Source Identification of $PM_{2.5}$ in Steubenville, Ohio Using a Hybrid Method for Highly Time-resolved Data

Ram Vedantham, Matthew S. Landis, David Olson, and Patrick Pancras

Supporting Information Figures

Figure 1: Samples of the episode finding algorithm for (a) Fe and (b) Ge. The red overlay shows episodic periods and the blue marks the non-episodic periods. The x-axis units show temporal index and y-axis units are in ppb. The black boxes show periods when Fe and Ge independently impacted the receptor site. Other times, the species impacted the site simultaneously. This showed that Fe and Ge had both related and unrelated sources.

Figure 2: Single episode of matching peaks of Cd and Pb (black boxes) that may have influenced Unmix pairing of Cd and Pb in a single source while there are many Pb peaks that do not have matching Cd peaks. ReSCUE did not pair Cd and Pb. Cd was identified as a single species source type by ReSCUE. Cd and Pb units are ppb. The bottom frame shows the wind feathers (positive wind speed (m s^{-1}) values are indicative of northerly winds; negative wind speed values are indicative of southerly winds with normed velocity units).

Figure 3: Possible issue associated with gaseous SO_2 conversion to particulate SO_4 (S), units are ppb. The bottom frame shows the wind feathers (positive wind speed (m s^{-1}) values are indicative of northerly winds; negative wind speed values are indicative of southerly winds with normed velocity units). Note: (i) when S and SO_2 have matching peaks, the winds are from the north, (ii) most of the S peaks are from the north, but matched rarely by SO_2 peaks, and (iii) there are multiple instances of S episodes without correcsponding SO_2 episodes. Thus, SO_2 is only moderately associated with S according to both ReSCUE and Unmix results. The red cicles show the peaks associated with SO_2 with matching values (not necessarily peaks) with S. The red lines in the wind plot matches the SO_2 peaks.

Supporting Information Tables

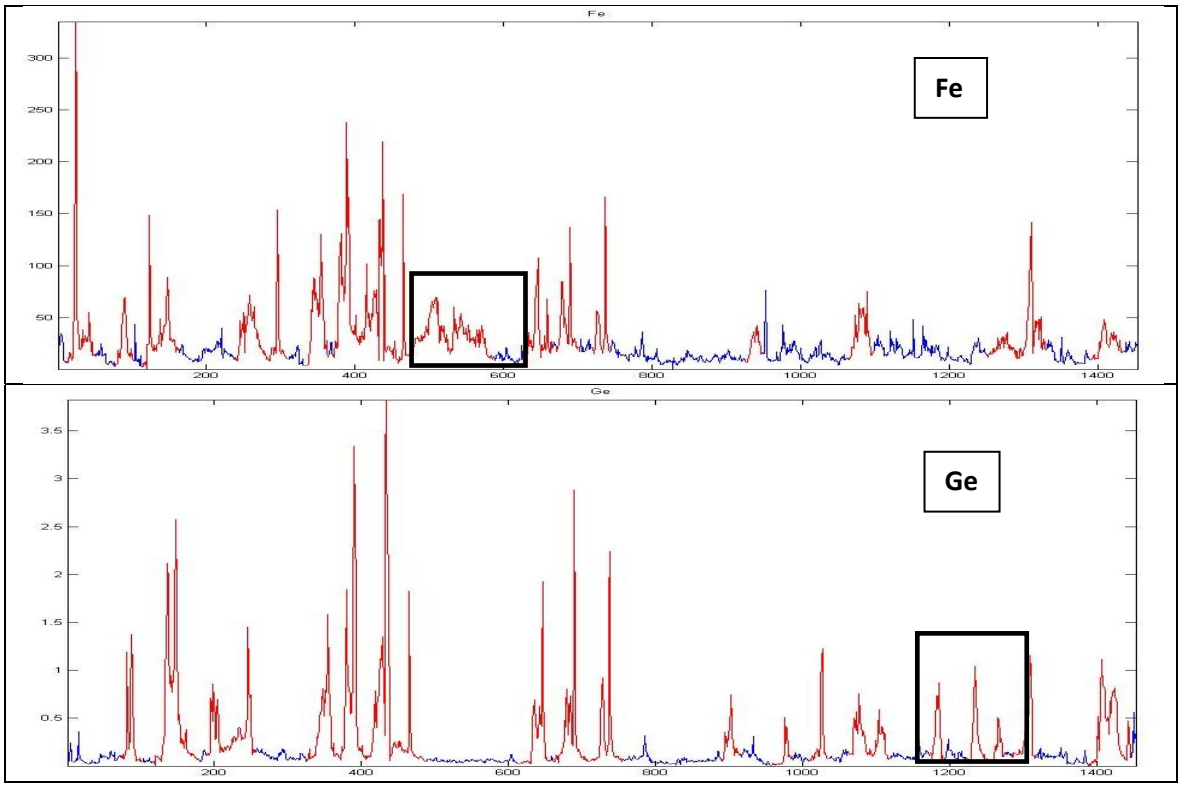
Table 1: Summary of Steubenville SEAS-III $PM_{2.5}$ species concentrations used in Rescue and Unmix analysis. All values are ppb except $PM_{2.5}$ (μg m^{-3}).

Table 2: EPA HR-ICPMS NIST Standard Reference Material (SRM) recovery summary for 1640a (Trace Elements in Natural Water) and 1643e (Trace Elements in Water) analyzed with Steubenville SEAS-III samples.

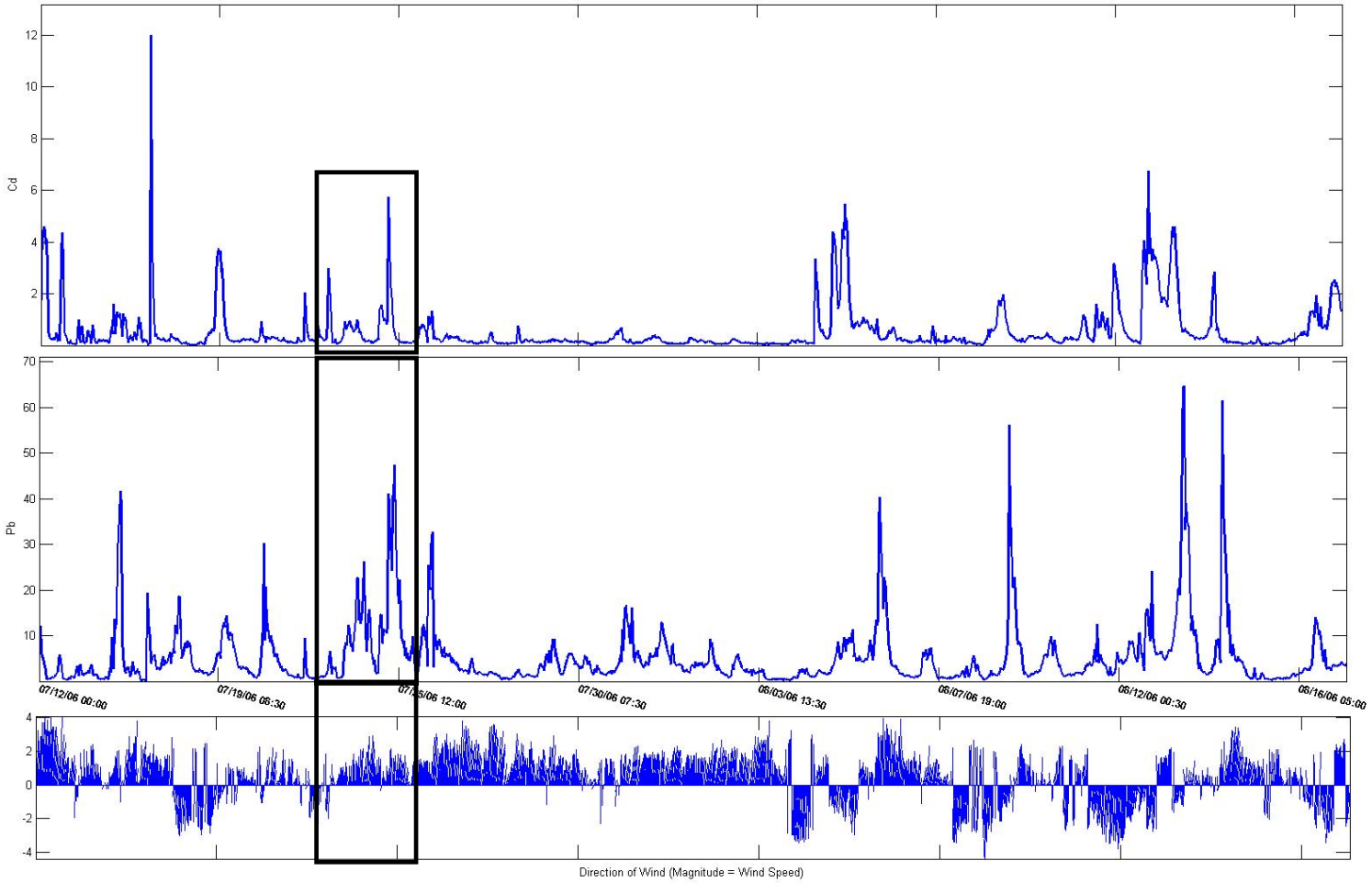
Table 3: Comparison between ReSCUE based results versus Simple Pearson Correlation using all data (not just episodes). Some of the species clusters are simply empty when simple correlation was used to create clusters.

Supporting Information MatLab Code (Version 2013A)

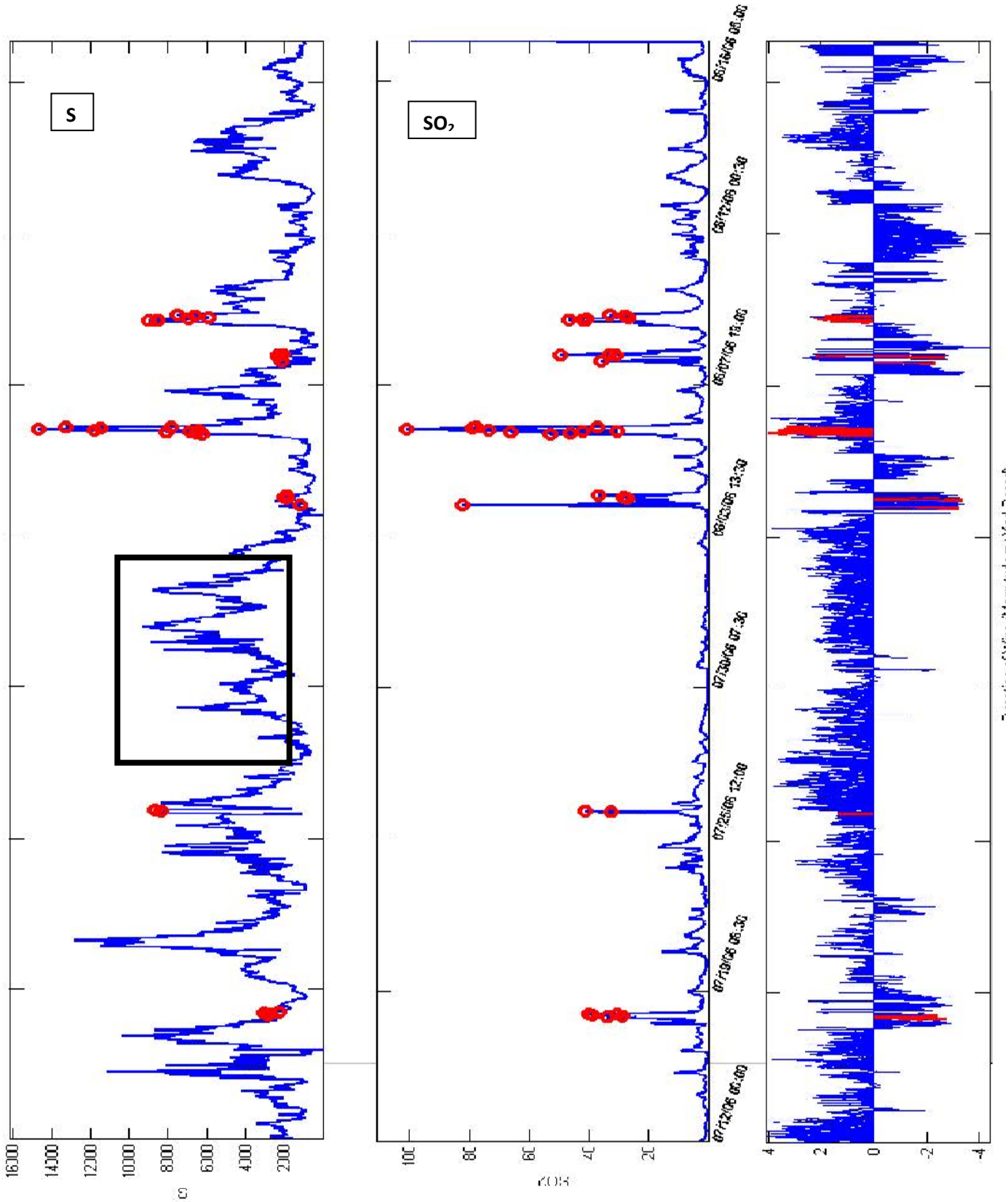
Episode Finding Algorithm
ReSCUE Algorithm
Determining Episodic Correlations (ESC) - Logic



SI Figure 1 Samples of the episode finding algorithm for (a) Fe and (b) Ge. The red overlay shows episodic periods and the blue marks the non-episodic periods. The x-axis units show temporal index and y-axis units are in ppb. The black boxes show periods when Fe and Ge independently impacted the receptor site. Other times, the species impacted the site simultaneously. This showed that Fe and Ge had both related and unrelated sources.



SI Figure 2: Single episode of matching peaks of Cd and Pb (black boxes) that may have influenced Unmix pairing of Cd and Pb in a single source while there are many Pb peaks that do not have matching Cd peaks. ReSCUE did not pair Cd and Pb. Cd was identified as a single species source type by ReSCUE. Cd and Pb units are ppb. The bottom frame shows the wind feathers (positive wind speed (m s^{-1}) values are indicative of northerly winds; negative wind speed values are indicative of southerly winds with normed velocity units).



SI Figure 3: Possible issue associated with gaseous SO_2 conversion to particulate SO_4 (S), units are ppb. The bottom frame shows the wind feathers (positive wind speed (m s^{-1}) values are indicative of northerly winds; negative wind speed values are indicative of southerly winds with normed velocity units). Note: (i) when S and SO_2 have matching peaks, the winds are from the north, (ii) most of the S peaks are from the north, but matched rarely by SO_2 peaks, and (iii) there are multiple instances of S episodes without correcsponding SO_2 episodes. Thus, SO_2 is only moderately associated with S according to both ReSCUE and Unmix results. The red cicles show the peaks associated with SO_2 with matching values (not necessarily peaks) with S. The red lines in the wind plot matches the SO_2 peaks.

Species	Mean	Standard Deviation	25th Percentile	Median	75th percentile
Rb	0.16	0.18	0.09	0.12	0.17
Sr	0.56	0.36	0.33	0.45	0.68
Mo	0.94	1.61	0.14	0.39	1.01
Cd	0.53	0.90	0.12	0.21	0.50
Ba	2.63	1.77	1.67	2.24	3.12
La	0.11	0.07	0.07	0.10	0.15
Ce	0.14	0.07	0.09	0.12	0.17
Sm	0.02	0.02	0.01	0.02	0.03
Pb	5.21	6.99	1.51	3.12	5.93
Mg	21.59	16.99	11.31	15.36	26.17
Al	31.69	41.23	12.85	20.64	35.85
S	3081.11	2102.85	1442.12	2593.75	4182.74
Ti	0.85	0.89	0.36	0.56	0.98
V	1.21	1.65	0.37	0.64	1.29
Cr	0.45	1.56	0.15	0.25	0.47
Mn	3.78	3.64	1.67	2.45	4.53
Fe	24.69	25.65	11.20	17.79	27.80
Ni	0.49	2.59	0.09	0.16	0.31
Cu	2.68	3.15	1.08	1.78	3.05
Zn	37.17	99.35	7.21	12.41	26.28
K	49.27	48.55	26.58	37.31	53.40
Ge	0.22	0.36	0.06	0.10	0.19
As	1.26	1.20	0.52	0.81	1.53
Se	2.09	3.16	0.93	1.48	2.29
SO_2	4.29	8.11	0.96	2.08	4.33
NO_X	13.73	10.84	5.69	9.48	20.22
$PM_{2.5}$	28.38	15.05	17.02	26.17	37.46

SI Table 1: Summary of Steubenville SEAS-III $PM_{2.5}$ species concentrations used in Rescue and Unmix analysis. All values are ppb except $PM_{2.5}$ (μg m^{-3}).

	10% SRM 1640 analysis				2% SRM1643 analysis			
Element	**Mean (ppb)**	**±**	**RSD* (%)**	**Error (%)**	**Mean(ppb)**	**±**	**RSD* (%)**	**Error (%)**
Al	5.66	±	21.07	8.9	2.90	±	2.79	2.2
As	2.66	±	5.48	0.4	1.20	±	3.71	0.8
Ba	14.19	±	4.17	4.1	10.40	±	1.29	4.5
Be	3.53	±	5.00	0.9	0.29	±	4.50	2.7
Bi	0.04	±	55.91		0.27	±	3.14	4.7
Ca	687.39	±	4.44	2.4	606.94	±	2.88	6.0
Cd	2.24	±	4.35	1.7	0.13	±	2.12	2.6
Co	2.02	±	4.07	0.3	0.54	±	3.00	0.3
Cr	3.73	±	4.71	3.4	0.42	±	3.15	2.9
Cu	8.62	±	4.70	1.1	0.46	±	2.79	2.1
Fe	3.11	±	8.08	9.3	2.01	±	3.44	2.5
K	94.42	±	5.25	5.0	41.95	±	3.10	3.1
Mg	583.66	±	4.51	0.3	161.69	±	3.42	0.6
Mn	12.21	±	4.86	0.5	0.80	±	3.02	2.5
Mo	4.62	±	5.05	1.1	2.53	±	2.23	4.3
Na	3020.69	±	4.97	2.9	425.53	±	3.75	2.6
Ni	2.69	±	4.08	1.7	1.20	±	2.44	3.7
Pb	2.63	±	4.05	5.6	0.37	±	1.76	5.2
Rb	0.22	±	8.36	12.1	0.30	±	2.85	6.0
Sb	1.33	±	4.97	3.9	1.14	±	2.42	2.0
Se	2.22	±	5.97	1.2	0.24	±	13.72	0.9
Si	505.23	±	5.67	6.8	0.60			
Sr	12.19	±	6.40	1.9	6.61	±	3.38	2.3
V	1.32	±	4.44	1.6	0.77	±	2.91	1.9
Zn	5.38	±	4.65	1.2	1.57	±	3.22	1.2

*Average and relative standard deviation (n = 25) from eight different sequences of analysis.

SI Table 2: EPA HR-ICPMS NIST Standard Reference Material (SRM) recovery summary for 1640a (Trace Elements in Natural Water) and 1643e (Trace Elements in Water) analyzed with Steubenville SEAS-III samples.

Base Species	ReSCUE based species clusters	Simple-correlation based species cluster
$PM_{2.5}$ Mass	**S**, Se	**S**
S	**$PM_{2.5}$ Mass**, Se, SO2, Fe	**$PM_{2.5}$ Mass**
K	**Rb**, **Mg**, **Fe**, **Zn**, **Mn**, **La**, Ce, Ge, Cu, Cd, NO_x, As, Ni, Ba,	**Rb**
Zn	**Rb**, **Mn**, **K**, **Fe**, **Mg**, Ge, NO_x, La, Ce,	**Rb**, Mn
Al	Sr, Mg	
Fe	**Ge**, **Rb**, **K**, **Mn**, **Mg**, **Zn**, La, Cu, Ce, Cr, Ni, NO_x, Ba, S, Sr, Pb, As	Mn, Rb, Mg, Ge
Mg	**Mn**, **Sr**, **Rb**, **K**, **Fe**, **Zn**, Ge, Ce, NO_x, Ba, La, Al	**Sr**, **Mn**, Rb, Fe
NO_X	**La**, **Ce**, **Rb**, **Mn**, **Ge**, **Mg**, As, K, Zn, Fe, Ti, Ba, V	Rb
Pb	Cu,SO_2,Fe	
SO_2	**Se**, S, Pb	Se
Mn	**Rb**, **Mg**, **Ge**, **Zn**, **NO_x**,**K**, **Fe**, Ce, La, Sr	**Mg**, Rb, Ge, Zn, Fe
Cu	Fe, K, Ge, Pb, Ce, La, Rb	
Ba	Fe, NO_X, Mg, Cr, Ni, Sr, K, Cd, Rb	
Se	**SO_2**,S,PM Fine	SO_2
As	La, NO_x, Rb, Ge, Ce, V, K, Cd, Fe	
V	Ti, As, Cr, Mo, NO_x	
Mo	Cr, V, Ti	
Ti	V, NO_x, Mo	
Sr	**Mg**, Mn, Al, Ce, Ba, Fe, Ge, Rb, Zn, K, NO_x	**Mg**
Cd	K, As, Ba	
Ni	**Cr**, Fe, K, Ba, Mo	**Cr**
Cr	**Ni**, Fe, Mo, V, Ba, La	**Ni**
Ge	**Rb**, **Mn**, **Fe**, **NO_x**, **Ce**, **La**, Mg, Zn, K, Cu, As, Sr	Rb, Mn, Fe
Rb	**Zn**, **Mn**, **K**, **Ge, Mg**, **La**, **NO_x**, **Fe**, **Ce**, As, Cu, Ba, Cd	**Zn**, **K**, Mn, Ge, Mg, Fe, NO_X
Ce	**La**, **NO_x**, **Rb**, **Ge**, Fe, K, Mg, Mn, As, Cu, Zn, Sr, Sm	**La**
La	**Ce**, **NO_x**, **Rb**, **Ge**, K, Fe, As, Mn, Cu, Zn, Mg, Cr	**Ce**
Sm	Ce	

SI Table 3: Comparison between ReSCUE based results versus Simple Pearson Correlation using all data (not just episodes). Some of the species clusters are simply empty when simple correlation was used to create clusters.

Episode Finding Algorithm

Matlab Code for identifying episodes and associating NaN (Not a Number – IEEE convention) to those values that do not fall within the episodes. Matlab version 2013A was utilized for all analysis detailed in this manuscript.

```
for specNum = 1:size(data,2)
    % Smooth the funtion first so peaks are easily found.
    filterData = filtfilt([1 0 1]/2,1,data(:,specNum));
    windowSize = 6;
    % The next two lines of code smoothes out data just enough to help identify
    % the start and end points of episodes. Care should be exercised so that
    % the smoothing technique only removes noisy substructures and not entire episodes.
    % In order to find robust episodes, a spline smoothing is applied to the time series.
    % This helps eliminate small scale variations from being identified as episodes.
    % The smoothed out curve is used only to identify the so-called “valley points” that
    % define episodes. This is the only use of the smoothed curve and is not part of any
    % further computation.
    filterData = filtfilt(ones(1,windowSize)/windowSize,1,data(:,specNum));
    fullFiltSplined = csaps(1:length(data),filterData,0.96,1:length(data));
    lowVals  = find(fullFiltSplined(nanVals)<prctile(data(:,specNum),25));
    fullFiltSplined(lowVals) = NaN;
    [peaks,locs] = findpeaks(fullFiltSplined,'minpeakheight',...
                        prctile(data(:,specNum),75),'minpeakdistance',3);
    if isempty(peaks)
        continue
    end
    % Use the peaks to find the episodes
    j = 1;
    diffVals = diff(fullFiltSplined(1:locs(1)));
    secondVal = find(diffVals(end-1:-1:1)<0,1,'first');
    if ~isempty(secondVal)
        prevBegin = locs(1)-secondVal;
    else
        prevBegin = 1;
    end
    for i = 1:length(locs)-1
        diffVals = diff(fullFiltSplined(locs(i):locs(i+1)));
        firstVal = find(diffVals>0,1,'first');
        episodicEnd = locs(i)+firstVal-1;
        while fullFiltSplined(episodicEnd) > prctile(fullFiltSplined,75)
            diffVals(1:firstVal) = [];
            firstVal = find(diffVals>0,1,'first');
            episodicEnd = episodicEnd+firstVal;
        end
        if length(diffVals)< 5
            continue
        end
        firstNaN = find(isnan(diffVals(end-1:-1:1)),6,'first');
        secondVal = find(diffVals(end-1:-1:1)<-0.00001,1,'first');
        if ~isempty(firstNaN) & firstNaN(end) < secondVal
            episodicBegin = locs(i+1)- firstNaN(1)+1;
            secondVal = firstNaN;
        else
            episodicBegin = locs(i+1)-secondVal;
            while fullFiltSplined(episodicBegin) > prctile(fullFiltSplined,75)
                diffVals(end-secondVal:end) = [];
                secondVal = find(diffVals(end-1:-1:1)<0,1,'first');
                episodicBegin = episodicBegin-secondVal;
            end
```

```
        end
        if episodicEnd - prevBegin < 5 & i > 1
            episodicBegin = round(mean([prevBegin episodicEnd]));
            episodicEnd = episodicBegin;
            prevBegin = episodicBegin;
            continue
        end
        episodicPair{specNum,j} = [prevBegin episodicEnd];
        included = [included prevBegin:episodicEnd];
        j = j + 1;
        prevBegin = episodicBegin;
    end
    diffVals = diff(fullFiltSplined(locs(end):end));
    firstVal = find(diffVals>0,1,'first');
    if isempty(firstVal)
        episodicEnd = length(fullFiltSplined);
    else
        episodicEnd = locs(end)+firstVal-1;
    end
    episodicPair{specNum,j} = [prevBegin episodicEnd];
    included = [included prevBegin:episodicEnd];
    included = unique(included);
    excluded = setdiff([1:size(data,1)],included);
    episodicData(excluded,specNum) = NaN;
end
```

ReSCUE Algorithm

```
clear varGroup varGroupNum varGroupVal
strongCut = 0.75;
weakCut = 0.60;
veryStrongCut = 0.80;
veryStrongStrong = 0.85;
veryStrongWeak = 0.70;
aveCut = 0.45;
colNum = 0;
for i = 1:length(drivers)
    strong = find(roundn(values(:,i),-2)>=strongCut);
    weak = setdiff(find(roundn(values(:,i),-2)>=weakCut),strong);
    grouping = [i indices(strong,i)'];
    grouping = [grouping indices(weak,i)'];
    strongGroupVal = nansum(values(strong,i));
    weakGroupVal = nansum(values(weak,i));
    strongVars = selectedTitles(indices(strong,i));
    veryStrong = find(roundn(values(:,i),-2)>=veryStrongCut);
    alsoStrong = [];
    for j = 1:length(veryStrong)
        currVar = indices(veryStrong(j),i);
        alsoStrong = setdiff(indices(find(roundn(values(:,currVar),-2)>=...
                        veryStrongStrong),currVar),grouping);
        grouping = [grouping indices(alsoStrong,currVar)'];
        alsoWeak = setdiff(find(indices(roundn(values(:,currVar),-2)>=...
                        veryStrongWeak),currVar),grouping);
        grouping = [grouping indices(alsoWeak,currVar)'];
        strongGroupVal = strongGroupVal + nansum(values(alsoStrong,currVar));
    end
    [uniqueGroup,index,uniqueIndex] = unique(grouping,'first');
    colNum = colNum + 1;
    varGroup{colNum} = [selectedTitles(grouping(sort(index)))]';
```

```
    varGroupVal(1,colNum) = strongGroupVal;
    varGroupVal(2,colNum) = weakGroupVal;
    varGroupNum(1,colNum) = length(strong)+length(alsoStrong);
    varGroupNum(2,colNum) = length(unique(grouping)) - varGroupNum(1,colNum) - 1;
end
```

Determining Episodic Correlations (ESC) - Logic

Episodic correlations (ESCs) between columns j and k are calculated by:

(1) finding those rows i where both X(i,j) and X(i,k) belong to their respective episodes.
(2) if there are no such rows, ESC(j,k)=0
(3) if there are such rows, then the Pearson correlation are calculated using only those rows (i). Such correlation is the desired ESC(j,k).
(4) the number N, present in their equation (1), is the number of rows where both X(i,j) and X(i,k) belong to their respective episodes. This number is generally different for all combinations of j and k.