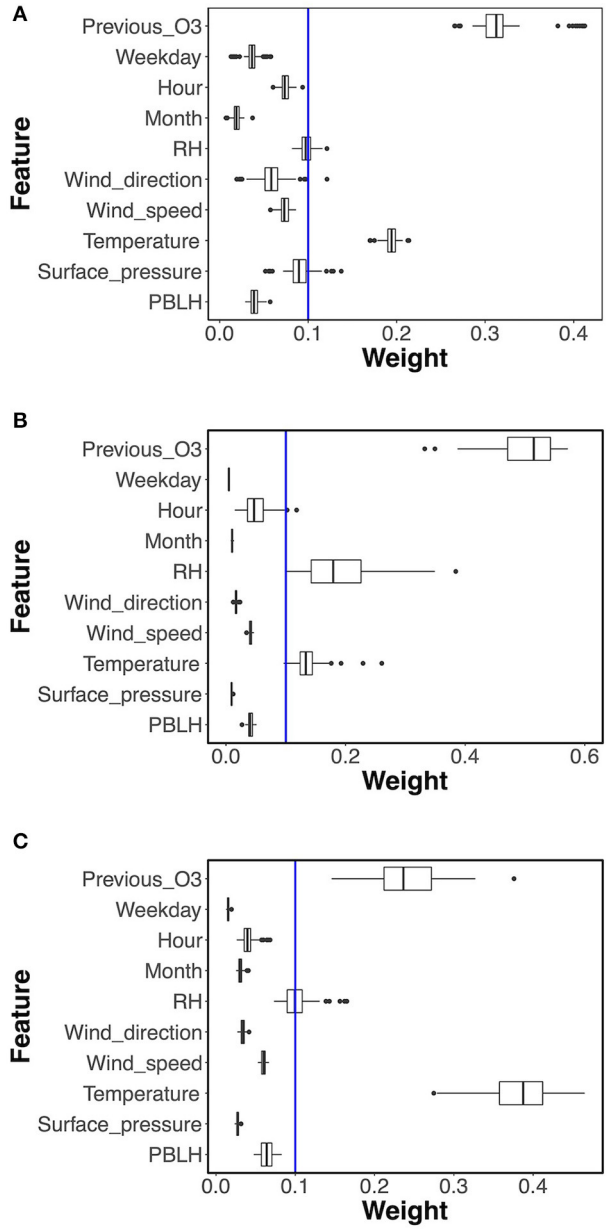


**FIGURE 3 |** Boxplot of feature weights from **(A)** RF classifier model in ML1, **(B)** the first and **(C)** the second RF regression model in ML2. The blue lines show the mean of the feature weights (0.1).

shown in **Figure 3B**, the previous day's $O_3$ observation was the most important feature, while the relative humidity was more important than temperature. The temperature became the most important feature in the second phase, while the previous day's $O_3$ observation ranked second and the relative humidity was selected in some cases.

For the MLR model used in ML1, the built-in feature selection function chose five features, which were AQI category, previous day's $O_3$ observation, relative humidity, and surface pressure for all training processes, while the fifth selected feature was either temperature, PBLH, or month.

## 10-Time, 10-Fold Cross-Validation

The *k*-fold cross-validation is one of the most commonly used techniques for machine learning model evaluation (Raschka, 2015). It first divides the dataset into *k* randomly chosen subsets. Then $k - 1$ subsets are used to train the model, while the remaining portion, which is not used in the training process, is used to test the model. This process is repeated k times to test all *k* subsets: every time, the "test" dataset is not used during the training process. In this study, we use $k = 10$, which is termed a 10-fold cross-validation. The *RepeatedKFold* function in the Python module *scikit-learn* is used to separate the dataset (Pedregosa et al., 2011). To avoid any bias from data separation, the 10-fold cross-validation is repeated 10 times (**Figure 4**) in this research.

The overall performance statistics of the 10-time, 10-fold cross-validations of the $O_3$ prediction are presented in **Table 4**. The mean normalized mean bias (NMB) and normalized mean error (NME) are $5.5 \pm 0.2$ and $16 \pm 0.1\%$ for ML1, $-0.14 \pm 0.05$ and $12 \pm 0.1\%$ for ML2, respectively. The low standard deviations show that there is no significant difference between each of the 10 times training conducted, indicating that the model performance is stable. The AIRPACT NMB and NME are 1.1% and 17% when using all data points, which is comparable to the ML performance. Interestingly, AIRPACT has eight extremely over-predicted $O_3$ days during the period used in this study. When these extreme values are excluded, its NMB and NME are changed to −2.2% and 14%, respectively. AIRPACT with all data points has a poor correlation ($R^2 = 0.070$), but without the eight extreme values, the $R^2$ is 0.38, which becomes comparable to results from the ML models (i.e., $R^2$ of 0.43 and 0.54). When comparing all models, ML2 has the highest $R^2$ and the lowest NMB and NME among the three models. We observe similar performance for the ML models using walk-forward method as the 10-time, 10-fold cross-validation (see **Supplementary Table 1**).

The CSI scores show the model performance for each AQI category. Based on the CSI values, ML2 performs better for the days with AQI 1 (which is the category that most of our $O_3$ data fall into), and ML1 performs better for higher $O_3$ (AQI > 1). AIRPACT and ML2 do not capture the days with AQI > 2, while ML1 captures 5 out of 10 high $O_3$ cases. The better performance of ML1's high $O_3$ predictions leads to higher HSS and KSS scores, especially for KSS, which is about two times of AIRPACT and ML2. This makes sense because KSS is sensitive to high-$O_3$ events.

ML1 performs better in the high-$O_3$ cases, and it is likely due to the linear relationship used in the MLR model that is not as sensitive to the range of data in the training data. Conversely, the RF model in ML2 may not work well when the input data exceeds the range of the training data as it uses an ensemble of decision trees and thus can be limited by the training dataset. The indoor air quality study by Yuchi et al. (2019) that used RF and MLR models drew a similar conclusion.

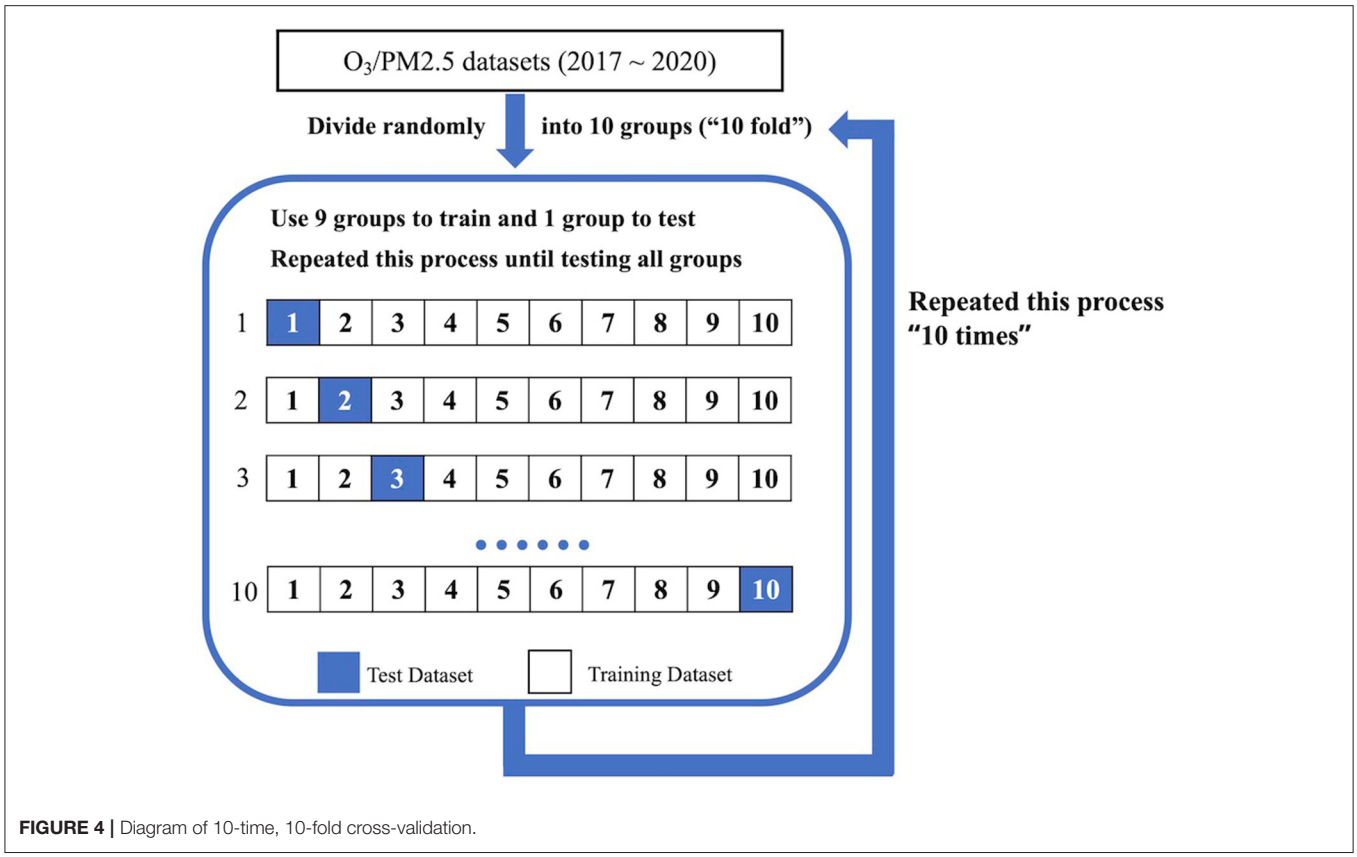


**FIGURE 4 |** Diagram of 10-time, 10-fold cross-validation.

**TABLE 4 |** Statistics and forecast verifications of the 10-time, 10-fold cross-validations of the simulated $O_3$ at Kennewick, WA during 2017–2020.

| | | AIRPACT | AIRPACT (w/o eight extreme values) | ML1 | ML2 |
|---|---|---|---|---|---|
| $R^2$ | | 0.070 | 0.38 | 0.43 | 0.54 |
| NMB (%) | | 1.1 | −2.2 | 5.2 | −0.22 |
| NME (%) | | 17 | 14 | 16 | 12 |
| HSS | | 0.34 | 0.34 | 0.42 | 0.4 |
| KSS | | 0.30 | 0.30 | 0.61 | 0.33 |
| CSI | 1 | 0.85 | 0.85 | 0.74 | 0.87 |
| | 2 | 0.24 | 0.24 | 0.34 | 0.27 |
| | 3 | 0 | 0 | 0.28 | 0 |

**Figures 5A–C** show the ratio of the model predictions to the observations vs. the observed MDA8 $O_3$ for the AIRPACT, ML1, and ML2 models. To better compare the performance of the three models, the y-axis is set to the same range for all figures, so some extreme values are excluded. Interestingly, all models show a similar systematic bias: over-prediction of low MDA8 $O_3$ and under-prediction of high MDA8 $O_3$. This figure also shows that ML1 tends to predict higher $O_3$ levels than AIRPACT and ML2 for all mixing ratio ranges.

The results above demonstrate that our ML-based forecasts are comparable to AIRPACT except for the high-$O_3$ cases where the ML models clearly perform better. This means the ML models may not outperform the AIRPACT model if there is no high-$O_3$ event. Additionally, given the systematic biases shown in **Figures 5A–C** are strongly associated with the $O_3$ levels, the model performance will definitely vary by the distribution of the observed $O_3$ levels. Since the average $O_3$ levels have decreased from 2017 to 2020 and the year 2019 and 2020 did not have any high-$O_3$ event (AQI $>$ 2; see Section $O_3$ Observations at Kennewick, WA for the details), we perform the 10-time, 10-fold cross-validations for each year from 2017 to 2020 to explore the changes in the model performances (see **Table 5**). In addition to the AIRPACT, ML1, and ML2 models, **Table 5** includes a "combined" model that is based on our forecast modeling framework that uses ML1 forecasts when the predicted MDA8 $O_3$ is higher than 70 ppb and ML2 forecasts for all other cases. The time series of MDA8 $O_3$ in **Figure 6** shows that both AIRPACT and the combined ML predictions follow the trend of observations. Machine Learning predictions are generally closer to the observations and do not largely over-predict the MDA8 $O_3$; however, AIRPACT generates several extremely over-predicted $O_3$ events in 2017 and 2020. It should be noted that the "combined" results are available for only 2017 and 2018 because there are no unhealthy $O_3$ events in 2019 and 2020, so that only the ML2 model is used for those years.

**Table 5** shows how the model performance can vary year-to-year due to changes in $O_3$ distribution. The changes in the model performance can be explained by the systematic biases

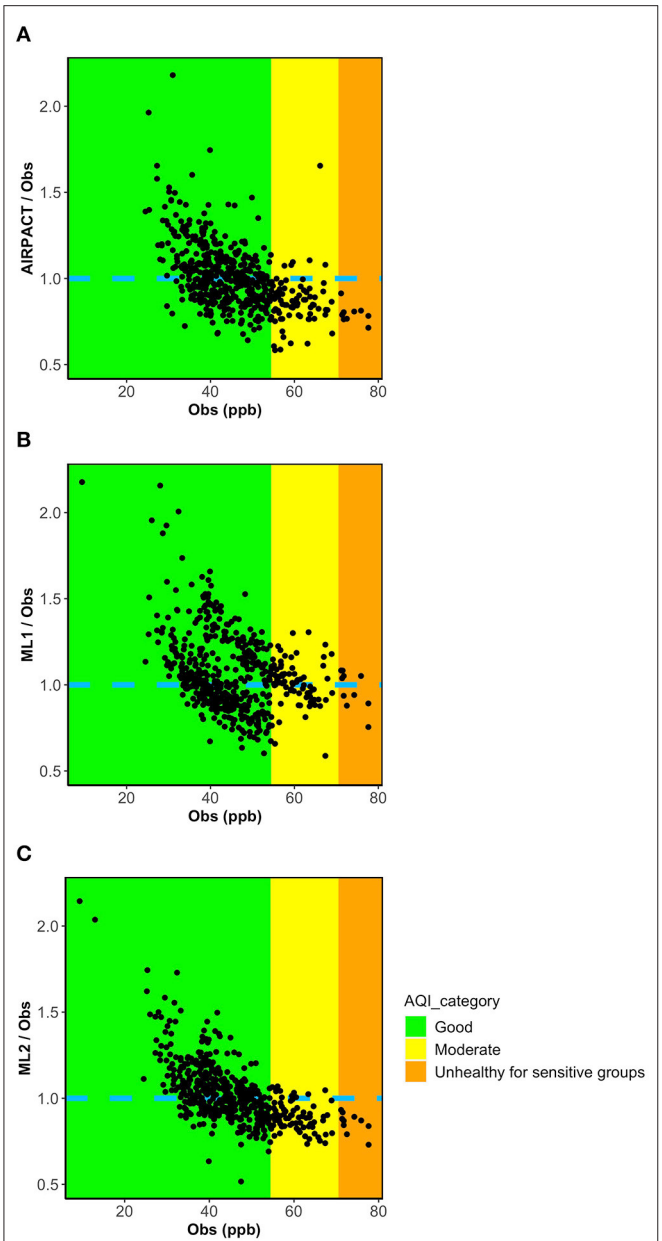


**FIGURE 5 |** Ratio plots of model prediction to observations vs. observations for three models **(A)** AIRPACT, **(B)** ML1, and **(C)** ML2.

trend. As the $O_3$ levels go down from 2017/2018 to 2019/2020, the model performance moves from under-prediction to over-prediction: models tend to over-predict the lower $O_3$ levels. The walk-forward method performs similarly to the 10-time, 10-fold cross-validation (see **Supplementary Figure 1**). Compared to the ML models, AIRPACT shows larger variations in the yearly performance, which is likely to be influenced by other changes in the AIRPACT simulations (Munson et al., 2021). The NMB of AIRPACT in 2017 is close to 0 (−1.7%). This is because of its extreme over-prediction in some cases. If they are excluded from the statistics, the NMB of AIRPACT is −12% in 2017. The same reason is attributed to the 12% over-prediction in 2020, and it is 7.3% after removing the extreme predictions. So, excluding the extreme predictions, the NMB from AIRPACT generally reveals the over-prediction of lower $O_3$ level and under-prediction of higher $O_3$ level. Similarly, ML1 and ML2 show higher NMB in 2019/2020 than in 2017/2018.

Despite these differences, the yearly validation results still show similar performance for the ML models: ML1 performs better for AQI > 2 while ML2 performs better for the other cases. There are unhealthy $O_3$ cases (AQI 3) in 2017 and 2018, and ML1 captures half of them. This leads to mostly better statistics than AIRPACT and ML2. The KSS score of ML1 is significantly higher than other models, which is because it is sensitive to the high-$O_3$ predictions. ML2 has a good performance for low-$O_3$ predictions, and the $CSI_1$ and $CSI_2$ scores are close or better than AIRPACT. Although the $R^2$ values of ML2 decrease in 2019 and 2020, the high $CSI_1$ scores (∼0.9) still show its accurate low-$O_3$ predictions.

ML2 performs better for the low-$O_3$ predictions and has higher $CSI_1$ scores than ML1, while ML1 can capture more high-$O_3$ events with good $CSI_3$ scores. The combined approach keeps the high $CSI_1$ scores as ML2 and captures some unhealthy $O_3$ events in 2017 and 2018. The $R^2$ of the combined model ($R^2$ = 0.57 and 0.58 in 2017 and 2018) is better than ML1 ($R^2 = 0.44$ and 0.46), but slightly worse than ML2 ($R^2 = 0.58$ and 0.64), because ML2 performs better for the low-$O_3$ days that are dominant in the observation datasets.

## Ensemble Forecasts in 2019 and 2020

Beginning in May 2019, the ML modeling framework has been used to provide 72-h "ensemble" operational $O_3$ forecasts each day for Kennewick, which uses 27 WRF ensemble forecasts from the University of Washington[2]. The ensemble WRF forecasts use multiple initial and boundary conditions, and various physical parameterizations and surface properties (Mass et al., 2003). We predict $O_3$ levels with each WRF member to compute a 72-h forecast and then these individual forecasts are combined to yield an ensemble mean forecast with an associated uncertainty range. The forecasts are available to the public[3], with the ability to sign up for email alerts if "unhealthy for sensitive groups" or worse AQI levels are forecasted. To increase the size of the training dataset and improve the forecast accuracy, we include the new observational data from the previous day and re-train the models daily.

We present the evaluation of the operational ensemble forecasts covering May to September in 2019 and 2020 in **Table 6**. The meteorology data used in the cross-validation is extracted from the WRF output that provided input data for AIRPACT, and it is named WRFRT. Most of the statistical variables in **Table 6** show that the performance of the ensemble mean is close to the single WRFRT forecasts. By using the ensemble WRF forecasts in the ML forecasting system, the variations of the meteorological forecasts are taken into consideration, although

[2]https://a.atmos.washington.edu/wrfrt/ensembles/info.html
[3]http://ozonematters.com

**TABLE 5 |** Annual statistics and forecast verifications of the 10-time, 10-fold cross-validations at Kennewick, WA.

| | | AIRPACT | | | | ML1 | | | | ML2 | | | | Combined* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 2017 | 2018 | 2019 | 2020 | 2017 | 2018 | 2019 | 2020 | 2017 | 2018 | 2019 | 2020 | 2017 | 2018 |
| $R^2$ | | 0.0053 | 0.46 | 0.43 | 0.029 | 0.44 | 0.46 | 0.34 | 0.33 | 0.58 | 0.64 | 0.43 | 0.44 | 0.57 | 0.58 |
| NMB (%) | | −1.7 | −7.5 | 2.5 | 12 | 2.3 | 4.3 | 6.6 | 7.1 | −6.3 | −1.8 | 1.4 | 5.3 | −4.5 | −0.36 |
| NME (%) | | 25 | 14 | 12 | 19 | 15 | 15 | 16 | 18 | 12 | 10 | 11 | 14 | 13 | 11 |
| HSS | | 0.30 | 0.31 | 0.47 | 0.28 | 0.41 | 0.55 | 0.31 | 0.31 | 0.30 | 0.51 | 0.32 | 0.35 | 0.30 | 0.52 |
| KSS | | 0.26 | 0.22 | 0.43 | 0.45 | 0.45 | 0.73 | 0.59 | 0.77 | 0.25 | 0.43 | 0.24 | 0.35 | 0.27 | 0.44 |
| CSI | 1 | 0.72 | 0.86 | 0.90 | 0.86 | 0.65 | 0.81 | 0.72 | 0.77 | 0.73 | 0.89 | 0.89 | 0.91 | 0.73 | 0.89 |
| | 2 | 0.24 | 0.17 | 0.35 | 0.23 | 0.37 | 0.45 | 0.27 | 0.24 | 0.24 | 0.35 | 0.22 | 0.25 | 0.14 | 0.32 |
| | 3 | 0 | 0 | – | – | 0.30 | 0.25 | – | – | 0 | 0 | – | – | 0.30 | 0.25 |

*Combined refers to using the ML1 predicted MDA8 $O_3$ predictions for high-$O_3$ days and the ML2 predictions for all other days.*

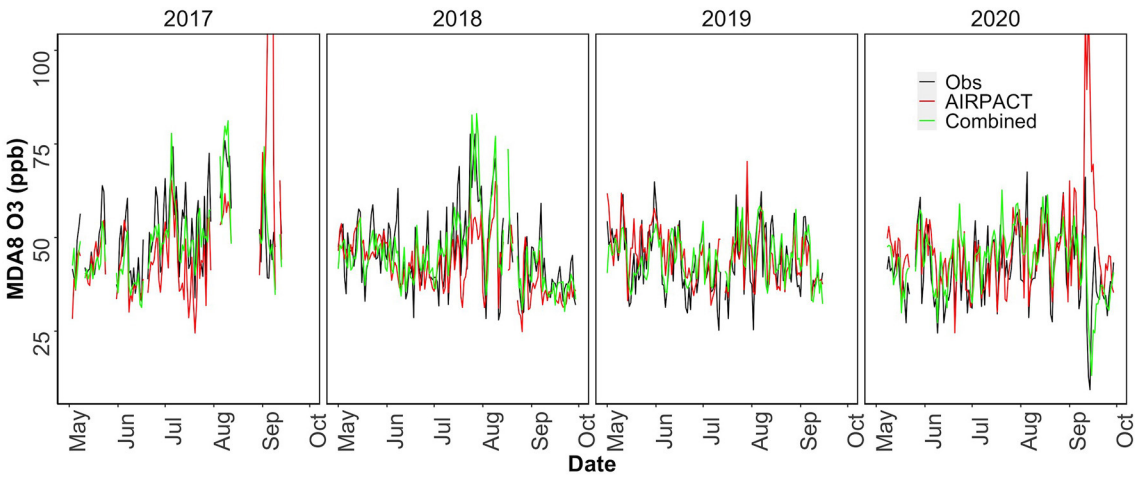


**FIGURE 6 |** Time series of MDA8 $O_3$ from observation, AIRPACT and combined ML model predictions from May to September in 2017–2020.

**TABLE 6 |** Statistics and forecast verifications in 2019–2020.

| | | ML1 (mean) | ML1 (WRFRT) | ML2 (mean) | ML2 (WRFRT) |
|---|---|---|---|---|---|
| $R^2$ | | 0.33 | 0.35 | 0.49 | 0.48 |
| NMB (%) | | 6.9 | 8.0 | 5.2 | 5.7 |
| NME (%) | | 17 | 18 | 12 | 13 |
| HSS | | 0.31 | 0.28 | 0.41 | 0.47 |
| KSS | | 0.64 | 0.66 | 0.39 | 0.44 |
| CSI | 1 | 0.75 | 0.70 | 0.90 | 0.91 |
| | 2 | 0.26 | 0.24 | 0.30 | 0.34 |

*Note that mean is the ensemble means of the MDA8 $O_3$ forecasts of the ensemble members, and WRFRT is the single WRF data that drives AIRPACT.*

the overall difference between the averaged MDA8 $O_3$ and the ensemble members is not significant (within 5%).

The distributions of the averaged ensemble MDA8 $O_3$ predictions are shown in **Figures 7A,B**. Due to the missing data for some ensemble members, 21 ensemble members are presented in total. The ML1 distributions have two peaks because it first classifies the AQI categories using the RF classifier model. The peaks from ensemble members are higher than the averaged distribution in **Figure 7A**, and the ensemble-averaged prediction can relatively weaken the bias from a single ensemble WRF member. The distribution of ML2 is close to AIRPACT as shown in **Figure 7B**. Both ML1 and ML2 do not over-predict MDA8 $O_3$ very much, while AIPRACT can severely over-predict some high MDA8 $O_3$ events.

## CONCLUSIONS

Chemical transport models are widely used for air quality modeling and forecasting, but they may fail to properly forecast pollution episodes, plus they are computationally expensive. AIRPACT is a CTM-based operational forecasting system for the Pacific Northwest, but it has a history of failing to predict high-$O_3$ events at Kennewick, WA during summer and fall. In this research, we developed machine learning models that use historical WRF meteorology and $O_3$ observation data to build a more reliable forecast system with much less computational burden. The new forecast framework consists of two ML models, ML1 and ML2, that predict the $O_3$ mixing ratios and AQI

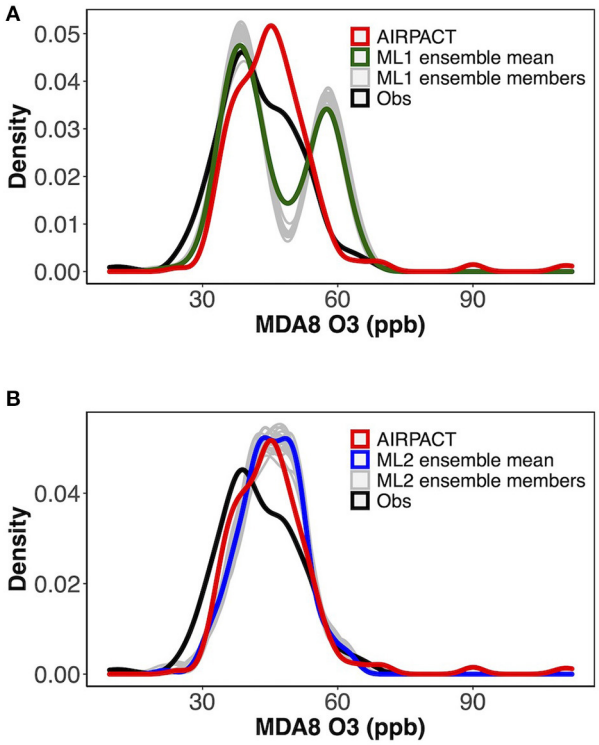


**FIGURE 7 |** Distributions of observed and **(A)** ML1, **(B)** ML2 model predicted MDA8 $O_3$ in 2019 and 2020.

categories. To evaluate and demonstrate this new forecast system, we applied the system to observations from Kennewick, WA over several years.

The $O_3$ observations and archived WRF meteorology data (temperature, surface pressure, relative humidity, wind speed, wind direction, and PBLH from 2017 to 2020) were used in the training dataset. ML1 uses both RF classifier and MLR models, and ML2 uses a two-phase RF regression model with weighting factors. The 10-time, 10-fold, and walk-forward cross-validation methods were used to evaluate the modeling framework, and the results agree with each other.

Comparing the statistics of the three models, ML2 has the highest $R^2$ (0.54) and lowest NMB (−0.22%) and NME (12%). The CSI values from the 10-time, 10-fold cross-validation showed that ML1 performs better for the high MDA8 $O_3$ prediction ($CSI_3 = 0.28$), and ML2 performs better for the low MDA8 $O_3$ predictions ($CSI_1 = 0.87$). Given this, our operational forecast system combines ML1 when $O_3$ is higher than 70 ppb with ML2 for all other cases.

The ML models provided improved predictions (most $R^2 > 0.5$) and correctly predicted 5 out of 10 high pollution events, while AIRPACT misses all these events. Also, the model performance of the ML modeling framework was more stable without extreme predictions: AIRPACT predicts eight extremely high MDA8 $O_3$ in 2017 and 2020.

Interestingly, we find similar systematic biases from all models; they tend to over-predict the low $O_3$ levels and under-predict high $O_3$ levels. Due to the systematic biases and decreasing trend of $O_3$ from 2017 to 2020, our ML modeling framework performs better than AIRPACT in 2017 and 2018, but shows no improvement in 2019 and 2020. Without unhealthy-$O_3$ events in 2019 and 2020, the ML modeling framework cannot demonstrate its superior capability for high $O_3$ events.

With about 4 min of CPU time, the ML modeling framework makes it possible to provide the ensemble daily forecast of $O_3$ level at Kennewick WA; AIRPACT needs 120 processors for 3 h (360 h of CPU time) throughout the PNW for one single WRF output. The 72-h "ensemble" operational $O_3$ forecasts have been provided by this ML modeling framework each day since May 2019. The ensemble mean forecasts take the ensemble model configurations of WRF forecasts into consideration.

Overall, our ML modeling framework is shown to be well-suited for predicting ground-level $O_3$ at a specific location using much less computational resources and fewer input datasets than CTMs. Our ML modeling framework has been successfully expanded to predict $O_3$ as well as PM2.5 at various AQS sites throughout the PNW region, which will be presented in a subsequent paper. We find that our ML models provide comparable predictability as CTMs (and even excels in some cases) at the locations we have studied (i.e., AQS monitoring sites). However, compared to CTMs, our ML models have a few obvious weaknesses. For instance, ML methods cannot provide predictions over a large domain where there are few monitoring stations, and these methods do not include physical and chemical processes. There are other exciting ML innovations that may help to overcome such weaknesses. We believe ML models can replace CTMs for some specific tasks (e.g., forecasts at specific locations) and a hybrid modeling approach of ML and CTM models could be very beneficial to overcome some of the continuing challenges in traditional atmospheric models.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Materials**, further inquiries can be directed to the corresponding author/s.

## AUTHOR CONTRIBUTIONS

YL and KH conceptualized the overall study. KF implemented the machine learning models and performed the experiments and validations/analysis with the support of YL, KH, BL, and RD. Datasets used in this study were curated by KF and RD. KF had the lead in writing the manuscript with contributions from YL, and all authors revised the final manuscript. RL participated in this study for a few months as an undergraduate researcher.

## FUNDING

This work was partially funded by the Center of Advanced Systems Understanding (CASUS) which is financed by Germany's Federal Ministry of Education and Research (BMBF) and by the Saxon Ministry for Science, Culture and

Tourism (SMWK) with tax funds on the basis of the budget approved by the Saxon State Parliament.

## ACKNOWLEDGMENTS

The authors acknowledge David Ovens from the University of Washington for his help to set up a data feed of WRF ensembles.

This manuscript was firstly preprinted at EarthArXiv on May 13, 2020 (doi: 10.31223/osf.io/bdpnm).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2022.781309/full#supplementary-material

## REFERENCES

Arganis, M. L., Val, R., Dominguez, R., Rodriguez, K., Dolz, J., and Eato, J. M. (2012). "Comparison between equations obtained by means of multiple linear regression and genetic programming to approach measured climatic data in a river," in Genetic Programming—New Approaches and Successful Applications, ed S. Ventura Soto (London: InTech), 239–254.

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Chaloulakou, A., Assimacopoulos, D., and Lekkas, T. (1999). Forecasting daily maximum ozone concentrations in the Athens basin. *Environ. Monit. Assess.* 56, 97–112.

Doswell, C. A. III, Davies-Jones, R., and Keller, D. L. (1990). On summary measures of skill in rare event forecasting based on contingency tables. *Weather Forecast.* 5, 576–585.

Feng, X., Li, Q., Zhu, Y., Hou, J., Jin, L., and Wang, J. (2015). Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation. *Atmos. Environ.* 107, 118–128. doi: 10.1016/j.atmosenv.2015.02.030

Freeman, B. S., Taylor, G., Gharabaghi, B., and Thé, J. (2018). Forecasting air quality time series using deep learning. *J. Air Waste Manage. Assoc.* 68, 866–886. doi: 10.1080/10962247.2018.1459956

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. (2017). Learning from class-imbalanced data: review of methods and applications. *Expert Syst. Appl.* 73, 220–239. doi: 10.1016/j.eswa.2016.12.035

Jiang, N., and Riley, M. L. (2015). Exploring the utility of the random forest method for forecasting ozone pollution in SYDNEY. *J. Environ. Protect. Sustain. Dev.* 1, 12–20.

Jobson, B. T., and, VanderSchelden, G. (2017). *The Tri-Cities Ozone Precursor Study (T-COPS)* [Final Report]. Washington Department of Ecology. Available online at: https://ecology.wa.gov/DOE/files/93/934a2f46-b000-4f9a-837ca286ccfa615e.pdf (accessed April 23, 2020).

Jolliffe, I. T., and Stephenson, D. B. (2012). Forecast *Verification: A Practitioner's Guide in Atmospheric Science*. Chichester: John Wiley and Sons.

Kam, H. T. (1995). "Random decision forest," in *Proceedings of the 3rd International Conference on Document Analysis and Recognition* (Montreal, QC).

Mass, C. F., Albright, M., Ovens, D., Steed, R., Maciver, M., Grimit, E., et al. (2003). Regional environmental prediction over the Pacific Northwest. *Bull. Am. Meteorol. Soc.* 84, 1353–1366. doi: 10.1175/BAMS-84-10-1353

Mitchell, T. M. (1997). *Machine Learning*. New York, NY: McGraw-Hill.

Moustris, K. P., Nastos, P. T., Larissi, I. K., and Paliatsos, A. G. (2012). Application of multiple linear regression models and artificial neural networks on the surface ozone forecast in the greater Athens Area, Greece. *Adv. Meteorol.* 2012, 1–8. doi: 10.1155/2012/894714

Munson, J., Vaughan, J. K., Lamb, B. K., and Lee, Y. (2021). *Decadal Evaluation of the AIRPACT Regional Air Quality Forecast System in the Pacific Northwest* from 2009−*2018*. Washington State University.

Murphy, K. P. (2012). Machine *Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Pernak, R., Alvarado, M., Lonsdale, C., Mountain, M., Hegarty, J., and Nehrkorn, T. (2019). Forecasting surface O3 in Texas Urban areas using random forest and generalized additive models. *Aerosol Air Qual. Res.* 9, 2815–2826. doi: 10.4209/aaqr.2018.12.0464

Raschka, S. (2015). *Python Machine Learning*. Birmingham: Packt Publishing Ltd.

Rybarczyk, Y., and Zalakeviciute, R. (2018). Machine Learning approaches for outdoor air quality modelling: a systematic review. *Appl. Sci.* 8, 2570. doi: 10.3390/app8122570

Seinfeld, J. H., and Pandis, S. N. (2016). Atmospheric *Chemistry* and *Physics: From Air Pollution to Climate Change*. Hoboken, NJ: John Wiley and Sons.

Sousa, S., Martins, F., Alvimferraz, M., and Pereira, M. (2007). Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environ. Modell. Softw.* 22, 97–103. doi: 10.1016/j.envsoft.2005.12.002

Sportisse, B. (2007). A review of current issues in air pollution modeling and simulation. *Comput. Geosci.* 11, 159–181. doi: 10.1007/s10596-006-9036-4

Washington State Office of Financial Management (2020). *April 1, 2020 Population of Cities, Towns and Counties Used for Allocation of Selected State Revenues State of Washington*. Washington State Office of Financial Management. Availabel online at: https://ofm.wa.gov/sites/default/files/public/dataresearch/pop/april1/ofm_april1_population_final.pdf (accessed March 10, 2020).

Weaver, C. P., Liang, X.-Z., Zhu, J., Adams, P. J., Amar, P., Avise, J., et al. (2009). A preliminary synthesis of modeled climate change impacts on U.S. regional ozone concentrations. *Bull. Amer. Meteorol. Soc.* 90, 1843–1864. doi: 10.1175/2009BAMS2568.1

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*, Vol. 100. Oxford: Academic Press.

Yu, R., Yang, Y., Yang, L., Han, G., and Move, O. (2016). RAQ–a random forest approach for predicting air quality in urban sensing systems. *Sensors* 16, 86. doi: 10.3390/s16010086

Yuchi, W., Gombojav, E., Boldbaatar, B., Galsuren, J., Enkhmaa, S., Beejin, B., et al. (2019). Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. *Environ. Pollut.* 245, 746–753. doi: 10.1016/j.envpol.2018.11.034

Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., and Talebiesfandarani, S. (2019). PM2.5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. *Atmosphere* 10, 373. doi: 10.3390/atmos10070373

Zhan, Y., Luo, Y., Deng, X., Grieneisen, M. L., Zhang, M., and Di, B. (2018). Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. *Environ. Pollut.* 233, 464–473. doi: 10.1016/j.envpol.2017.10.029

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Copyright © 2022 Fan, Dhammapala, Harrington, Lamastro, Lamb and Lee. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*