



Using Satellite Data and Interpretable Machine Learning for Ozone Source Apportionment and Process Attribution

Patrick J. Reddy, Chandler Wells, Tracey Holloway

Holloway Group, University of Wisconsin-Madison - Funded by the Utah Division of Air Quality

NASA Health and Air Quality Applied Sciences Team (HAQAST) Meeting, University of Wisconsin-Madison, May 13, 2026

pjreddy@wisc.edu | preddyresearch@gmail.com | HollowayGroup.org



Introduction and Motivation

- Our Utah Science for Solutions Project: **Comprehensive Google Earth Engine and Satellite Data Analysis Tools to Assess the Impacts of VOC and NOx Sensitivity, Smoke, Heat, Drought, and Plant Stress on Ozone Concentrations in Utah and the Northern Wasatch Front.** Our machine learning system is generalizable, and we have validated it for multiple cities or regions.

Random Forest and SHAP Analysis

- The Google Earth Engine Ozone Random Forest Model platform uses machine learning and 9 features that include satellite observations, land-biosphere state variables, and meteorology for a user-selected set of ozone monitors and region of interest to estimate source and process contributions to daily maximum 8-hour O₃.
- A SHapley Additive exPlanation (SHAP) algorithm computes daily and seasonal contributions to O₃ for each feature and group. SHAP originates from game theory and is a way to fairly and accurately distribute credit to actors or features.

Model Features and Data Sources

NOx Chemistry: NO₂ - TROPOMI/TEMPO L3 tropospheric column.

VOC Chemistry: HCHO - TROPOMI/TEMPO L3 column.

Heat: 2-meter temperature - ERA5-Land hourly.

Land-Biosphere:

Soil moisture - ERA5 hourly surface volumetric soil water.
Specific humidity anomaly - gridMET specific humidity (SPH) (SPH replaces Vapor Pressure Deficit to reduce autocorrelations).
NDVI - MODIS Terra 16-day 1 km composite.

Transport/Ventilation:

500 mb wind speed - MERRA-2 U500/V500 vector magnitude, and PBL anomaly - GEOS-CF boundary layer height.

Smoke Events: Smoke flag - Sum-of-flags (>=3); TROPOMI CO & MERRA-2 surface black carbon vs P95/P99 thresholds.

Random Forest: Tuned for Attribution

- Random Forest (RF) models can account for complex nonlinear processes in the atmosphere/biosphere and reproduce response curves consistent with atmospheric chemistry and physics. SHAP can explain model responses and feature interrelationships. To minimize variance leakage across features we test for low variance inflation factors and use k-fold cross-validation. RF for this project is tuned for retrospective attribution and is not optimized for forecasts.

SHAP Results Align with OSAT O₃ Source Apportionment

- For seven Salt Lake City sites, the RF/SHAP model has an R² of 0.84 and a relatively low underprediction bias for 2022-2024.
- SHAP estimates additive contributions to O₃ for broad categories of sources and process groups.
- In Table 1 we have compared our 2022-2024 SHAP results with Utah/Ramboll CAMx Ozone Source Apportionment Technology (OSAT) results for 2023 based on 2017 meteorology.

Table 1. OSAT versus SHAP O₃ for exceedance events: Salt Lake City Bountiful and Hawthorne sites.

Category	OSAT (ppb)	SHAP (ppb)
Predicted MDA8 O ₃ (exceedance days)	66.8	73.1
Anthropogenic (state/local)	16.6	14.0
Biogenic / Land-Biosphere	6.6	5.9
Heat	-	5.5
Ventilation	-	5.1
Smoke	-	0.7
Background / P05 Baseline	43.4	41.9

OSAT anthropogenic O₃ (whole state) is 2.6 ppb higher than local SHAP SLC anthropogenic O₃.

Multi-City CAMx OSAT/APCA vs SHAP

- In Figure 1, comparable categories show good agreement across Salt Lake City, Chicago, Las Vegas, and Denver. SHAP anthropogenic O₃ is 20-25% lower than OSAT and the Anthropogenic Precursor Culpability Assessment (APCA) - which are deterministic, process-based models.

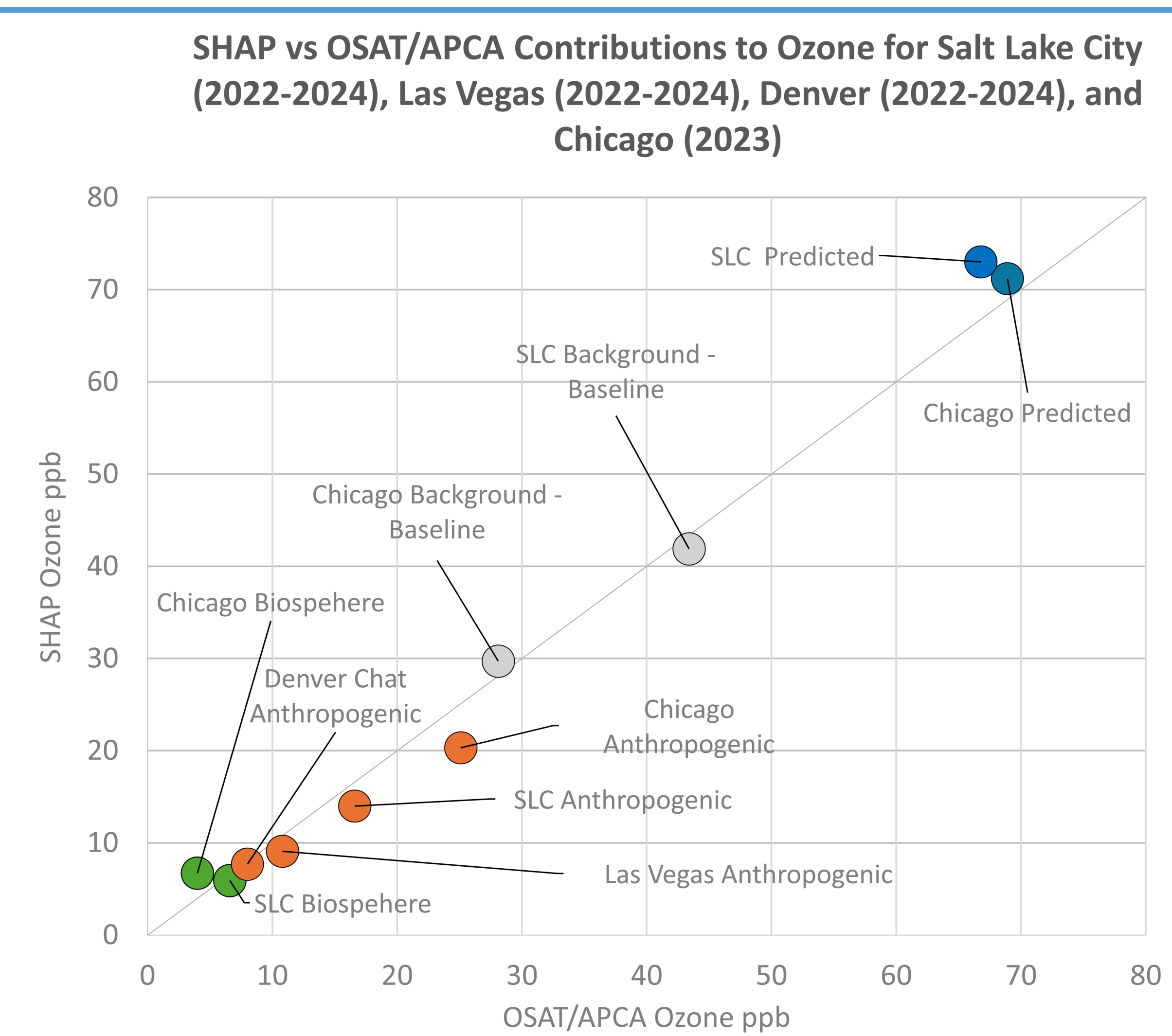


Figure 1. SHAP vs OSAT/APCA contributions to ozone for Hawthorne and Bountiful monitors in SLC, Joe Neal Monitor in Las Vegas, Chatfield near Denver, and 6 Chicago-area monitors (2022-2024). One-to-one line is shown.

SHAP O₃ from Smoke

- A sampling routine designed to optimize representation of time periods and atmospheric conditions results in the selection of about 33% of days in each season for SHAP analysis.
- Jing et al. (2026) estimate peak daily smoke contributions to O₃ in Chicago at 6.7 ppb during 2023. Our SHAP analysis for Chicago O₃ estimates peak daily smoke contributions of ~10 ppb in 2023.
- On August 7, 2021, a relatively narrow plume of smoke was associated with a 25 ppb increase in O₃ in southwest Utah compared with the mean for the day before and after. SHAP analysis shows a mean smoke contribution of 18 ppb for southwest Utah on this day.
- RF underprediction bias is very low for days with significant smoke. Figure 2 shows SHAP smoke O₃ in Salt Lake City in 2021.

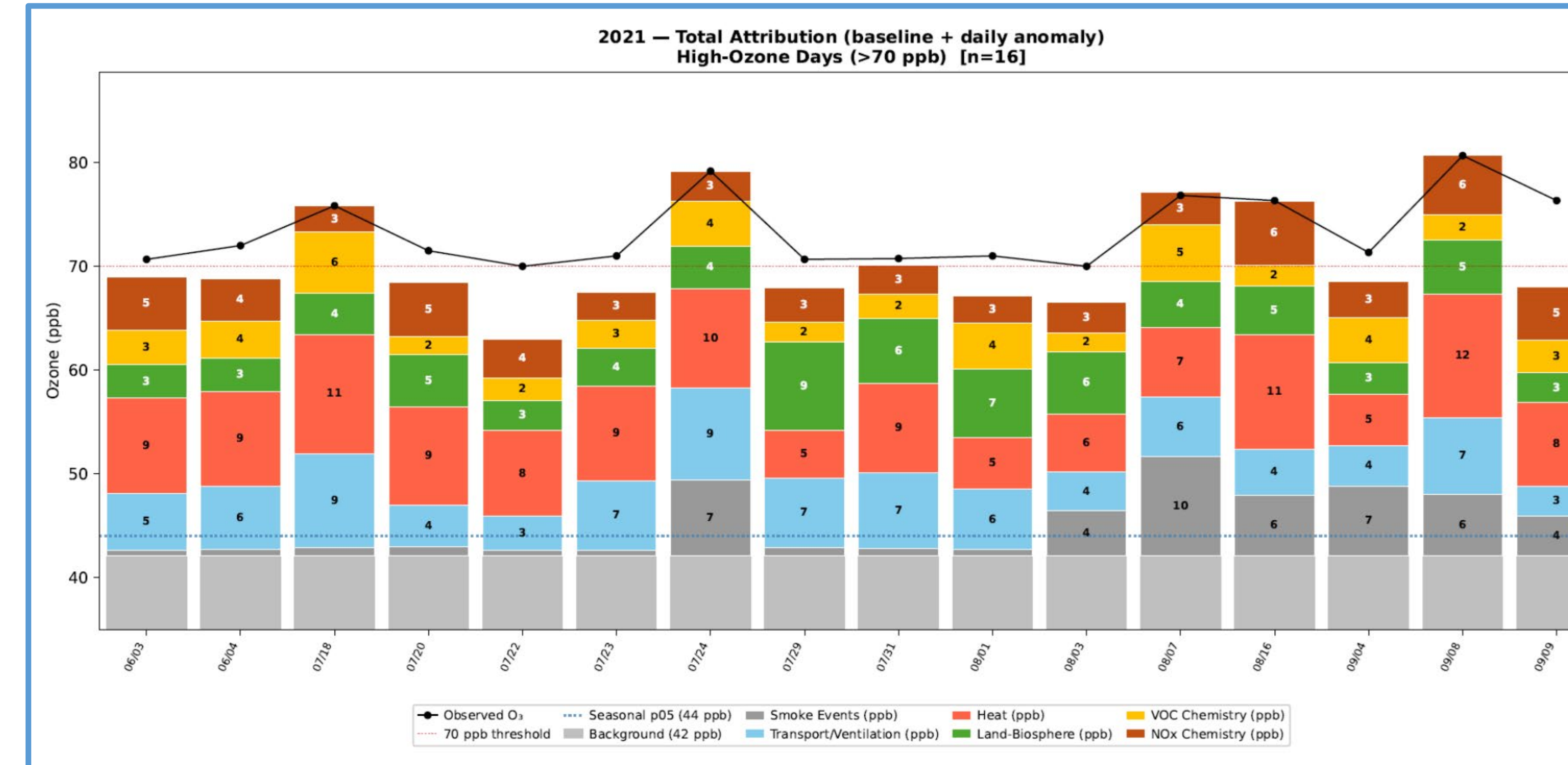


Figure 2. SHAP Contributions to O₃ for six monitors in Salt Lake City in 2021 - with smoke O₃ in dark grey. Smoke contributed 10 ppb on a day with widespread smoke from fires in California - 2021-08-07.

Emergent Properties from RF/SHAP Suggest Missing Pathways in MEGAN and BEIS on Days When Very Dry Air Enhances BVOCs from Monoterpene Emitters in the Arid West

- Sagebrush, Gambel oak (leaf litter?), piñon-juniper, mixed conifers, creosote bush, and desert scrub may be significant sources of monoterpene emissions from stored pools, volatilization through leaf and needle cuticles, as well as physiological responses to protect the plants. SPH anomaly is a surrogate for VPD in our models. MEGAN/BEIS are biogenic emissions models.

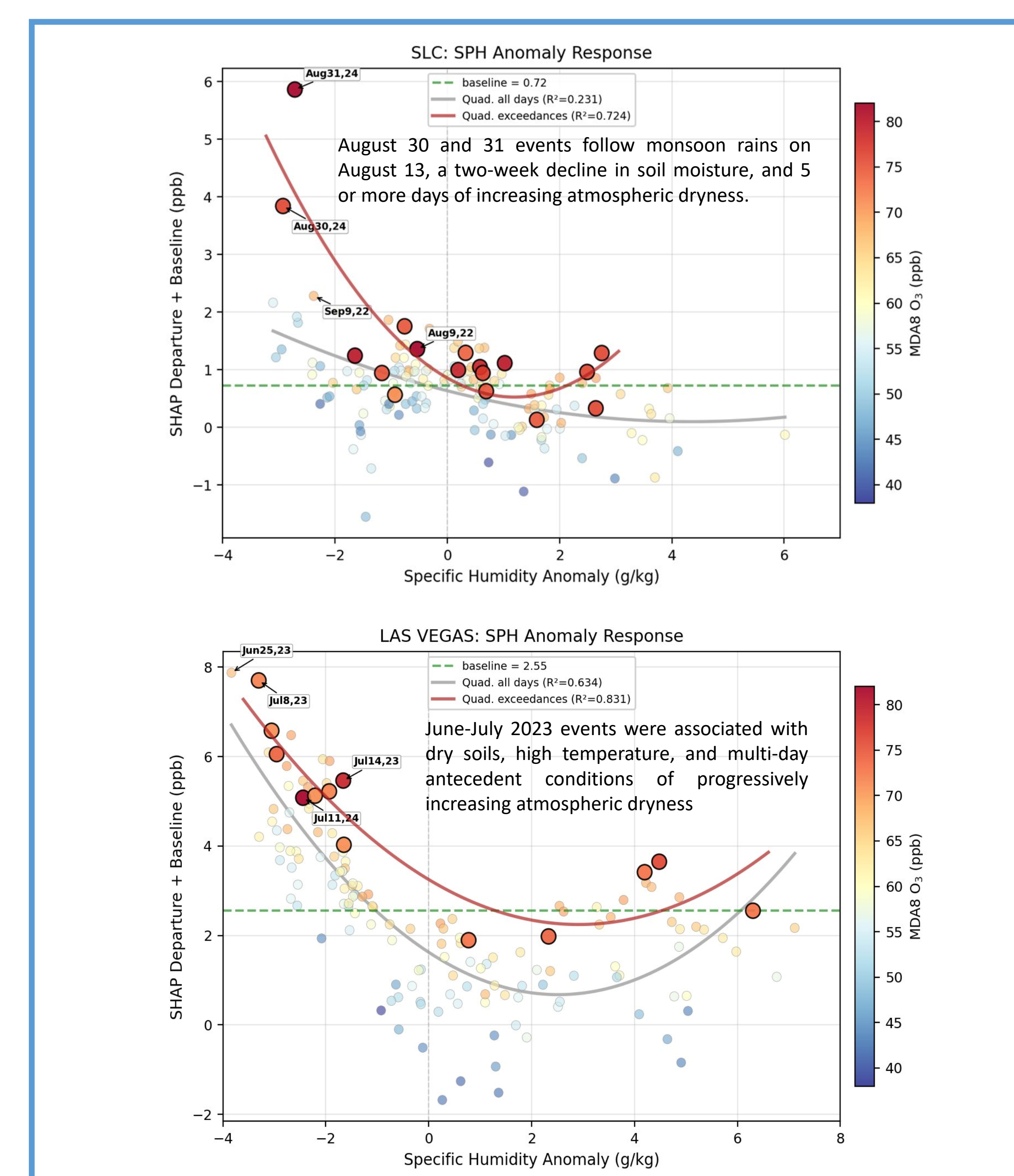


Figure 3. The quasi-independent response of O₃ to SPH anomaly suggests that 5- to 6-day episodes of increasingly dry air are associated with the release of monoterpenes that can enhance O₃ by up to 6-8 ppb on exceedance days in Salt Lake City (top) and Las Vegas (bottom). Dark red circles are exceedance days.

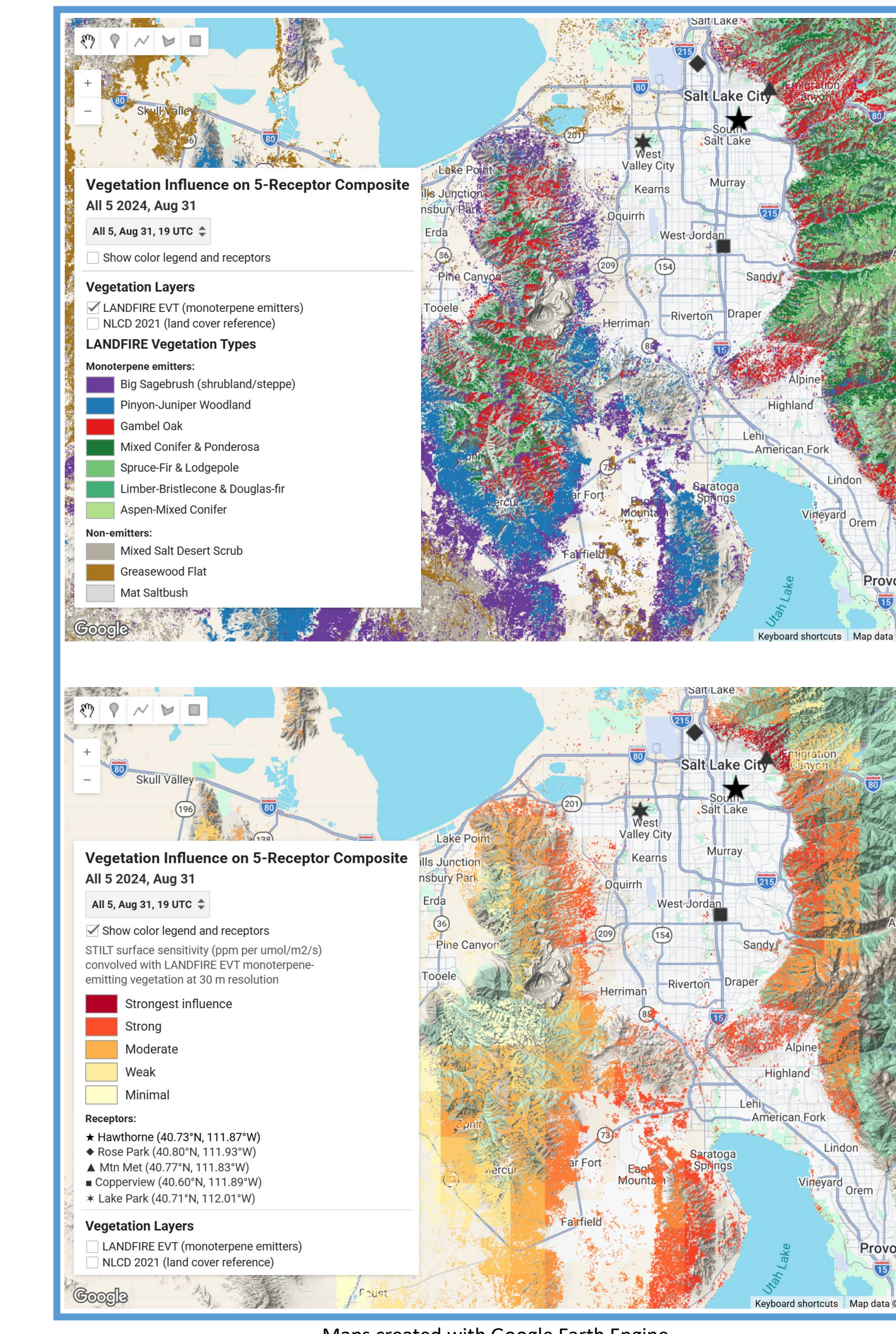


Figure 4. LANDFIRE EVT vegetation in the Salt Lake City region for potential monoterpene emitters (top) weighted by the Stochastic Time-Inverted Lagrangian Transport (STILT) source area footprint for 19 UTC August 31, 2024 (bottom, from the USOS field catalogue, Fasoli et al., 2018, and Lin et al., 2003). This supports the plausible influence of monoterpene emitters on a day with an SPH SHAP O₃ of 6 ppb.

- Figure 3 shows the non-linear response of O₃ to abnormally dry air in Salt Lake City and Las Vegas. While this SHAP category shares some variance with Heat, NDVI, and soil moisture SHAP, it is largely independent and points to possible dry-side contributions by plants. **MEGAN and BEIS biogenic emission models do not account for this emission pathway, which may be a significant contributor to ozone during exceedance events (up to 6-8 ppb).** The two panels in Figure 4 demonstrate the plausibility of this influence on the highest-concentration day in Salt Lake City in 2024, with MDA8 > 80 ppb. The STILT source area footprint highlights possible monoterpene emitters.

Conclusions

- The RF/SHAP modeling system is generalizable and shows good agreement with conventional ozone source apportionment methods. It shows promise as an estimator for ozone from smoke, with notably low underpredictions when smoke is present.
- The RF/SHAP suite provides an inexpensive, SIP-quality, satellite-based, attribution toolkit for major source groups that complements conventional modeling and source apportionment methods. Satellite observations of NO₂, HCHO, CO, NDVI, and aerosol properties drive the RF/SHAP models.
- The toolkit can be used annually with short set-up times and can provide a snapshot of conditions in between O₃ modeling years and field campaigns.
- RF/SHAP's reproduction of non-linear processes has identified a novel pathway for BVOC-enhanced O₃ in the arid Southwest.