Electronic Supplementary Material (ESI) for Environmental Science: Atmospheres.

Supplemental Information for:

Performance of Machine Learning for Ozone Modeling in Southern California during the COVID-19 Shutdown

Khanh Do[1,2], Arash Kashfi Yeganeh[1,2], Ziqi Gao[3], Cesunica E. Ivey[1,2,4*]

*Corresponding Author: iveyc@berkeley.edu

[1]Department of Chemical and Environmental Engineering, University of California Riverside, Riverside, CA, USA

[2]Center for Environmental Research and Technology, Riverside, CA, USA

[3]Department of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, USA

[4]Now at Department of Civil and Environmental Engineering, University of California, Berkeley, Berkeley, CA, USA

Evaluation Metrics

Correlation coefficient

$$CC = \frac{\sum_{i=1}^{N}(M_i - \bar{M})(O_i - \bar{O})}{[\sum_{i=1}^{N}(M_i - \bar{M})^2 \sum_{i=1}^{N}(O_i - \bar{O})^2]^{\frac{1}{2}}} \#(1)$$

Mean bias:

$$MB = \frac{1}{N}\sum_{i=1}^{N}(M_i - O_i) \#(2)$$

Mean absolute error:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|M_i - O_i| \#(3)$$

Root mean square error:

$$RMSE = [\frac{1}{N}\sum_{i=1}^{N}(M_i - O_i)^2]^{\frac{1}{2}} \#(4)$$

Relative root mean square error:

$$rRMSE = \frac{[\frac{1}{N}\sum_{i=1}^{N}(M_i - O_i)^2]^{\frac{1}{2}}}{\frac{1}{N}\sum_{i=1}^{N}O_i} \#(5)$$

Mean normalized bias:

$$MNB = \frac{1}{N}\sum_{i=1}^{N}\frac{M_i - O_i}{O_i} \#(6)$$

Mean normalized absolute error:

$$MNAE = \frac{1}{N}\sum_{i=1}^{N}\frac{|M_i - O_i|}{O_i} \#(7)$$

Normalized mean bias:

$$NMB = \frac{\sum_{i=1}^{N}(M_i - O_i)}{\sum_{i=1}^{N} O_i}\#(8)$$

Normalized mean absolute error:

$$NMAE = \frac{\sum_{i=1}^{N}|M_i - O_i|}{\sum_{i=1}^{N} O_i}\#(9)$$

Fractional bias:

$$FB = \frac{1}{N}\sum_{i=1}^{N}\frac{M_i - O_i}{(M_i + O_i)/2}\#(10)$$

Fractional absolute error:

$$FAE = \frac{1}{N}\sum_{i=1}^{N}\frac{|M_i - O_i|}{(M_i + O_i)/2}\#(11)$$

Model mean:

$$\bar{M} = \frac{1}{N}\sum_{i=1}^{N} M_i\#(12)$$

Observational mean:

$$\bar{O} = \frac{1}{N}\sum_{i=1}^{N} O_i\#(13)$$

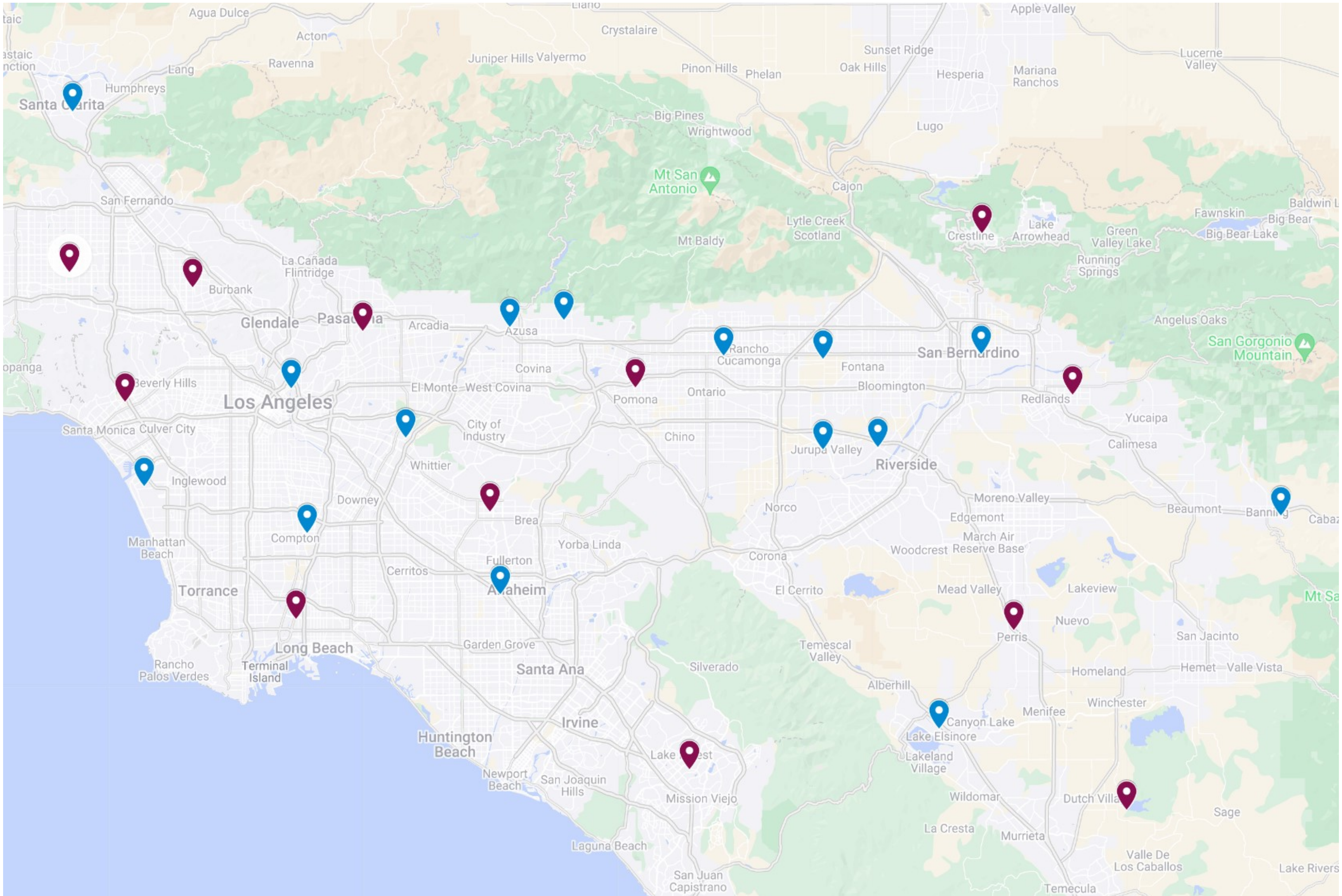


Figure S1. The map shows 27 air monitoring sites in the SoCAB. Blue labels were used for interpolation points, and red labels were used for interpolation performance evaluation.

Table S1. The R2 for CMAQ was computed for the entire year of 2020.

Sites	2020 CMAQ R^2
Anaheim	0.24
Azusa	0.39
Banning	0.17
Compton	0.38
Fontana	0.36
Glendora	0.37
Lake Elsinore	0.43
LA North Main ST	0.26
LAX	0.17
Mira Loma	0.48
Rubidoux	0.52
San Bernardino	0.50
San Gabriel	0.48
Santa Clarita	0.43
Upland	0.36
Crestline	0.22
La Habra	0.27
Long Beach	0.15
Mission Viejo	0.19
North Hollywood	0.36
Pasadena	0.32
Perris	0.47
Pomona	0.55
Redlands	0.43
Reseda	0.36
West LA	0.18
Winchester	0.30

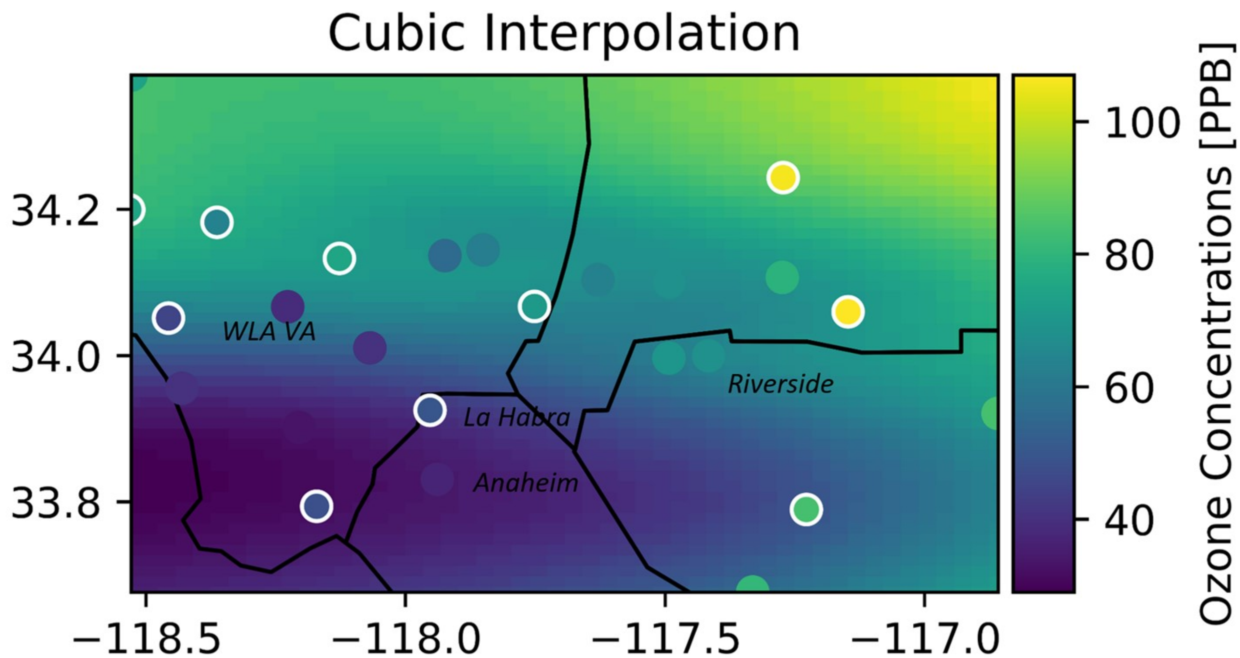


Figure S2. Hourly ozone heatmap (@16pm June 22, 2020) using cubic interpolation.

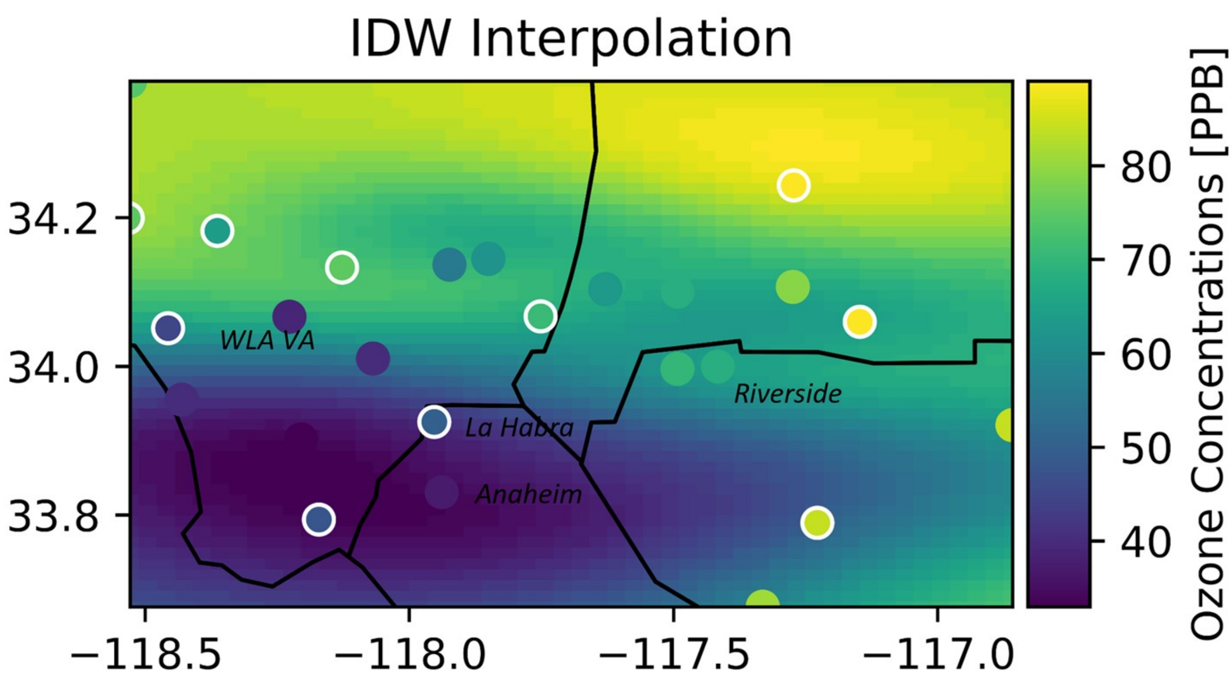


Figure S3. Hourly ozone heatmap (@16pm June 22, 2020) using IDW interpolation.

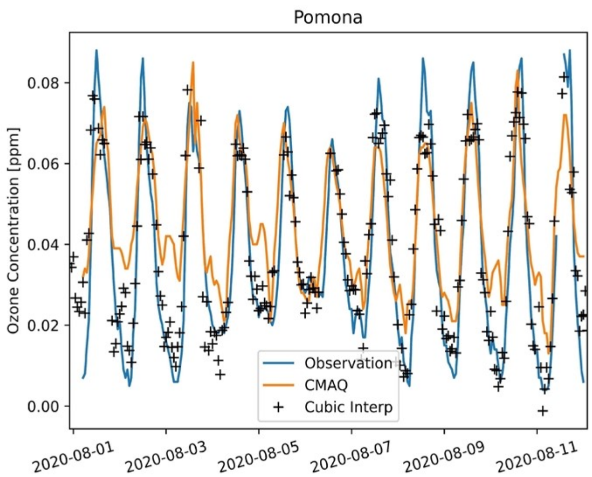


Figure S4. Time series plotting ozone concentrations for CMAQ model, cubic interpolation, and observation

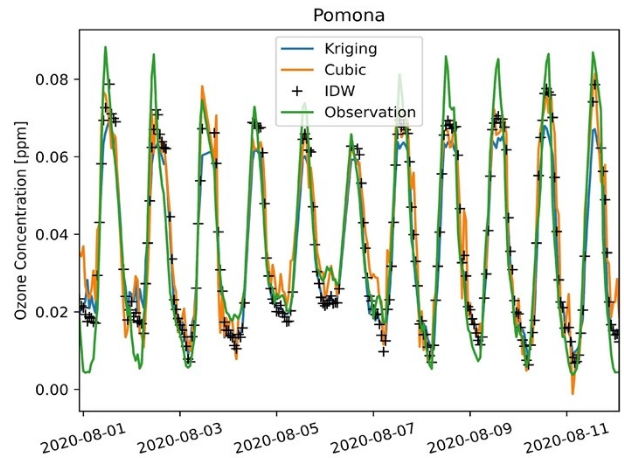


Figure S5. Time series plotting ozone concentrations for three different interpolation methods (kriging, cubic, and IDW) with observation

Table S2. Evaluation was computed using the average of 12 evaluation sites.

Month	CC	MB	MAGE	RMSE	MNB	MNAE	NMB	NMAE	FB	FAE	MM	OM
1	0.40	0.01	0.01	0.02	7.26	7.42	0.58	0.88	0.71	0.93	0.03	0.02
2	0.26	0.00	0.01	0.02	4.56	4.77	0.16	0.44	0.33	0.59	0.03	0.03
3	0.75	0.01	0.01	0.02	6.71	6.79	0.78	0.97	0.94	1.04	0.03	0.01
4	0.62	0.00	0.01	0.01	1.73	1.98	0.00	0.30	0.10	0.40	0.03	0.03
5	0.85	0.00	0.01	0.01	0.44	0.61	0.05	0.25	0.13	0.34	0.03	0.03
6	0.83	0.00	0.01	0.01	0.47	0.73	-0.03	0.23	-0.01	0.31	0.04	0.04
7	0.80	0.01	0.01	0.02	1.08	1.24	0.13	0.34	0.28	0.48	0.04	0.04
8	0.81	0.00	0.01	0.02	1.96	2.21	0.14	0.36	0.20	0.50	0.04	0.03
9	0.62	0.01	0.02	0.02	5.88	6.16	0.19	0.59	0.40	0.76	0.04	0.03
10	0.35	0.01	0.02	0.02	5.94	6.19	0.35	0.71	0.45	0.76	0.03	0.02
11	0.49	0.01	0.01	0.02	5.61	5.76	0.43	0.71	0.63	0.82	0.03	0.02
12	0.39	0.00	0.01	0.01	3.15	3.37	0.22	0.55	0.40	0.67	0.03	0.02

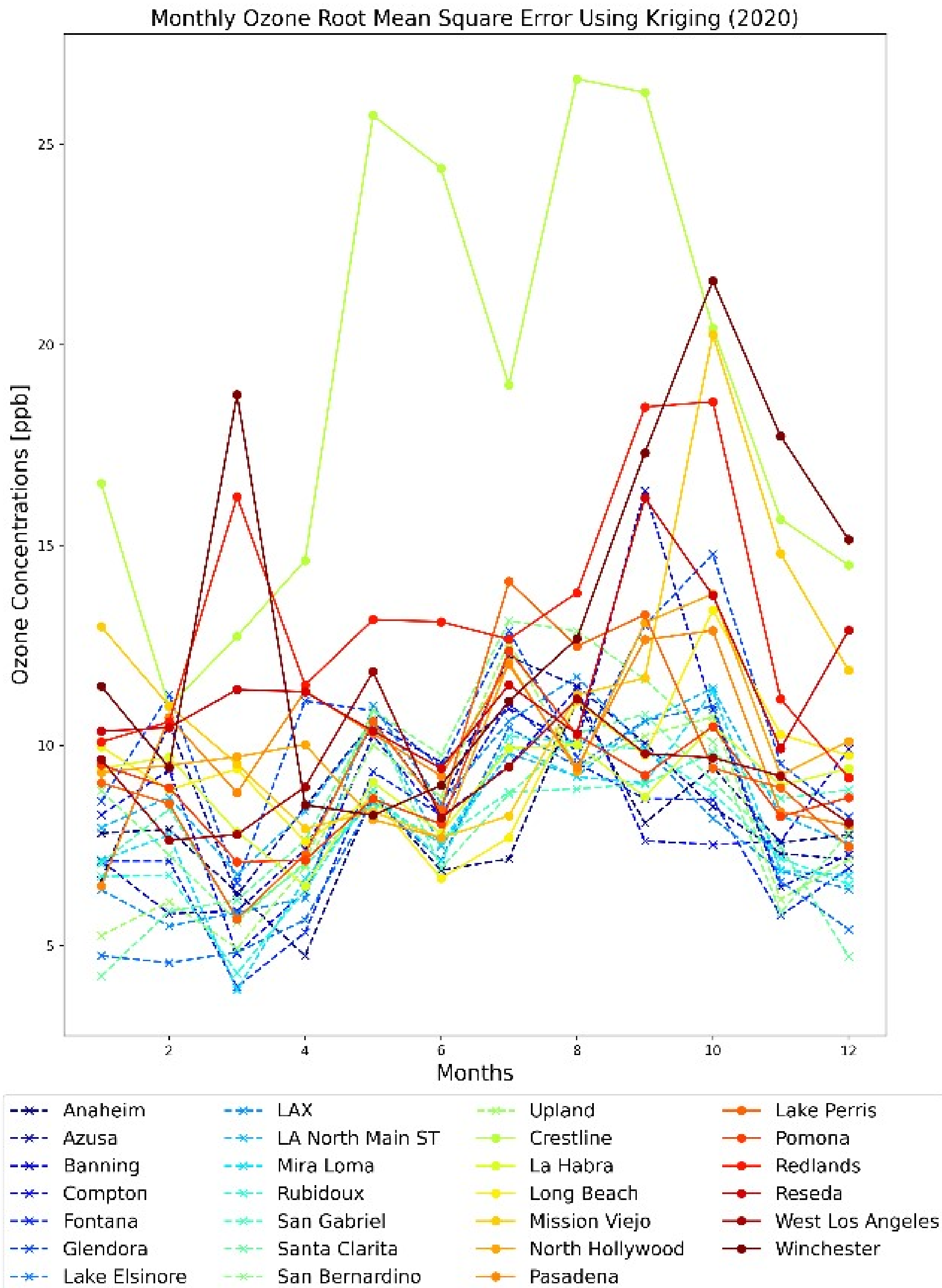


Figure S6. Monthly ozone mean bias from spatial interpolation for 2020 using the Kriging method. The dash lines with x-markers are fifteen building sites, and solid lines with filled dots are the evaluation sites.

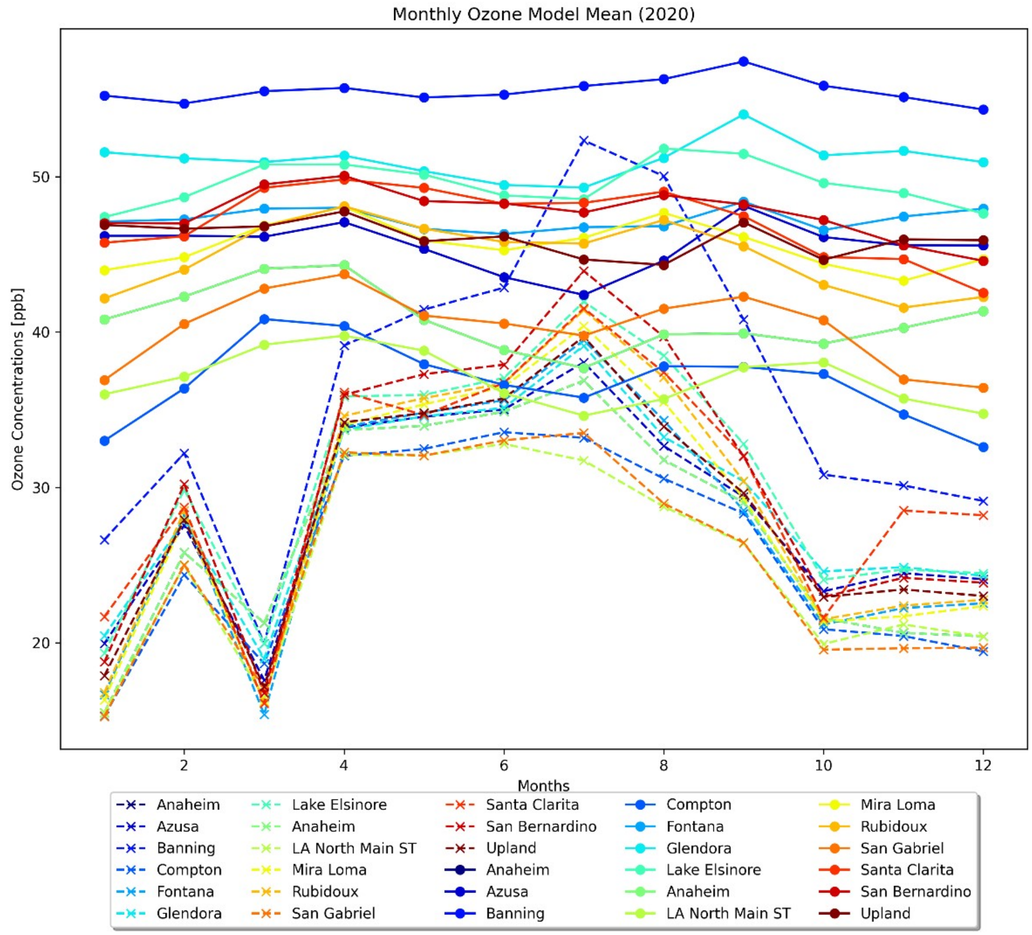


Figure S7. CMAQ (solid lines) vs. ML building sites (dash lines) model mean.

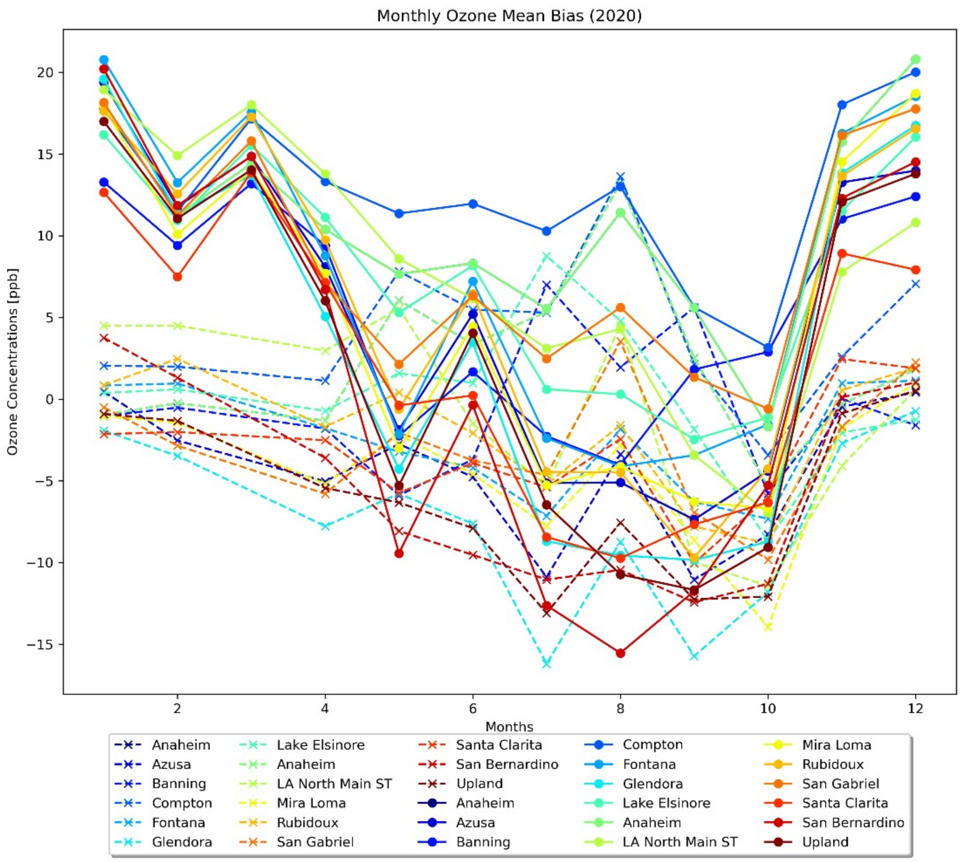


Figure S8. Monthly mean bias computed for 2020 from 9AM to 4PM. Monthly mean bias for 15 sites using kriging interpolation method (dash lines), and CMAQ simulation (solid lines). The colors of the lines corresponded to the evaluation locations.

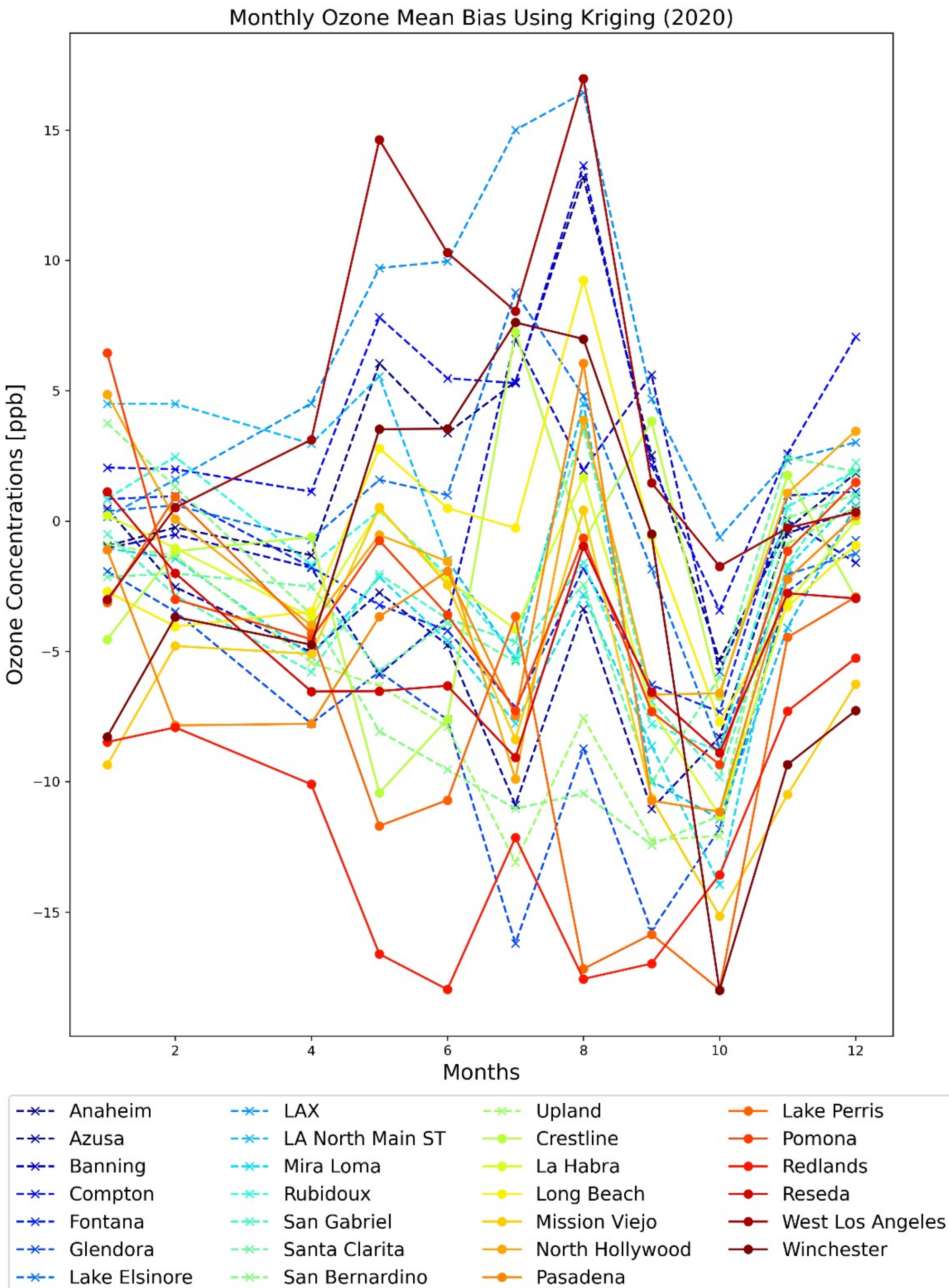


Figure S9. Monthly ozone mean bias from spatial interpolation for 2020 using the Kriging method calculated from 9AM to 4PM. The dash lines with x-markers are fifteen building sites, and solid lines with filled dots are the evaluation sites.

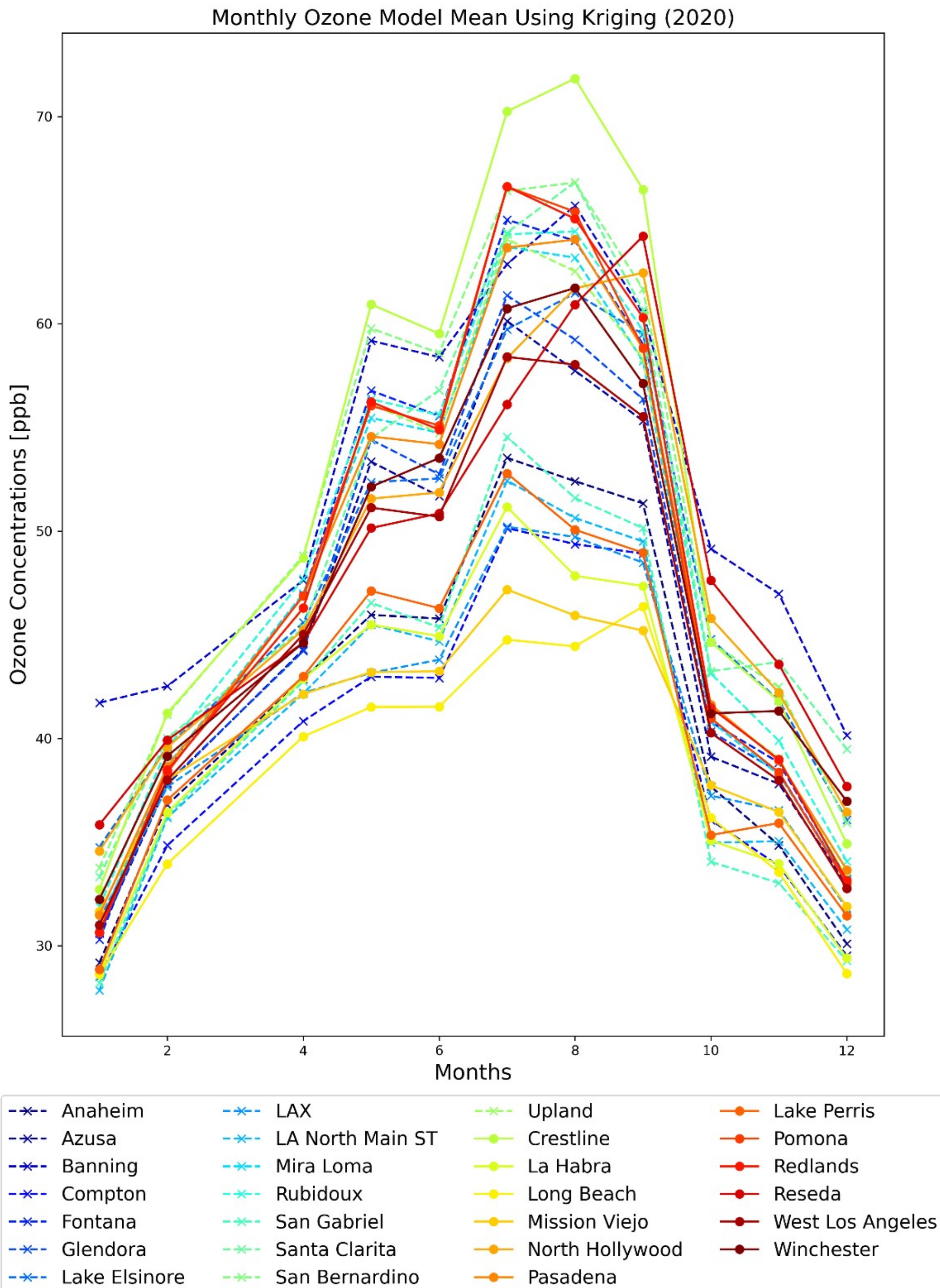


Figure S10. Building sites vs interpolated sites from 9AM to 4PM

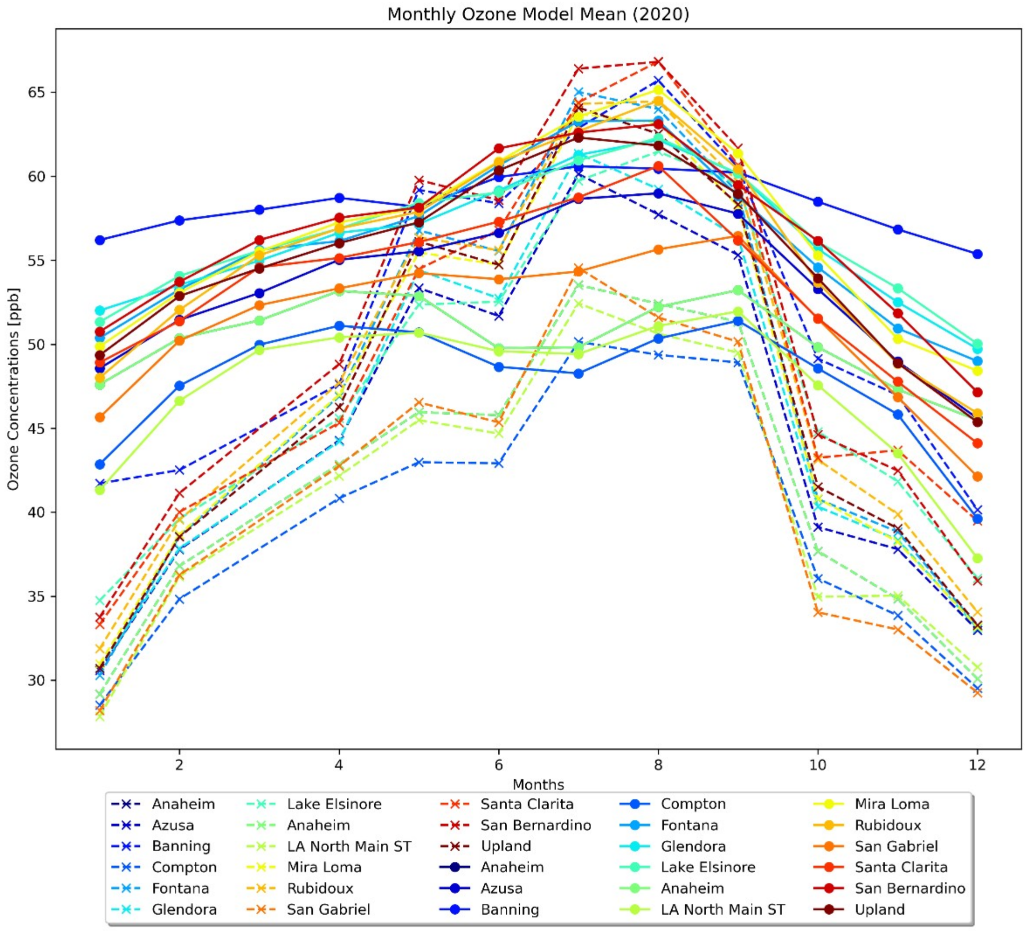


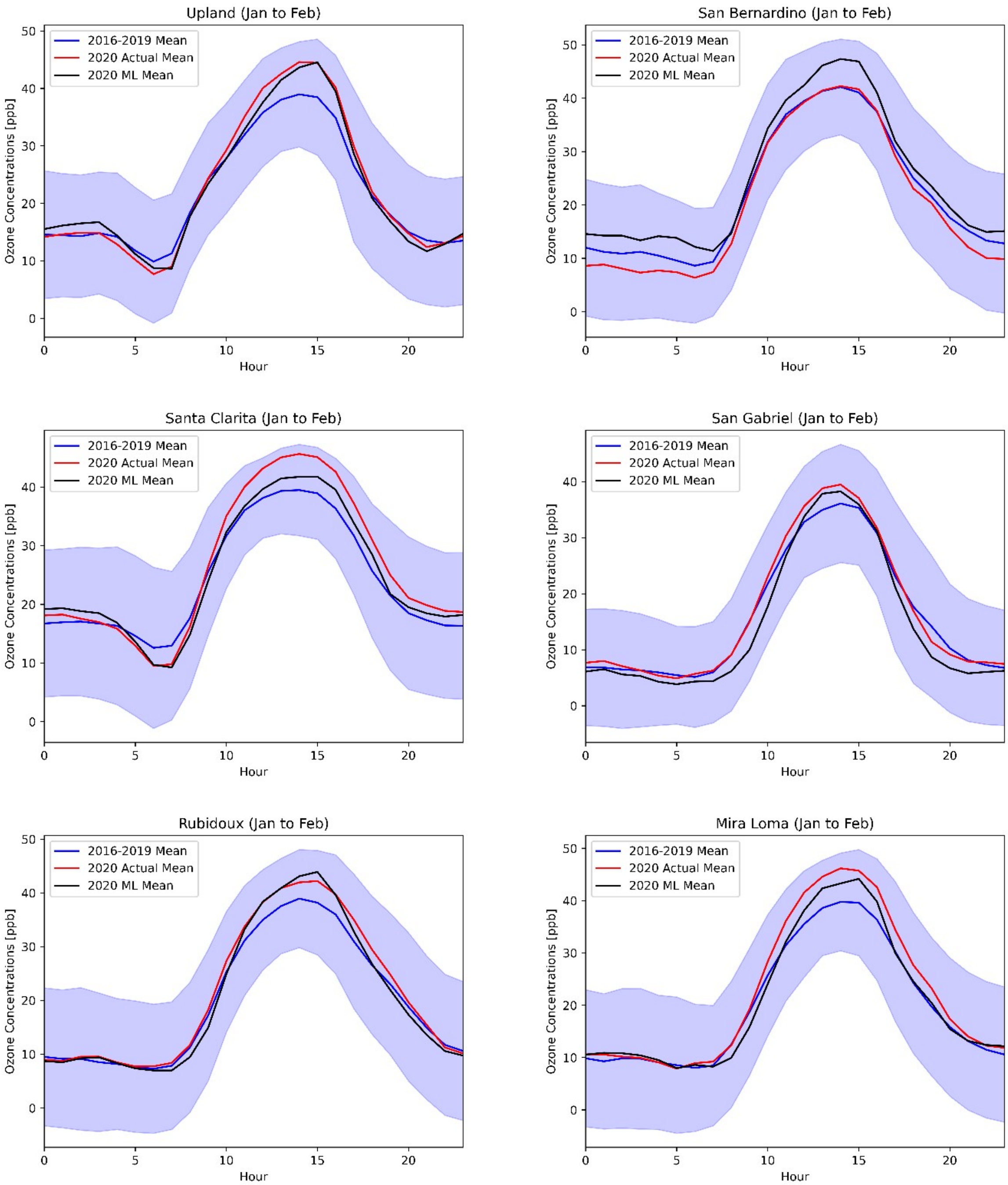
Figure S11. CMAQ (solid lines) vs. ML building sites (dash lines) from 9AM to 4Pm.

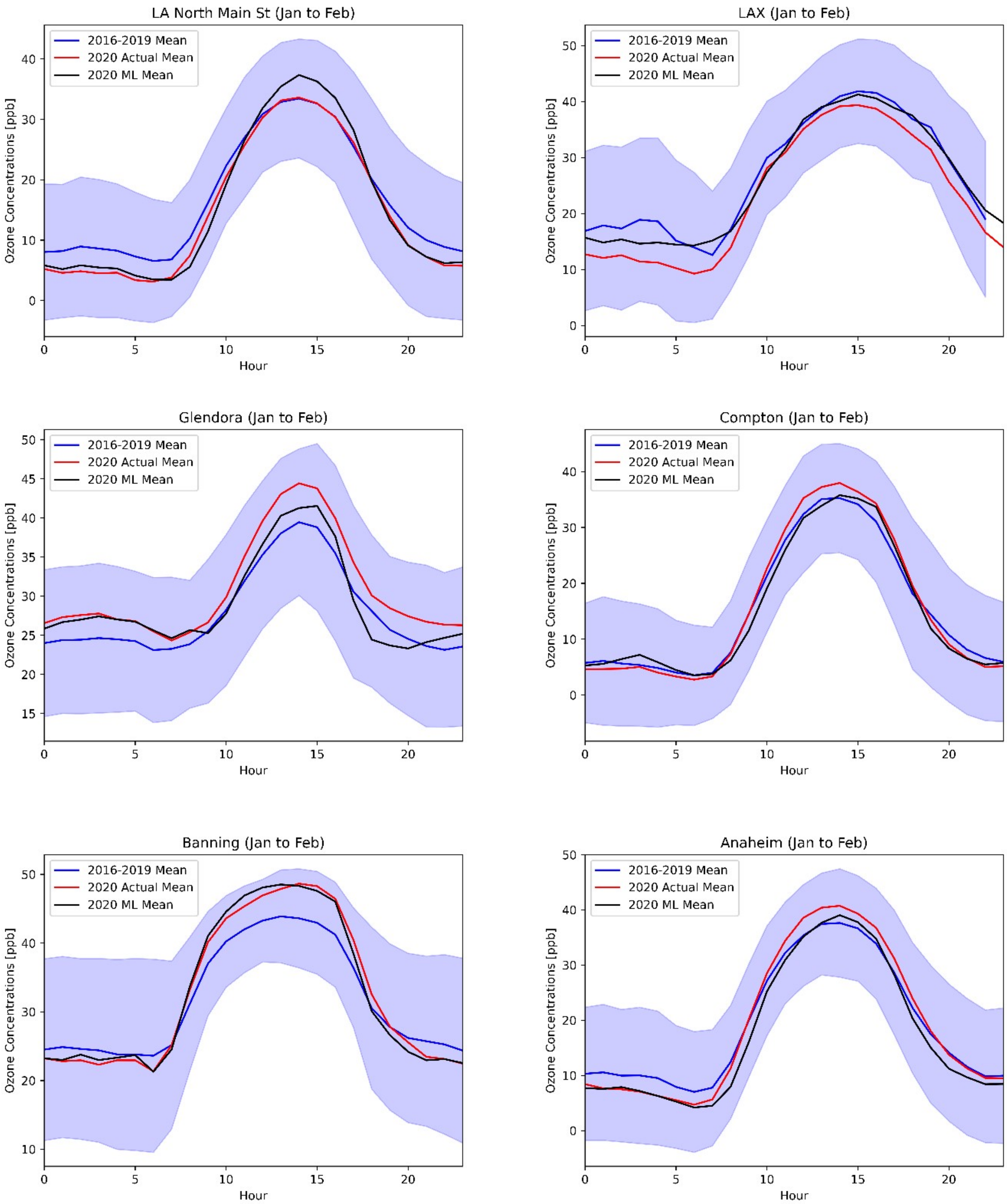
Table S3. The monthly correlation coefficient for fifteen building sites using the Kriging method.

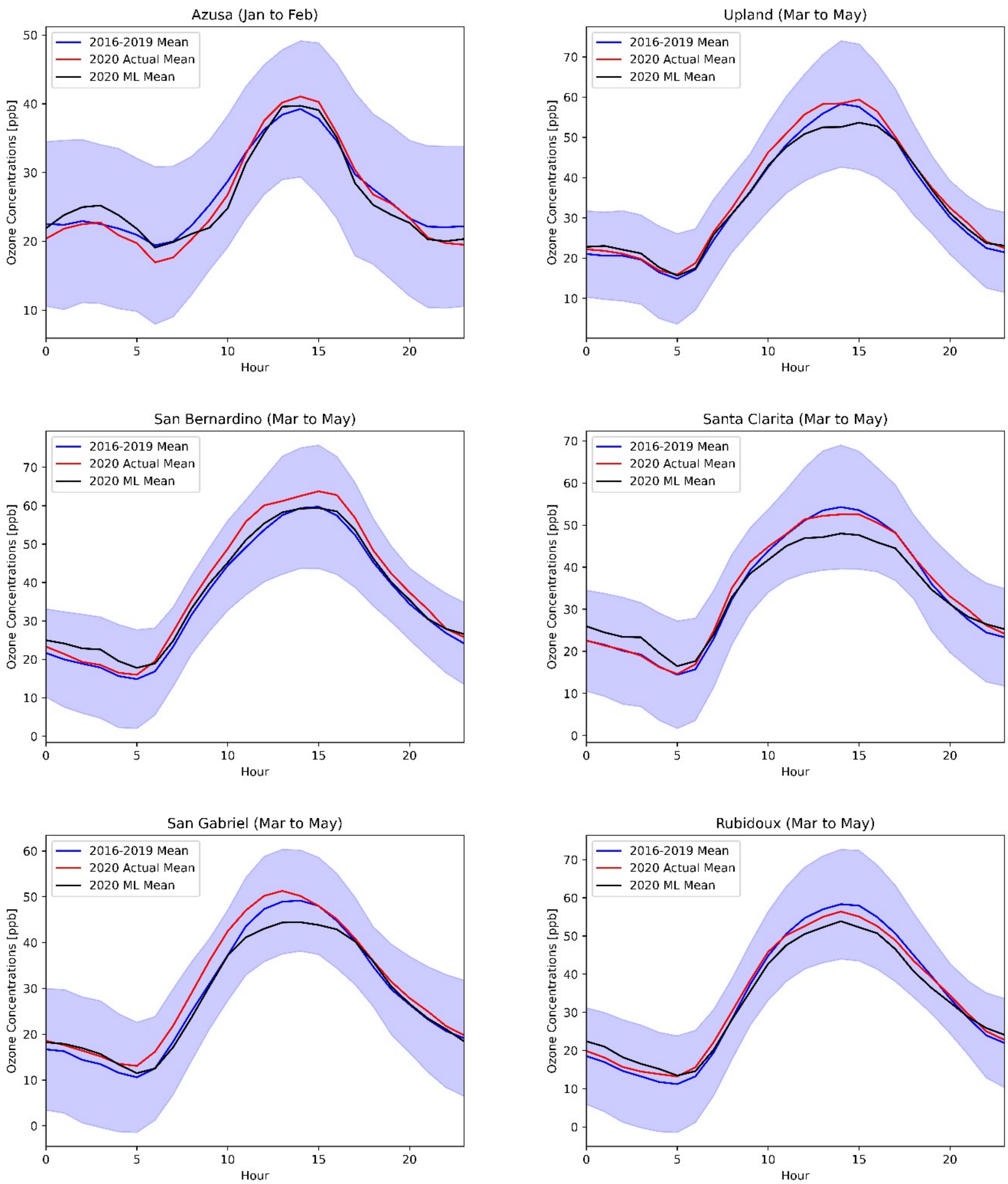
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Anaheim	0.85	0.91	0.90	0.91	0.71	0.81	0.89	0.86	0.88	0.87	0.89	0.86
Azusa	0.80	0.66	0.96	0.89	0.85	0.91	0.93	0.91	0.95	0.90	0.86	0.75
Banning	0.89	0.87	0.94	0.81	0.86	0.88	0.80	0.84	0.66	0.86	0.92	0.87
Compton	0.87	0.91	0.92	0.90	0.77	0.80	0.91	0.86	0.88	0.89	0.90	0.88
Fontana	0.86	0.88	0.98	0.95	0.93	0.94	0.92	0.95	0.96	0.91	0.94	0.88
Glendora	0.75	0.63	0.95	0.81	0.85	0.89	0.87	0.91	0.92	0.76	0.81	0.70
Elsinore	0.93	0.94	0.96	0.89	0.83	0.90	0.91	0.91	0.87	0.89	0.93	0.92
LA	0.87	0.90	0.93	0.89	0.84	0.87	0.96	0.92	0.92	0.87	0.90	0.89
LAX	0.90	0.92	0.89	0.86	0.58	0.70	0.86	0.79	0.75	0.83	0.89	0.87
Mira Loma	0.90	0.90	0.98	0.95	0.91	0.93	0.92	0.93	0.95	0.91	0.93	0.90
Rubidoux	0.90	0.92	0.98	0.94	0.91	0.93	0.93	0.94	0.95	0.94	0.92	0.90
San Bernardino	0.90	0.92	0.96	0.93	0.91	0.94	0.89	0.92	0.94	0.93	0.95	0.87
San Gabriel	0.86	0.92	0.94	0.92	0.85	0.86	0.93	0.91	0.92	0.87	0.88	0.87
Santa Clarita	0.95	0.91	0.95	0.88	0.89	0.95	0.94	0.91	0.92	0.89	0.89	0.95
Upland	0.93	0.88	0.97	0.93	0.91	0.93	0.90	0.94	0.96	0.91	0.93	0.87

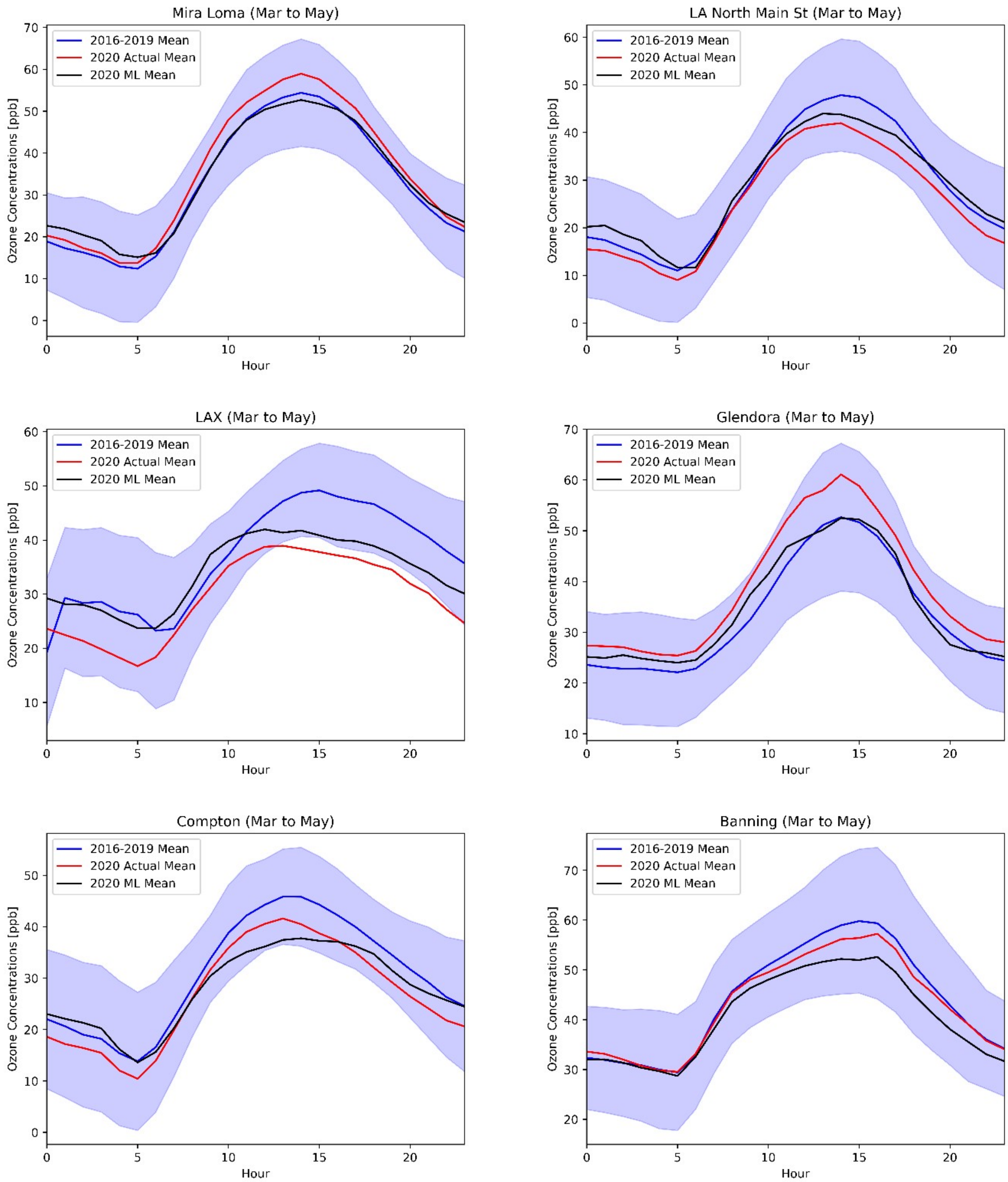
Table S4. The monthly correlation coefficient for twelve evaluation sites using the Kriging method.

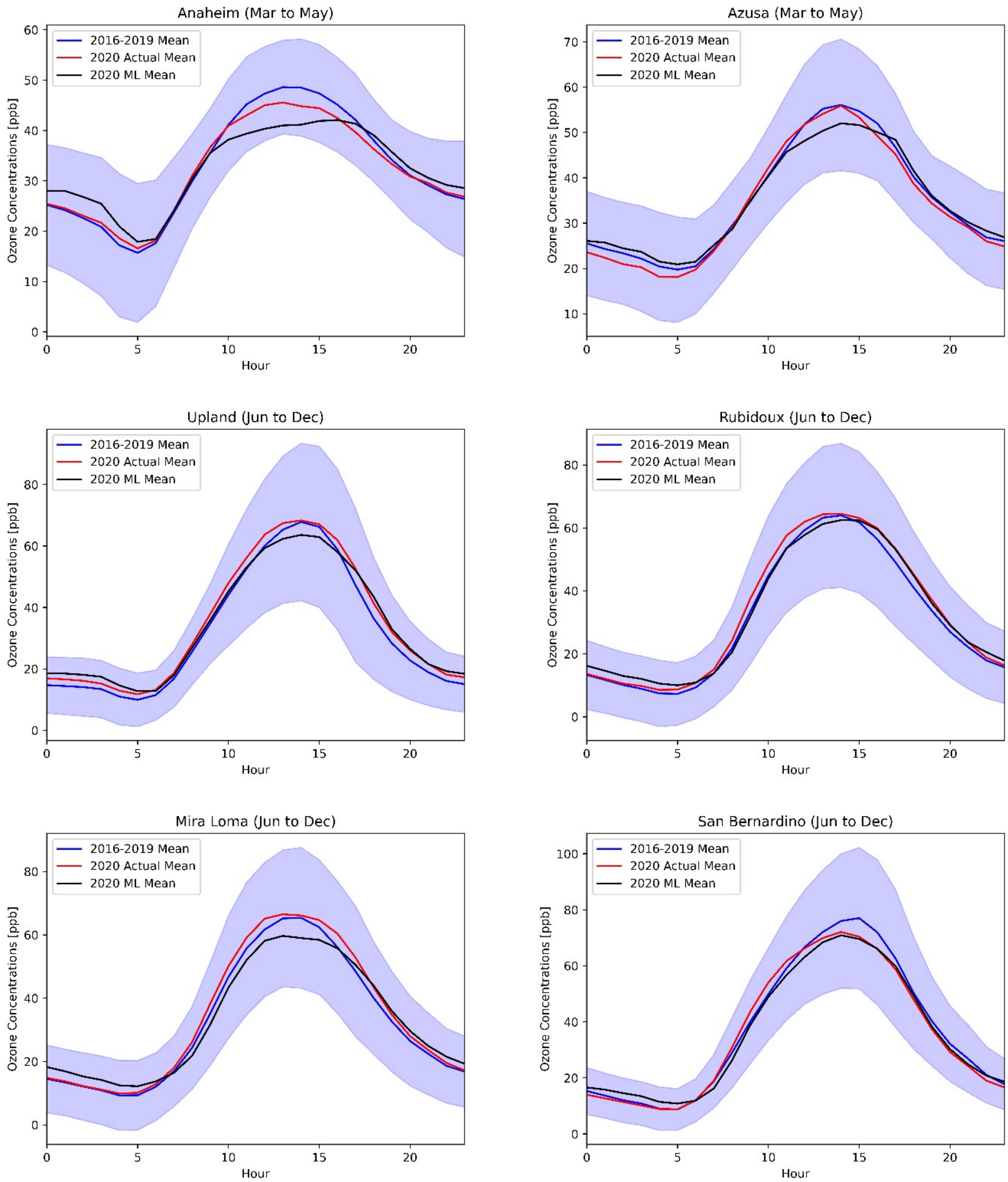
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Crestline	0.38	0.67	0.72	0.52	0.60	0.63	0.62	0.62	0.57	0.57	0.53	0.51
La Habra	0.85	0.90	0.86	0.90	0.82	0.83	0.91	0.89	0.91	0.85	0.88	0.81
Long Beach	0.75	0.84	0.74	0.74	0.57	0.71	0.79	0.75	0.80	0.75	0.79	0.73
Mission Viejo	0.79	0.74	0.84	0.79	0.71	0.76	0.89	0.84	0.82	0.63	0.76	0.69
North Hollywood	0.84	0.88	0.89	0.81	0.88	0.89	0.94	0.90	0.94	0.86	0.91	0.83
Pasadena	0.86	0.90	0.90	0.85	0.86	0.86	0.91	0.91	0.93	0.85	0.88	0.84
Lake Perris	0.86	0.88	0.96	0.91	0.86	0.92	0.84	0.90	0.88	0.91	0.87	0.89
Pomona	0.83	0.88	0.93	0.94	0.90	0.92	0.90	0.91	0.95	0.87	0.90	0.87
Redlands	0.77	0.80	0.89	0.83	0.90	0.92	0.86	0.88	0.83	0.61	0.82	0.71
Reseda	0.76	0.82	0.85	0.72	0.87	0.88	0.90	0.87	0.89	0.89	0.87	0.71
West LA	0.75	0.82	0.84	0.74	0.74	0.74	0.93	0.85	0.87	0.81	0.80	0.76
Winchester	0.77	0.71	0.86	0.73	0.79	0.78	0.75	0.69	0.58	0.56	0.70	0.71


Upland (Jan to Feb)
San Bernardino (Jan to Feb)
Santa Clarita (Jan to Feb)
San Gabriel (Jan to Feb)
Rubidoux (Jan to Feb)
Mira Loma (Jan to Feb)
2016-2019 Mean
2020 Actual Mean
2020 ML Mean
Ozone Concentrations [ppb]
Hour

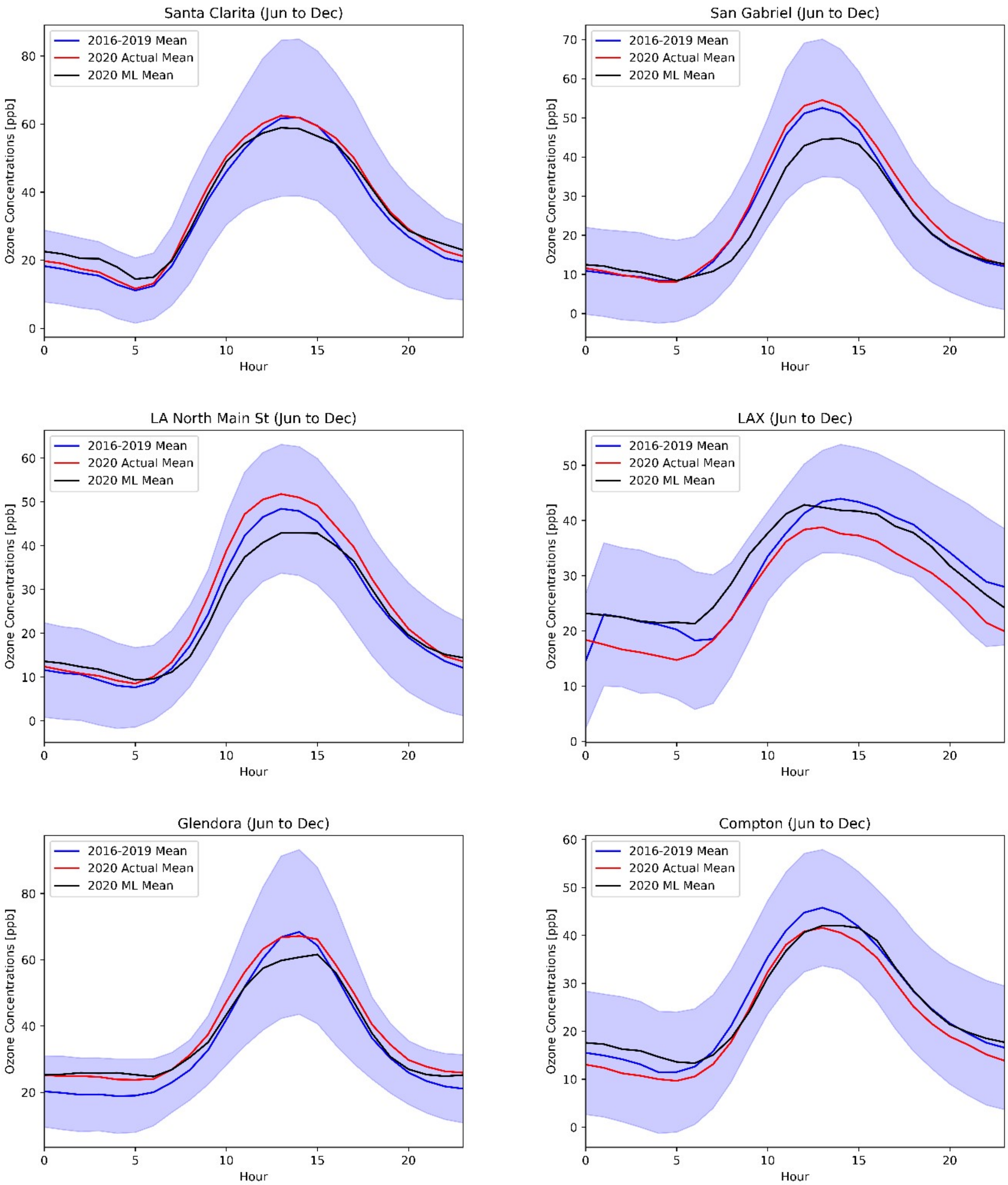

LA North Main St (Jan to Feb)
LAX (Jan to Feb)
Glendora (Jan to Feb)
Compton (Jan to Feb)
Banning (Jan to Feb)
Anaheim (Jan to Feb)
2016-2019 Mean
2020 Actual Mean
2020 ML Mean
Ozone Concentrations [ppb]
Hour


Azusa (Jan to Feb)
Upland (Mar to May)
San Bernardino (Mar to May)
Santa Clarita (Mar to May)
San Gabriel (Mar to May)
Rubidoux (Mar to May)
2016-2019 Mean
2020 Actual Mean
2020 ML Mean
Ozone Concentrations [ppb]
Hour


Mira Loma (Mar to May)
2016-2019 Mean
2020 Actual Mean
2020 ML Mean
Ozone Concentrations [ppb]
0
10
20
30
40
50
60
70
Hour
0
5
10
15
20
LA North Main St (Mar to May)
2016-2019 Mean
2020 Actual Mean
2020 ML Mean
Ozone Concentrations [ppb]
0
10
20
30
40
50
60
Hour
0
5
10
15
20
LAX (Mar to May)
2016-2019 Mean
2020 Actual Mean
2020 ML Mean
Ozone Concentrations [ppb]
10
20
30
40
50
60
Hour
0
5
10
15
20
Glendora (Mar to May)
2016-2019 Mean
2020 Actual Mean
2020 ML Mean
Ozone Concentrations [ppb]
10
20
30
40
50
60
70
Hour
0
5
10
15
20
Compton (Mar to May)
2016-2019 Mean
2020 Actual Mean
2020 ML Mean
Ozone Concentrations [ppb]
0
10
20
30
40
50
Hour
0
5
10
15
20
Banning (Mar to May)
2016-2019 Mean
2020 Actual Mean
2020 ML Mean
Ozone Concentrations [ppb]
20
30
40
50
60
70
Hour
0
5
10
15
20



Anaheim (Mar to May)
Azusa (Mar to May)
Upland (Jun to Dec)
Rubidoux (Jun to Dec)
Mira Loma (Jun to Dec)
San Bernardino (Jun to Dec)
2016-2019 Mean
2020 Actual Mean
2020 ML Mean
Ozone Concentrations [ppb]
Hour


Santa Clarita (Jun to Dec)
San Gabriel (Jun to Dec)
LA North Main St (Jun to Dec)
LAX (Jun to Dec)
Glendora (Jun to Dec)
Compton (Jun to Dec)
2016-2019 Mean
2020 Actual Mean
2020 ML Mean
Ozone Concentrations [ppb]
Hour

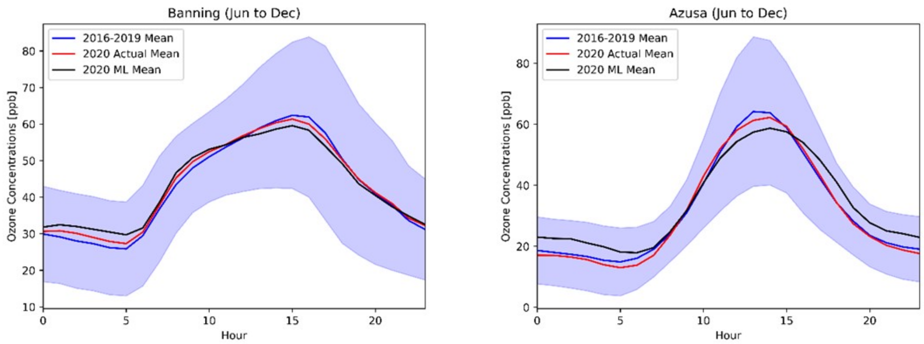
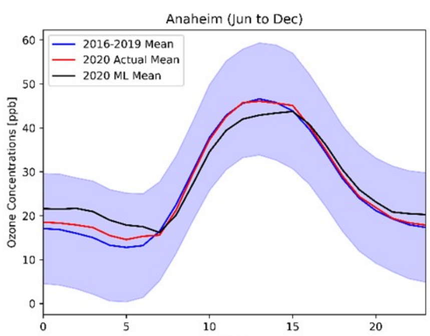
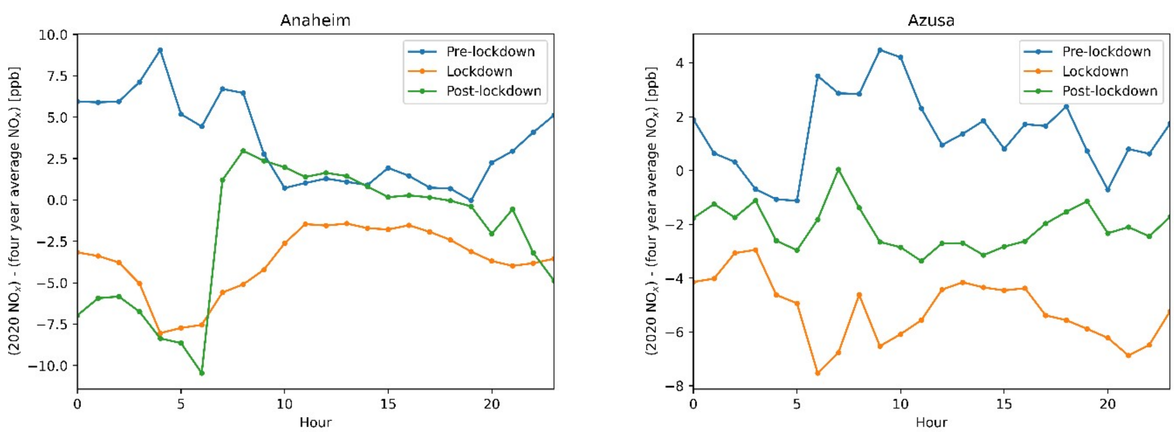


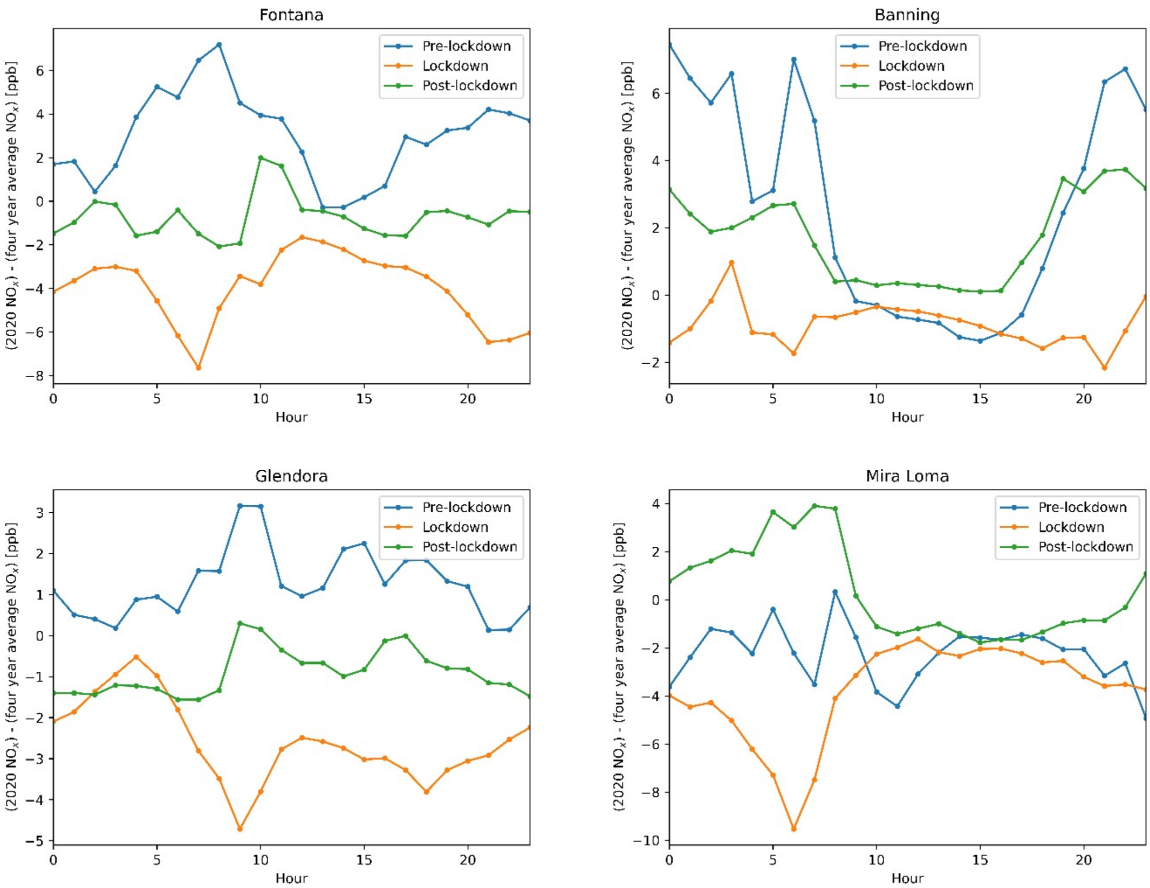
Figure S12. Averaged diurnal profiles of 2016 - 2019 (blue), actual 2020 (red), and ML predicted 2020 (black) ozone concentrations (ppb) for three different periods, the pre lockdown (Jan to Feb), the lockdown (Mar to May), and the post lockdown period (after May). The shaded area is the standard deviation of the 2016 - 2019 measurements.



Comment [KD]: Changed Figure S12.



Anaheim
Azusa
Pre-lockdown
Lockdown
Post-lockdown
(2020 NO$_x$) - (four year average NO$_x$) [ppb]
Hour



Fontana
Banning
Glendora
Mira Loma
Pre-lockdown
Lockdown
Post-lockdown
(2020 NO$_x$) - (four year average NO$_x$) [ppb]
Hour

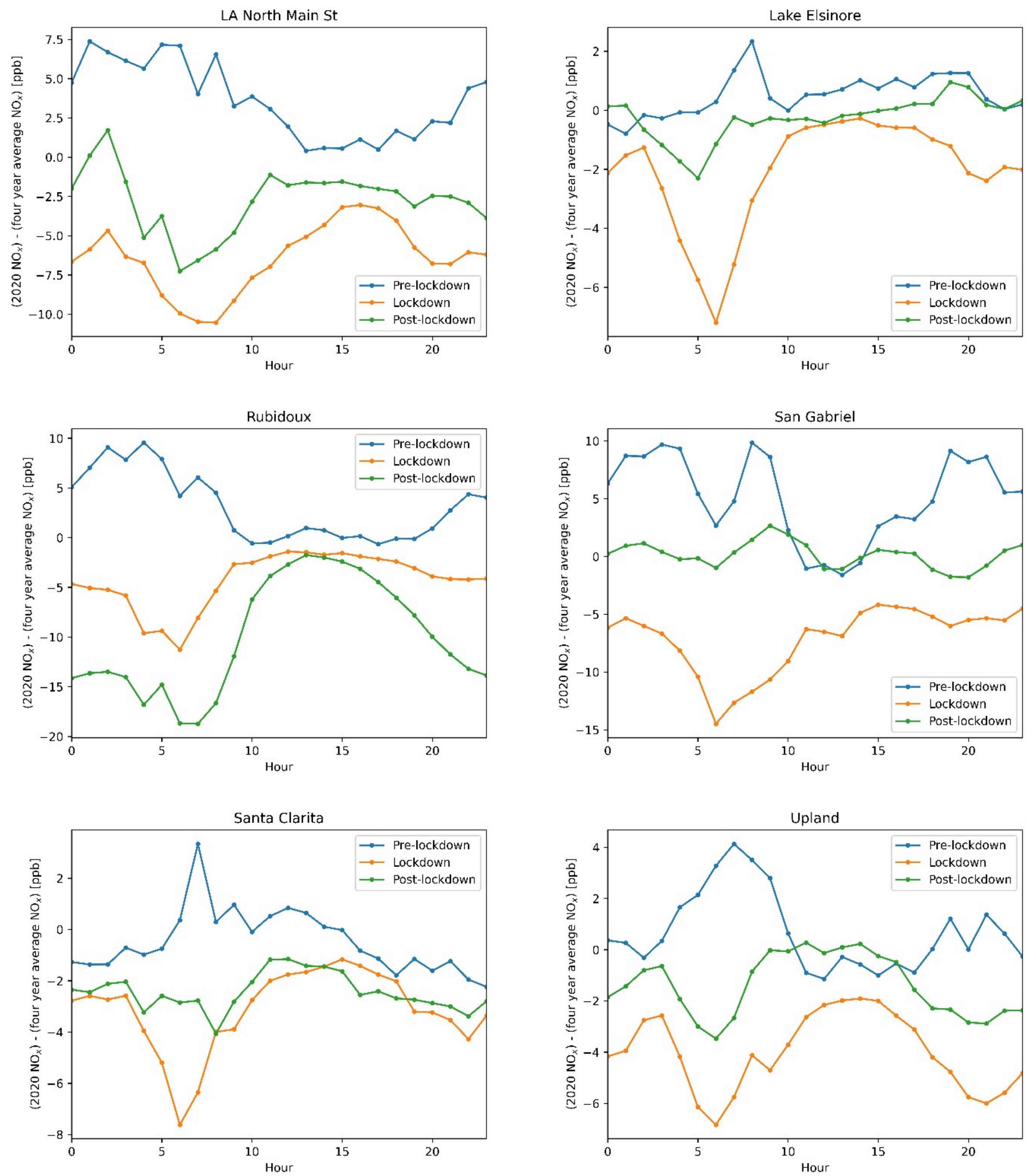
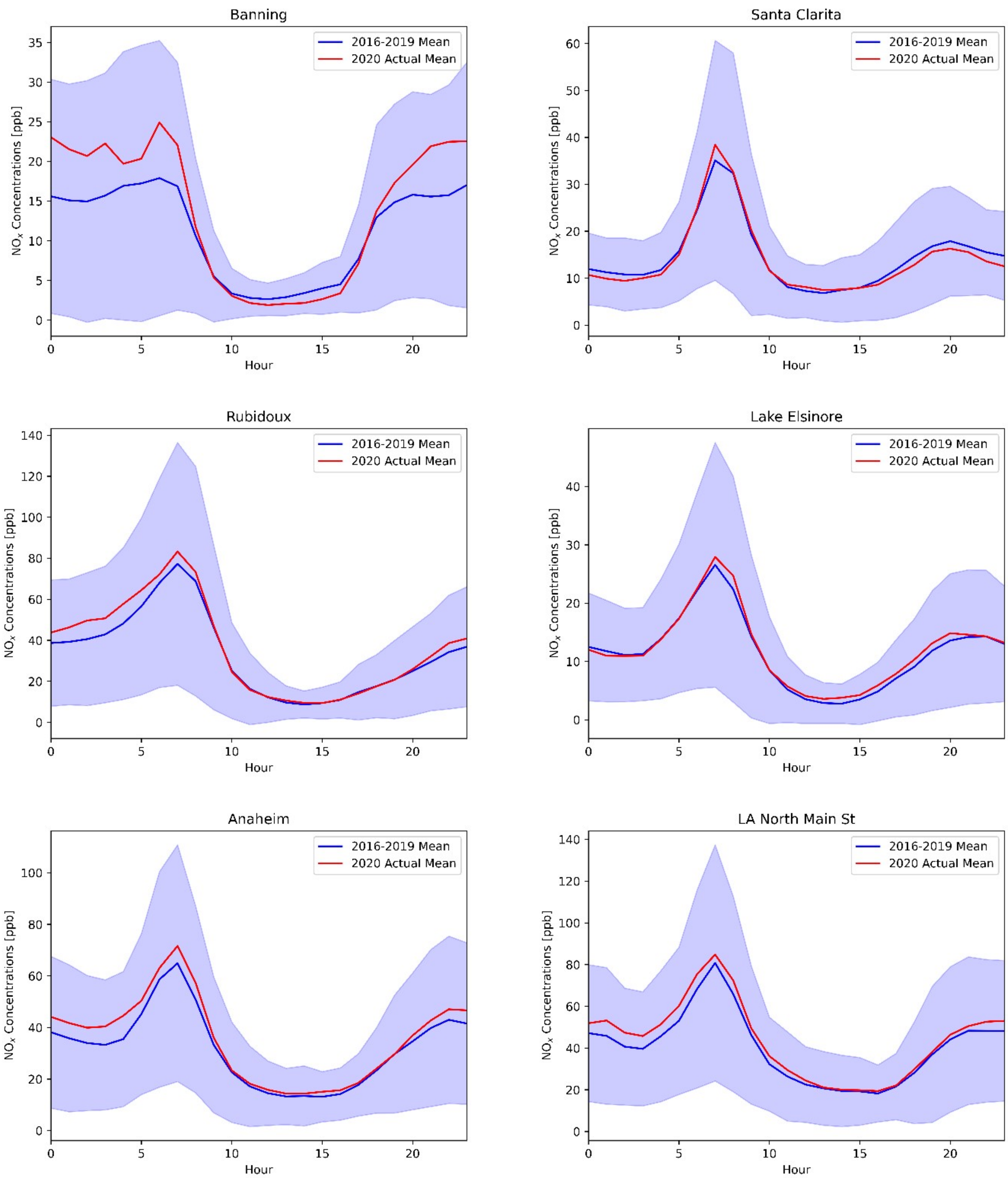


Figure S13. Averaged diurnal profiles of the differences between 2020 NO$_x$ and the average 2016 - 2019 NO$_x$ for three periods, pre-lockdown (blue), lockdown (orange), and post-lockdown (green).

Pre-lockdown



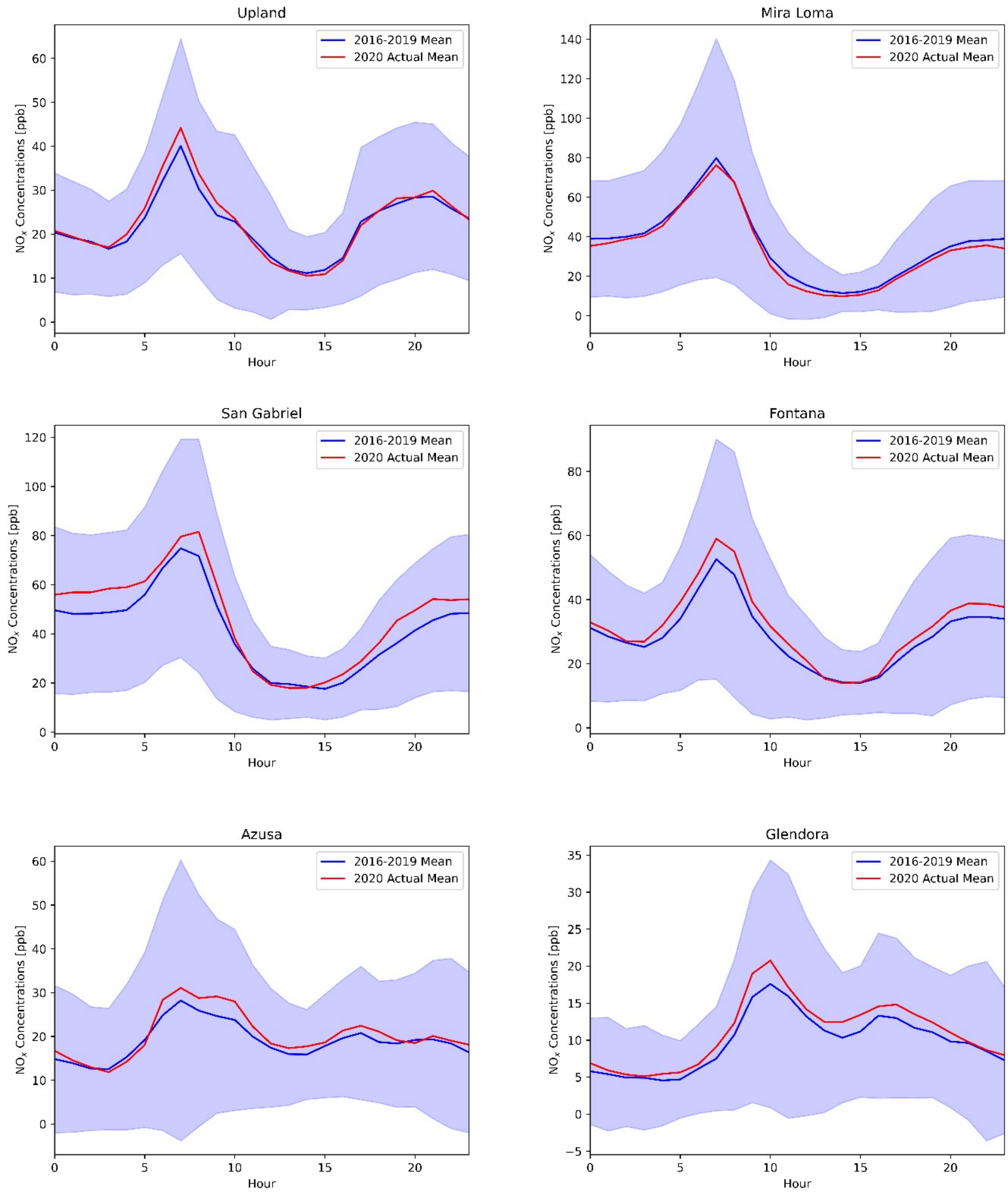
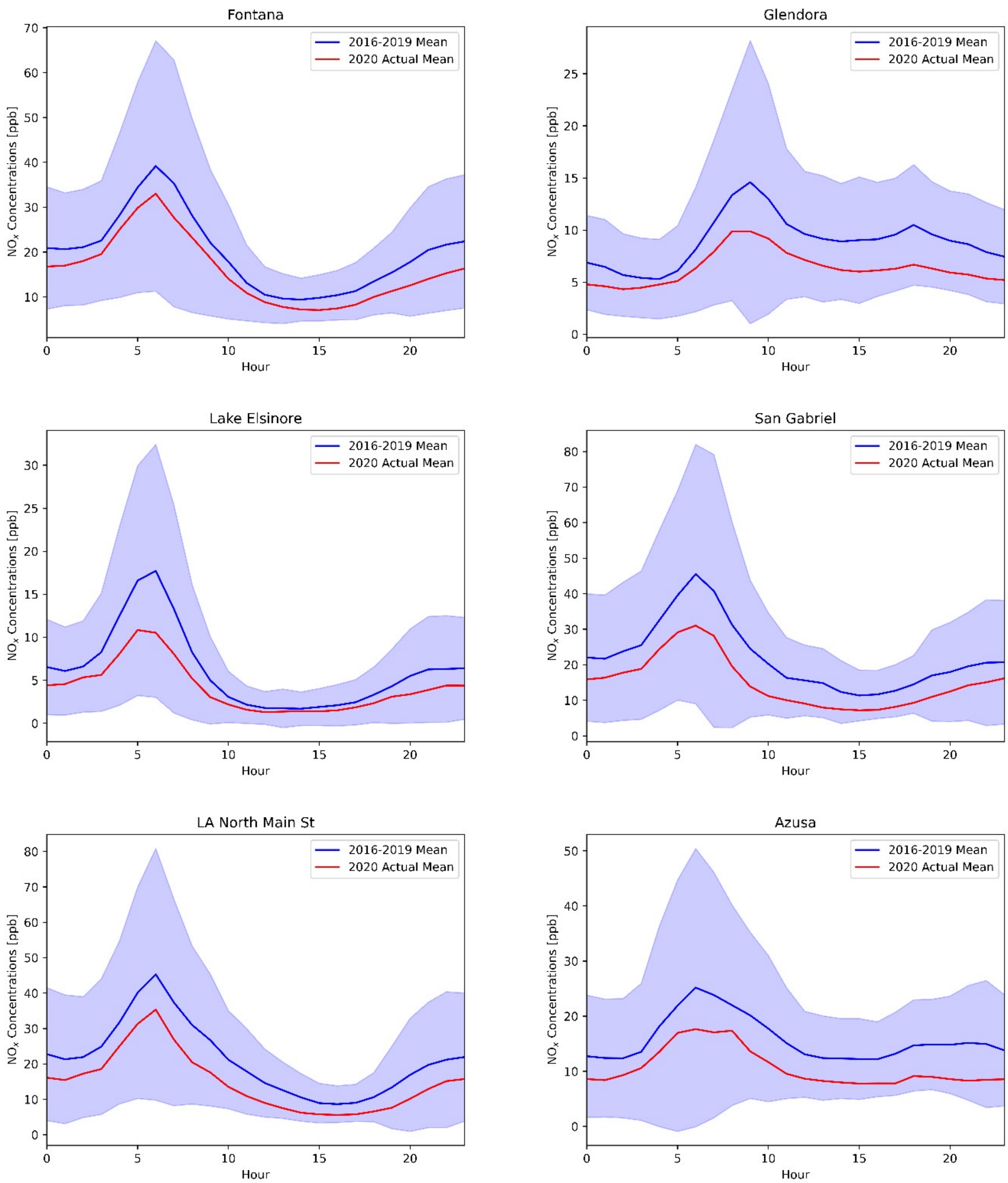


Figure S14. Averaged diurnal profiles of 2016 - 2019 (blue), actual 2020 (red) NO_x concentrations (ppb) for before the lockdown (Jan to Feb).

During Lockdown



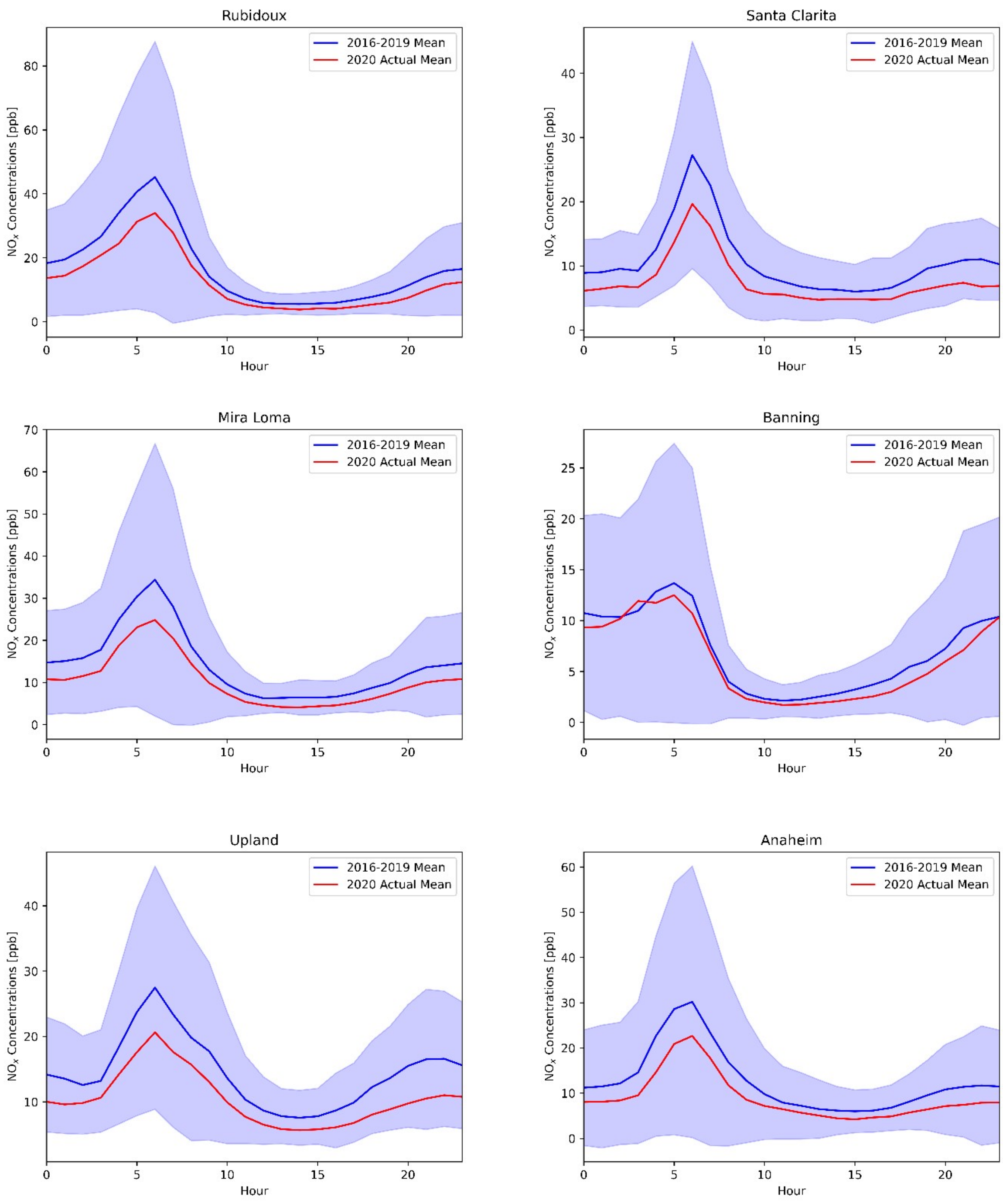
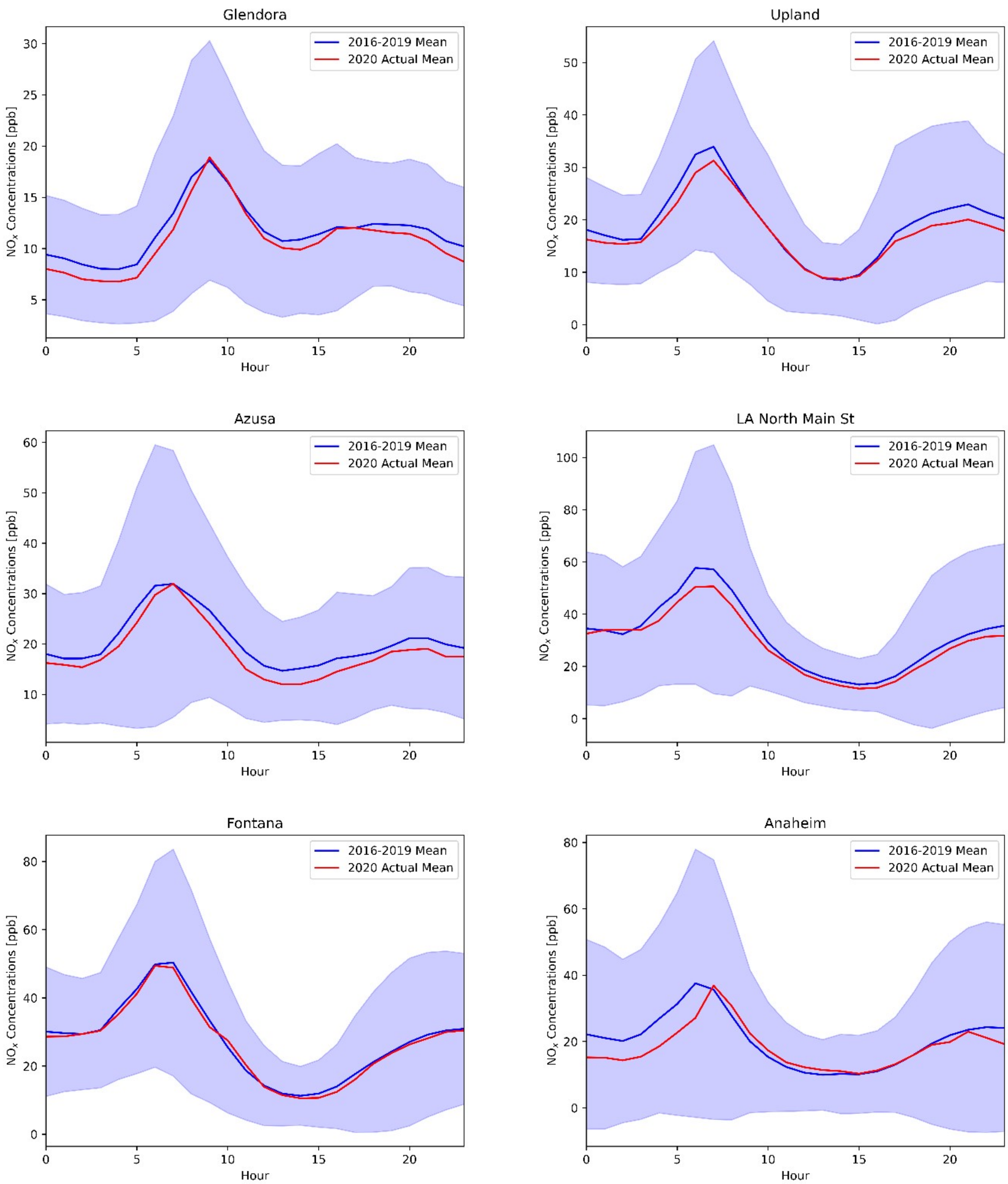


Figure S15. Averaged diurnal profiles of 2016 - 2019 (blue), actual 2020 (red) NOx concentrations (ppb) during the lockdown (Mar to May).

Post-Lockdown



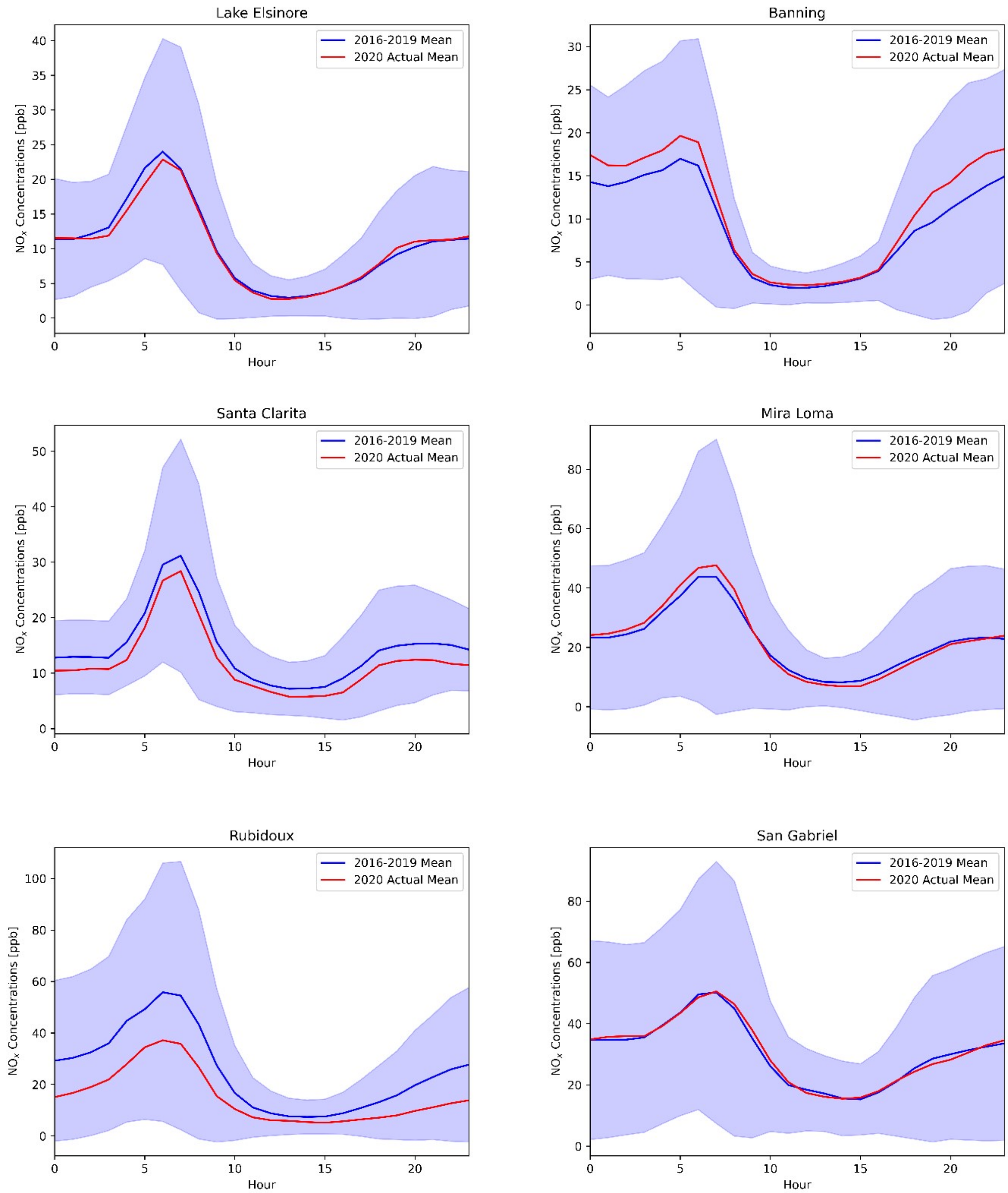
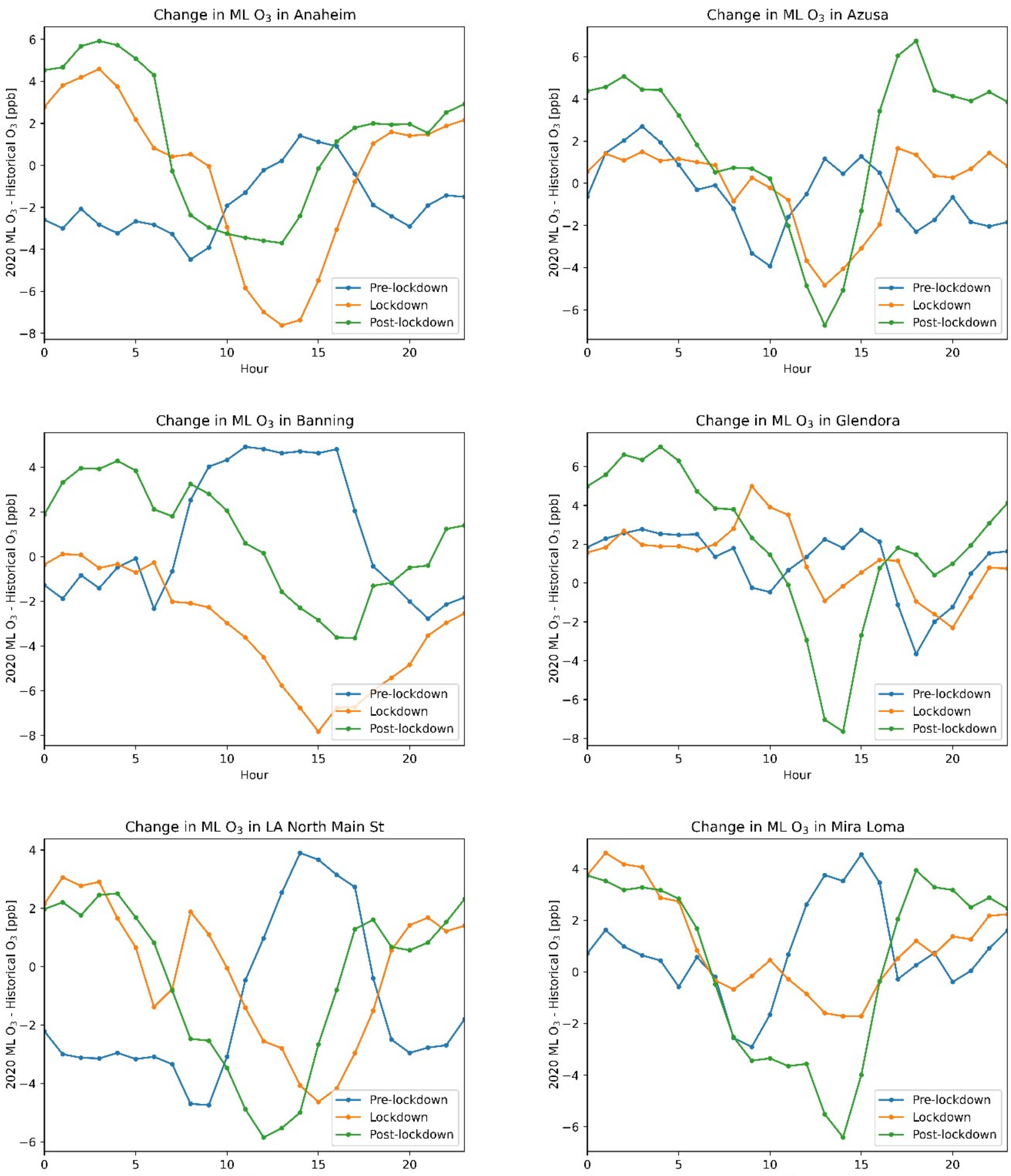


Figure S16. Averaged diurnal profiles of 2016 - 2019 (blue), actual 2020 (red) NOx concentrations (ppb) after the lockdown (Jun to Dec).

Changes in O_3 between 2020 ML O_3 and average of 2016 to 2019 O_3



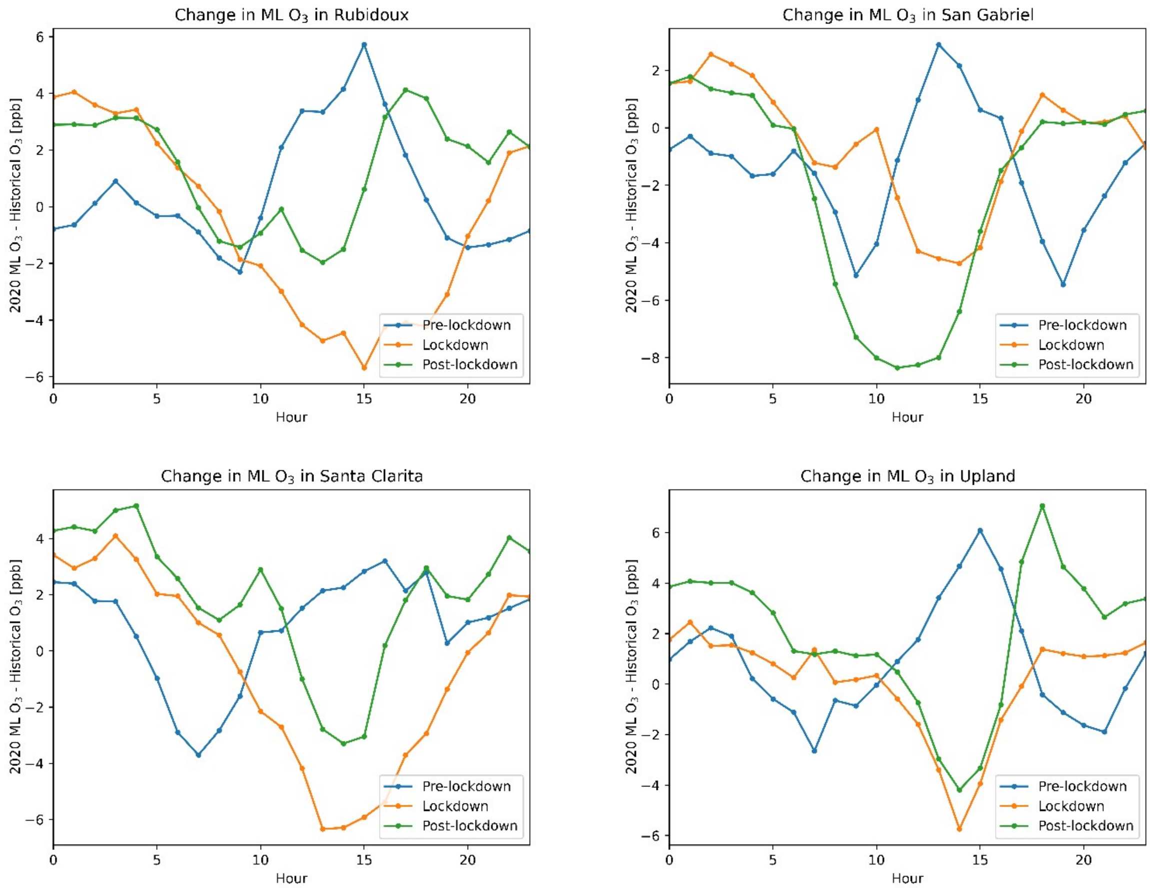
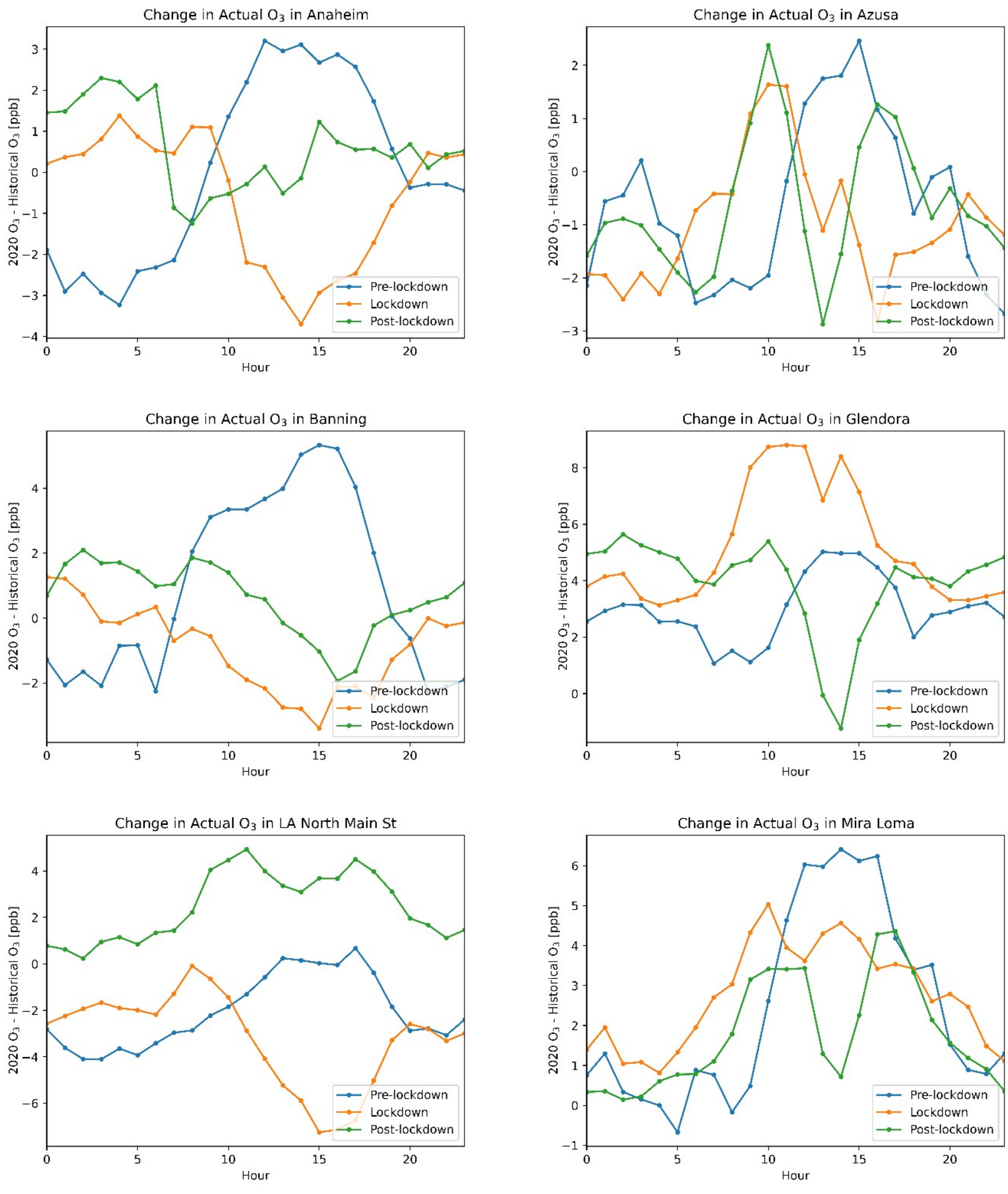


Figure S17. Diurnal changes in O_3 between 2020 machine learning prediction and the average of 2016 to 2019 O_3.

Changes in O_3 between 2020 actual O_3 and average of 2016 to 2019 O_3



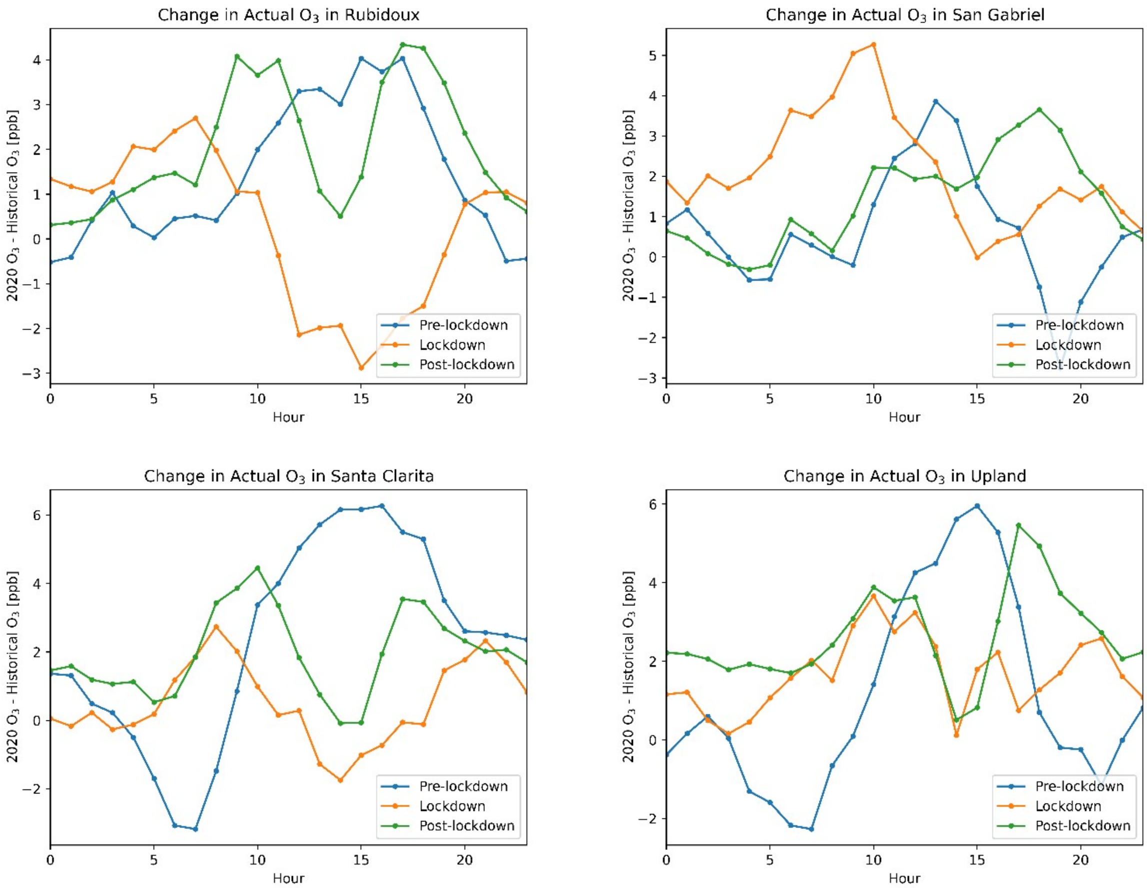


Figure S18. Diurnal changes in O_3 between 2020 actual O_3 and the average of 2016 to 2019 O_3.

Comparing between three interpolation methods (Kriging, Bicubic, IDW interpolation)

The bicubic R^2 indicates the poorest performance of the three interpolation methods. The lowest R^2 values for the 12 evaluation sites were 0.15 and 0.29, Mission Viejo and West LA, respectively (Table 4). The poor performance resulted from the method used to calculate the coefficients a_{ij} (Eq. 5), for which the values of coefficients did not depend on the distance between interpolating points but were dependent on the formation of a smooth curve. Bicubic is best for evenly distributed points, such as interpolating image pixels. IDW showed a significant improvement compared to bicubic interpolation. The lowest R^2 was 0.36 for Mission Viejo, and the highest R^2 was 0.83 for Pomona. Since IDW accounts for the distances between the interpolation points and the data points, farther data points have less influence on the interpolation points. Ordinary kriging resulted in the best interpolation method, with the lowest R^2 of 0.39 for Winchester and the highest R^2 of 0.84 for Pomona. Kriging not only accounts for the distance between building points and interpolated data by assigning larger weight λ_i to the nearest neighbors, but it also considers the variability of data by considering the variance of input data, σ^2. The basis of the variogram function represents the spatial variability of data. The variance depends not on observation values but on the variogram model and geometry (Eq. 2) (1).

1. Kebaili Bargaoui Z, Chebbi A. Comparison of two kriging interpolation methods applied to spatiotemporal rainfall. Journal of Hydrology [Internet]. 2009 Feb [cited 2023 Jan 30];365(1–2):56–73. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0022169408005726