



1

2 **Using wavelet transform and dynamic time warping to identify the limitations**  
3 **of the CNN model as an air quality forecasting system**

4

5 Ebrahim Eslami<sup>1</sup>, Yunsoo Choi<sup>1,\*</sup>, Yannic Lops<sup>1</sup>, Alqamah Sayeed<sup>1</sup>, Ahmed Khan Salman<sup>1</sup>

6 <sup>1</sup>Department of Earth and Atmospheric Sciences, University of Houston, Houston, TX 77204,  
7 United States

8

9 \*Corresponding author: [ychoi6@uh.edu](mailto:ychoi6@uh.edu)

10



11 **Abstract:**

12 As the deep learning algorithm has become a popular data analytic technique, atmospheric  
13 scientists should have a balanced perception of its strengths and limitations so that they can provide  
14 a powerful analysis of complex data with well-established procedures. Despite the enormous  
15 success of the algorithm in numerous applications, certain issues related to its applications in air  
16 quality forecasting (AQF) require further analysis and discussion. This study addresses significant  
17 limitations of an advanced deep learning algorithm, the convolutional neural network (CNN), in  
18 two common applications: (i) a real-time AQF model, and (ii) a post-processing tool in a dynamical  
19 AQF model, the Community Multi-scale Air Quality Model (CMAQ). In both cases, the CNN  
20 model shows promising accuracy for ozone prediction 24 hours in advance in both the United  
21 States and South Korea (with an overall index of agreement exceeding 0.8). For the first case, we  
22 use the wavelet transform to determine the reasons behind the poor performance of CNN during  
23 the nighttime, cold months, and high ozone episodes. We find that when fine wavelet modes  
24 (hourly and daily) are relatively weak or when coarse wavelet modes (weekly) are strong, the CNN  
25 model produces less accurate forecasts. For the second case, we use the dynamic time warping  
26 (DTW) distance analysis to compare post-processed results with their CMAQ counterparts (as a  
27 base model). For CMAQ results that show a consistent DTW distance from the observation, the  
28 post-processing approach properly addresses the modeling bias with predicted IOAs exceeding  
29 0.85. When the DTW distance of CMAQ-vs-observation is irregular, the post-processing approach  
30 is unlikely to perform satisfactorily. Awareness of the limitations in CNN models will enable  
31 scientists to develop more accurate regional or local air quality forecasting systems by identifying  
32 the affecting factors in high concentration episodes.

33  
34 **Keyword:** machine learning, neural networks, atmospheric chemistry, air quality modeling.



35 **1. Introduction:**

36 Currently, atmospheric scientists have shown significant interest in applying machine  
37 learning (ML) algorithms in their field, specifically for air quality forecasting, remote sensing data  
38 retrieval, and hurricane tracking. ML is a technique used for developing data-driven algorithms  
39 that learn to mimic human behavior on the basis of a prior example or experience. It is a tool that  
40 allows systems to more effectively deal with knowledge-intensive problems in complex domains,  
41 which occurs via learning that involves gathering information from a training dataset and using  
42 certain logic to purposefully detect a pattern of behavior. The fundamental goal of ML models is  
43 to apply the detected patterns to make generalizations beyond the examples in the training set.

44 Generalizations stemming from ML models provide a scope of improvement in a number  
45 of physical applications. Evidence of the growing interest in applying ML is the rapid increase in  
46 the number of scientific publications in this area, illustrated in Fig. S1. Inevitably, a consequence  
47 of such enthusiasm in the field is the risk of exaggerated expectations, fueled by results focusing  
48 on the general performance of ML models compared to that of conventional statistical models.  
49 Such examples can be found in studies by Eslami et al. (2019a, 2019b, 2019c), Choi et al. (2019),  
50 Sayeed et al. (2020), and Lops et al. (2019). To assume more reasonable expectations, we must  
51 first explore the current challenges we face when forecasting ambient air quality and then assess  
52 how or even whether ML models can address the challenges to produce more accurate forecasting.

53 To develop a capable air quality forecasting tool, atmospheric scientists often turn to  
54 chemical transport models (CTMs) and statistical models, both of which use meteorological  
55 parameters and chemical precursors from previous atmospheric conditions to estimate the  
56 following conditions. A brief summary of these models appears in Zhang et al. (2012). Although  
57 CTMs, with their dynamical implementation of atmospheric chemistry and physics, have shown  
58 promise in forecasting, they are too computationally intensive for operational real-time forecasts.  
59 Thus, computationally efficient statistical models such as ML have emerged as alternative  
60 approaches. Unlike CTMs, however, these models mainly rely on data from a network of  
61 monitoring stations that are sparsely distributed and measure a limited number of meteorology and  
62 air quality variables (Eslami et al., 2019a). Given the complexity of the formation/depletion of air  
63 pollutants such as ozone, this limitation may be vital in predicting extreme events (Eslami et al.,  
64 2019b).

65 Another challenge in predicting ozone concentration is the “external” relationships among  
66 predictors. For instance, as important meteorological parameters, temperature and solar radiation  
67 are synoptic factors while the wind field is influenced by regional factors such as geography and  
68 urbanization. Such conditions particularly affect ozone variability since locally-produced NO<sub>2</sub>  
69 emissions under certain meteorological circumstances lead to the formation of ozone that is later  
70 transported by the wind and detected by monitoring stations (Pan et al., 2015). Nevertheless,  
71 station-specific ML models use such chemical and meteorological variables as a footprint of local  
72 conditions.

73 Although local emissions of ozone precursors are the dominant source of ozone,  
74 particularly in urban areas, ozone pollution arising from sources outside of a target region, such as  
75 background ozone, inevitably degrade local air quality (Camalier et al., 2007). The lack of  
76 measurable environmental variables that indicate the potential long-range transport of air  
77 pollutants poses an unprecedented challenge for a ML model to estimate ozone concentrations  
78 over downwind communities (Eslami et al., 2019a). Because of the nonlinear spatial relationships  
79 between neighboring monitoring stations, ML models as operational real-time forecasting systems  
80 produce relative uncertainty.



81 A number of studies have proposed solutions addressing the above limitations of ML  
82 models. Eslami et al. (2019a) implement a deep convolutional neural network (CNN) (Krizhevsky  
83 et al., 2012) model that uses hourly values of several meteorological and air pollution variables to  
84 predict hourly ozone concentrations 24 hours in advance. Even though the accuracy of the  
85 forecasting system guarantees a reasonable level of accuracy, it fails to address high ozone  
86 episodes owing to the infrequent occurrences of such events, which lead to the undertraining of  
87 the CNN model. In another study, Eslami et al. (2019b) propose a data ensemble approach that  
88 mitigates this issue by regularizing the training dataset toward capturing high ozone episodes.  
89 While the authors remove a significant portion of the underprediction biases of the CNN model,  
90 its predictions of ozone during the nighttime and on rainy days are unreliable. Sayeed et al. (2020)  
91 use historical data covering a longer period within a diverse geographical domain (Texas) to train  
92 a similar CNN model. Their results from stations for which fewer measurements are available,  
93 while more accurate, are prone to uncertainty. Using the outputs of air quality and meteorological  
94 forecast models to map the hourly ozone concentrations at station locations, Choi et al. (2019)  
95 train a similar deep CNN model, a spatially generalized model that bias-corrects ozone forecasts  
96 of the community multi-scale air quality (CMAQ) model for all monitoring stations in the EPA  
97 AirNow network. Even though the model significantly improved CAMQ forecasts, the bias-  
98 correction process and the unbalanced CMAQ modeling outputs are unclear.

99 that uses wavelet transform and dynamic time warping (DTW) to  
100 This paper discusses the general inability of the machine learning model using wavelet  
101 transform and dynamic time warping (DTW). Wavelet transform is a powerful technique for  
102 analyzing the temporal variation of a time-series (Grinsted et al., 2004). Wavelet analysis uses an  
103 adjustable resolution to translate time-series data and then decomposes the data into a certain  
104 frequency level that cannot be achieved by other conventional methods such as Fourier analysis  
105 (Huang et al., 2010). DTW is a non-linear technique that measures any alignment between two  
106 times-series (i.e., model prediction and observation in this study) by warping them to match their  
107 similarities (Berndt and Clifford, 1994). By introducing two applications of CNN in the real-time  
108 ozone forecasting system, we use these analytical tools to identify the source of the prediction  
109 biases of the CNN model. In this paper, we do not describe the forecasting results in detail but  
110 instead refer the reader to studies by Eslami et al. (2019a, 2019b), Choi et al. (2019), and Sayeed  
111 et al. (2020).  
112

## 113 2. Materials and Methods

### 114 2.1. Deep convolutional neural networks:

115 The deep CNN model (Krizhevsky et al., 2012) is a common deep learning architecture  
116 that has long been used in numerous applications (Deng and Yu, 2014; Schmidhuber, 2015;  
117 Goodfellow et al., 2016; Litjens et al., 2017; Chen et al., 2018; Kamilaris and Prenafeta-Boldú,  
118 2018; Higham and Higham, 2019). Unlike other methods, the CNN model is capable of analyzing  
119 joint features and attaining greater accuracy on large-scale datasets. Deep CNNs can be trained to  
120 approximate smooth, highly nonlinear functions (LeCun et al., 2015), rendering them appropriate  
121 for analyzing nonlinear processes in the atmosphere. In addition, feature extraction using deep  
122 learning algorithms is more efficient than using other neural network methods, particularly when  
123 multiple hidden layers are structured (Krizhevsky et al., 2012).

124 A schematic for the deep CNN used in this paper appears in Fig. 3. The figure shows the  
125 input layer of the CNN algorithm, which represents the normalized time series of all input



126 variables. The normalization process prevents a steep cost function and averts one feature  
127 overbearing others. A filter passes through a set of units located in a small neighborhood in the  
128 previous convolutional layer. With local receptive fields, neurons can extract the elementary  
129 features of inputs that are then combined with those of higher layers. The outputs of such a set of  
130 neurons constitute a feature map (see Fig. 3). At each position, various types of units in different  
131 feature maps compute various types of features. A sequential implementation of this procedure for  
132 each feature map is used for scanning the input data with a single neuron in a local receptive field  
133 and storing the states of this neuron at corresponding locations in the feature map. The constrained  
134 units in a feature map perform the same operation on different instances in a time series, and  
135 several feature maps (with different weight vectors) can comprise one convolutional layer. Thus,  
136 multiple features can be extracted in each instance. Once a feature is detected, its exact “location”  
137 becomes less important as long as its approximate position relative to the other features is  
138 preserved (Krizhevsky et al., 2012; LeCun et al., 2015).

139 CNN uses a kernel of a given size to capture changes in the temporal variation of the input  
140 data by sweeping through time series. The various sections of the data are represented by feature  
141 maps. An additional layer performs local averaging, called “pooling,” and subsampling reduces  
142 the resolution of the feature map and the sensitivity of the output to possible shifts and distortions.  
143 This step could potentially discard important information (e.g., sudden ozone peaks) as explained  
144 in Sabour et al. (2017). Hence, this study uses the convolution layer without pooling. The feature  
145 maps are connected to a fully-connected layer, which helps us to map each feature of multiple  
146 inputs to the hourly ozone output (see Fig. 1).

147 Compared to fully-connected multilayer perceptrons (MLPs) and recurrent neural  
148 networks (RNN), which have been extensively used as regression models, CNNs are attractive for  
149 several reasons. MLPs and RNNs are not explicitly designed to model variance within an  
150 estimation that results from a complex interaction between several inputs and outputs. While MLPs  
151 of sufficient size could indeed capture invariance, they require large networks with a large training  
152 set. Compared to the CNNs proposed in this study, RNNs are challenging to implement and  
153 computationally expensive (Eslami et al., 2019a; Sayeed et al., 2020; Lops et al., 2019).

154

## 155 **2.2. Wavelet transform:**

156 Wavelet transformation decomposes a signal into a scale frequency space, allowing the  
157 determination of the relative contributions of each temporal scale present within a signal (Mallet,  
158 1989). Wavelet decompositions are powerful tools for analyzing the variation in signal properties  
159 across different resolutions of geophysical variables (Mallet, 1989; Grinsted et al., 2004; Foufoula-  
160 Georgiou and Kumar, 2014). Using a fully scalable modulated window that shifts along with the  
161 signal, the wavelet transform overcomes the inability of the Fourier transform to represent a signal  
162 in the time and frequency domain at the same time (see Fig. S2 in the supplementary document).  
163 The spectrum is calculated for every position. After repeating the process, each time with a  
164 different window size, the results constitute a collection of time-frequency representations of the  
165 signal, all with different resolutions. The data are separated into multiresolution components, each  
166 of which is studied with a resolution that matches its scale (Aiazzi et al., 2002). While high-  
167 resolution components capture fine-scale features in the signal, low-resolution components capture  
168 the coarse-scale features.

169 As wavelet analysis represents any arbitrary (nonlinear) function by a linear combination  
170 of a set of wavelets or alternative basis functions, they are highly suitable for use as both an  
171 integration kernel for analysis to extract information about the process and a basis for



172 representation or characterization of processes (Kaheil et al., 2008). Figure S3 in the  
173 supplementary document shows the hourly ozone time series of a monitoring station in downtown  
174 Seoul, South Korea, with wavelet transform for the year 2017. Here, the wavelet transform exhibits  
175 strong power levels associated with period=24 and period=168 in the middle of the year, indicating  
176 dominant daily (24 hours) and weekly variation (168 hours).

177

### 178 **2.3. Dynamic time warping:**

179 To assess the similarity between two time series, DTW expands or contracts a given time  
180 series to minimize the difference between the two of them (Berndt and Clifford, 1994). Advantage  
181 it has over Euclidean distance, a conventional distance analysis method, is that it highlights when  
182 a shift (e.g., a time lag) occurs between two time-steps in two time series (see Fig. S4 in the  
183 supplementary document). Euclidean distance takes pairs of data within the time series and  
184 compares them. DTW calculates the smallest distance between all points, matching one time-step  
185 to many counterpart steps on the linked time series (see Fig. S4). Owing to its non-linear mapping  
186 capability, it is widely used in various domains from time-series classification (Jeong et al., 2011)  
187 to bioinformatics (Giorgino, 2009), health signal processing (Tormene et al., 2009), and speech  
188 recognition (Berndt and Clifford, 1994).

189 One benefit of DTW is that it will classify two time series of the same shape as similar  
190 even if their absolute values differ or if one time series contains large variability. Figure S5  
191 compares the DTW distance between the observation time series and two prediction models for an  
192 ozone monitoring station in Texas. DTW detects the differences between CMAQ estimation and  
193 observation with the highest difference in the middle of 2014.

194

## 195 **3. Results and Discussion**

### 196 **3.1. Case 1: CNN as a real-time ozone forecasting system**

197 In this case, we used the modeling experience reported in Eslami et al. (2019a). Briefly,  
198 the system employs a deep CNN model that uses an hourly variation of seven meteorological and  
199 two air quality parameters from the day before as inputs to predict hourly ozone concentrations on  
200 the following day for 25 monitoring stations in Seoul, South Korea. Figures S7 and S8 show the  
201 accuracy of the CNN model (using the index of agreement (IOA)) and the time series comparison  
202 of average ozone concentrations between the observation and the CNN prediction, respectively.  
203 While the model maintained a proper level of prediction accuracy, it was prone to two main  
204 limitations: (i) Its performance at various times of the year varied (see Fig. S6); and (ii) nighttime  
205 predictions showed higher relative bias and lower modeling performance than daytime predictions  
206 (see Fig. S7). In general, wavelet transform can explain varying, time-dependent modeling  
207 performance; nevertheless, the significant difference between modeling performance during the  
208 daytime and the nighttime indicates an undertrained CNN model.

209

#### 210 **3.1.1. Time-dependent model performance:**

211 The performance of the CNN model is directly dependent on how well the model  
212 understands the relationship between the inputs (meteorology and ozone precursors) and output  
213 (ozone concentration). While emission sources from volatile organic compounds (VOCs) and NO<sub>x</sub>  
214 are relatively constant in time, meteorological variables govern the variation of the ozone at  
215 different times throughout the year (Choi, 2014; Pan et al., 2019). Temperature, wind speed, and  
216 relative humidity (RH) are among the most important meteorological parameters affecting ozone  
217 variation.





218 Figure 2 shows the wavelet power transform of the aforementioned meteorological  
219 variables for 2017. Since we used an hourly time series to calculate the wavelet powers, both the  
220 index and the period are in hours. The figure also locates five time periods, which indicates  
221 significant performance variations. From Fig. S6, the CNN model underperformed during weeks  
222 3-9 and 44-51, labeled the “Worst CNN results” in Fig. 2. For weeks 14-22 and 42-44, the CNN  
223 model showed the best forecasting results. Between weeks 29 and 33, the CNN model produced  
224 significant underestimations, labeled “Large under-prediction” in Fig. 2. The figure shows strong  
225 wavelet powers during a 24-hour (daily) period for all variables, the results of strong diurnal  
226 variation of these parameters, which are directly or indirectly controlled by sunlight (e.g.,  
227 temperature, relative humidity, etc.). While the wavelet powers for wind speed were generally  
228 larger than RH, the temperature showed lower, but more consistent daily modes. This finding is  
229 important since the CNN model can more accurately detect specific “patterns” in the temperature  
230 than those in the wind speed and RH. Thus, when the daily modes are stronger in temperature, the  
231 CNN model likely performs better. In contrast, when the daily modes of the meteorological  
232 variables are relatively weak, the CNN model performs poorly (see Fig. 2).

233 The large coarse modes in the wind speed and RH lead to significant over and  
234 underestimation of the CNN model. Figure S8 shows the polar frequency (influenced by the wind  
235 speed) of the CNN modeling bias in various months. As the figure shows, while southwesterly  
236 winds in August 2017 were associated with relatively large underpredictions boosted by pollution  
237 transport from the Incheon area, north-northwesterly winds with air coming from less urbanized  
238 regions were allied with notable overpredictions.

239 Figure S9 compares the CNN model predictions with observational data for the seasons  
240 with respect to levels of RH. The figure shows the largest differences in the CNN model  
241 predictions (both over and underpredictions) when the level of RH was close to the extreme (very  
242 high and very low). This finding was particularly evident for the summer months when the model  
243 showed poor performance at capturing high ozone episodes. This finding underscores the  
244 importance of coarse models from the wavelet analysis during the warm months. Directly  
245 indicating the over or underpredictions by the model through these modes, however, is  
246 challenging. For instance, Fig. S10 shows one high ozone episode in July 2017, when the daily  
247 ozone peak exceeded 90ppb on two continuous days at most stations. Here, the overprediction of  
248 the CNN model was associated with high RH while the underprediction was linked to low RH,  
249 indicating more complexity among the relationships between meteorological factors and ozone  
250 formation or depletion.

251 Another reason for the poor performance of the CNN model during the selected time  
252 period was the relatively large coarse modes (period > 24 hours). The CNN model received  
253 information about only the last day; hence, it was unable to address the bi-daily and weekly trends  
254 with the input data. For instance, for time periods with large underpredictions, coarse modes in the  
255 wind speed were even larger than the daily modes. Thus, employing a longer history would  
256 adequately explain the relationship between wind speed and ozone. In the comparison of the  
257 average wavelet powers in various periods (from daily to weekly modes) of CNN predictions and  
258 observational data, Fig. 3 shows that the powers for both time series match periods of  
259 approximately 24 hours. After 32 hours, however, the wavelet power of the CNN model shrinks  
260 to a relatively constant power while that for the observation reaches local extremums at around 3,  
261 5, and 7 days.

262 Although wavelet analysis indicates that modes coarser than 24 hours are important  
263 components of the ozone time series, their relationship to CNN model accuracy can be



264 complicated. Figure 4 compares wavelet powers for both fine and coarse modes with a correlation  
265 coefficient ( $r$ ) in 25 ozone stations in Seoul. For stations closer to the downtown area (i.e., those  
266 with station numbers under 11), the fine modes had fewer wavelet powers than those for stations  
267 in less urbanized areas, indicating that the relationship between ozone concentrations with local  
268 emissions was evident in the less urbanized areas than it was in the other areas. The coarse modes,  
269 however, varied from station-to-station with relatively higher coarse wavelet power for those in  
270 less urbanized areas. Nonetheless, no evidence points to a clear relationship between either coarse  
271 or fine wavelet modes and the accuracy of the model. Figure 4 shows that the CNN model generally  
272 performed better for stations close to downtown Seoul. Because Seoul has only one meteorological  
273 station, these stations had accessed to more realistic weather parameters in their training/prediction  
274 process.

275

### 276 3.1.2. Low modeling performance during the nighttime:

277

278 In their discussion of several air quality forecasting models that incorporated machine  
279 learning algorithms, including CNN, deep neural networks, and decision trees, Eslami et al.  
280 (2019a) and Eslami et al. (2019b) claimed that the algorithms encounter a significant modeling  
281 bias while estimating air quality concentrations during the nighttime. This bias reduced the  
282 prediction accuracy of nighttime ozone concentrations, compared to daytime concentrations, by  
283 more than 20%. A similar issue is also encountered by CTMs, even those with complex physical  
284 and chemical equations that explain the diurnal variation of ozone concentrations.

285

286 One reason for this modeling bias was likely the result of variation among the  
287 meteorological inputs during the nighttime. Although their absolute values were generally higher  
288 they were during the daytime, the relative frequency of variation was more pronounced during the  
289 nighttime, causing a discontinuity in the learning process of the CNN model. Since both daytime  
290 and nighttime hours were inputs, the CNN model minimized the cost function that contained  
291 “normalized” errors during both daytime and nighttime hours (the cost function was the mean  
292 squared errors or 24-hour ozone predictions at each step). Generally, there are more daytime hours  
293 than nighttime hours (see Fig. S11). Also, the accumulation of  $\text{NO}_2$  concentrations for these  
294 extreme cases was mainly due to stagnant atmospheric conditions with wind speeds close to their  
295 yearly minimum values (see Fig. S12a for scatter plots with levels of wind speeds). As a result,  
296 the CNN model was vulnerable to characteristic bias in nighttime ozone estimations. As a  
297 customized cost function could be a potential solution to this limitation, it requires further  
298 investigation.

299

300 The performance of the CNN model in predicting nighttime ozone concentrations also  
301 suffered because of the misinterpretation of extreme conditions of the input parameters. Figure 5  
302 shows scatter plots that compare CNN predictions and observations by the levels of two important  
303 ozone precursors ( $\text{NO}_2$  concentrations) and meteorological variables (RH%) separated into  
304 daytime and nighttime. The  $\text{NO}_2$  concentration was generally higher during the nighttime, when  
305 the ozone concentration was near zero for extreme  $\text{NO}_2$  values because of conditions amenable to  
306 ozone depletion with the absence of sunlight. Unable to capture this relationship, however, the  
307 CNN model overestimated these cases (See Fig. 5a).

308

309 In contrast to the above-mentioned overestimated events, Fig. 5b shows an underestimation  
of nighttime ozone when the level of RH% was generally high, primarily during warm days. A  
similar pattern occurred when the surface pressure was accounted for (Fig. S12b). Such  
underestimated events occurred for two reasons. One is that high (or low) levels of RH% and  
surface pressure generally occur at about the same time during the early morning (or late





310 afternoon), when the planetary boundary layer (PBL) is at its lowest (or highest) level during the  
311 day. In these extreme conditions, the earlier sunrise (or later sunset) during the summer months  
312 established a condition that elevated ozone concentrations. As these events normally occurred only  
313 during short periods of time, the CNN model was not sufficiently trained to capture these  
314 relationships.

315

### 316 **3.2. Case 2: CNN as a post-processing tool in a real-time ozone forecasting system:**

317 In this case, a generalized bias-correction CNN model introduced by Choi et al. (2019) was  
318 used. Their model is a computationally efficient deep learning-based model that produces more  
319 reliable numerical results. The authors used a deep CNN model to map ozone precursors from  
320 CMAQ and meteorological parameters from the weather research and forecasting (WRF) model  
321 (as input variables) to observe hourly ozone concentrations at a monitoring station (as a target).  
322 Their model, the CMAQ-CNN model, significantly improves the performance of the CMAQ  
323 model in both accuracy and bias. Figures S13 shows the statistical improvements (in correlation,  
324 root mean squared error, and standard deviation) of the CMAQ-CNN model over the CMAQ  
325 model (as a base model) in different months. Figure S14 compares the daily maximum ozone  
326 estimated by CMAQ and CMAQ-CNN in 48 states for which the CMAQ-CNN significantly  
327 moderated the overpredictions of the CMAQ.

328 It was clear that the likelihood of the CMAQ-CNN model producing accurate results was  
329 strongly associated with the quality of CMAQ forecasts; when CMAQ forecasted hourly ozone  
330 concentrations with a station-specific yearly IOA more than 0.5, the IOA of the CMAQ-CNN  
331 model was more than 0.8 for most cases. The probability of such accuracy was generally unrelated  
332 to that of the CMAQ model. For instance, the CMAQ-CNN model was unable to reach the yearly  
333 IOA=0.8 even though the CMAQ IOA was more than 0.7 (e.g., EPA #101 Tennessee: CMAQ  
334 IOA=0.7; CMAQ-CNN IOA=0.78). In some cases, however, the yearly IOA following the post-  
335 processing approach was less than 0.7 (e.g., EPA #1011 California: CMAQ-CNN IOA=0.63).  
336 Here, we used the distance analysis from DTW to explain (i) why CMAQ-CNN produced  
337 satisfactory results at some stations but not others, and (ii) why it performed poorly at some  
338 stations.

339

#### 340 **3.2.1. Satisfactory post-processing scenarios:**

341 Figure 6 shows the time-series of CMAQ, CMAQ-CNN, and observed daily ozone  
342 concentrations at three EPA stations. These stations were selected because the IOA accuracy of  
343 the CMAQ-CNN model was either more than 0.9 (Fig. 6a and 6b) or 20% more than that of CMAQ  
344 (Fig. 6c). Figure 7 compares the DTW distance analysis of CMAQ and CMAQ-CNN for the same  
345 stations. These are three typical cases of satisfactory improvement by the CMAQ-CNN post-  
346 processing approach:

347 Figures 6-7(a): Observed ozone concentrations in this California location were higher at the  
348 beginning of the ozone season, followed by relatively steady values ranging  
349 between 20-40ppb. After May, however, CMAQ significantly overestimated daily  
350 ozone concentrations. The overestimation was more pronounced at the end of the  
351 ozone season, resulting in an overall IOA accuracy of 0.73. The DTW distance  
352 analysis showed a consistent distance between CMAQ predictions and observed  
353 values. Because of this consistency, the CMAQ-CNN model recognized the bias  
354 trends in CMAQ, boosting its prediction accuracy by 0.17, even though the large



355 distance from the CMAQ predictions (mean distance=0.52) mirrored a relatively  
356 significant overestimation in the CMAQ-CNN post-processed results.  
357 Figures 6-7(b): Here, the trend in ozone concentrations followed a U-shaped curve in the ozone  
358 season because of strong summer winds coming from the large bodies of water  
359 near Florida (the North Atlantic Ocean and the Gulf of Mexico). For this station,  
360 CMAQ accurately predicted this trend throughout the ozone season with a  
361 relatively constant bias from July to September. As a result, the overall accuracy  
362 of the IOA was 0.84 for the CMAQ prediction. The CMAQ was also consistent  
363 with the DTW analysis, with two distance gaps in July and September (at the  
364 beginning and the end of the CMAQ overestimation period). The CMAQ-CNN  
365 model, recognizing the adequate performance of the base model in its post-  
366 processing algorithm, further improved the IOA accuracy of CMAQ by around  
367 10%.

368 Figures 6-7(c): The trend of observed ozone showed a steady decrease in this northeastern state  
369 because of the significantly cooler summer and fall months. This trend, along with  
370 the fewer ozone emission sources surrounding this station resulted in the formation  
371 of less ozone during the ozone season. The CMAQ model overestimated ozone  
372 concentrations by more than 50% during most of the season with a relatively large  
373 mean DTW distance (0.62). The CMAQ-CNN model was able to address this issue  
374 because of the consistency of the bias trend in CMAQ predictions (see left panel  
375 for DTW distance). Thus, overall, the accuracy of IOA improved by 0.2.

376 The satisfactory post-processing results using the CMAQ-CNN model were mainly  
377 characterized by the regularity of the bias trend in CMAQ as the base model for training the CNN  
378 model. As shown by the DTW distance analysis, when the DTW distance of CMAQ predictions  
379 from observed values was consistent throughout the ozone season, the CNN model was able to  
380 improve the CMAQ results to a reliable level (IOA>0.8). To test this hypothesis, we used the  
381 CMAQ-CNN post-processing approach in typical unsatisfactory scenarios.

382

### 383 3.2.2. Unsatisfactory post-processing scenarios:

384 Figure 8 compares the time series of ozone observations with the CMAQ and CMAQ-CNN  
385 models at three selected EPA stations. For all of these stations, the CMAQ-CNN model failed to  
386 reach a reliable IOA accuracy level of 0.8 while the accuracy of the CMAQ model improved.  
387 Figure 9 represents the DTW distance analysis of the two models and the ozone observation for  
388 the same stations. Unsatisfactory improvement by the CMAQ-CNN model occurred in the  
389 following three cases:

390 Figures 8-9(a): The ozone trend in this station fluctuated throughout the ozone season with  
391 frequent spikes in May, July, and October, primarily the result of biomass burning  
392 (Choi et al., 2016). While the CMAQ model predicted ozone concentrations with  
393 a relatively small bias (IOA=0.7), the bias trend varied from time to time—that is,  
394 trends of under and overpredictions changed frequently. A footprint of these  
395 trends, that is, changes in the path of the distance trend, is evident in the DTW  
396 analysis. This inconsistency was mirrored in the equivalent DTW analysis for the  
397 CMAQ-CNN model by a consistent distance trend, resulting in an unsatisfactory  
398 IOA accuracy level (IOA=0.78) with an increased mean DTW distance (0.89  
399 compared to 0.74 for the CMAQ time series).



400 Figures 8-9(b): The trend in this California location was a relatively constant concentration of  
401 ozone generally ranging between 10-30ppb. The CMAQ model significantly  
402 overpredicted ozone concentrations throughout the entire time period, mostly the  
403 result of the proximity of this station to the Pacific Ocean (San Diego County),  
404 which controls the variation in the daily ozone concentration (Pan et al., 2017).  
405 The DTW distance analysis shows a significant, yet steady spike in distance  
406 between CMAQ and the observation. Thus, even though the CMAQ-CNN  
407 significantly improved the accuracy of the CMAQ model (IOA=0.63 compared to  
408 CMAQ IOA=0.44), the large distance accounted for the underperformance of the  
409 post-processing approach. That also mirrored the consistent distance in the  
410 CMAQ-CNN distance trend (see the right panel).

411 Figures 8-9(c): In this station, the ozone concentration followed an infrequent trend with lows and  
412 highs spread indiscriminately across the ozone season, the result of several factors  
413 affecting air pollution in this region, including biomass burning, a strong frontal  
414 system, and other conditions. As a result, the CMAQ model underperformed with  
415 substantial overestimation during most of the time period (IOA=0.55). In addition,  
416 the bias of the CMAQ model did not follow as clear a trend as the DTW distance  
417 analysis. The CMAQ-CNN model improved the prediction results by more than  
418 10% with a reduced DTW distance (0.27 vs. 0.35 for the CMAQ time series).  
419 Nevertheless, the varying ozone trend accompanying the inconsistency in the  
420 prediction bias trend resulted in the low overall accuracy of the IOA of the CMAQ-  
421 CNN for this station (IOA=0.67).

422 Unlike the satisfactory cases, the unsatisfactory post-processing results using the CMAQ-  
423 CNN model stemmed from the inconsistency in the bias trend found by the DTW distance analysis.  
424 Another influential factor was the variability of observed ozone concentrations. Because of the  
425 frequent variation in the observational data, it was more complicated to train the CMAQ-CNN  
426 model so that it addressed the bias in the CMAQ model. The geographical location of a station  
427 was also an important factor in the improvement level of the post-processing approach. Proximity  
428 to the large body of water and/or sources from biomass burning during the ozone season were  
429 among the influential geographical features. Also, as Figs. 8-9 show, the DTW distances of the  
430 CMAQ-CNN predictions from the observed ones followed a consistent trend. Therefore, the  
431 information in Figs. 6-7 indicate that a secondary post-processing model might be a possible  
432 solution to boosting prediction accuracy.  
433

#### 434 4. Conclusion:

435 Various applications of deep learning algorithms, particularly convolutional neural  
436 networks, have universally been applied in the field of atmospheric sciences, especially in air  
437 quality forecasting systems. Although such applications supported easy-to-use, computationally-  
438 efficient frameworks and flexible capabilities appeared to generate accurate prediction results, the  
439 risk of exaggerated expectations may be a cause for concern. In an effort to elucidate both the  
440 advantages and limitations of deep learning models in air quality forecasting (AQF) systems, this  
441 paper addressed several common issues raised by the use of these models.

442 To explore the limitation, we chose two applications of two similar CNN models. (i) CNN  
443 as an independent real-time AQF; and (ii) CNN as a post-processing model of a state-of-the-art  
444 dynamical model, the Community Multi-scale Air Quality Model (CMAQ). For both cases, the  
445 CNN model resulted in an acceptable 24-hour in advance, hourly ozone concentration prediction



446 with an index of agreement (IOA) of more than 0.8 for two networks of monitoring stations in  
447 South Korea and the United States. We selected two powerful statistical data analytic techniques—  
448 wavelet transform and dynamic time warping (DTW)—to identify the limitations of the proposed  
449 models in both cases. By applying these techniques, researchers find discrepancies in the input  
450 data and their temporal trends and thus gain awareness of the limitations of deep learning models.

451 When the CNN model was used as a real-time AQF system in South Korea, it  
452 underperformed during both cold months and high ozone episodes. In these scenarios, we found  
453 that the fine wavelet modes (daily and hourly) were relatively weaker than they were in other  
454 conditions. Also, when the coarse modes were strong, the predictions of the CNN model were  
455 fraught with a large number of errors. We also found that the model underperformed during the  
456 nighttime hours, the results of an undertrained model and extreme values of the input parameters  
457 during the nighttime.

458 For the post-processing CNN model, the level of improvement depended on the DTW  
459 distance of the CMAQ model to the observations. When the calculated distance followed a  
460 consistent trend, the post-processing model was able to address the bias of CMAQ, independent  
461 from its accuracy level or error range. When such consistency was absent or when observed ozone  
462 varied frequently, however, the errors in the CMAQ model were mirrored in the results of the post-  
463 processing model.

464 Given this discussion of the limitations of deep learning models, we suggest that  
465 researchers configure their deep learning models based on temporal trends within the input  
466 parameters, geographical locations, and variation frequency of target pollutants. To predict  
467 ambient hourly ozone concentrations, we have restricted our discussions to a multi-output  
468 regression problem in supervised settings. While our study approach might be valid for other  
469 supervised algorithms, we leave a detailed study of other supervised methods for future work.

470

471 **Code availability.** The code for the algorithm development, evaluation, and statistical analysis is  
472 freely available for non-commercial research purposes by contacting the corresponding author.

473

474 **Supplement.** The supplementary document related to this article is available.

475

476 **Author contribution.** E.E., Y.C, Y.L., A.S., and A.K.S. contributed to the design and  
477 implementation of the research, to the analysis of the results. E.E. took the lead in writing the  
478 manuscript with inputs from Y.C, Y.L., A.S., and A.K.S.. Y.C. supervised the project. E.E. and  
479 A.S. prepared the modeling input data and optimized the python codes. All authors discussed the  
480 results and commented on the manuscript and contributed to the final version of the manuscript.

481

482 **Competing interest.** The authors declare no competing financial and/or non-financial interests  
483 in relation to the work described.

484

#### 485 **Acknowledgment**

486 This study was supported by the High Priority Area Research Seed Grant of the University of  
487 Houston. The authors also express their gratefulness to Drs. Wonbae Jeon and Shuai Pan whose  
488 prepared the 4-year CMAQ and SMOKE runs for the TCEQ Project No. 582-15-54181-09,  
489 which were used in this study.

490



## 491 **References**

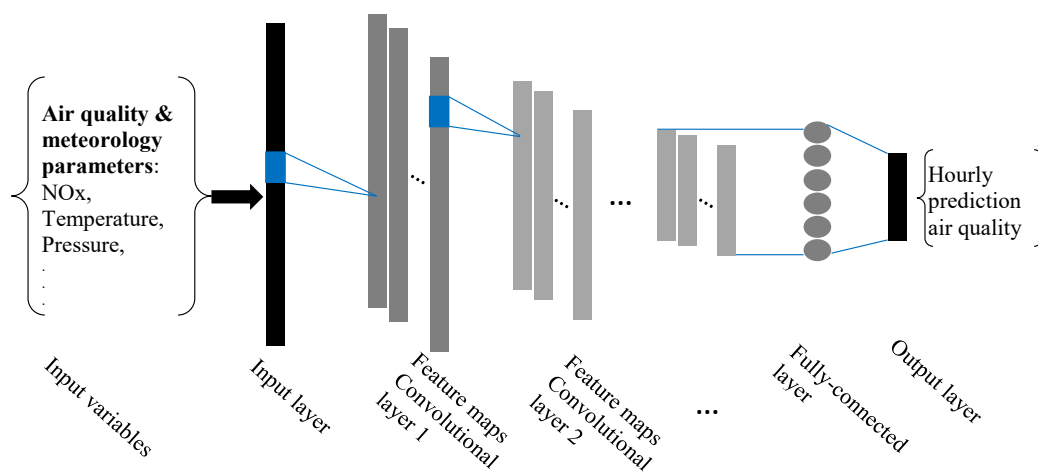
- 492 Aiazzi, B., Alparone, L., Baronti, S., and Garzelli, A.: Context-driven fusion of high spatial and  
493 spectral resolution images based on oversampled multiresolution analysis. *IEEE T Geosci.*  
494 *Remote.*, 40(10), 2300-2312, 2002.
- 495 Berndt, D. J., and Clifford, J.: Using dynamic time warping to find patterns in time series, in: *KDD*  
496 *workshop*, 10(16), 359-370, 1994.
- 497 Camalier, L., Cox, W., and Dolwick, P.: The effects of meteorology on ozone in urban areas and  
498 their use in assessing ozone trends, *Atmos. Environ.*, 41(33), 7127-7137, 2007.
- 499 Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T.: The rise of deep learning in  
500 drug discovery, *Drug Discov. Today*, 23(6), 1241-1250, 2018.
- 501 Choi, Y.: The impact of satellite-adjusted NO<sub>x</sub> emissions on simulated NO<sub>x</sub> and O<sub>3</sub> discrepancies  
502 in the urban and outflow areas of the Pacific and Lower Middle US, *Atmos. Chem. Phys.*,  
503 14, 675-690, 2014.
- 504 Choi, Y., Eslami, E., Sayeed, A., and Lops, Y.: CAMQ-AI: A computationally efficient deep  
505 learning model to improve CMAQ performance over the United States. In: *2019 AGU Fall*  
506 *Meeting*, San Francisco, CA, December 2019, 2019.
- 507 Choi, Y., W. Jeon, A. Roy, A. H. Souri, L. Diao, S. Pan, and Eslami, E.: CMAQ Modeling Archive  
508 for Exceptional Events Analysis, Texas Commission on Environmental Quality (TCEQ).  
509 [https://www.tceq.texas.gov/assets/public/implementation/air/am/contracts/reports/pm/5821](https://www.tceq.texas.gov/assets/public/implementation/air/am/contracts/reports/pm/5821554181FY1609-20160829-uh-MAQModelingArchiveForExceptionalEventsAnalyses.pdf)  
510 [554181FY1609-20160829-uh-MAQModelingArchiveForExceptionalEventsAnalyses.pdf](https://www.tceq.texas.gov/assets/public/implementation/air/am/contracts/reports/pm/5821554181FY1609-20160829-uh-MAQModelingArchiveForExceptionalEventsAnalyses.pdf),  
511 2016.
- 512 Deng, L., and Yu, D: Deep learning: methods and applications, *Found. Trends Signal Process.*, 7(3-  
513 4), 197-387, 2014
- 514 Eslami, E., Choi, Y., Lops, Y., Sayeed, A.: A real-time hourly ozone prediction system using deep  
515 convolutional neural network, *Neural Comput. Appl.*, 1-15, 2019a.
- 516 Eslami, E., Salman, A.K., Choi, Y., Sayeed, A. and Lops, Y.: A data ensemble approach for real-  
517 time air quality forecasting using extremely randomized trees and deep neural networks,  
518 *Neural Comput. Appl.*, 1-17, 2019b.
- 519 Eslami, E., Choi, Y., Lops, Y., and Sayeed, A.: A Deep Learning Driven Improved Ensemble  
520 Approach for Hurricane Forecasting. In: *2019 ESIP Annual Meeting*, Bethesda, MD, January  
521 2019, 10.6084/m9.figshare.7591775.v1, 2019c.
- 522 Foufoula-Georgiou, E., and Kumar, P. (Eds.): *Wavelets in geophysics* (Vol. 4). Elsevier, USA,  
523 2014.
- 524 Giorgino, T: Computing and visualizing dynamic time warping alignments in R: the dtw package, *J*  
525 *Stat. Softw.*, 31(7), 1-24, 2009.
- 526 Goodfellow, I., Bengio, Y., and Courville, A.: *Deep learning*, MIT press, USA, 2016
- 527 Grinsted, A., Moore, J. C., and Jevrejeva, S.: Application of the cross wavelet transform and  
528 wavelet coherence to geophysical time series, *Nonlinear Proc. Geoph.*, 11(5/6), 561-566,  
529 2004.
- 530 Higham, C. F., and Higham, D. J.: *Deep learning: An introduction for applied*  
531 *mathematicians*, *SIAM Rev.*, 61(4), 860-891, 2019
- 532 Huang, L., Kemao, Q., Pan, B., and Asundi, A. K.: Comparison of Fourier transform, windowed  
533 Fourier transform, and wavelet transform methods for phase extraction from a single fringe  
534 pattern in fringe projection profilometry, *OPT. Laser Eng.*, 48(2), 141-148, 2010.



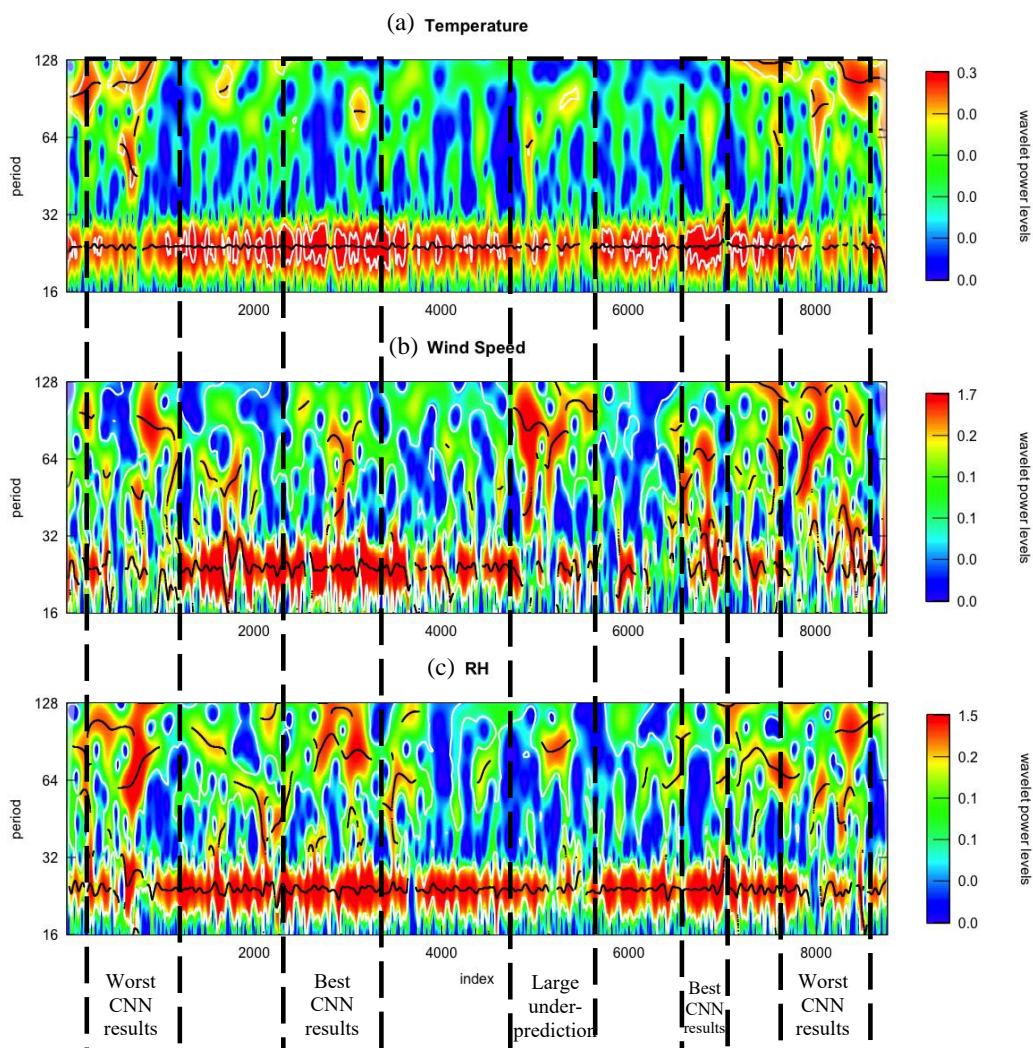


- 535 Jeong, Y. S., Jeong, M. K., and Omitaomu, O. A.: Weighted dynamic time warping for time series  
536 classification, *Pattern Recogn.*, 44(9), 2231-2240, 2011.
- 537 Kaheil, Y. H., Rosero, E., Gill, M. K., McKee, M., and Bastidas, L. A.: Downscaling and  
538 forecasting of evapotranspiration using a synthetic model of wavelets and support vector  
539 machines, *IEEE T Geosci. Remote.*, 46(9), 2692-2707, 2008.
- 540 Kamilaris, A., and Prenafeta-Boldú, F. X.: Deep learning in agriculture: A survey, *Comput.*  
541 *Electron. Agr.*, 147, 70-90, 2018.
- 542 Krizhevsky, A., Sutskever, I., and Hinton, G. E.: Imagenet classification with deep convolutional  
543 neural networks, *Adv. Neur. In.*, 25(2), 1097-1105, 2012.
- 544 LeCun, Y., Bengio, Y., Hinton, G.: Deep learning, *nature*, 521(7553), 436-444, 2015.
- 545 Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al.: A survey  
546 on deep learning in medical image analysis, *Med. Image Anal.*, 42, 60-88, 2017,
- 547 Lops, Y., Choi, Y., Eslami, E., and Sayeed, A.: Real-time 7-day forecast of pollen counts using a  
548 deep convolutional neural network, accepted in: *Neural Comput. Appl.*, 2019.
- 549 Mallat, S. G.: A theory for multiresolution signal decomposition: the wavelet representation, *IEEE*  
550 *T. Pattern Anal.*, (7), 674-693, 1989.
- 551 Pan, S., Choi, Y., Roy, A., Li, X., Jeon, W., Souri, A. H.: Modeling the uncertainty of several VOC  
552 and its impact on simulated VOC and ozone in Houston, Texas, *Atmos. Environ.*, 120, 404-  
553 416, 2015.
- 554 Pan, S., Choi, Y., Jeon, W., Roy, A., Westenbarger, D. A., and Kim, H. C.: Impact of high-  
555 resolution sea surface temperature, emission spikes and wind on simulated surface ozone in  
556 Houston, Texas during a high ozone episode, *Atmos. Environ.*, 152, 362-376, 2017.
- 557 Pan, S., Roy, A., Choi, Y., Eslami, E., Thomas, S., Jiang, X., and Gao, H. O.: Potential impacts of  
558 electric vehicles on air quality and health endpoints in the Greater Houston Area in  
559 2040, *Atmos. Environ.*, 207, 38-51, 2019.
- 560 Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules, *Adv. Neur.*  
561 *In.*, 3856-3866, 2017.
- 562 Sayeed, A., Choi, Y., Eslami, E., Lops, Y., Roy, A., Jung, J.: Using a Deep Convolutional Neural  
563 Network to Predict 2017 Ozone Concentrations, 24 Hours in Advance, *Neural Networks*,  
564 121, 396-408, 2020.
- 565 Schmidhuber, J.: Deep learning in neural networks: An overview, *Neural Networks*, 61, 85-117,  
566 2015
- 567 Tormene, P., Giorgino, T., Quaglino, S., and Stefanelli, M.: Matching incomplete time series with  
568 dynamic time warping: an algorithm and an application to post-stroke rehabilitation, *Artif.*  
569 *Intell. Med.*, 45(1), 11-34, 2009.
- 570 Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., and Baklanov, A.: Real-time air quality  
571 forecasting, part I: History, techniques, and current status, *Atmos. Environ.*, 60, 632-655, 2012.

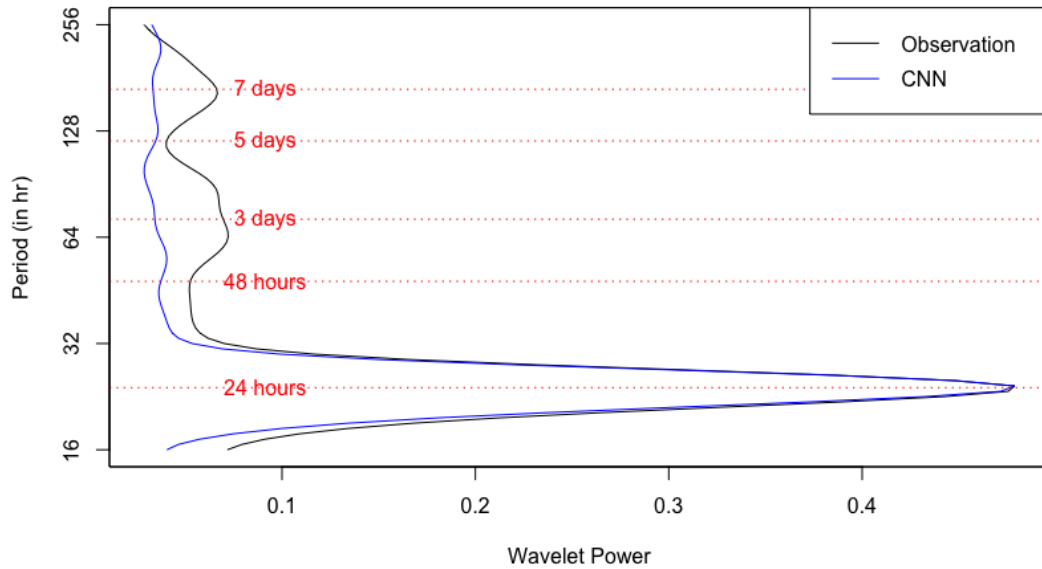




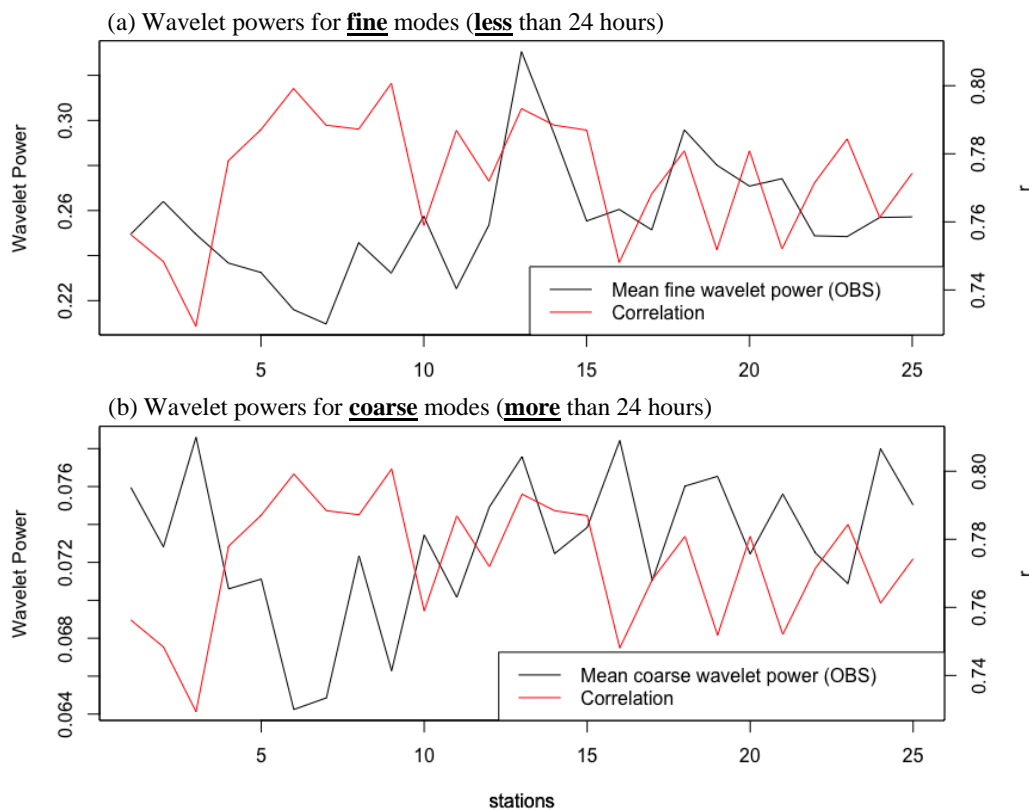
**Figure 1.** Schematic of the deep CNN model in our approach.



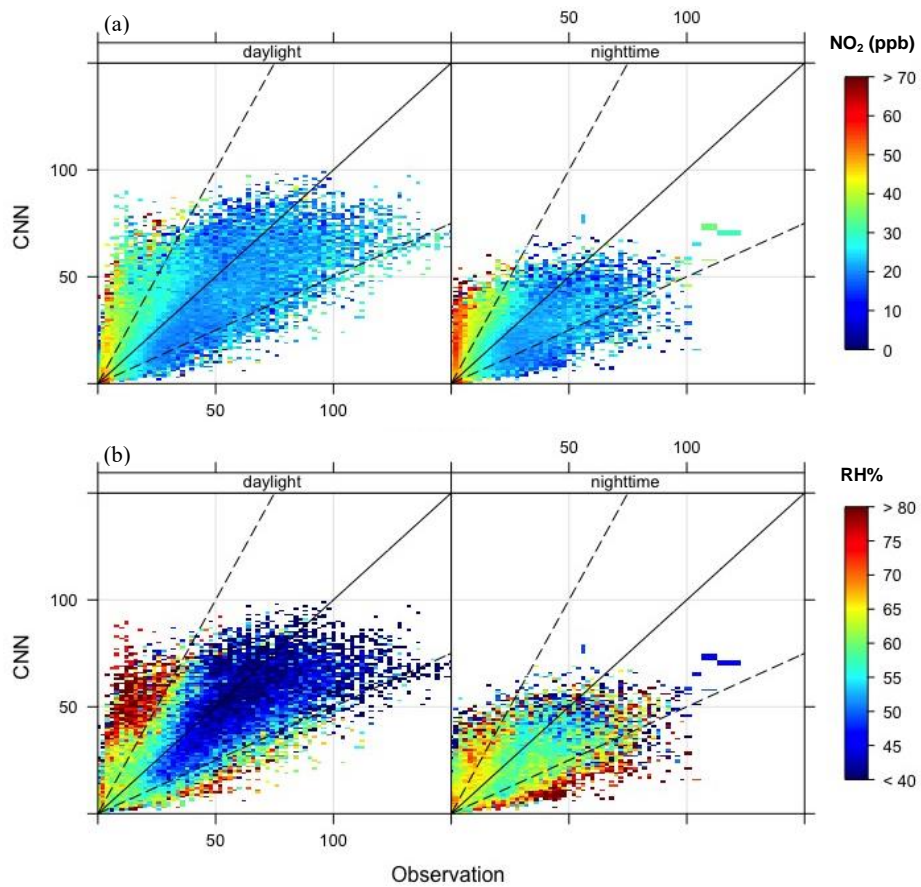
**Figure 2.** Wavelet power transform of (a) temperature, (b) wind speed, and (c) RH% for 2017 in Seoul, South Korea.



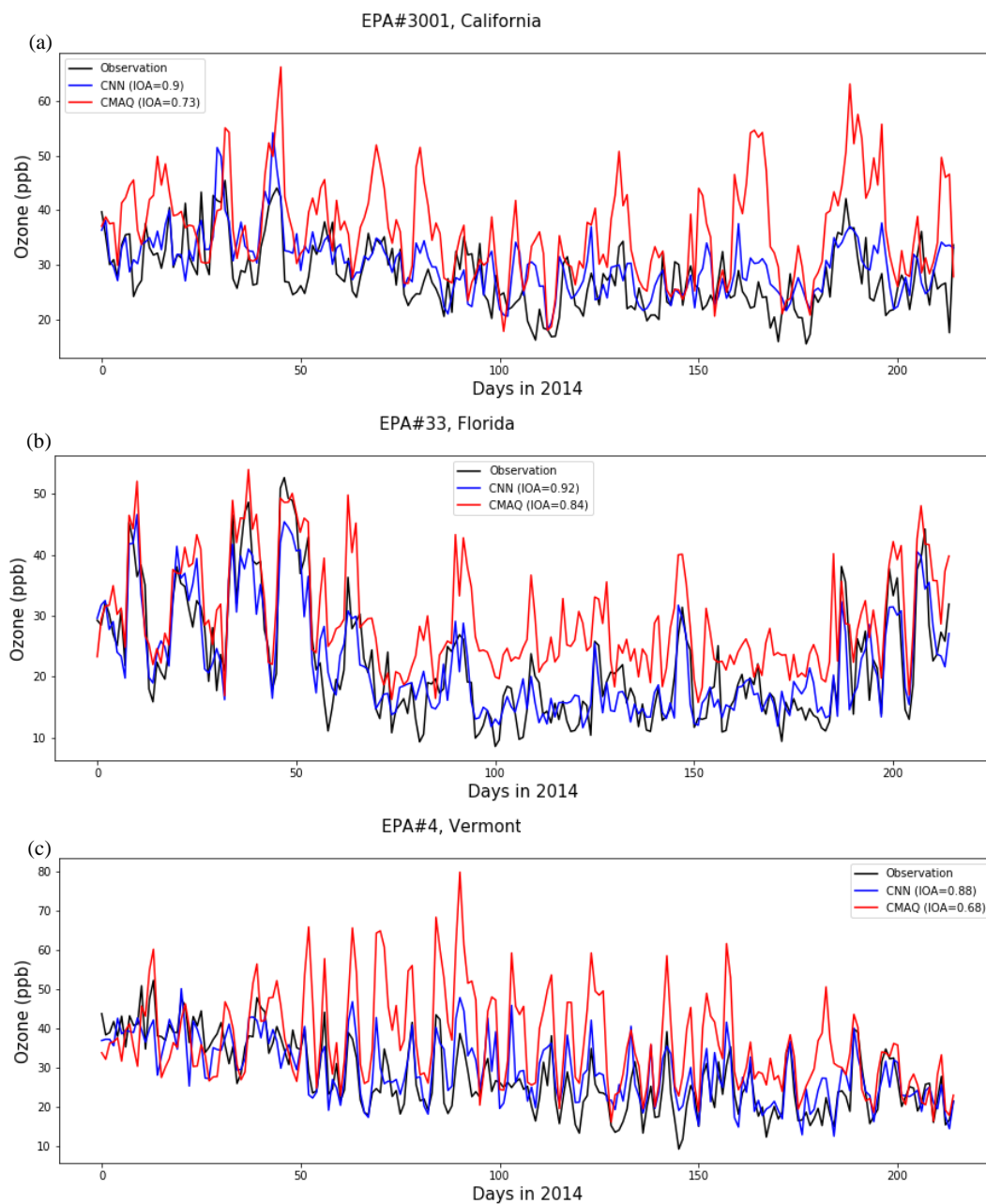
**Figure 3.** Wavelet power for various time periods (modes) for CNN predictions and observations.



**Figure 4.** Relationship between (a) fine and (b) coarse wavelet power modes and correlation coefficients in all stations in Seoul, South Korea.

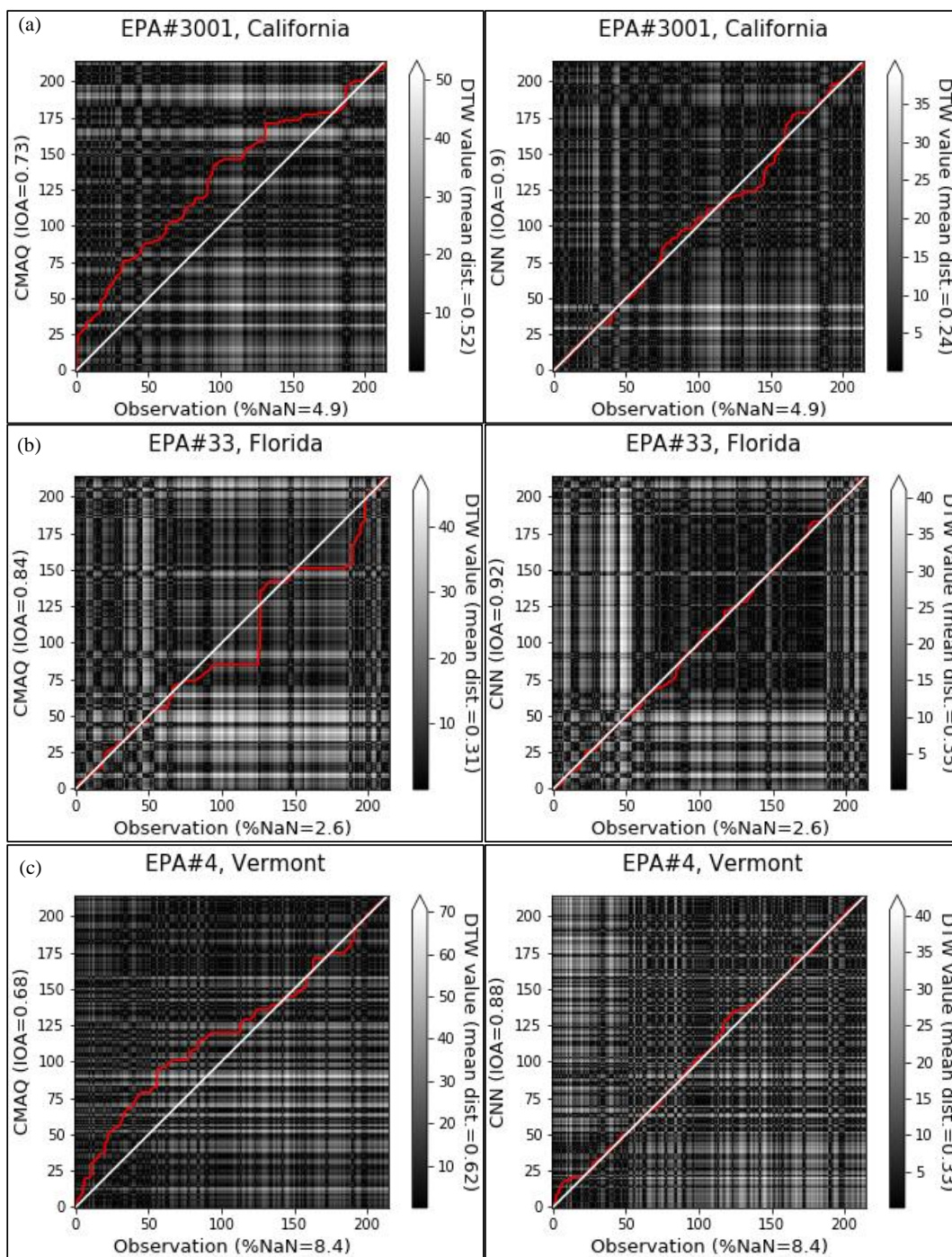


**Figure 5.** Scatter plots comparing CNN predictions and observations with respect to levels of (a) NO<sub>2</sub> concentrations and (b) RH%.

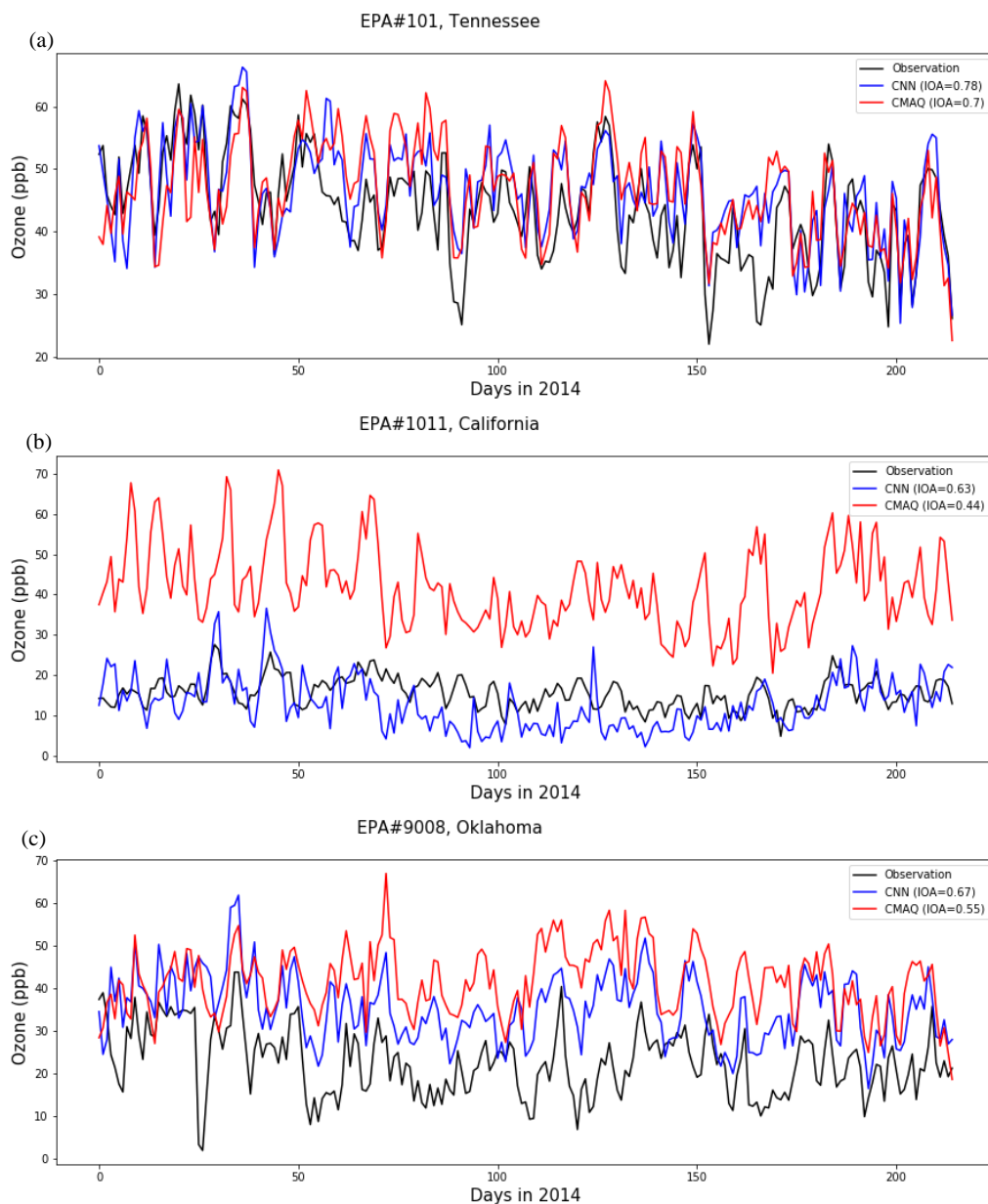


**Figure 6.** Comparison of the time series of CMAQ and CMAQ-CNN predictions for EPA stations (a) #3001 (California), (b) #33 (Florida), and (c) #4 (Vermont).

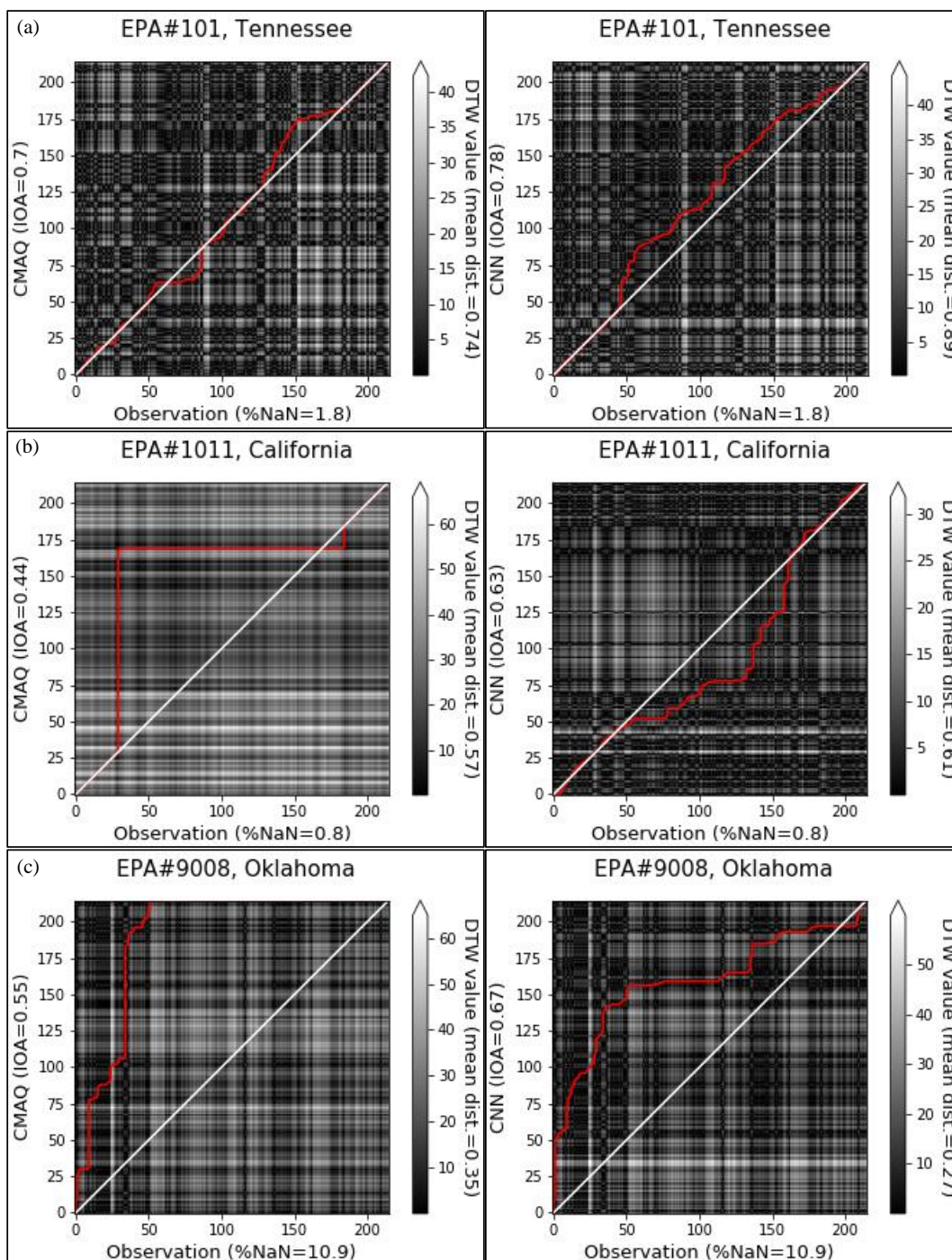




**Figure 7.** Comparison of the distance analysis of CMAQ and CMAQ-CNN predictions for EPA stations (a) #3001 (California), (b) #33 (Florida), and (c) #4 (Vermont).



**Figure 8.** Comparison of the distance analysis of CMAQ and CMAQ-CNN predictions for EPA stations (a) #101 (Tennessee), (b) #1011 (California), and (c) #9008 (Oklahoma).



**Figure 9.** Comparison of the distance analysis of CMAQ and CMAQ-CNN predictions for EPA stations (a) #101 (Tennessee), (b) #1011 (California), and (c) #9008 (Oklahoma).