List of Tables

List of Figures

S1. Geostatistical model validation

To validate and visualize the performance of spatial mappings for ozone extreme values, we use the 90th percentile for JJA (abbreviated as JJA90) at each individual station in 2023 as a demonstration (the upper panel of Fig S2), and different approaches are compared under the framework of generalized additive models (GAM), including:

- GP (Gaussian process) for block extremes (Wood, 2006): the JJA90 value is calculated at each station, and then the GP is used in the model fit.
- GEV (generalized extreme value location-scale model) link for block extremes (Wood and Fasiolo, 2017): the JJA90 value is calculated at each station, and then the GEV is used in the model fit.
- Quantile GAM (Fasiolo et al., 2020): the seasonal percentile mapping is directly derived from all daily values.

In terms of the computational costs, fast computations can be achieved for the GP, followed by the GEV (these two approaches deal with the block extremes at each station, e.g., only a sample size $n \sim 1200$ stations needs to be considered in 2023), and the quantile GAM is the most computationally intensive (this approach takes all available MDA8 values in JJA 2023 into account). One important technical detail is that the roughness parameter needs to be specified in advance to ensure fair comparisons between different approaches (for determining the degree of smoothing in the interpolated surface: The higher the parameter, the smoother the interpolated surface). The default setting for the GAM implementation adopts a measure to avoid overfitting and thus penalize smaller local features (Chang et al., 2020). However, surface ozone tends to be highly variable in time and space (e.g., across urban, suburban and rural areas), and the measurements are often of high quality with a high sampling frequency, therefore the local features and spatial gradients in Fig S2(a1) should be considered to be mostly real, and the roughness parameter is adjusted to allow more smaller scale variability (see the lower panel of Fig S2 for a demonstration): 5% of the default roughness value estimated through the generalized cross validation criterion is used, the results are less sensitive to a further reduction of the roughness parameter.

The interpolated surfaces in Fig S2 are well representative of the observed locations and the general patterns are fairly consistent, but some differences at (unsampled) spatial gaps can be observed. To quantitatively evaluate the fitted quality, we use root-mean-square error (RMSE) to compare different approaches:

$$\text{RMSE} = \left(\frac{\sum_{s=1}^{n} (\hat{y}_s - y_s)^2}{n} \right)^{1/2}, \text{ for the units of ppbv,}$$

where y_s is the ozone value at a site s (with a total of n sites), and $\hat{y}_s$ is the fitted value of y_s. Another useful measure is the leave-one-out generalized cross validation (GCV) score, which can be understood as the overall prediction error when each site value is omitted and predicted from the remaining sites, and can be approximated by (Golub et al., 1979):

$$\text{GCV} = \left(\frac{\sum_{s=1}^{n} (\hat{y}_s - y_s)^2}{n - EDF} \right)^{1/2}, \text{ for the units of ppbv,}$$

where EDF is the effective degree of freedom (a quantity to represent model complexity, i.e. a higher EDF implies a more complex model). Overall, the GP has the best performance, followed by quantile GAM and GEV. It should be noted that the GP is by no means claimed to be superior to the GEV in terms of extreme values interpolation. However, the GP generally has a greater numerical stability and consistency (i.e. a single configuration can achieve reasonable results). In contrast, the GEV relies more heavily on customization (parameters tuning for different percentiles, seasons and years) to optimize the results. In this analysis we use the same roughness parameter for GP, GEV and quantile GAM for a fair comparison, albeit the optimized roughness parameter for GP might not be completely suitable for GEV or quantile GAM. This issue warrants further detailed investigation, but it is beyond our scope. In summary, in our analysis the GP provides a fast and robust approximation of ozone fields, and is used in our regional analysis.

S2. Correlation between two probabilistic events

Based on the notations in Section 2.3, the correlation between P_E and P_H can be expressed as

$$\rho = \frac{\mathrm{Cov}(E,H)}{\sqrt{\mathrm{Var}(E)\mathrm{Var}(H)}} = \frac{P_{E\cap H} - P_E P_H}{\sqrt{P_E(1-P_E)P_H(1-P_H)}}.$$

Hence, if $P_{E|H} > P_E$, then

$$\begin{aligned}&\frac{P_{E\cap H}}{P_H} > P_E,\\ \Rightarrow &P_{E\cap H} > P_E P_H,\\ \Rightarrow &\rho > 0.\end{aligned}$$

S3. Exceedance trends in normal and heatwave conditions

Figs S12-S14 show ozone exceedance trends in normal and heatwave conditions, based on the TX90pct, TX95pct and TX35deg metrics, respectively. We can see a contrast between more reliable decreasing trends in normal days and less reliable decreasing trends in heatwave days, especially for the threshold of 35 ppbv where consistent negative trends are largely vanished. Site percentages of trends categorized by different reliability scales and different heatwave metrics are provided in Table S2. This analysis can be summarized as follows: When all daily data in May-Sep are considered (Fig 6), 82.4%, 83.8%, 78.4%, and 55.7% of sites show reliably negative trends at 70, 60, 50, and 35 ppbv exceedances, respectively. If we exclude heatwave observations, these percentages are generally similar. Above percentages are substantially reduced to 55.9%, 64.7%, 60.2%, and 18.5%, respectively, if the TX90pct heatwave condition is considered. These percentages are further reduced to 44.6%, 57.2%, 49.3%, and 11.9%, if a stricter TX95pct heatwave condition is considered. Even though the TX35deg metric is merely provided for a reference, it identifies which sites have consistently reached the high temperature scenario. By comparing the common sites with other metrics, trend patterns are similar in general.

Table S1: Site percentages of ozone accumulation rates, by different regions, different reliability scales, and different heatwave metrics (as in Fig 7, trends with $p \leq 0.01$ are merged into $p \leq 0.05$). Note that for each row the relative percentages are shown (i.e., sum to 100%), but much fewer sites are available for the analysis of the TX35deg metric, so the results for TX35deg metric are not comparable to the TX90pct and TX95pct metrics.

Metric	SNR≥2	2>SNR≥1	|SNR| <1	-2<SNR≤-1	SNR≤-2	Confidence level
	$p \leq 0.05$	$0.33 \leq p < 0.05$	$p < 0.33$	$0.33 \leq p < 0.05$	$p \leq 0.05$	
Western USA						
TX90pct	30.5	18.4	35.1	8.0	8.0	Medium agreement & limited evidence
TX95pct	29.5	16.8	32.9	11.6	9.3	Low agreement & limited evidence
TX35eg	19.0	12.0	20.0	16.0	33.0	Medium agreement & robust evidence
Eastern USA						
TX90pct	10.5	10.3	46.6	20.3	12.3	Medium agreement & limited evidence
TX95pct	9.2	10.3	52.7	18.9	8.9	High agreement & limited evidence
TX35eg	40.1	9.9	24.7	17.3	8.0	Medium agreement & robust evidence
Conterminous USA						
TX90pct	16.2	12.6	43.3	16.8	11.1	Medium agreement & limited evidence
TX95pct	14.9	12.1	47.1	16.9	9.0	Medium agreement & limited evidence
TX35eg	32.0	10.7	22.9	16.8	17.5	Low agreement & robust evidence

Table S2: Site percentages of exceedance probability trends (as in Figs 6 and S12-S14), by different reliability scales and different heatwave metrics (trends with $p \leq 0.01$ are merged into $p \leq 0.05$): The first part is based on all observations. The second and third parts are based on observations in heatwave and normal conditions, partitioned by the TX90pct, TX95pct, and TX35deg metrics, respectively. Note that for each row the relative percentages are shown (i.e., sum to 100%), but much fewer sites are available for the analysis of the TX35deg metric, so the results for TX35deg metric are not comparable to the TX90pct and TX95pct metrics. The only site with reliable increasing 70 ppbv exceedance trends ($p \leq 0.05$) is shown in Fig S4.

Threshold	Metric	SNR≥2	2>SNR≥1	|SNR| <1	-2<SNR≤-1	SNR≤-2	Confidence level
[ppbv]		$p \leq 0.05$	$0.33 \leq p < 0.05$	$p < 0.33$	$0.33 \leq p < 0.05$	$p \leq 0.05$	
All daily observations							
70	-	0.2	0.7	8.3	8.4	82.4	High agreement & robust evidence
60	-	0.8	2.5	7.1	5.8	83.8	High agreement & robust evidence
50	-	3.5	3.1	9.9	5.1	78.4	High agreement & robust evidence
35	-	7.1	6.3	19.3	11.6	55.7	High agreement & robust evidence
Heatwave observations							
70	TX90pct	0.2	0.8	23.9	19.3	55.9	High agreement & robust evidence
60	TX90pct	0.4	1.8	17.8	15.4	64.7	High agreement & robust evidence
50	TX90pct	0.8	2.0	22.6	14.4	60.2	High agreement & robust evidence
35	TX90pct	0.7	3.9	48.0	29.0	18.5	Medium agreement & limited evidence
70	TX95pct	0	0.7	29.5	25.2	44.6	Medium agreement & robust evidence
60	TX95pct	0	1.5	22.0	19.3	57.2	High agreement & robust evidence
50	TX95pct	0.9	1.8	27.5	20.5	49.3	Medium agreement & robust evidence
35	TX95pct	0	5.7	56.9	25.5	11.9	High agreement & limited evidence
70	TX35deg	0	0.8	21.2	19.7	58.3	High agreement & robust evidence
60	TX35deg	0	1.9	22.3	16.9	58.8	High agreement & robust evidence
50	TX35deg	0.8	1.9	36.5	19.6	41.2	Medium agreement & robust evidence
35	TX35deg	2.7	6.2	65.0	11.9	14.2	High agreement & limited evidence
Normal observations							
70	TX90pct	0.3	0.8	7.8	7.8	83.2	High agreement & robust evidence
60	TX90pct	1.2	2.1	5.9	6.2	84.6	High agreement & robust evidence
50	TX90pct	3.6	2.9	7.2	5.7	80.6	High agreement & robust evidence
35	TX90pct	7.2	5.7	19.6	11.5	55.9	High agreement & robust evidence
70	TX95pct	0.4	0.7	7.2	8.4	83.4	High agreement & robust evidence
60	TX95pct	1.0	2.3	5.7	5.7	85.2	High agreement & robust evidence
50	TX95pct	3.3	3.2	8.1	5.7	79.7	High agreement & robust evidence
35	TX95pct	7.2	5.7	19.1	11.2	56.7	High agreement & robust evidence
70	TX35deg	0	0.8	4.9	10.6	83.7	High agreement & robust evidence
60	TX35deg	0.4	3.1	6.2	8.1	82.3	High agreement & robust evidence
50	TX35deg	4.7	3.1	11.5	5.8	75.0	High agreement & robust evidence
35	TX35deg	5.7	6.5	21.9	10.0	55.8	High agreement & robust evidence

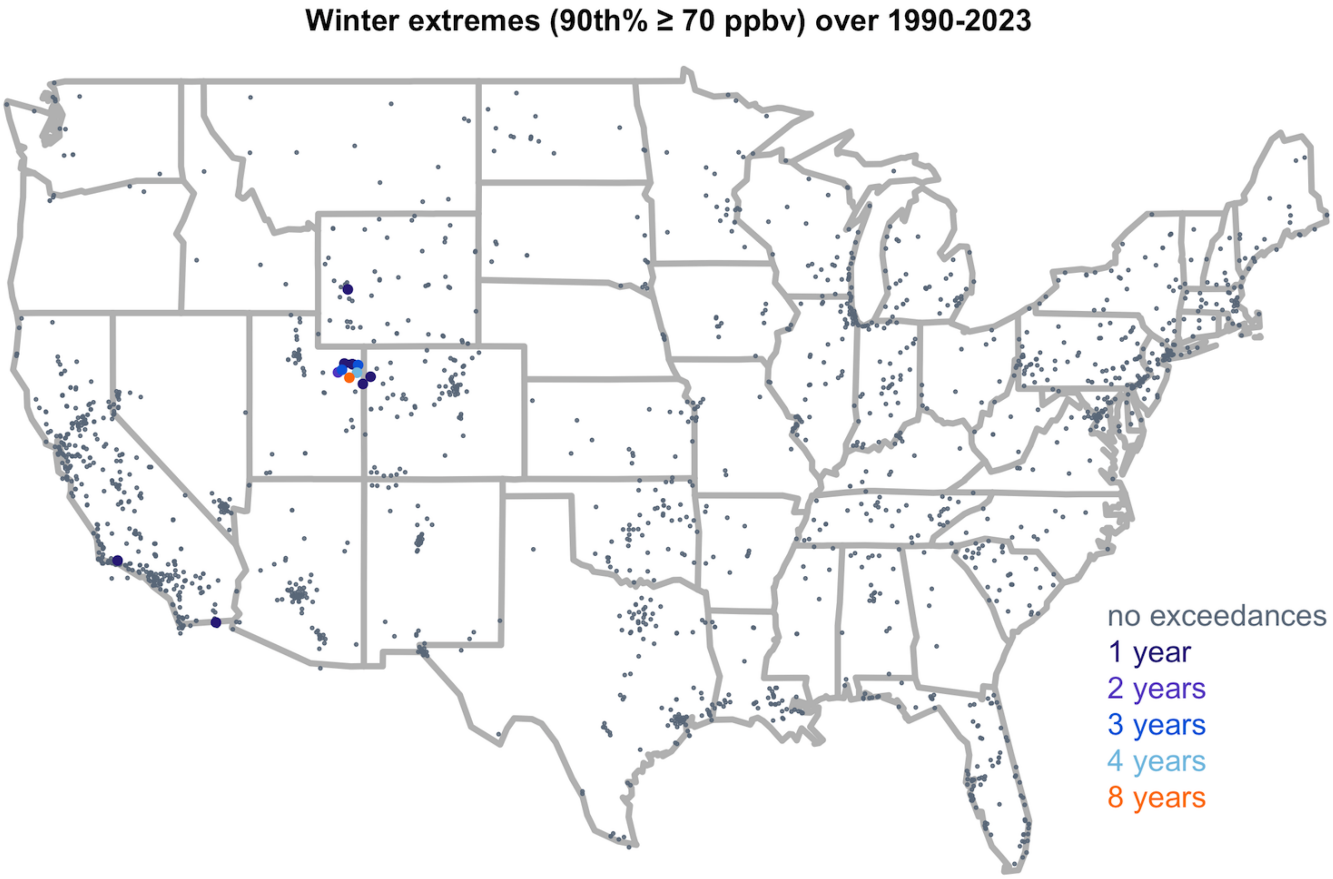


Figure S1: Locations of winter ozone exceedances, colored by the frequency of the seasonal 90th percentile exceeds 70 ppbv over 1995-2022. Two records are observed in California in the 1990s. After 2008, ozone exceedances are mainly observed in the snow-covered oil and gas basin of northeastern Utah.

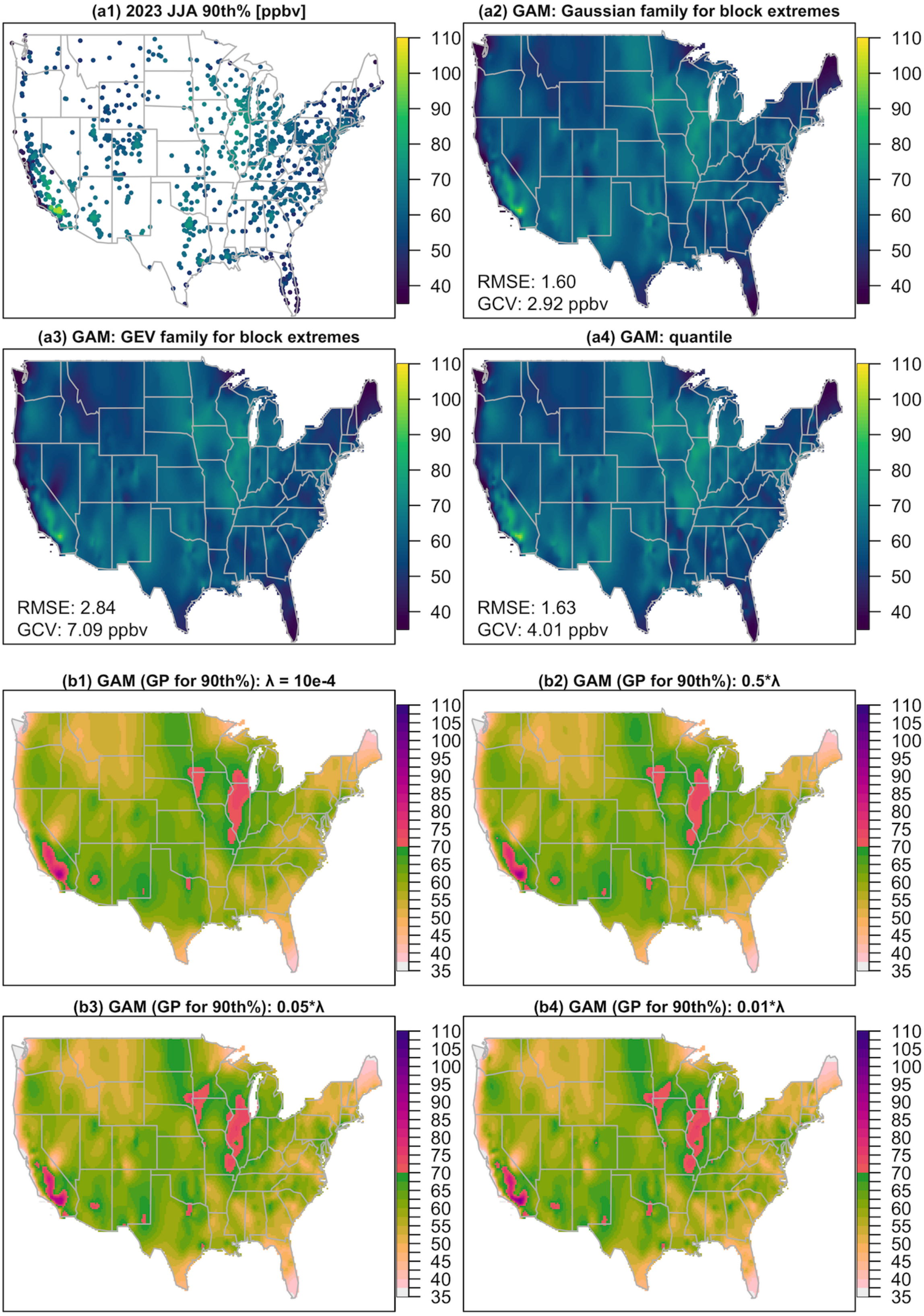

(a1) 2023 JJA 90th% [ppbv]
(a2) GAM: Gaussian family for block extremes
RMSE: 1.60
GCV: 2.92 ppbv
(a3) GAM: GEV family for block extremes
RMSE: 2.84
GCV: 7.09 ppbv
(a4) GAM: quantile
RMSE: 1.63
GCV: 4.01 ppbv
110
100
90
80
70
60
50
40
(b1) GAM (GP for 90th%): λ = 10e-4
(b2) GAM (GP for 90th%): 0.5*λ
(b3) GAM (GP for 90th%): 0.05*λ
(b4) GAM (GP for 90th%): 0.01*λ
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35

Figure S2: The upper panel shows observed values and spatially interpolated surfaces of the seasonal 90th percentile MDA8 ozone in JJA 2023 from different approaches (see Section S1 for details). The lower panel shows the impact of the roughness parameter λ on GP interpolations: the initial input $\lambda = 10^{-4}$ is generated by the generalized cross validation criterion. The higher the roughness parameter, the smoother the resulting surface.

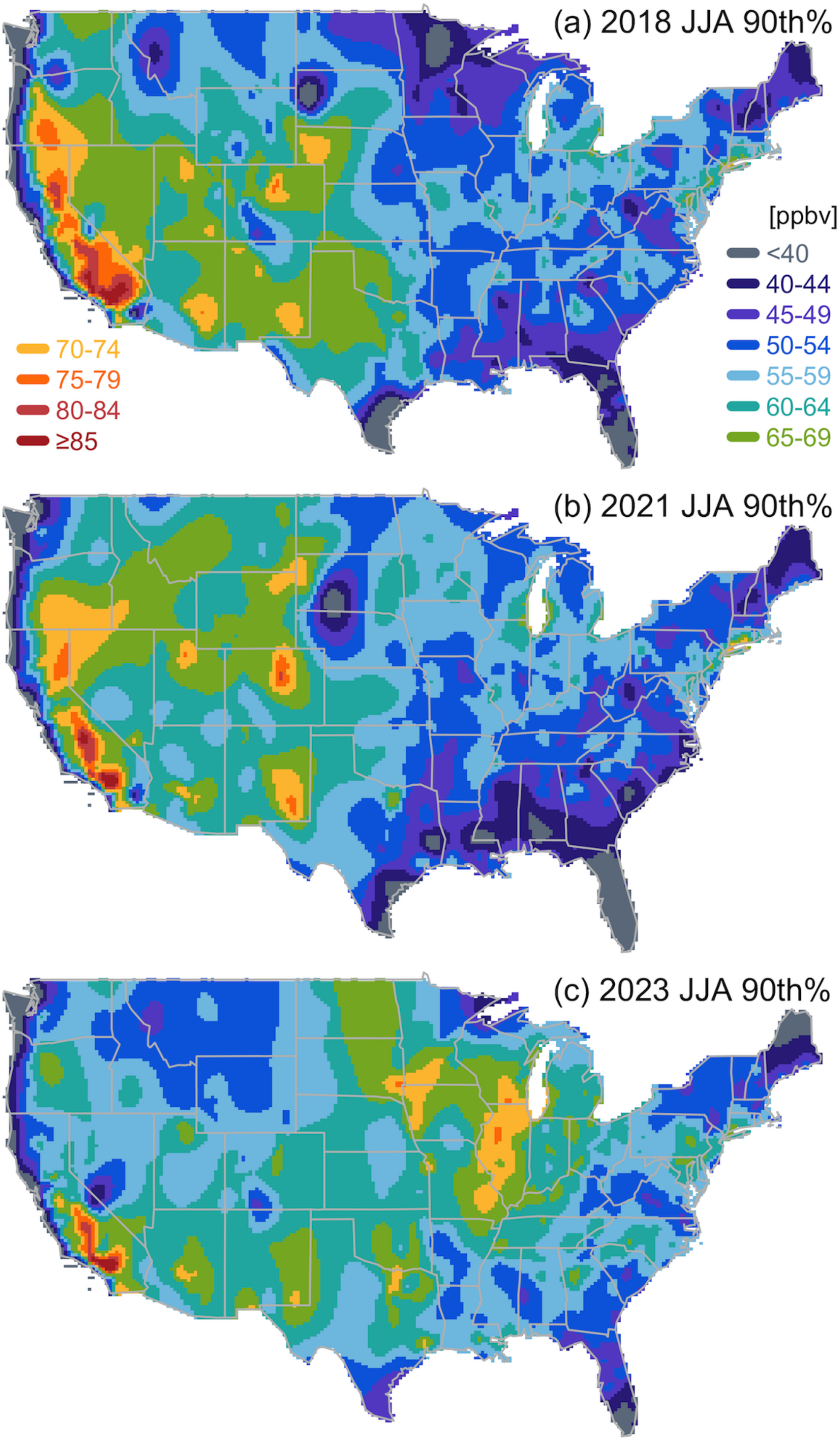


Figure S3: Spatially interpolated surfaces of the seasonal 90th percentile MDA8 ozone in JJA 2018, 2021, and 2023.

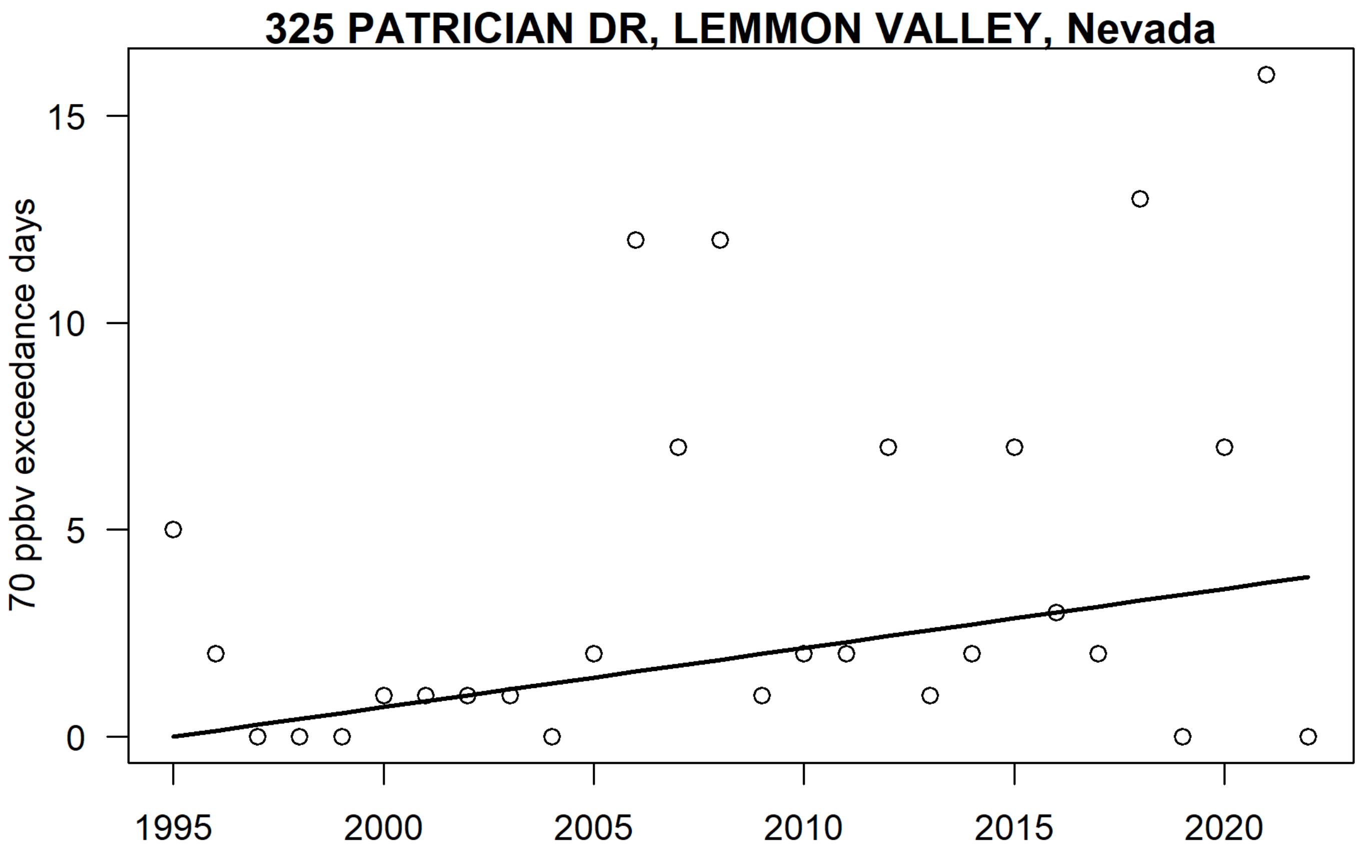
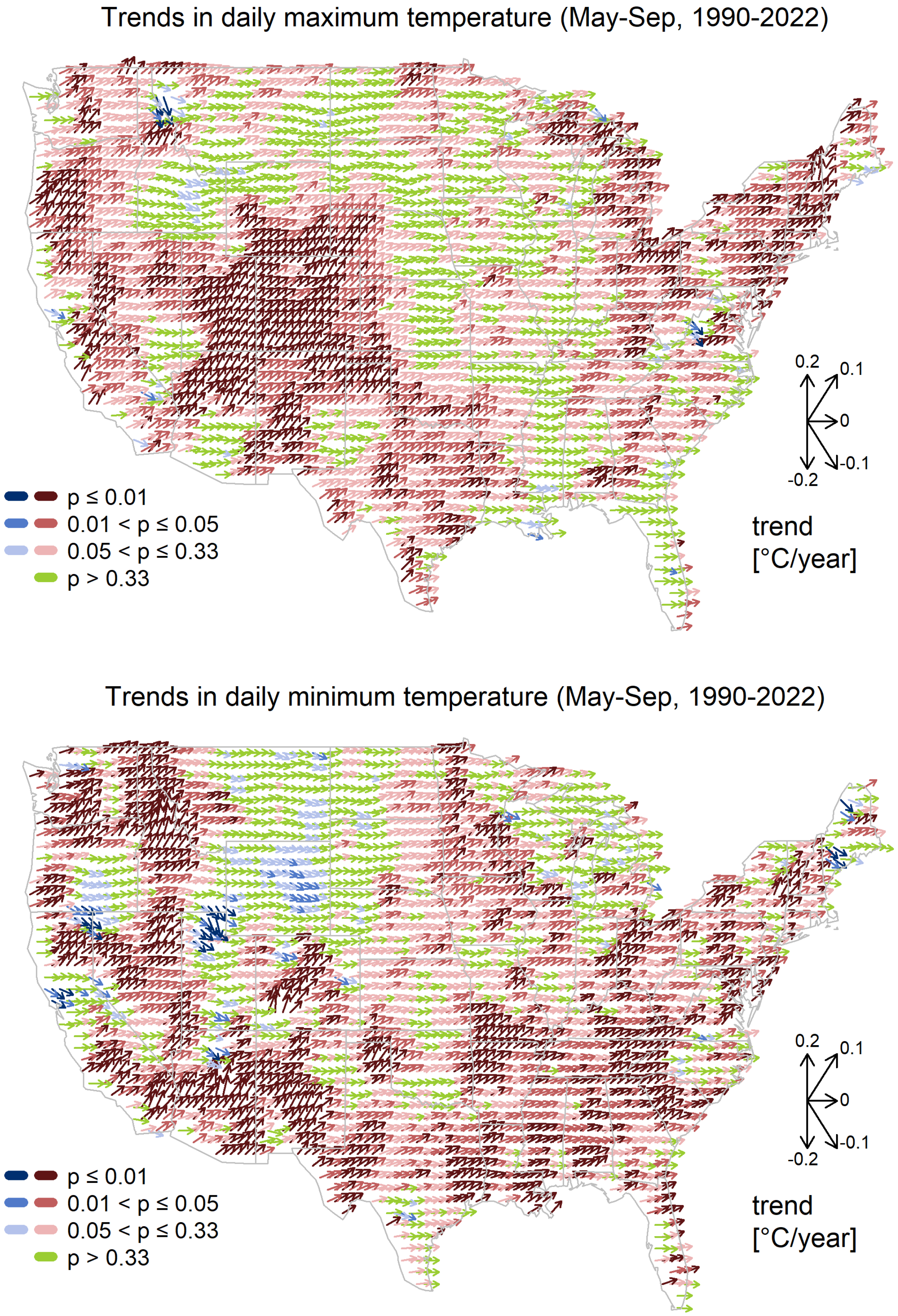


Figure S4: The only site with reliable increasing 70 ppbv exceedance trends ($p \leq 0.05$) over 1995-2022.


Trends in daily maximum temperature (May-Sep, 1990-2022)

Trends in daily minimum temperature (May-Sep, 1990-2022)

Figure S5: Trends in daily maximum and minimum temperatures based on the gridded dataset (May-Sep, 1990-2022).

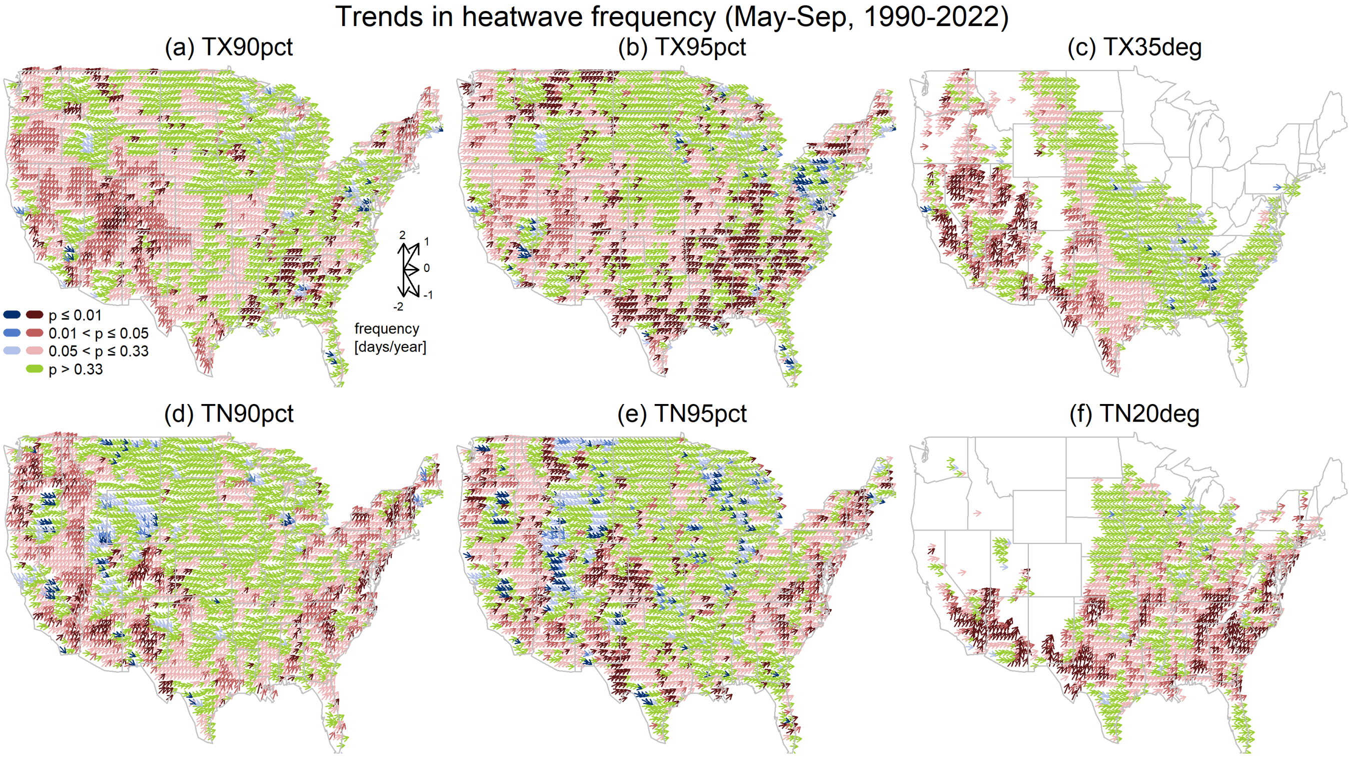
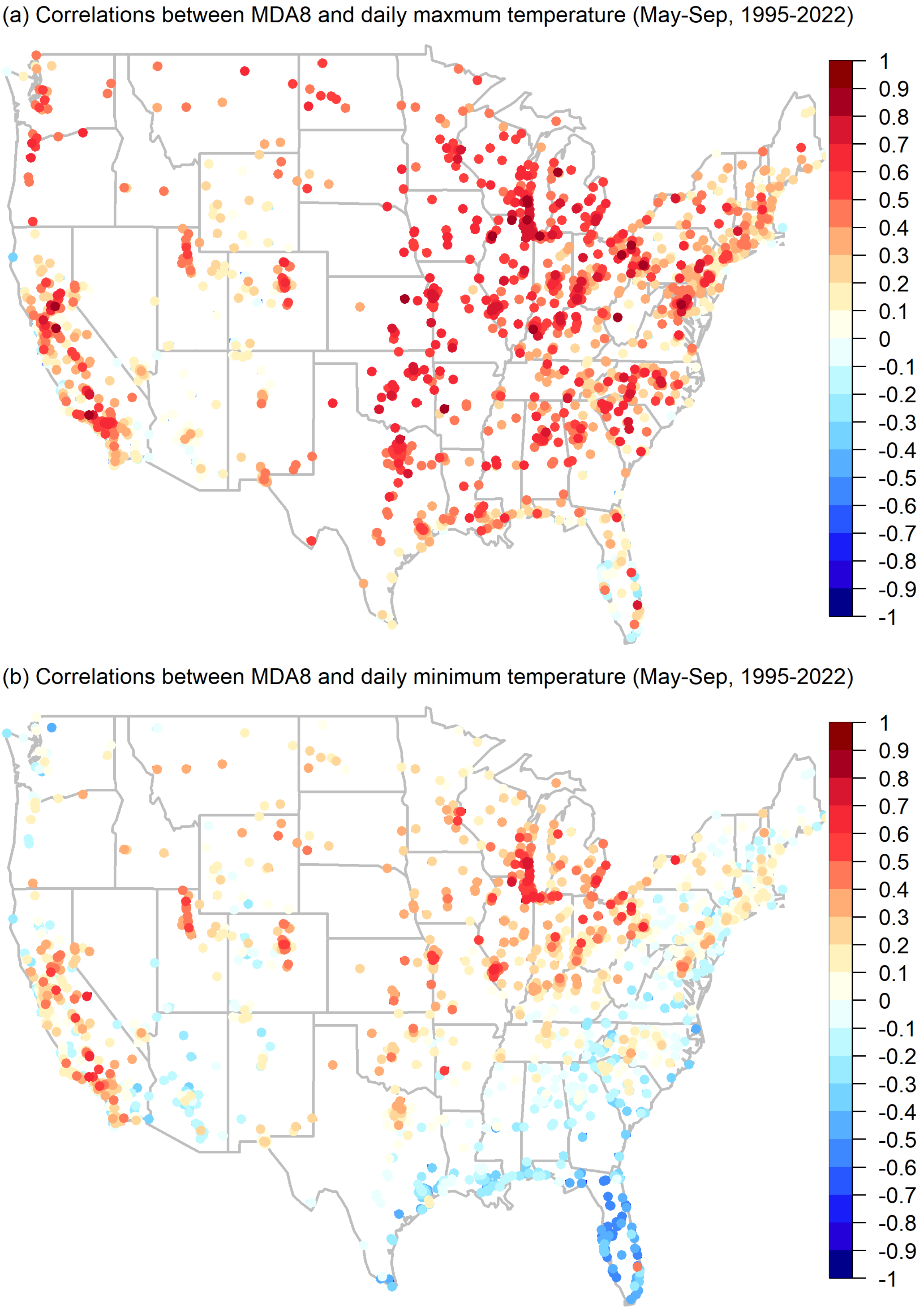


Figure S6: Trends in heatwave frequency [days/year] based on various heatwave metrics (May-Sep, 1990-2022). Due to stricter criteria for TX35deg and TN20deg, trends are not estimated if heatwaves occur less than 10 years between 1990-2022.



(a) Correlations between MDA8 and daily maxmum temperature (May-Sep, 1995-2022)

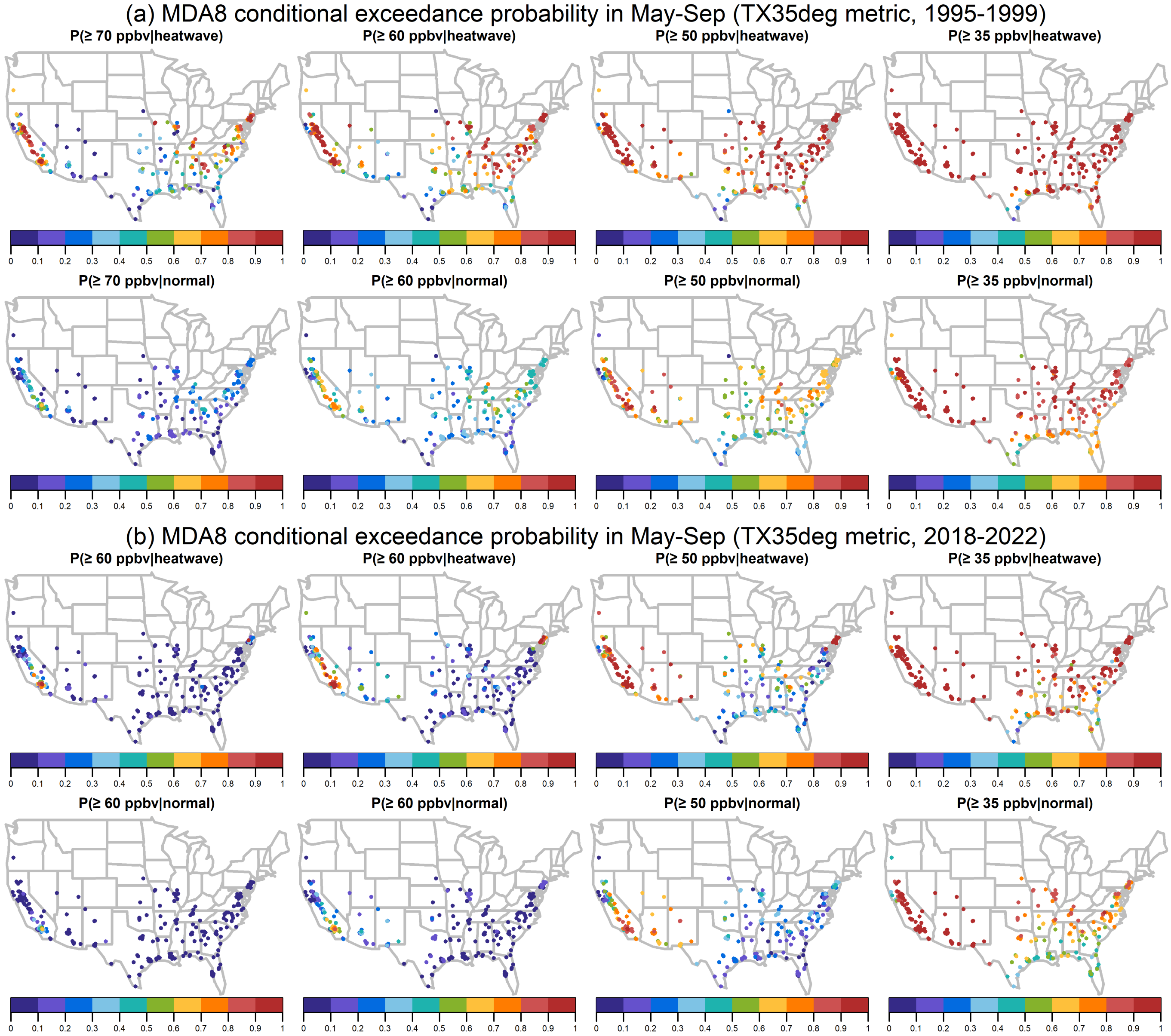
(b) Correlations between MDA8 and daily minimum temperature (May-Sep, 1995-2022)

Figure S7: Correlations between MDA8 and daily maximum/minimum temperature (May-Sep, 1990-2022).

(a) MDA8 conditional exceedance probability in May-Sep (TX95pct metric, 1995-1999)

(b) MDA8 conditional exceedance probability in May-Sep (TX95pct metric, 2018-2022)

Figure S8: Same as Fig 8, but for the TX95pct metric.


(a) MDA8 conditional exceedance probability in May-Sep (TX35deg metric, 1995-1999)

(b) MDA8 conditional exceedance probability in May-Sep (TX35deg metric, 2018-2022)

Figure S9: Same as Fig 8, but for the TX35deg metric.

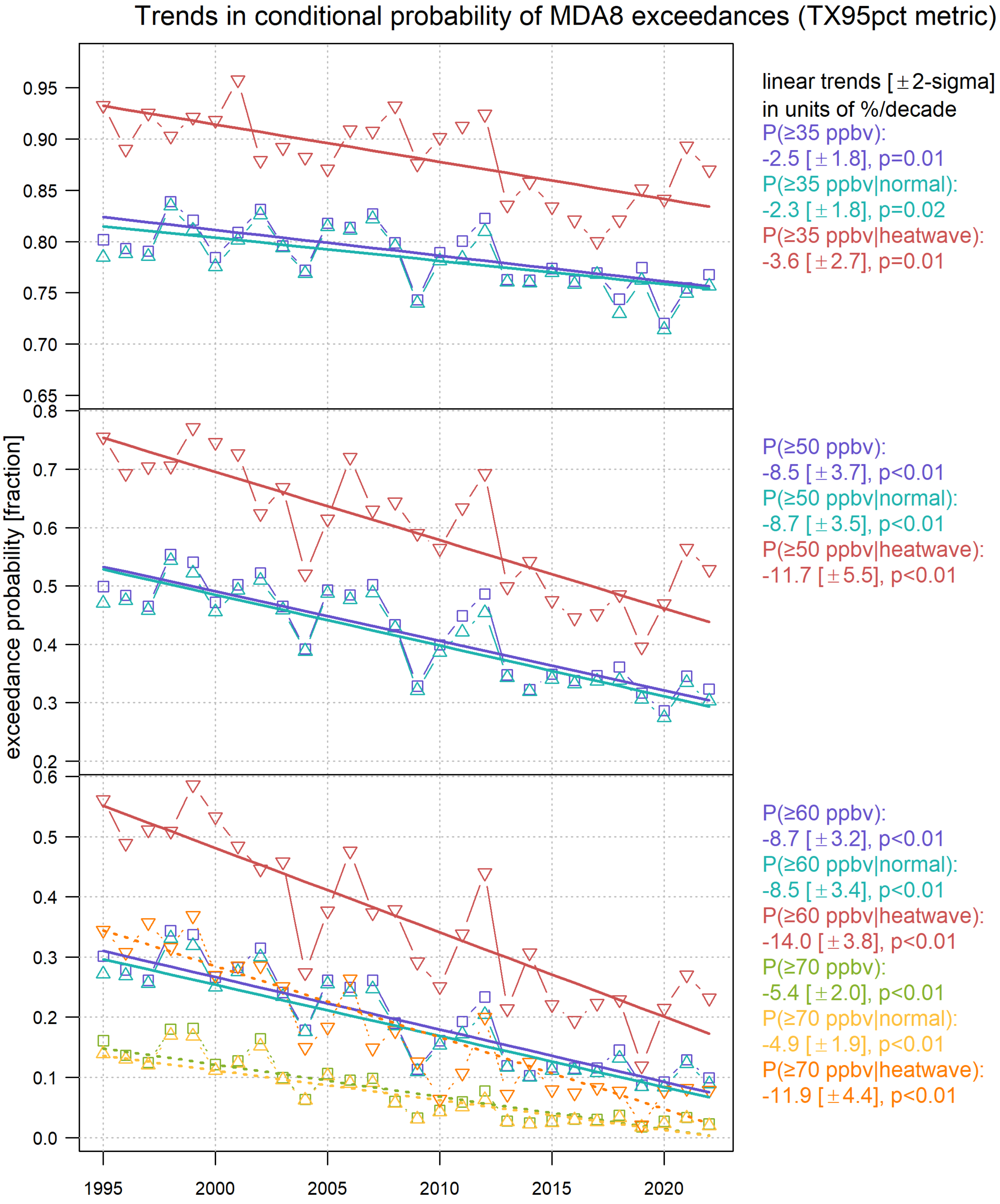


Figure S10: Same as Fig 9, but for the TX95pct metric.

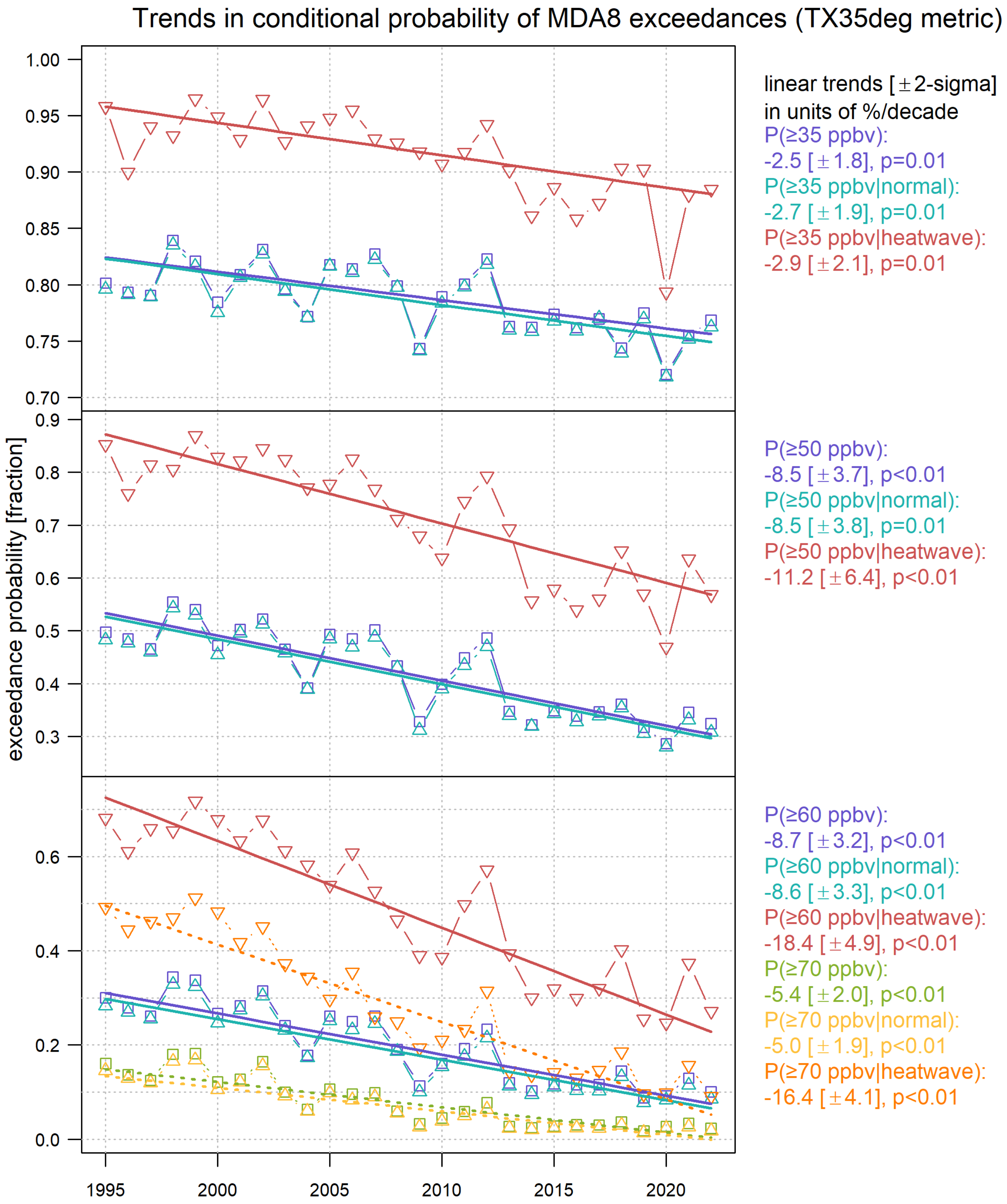


Figure S11: Same as Fig 9, but for the TX35deg metric.

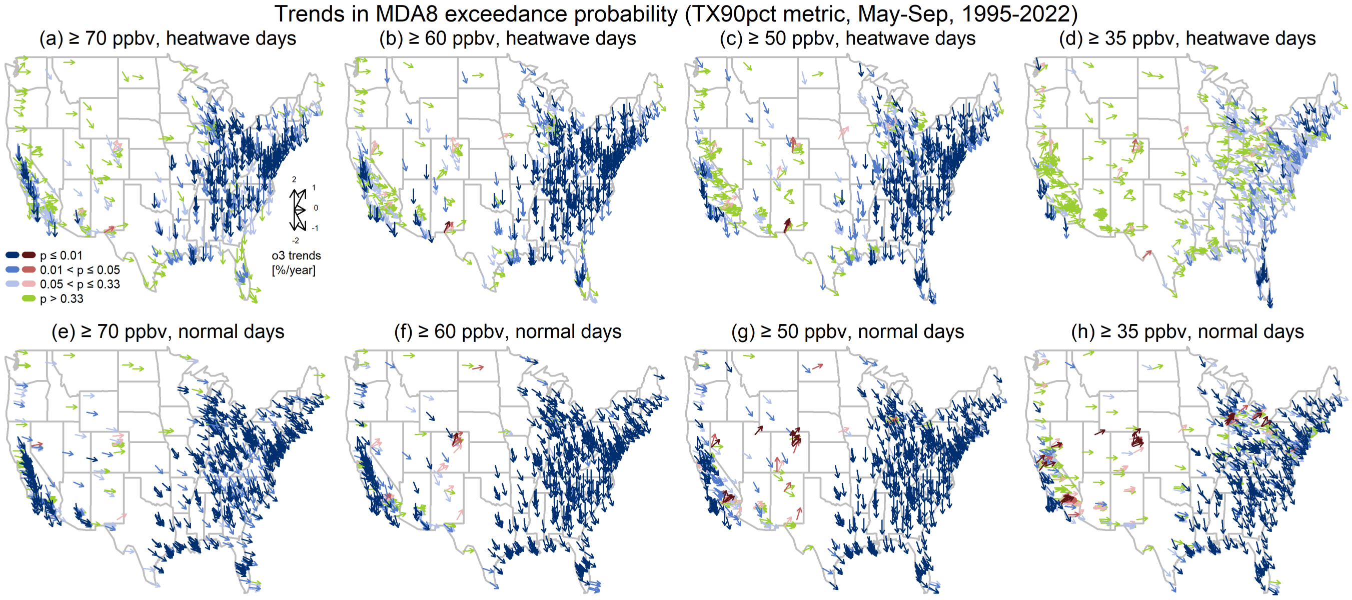

Trends in MDA8 exceedance probability (TX90pct metric, May-Sep, 1995-2022)
(a) ≥ 70 ppbv, heatwave days
(b) ≥ 60 ppbv, heatwave days
(c) ≥ 50 ppbv, heatwave days
(d) ≥ 35 ppbv, heatwave days
2
1
0
-1
-2
o3 trends
[%/year]
p ≤ 0.01
0.01 < p ≤ 0.05
0.05 < p ≤ 0.33
p > 0.33
(e) ≥ 70 ppbv, normal days
(f) ≥ 60 ppbv, normal days
(g) ≥ 50 ppbv, normal days
(h) ≥ 35 ppbv, normal days

Figure S12: Trends in ozone exceedances in heatwave and normal days, based on various ozone thresholds and the TX90pct metric (May-Sep, 1995-2022).

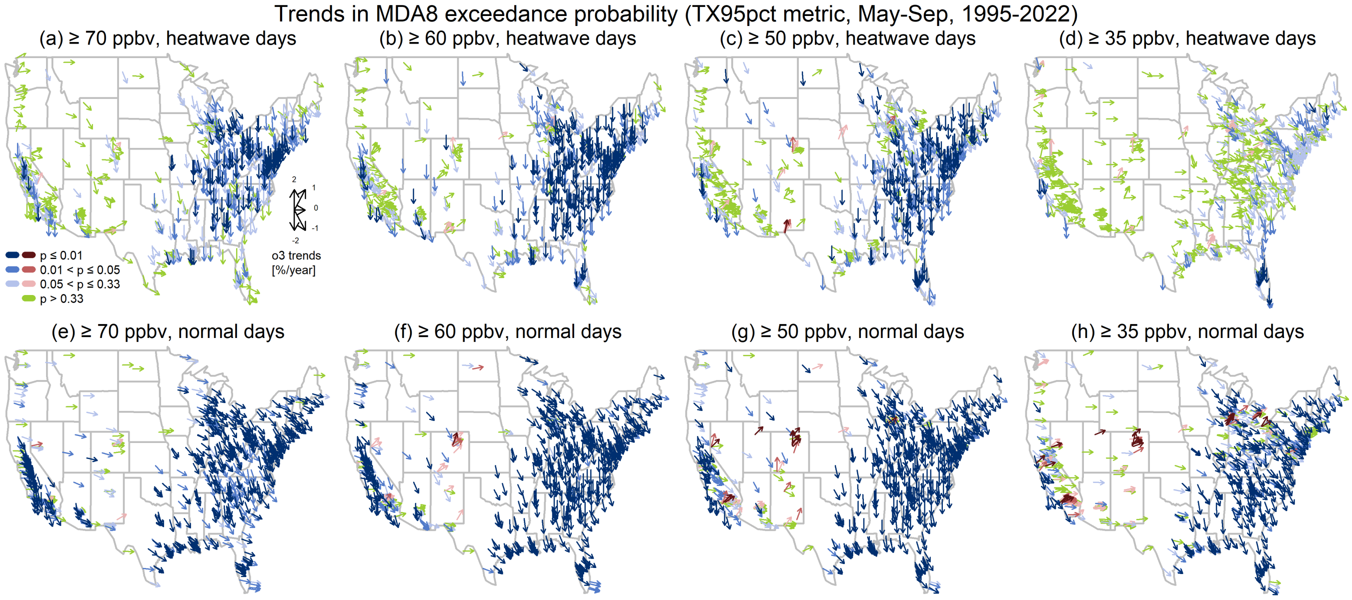

Trends in MDA8 exceedance probability (TX95pct metric, May-Sep, 1995-2022)
(a) ≥ 70 ppbv, heatwave days
(b) ≥ 60 ppbv, heatwave days
(c) ≥ 50 ppbv, heatwave days
(d) ≥ 35 ppbv, heatwave days
2
1
0
-1
-2
o3 trends
[%/year]
p ≤ 0.01
0.01 < p ≤ 0.05
0.05 < p ≤ 0.33
p > 0.33
(e) ≥ 70 ppbv, normal days
(f) ≥ 60 ppbv, normal days
(g) ≥ 50 ppbv, normal days
(h) ≥ 35 ppbv, normal days

Figure S13: Same as Fig S12, but for the TX95pct metric.

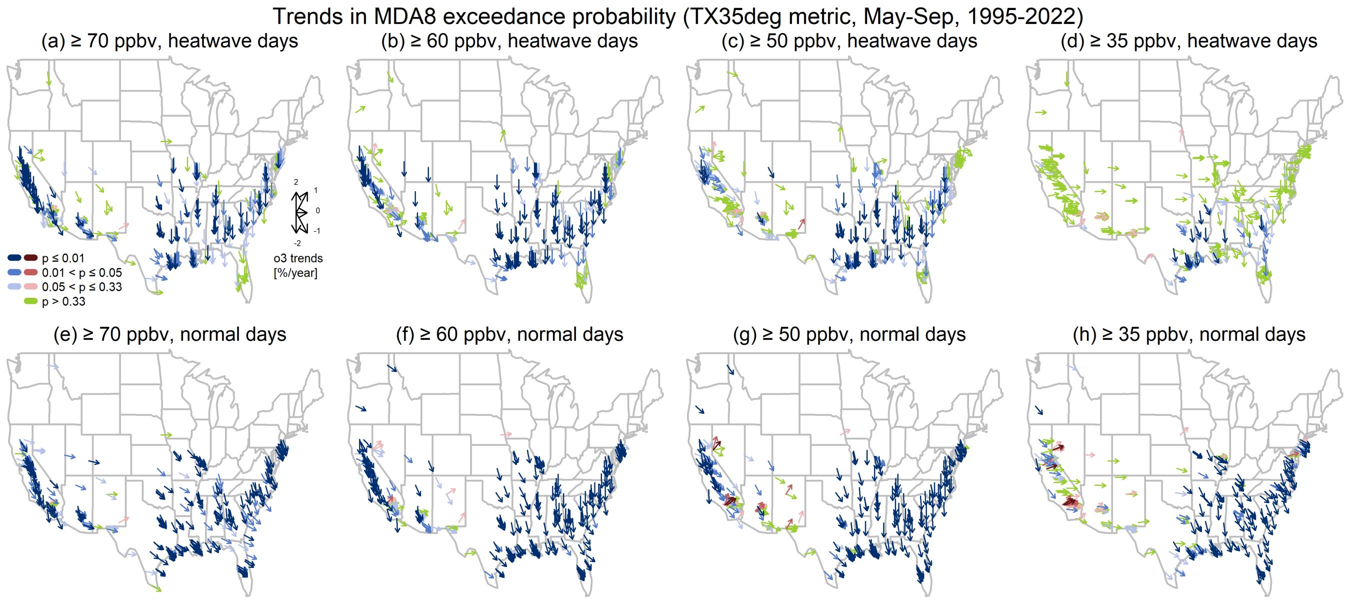

Trends in MDA8 exceedance probability (TX35deg metric, May-Sep, 1995-2022)
(a) ≥ 70 ppbv, heatwave days
(b) ≥ 60 ppbv, heatwave days
(c) ≥ 50 ppbv, heatwave days
(d) ≥ 35 ppbv, heatwave days
(e) ≥ 70 ppbv, normal days
(f) ≥ 60 ppbv, normal days
(g) ≥ 50 ppbv, normal days
(h) ≥ 35 ppbv, normal days
p ≤ 0.01
0.01 < p ≤ 0.05
0.05 < p ≤ 0.33
p > 0.33
2
1
0
-1
-2
o3 trends
[%/year]

Figure S14: Same as Fig S12, but for the TX35deg metric.

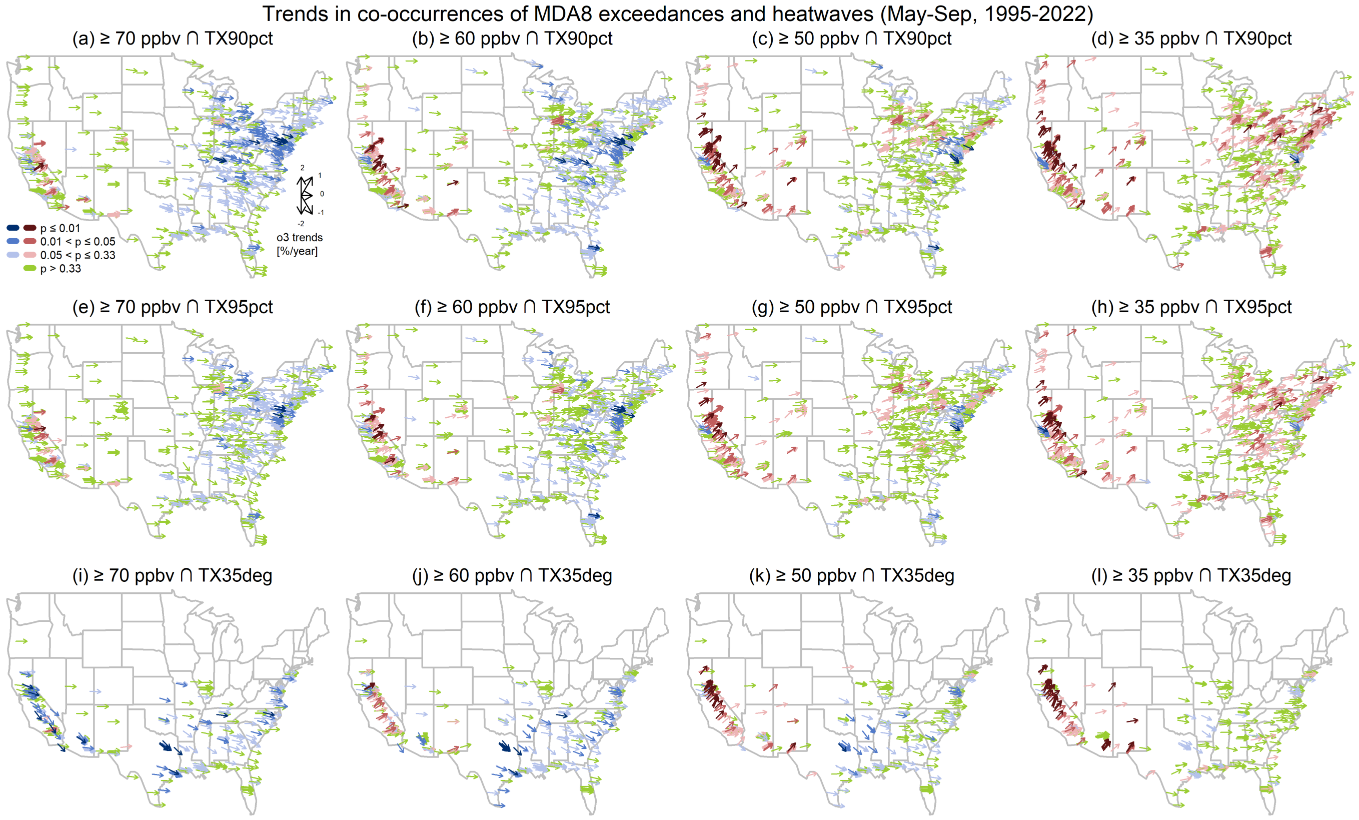


Figure S15: Trends in co-occurrences of ozone exceedances and heatwave events, based on various ozone thresholds and heatwave metrics (May-Sep, 1995-2022).

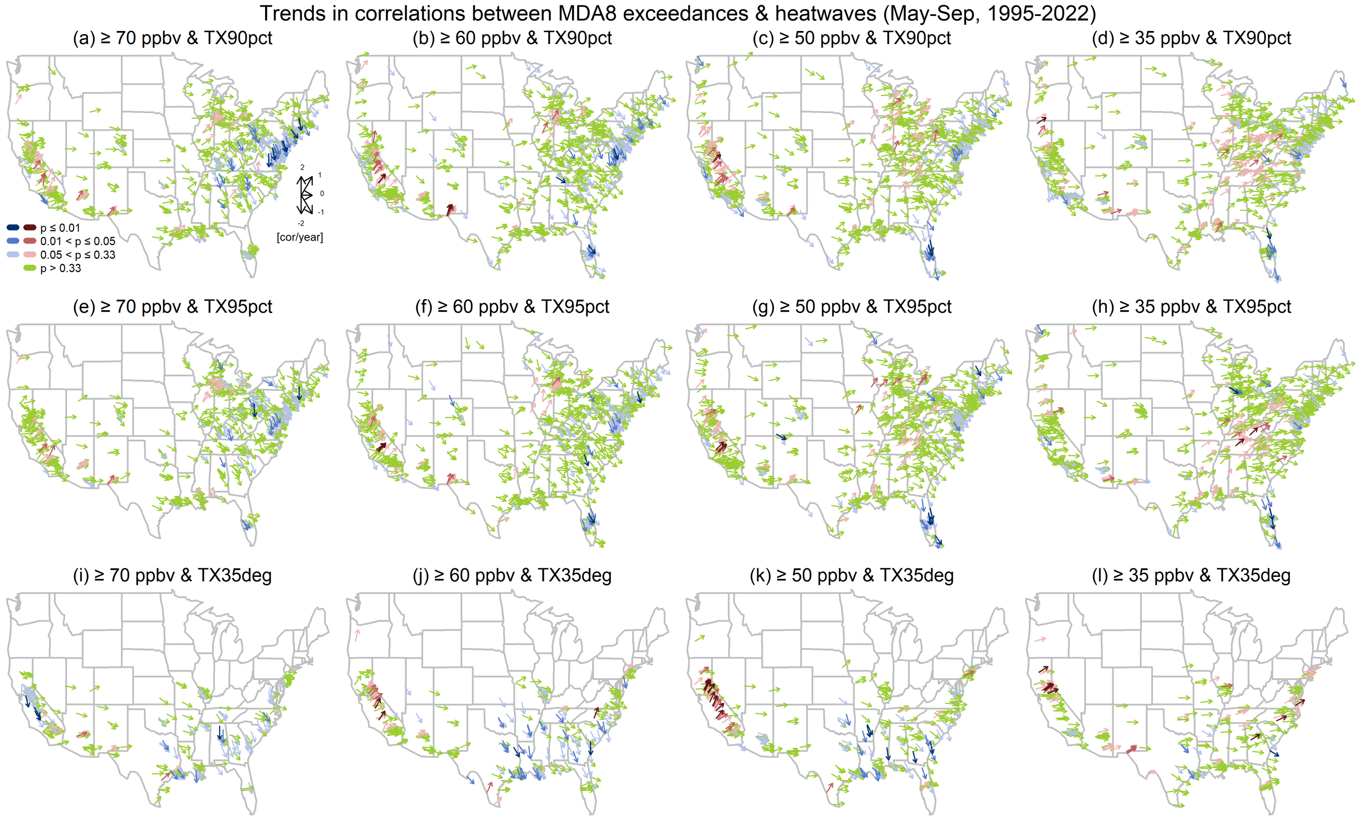


Figure S16: Trends in correlations between ozone exceedances and heatwave events, based on various ozone thresholds and heatwave metrics (May-Sep, 1995-2022).

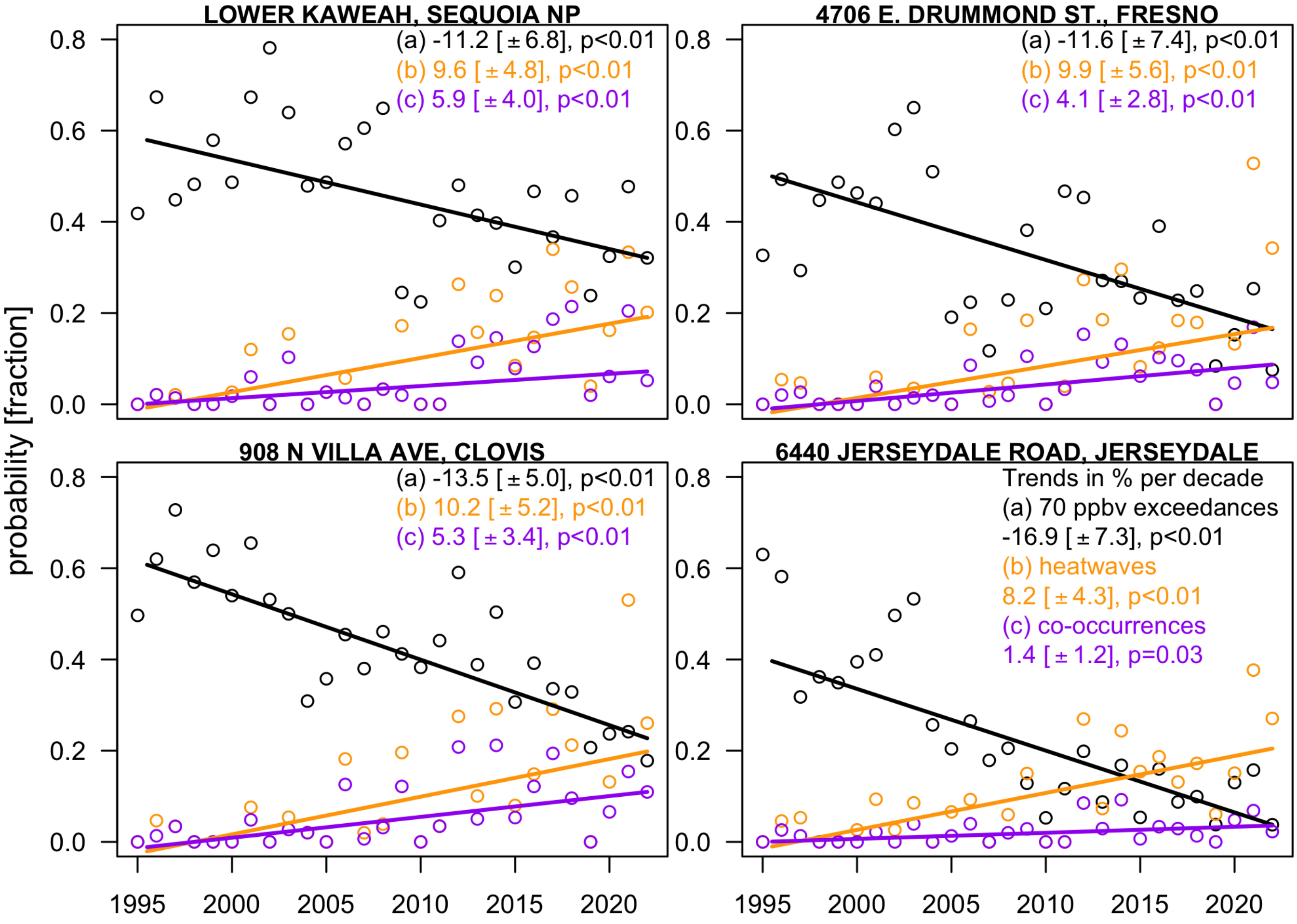


Figure S17: Four sites identified in Fig 11 with decreasing ozone exceedances above 70 ppbv (black), but co-occurrences of ozone exceedances and (TX90pct) heatwaves are increasing (purple). Data points represent the percentages of event days per May-Sep. Also shown are percentages of the TX90pct heatwave days (orange).

References

Chang, K.-L., Cooper, O. R., Gaudel, A., Petropavlovskikh, I., and Thouret, V. (2020). Statistical regularization for trend detection: An integrated approach for detecting long-term trends from sparse tropospheric ozone profiles. *Atmospheric Chemistry and Physics*, 20(16):9915–9938.

Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R., and Goude, Y. (2020). qgam: Bayesian non-parametric quantile regression modelling in R. *Journal of Statistical Software*, 100(9):1–31.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Wood, S. N. (2006). *Generalized additive models: an introduction with R.* CRC press.

Wood, S. N. and Fasiolo, M. (2017). A generalized fellner-schall method for smoothing parameter optimization with application to tweedie location, scale and shape models. *Biometrics*, 73(4):1071–1081.