Data Validation Overview

Hilary R. Hafner
Sonoma Technology, Inc.
Petaluma, CA

Presented at
2009 National Ambient Air Monitoring Conference
Breakout Session: Data Validation
Nashville, TN

November 2-5, 2009

STi
Sonoma Technology, Inc.
Air Quality Research and Innovative Solutions

Outline

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

Data validation is defined as the process of determining the quality of observations and identifying their validity

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 result from erroneous data values.
- Accurate information helps you respond to community concerns.

Objectives and Benefit

- The objectives of the data validation process are to
  - produce a database with values that are of a known quality
  - evaluate the internal, spatial, temporal, and physical consistency of the data
  - identify errors, biases, and outliers
- The benefit for the data analyst who performs data validation is enhanced familiarity with the unique features of the data set.

Data Validation Levels

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

Level 1: Field and Laboratory Checks

- Verify computer file entries against data sheets.
- Flag samples when significant deviations from measurement assumptions have occurred.
- Eliminate values for measurements that are known to be invalid because of instrument malfunctions.
- Replace data from a backup data acquisition system in the event of primary system failure.
- Adjust measurement values for quantifiable calibration or interference bias.

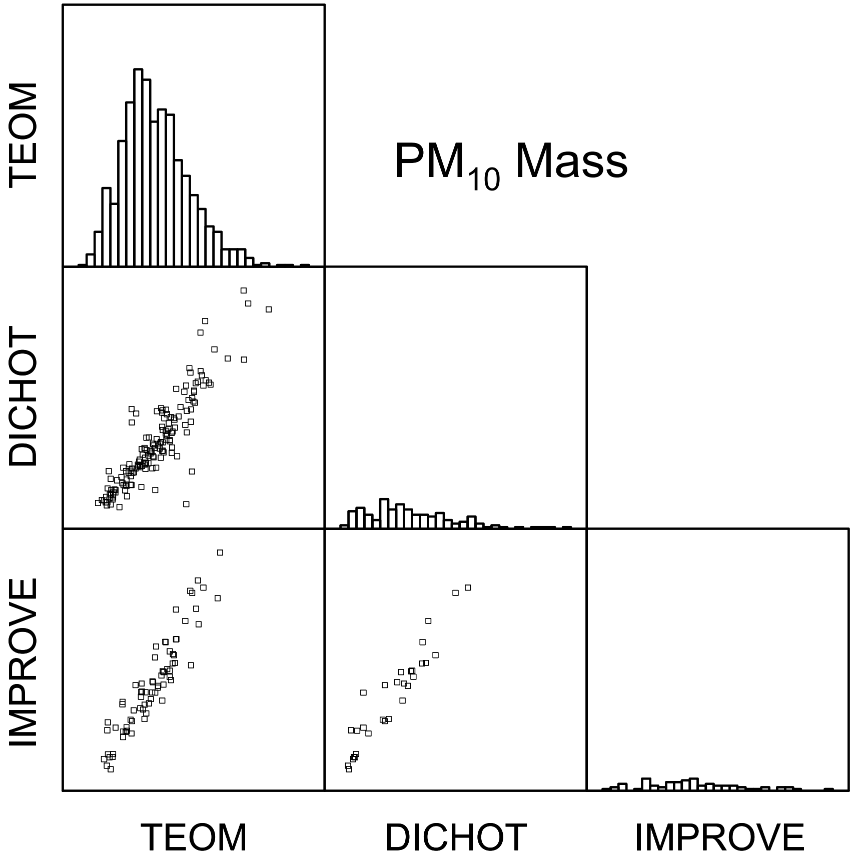
Level II: Internal Consistency Checks

- Inspect time series to see if concentrations vary by time of day, day of week, and season as expected.
- Compare pollutant concentrations for expected relationships.
- 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."
>
> *Judy Chow, Desert Research Institute*

Level II+: External Consistency Checks

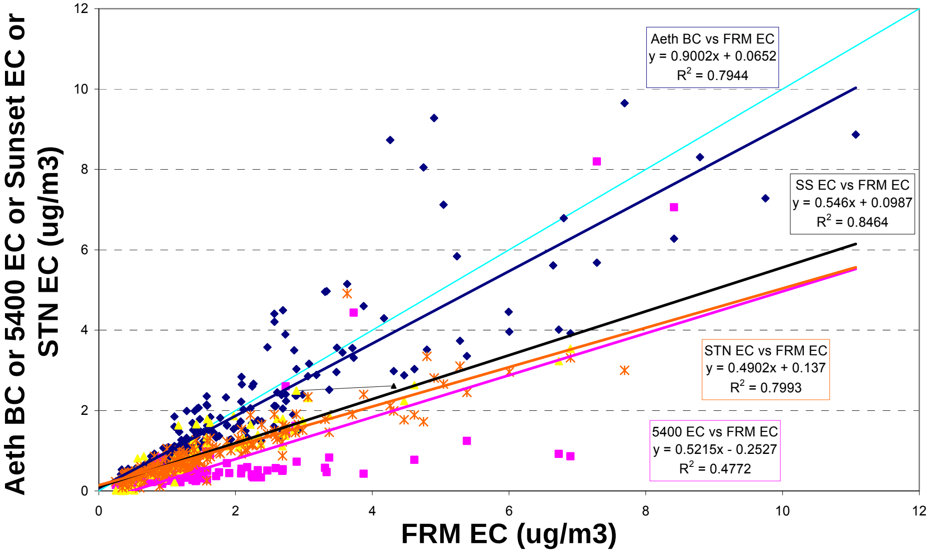
- Compare collocated measurements.
- Compare relationships (e.g., temporal, among species) observed in the current data set to relationships observed at other sites or in previous years.
- Compare pollutant concentrations to meteorology.



General Approach to Data Validation

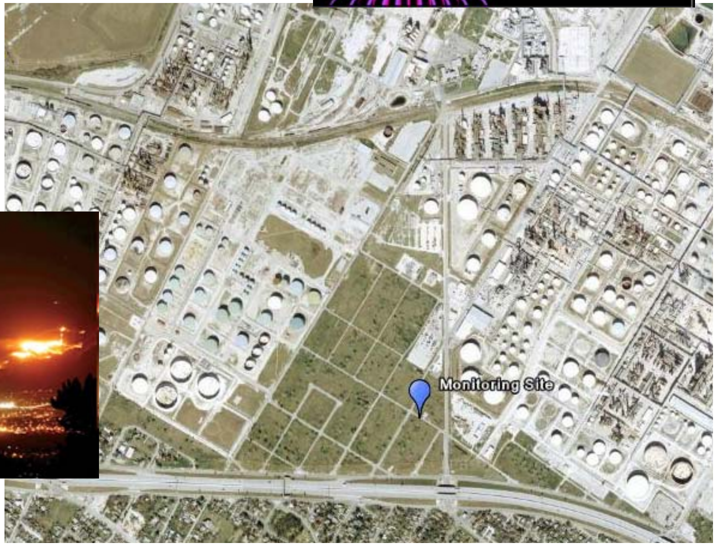
- Look at your data.
- Manipulate your data—sort it, graph it, map it—so it begins to tell a story.
- Often, important issues with or errors in data will become apparent only after the data are being used for something.

- Examples
  - Scatter plots
  - Time series plots
  - Fingerprint plots
  - Box whisker plots
  - Summary statistics



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

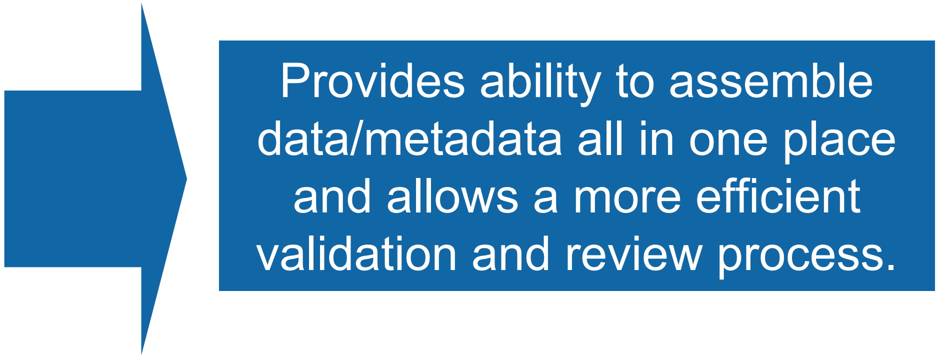


Steps in Data Validation

1. Assemble data set.
2. Apply general screening criteria.
3. Prepare and inspect summary statistics for unrealistic maxima and minima and other factors.
4. Investigate internal consistency.
5. Flag data and document data modifications.
6. Perform spatial and temporal comparisons, compare data from different instruments (i.e., begin Level II).

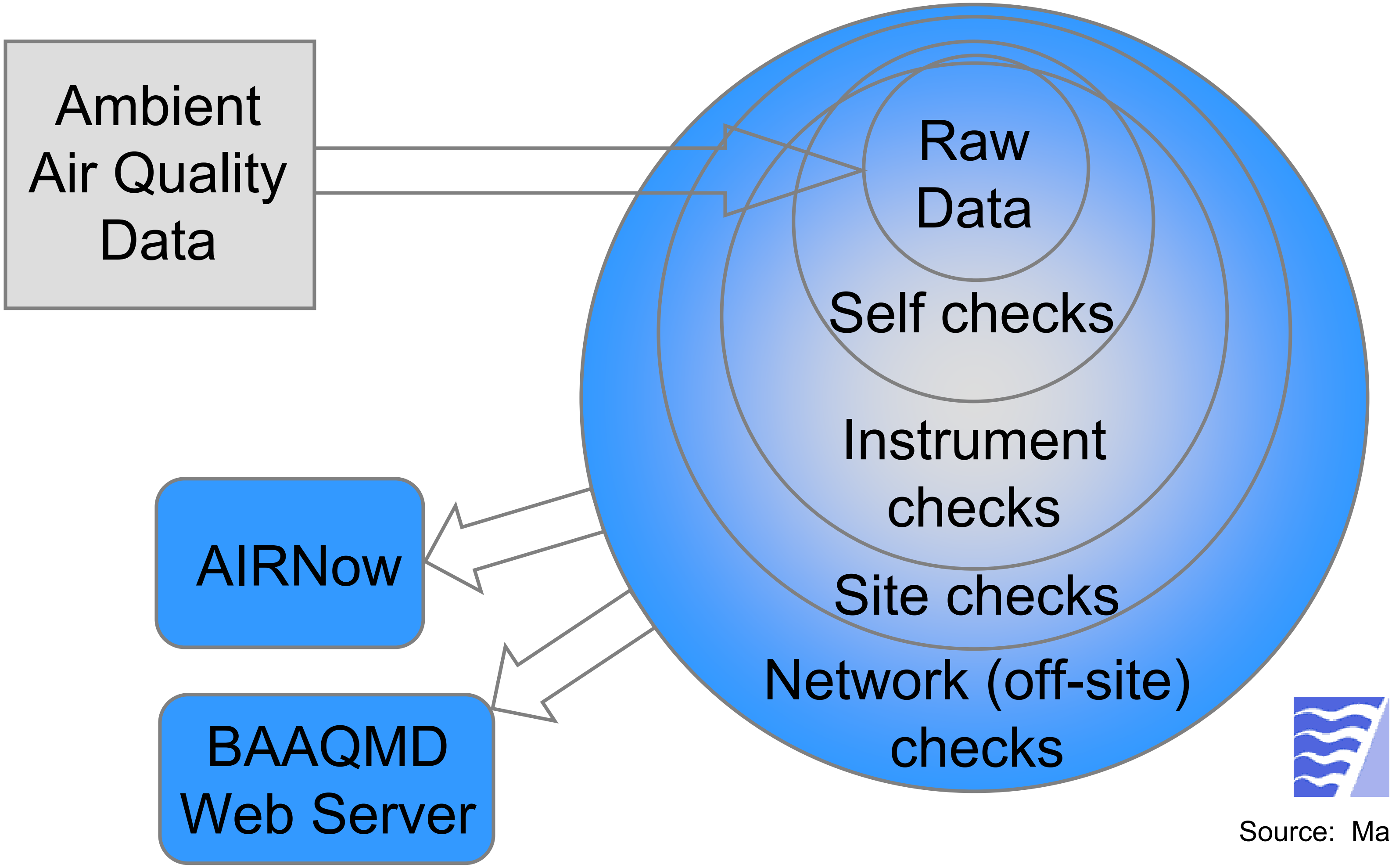
How Has the Data Validation Process Changed?

- More data being collected
- New instruments
- Better computing
- Better tools
  (e.g., visualization)
- Improved communication
  (allows remote access and frequent review)



Provides ability to assemble data/metadata all in one place and allows a more efficient validation and review process.

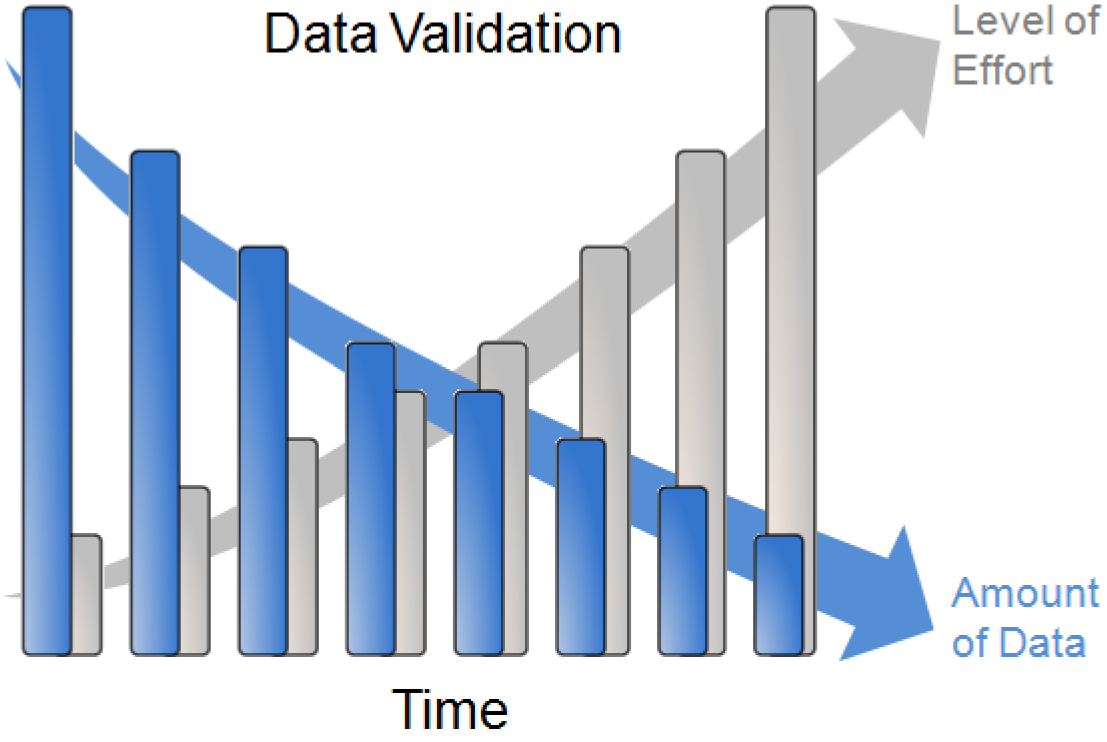
Automated Quality Assurance Checks



Source: Mark Stoelting

Changes Should Lead to…

- More timely review
- Better data capture
- Higher quality data



Importance of Supplemental Data
Examples

- Sample collection specifications – sampler type, sampling media, inlet type, etc.
- Sampling location description – nearby sources, topography, distance to roadways, etc.
- Audit, blank collection, and collocated sampler descriptions (accuracy and precision)
- Sample analysis and instrument calibration descriptions
- Replication and duplication of sample results
- Sample schedule

Screening Criteria:
Singling Out Unusual Data

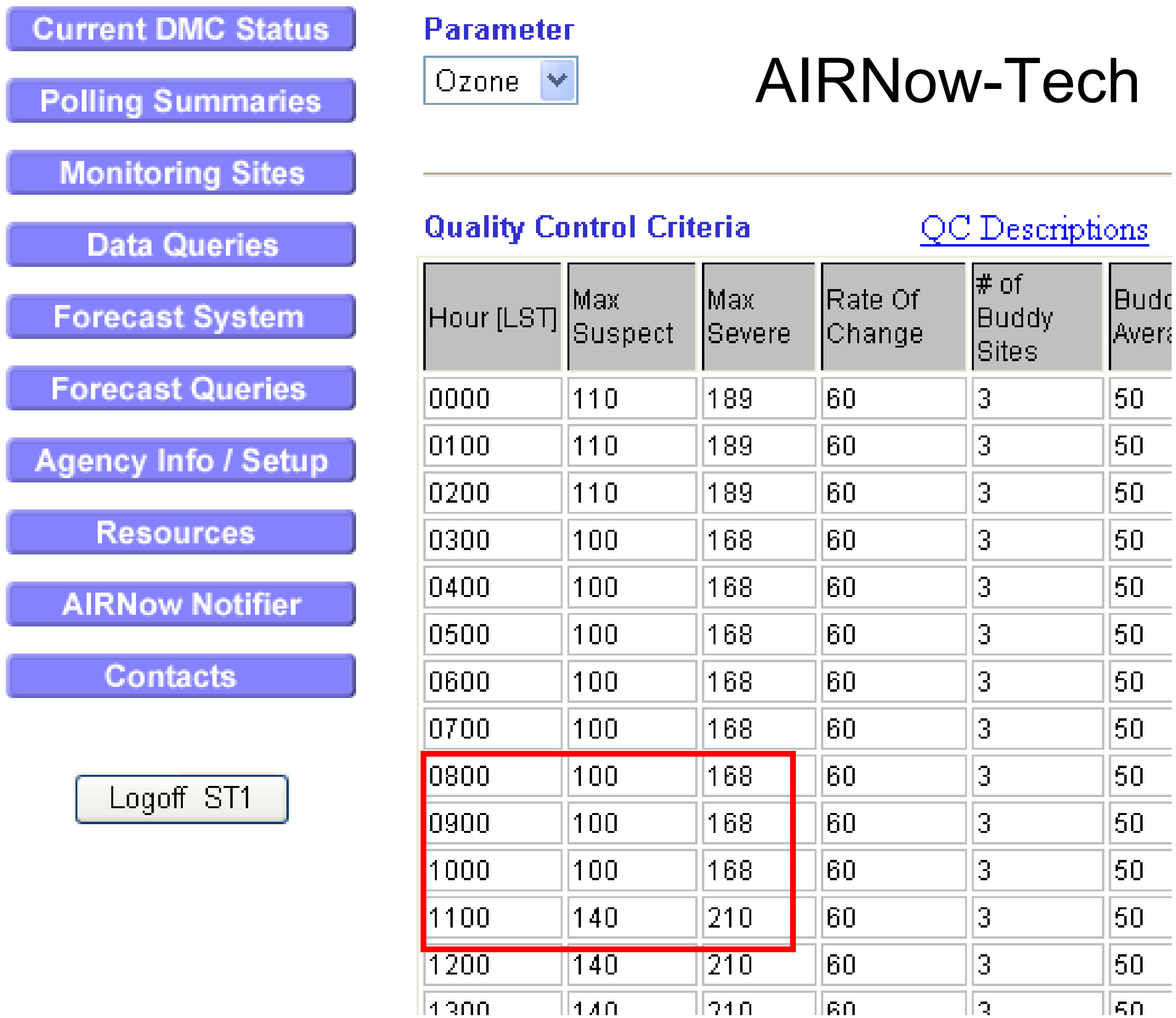
- Range checks: minimum and maximum concentrations
- Temporal consistency checks: maximum hour test
- Rate of change or spike check
- Buddy site check: comparison to nearby sites
- Sticking check: consecutive equal data values

Sample Screening Criteria
Ozone

- Checks:
  - Are often site-specific
  - May be hour-specific
  - May be automated
- But data should still 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

Example – Ozone Screening



AIRNow-Tech

Parameter

Ozone

Quality Control Criteria

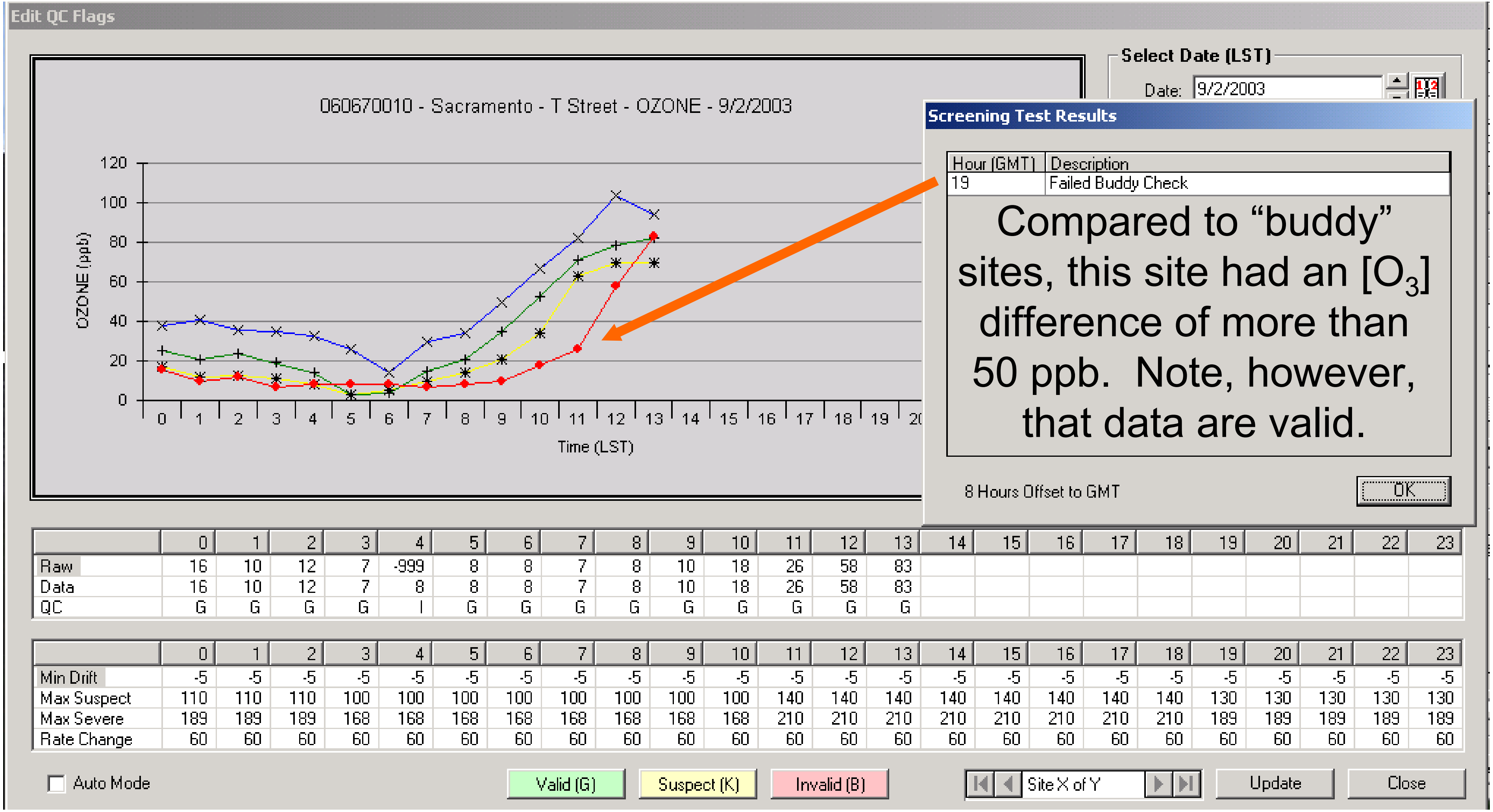
Hour [LST]	Max Suspect	Max Severe	Rate Of Change	# of Buddy Sites	Budd Avera
0000	110	189	60	3	50
0100	110	189	60	3	50
0200	110	189	60	3	50
0300	100	168	60	3	50
0400	100	168	60	3	50
0500	100	168	60	3	50
0600	100	168	60	3	50
0700	100	168	60	3	50
0800	100	168	60	3	50
0900	100	168	60	3	50
1000	100	168	60	3	50
1100	140	210	60	3	50
1200	140	210	60	3	50
1300	140	210	60	3	50

Max Suspect:
Still used in
spatial mapping

Max Severe:
Not used in maps

Note hour-specific screening

Example – Ozone Screening



Edit QC Flags

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

	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

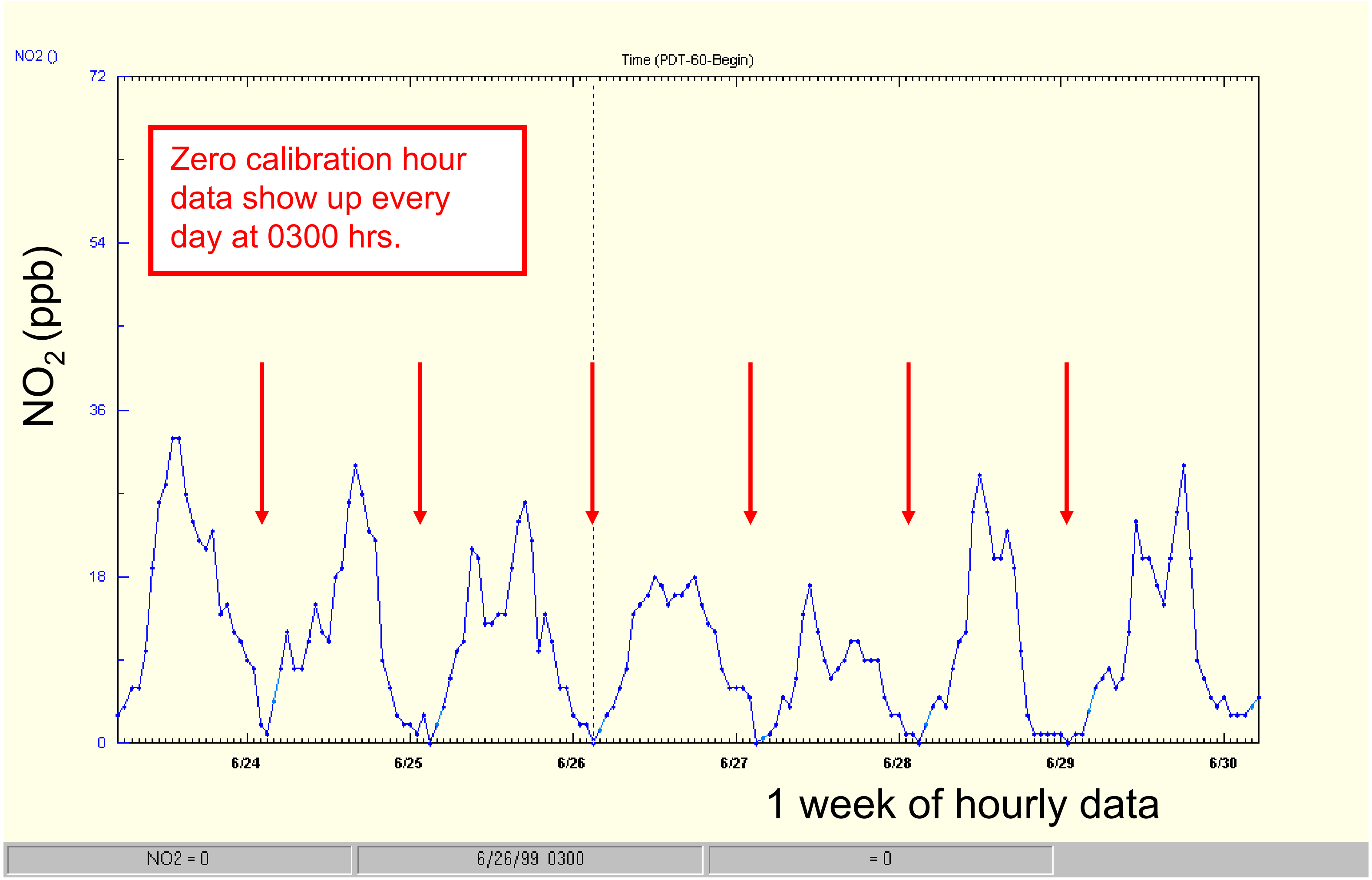
Sample Screening Criteria
$NO/NO_x/NO_y$

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

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

STI Sonoma Technology, Inc.

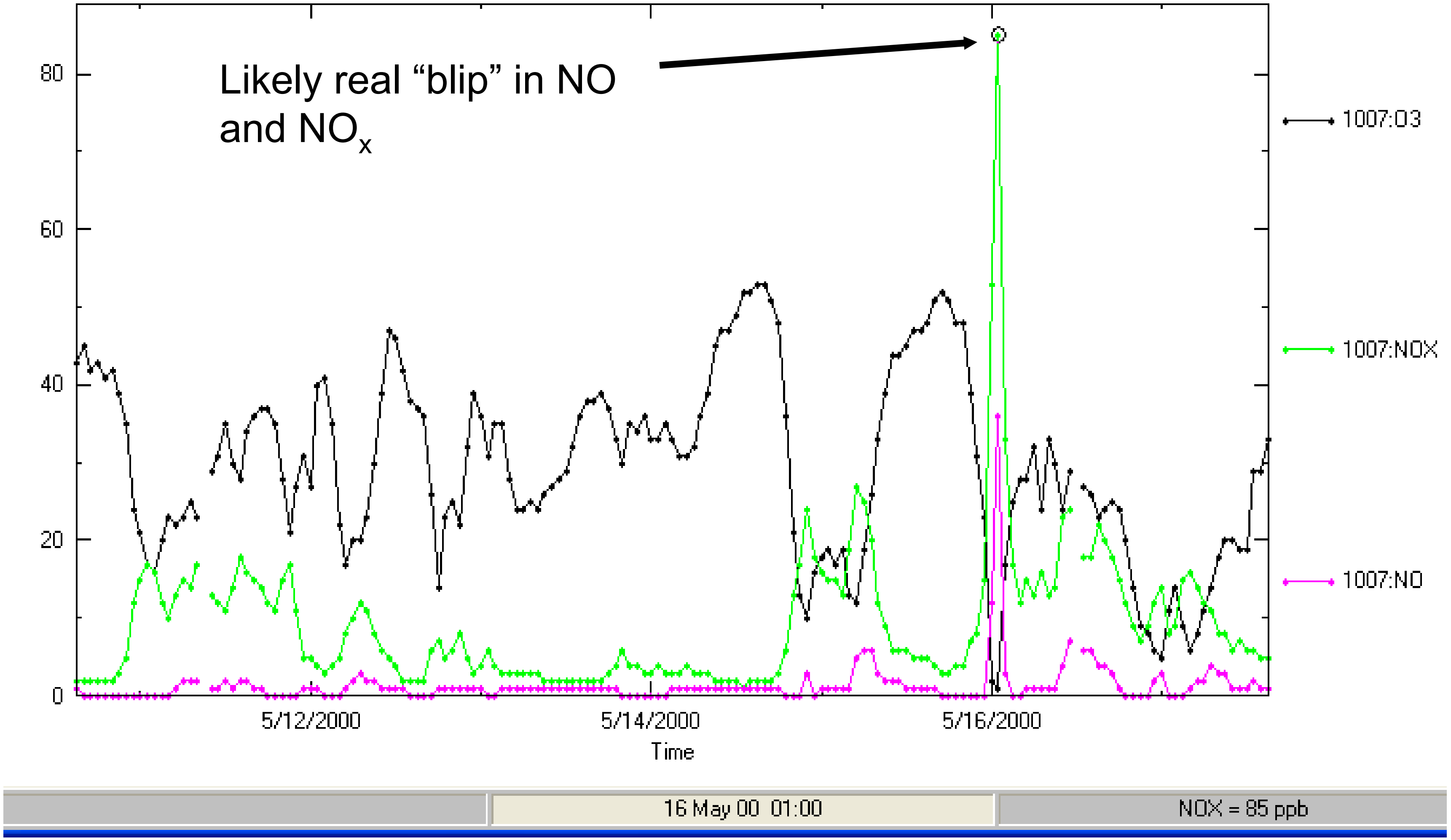
Example – Unusual Patterns



Zero calibration hour data show up every day at 0300 hrs.

1 week of hourly data

Example – Ozone, NO_x, NO



Likely real "blip" in NO and NO_x

Sample Screening Criteria
Carbon Monoxide

- Checks:
  - Select buddy check criteria
- Checks may vary with 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

Sample Screening Criteria

SO_2

- Checks:
  - Select buddy check criteria
- Checks may vary with 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

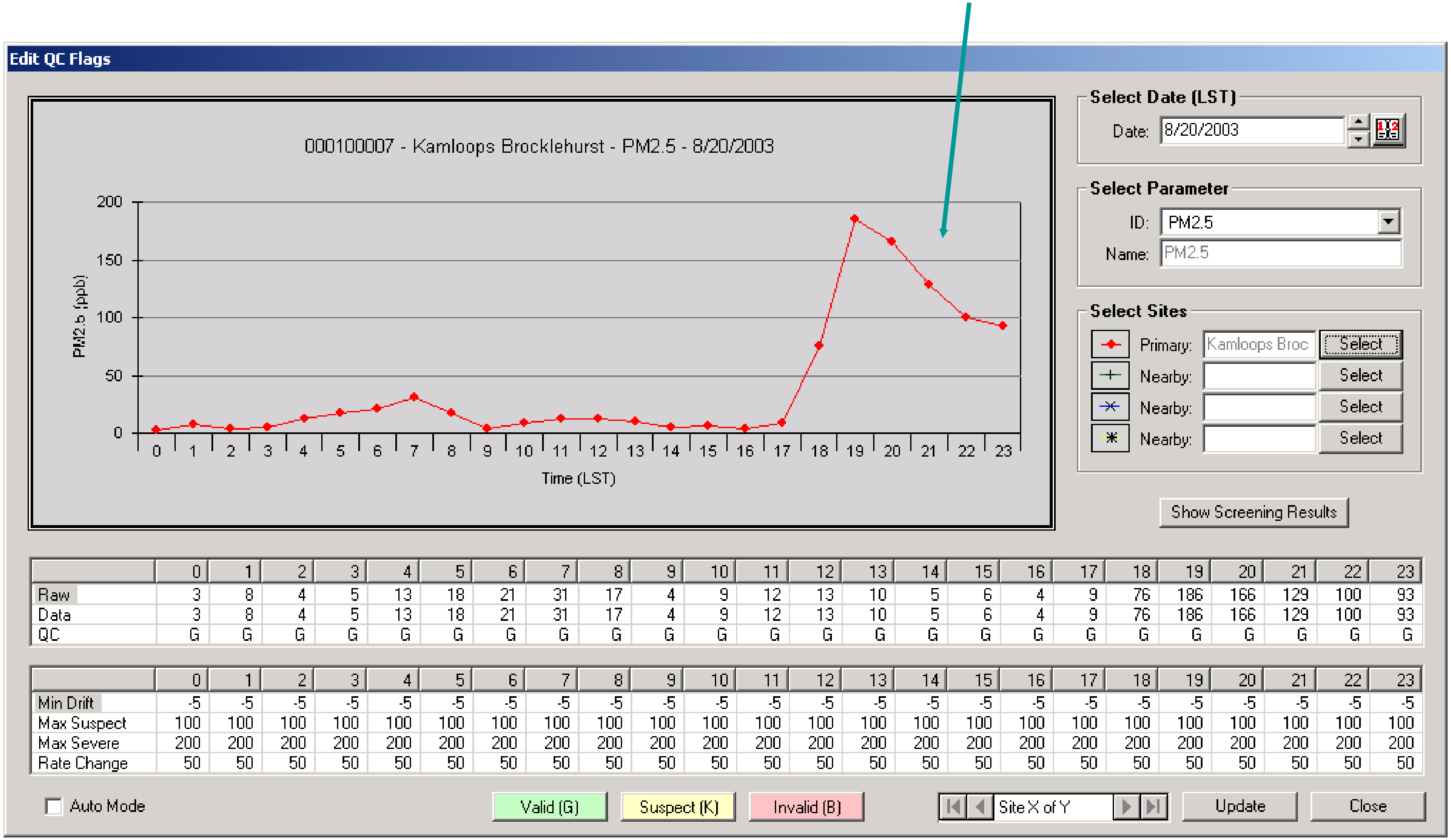
Sample Screening Criteria

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}

Example – Wildfire Events



Edit QC Flags

Select Date (LST)

Date: 8/20/2003

Select Parameter

ID: PM2.5

Name: PM2.5

Select Sites

Primary: Kamloops Broc — Select

Nearby: — Select

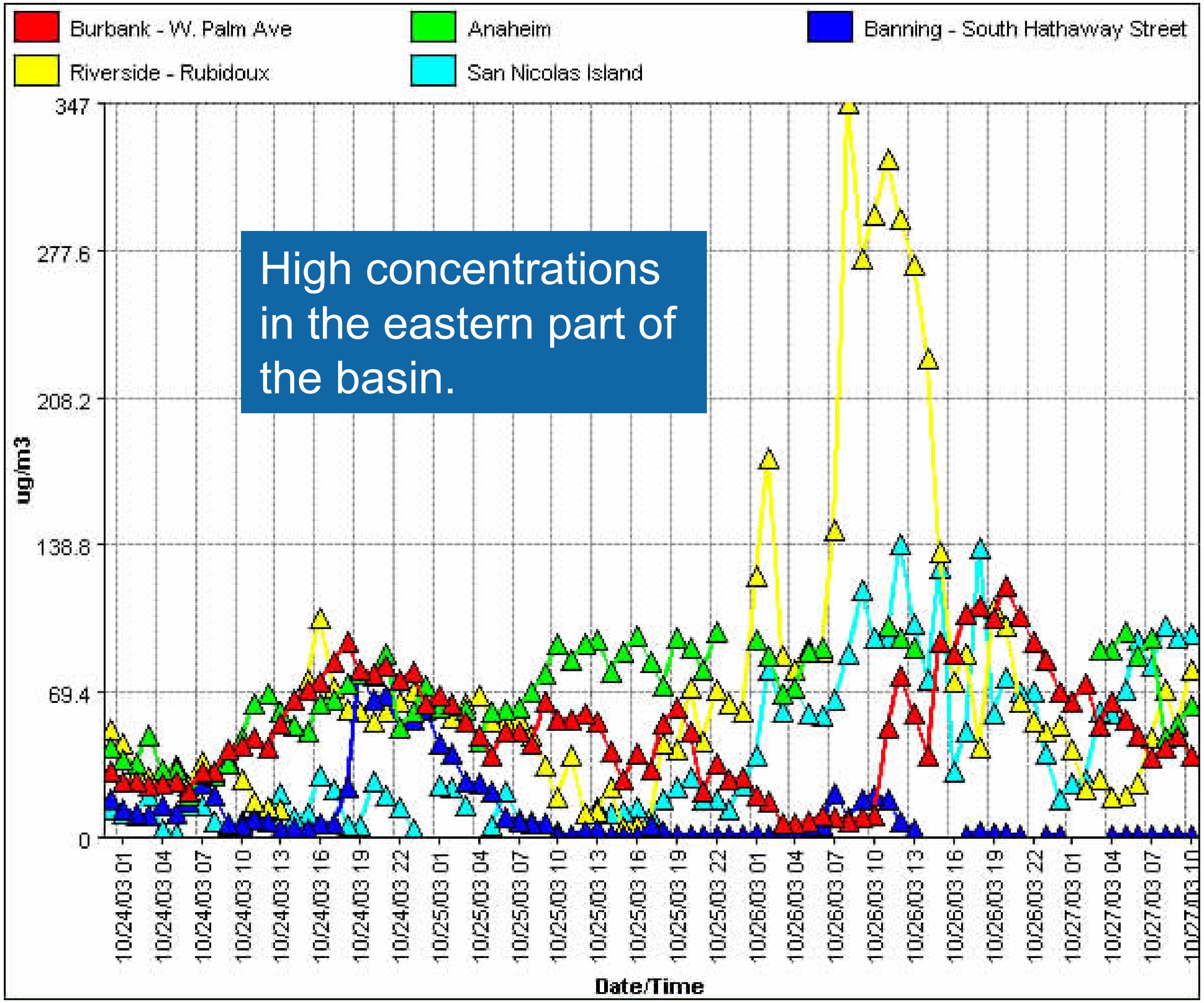
Nearby: — Select

Nearby: — Select

	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

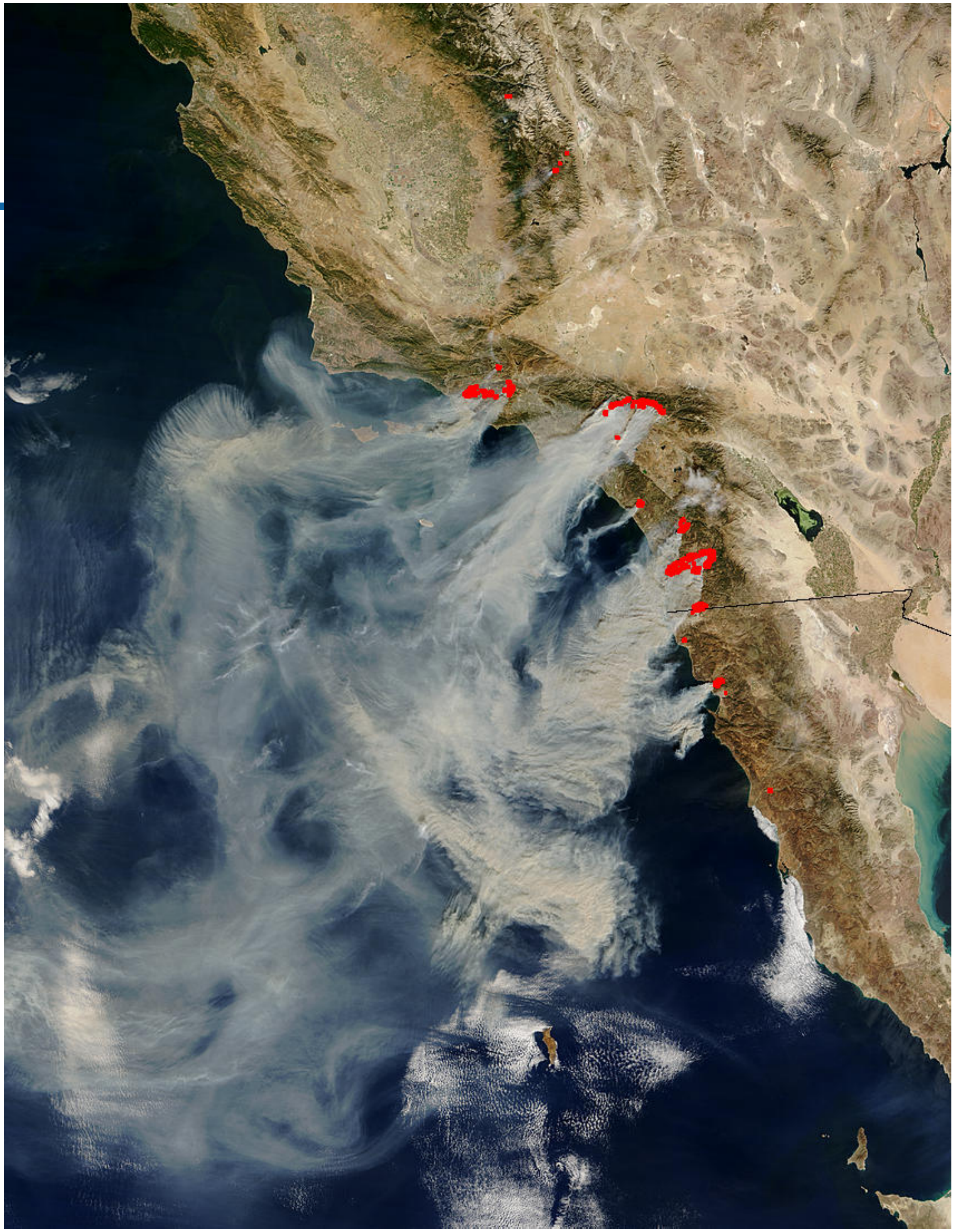
☐ Auto Mode



High concentrations in the eastern part of the basin.

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

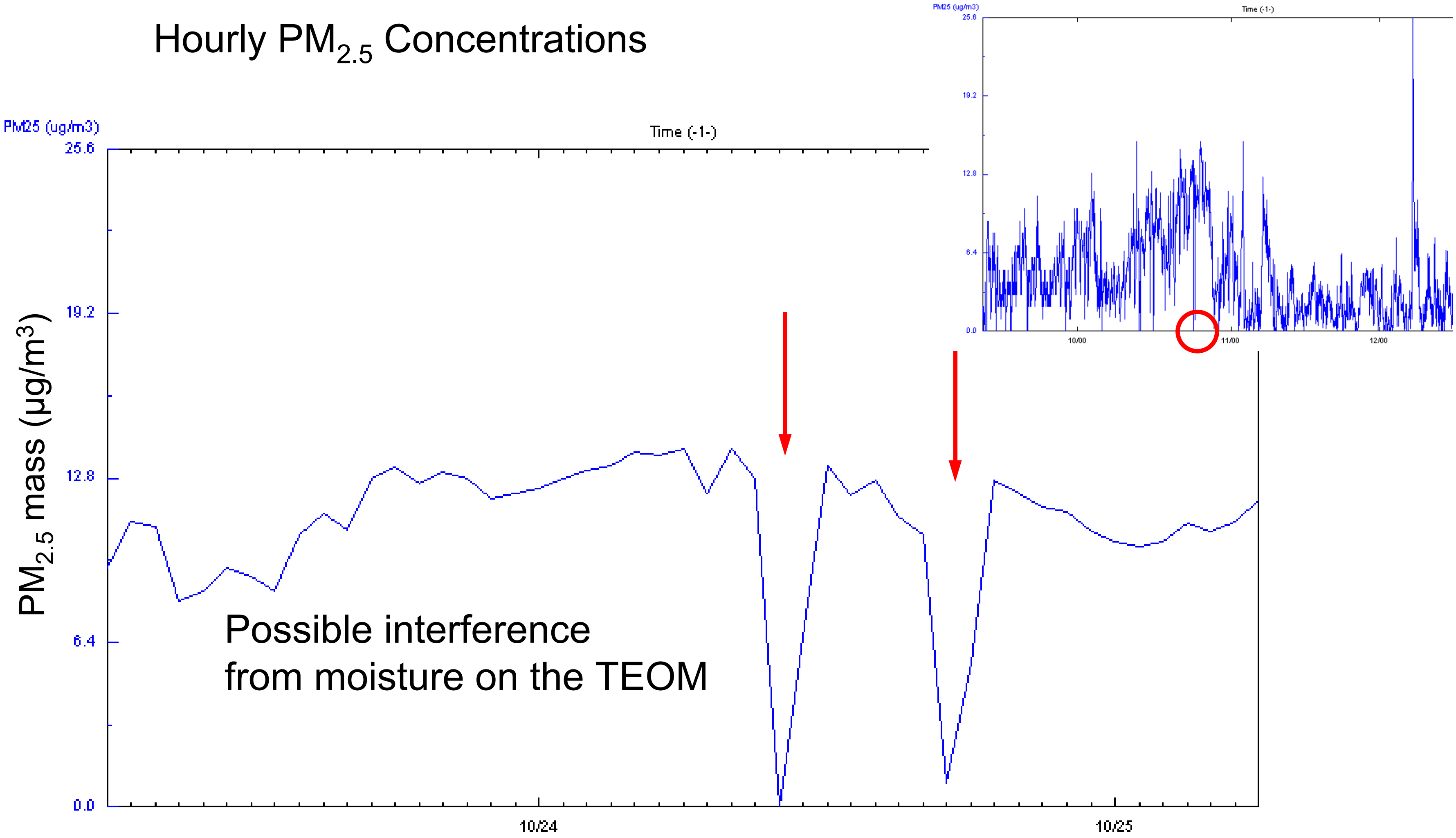
High concentrations are consistent with wildfire smoke as shown on this satellite photo from http://rapidfire.sci.gsfc.nasa.gov/gallery/



Sonoma Technology, Inc.

Example – Unusually Low Concentrations

Hourly $PM_{2.5}$ Concentrations



Possible interference
from moisture on the TEOM

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

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

* Dichotomous data may be an exception to this check

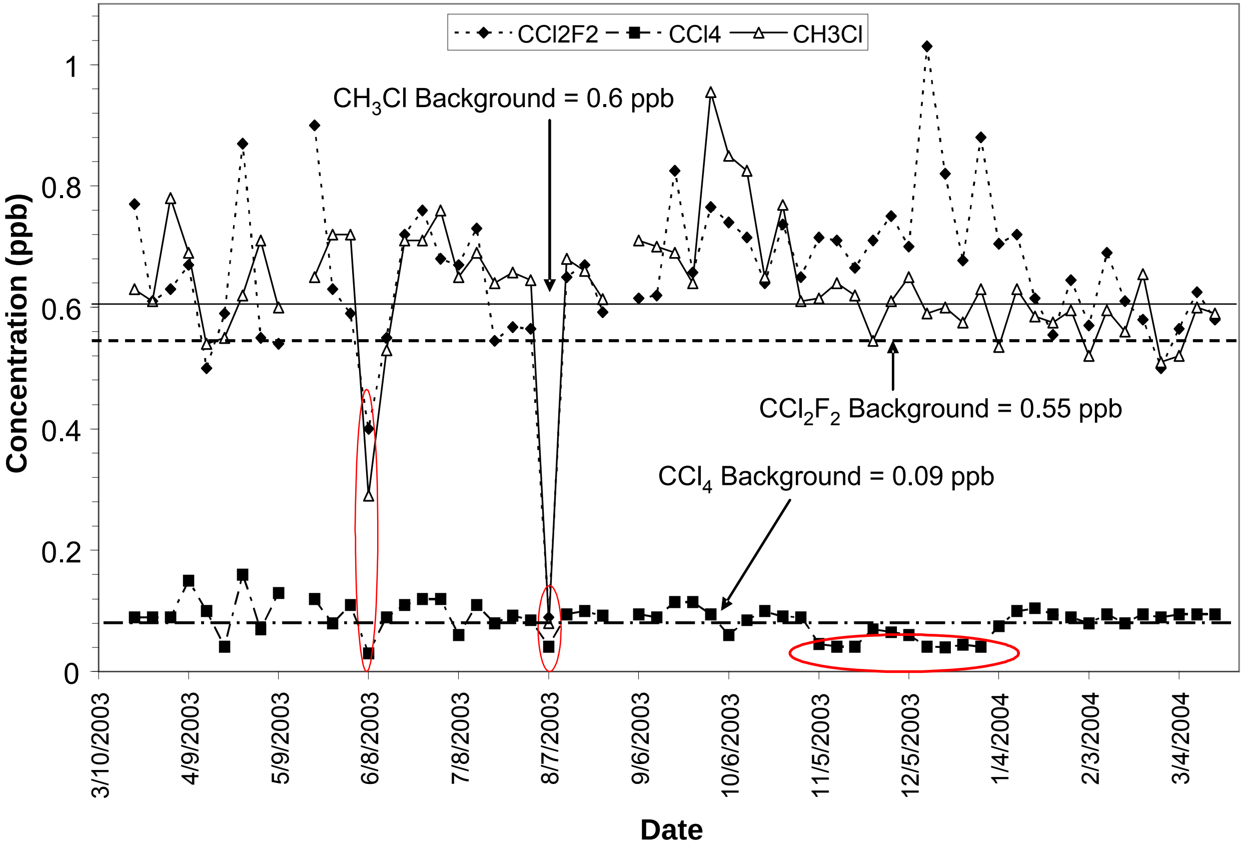
Chow, 1998

Sample Screening Criteria
Speciated Gaseous Data

- Apply screening checks (background, typical concentrations)
- Check that the data meet expected relationships
- Check for step changes/abrupt data shifts (concentration, species relationships, etc.)
- Review method detection limits (magnitude, changes over time, substitution)
- Further investigate maxima — high concentrations may be real

Screening Data Using Remote Background Concentrations

- Time series plot of concentrations of long-lived species compared to background concentrations measured at remote sites in the Northern Hemisphere.
- Significant dips in concentrations are circled.
- Concentrations more than 20% below the background level were identified as suspect for further review.



Concentrations (ppb) of carbon tetrachloride (CCl4), dichlorodifluoromethane (CCl2F2), and methyl chloride (CH3Cl) from 2003 and 2004.

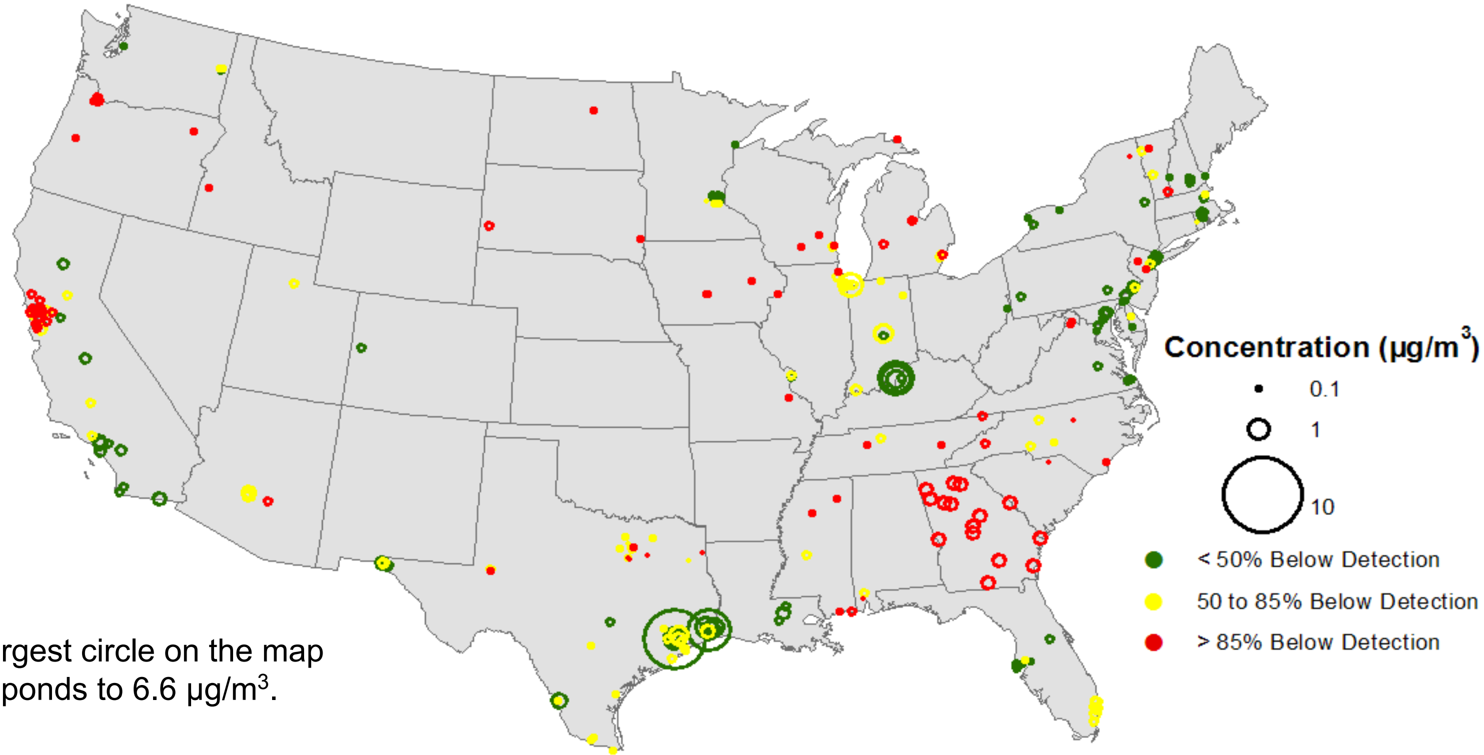
Sonoma Technology, Inc.

Concentration Spikes

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 indicate anomalous concentrations of other species.
- Real spikes in ambient concentrations are likely due to nearby point sources.
- A combination of maps, the Toxic Release Inventory, and local knowledge is likely required (but may not be sufficient) to explain spikes in ambient concentrations.
- Fugitive emission/upsets data are needed (and may be difficult to obtain).

Visualization Is Key!



The largest circle on the map corresponds to 6.6 µg/m^3.

- The ability to obtain 1,3-butadiene concentration measurements above the MDL across the United States varies (note all the red circles and their varying sizes).
- Higher concentrations generally coincide with locations of known point source emissions.
- Differences in monitoring methods and methods application have resulted in large differences in reported MDLs across the United States.

Data Validation Summary

For pollutant data validation

- Understand formation, emissions, and transport
- Establish and apply screening criteria to identify potentially suspect data
- Investigate suspect data
- Invalidate data only if there is sufficient evidence
- Document invalid data (so others can learn)

Data validation is very important!

Resources

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

Some 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
- EPA Supersite Overview: http://www.epa.gov/ttn/amtic/supersites.html
- Air Toxics Data Analysis Workbook: http://www.epa.gov/ttnamti1/toxdat.html