Level 2 Data Validation: Whoops! What the Grass Roots Level Missed

Prepared by:
Hilary R. Hafner
Sonoma Technology, Inc.
Petaluma, CA

Kevin Cavender
U.S. EPA OAQPS

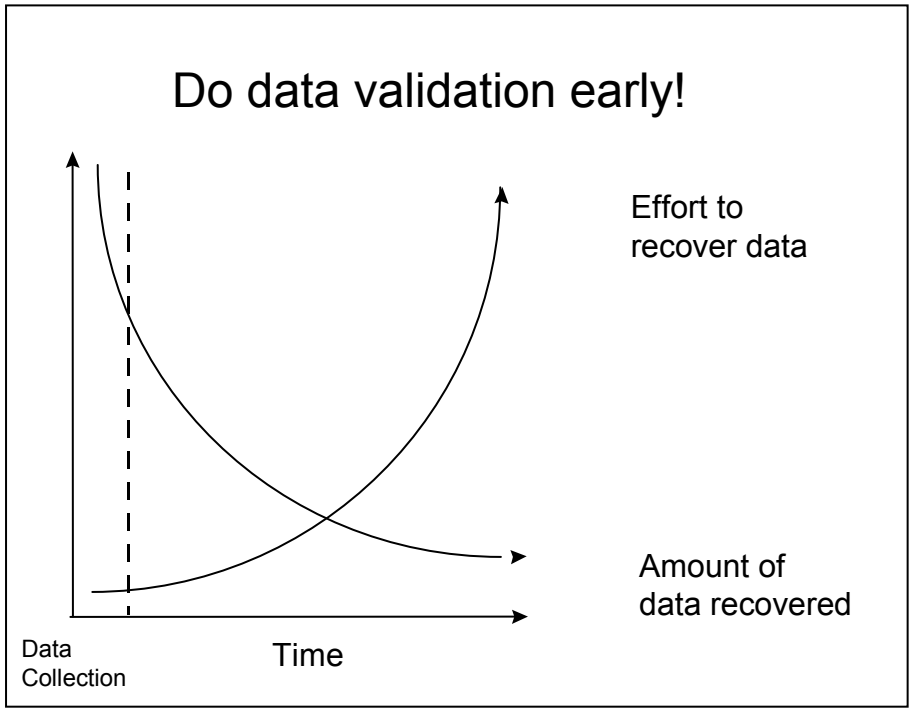
Presented to:
2006 National Air Monitoring Conference
Las Vegas, NV
November 6-9, 2006

Introduction and Overview

- Importance of Data Validation
- Data Validation Levels
- General Approach to Data Validation
- Examples
- Resources

Why Should You Validate Your Data?

- It is the monitoring agency's responsibility to prevent, identify, correct, and define the consequences of monitoring difficulties that might affect the precision and accuracy, and/or the validity, of the measurements.
- Serious errors in data analysis and modeling (and subsequent policy development) can be caused by erroneous data values.
- Accurate information helps you respond to community concerns.



Examples of Problems in Criteria Pollutant Databases (*and Validation Actions*)

- Air quality data reported during calibration runs. For example, ozone data with values of 0 ppb (or the calibration gas level) reported during hours when instruments are known to be automatically calibrated. *Data were flagged as calibration.*
- Nitrogen oxides data found to have a constant offset based on comparisons of NO_x to $NO+NO_2$. *Data were adjusted.*
- Ozone concentrations "capped" at 100 ppb. Investigation showed that the instrument maximum concentration setting was incorrect. *Data at 100 ppb were flagged as suspect low, and the instrument settings were adjusted.*

Data Validation Levels: Summary of Types of Checks

- Level I
  - Routine checks during the initial data processing and generation of data including proper data file identification; review of unusual events, field data sheets, and result reports; and instrument performance checks.
- Level II
  - Internal consistency tests to identify values in the data that appear atypical when compared to values from the entire data set.
  - Comparisons of current data with historical data to verify consistency over time.
  - Parallel consistency tests with data sets from the same population (e.g., region, period of time, air mass) to identify systematic bias.

U.S. EPA, 1999a

Level II: Internal Consistency Checks

- Inspect time series. Are concentrations consistent with time of day, day of week, and season?
- Compare pollutant concentrations. Are expected relationships observed?
- Identify and flag unusual values including
  - Values that normally follow a qualitatively predictable spatial or temporal pattern
  - Values that normally track the values of other variables in a time series
  - Extreme values, outliers

The first assumption upon finding a measurement that is inconsistent with physical expectations is that the unusual value is due to a measurement error. If, upon tracing the path of the measurement, nothing unusual is found, the value can be assumed to be a valid result of an environmental cause.

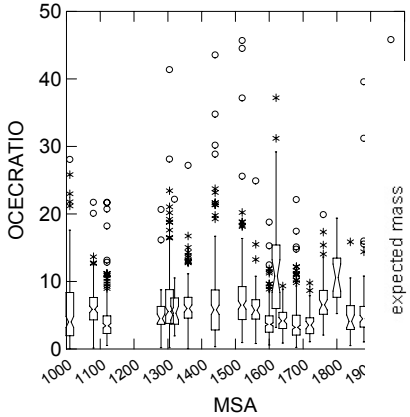
Chow et al., 1996

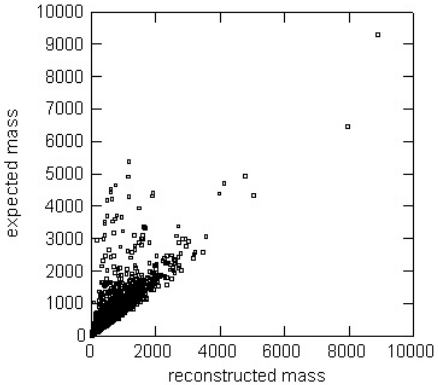
Level II+: Comparisons to Other Data Sets

- Compare collocated measurements.
- Compare relationships (e.g., temporal, among species) observed in the current data set to relationships observed at other sites or in previous years.
- Compare pollutant concentrations to meteorology.

General Approach to Data Validation

- Look at your data.
- Manipulate your data—sort it, graph it, map it—so that it begins to tell a story.
- Often, important issues or errors with the data will become apparent only after someone begins to use the data for something.
- Examples
  - Scatter plots
  - Time series plots
  - Fingerprint plots
  - Box whisker plots
  - Summary statistics





Example Validation Steps (Page 1 of 2)

- Assemble the database.
- Place data in a common data format with descriptive information concerning variables, validation level, QC codes, detection limits, time standard, standard units, and metadata (site information, etc.).
- Ensure that results of and suggestions from all audit reports have been incorporated into the database.

Example Validation Steps (Page 2 of 2)

- Review summary statistics for unrealistic maxima or minima and for consistency with nearby stations (data are still Level I).
- Perform spatial and temporal comparisons of the data (begin Level II).
- Perform intercomparisons of the data (e.g., from two different instruments).

Example Data Overview

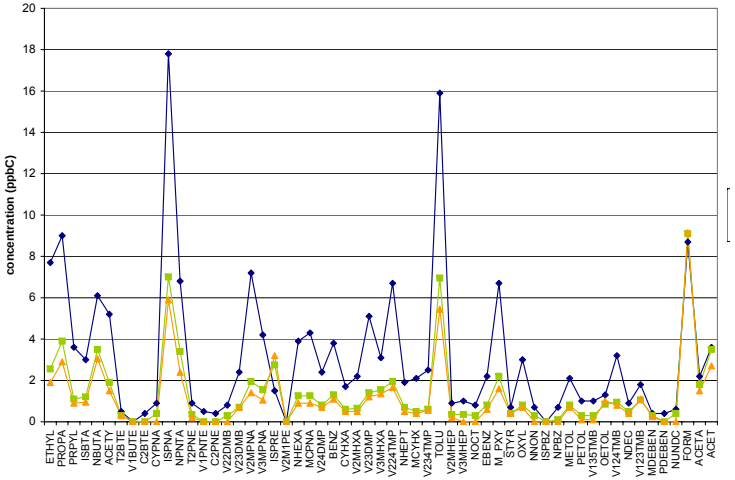
Site	Year	Ozone	CO	NO_2	SO_2	PM_{10} HiVOL	PM_{10} Dichot	PM_{10} BAM	$PM_{2.5}$ Dichot	$PM_{2.5}$ Dichot	$PM_{2.5}$ BAM
Chicago-Jardine	1999	8432	8212								
Chicago-Jardine	2000	8401	…								

Be sure to split aerosol measurements into different size and analytical groups.

- As a part of validation, it is useful to prepare a summary of the monitoring network by year: summarize which sites have data and how much data for which years.
- Use this summary to detect potential problems and to determine what types of analyses are possible.

Considerations in Evaluating Your Data

- Levels of other pollutants
- Time of day/year
- Observations at other sites
- Audits and inter-laboratory comparisons
- Instrument performance history
- Calibration drift
- Site characteristics
- Meteorology
- Exceptional events



Singling Out Odd Data

- Range checks: minimum and maximum concentrations
- Temporal consistency checks: maximum hour test
- Rate of change or spike check
- Buddy site check: comparison to nearby sites
- Sticking check: consecutive equal data values

Example Screening Criteria – Ozone

- Checks:
  - Are often site-specific
  - May be hour-specific
  - May be automated
- But, data should be graphically reviewed!

Check	Criteria
Maximum	~170 to 225 ppb
Minimum	–5 ppb
Rate of change	>50 to 60 ppb/hr
Buddy sites	±50 ppb up to 5 sites
Sticking check	@≥40 ppb for 5 hours
Co-pollutant	NO, NO_x

U.S. EPA, 2003

Example – Ozone Screening (1 of 2)

AIRNowTech

Current DMC Status
Polling Summaries
Monitoring Sites
Data Queries
Forecast System
Forecast Queries
Agency Info / Setup
Resources
AIRNow Notifier
Contacts

Logoff ST1

Parameter

Ozone

Buddy Sites

No buddy sites selected

Quality Control Criteria QC Descriptions

Hour [LST]	Max Suspect	Max Severe	Rate Of Change	# of Buddy Sites	Buddy Average	# of Sticking Hours	Sticking Value	Minimun Drift
0000	110	189	60	3	50	5	40	-5
0100	110	189	60	3	50	5	40	-5
0200	110	189	60	3	50	5	40	-5
0300	100	168	60	3	50	5	40	-5
0400	100	168	60	3	50	5	40	-5
0500	100	168	60	3	50	5	40	-5
0600	100	168	60	3	50	5	40	-5
0700	100	168	60	3	50	5	40	-5
0800	100	168	60	3	50	5	40	-5
0900	100	168	60	3	50	5	40	-5
1000	100	168	60	3	50	5	40	-5
1100	140	210	60	3	50	5	40	-5
1200	140	210	60	3	50	5	40	-5
1300	140	210	60	3	50	5	40	-5
1400	140	210	60	3	50	5	40	-5
1500	140	210	60	3	50	5	40	-5
1600	140	210	60	3	50	5	40	-5
1700	140	210	60	3	50	5	40	-5
1800	140	210	60	3	50	5	40	-5

Note hour-specific screening

Max Suspect:
Still used in spatial mapping

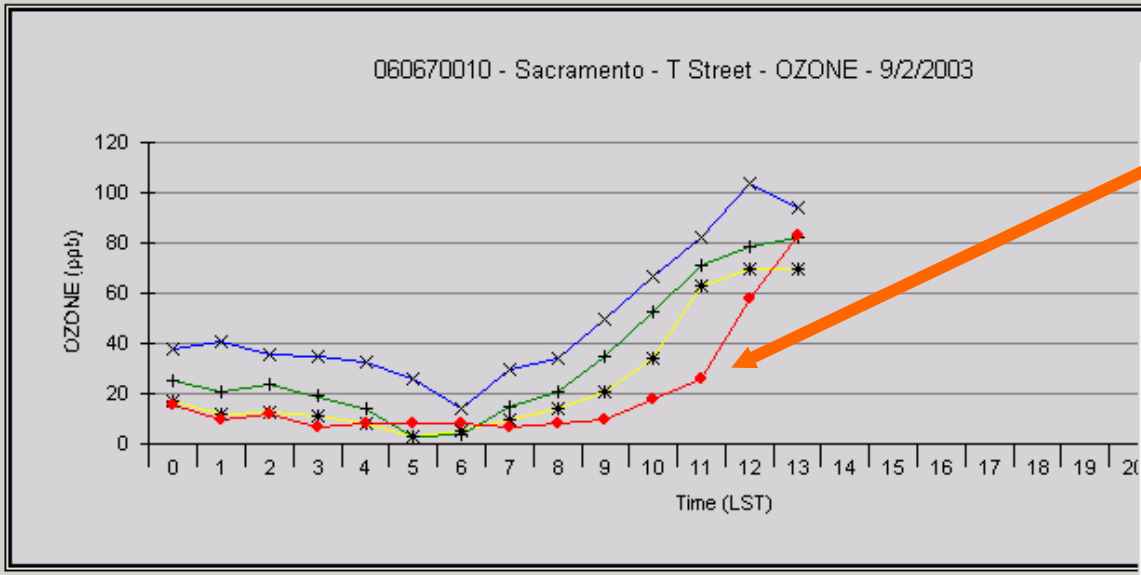
Max Severe:
Not used in maps

Example – Ozone Screening (2 of 2)

Edit QC Flags

Select Date (LST)

Date: 9/2/2003



Screening Test Results

Hour (GMT)	Description
19	Failed Buddy Check

Compared to "buddy" sites, this site had an [O_3] difference of more than 50 ppb. Note, however, that data are valid.

8 Hours Offset to GMT

OK

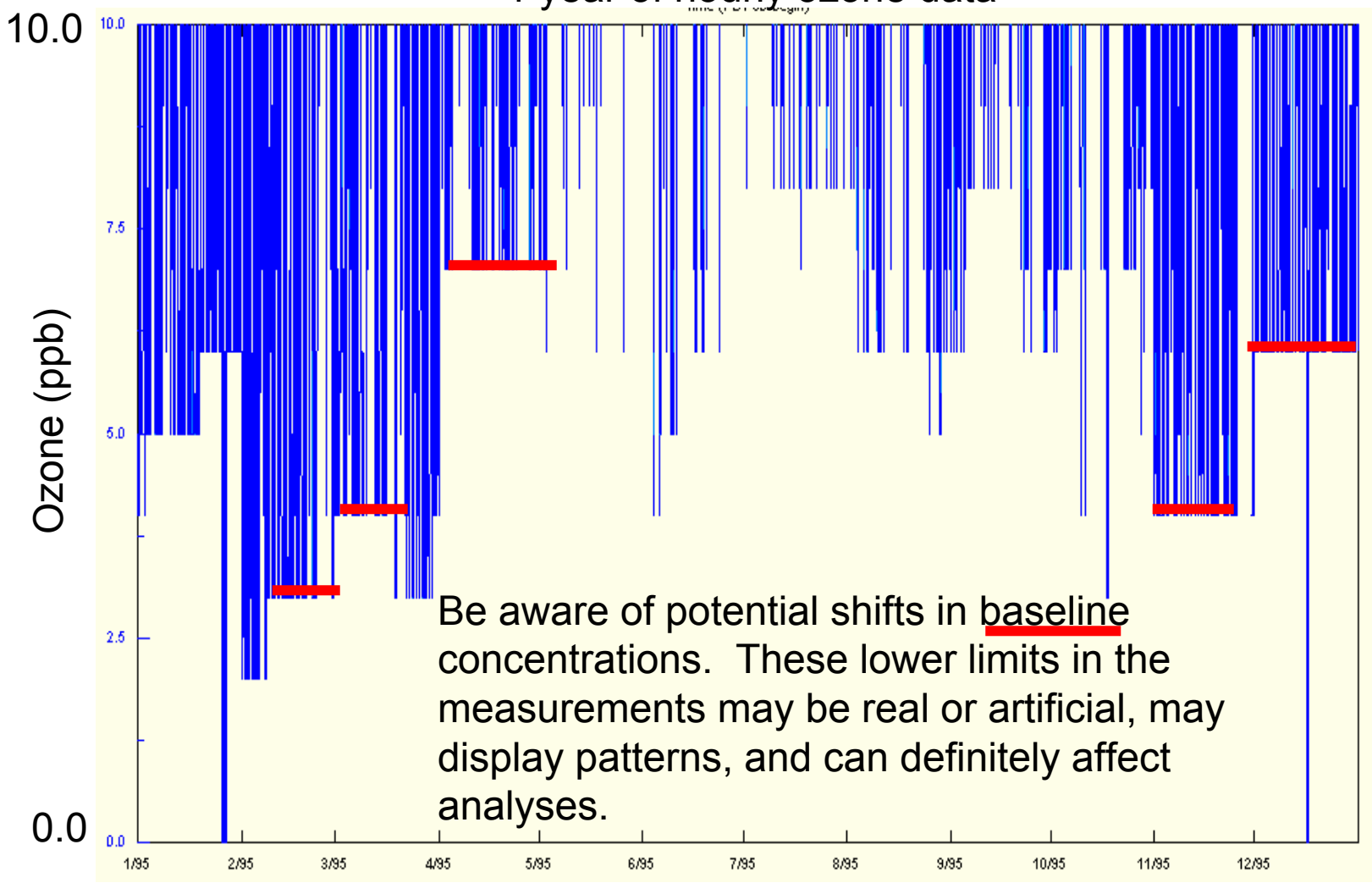
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Raw	16	10	12	7	-999	8	8	7	8	10	18	26	58	83										
Data	16	10	12	7	8	8	8	7	8	10	18	26	58	83										
QC	G	G	G	G	I	G	G	G	G	G	G	G	G	G										

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Min Drift	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5
Max Suspect	110	110	110	100	100	100	100	100	100	100	100	140	140	140	140	140	140	140	140	130	130	130	130	130
Max Severe	189	189	189	168	168	168	168	168	168	168	168	210	210	210	210	210	210	210	210	189	189	189	189	189
Rate Change	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60

Auto Mode | Valid (G) | Suspect (K) | Invalid (B) | Site X of Y | Update | Close

Example – Ozone Validation

1 year of hourly ozone data



Be aware of potential shifts in baseline concentrations. These lower limits in the measurements may be real or artificial, may display patterns, and can definitely affect analyses.

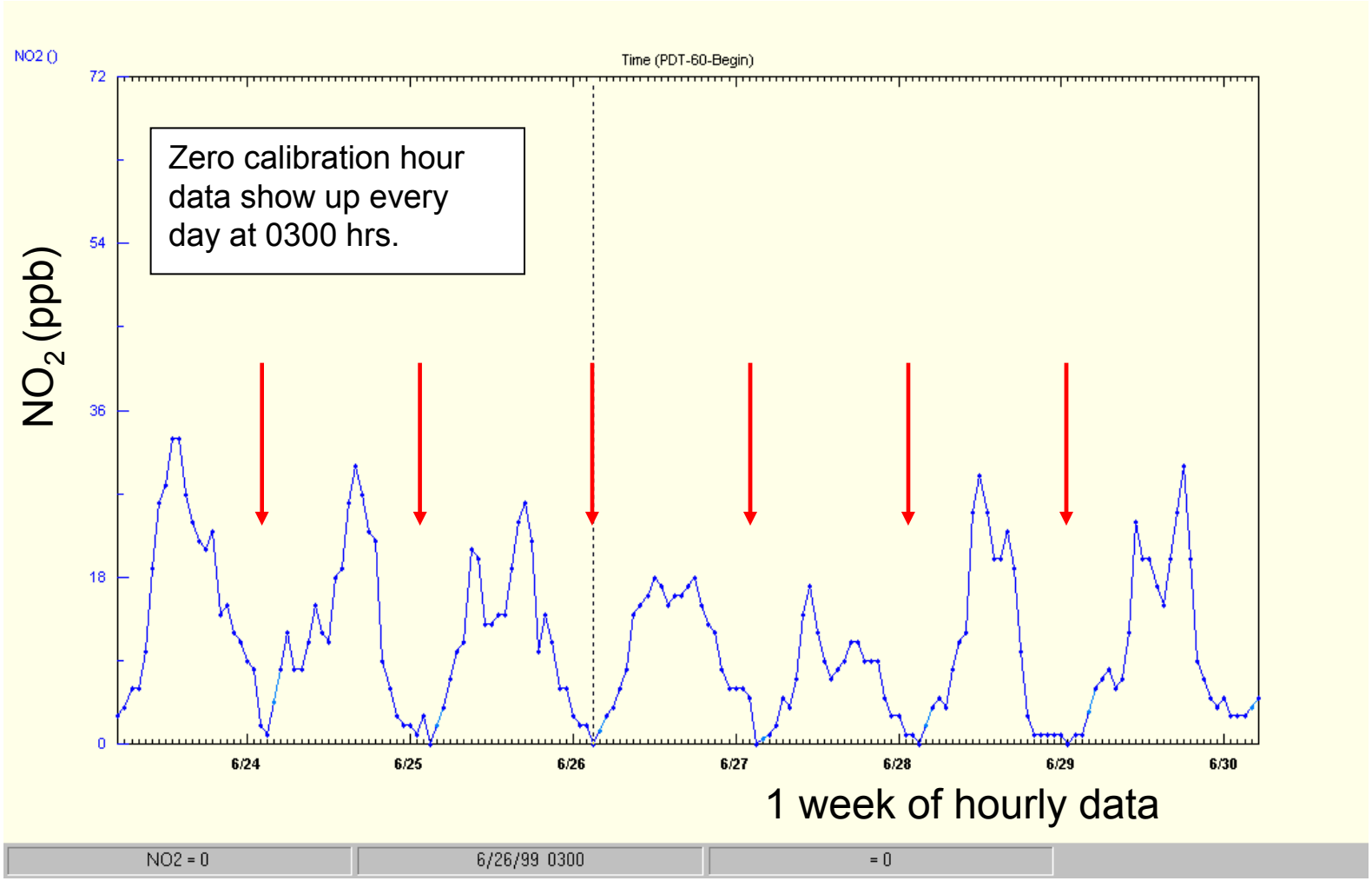
Example Screening Criteria – $NO/NO_x/NO_y$

- Checks:
  - Select buddy check criteria
- Collocated ozone can be used to assess NO, NO_x, NO_y
- Checks may vary depending upon instrument sensitivity

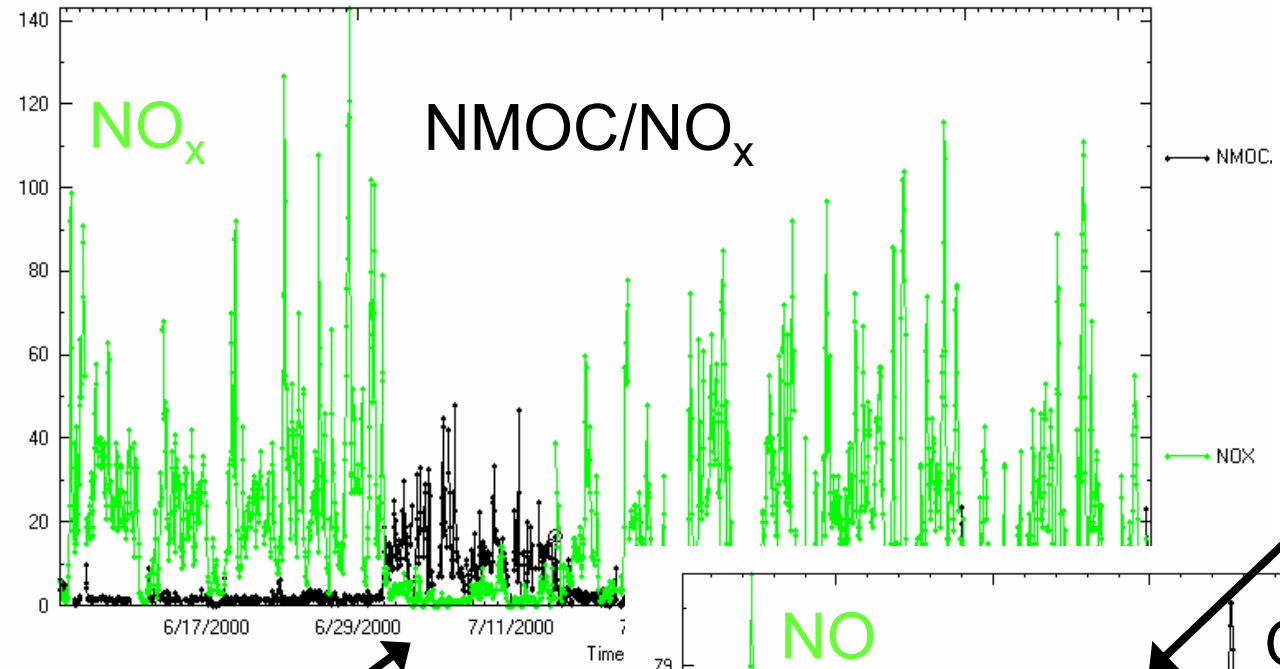
Check	Criteria
Maximum	>700 ppb urban >300 ppb rural
Minimum	–1 ppb
Rate of change	>30 ppb/hr
Sticking check	5 hours
Co-pollutant	NO should not exceed NO_x or NO_y

Hafner et al., 2003

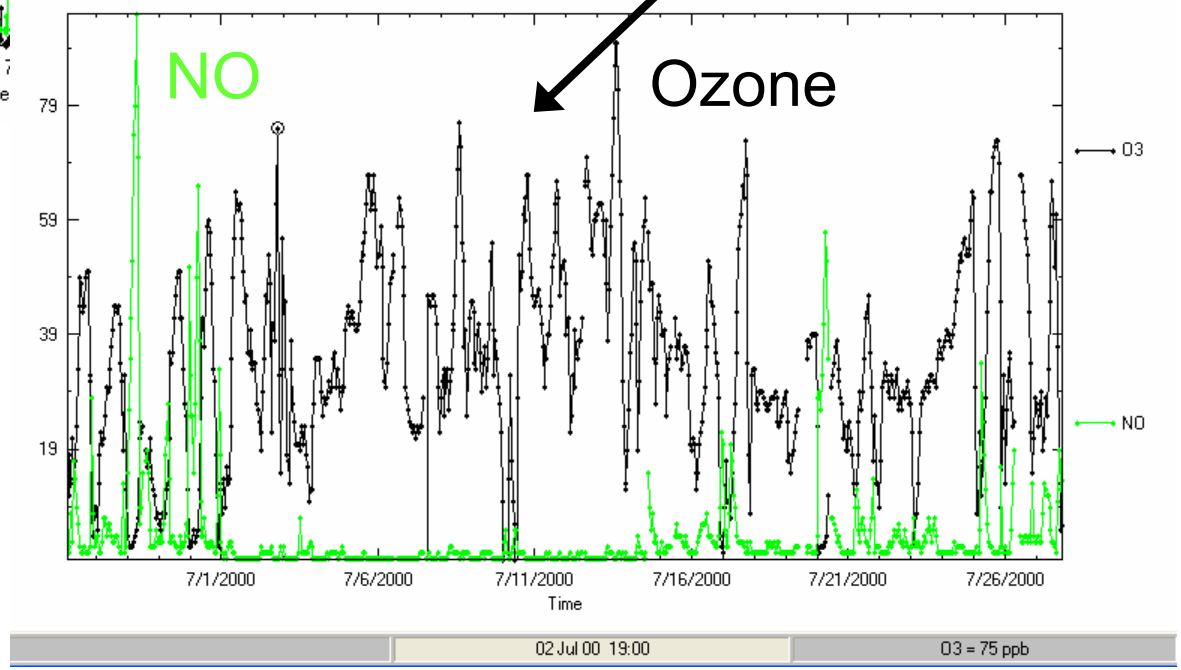
Example – Odd Patterns (1 of 2)



Example – Odd Patterns (2 of 2)

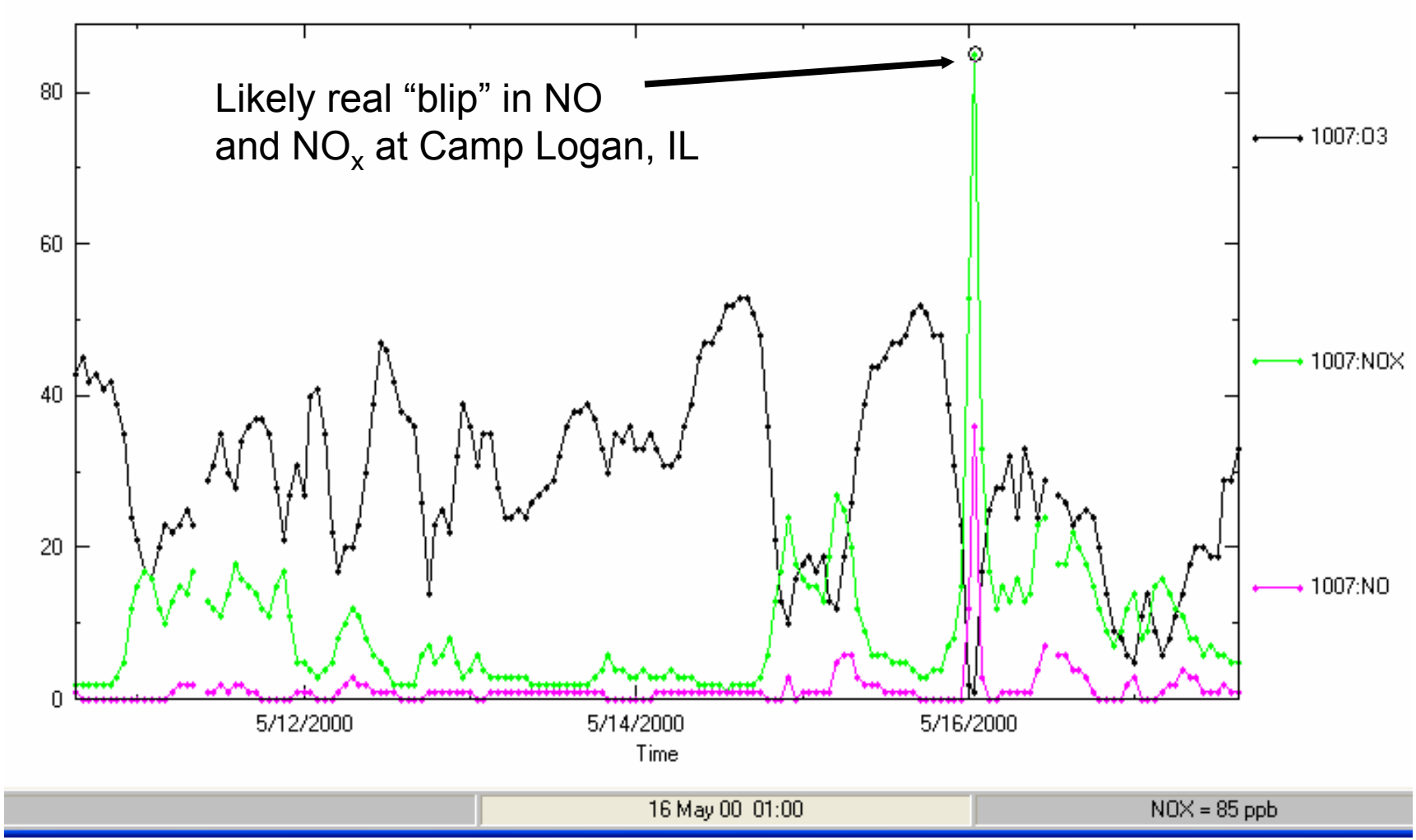


Other pollutants are useful in identifying periods of different behavior.



Note period of lower NO_x, NO concentrations. Is this real or an instrument problem?

Example – Ozone, NO_x, NO

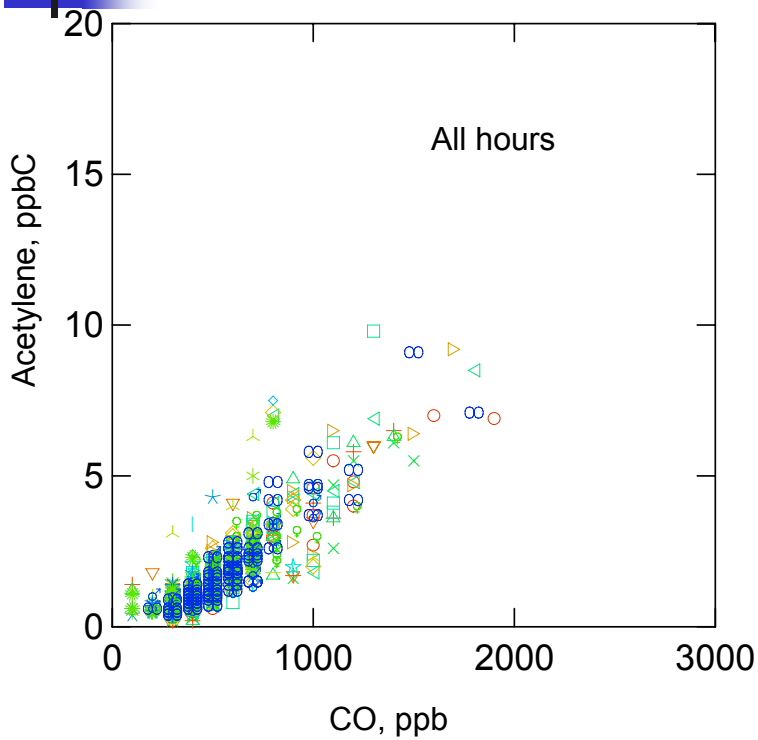


Example Screening Criteria – CO

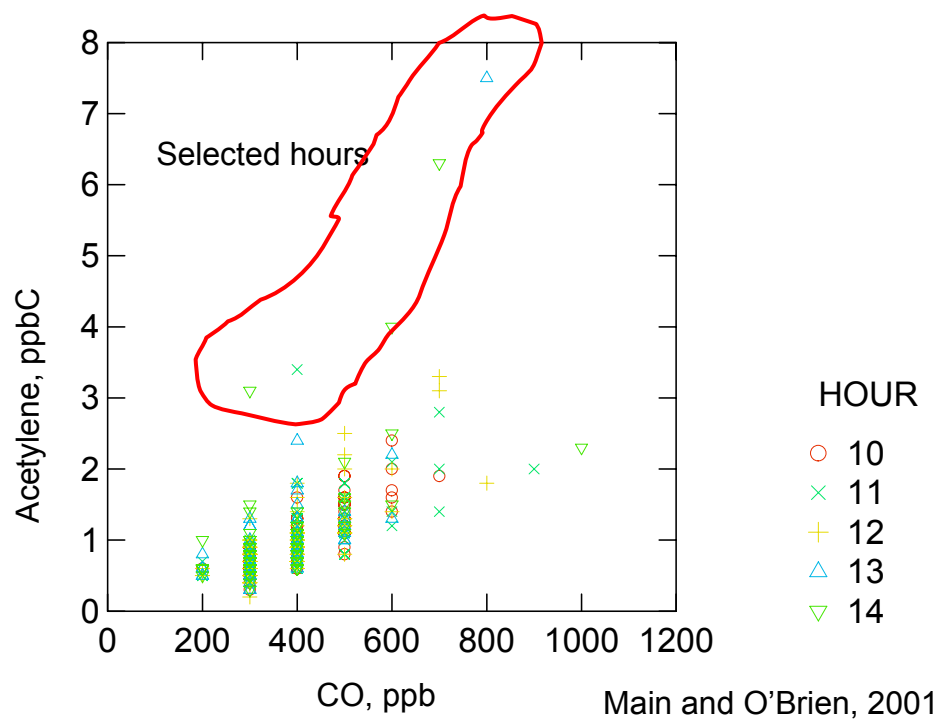
- Checks:
  - Select buddy check criteria
- Checks may vary depending upon instrument sensitivity

Check	Criteria
Maximum	>15 ppm
Minimum	–1 ppm
Rate of change	>10 ppm/hr
Sticking check	> 0 ppm for 5 hours
Co-pollutant	NO, acetylene

Example – PAMS Data



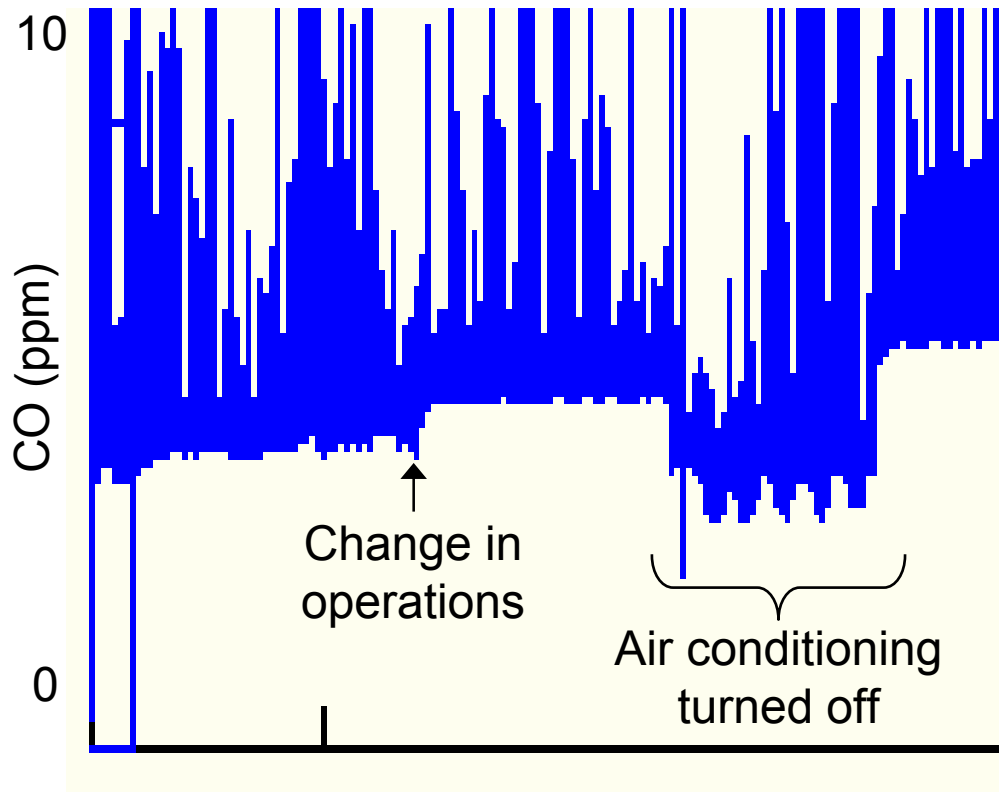
PAMS VOC data at Camden, New Jersey, were compared to CO data for all hours and for selected hours. Some of the midday CO concentrations should be investigated further.



Main and O'Brien, 2001

Example – CO Baseline

- In this example, CO data were collected with a deliberate offset of about 5 ppm so that changes in baseline concentrations could be observed.
- Note the lower baseline concentrations and diurnal variation for the indicated time period. The changes were a result of no air conditioning in the instrument shelter.



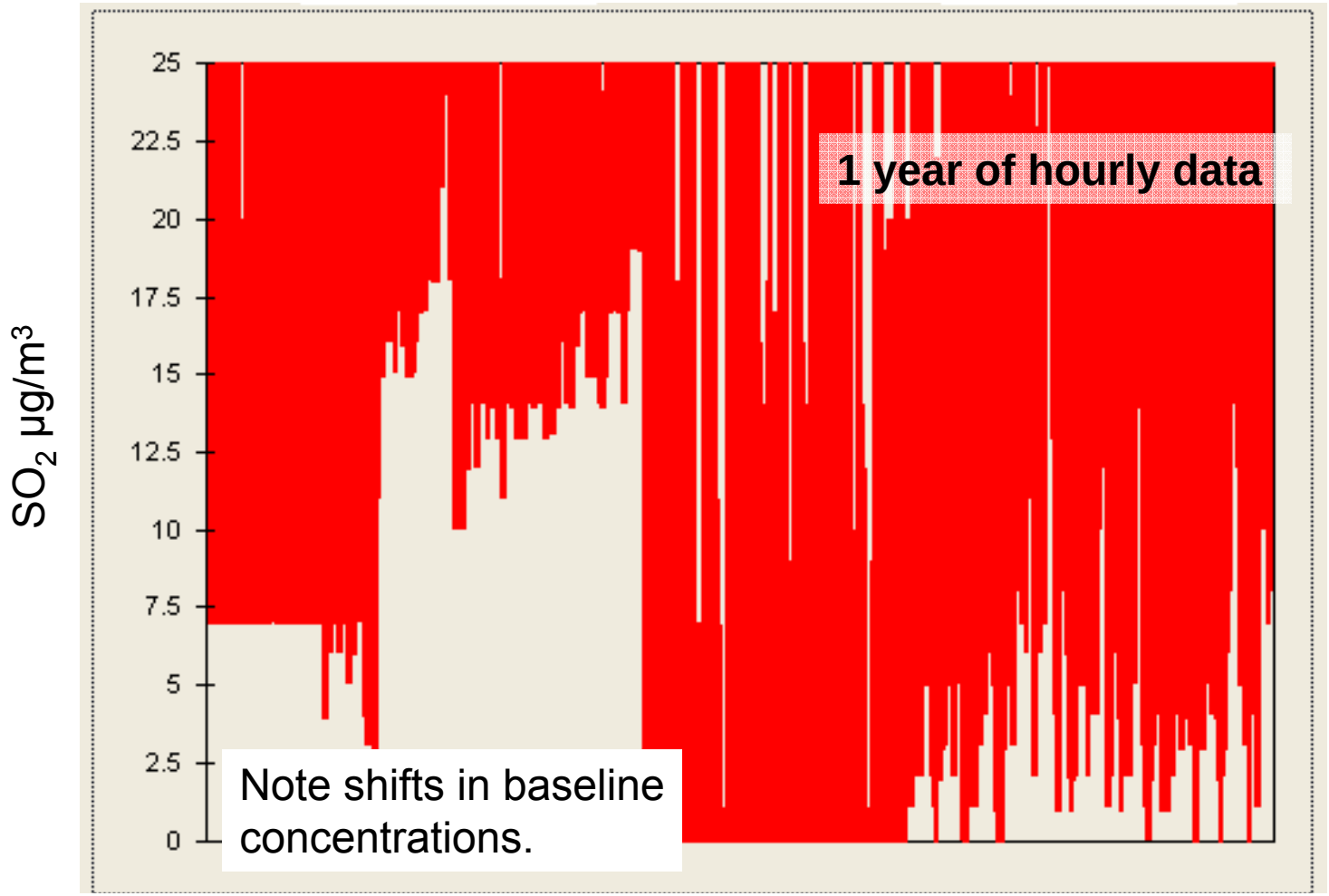
Hourly data for more than one month

Example Screening Criteria – SO_2

- Checks:
  - Select buddy check criteria
- Checks may vary depending upon instrument sensitivity
- Regional issues
  - Rural/urban differences
  - Southeast vs. west

Check	Criteria
Maximum	400 μg/m^3 (or 150 ppb)
Minimum	–5 μg/m^3 (or -2 ppb)
Rate of change	>100 μg/m^3/hr (or 40 ppb/hr)
Sticking check	>0 for 5 hours
Co-pollutant	NO_x

SO_2 Baseline Check



Typical Pollutant Groupings Used in Validation

- Ozone, NO, and NO_x or NO_y
  - Also useful: particle scattering
- CO, NO, and NO_x
  - Also useful: TNMOC
- SO_2 and NO_x
  - Also useful: continuous mass or sulfate

TNMOC = total nonmethane organic compounds

Example Baseline Investigation Criteria

Example criteria for investigating baseline changes (especially important to exposure assessments)

- Use 1 year of hourly data
- Use the following data ranges
  - Ozone: 0 to 10 ppb
  - NO_2: 0 to 15 ppb
  - CO: 0 to 2 ppm
  - SO_2: 0 to 10 ppb
- Look for either step functions or gradual drift in the baseline resulting from improper maintenance, post-processing of the data, etc.

Alcorn, 2003

Example Screening Criteria – 1-hr $PM_{2.5}$ Mass

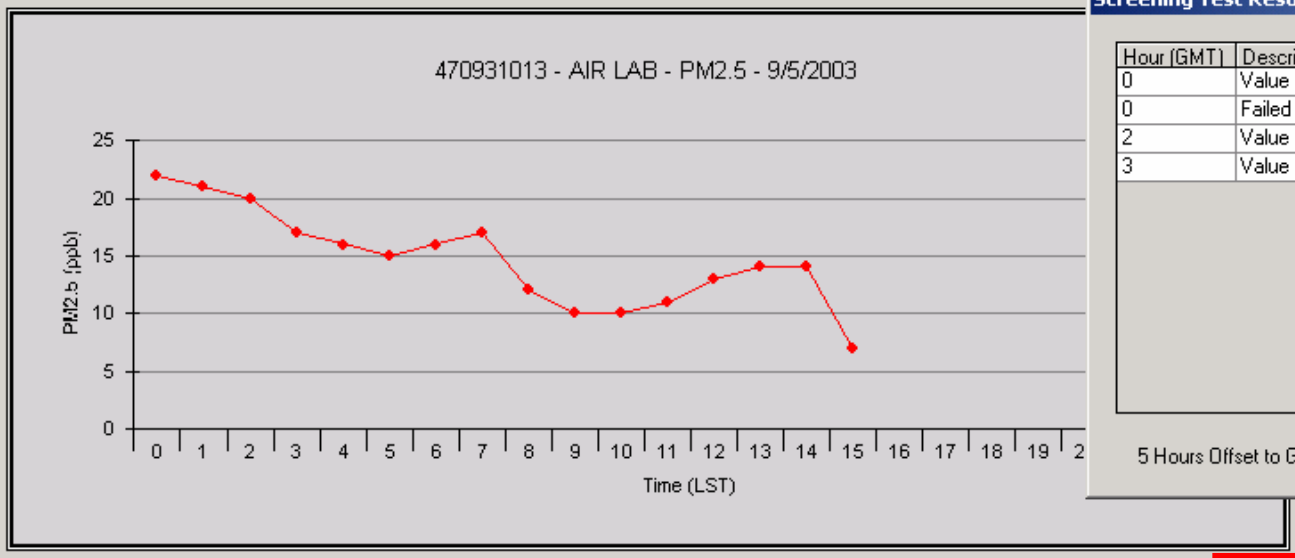
- Checks
  - Are often site-specific
  - May be hour-specific
  - May be automated
- But, data should be graphically reviewed!

Check	Criteria
Maximum	>200 µg/m^3
Minimum	–5 µg/m^3
Rate of change	>50 µg/m^3/hr
Buddy Sites	± 50 µg/m^3 up to 5 sites
Sticking check	>50 µg/m^3 for 5 hours
Co-pollutant	PM_{10}

U.S. EPA, 1999a

Example – Erroneous Data in Tennessee

Edit QC Flags



Screening Test Results

Hour (GMT)	Description
0	Value exceeds MAX suspect value
0	Failed Rate Change Check
2	Value exceeds MAX severe value
3	Value exceeds MAX severe value

5 Hours Offset to GMT

OK

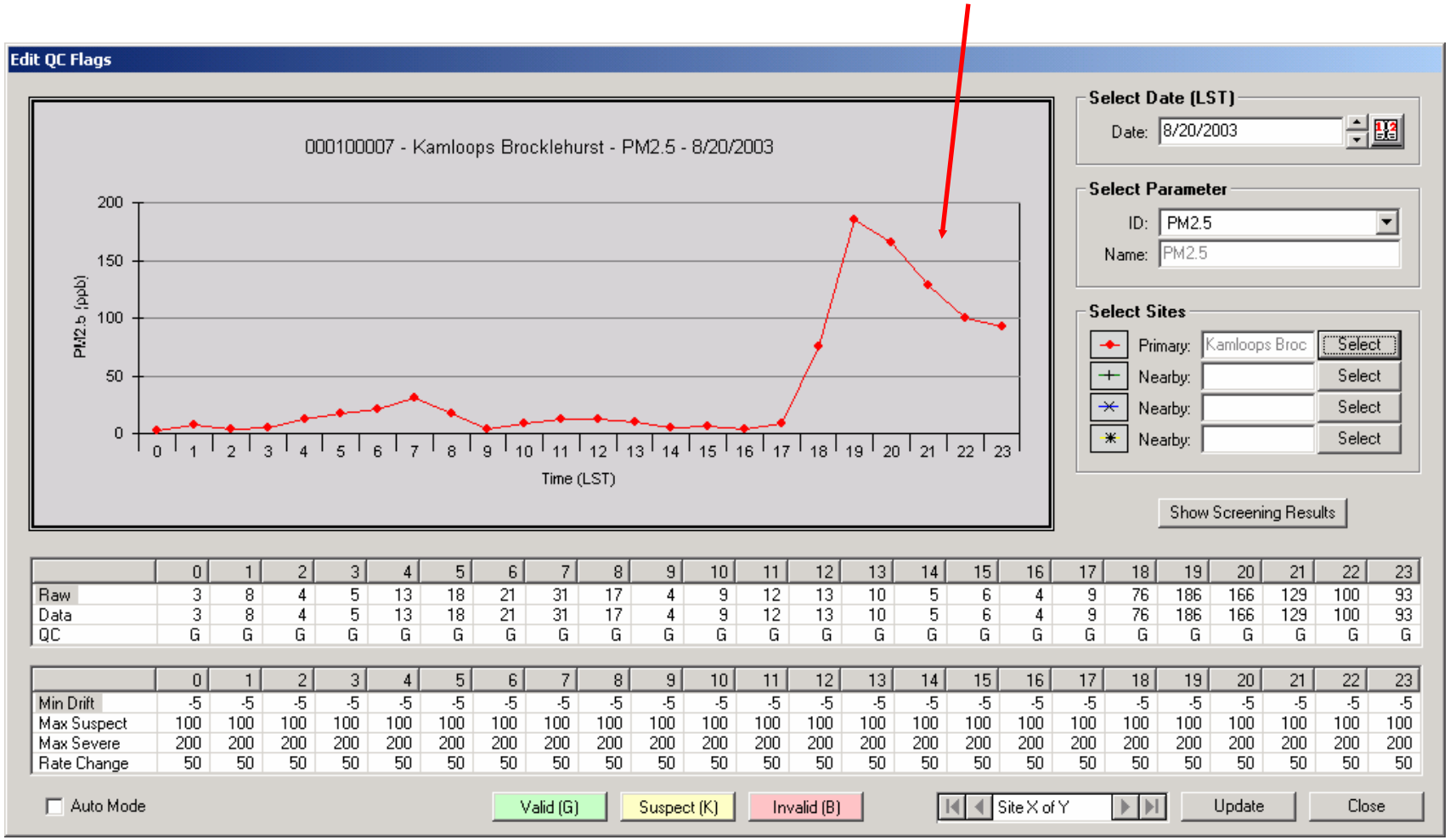
Show Screening Results

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Raw	22	21	20	17	16	15	16	17	12	10	10	11	13	14	14	7	0	0	4	131	30	236	487	13
Data	22	21	20	17	16	15	16	17	12	10	10	11	13	14	14	7	-980	-980	-980	-980	-980	-980	-980	-980
QC	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	B	B	B	B	B	B	B	B

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Min Drift	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5
Max Suspect	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Max Severe	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200
Rate Change	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50

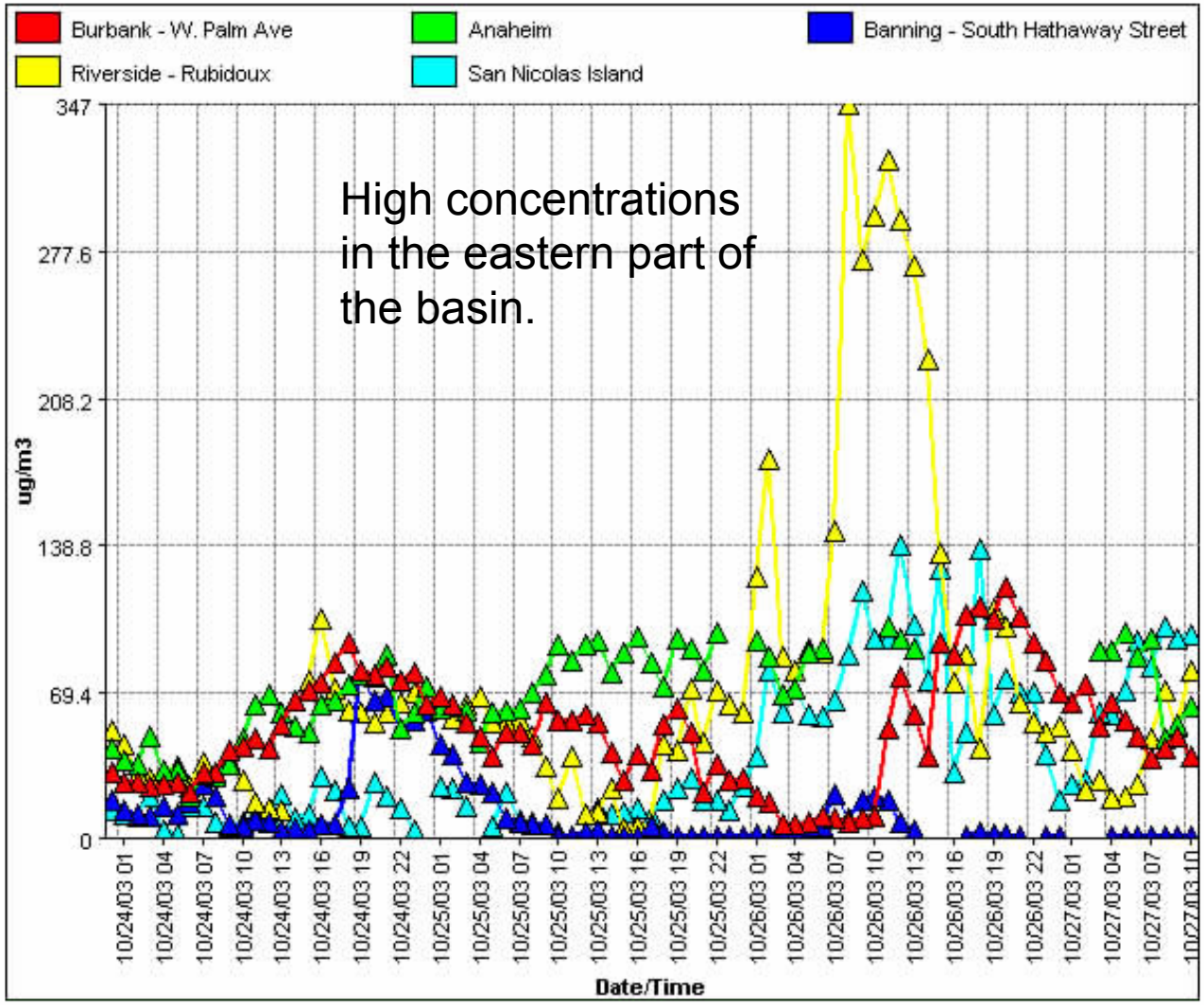
Auto Mode | Valid (G) | Suspect (K) | Invalid (B) | Site X of Y | Update | Close

Examples – Wildfire Events



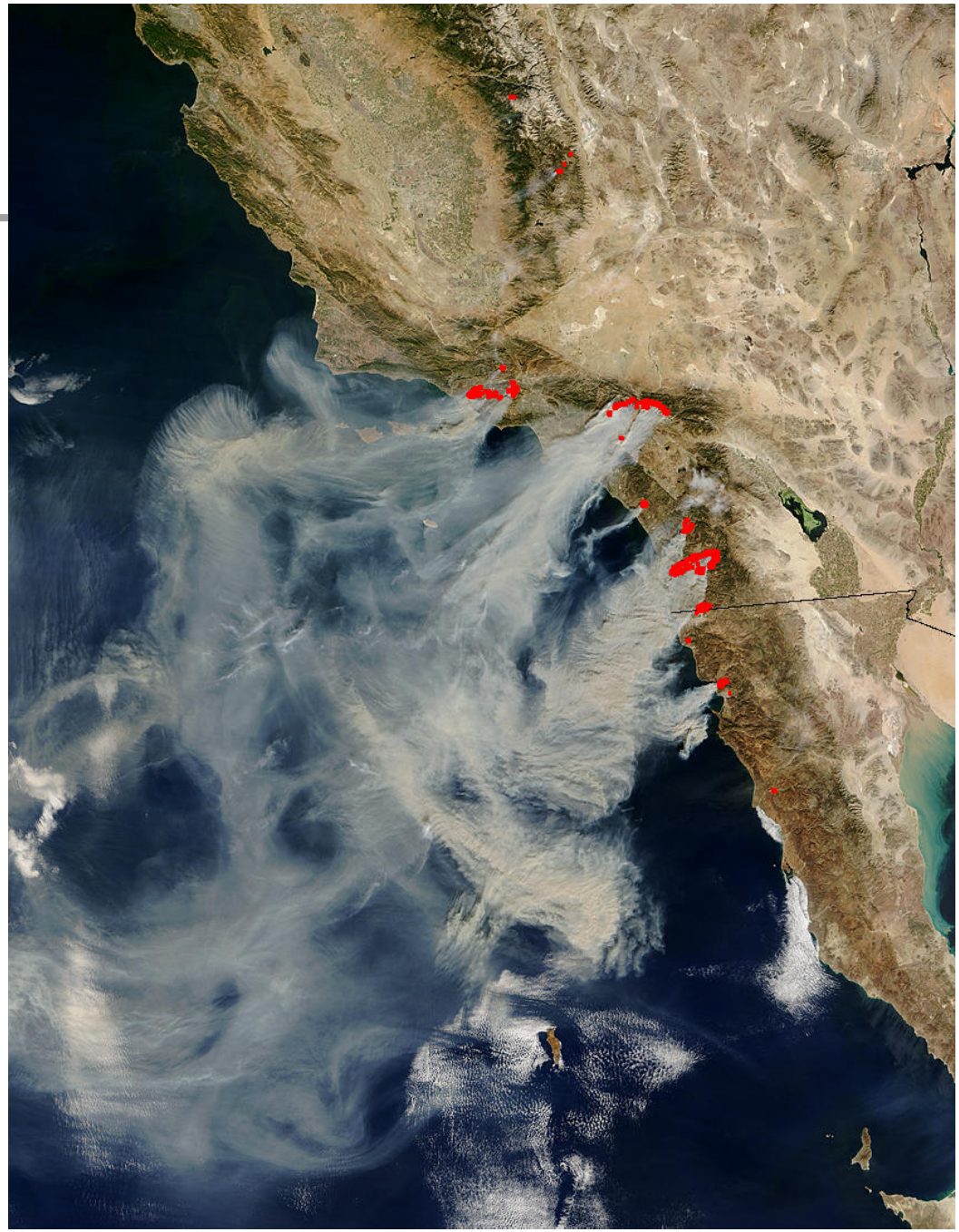
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Raw	3	8	4	5	13	18	21	31	17	4	9	12	13	10	5	6	4	9	76	186	166	129	100	93
Data	3	8	4	5	13	18	21	31	17	4	9	12	13	10	5	6	4	9	76	186	166	129	100	93
QC	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Min Drift	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5
Max Suspect	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Max Severe	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200
Rate Change	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50



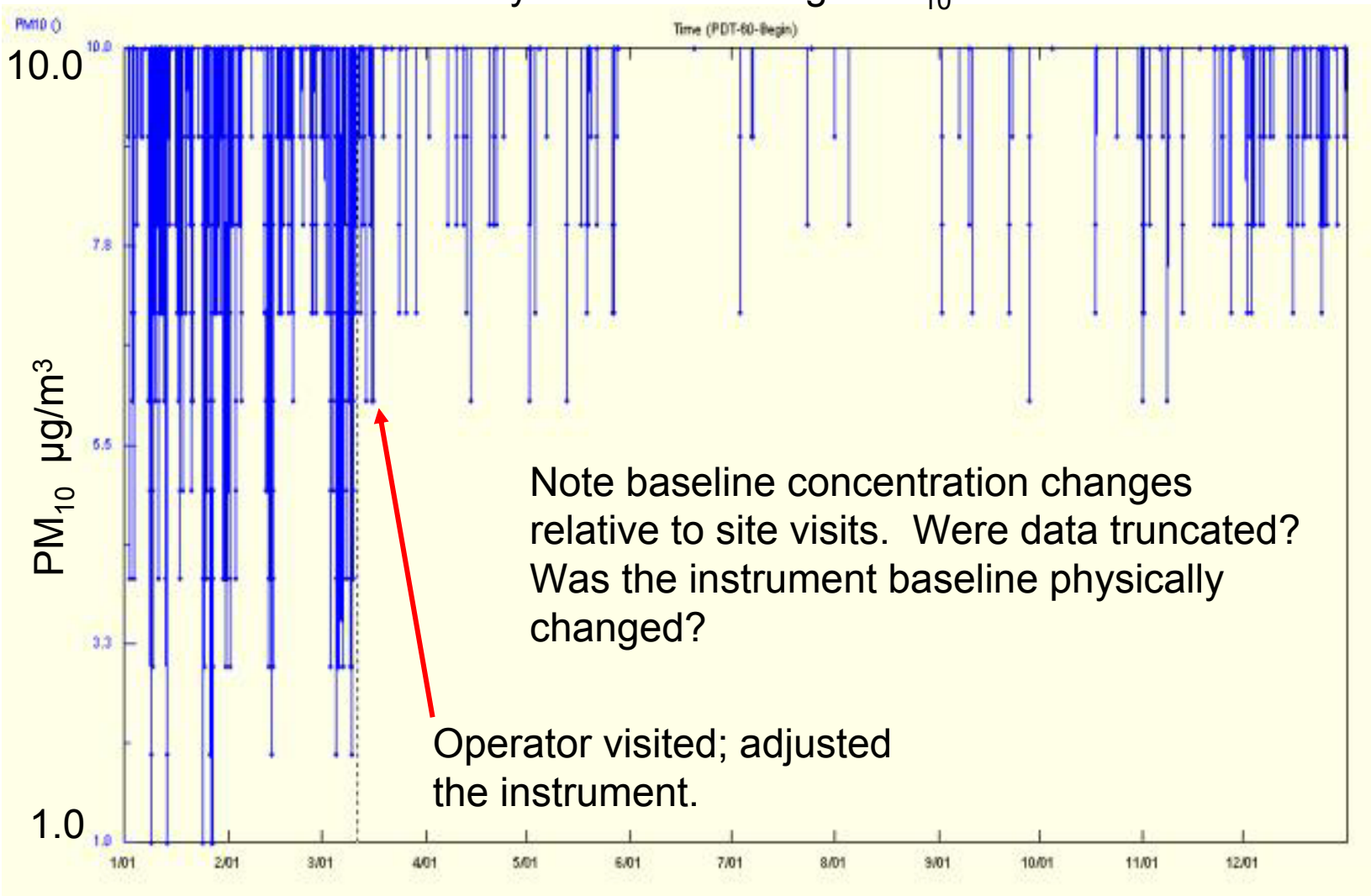
Los Angeles continuous $PM_{2.5}$ mass concentrations on 10/24/03 to 10/27/03 (raw data – USEPA AirNow)

High concentrations are consistent with wildfire smoke as shown on this satellite photo from http://rapidfire.sci.gsfc.nasa.gov/gallery/

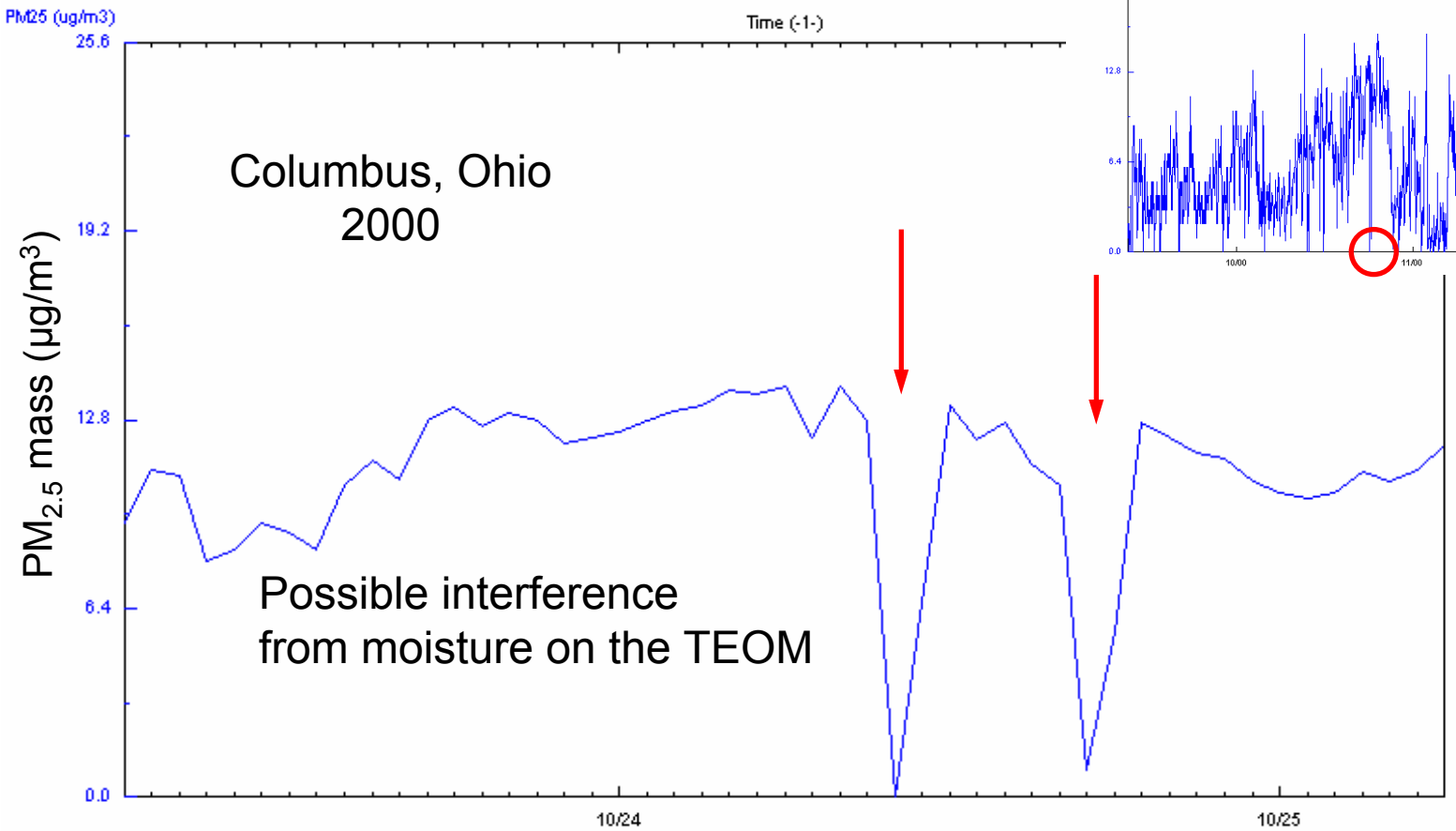


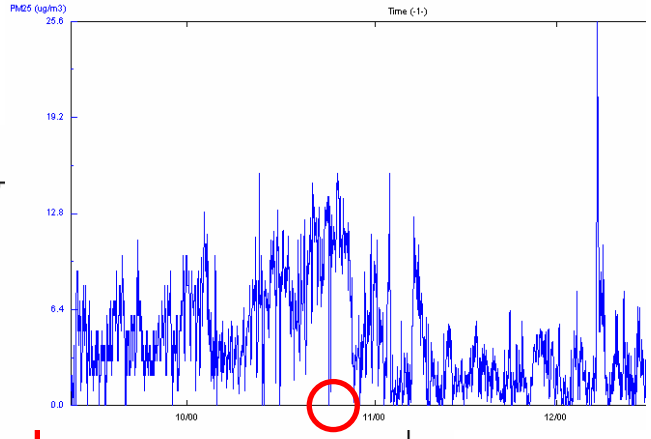
Example – PM_{10} Baseline Changes

1 year 24-hr average PM_{10} data



Example – Odd Low Concentrations





Speciated $PM_{2.5}$ Internal Consistency Checks (1 of 2)

- Check sum of chemical species versus $PM_{2.5}$ mass (multi-elements Al to U + sulfate + nitrate + ammonium ions + OC + EC)
- Check physical and chemical consistency (sulfate vs. total sulfur, soluble potassium versus total potassium, soluble chloride vs. chlorine, b_{abs} versus elemental carbon)
- Balance charge (cations and anions)
- Balance ammonium

Chow, 1998

Speciated $PM_{2.5}$ Internal Consistency Checks (2 of 2)

- Investigate nitrate volatilization and adsorption of gaseous organic carbon (compare front and backup filter concentrations).
- Prepare crude mass balance (sum of geologic mass, combustion-related mass, and sulfate)
- Compare to collocated or near-collocated FRM mass.

PM Consistency Checks and Expectations

Consistency Check	Expectation
Difference between PM_{10} and $PM_{2.5}$*	$PM_{2.5} \leq PM_{10}$
Sum of individual chemical species and $PM_{2.5}$	species sum < $PM_{2.5}$
Ratio of water-soluble sulfate by IC to total sulfur by XRF	~ 3
Ratio of chloride by IC to chlorine by XRF	< 1
Ratio of water-soluble potassium by AAS to total potassium by XRF	< 1
b_{abs} compared to elemental carbon	good correlation

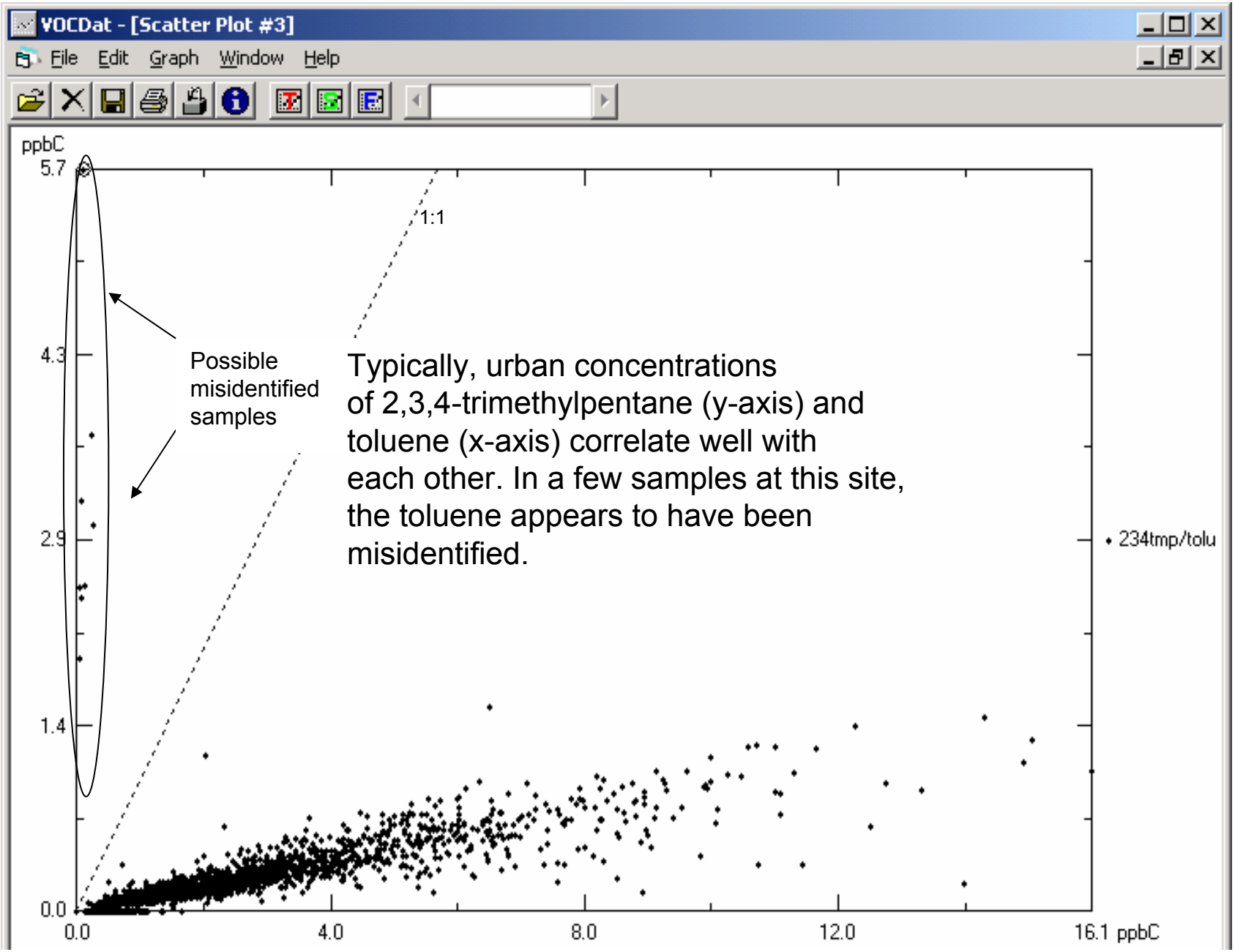
IC = ion chromatography
XRF = energy dispersive X-ray fluorescence
AAS = atomic absorption spectrophotometry

* Dichotomous data may be an exception to this check

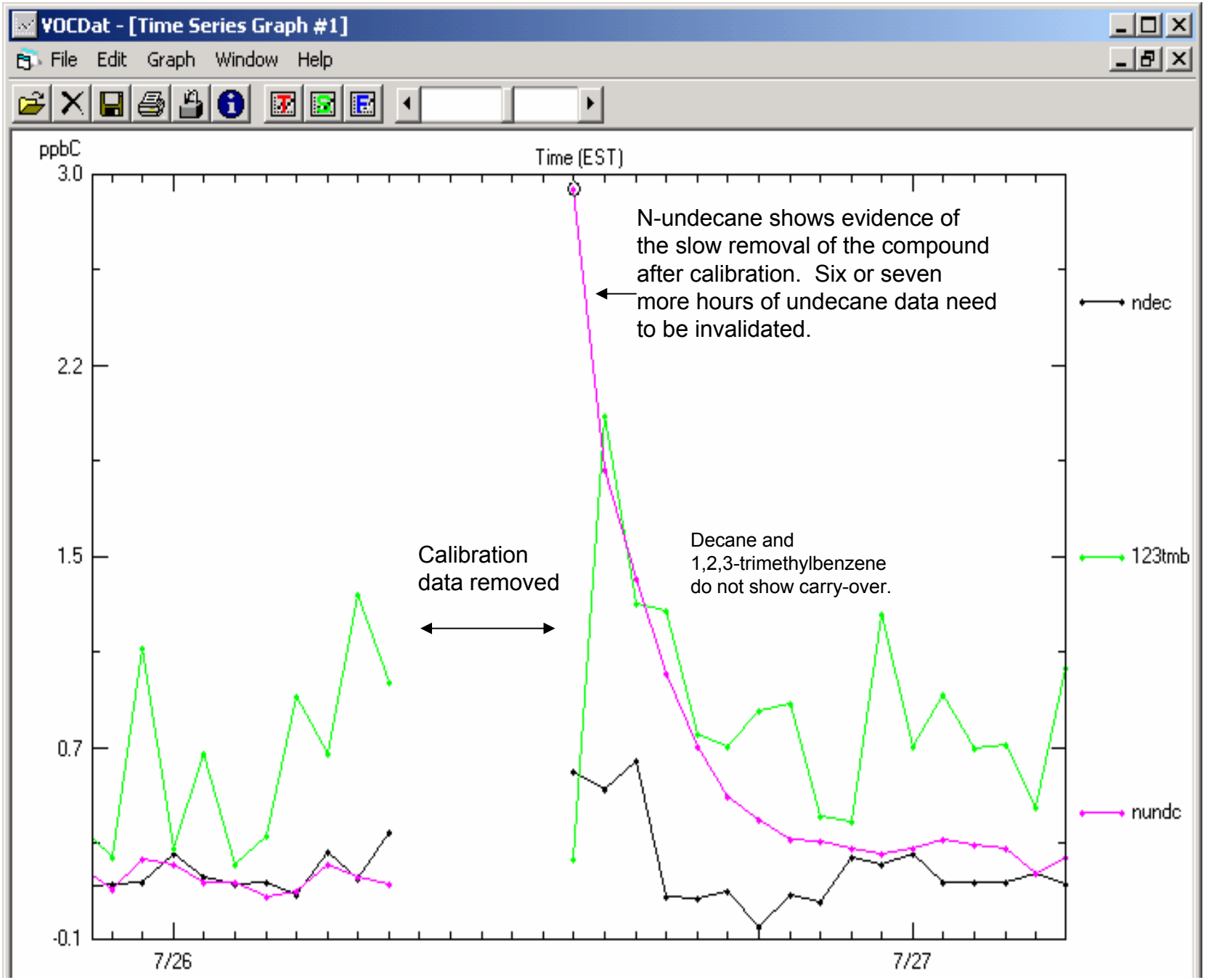
Chow, 1998

PAMS Data Validation
Example Screening Criteria

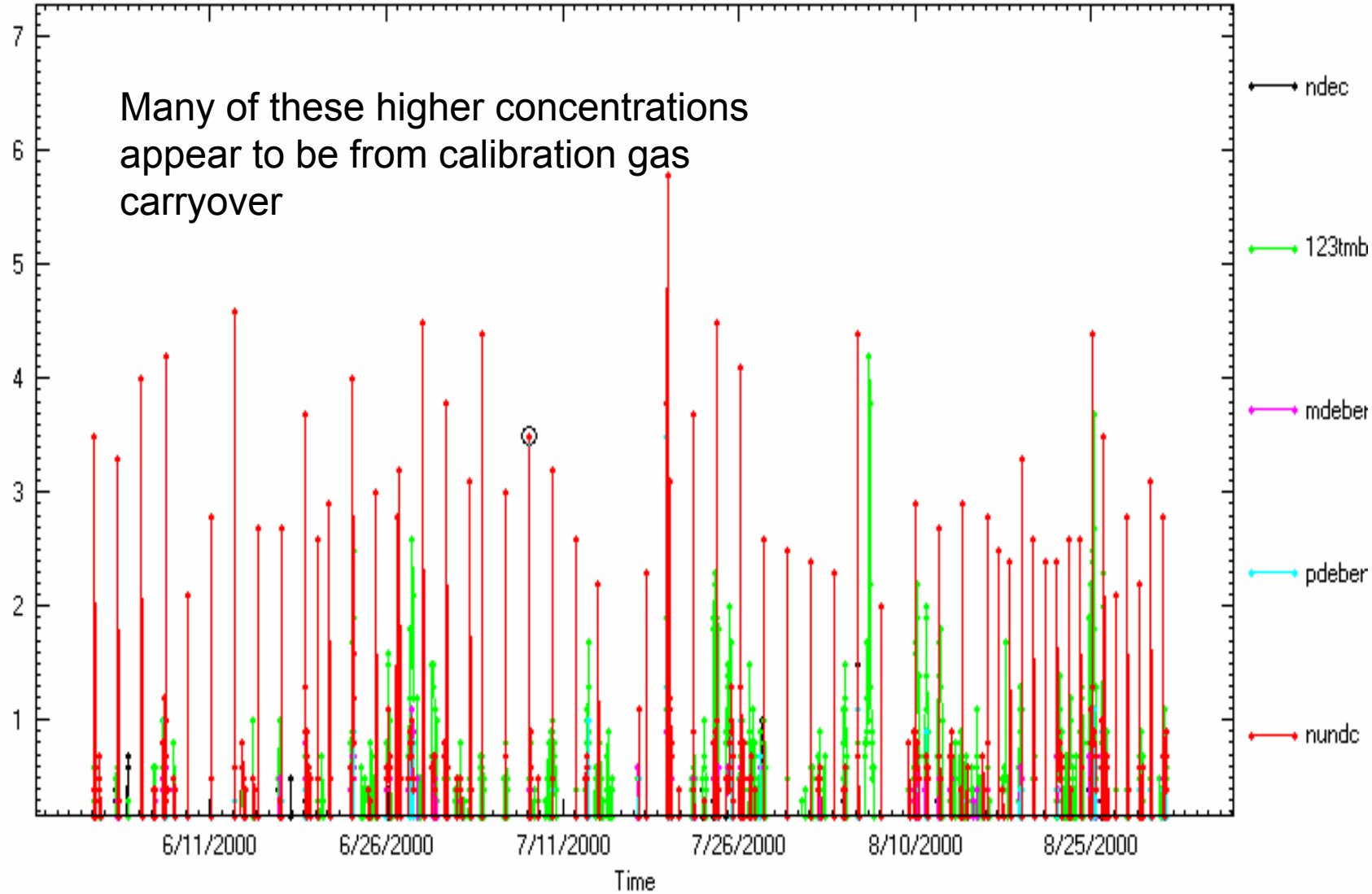
- Check that abundant hydrocarbons (e.g., acetylene, ethane, propane, n-butane, i-pentane, n-pentane, n-hexane, benzene, toluene, and m-&p-xylenes) are present in the same samples.
  - This check helps identify “missing” abundant hydrocarbons.
  - Set the screening concentrations sufficiently higher than the detection limit (e.g., 10 times) to limit the number of data “failing” these criteria.
- Check that the data meet expected relationships.
  - For example, n-pentane concentrations are usually less than i-pentane concentrations.
  - Other possible screens include o-xylene < m-&p-xylenes and benzene < toluene.
- Check for unusual sample compositions including
  - ethane concentration < 2 ppbC but benzene > 2 ppbC (may indicate cold trap problems in auto-GC)
  - unidentified fraction of TNMOC > 50% (the less known about a sample’s composition, the less useful the sample).



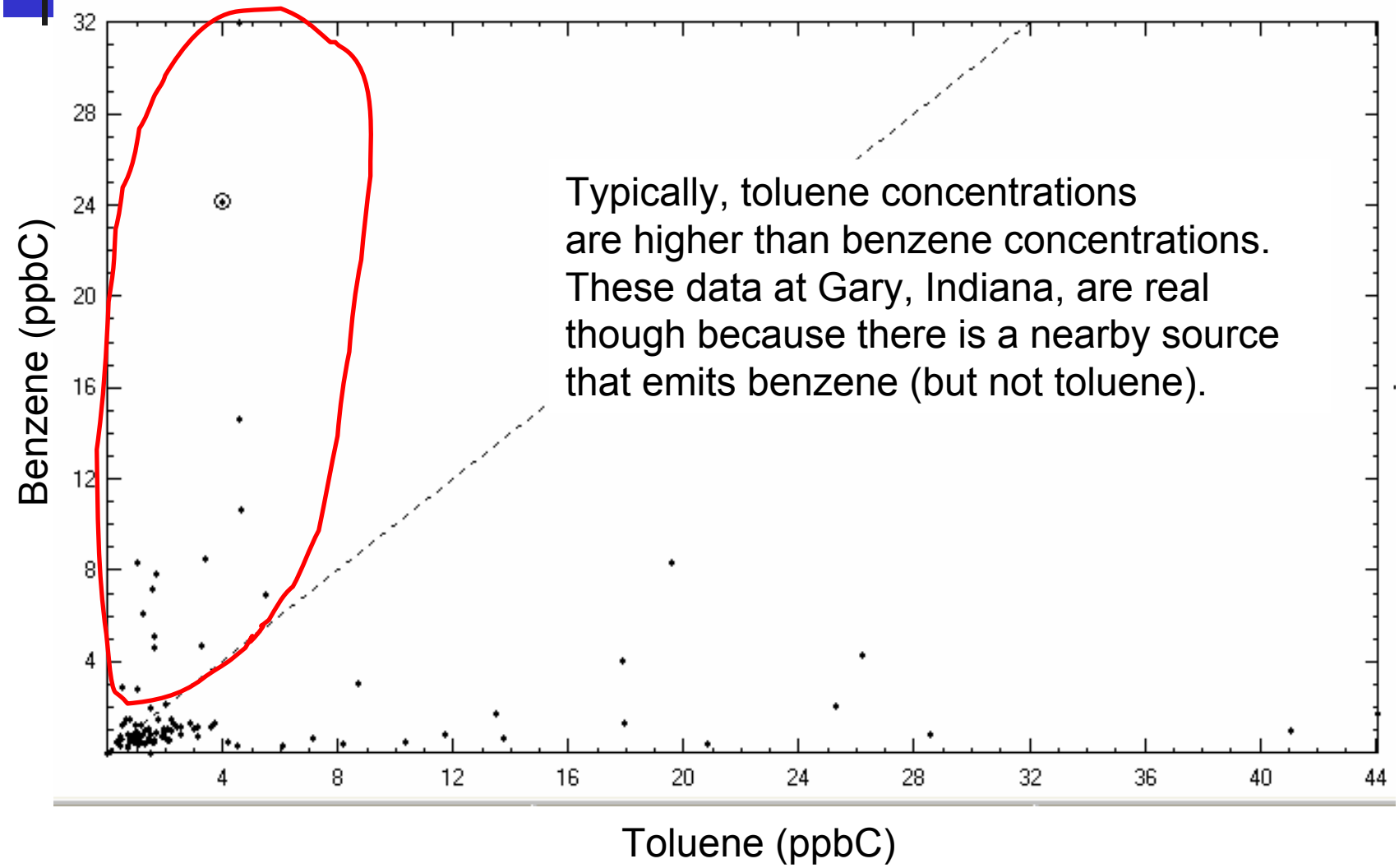
Typically, urban concentrations of 2,3,4-trimethylpentane (y-axis) and toluene (x-axis) correlate well with each other. In a few samples at this site, the toluene appears to have been misidentified.


VOCDat - [Time Series Graph #1]
File Edit Graph Window Help
ppbC
Time (EST)
3.0
2.2
1.5
0.7
-0.1
7/26
7/27
N-undecane shows evidence of the slow removal of the compound after calibration. Six or seven more hours of undecane data need to be invalidated.
Calibration data removed
Decane and 1,2,3-trimethylbenzene do not show carry-over.
ndec
123tmb
nundc

Another Carryover Example

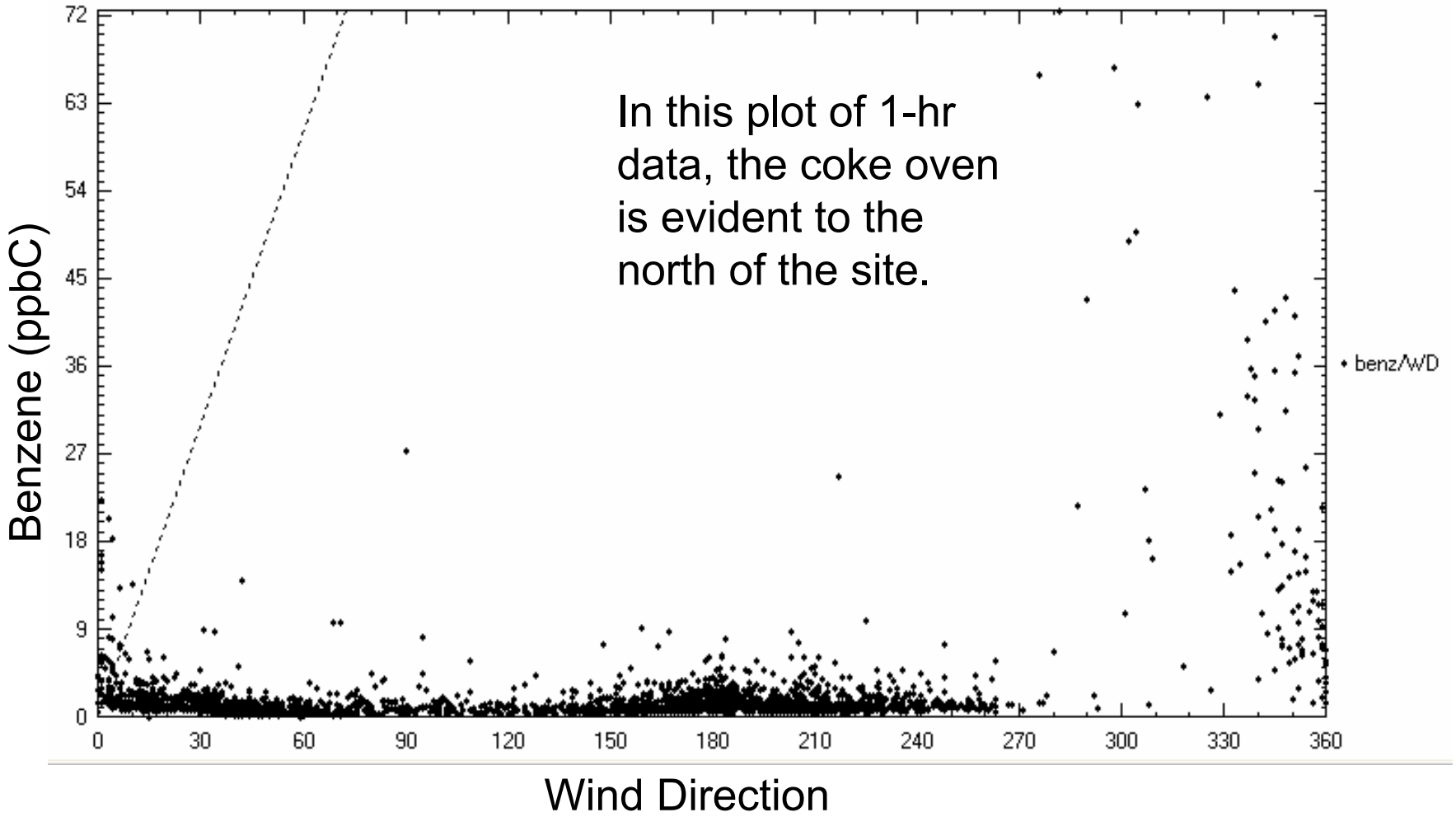


High Benzene (1 of 2)

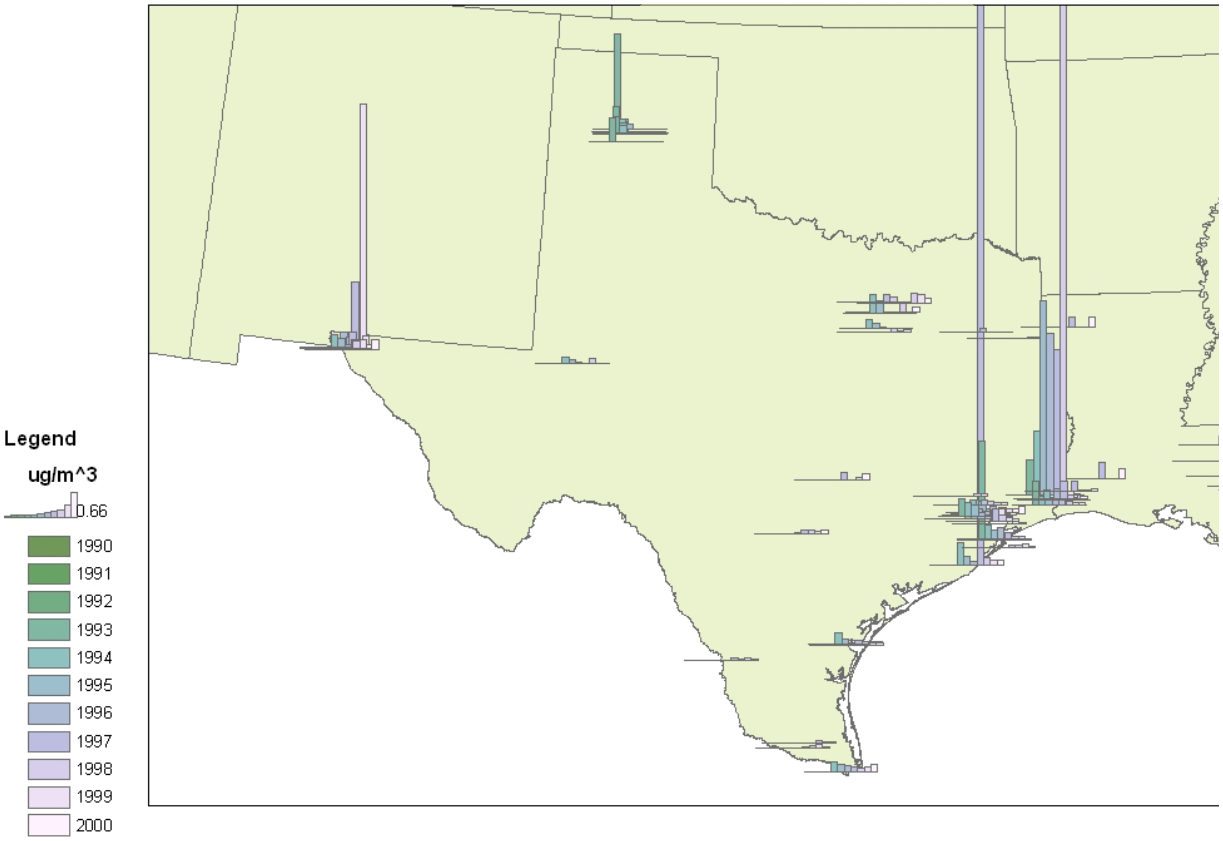


Typically, toluene concentrations are higher than benzene concentrations. These data at Gary, Indiana, are real though because there is a nearby source that emits benzene (but not toluene).

High Benzene (2 of 2)



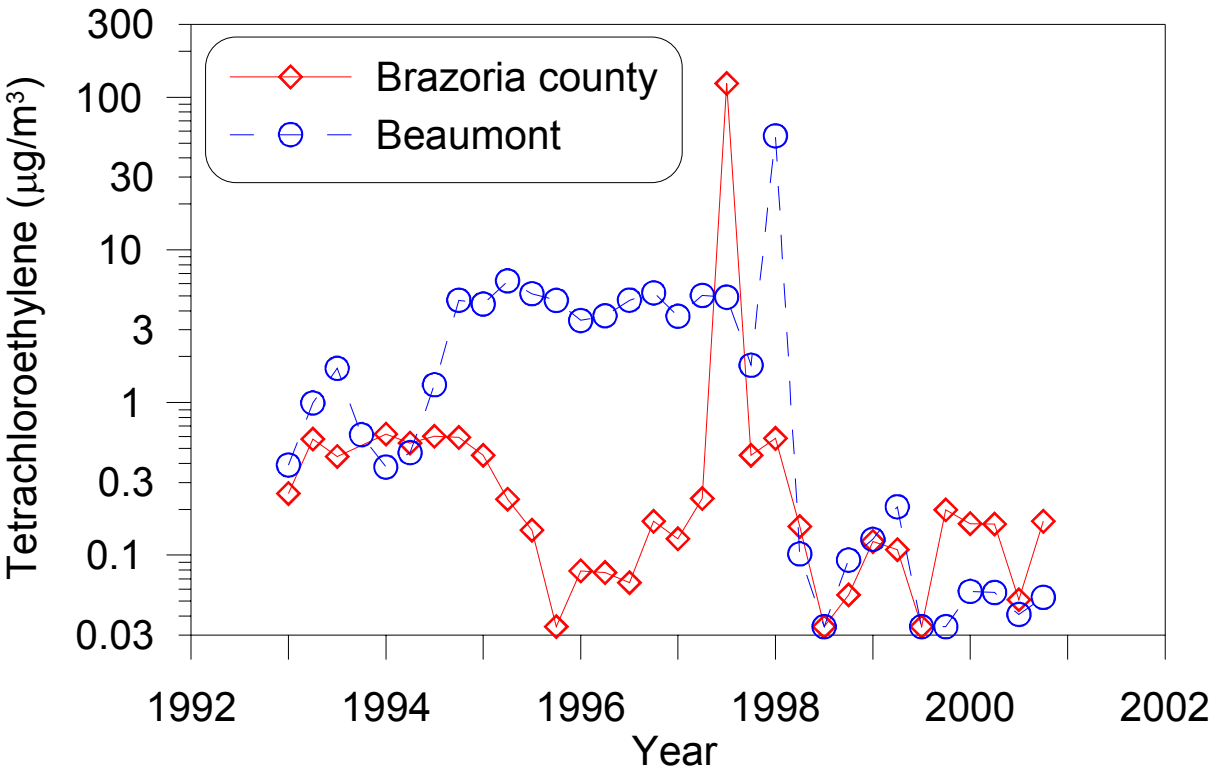
Air Toxics:
Characterizing Spikes – Tetrachloroethylene



Annual averages of tetrachloroethylene

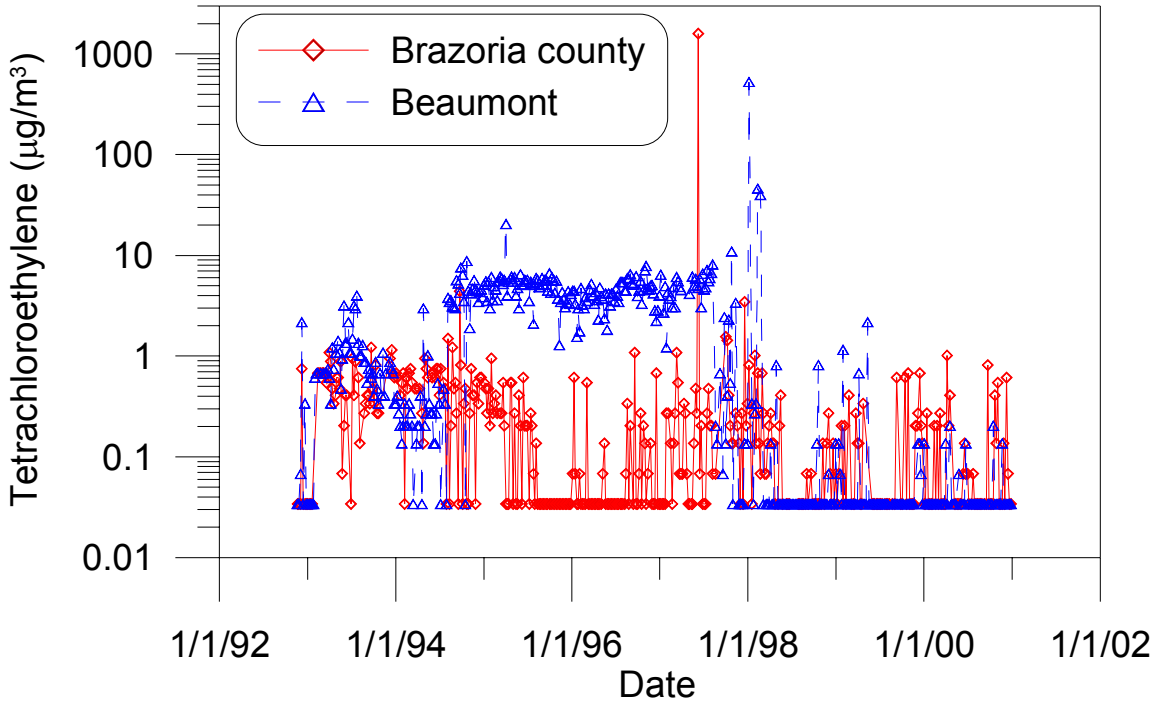
- Two sites on the Gulf of Mexico have spikes in tetrachloroethylene concentrations in 1997 and 1998.
- Are these spikes real or measurement artifacts?
- Should these data be used to understand trends or health risks?

Tetrachloroethylene Spikes (1 of 4)



- Time series of seasonal average concentrations of tetrachloroethylene at two sites in Texas.
- Both sites have a single spike in concentration that is at least ten times the typical seasonal average.

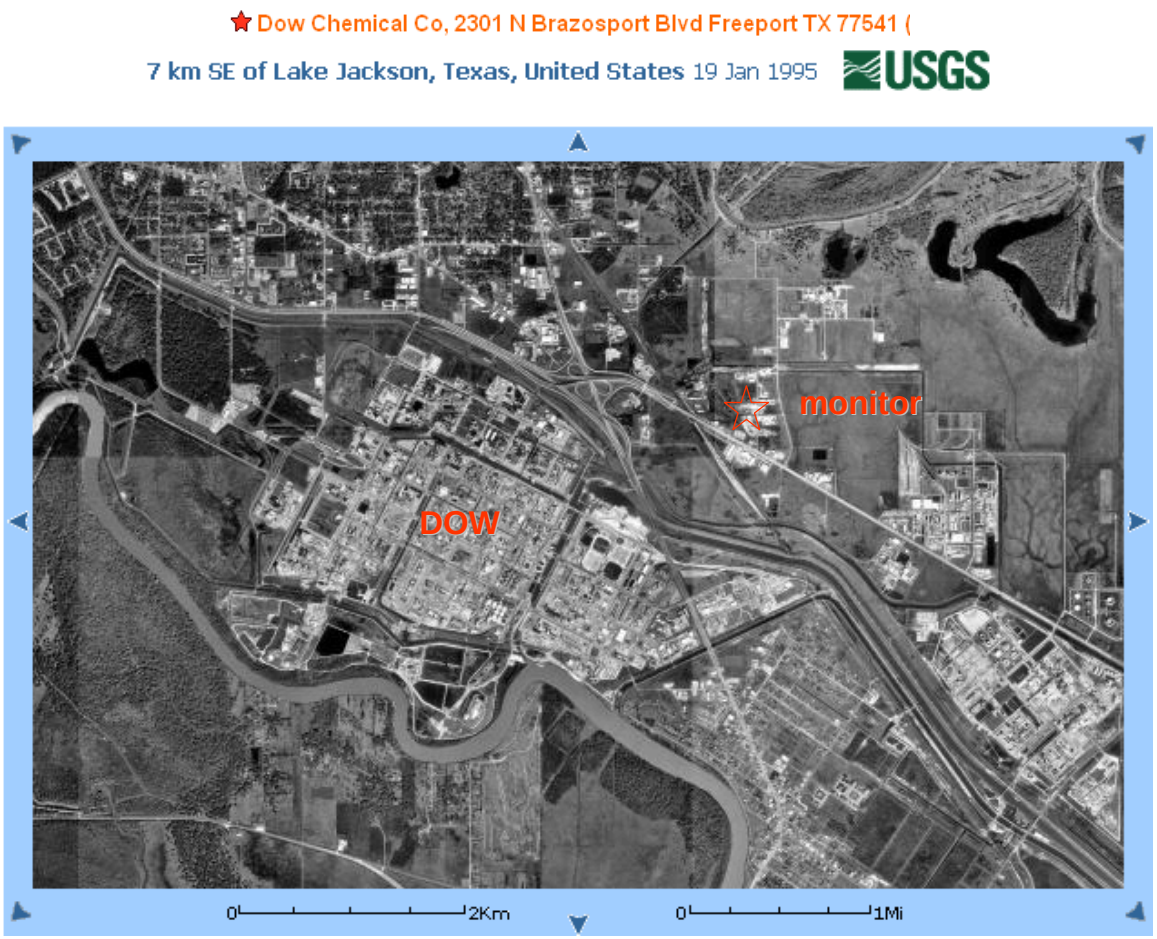
Tetrachloroethylene Spikes (2 of 4)



24-hr measurements of tetrachloroethylene at the two sites in Texas.

- The Brazoria County spike on June 9, 1997, was more than three orders of magnitude larger than any other in the site's history.
- The Beaumont site had four years with concentrations at 5 $\mu g/m^3$ before decreasing in 1997. Three spikes in 1998 were larger than 50 $\mu g/m^3$. After that, concentrations were usually below 0.1 $\mu g/m^3$.

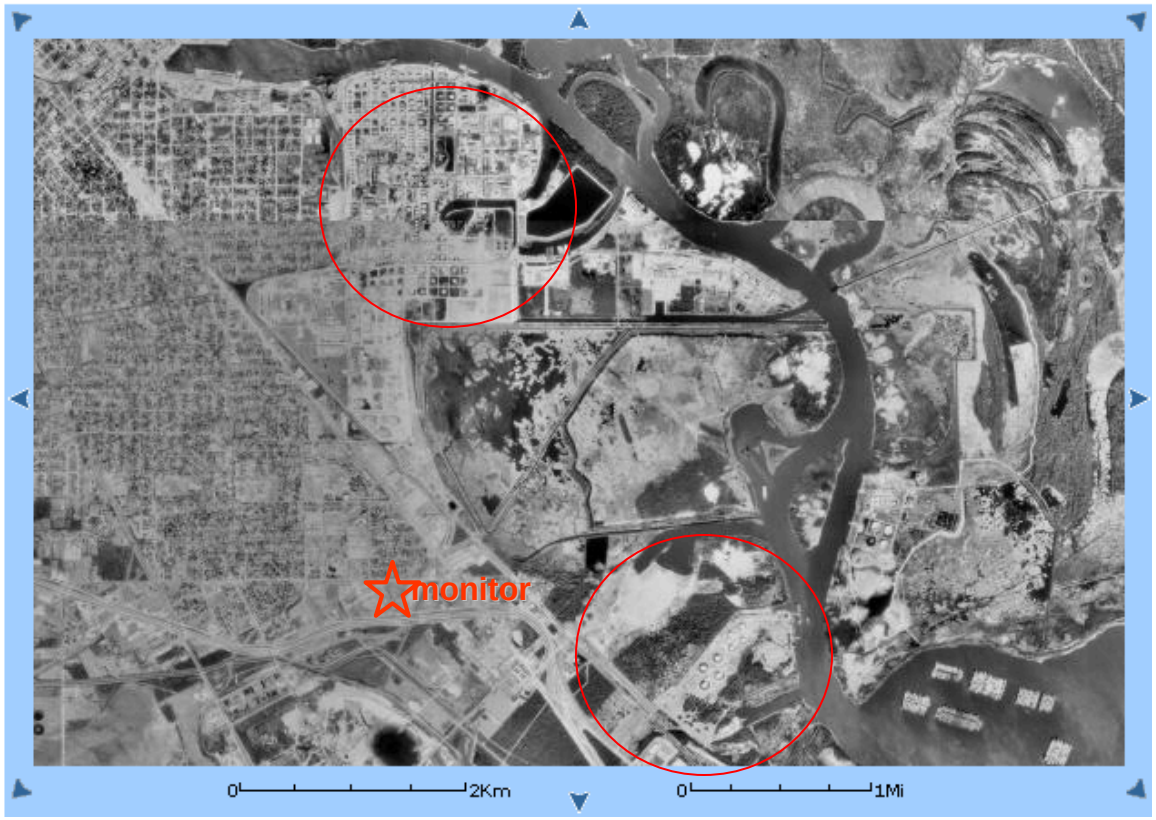
Tetrachloroethylene Spikes (3 of 4)



- The Brazoria County site is located within 1 km of a Dow Chemical Company industrial facility.
- The TRI documents that this site has fugitive releases of tetrachloroethylene.
- 17,000 pounds of tetrachloroethylene were released in 1997.

Tetrachloroethylene Spikes (4 of 4)

8 km SE of Beaumont, Texas, United States 22 Feb 1989 USGS



Note date of imagery, over 13 years old

- The Beaumont site is located within 2 miles of liquid storage tanks to the north, east, and south.
- The NEI documents oil-tanking operations 3.5 miles from the monitoring station.
- The TRI showed that the Exxon facility two miles to the north of the monitor (circled) released 9,000 pounds of tetrachloroethylene in 1998.

Spikes Summary (1 of 3)

- The Brazoria County seasonal spike was an anomaly
  - No other measurement at that site was nearly that high.
  - No other co-measured species were abnormally high.
  - Rough calculations indicate the nearby source could account for a spike of that magnitude (3000 pounds).
  - However, sampling or analytical error may also explain the spike.
  - TCEQ does not have readily available records of upsets or emissions older than five years to provide additional information (per Dick Flannery, TCEQ).
  - This spike should not be used for understanding trends, because it is completely atypical for this site.

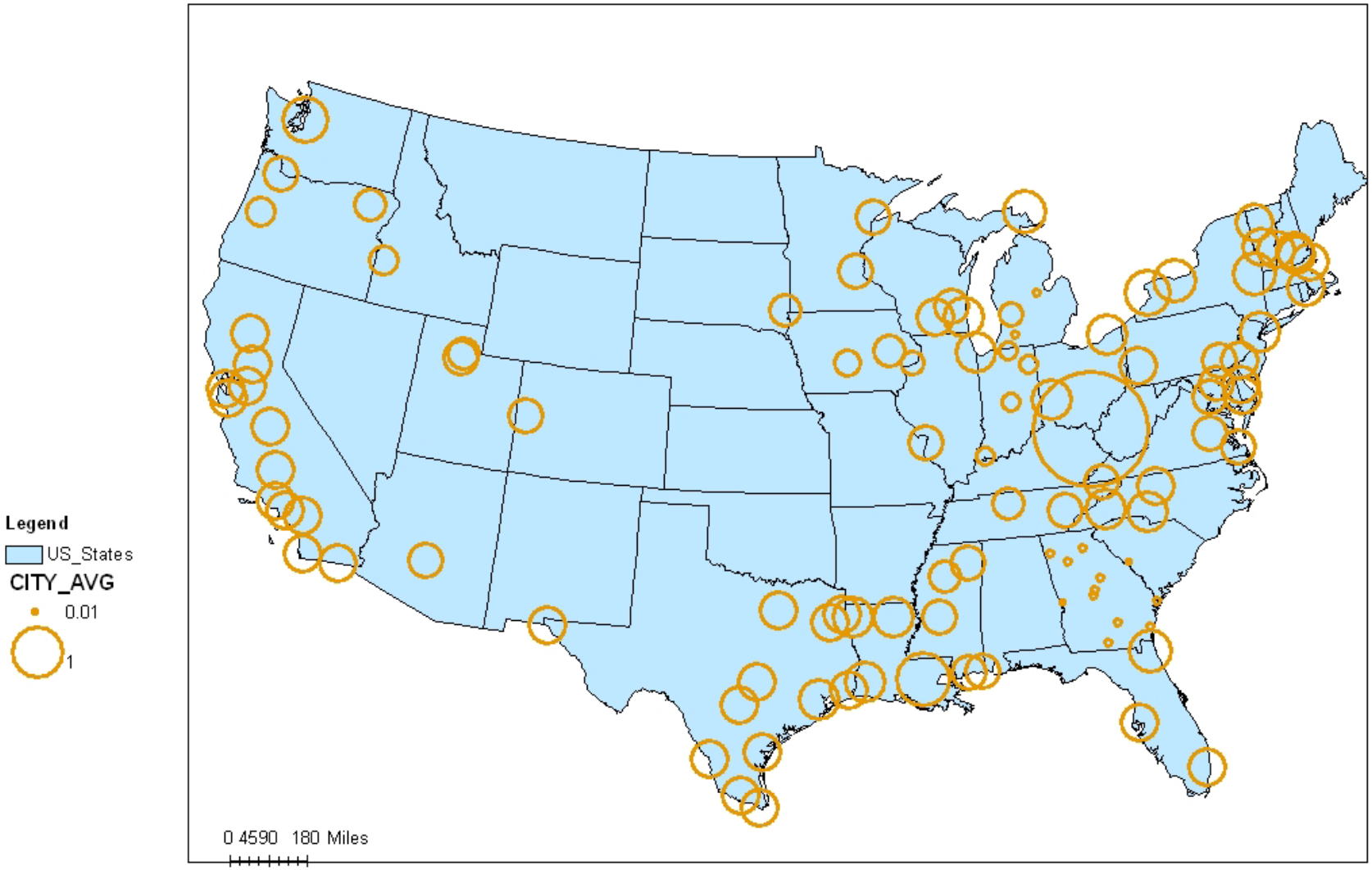
Spikes Summary (2 of 3)

- The Beaumont seasonal spike is not anomalous
  - Three measurements were ten times normal concentrations. However, only tetrachloroethylene concentrations were high; no other co-measured species were abnormally high.
  - Concentrations at the site were elevated relative to typical levels from 1994-1997.
  - This seasonal average concentration should be used for both risk and trend assessment, since it likely reflects real concentrations.

Spikes Summary (3 of 3)

- Characterizing spikes requires significant work
  - Identifying the spikes is straightforward using visual plots of the data (e.g., maps or time series).
  - Spikes caused by analytical or sampling error may have anomalous concentrations of other species.
  - Real spikes in ambient concentrations are likely due to nearby point sources.
  - A combination of maps, the TRI, and local knowledge is likely required (but may not be sufficient) to explain spikes in ambient concentrations.
  - Fugitive emission/upsets data are needed!

Visualization Is Key!



Carbon Tetrachloride – annual averages circa 2004

Data Validation Summary

- For pollutant data validation,
  - Understand formation, emissions, and transport
  - Establish screening criteria to identify potentially suspect data
  - Investigate suspect data
  - Invalidate data only if there is sufficient evidence
- Data validation is very important!

Resources

- Operator knowledge
- Previous documentation for the site and past data validation results
- EPA guidance documents (available on AMTIC web site)
- Workbooks (e.g., PAMS and $PM_{2.5}$ Data Analysis Workbooks)
- Web sites (e.g., IMPROVE, EPA Supersite)
- Journal articles and conference presentations (e.g., *Atmospheric Environment, Environmental Science and Technology*, Air and Waste Management Association)
- Academia

Key Internet Sites

- Ambient Monitoring Technology Information Center: http://www.epa.gov/ttn/amtic/
- IMPROVE QA/QC: http://vista.cira.colostate.edu/improve/Data/QA_QC/qa_qc_Branch.htm
- EPA Quality Assurance: http://www.epa.gov/oar/oaqps/qa/index.html#back
- PAMS Data Analysis Workbook (old):
- http://www.epa.gov/oar/oaqps/pams/analysis/
- EPA Supersite Overview: http://www.epa.gov/ttn/amtic/supersites.html