Development and Implementation of Machine Learning Tools for Ozone Formation in the LADCO Region

Hantao Wang

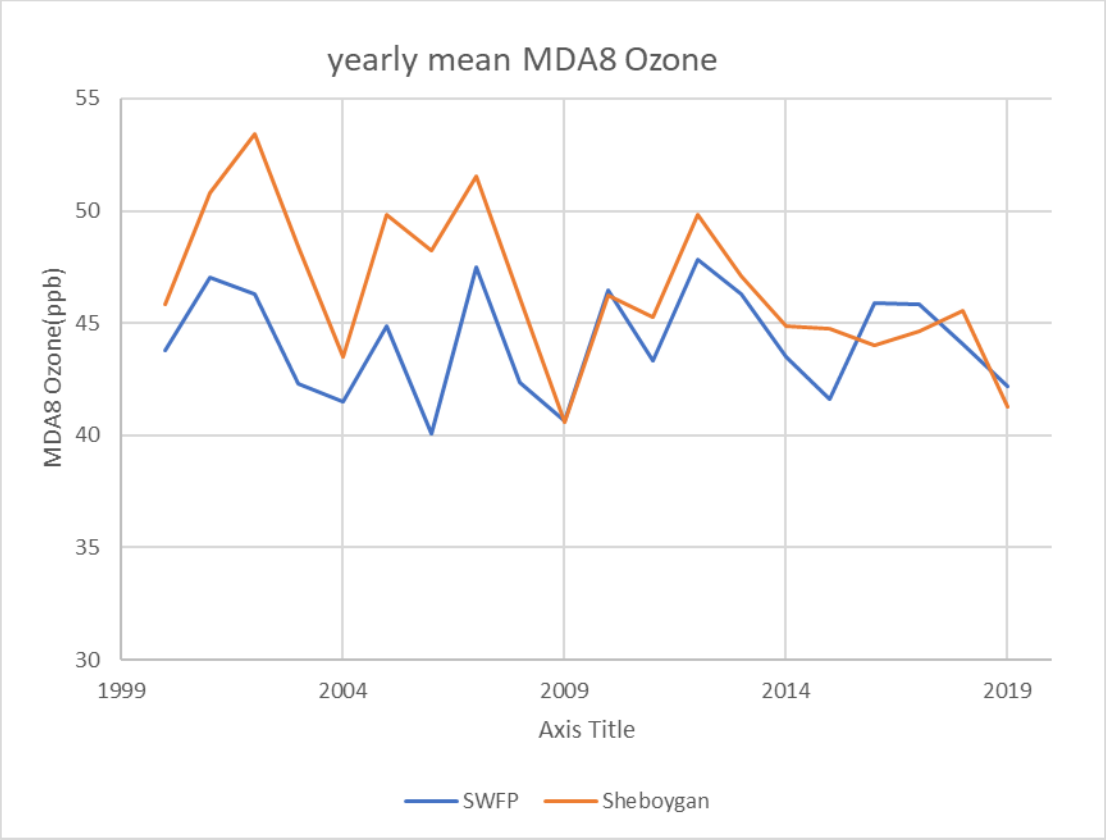
Objectives

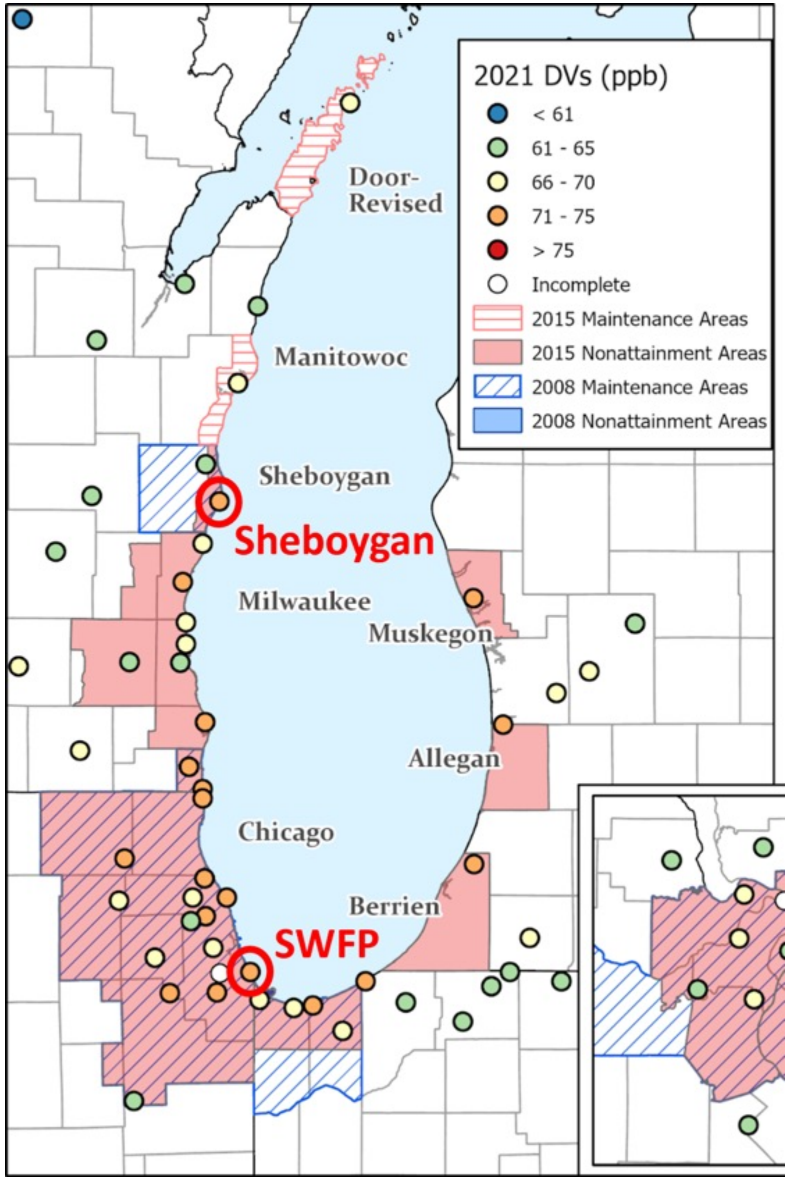
- Reproduce the GAM analysis for Sheboygan
- Refine the GAM analysis for Sheboygan
- Adjust the annual ozone trends for meteorology
- Apply the GAM analysis to other areas in the LADCO region(SWFP)

Background

- Camalier et al., 2007 developed a generalized additive model (GAM) to assess the impacts of meteorology on ozone. The method is cited by EPA (2018) for use as part of weight-of-evidence analysis for ozone attainment demonstrations. Wells et al., 2021, further refined this model.
- Dr. Charles L. Blanchard develop and extend this EPA GAM to describe the relative influences of weather, emissions on ozone in the southern Lake Michigan area (under a contract for LADCO and WDNR).
- We develop and extend a quantile regression to Dr. Blanchard’s GAM to analyze the relative influences of meteorology on ozone at two sites in the Lake Michigan area.
- By replacing the imported initial data, our model can analyze the impacts of meteorological conditions on ozone in different regions.

Apply the GAM analysis in the LADCO region (Sheboygan and SWFP)





Methods

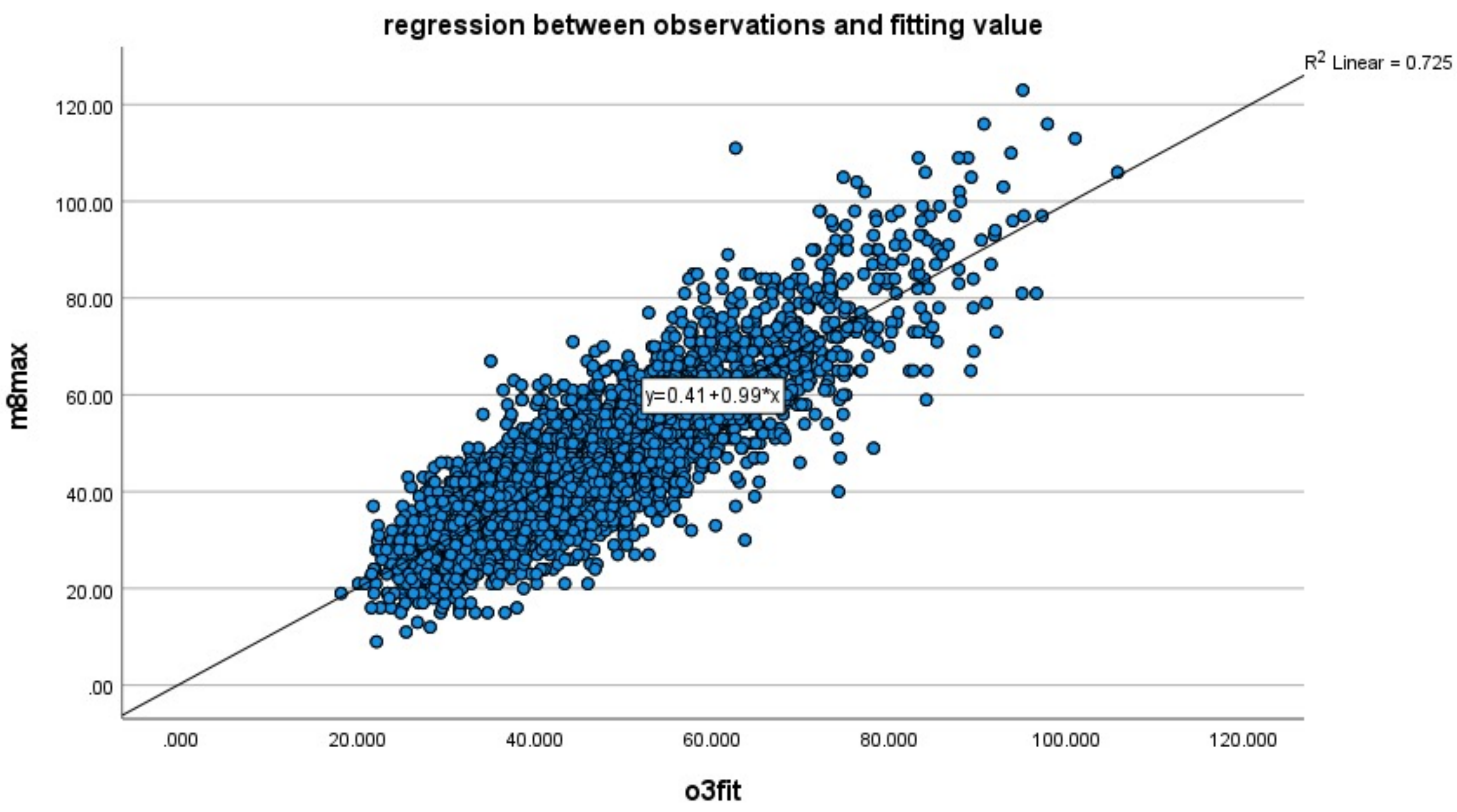
- GAM(generalized additive model) analysis
- With Log Link function and the Gaussian Distribution

$$l(O_3)_i = \mu + f_1(x_1)_i + \ldots + f_m(x_m)_i + g_1(y_1)_i + \ldots + g_n(y_n)_i + h_1(z_1) + \ldots + h_p(z_p) + e_i$$

$l(O_3)$ is the logarithm of the peak 8-hour O3 on day "i"
μ represent the intercept of the regression
x parameterize the associations of meteorological variables
y parameterize associations of ambient concentrations of O_3 precursors
z represent temporal variables, including "day of week" and "year"
e is the difference between observed and predicted O_3 (error).
f,g,h are the functions, which are generated by the GAM

- The GAM method was used to analyze the trend of MDA8 Ozone concentration and the effect of different meteorological conditions on ozone.

Reproduce the GAM analysis



Model Summary[b]

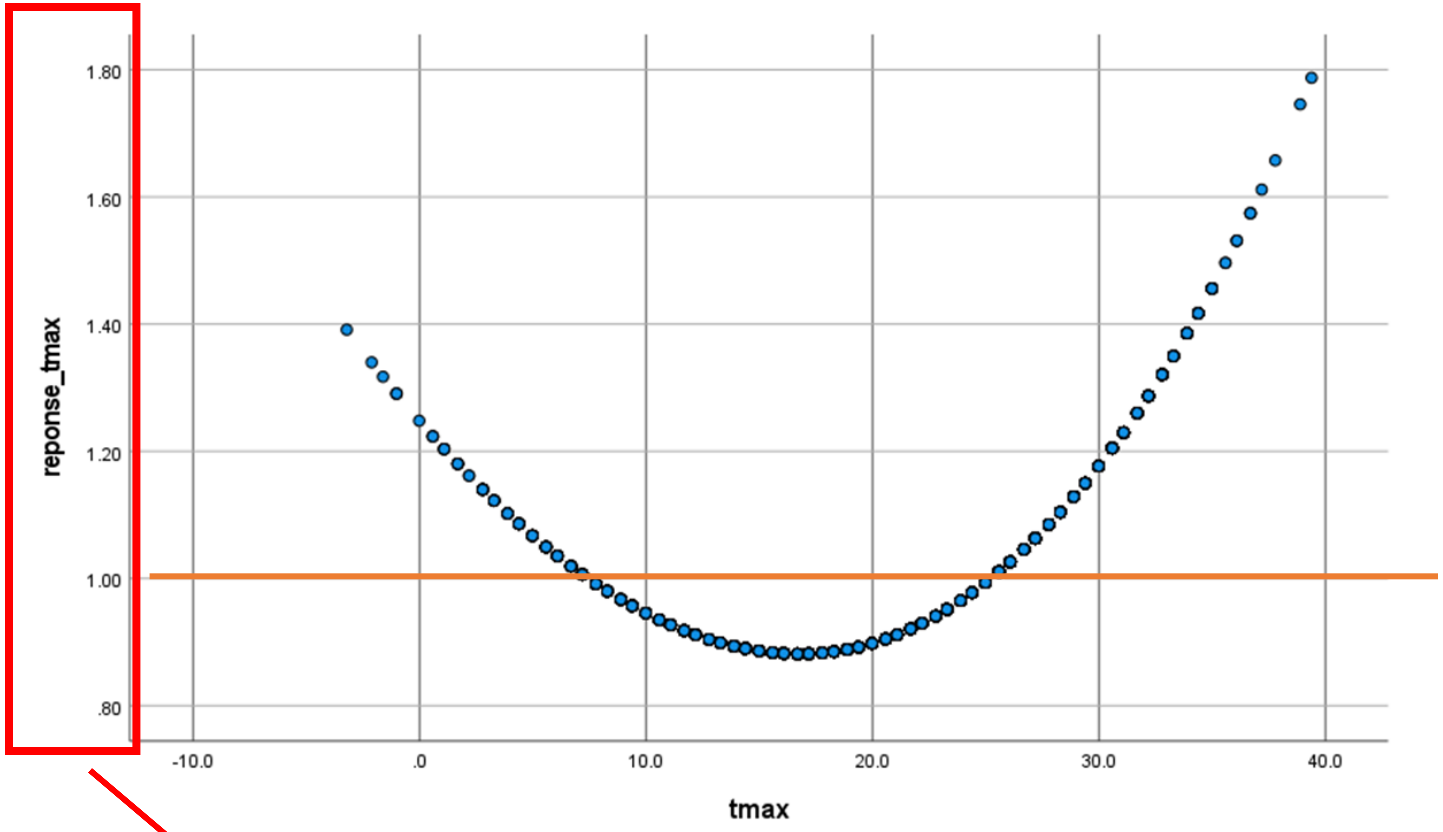
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.851[a]	.725	.725	8.06461	1.973

a. Predictors: (Constant), o3fit

b. Dependent Variable: m8max.x

The regression between ozone observations and fitting values from GAM, shows a high R-square(0.725). Means GAM works well.

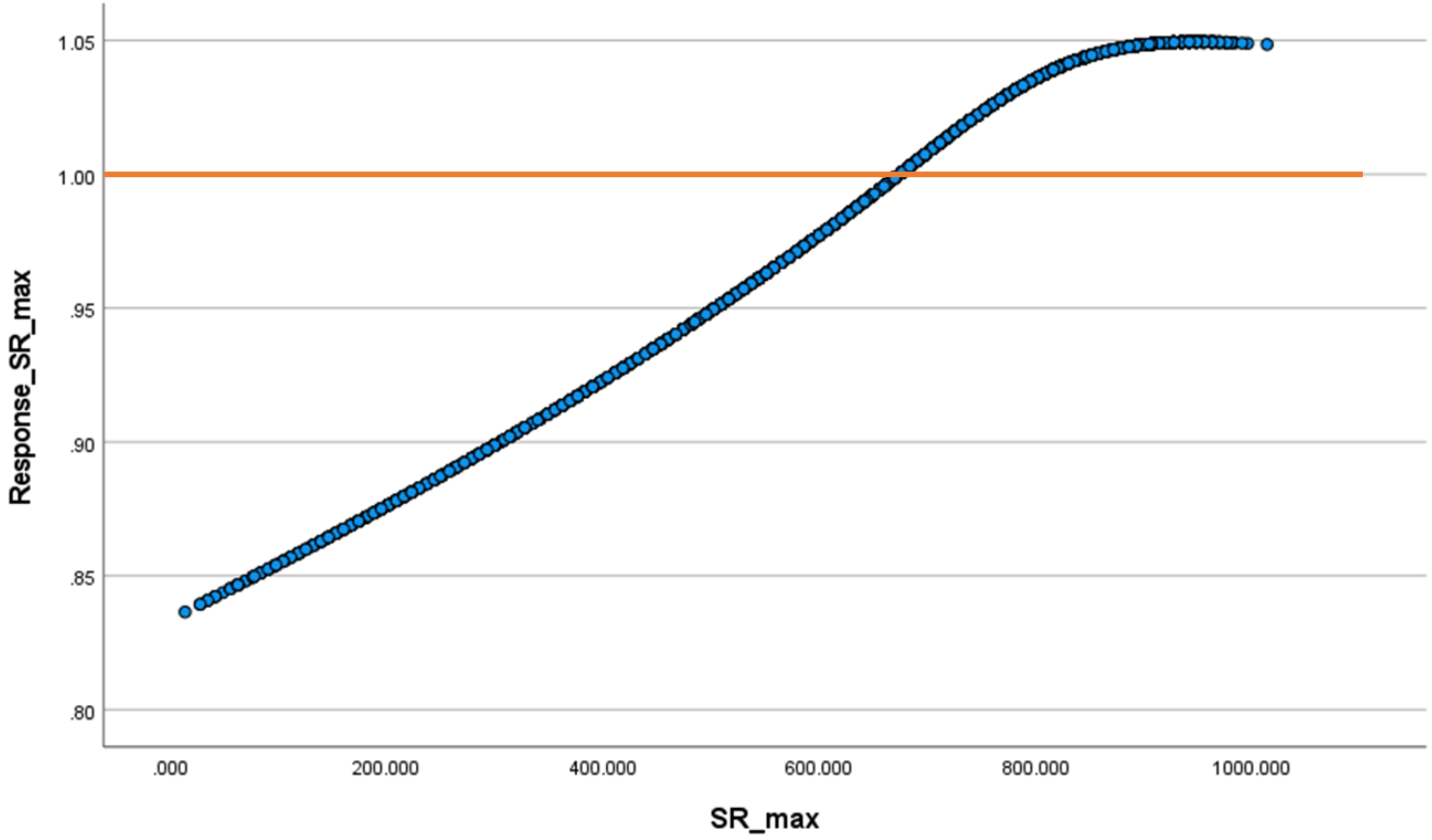
relative influences of meteorology on ozone



The tmax represents the maximum temperature on this day.
From the plot that as the temperature increases, the effect on ozone decreases from positive to negative and then increase again. The inhibition of ozone production is greatest at around 17 degree.

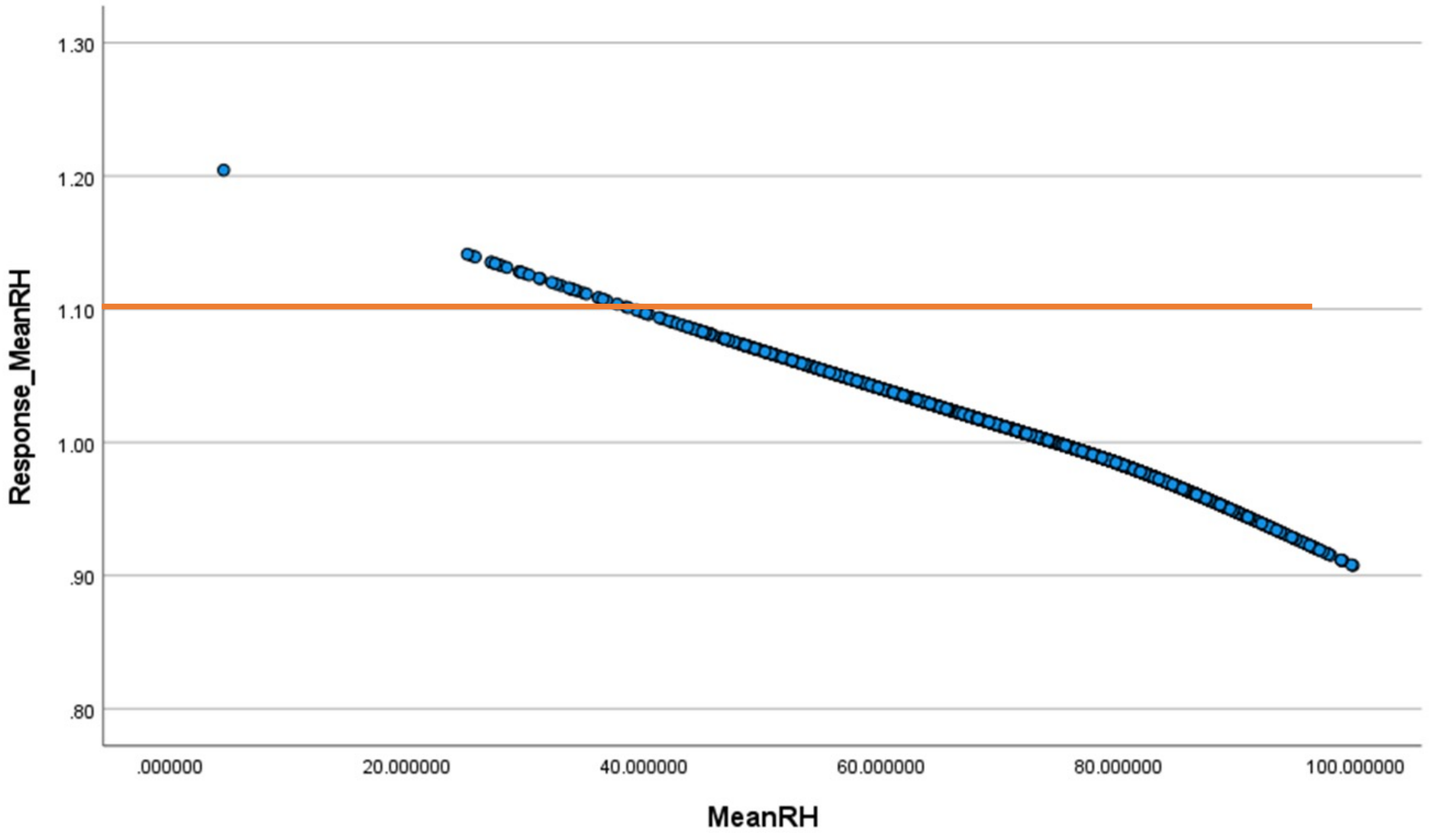
The vertical coordinate represents the effect on ozone and is the ratio of the ozone concentration at that condition to the average MDA8 ozone concentrations for the year, with greater than 1 indicating an increase and less than 1 indicating a decrease.

relative influences of meteorology on ozone



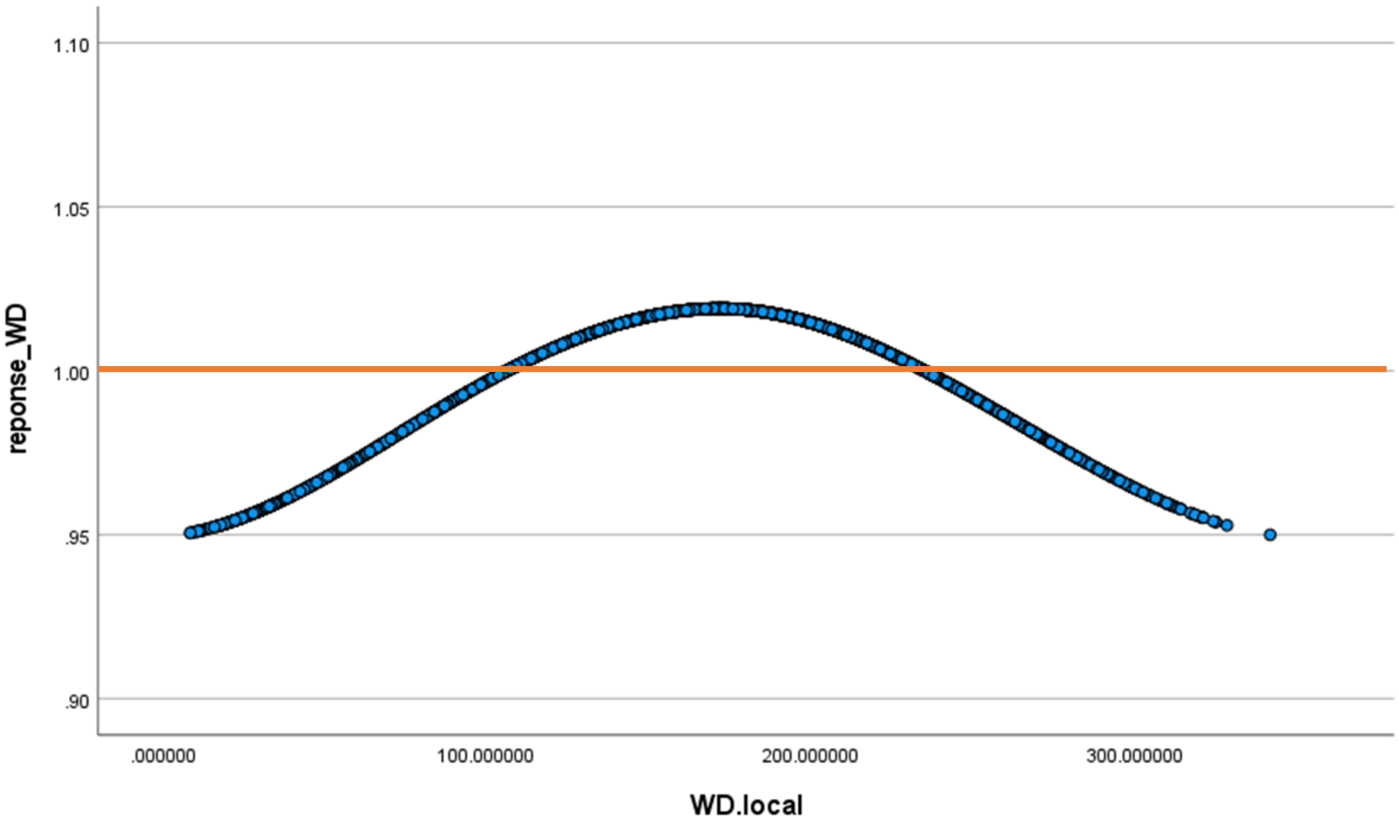
SRmax represents the maximum solar radiation on this day.
From the plot that the effect on ozone is increasing all the time as the solar radiation increases, but the limit is 1.05.

relative influences of meteorology on ozone



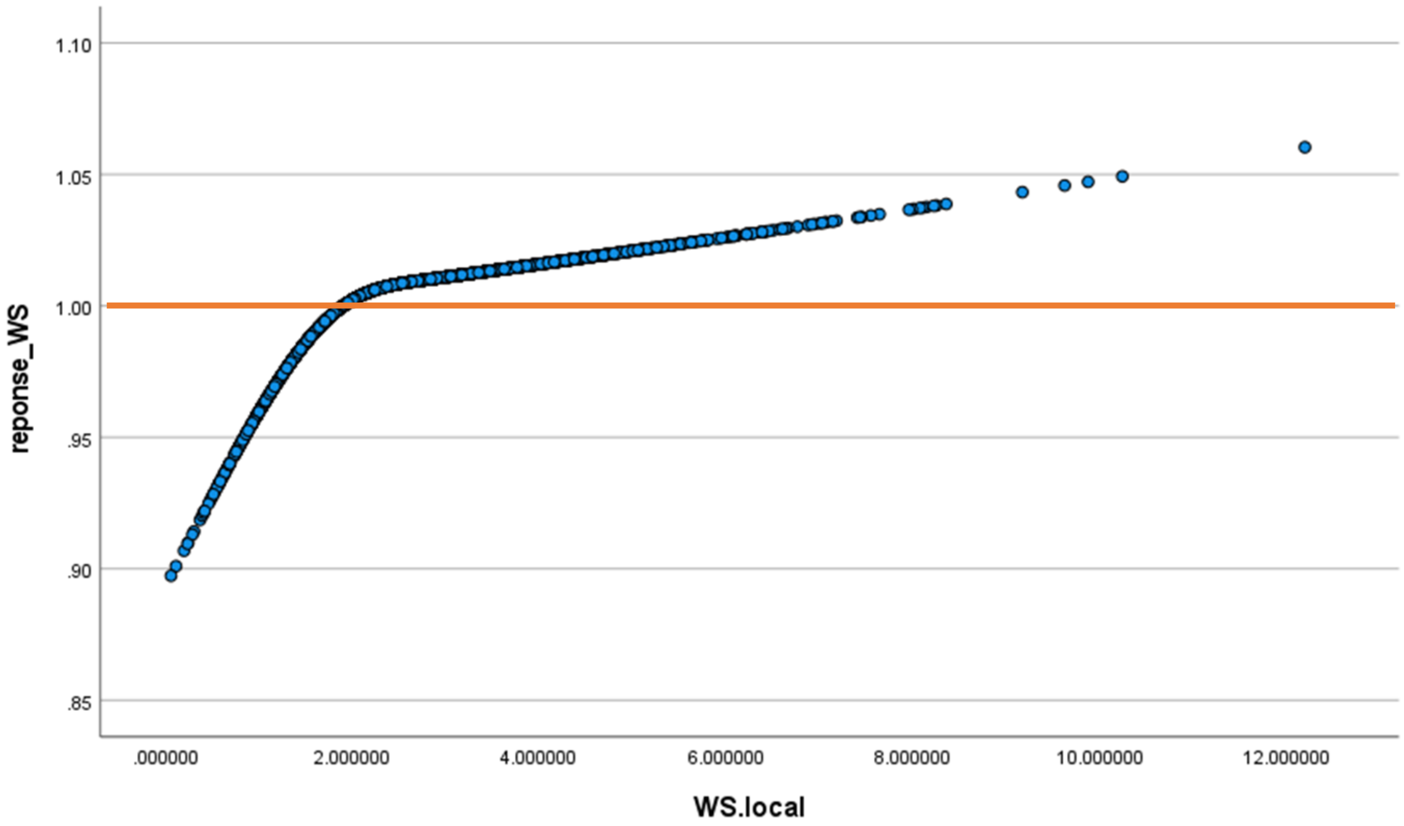
MeanRH represents the average relative humidity of the day.
From the plot, the relative humidity has a negative effect on ozone, and the negative effect increases with increasing humidity

relative influences of meteorology on ozone



WD.local represents the local wind direction. From the plot, ozone increases in a specific range of wind directions, while in other wind direction ranges, ozone keeps decreasing. This is related to the direction of the lake breeze at this site.

relative influences of meteorology on ozone



WS.local represents the local wind speed. From the figure, it can be seen that as the wind speed increases, its negative effect on ozone decreases, while the positive effect keeps increasing.

Refine the GAM analysis

Variables selection

Akaike Information Criterion(AIC) and Analysis of Variance(ANOVA)

```
                                          Df Deviance   AIC  F value     Pr(F)
<none>                                        231450 25128
ns(tmax, 3)                                3  264683 25600 166.3239 < 2.2e-16 ***
ns(tmin, 3)                                3  232281 25135   4.1611  0.005955 **
ns(LM_surf_T, 3)                           3  231964 25130   2.5718  0.052446 .
ns(MeanRH, 3)                              3  235852 25190  22.0329 3.945e-14 ***
ns(SR_max, 3)                              3  239705 25247  41.3136 < 2.2e-16 ***
ns(MeanSondeBP, 3)                         3  233898 25160  12.2519 5.679e-08 ***
ns(WS.local, 3)                            3  232888 25145   7.1995 8.151e-05 ***
bc(WD.local, period = 360, nknots = 4)     3  232863 25144   7.0736 9.756e-05 ***
bc(WD.midday, period = 360, nknots = 4)    3  235814 25189  21.8405 5.215e-14 ***
bc(Mean_WD850, period = 360, nknots = 4)   3  234599 25171  15.7602 3.530e-10 ***
bc(Mean_WD500, period = 360, nknots = 4)   3  231979 25131   2.6471  0.047411 *
ns(WS850mb, 3)                             3  232280 25135   4.1553  0.006004 **
ns(WS500mb, 3)                             3  236810 25204  26.8289 < 2.2e-16 ***
ns(Ht850mb, 3)                             3  233597 25155  10.7490 4.979e-07 ***
ns(Ht500mb, 3)                             3  231695 25126   1.2257  0.298695
ns(GB_Li_BPsurf, 3)                        3  234060 25162  13.0634 1.756e-08 ***
ns(GB_Dt_BPsurf, 3)                        3  234835 25174  16.9429 6.352e-11 ***
ns(jday, 3)                                3  235199 25180  18.7629 4.532e-12 ***
as.factor(X1st.Max.Hour)                  16  255532 25449  22.5983 < 2.2e-16 ***
dowf                                       6  235212 25174   9.4143 2.825e-10 ***
sumemissions                               1  231519 25128   1.0441  0.306951
ns(MultiNOx, 3)                            3  238362 25227  34.5950 < 2.2e-16 ***
ns(MultiCO, 3)                             3  233807 25159  11.7995 1.092e-07 ***
```

```r
predval <- function(mnam,fitm,fitr)
 {
  const <- mean(fitm$y)
  tempf <- predict(fitm, type = "terms")
  tempr <- predict(fitr, type = "terms")
  pred1 <- tempf[,1]
  pred2 <- tempf[,3]
  pred3 <- tempf[,4]
  pred4 <- tempf[,7]
  pred5 <- tempf[,8]
  pvf   <- const*exp(pred1)*exp(pred2)*exp(pred3)*exp(pred4)*exp(pred5)
  pvr   <- const*exp(tempr[,"yrf"])
  dat2 <- cbind(year=dat$year,pvf=pvf,pvr=pvr)
  pboth <- aggregate(dat2[,c("pvf","pvr")],list(year=dat2[,"year"]),mean)
  pboth$year <- as.numeric(as.character(pboth$year))
  zz <- data.frame(mnam=mnam,pboth)
  zz
 }
pval <- predval(mnam,fitm,fitr)
```

Our model automatically outputs the results of the analysis of the each independent variables and selects the meteorological variables that have the greatest impact on ozone. It is possible to freely select the variables when calculating the adjusted mean.

Variables selection for Sheboygan and SWFP

	Df	Deviance	AIC	F value	Pr(F)	
<none>		268909	28318			
ns(tmax, 3)	3	281007	28487	58.6362	< 2.2e-16	***
ns(tmin, 3)	3	269270	28317	1.7536	0.1538464	
ns(LM_surf_T, 3)	3	269031	28313	0.5919	0.6202837	
ns(MeanRH, 3)	3	290436	28619	104.3403	< 2.2e-16	***
ns(SR_max, 3)	3	277315	28435	40.7446	< 2.2e-16	***
ns(MeanSondeBP, 3)	3	272477	28364	17.2967	3.725e-11	***
ns(WS.local, 3)	3	269048	28314	0.6763	0.5664575	
bc(WD.local, period = 360, nknots = 4)	3	271727	28353	13.6623	7.286e-09	***
bc(WD.midday, period = 360, nknots = 4)	3	271549	28351	12.7953	2.559e-08	***
bc(Mean_WD850, period = 360, nknots = 4)	3	270095	28329	5.7497	0.0006388	***
bc(Mean_WD500, period = 360, nknots = 4)	3	269378	28319	2.2735	0.0780226	.
ns(WS850mb, 3)	3	269601	28322	3.3550	0.0181182	*
ns(WS500mb, 3)	3	270872	28341	9.5149	2.929e-06	***
ns(Ht850mb, 3)	3	272298	28362	16.4290	1.314e-10	***
ns(Ht500mb, 3)	3	269364	28318	2.2096	0.0849019	.
ns(GB_Li_BPsurf, 3)	3	269765	28324	4.1513	0.0060307	**
ns(GB_Dt_BPsurf, 3)	3	270876	28341	9.5348	2.847e-06	***
ns(jday, 3)	3	275780	28412	33.3021	< 2.2e-16	***
as.factor(X1st.Max.Hour)	16	285659	28527	15.2219	< 2.2e-16	***
dowf	6	270557	28330	3.9945	0.0005435	***
sumemissions	1	270488	28339	22.9649	1.711e-06	***
ns(MultiNOx, 3)	3	273086	28373	20.2481	5.109e-13	***
ns(MultiCO, 3)	3	274261	28390	25.9398	< 2.2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Most important variance: SWFP

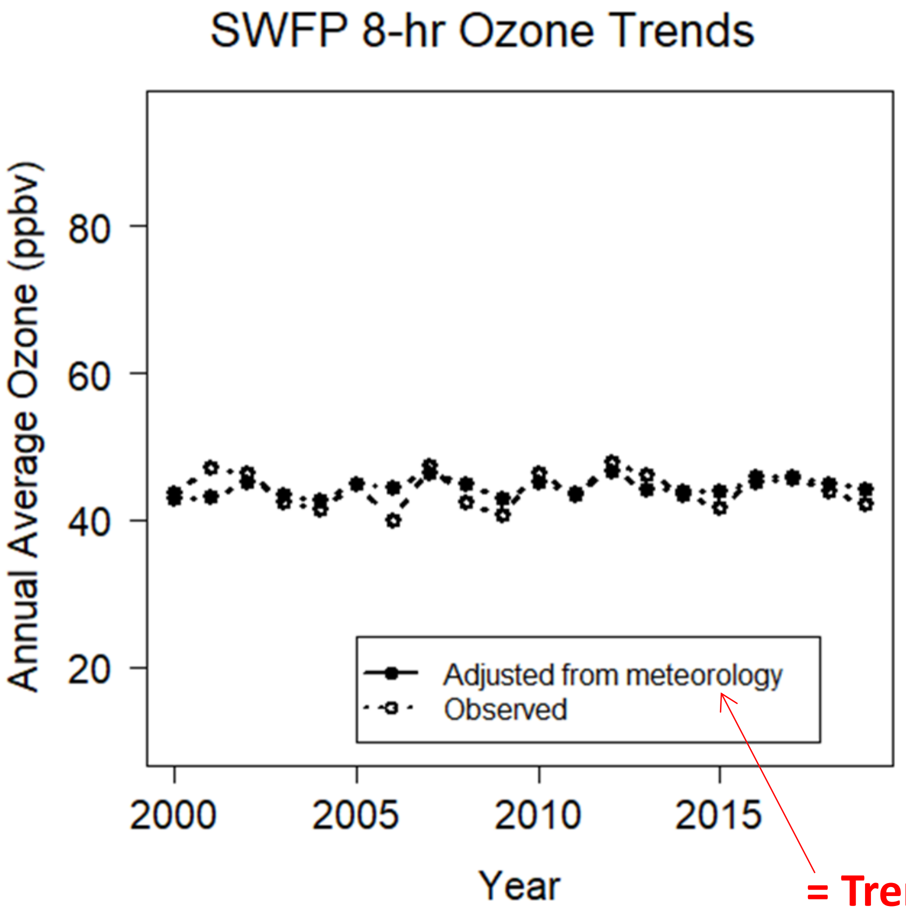
1. MeanRH
2. Tmax
3. SR_max
4. Ht850mb
5. WD.local

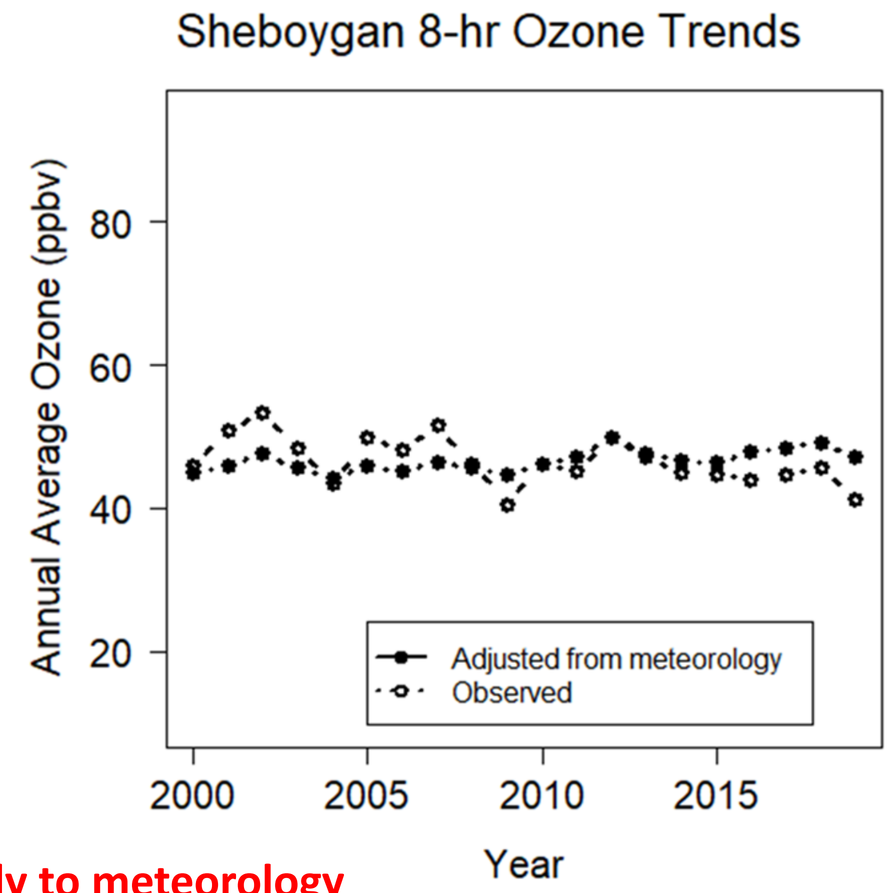
	Df	Deviance	AIC	F value	Pr(F)	
<none>		231450	25128			
ns(tmax, 3)	3	264683	25600	166.3239	< 2.2e-16	***
ns(tmin, 3)	3	232281	25135	4.1611	0.005955	**
ns(LM_surf_T, 3)	3	231964	25130	2.5718	0.052446	.
ns(MeanRH, 3)	3	235852	25190	22.0329	3.945e-14	***
ns(SR_max, 3)	3	239705	25247	41.3136	< 2.2e-16	***
ns(MeanSondeBP, 3)	3	233898	25160	12.2519	5.679e-08	***
ns(WS.local, 3)	3	232888	25145	7.1995	8.151e-05	***
bc(WD.local, period = 360, nknots = 4)	3	232863	25144	7.0736	9.756e-05	***
bc(WD.midday, period = 360, nknots = 4)	3	235814	25189	21.8405	5.215e-14	***
bc(Mean_WD850, period = 360, nknots = 4)	3	234599	25171	15.7602	3.530e-10	***
bc(Mean_WD500, period = 360, nknots = 4)	3	231979	25131	2.6471	0.047411	*
ns(WS850mb, 3)	3	232280	25135	4.1553	0.006004	**
ns(WS500mb, 3)	3	236810	25204	26.8289	< 2.2e-16	***
ns(Ht850mb, 3)	3	233597	25155	10.7490	4.979e-07	***
ns(Ht500mb, 3)	3	231695	25126	1.2257	0.298695	
ns(GB_Li_BPsurf, 3)	3	234060	25162	13.0634	1.756e-08	***
ns(GB_Dt_BPsurf, 3)	3	234835	25174	16.9429	6.352e-11	***
ns(jday, 3)	3	235199	25180	18.7629	4.532e-12	***
as.factor(X1st.Max.Hour)	16	255532	25449	22.5983	< 2.2e-16	***
dowf	6	235212	25174	9.4143	2.825e-10	***
sumemissions	1	231519	25128	1.0441	0.306951	
ns(MultiNOx, 3)	3	238362	25227	34.5950	< 2.2e-16	***
ns(MultiCO, 3)	3	233807	25159	11.7995	1.092e-07	***

Most important variance: Sheboygan

1. Tmax
2. SR_max
3. MeanRH
4. WS.local
5. WD.local

Apply the GAM analysis in the LADCO region(Sheboygan and SWFP)





= Trends due only to meteorology

Refine the GAM analysis

- Add quantile regression for meteorology variables with ozone observations

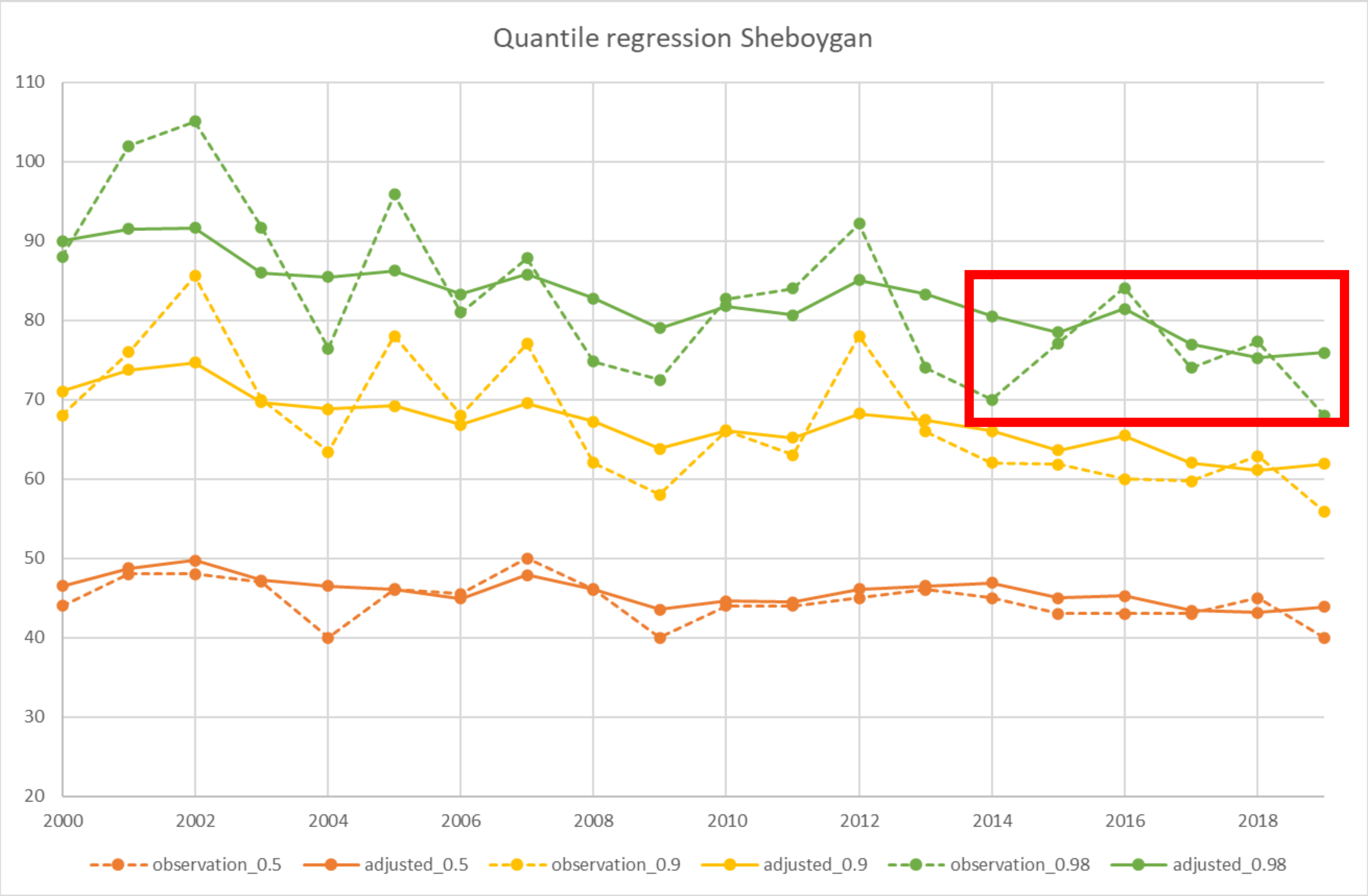
Model Qualitya,b,c

	q=0.5	q=0.9	q=0.98
Pseudo R Squared	.296	.288	.305
Mean Absolute Error (MAE)	.1899	.3107	.4433

a Dependent Variable: logm8max
b Model: (Intercept), tmax , tmin , yrf , MeanRH , LM_surf_T, WS.local
c Method: Interior Point non-linear optimization

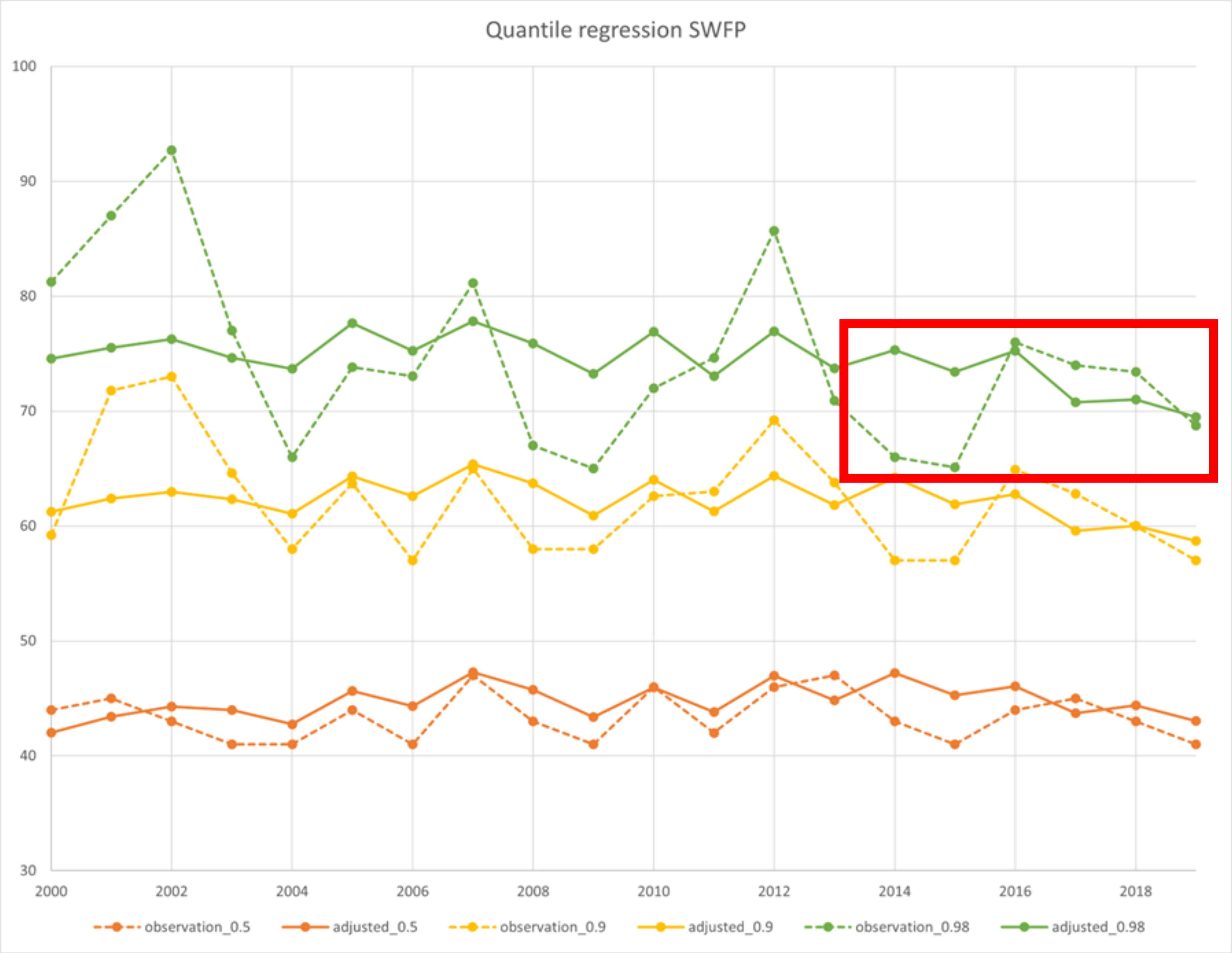
Our model uses no linear optimization and supports different quantiles and are able to change the weather conditions in the regression if needed.

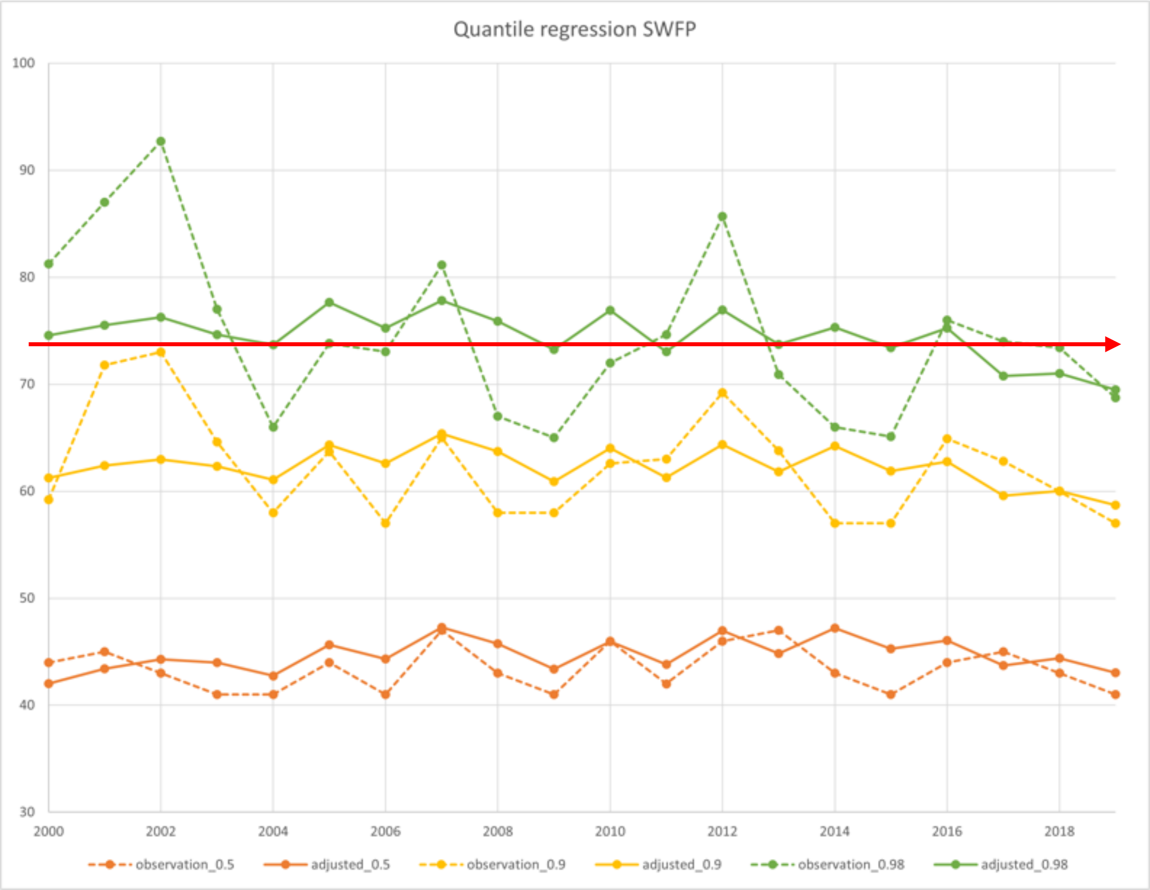
Apply the quantile regression to Sheboygan region

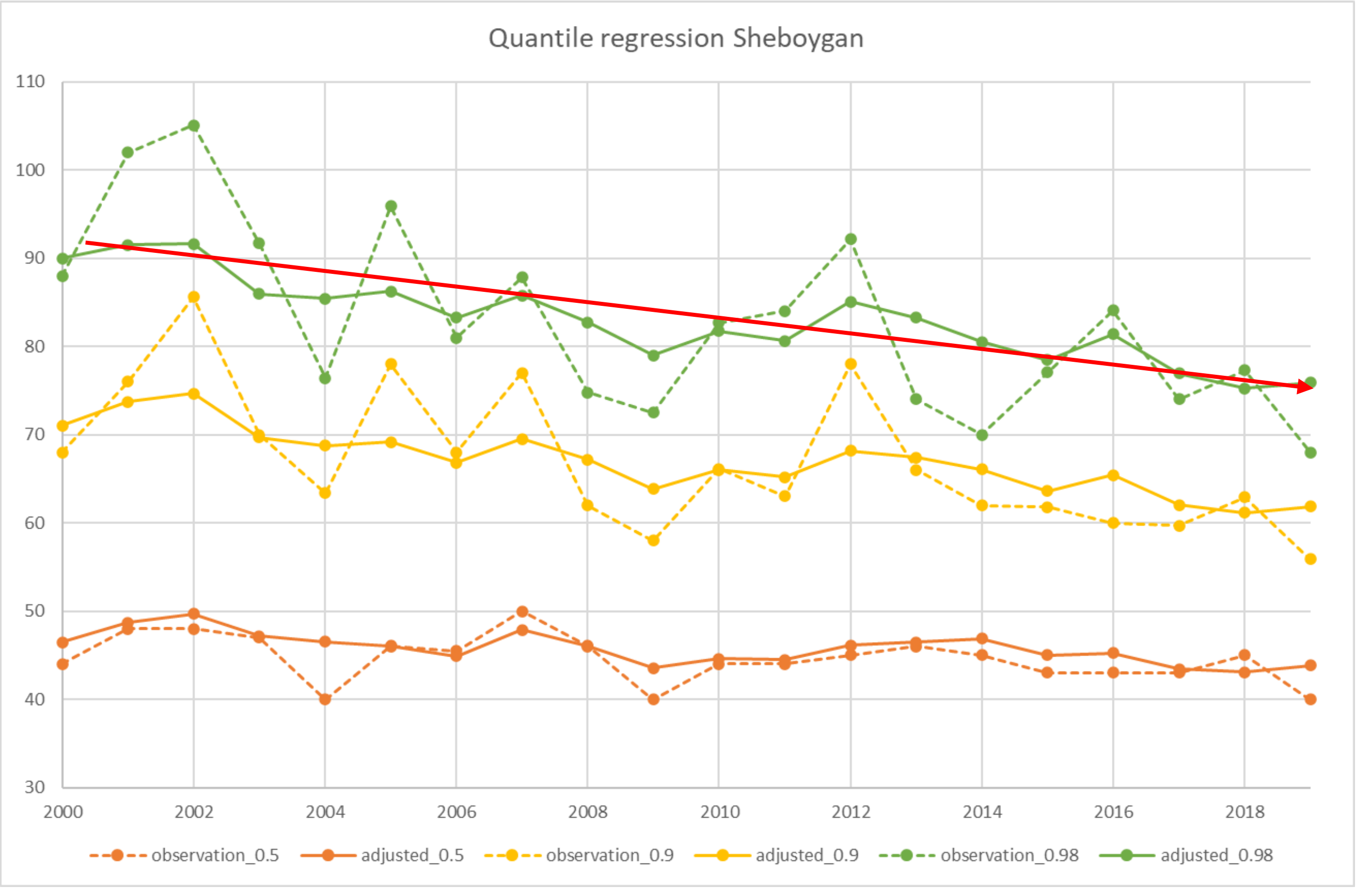


The images show the change in ozone concentration after excluding the effect of meteorological changes. After adjusting by meteorological variables, the fluctuation of ozone concentration becomes smaller. The effect of adjusting by meteorological variables can be seen by quantile regression for peak MDA8 ozone concentrations (0.9, 0.98).

Apply the quantile regression to SWFP







90th and 98th percentile ozone is flatter at SWFP but has decreased a lot at Sheboygan
Suggests ozone concentrations are continuing to decrease at Sheboygan much more than at SWFP (this matches other LADCO analyses)

Overall much less interannual variability at the 50th percentile level

Discussion

- The average ozone concentration trends after adjustment of meteorological conditions can help to understand the influence of meteorological conditions on the average ozone observations and to reduce the fluctuating interference of meteorological conditions when study the influence of precursors on ozone.
- Trends in peak MDA8 Ozone concentrations are adjusted by implementing quantile regression methods. These trends can help air quality modelers understand the overall impact of meteorological conditions that contribute to peak Ozone levels in a given year.
- Observed much greater reductions in 90th and 98th percentile meteorologically adjusted ozone at the Sheboygan site than at SWFP