Comparison of Trajectory Clustering Techniques with Source-Tagging Chemical Transport Models

Gary Kleiman, John Graham, Iyad Kheirbek, Nicolas Hamel, Ingrid Ulbrich and Jaime Lehner

Northeast States for Coordinated Air Use Management (NESCAUM)
Boston, MA 02114.

ABSTRACT

Back trajectories have been calculated (8 per day) for the five-year period including 1998 through 2002 using the HY-SPLIT modeling system for 17 Eastern sites including 10 rural locations near Class 1 areas subject to EPA's Regional Haze Rule and 7 urban location which have annual average $PM_{2.5}$ concentrations above or near the National Ambient Air Quality Standard (NAAQS). The back trajectories have been clustered based on 3-dimensional similarity to identify the predominant meteorological pathways influencing each site. Trajectories have also been associated with the nearest temporal value of 24-hr average concentration of $PM_{2.5}$ and IMPROVE SO_4^{2-} values measured at or near each site. By calculating and summing sulfate-weighted regional contributions to each cluster, results are compared with tagged runs of the REMSAD model and provide an independent check on contribution assessments developed through both techniques.

INTRODUCTION

The 1999 Regional Haze Rule (RHR) contains requirements for a site-specific pollution apportionment as part of each mandatory Federal Class I area's long-term emissions management strategy. A variety of techniques have been explored for conducting such pollution apportionments, but tagged chemical transport modeling is one of the few techniques which provide quantitative assessments of individual state or regional contributions to ambient concentrations. Given the importance of accurate pollution apportionment assessments, it is highly desirable to have independent techniques provide confirmation of transport model results.

Traditional trajectory analyses that associate an ambient measurement of air quality with the geographical region upwind prior to the observation are limited in that they demonstrate the relationship between ambient air quality and the integrated path along the length of a back trajectory. It is difficult to distinguish the contribution of a specific point along a single back trajectory from the contribution of other points along that path. Large numbers of back trajectories have been used in a variety of ways to try to "triangulate" by taking advantage of the variation in meteorology and paths that an ensemble of back trajectories offers.[1-5] Combining results from multiple receptor sites offers a more robust method of triangulation and can yield very specific source regions

associated with unique chemical signatures available with source apportionment techniques.

Moody et al.,[6] following methods of Dorling,[7] have applied the Patterns in Atmospheric Transport History (PATH) clustering algorithm, to large numbers of back trajectories in order to group trajectories by three-dimensional similarity. Calculation of average pollution levels corresponding to the members of a cluster of back trajectories of similar three-dimensional structure provides a robust technique of associating air pollutants with typical meteorological pathways,[8,9] but remains limited in its ability to distinguish individual points along an atmospheric pathway defined by a cluster of back trajectories.

The definition of an individual cluster of back trajectories in PATH analysis is dependent on a subjective choice of the "Radius of Proximity." This threshold defines the limiting difference between the three-dimensional coordinates of two back trajectories and determines if they are in the same cluster or different clusters. Selection of a smaller radius of proximity, in effect, will split clusters into component sub-clusters. Thus in the limiting case (radius of proximity = 0) the analysis reverts to a traditional trajectory analysis with each trajectory representing its own cluster. In this sense, PATH analysis offers a trade-off between uncertainty and fine scale structure of a trajectory analysis. By using a smaller radius of proximity – and thus a much larger number of clusters (100-200 clusters representing the 10,000+ back trajectories for each site over the 5-year period) – we have applied the PATH techniques to develop relatively well resolved (spatially) trajectory clusters. These have been weighted by pollution measurements and attributed to geographic areas.

An independent method of associating emissions with downwind air quality impacts involves the use of chemical transport models, or *source* modeling, rather than the *receptor* based approaches used in trajectory analysis. Here we use a large database of back trajectories and corresponding air pollution measurements to develop metrics related to Wishinski and Poirot's "incremental probability"[10] which is reflective of the increase in probability – relative to the everyday probability – of a geographic region being associated with a predominant meteorological pathway for sulfate transport (as opposed to a source region itself). Here we use these metrics in two ways. First, a sulfate-weighted probability is used to apportion observed sulfate as an independent check on source modeling results that have identified state-specific contributions of elevated point emission sources to sulfate formation. The weighted probability developed here is then compared to other incremental probability metrics to better understand how this technique compares to more traditional methods.

METHODS

The Hybrid Single Particle Lagrangian Intregrated Trajectory (HY-SPLIT) model[11,12] was used to calculate back trajectories for 17 sites in the Northeast U.S. The locations correspond to Class I areas subject to the RHR as well as several sites where potential nonattainment issues with the $PM_{2.5}$ NAAQS warranted analysis. Results are presented

here for Acadia National Park, Maine, Lye Brook Wilderness Area, Vermont, and Brigantine Wilderness Area, New Jersey.

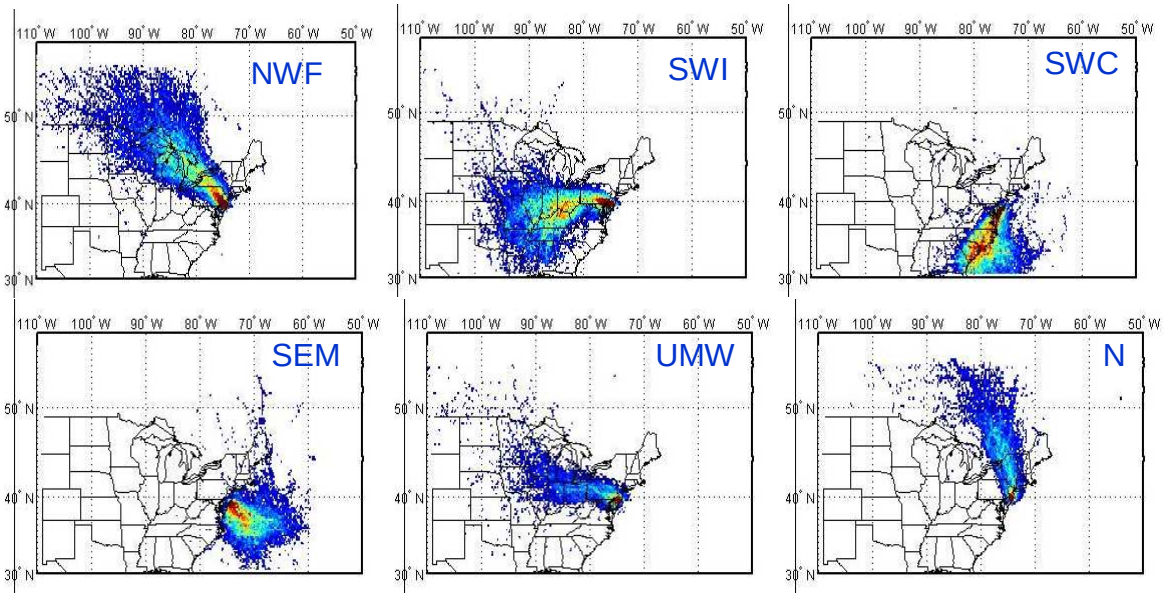
Back trajectories were calculated eight times per day for starting heights of 200, 500 and 1000m above ground level using two different sets of meteorological wind fields for the five year period 1998-2002. NOAA ARL archives analyzed meteorological products for use with the HYSPLIT model including the Eta Data Assimilation System (EDAS) wind fields, which cover North America with an 80 km spatial resolution and are based on 3-hourly variational analyses as well as wind fields based on the final run of the Global Data Assimilation System (FNL) which has 6-hourly temporal and 190 km horizontal resolution over the entire globe.[13]

Clusters were calculated using the PATH approach.[6] Trajectories are grouped based on Euclidean distance between three-dimensional normalized coordinates of the respective trajectories. Clusters are formed by finding the "central" trajectory which has the greatest number of neighboring trajectories within a subjectively selected "radius of proximity." There is a trade-off between the "resolution" of various modes of atmospheric transport identified by PATH and the number of clusters. While using a small radius of proximity as the threshold criterion for membership in a cluster results in generally more defined clusters that are easily identifiable with a specific class of meteorological transport (e.g. fast flow from the Northwest, shallow coastal flow, etc.), it also results in a large number of clusters at each site.

In order to better define specific meteorological pathways that might be associated with pollutant transport, we used a radius of proximity of 6 (this is a unit-less value since the coordinates have all been normalized prior to clustering). This typically resulted in 100-200 clusters at each site. Figure 1 shows typical meteorological patterns among the most frequent clusters calculated for Brigantine Wilderness Area, New Jersey. Results are plotted as a residence-time density for each cluster, which is a measure of the total time spent in a particular grid cell. Clusters in the figure have been associated with specific atmospheric "modes" or meteorological patterns that are commonly observed at multiple sites. The modes pictured correspond to Northwest Fast flow (NWF), Southwest Interior (SWI), Southwest Coastal (SWC), Southeast Maritime (SEM), Upper Midwest (UMW), and Northerly flow (N).

Trajectories were then associated with corresponding monitoring data measured as close in time as possible to the "start" time of the back trajectory calculation. Associations were made for $PM_{2.5}$, Ozone, and all PM components routinely measured as part of the IMPROVE program, although results are presented here only for 24-hr integrated sulfate ion mass.

Figure 1. Residence-time density for 6 back trajectory clusters observed at Brigantine Wilderness Area, New Jersey between 1998 and 2002.



IDENTIFYING INCREASED PROBABILITY OF SULFATE TRANSPORT PATHWAYS

The residence-time densities and corresponding sulfate measurements can be combined in a number of ways to yield various metrics which may help identify specific meteorological pathways that are more likely than others to contribute to sulfate transport to a specific receptor site.

One method begins with residence-time probabilities, which are a measure of the time spent in a specific grid cell relative to the total time spent in any grid cell.[14] When calculated for all trajectories considered in an analysis, this defines the *everyday* probability as shown in Equation 1.

Equation 1. Everyday Residence-time Probability

$$EP = \left(n_{ij} / N \right)$$

n_{ij} = total endpoints passing through grid cell i, j

N = total endpoints passing through all grid cells from all trajectories

These residence time probabilities can be calculated for any subset of trajectories, and have been traditionally applied to a subset corresponding to high concentrations of pollutants resulting in a *high-day* probability as shown in Equation 2.

Equation 2. High Day Residence-time Probability

$$HP = \left(m_{ij} \middle/ M \right)$$

m_{ij} = total high day endpoints passing through grid cell i, j

M = total high day endpoints passing through all grid cells from high day trajectories

The difference between these two sets of probabilities has been referred to as the *incremental* probability and identifies areas where the probability of poor air quality is greater than the average probability associated with typical meteorological patterns (see equation 3).

Equation 3. Incremental Probability

$$IP = HP - EP$$

In order to take advantage of the PATH analysis, two new metrics have been derived using the concept of incremental probability to investigate meteorological pathways that influence sulfate transport. First, a *clustered incremental probability* is defined by subtracting the everyday probability from a sum of the worst day clusters. Rather than choosing a subset of the trajectories, we have selected a subset of the clusters which are chosen based on their associated average sulfate concentrations. Clusters are ranked in order of their associated average sulfate value and clusters are summed until 20 percent of the overall trajectory population are included. Those clusters define the worst day conditions and their sum represents the worst day probability which can be applied in a modified incremental probability as shown in equation 4.

Equation 4. Clustered Incremental Probability

$$CIP = \sum_{k \in K} RP_k - EP$$

K = Subset of clusters encompassing 20% worst pollution days

RP_k = Residence - time probability for cluster k

Finally, we present the weighted-cluster probability. Each PATH-derived cluster's residence-time probability is weighted by the average sulfate value for any measurements corresponding to a trajectory which is a member of that cluster. The weighted residence-time probability is summed over *all* clusters calculated for a site. The everyday probability is subtracted from the sum of weighted-cluster probabilities to identify areas of increased (or in the case of negative values, decreased) probability of being associated with a meteorological pathway for pollutant transport. Equation 5 presents the weighted-cluster probability.

Equation 5. Weighted-Cluster Probability

$$WCP = \frac{1}{\overline{C}}\left(\sum_{i=1}^{L}(\overline{C})_i \cdot RP_i - \overline{C} \cdot EP\right)$$

L = total number of clusters calculated

$(\overline{C})_i$ = Average pollutant concentration (based on observations associated with cluster i)

$\overline{C}$ = Average pollutant concentration (based on all days)

Here, $(\overline{C})_i$ represents the average sulfate value for all trajectories within cluster i which had an associated SO_4^{2-} measurement (roughly 25-30 percent, given 1-in-3 day sampling schedules). By weighting the residence-time probability for cluster i by this quantity, we are implicitly assuming that similar trajectories (i.e. traversing similar source regions under similar meteorological conditions) will have similar resulting ambient concentrations at the receptor. The quantity $(\overline{C})$ represents the average value for sulfate measurements associated with trajectories in any of the clusters and acts to normalize the sum of the residence-time probabilities.

Figures 2 through 4 show the results of the weighted-cluster probability calculation for sulfate at Acadia National Park, Maine, Brigantine Wilderness Area, New Jersey and Lye Brook Wilderness Area, Vermont.

Figure 2. Areas of increased (yellow/red) or decreased (cyan/blue) probability of being associated with sulfate transport to Acadia National Park, Maine.

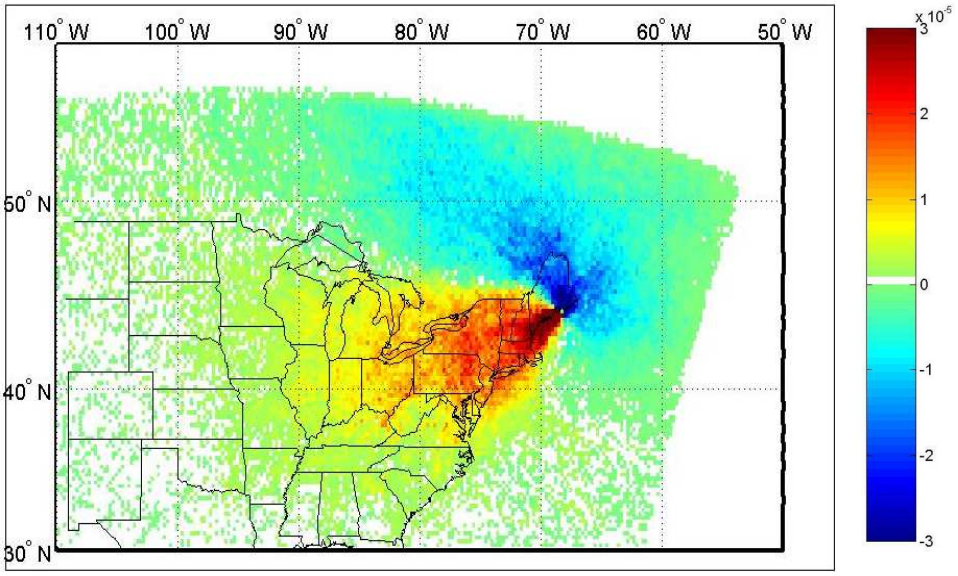


Figure 3. Areas of increased (yellow/red) or decreased (cyan/blue) probability of being associated with sulfate transport to Lye Brook Wilderness Area, Vermont.

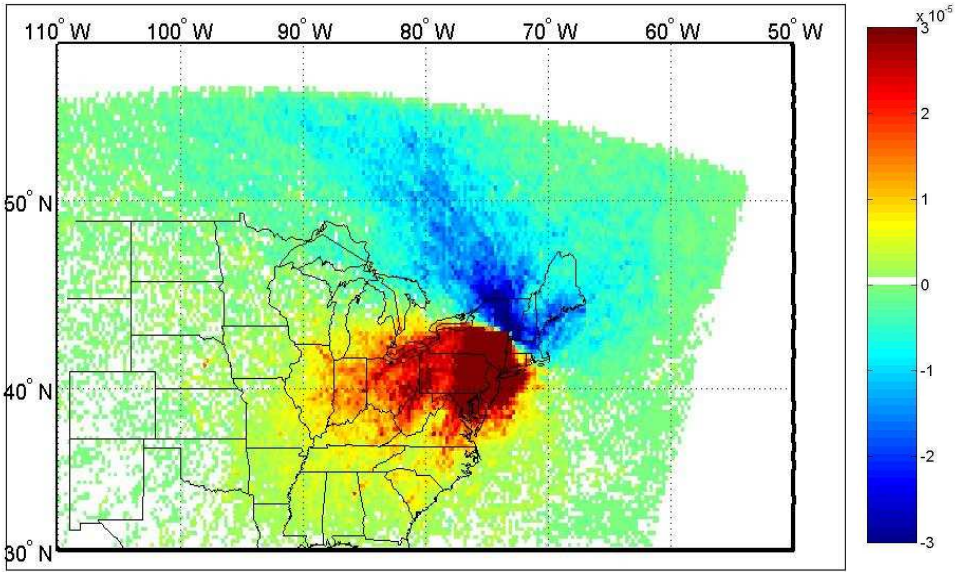
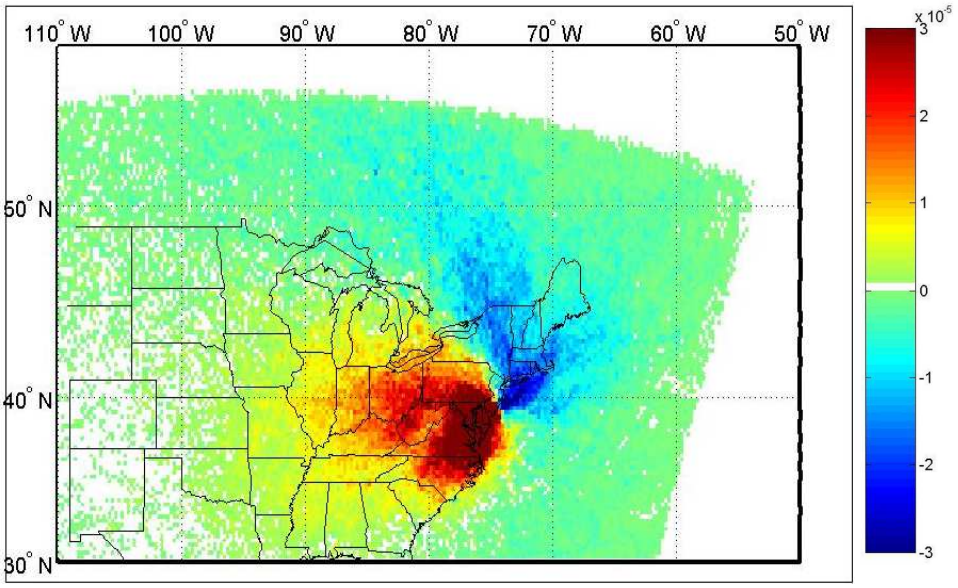
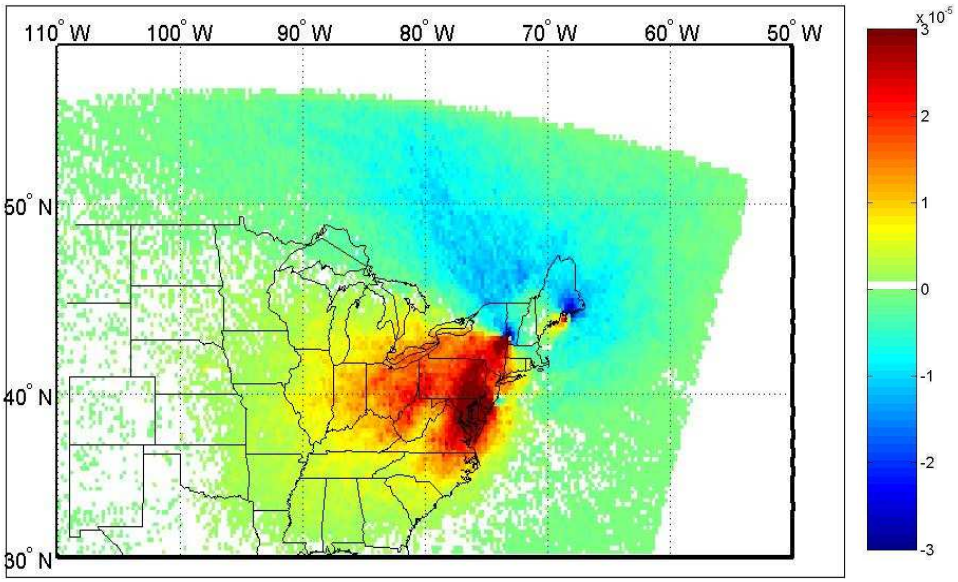


Figure 4. Areas of increased (yellow/red) or decreased (cyan/blue) probability of being associated with sulfate transport to Brigantine Wilderness Area, New Jersey.



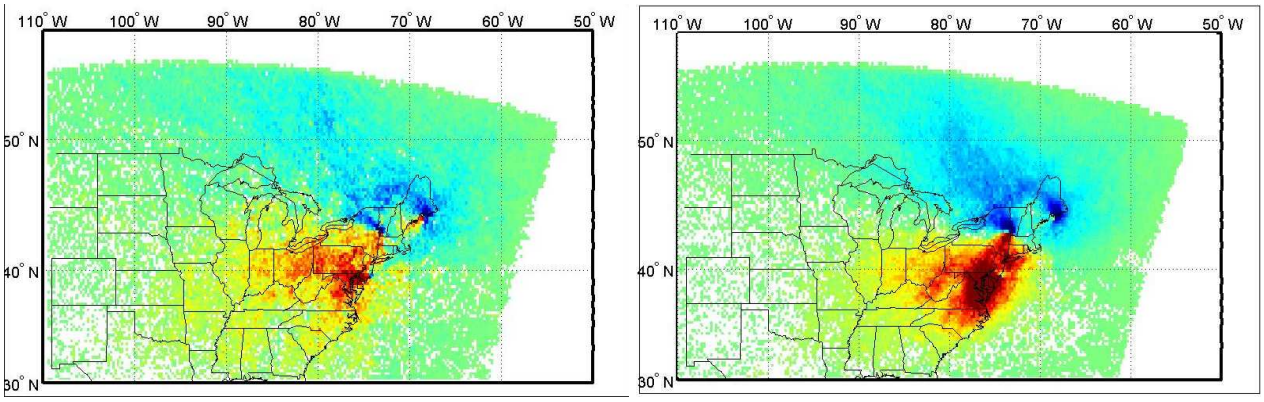
By averaging the weighted-cluster probability fields across sites, Figure 5 demonstrates that the regions most likely to be associated with sulfate transport to Lye Brook, Acadia *and* Brigantine include Virginia, Maryland and Eastern Pennsylvania followed by the Ohio River Valley.

Figure 5. weighted-cluster probability showing areas of increased (yellow/red) or decreased (cyan/blue) probability of being associated with sulfate transport to Acadia, Brigantine, and Lye Brook.



In trying to interpret this new metric, it is useful to contrast these results against the alternative formulations in order to better understand how the clusters influence the results. Figures 6 shows the three site average probability fields for the incremental probability as well as the clustered incremental probability, as defined in equations 3 and 4 above, for the twenty percent worst sulfate values.

Figure 6. Areas of increased (yellow/red) or decreased (cyan/blue) probability of being associated with sulfate transport to Acadia, Brigantine, and Lye Brook as measured by incremental probability (left) and clustered incremental probability (right).



One obvious difference between the metrics shown in figures 5 and 6 is that the incremental probability has values closer to zero than either of the other two metrics which may be a function of the sample size for each analysis. Whereas the cluster techniques use all trajectories in the sample, the incremental probability is limited to the

trajectories for which a corresponding measurement exists and thus uses only about 25-30 percent of the trajectories. From a qualitative perspective, the three metrics are quite similar showing significant sulfate transport (on an annual average basis) along the Eastern corridor from Virginia and North Carolina up through Maryland and Eastern Pennsylvania. A second area of influence along the Ohio River valley between Ohio, Pennsylvania and West Virginia may play a role as well.

As noted previously, the PATH-derived clusters and associated techniques may not be appropriate for distinguishing *source* regions from regions associated with sulfate transport (i.e. the integrated path from source to receptor). This may explain why the cluster-based techniques show greater influence in the region between Albany, NY and Long Island, NY which is closer to the Lye Brook and Brigantine source regions. However, it is difficult to discern given the difference in statistical samples between the techniques as discussed above.

COMPARISON OF SOURCE AND RECEPTOR TECHNIQUES FOR SULFATE CONTRIBUTION

Trajectory cluster techniques were used to provide an independent confirmation of REMSAD (v. 7.10 with source tagging) calculated contributions of various source regions to ambient concentration of sulfate ion at receptor sites in the Northeast U.S.

1996 meteorology and a 2001 "proxy" inventory developed by the U.S. EPA for analysis of the 2003 Clear Skies Act were used to generate annual average contributions from tagged elevated point source emissions of SO_2 in 31 eastern states. The elevated point source emissions were grouped by state allowing for the individual contribution from each state's elevated point sources to be individually tracked in the model. The picture of each state's total contribution to $PM_{2.5}$ is incomplete because area and mobile sources of SO_2 were not included in these tags and other components of fine particulate were not included. However, this does give a good sense of the relative ranking of various states point sources (>80% of the national SO_2 inventory). State-specific contributions to annual average sulfate concentrations are then ranked and grouped into quintiles.

Residence-times for the PATH derived clusters were then broken down by state to calculate the percent of each site's weighted-cluster probabilities that lay within a specific state's boundaries. That fraction gives a measure of each state's sulfate-weighted residence time that is attributable to a given sites transport. It is important to note that the average sulfate value derived for each cluster is based on only those members of the cluster for which there is a corresponding measurement. Given the 1-in-3 day schedule of the IMPROVE program, this means that the average is based on approximately one third of the population within the cluster. Here we assume that the distribution of measurement days is random with respect to actual concentrations and that no bias is incurred.

The result provides an indication of the state-specific contribution to the sulfate weighted residence-time for the most frequently occurring meteorological patterns. These contributions are ranked and grouped into quintiles similar to the REMSAD results.

Figures 7 through 9 show the results for the two techniques with the states shaded black representing the top quintile contributor, dark gray representing the second quintile, medium gray representing the middle quintile, light gray representing the fourth quintile and off-white for the bottom quintile. White states where not included in the REMSAD tagging scheme and thus were not included in the ranking schemes.

Figure 7. Ranked contributions of states to ambient sulfate concentration at Acadia National Park, Maine derived by REMSAD with source tagging (left) and weighted-cluster probability derived by PATH (right).

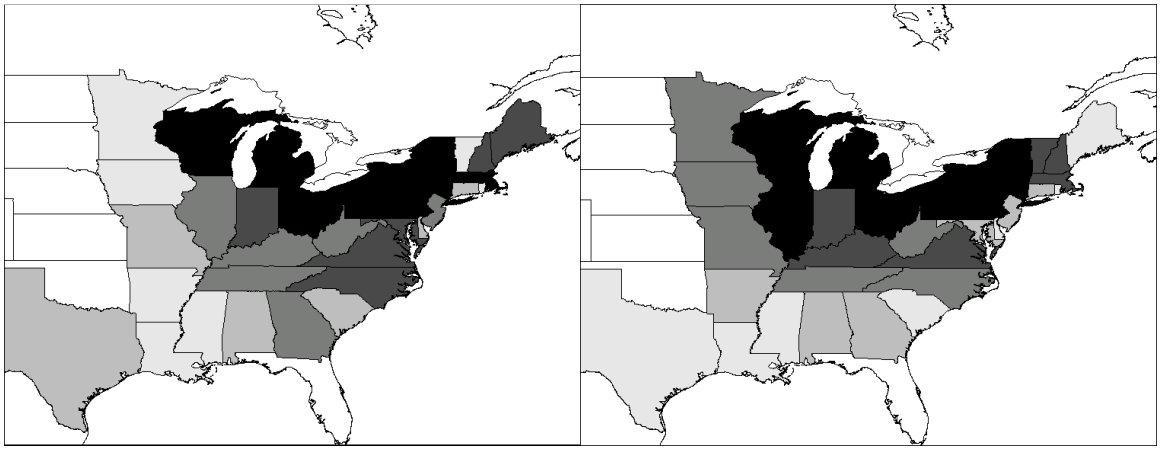


Figure 8. Ranked contributions of states to ambient sulfate concentration at Lye Brook Wilderness Area, Vermont derived by REMSAD with source tagging (left) and trajectory clusters derived by PATH (right).

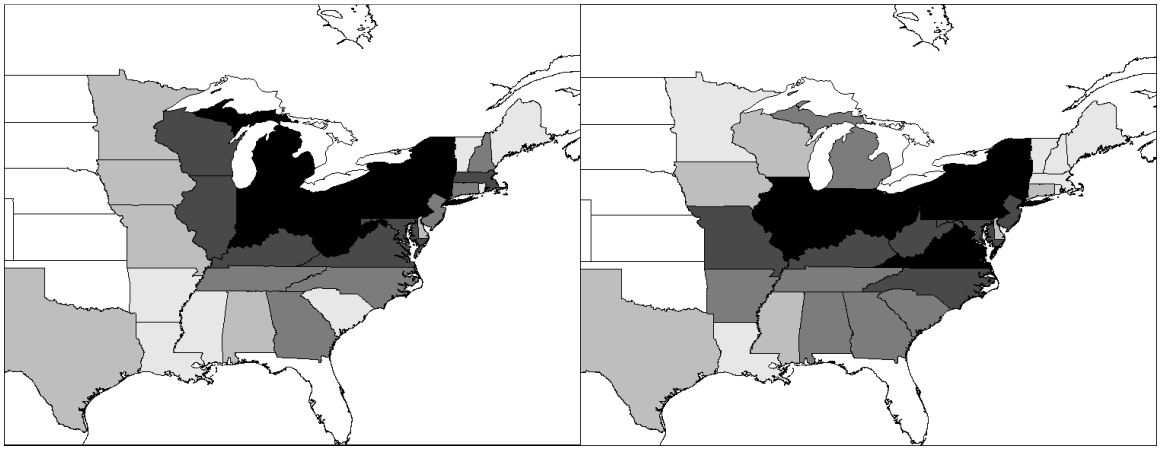
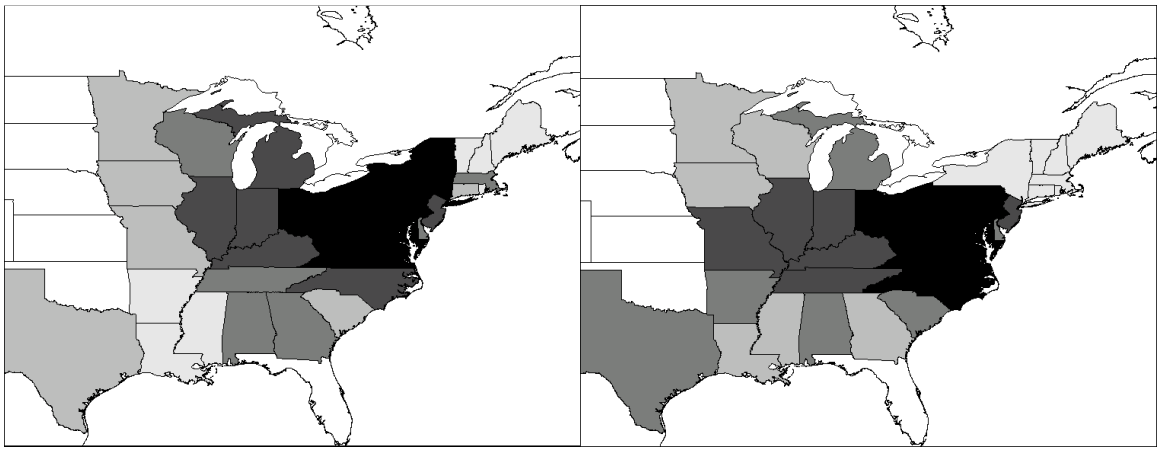


Figure 9. Ranked contributions of states to ambient sulfate concentration at Brigantine Wilderness Area, New Jersey derived by REMSAD with source tagging (left) and trajectory clusters derived by PATH (right).



In comparing these techniques, we recognize that, given the nature of PATH derived clusters, this approach is likely to give undue weight to nearby geographical areas simply due to the inability of the technique to distinguish contributions from specific locations along an integrated trajectory path. In fact, this is borne out in the results which show a bias toward contribution from nearby states relative to the REMSAD calculated sulfate contributions.

Nonetheless, given the similar patterns of contribution and the completely independent methodology for deriving these results, they do appear to provide an important check on source-based contribution techniques.

CONCLUSIONS

A large database of back trajectories and corresponding air pollution measurements have been used with trajectory cluster analysis techniques to apportion observed sulfate mass concentrations as an independent check on source modeling results. Both techniques have identified state-specific contributions of SO_2 emission sources to sulfate formation. Results indicate that cluster-based trajectory techniques can provide a semi-quantitative check on chemical transport model results, although greater effort must be taken to develop a tagged model run that incorporates all emission sources of SO_2 in order to directly compare with receptor approaches.

Clusters have also been used here to develop metrics related to Wishinski and Poirot's "incremental probability"[10] which provides the relative increase in probability of a geographic region being associated with a predominant meteorological pathway associated with pollutant transport (as opposed to a source region itself).

This analysis demonstrates a potentially novel way of identifying regions that play a role in pollutant transport (as opposed to regions which may host source emissions).

Combined with meteorological information and source apportionment model results, the approach may yield a more comprehensive picture of source emissions and the circumstances under which they are transported to specific receptor sites.

ACKNOWLEDGEMENTS

We thank Jung-Hun Woo who ably assisted with several graphics. Emily Savelli generated the REMSAD results which were compared to the receptor approaches. We also thank the U.S. EPA who funded this work under a U.S. EPA grant to the Mid-Atlantic/Northeast Visibility Union (MANE-VU). The HYSPLIT4 model was provided by NOAA. (HYbrid Single-Particle Lagrangian Integrated Trajectory Model, 1997. Web address: http://www.arl.noaa.gov/ready/hysplit4.html, NOAA Air Resources Laboratory, Silver Spring, MD).

REFERENCES

1. Seibert, P.; Kromp-Kolb, H; Baltensperger, U.; Jost, D.T.; Schwikowski, M.; Kaspar, A.; Puxbaum, H., "Trajectory Analysis of Aerosol Measurements at High Alpine Sites," in *Transport and Transformation of Pollutants in the Troposphere*, eds. Borell, P.M.; Borell, P.; Cvitas, T.; Seiler, W., Academic Publishing, Den Haag, **1994**, 689-693.

2. Hopke, P.K.; Li, C.L.; Cizek, W.; Landsberger, S., "The Use of Bootstrapping to Estimate Conditional Probability Fields for Source Locations of Airborne Pollutants," *Chemom. Intell. Lab. Syst.* **1995**, 30, 69-79.

3. Stohl, A., "Trajectory statistics: A new method to establish source-receptor relationships of air pollutants and its application to the transport of particulate sulfate in Europe," *Atmos. Environ.*, **1996**, 30, 579-587.

4. Kleiman, G.; Prinn, R.G., "Measurement and Deduction of emissions of trichloroethene, tetrachloroethene, and trichloromethane (chloroform) in the northeastern United States and southeastern Canada," *J. Geophys. Res*., **2000**, 105, 28875-28893.

5. Hsu, Y.K.; Holsen, T.M.; Hopke, P.K., "Comparison of hybrid receptor models to locate PCB sources in Chicago," *Atmos. Env*., **2003**, 37, 545-562.

6. Moody, J.L.; Munger, J.W.; Goldstein, A.H.; Jacob, D.J.; Wofsy, S.C., "Harvard Forest regional-scale air mass composition by Patterns in Atmospheric Transport History (PATH)," *J. Geophys. Res.*, **1998**, 103, 13181-13194.

7. Dorling, S.R.; Davies, T.D.; Pierce, C.E., “Cluster analysis: A technique for estimating synoptic meteorological controls on air and precipitation chemistry – Method and Application,” Atmos. Env., **1992**, 26A, 2575-2581.

8. Dorling, S.R.; Davies, T.D.; Pierce, C.E., “Cluster analysis: A technique for estimating synoptic meteorological controls on air and precipitation chemistry – results from Eskdalemuir, south Scotland,” Atmos. Env., **1992**, 26A, 2583-2602.

9. Kahl, J.D.W.; Liu, D.; White, W.H.; Macias, E.W.; Vasconselos, L., “The Relationship Between Atmospheric Transport and the Particle Scattering Coefficient at the Grand Canyon,” *J. Air & Waste Manage. Assoc*., **1997**, 47, 419-425.

10. Wishinski, P.R.; Poirot, R. L., “Long-Term Ozone Trajectory Climatology for the Eastern US, Part I: Methods,” paper 98-A613, Vermont Dept. of Env. Conservation (1998).

11. Draxler, R.D.; Hess, G.D., “Description of the HYSPLIT-4 Modeling System,” *NOAA Technical Memorandum ERL, ARL-224*, Air Resources Laboratory, Silver Springs, Maryland, 24 pgs., **1997**.

12. Draxler, R.D.; Hess, G.D., “An Overview of the HYSPLIT-4 Modelling System for Trajectories, Dispersion, and Deposition,” *Australian Meteorological Magazine*, **1998**, 47, 295-308.

13. Rolph, G.D., Real-time Environmental Applications and Display sYstem (READY) Website (http://www.arl.noaa.gov/ready/hysplit4.html). NOAA Air Resources Laboratory, Silver Spring, MD., **2003**.

14. Poirot, R. L.; Wishinski, P.R., “Visibility, sulfate and air mass history associated with the summertime aerosol in Northern Vermont,” *Atmos. Environ*. **1986**, 20, 1457-1469.

KEY WORDS

Trajectories
Receptor models
Clusters
Conditional probabilities
Source region
Transport pathways
Clean air corridors