ANSTO/ENV/TN03-2008

AUSTRALIAN NUCLEAR SCIENCE AND TECHNOLOGY ORGANISATION

LUCAS HEIGHTS SCIENCE AND TECHNOLOGY CENTRE

HYSPLIT Back Trajectories For Potential Source Contribution Function studies of

Trans boundary pollution Episodes (pscfV6)

by

J. Crawford
D. Cohen

ANSTO Institute for Environmental Research

August 2008

Table of Contents

Summary

A program that calculates the parameters needed to generate a Potential Source Contribution Function (called PSCF) has been implemented. When using PSCF the user specifies two values through the use of the standard deviation of the dataset to subdivide the data into 3 sets. Following this, the parameters that are required for the generation of a Potential Source Contribution Function are estimated by the program for each of the three regions.

The program and its input and output files are described in this report. It can generate either forward or backward trajectories through the specification of the length of the trajectory to consider, forward is specified by a positive length and backward is specified by a negative length.

1. Introduction

Atmospheric back trajectories have been widely used to quantify the influence of air transport on the pollution at a site (Stohl 1998). There are two major ways to visualise air quality data (Owega et al. 2006). The first is a probability map which identifies areas around the receptor site that contribute to the pollution observed at the site (the Potential Source Contribution Function, PSCF; Hopke et al., 1995, Crawford et al., 2007). The second is cluster analysis where the data is split into a number of groups representing distinct fetch areas and atmospheric transport patterns.

The program described here implements a variant of Potential Source Contribution Function (see section 3).

2. Trajectory calculation

Air mass trajectories give an approximation of the path of polluted air parcels over a period of time. Hence, a trajectory model can be used to identify the potential source regions for pollutants measured at the receptor site (e.g. Brankov et al., 1997). Various methods to compute trajectories, based on different assumptions, have been developed (Stohl, 1998). The accuracy of an individual trajectory is limited by the temporal and spatial resolutions of meteorological observations, measurement and analysis errors, and any simplifying assumptions used in the trajectory model (Stohl, 1998, Brankov et al., 1997). An assumption that is often made is that the errors are random, and instead of relying on the analysis of any individual trajectory, a large number of trajectories are used (Harris and Kahl 1990, Brankov et al., 1997, Hopke et al., 1995), which results in the cancelling-out of errors from individual trajectories.

PSCF uses the PC based version of HYSPLIT v4.0 (Draxler and Rolph, 2003) to generate the back trajectories. Small-scale features, such as the impact of local terrain, cannot be resolved by the data assimilation system that produces the global-grided wind fields from which trajectories are calculated. The trajectories therefore reflect the large-scale atmospheric transport characteristics of the air (Harris and Kahl 1990).

3. Potential Source Contribution Function

PSCF was developed to identify geographical regions giving rise to observed concentrations (Hopke et al., 1995). In this model, air mass back trajectories are combined with the elemental concentration measurements or identified source factor contribution to produce conditional probabilities over the region, where the region is subdivided into a number of grid cells. The number of trajectory endpoints (i.e. coordinates of the back trajectory for each hour before arriving at the receptor site) falling within grid cell *i,j* over the whole set of samples, $n_{i,j}$, are then counted. Then, the subset of trajectories associated with high concentration samples are identified by comparing the measured concentration to a threshold level. The number of endpoints in each grid cell associated with these 'contaminated' samples, $m_{i,j,}$ are determined. The PSCF for grid cell *i,j* (derivation is detailed in Hopke et al., 1995), is given by:

$$PSCF_{i,j} = \frac{m_{i,j}}{n_{i,j}}$$

PSCF is an indication of the likelihood that a given region contributed to those measurements that had higher than a selected threshold concentration at the receptor. In the context of each chemical species, Hopke et al., (1995) suggest using the average value as the threshold level, whereas others have used the top 25% of their measurements.

One complication with our PSCF analysis is that while the aerosol data is integrated over 24-hours, back trajectories are available hourly. Consequently this integrated value has to be allocated to all the trajectories for that day despite possible changes in the direction from which air masses could arrive over that time. However, an increased number of trajectories reduce the fetch uncertainty and bootstrapping has also been investigated to improve the quality of the PSCF results (e.g. Hopke et al., 1995).

Another complication of PSCF analysis is that some cells have a limited number of trajectories passing over them in which case PSCF might erroneously report large probabilities. Common practice is to down weigh these locations.

4. Program pscf

The program reads a data matrix and generates trajectory end point counts over a rectangular region, i.e. it generates $n_{i,j}$ and $m_{i,j}$ as described in section 3. The format of the data matrix is given in Appendix 1. It contains the date of measurement as well as total mass and either elemental mass or the mass of source fingerprints such as those identified using programs such as PMF (Paatero and Tapper 1994). The end result is the generation of fetch maps such as:

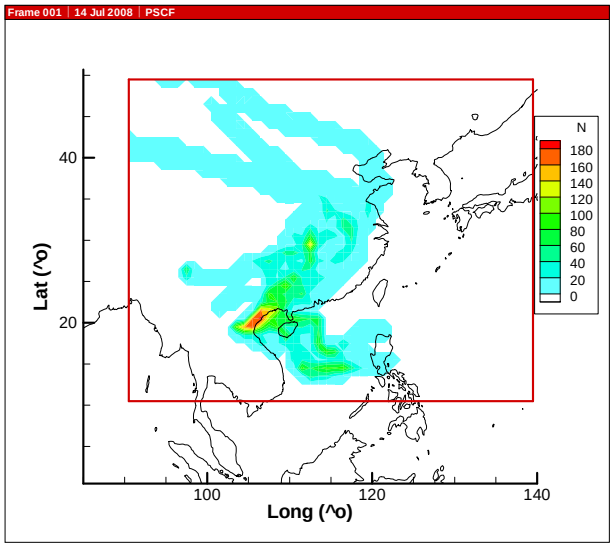


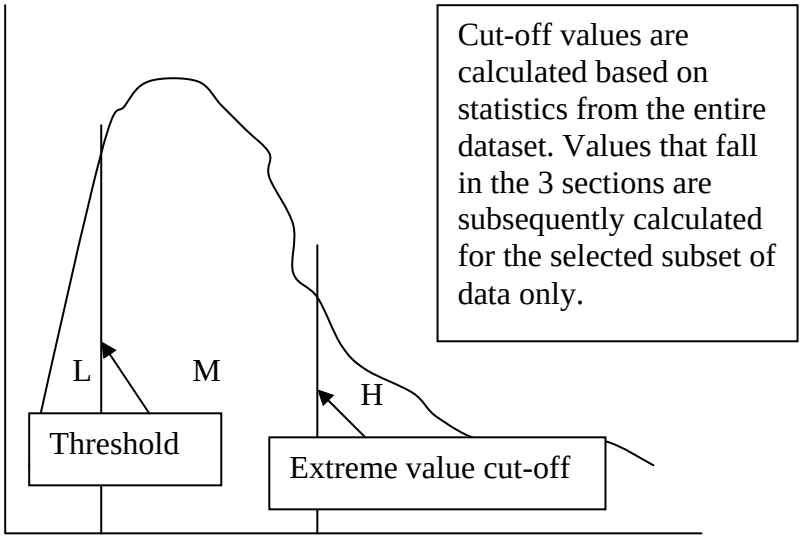
Figure 1: Trajectory coverage.

Figure 1 shows a map of the total number of trajectories crossing grid cells over the selected region (shown by the red rectangle), i.e. it plots the $n_{i,j}$s over the region.

The choice of the grid dimensions is specified in an input file called “control.txt”. This file also contains information such as the location of the study site and the directory structure for running of the program. A copy of the file “control.txt” is given in Appendix 2.

The input from the screen specifies two cut-off points which subdivide each column of the dataset into 3 segments (low, mid, and high) based on the mass. An example of the screen dialog is given in Appendix 3. The entire data is read and descriptive statistics are generated (mean, median, maximum and standard deviation) for each column entry. The user is then prompted to enter the number of standard deviations (through the “`Input m`” prompt, see Appendix 3), which is then used to compute the high end cut-off (which is calculated as the mean + value * standard deviation). The intention of this cut-off is to remove the tail from the distribution, by removing values larger than the calculated cut-off.

Once the high (or extreme) values are removed the distribution parameters are calculated for the reduced data set and a threshold value is requested for (through the “`Threshold std`” prompt). This value is again specified as a multiple of the standard deviation of the reduced dataset. The intention here is to select the peak of the distribution. Thus by specifying the two cut-off values the data space is subdivided into 3 regions.



While the cut-off values are calculated on the whole dataset the data to be processed is selected by the start date and end date as specified in file control.txt. These two dates are read in and the input matrix is searched and only those entries falling between the start date and end date (inclusive) are further considered. The number of entries in the selected subset of data falling in the 3 bands (L, M, H) is calculated and this is reported in the columns labelled (L, M, H) on the screen and in the file cutoff.txt.

5. Meteorological data files

The program requires the FNL and GDAS files to run hysplit to generate either forward or backward trajectories. For each day of the selected subset of data to analyse 24 back trajectories are generated. An intermediate file CONTROL is generated by the PSCF program which is then read in by a hysplit module called hymodelt (which is invoked by a system call). The contents of the CONTROL file are specified in Appendix 5. This file is generated by the program and used by hysplit, however, it can be useful to look at if a problem arrises.

The PSCF program constructs the file name (see end of Appendix 4 for the naming convention) for each trajectory to be generated and then checks if the file exists. If the file exists and the latitude, longitude and height at the receptor site match those required by the current run of the program it uses the existing data in the file otherwise it calls hymodelt to regenerate the file. The content and format of the generated file containing the back/forward trajectory information is given in Appendix 4.

6. Output files

Three output files are generated: pscfl.dat, pscfm.dat, pscfh.dat, which contain the cell trajectory counts for data falling in each of the three selected regions (L,M and H as specified in section 3). Note that this is for the selected subset of the input data, as specified by lines 6 and 7 in the control.txt file. Each of the 3 files has the same format. An example of pscfl.dat follows:

```
    Long     Lat      N   MassL Na      SI      S       CL      K       CA      FE
ZN      PBL
       0       0      40      40      40      40      40      40      40      40
40      40      40
       0       0     636     523     636     636     523     636     636     636
523     636     636
  130.50  -44.50       0       0       0       0       0       0       0       0
0       0       0
  130.50  -43.50       0       0       0       0       0       0       0       0
0       0       0
  130.50  -42.50       0       0       0       0       0       0       0       0
0       0       0

…..
….

   154.50  -31.50      83      40      83      83      40      83      83      83
40      83      83
  154.50  -30.50     141      85     141     141      85     141     141     141
85     141     141

…….
….
```

The first line contains the headers of the columns which are obtained from the input file “input.txt”. The 2nd line contains the column median and the 3rd line contains the column maximum. The subsequent rows give the number of trajectories that crossed the grid cell cantered at the longitude, latitude given in column 1. The column labelled “N” gives the total number of trajectories crossing the grid cell. Whereas, the subsequent columns give the number of trajectories that passed through the grid cell

for those days that satisfied the condition on mass, i.e. for file pscfl.dat the mass for each component had to be in the lowest band, for pscfm.dat the mass had to be in the middle band and for pscfh.dat the mass has to be in the high band.

7. Procedure in using PSCF

The following needs to be carried out in order to user PSCF:

1. Create file input.txt (Appendix 1)
2. Create file control.txt (Appendix 2)
3. Run the program choosing the two values separating low, medium and high mass (section 4 and Appendix 3)
4. View and plot results (Section 6).

5. References

Brankov E., Rao S.T. and Porter P.S., 1997: A Trajectory-clustering-correlation methodology for examining the long-range transport of air pollutants. Atmospheric Environment 32(9), 1525-1534.

Crawford J., Chambers S., Cohen D.dD., Dyer L., Wang T. and Zahorowski W., 2007: Receptor modelling using Positive Matrix Factorisation, back trajectories and Radon-222, Atmospheric Environment 41, 6823-6837.

Harris J.M. and Kahl J. 1990: A descriptive Atmospheric Transport Climatology for the Manua Loa Observatory, Using Clustered Trajectories. Journal of Geophysical Research 95 (D9) 13,651-13,667.

Harris J.M., Draxler R.R. and Oltmans S.J. 2005: Trajectory model sensitivity to differences in input data and vertical transport method, Journal of Geophysical Research, 110 D14109.

Hopke P.K., Li C.L., Ciszek W. and Landsberger S., 1995: The use of bootstrapping to estimate conditional probability fields for source locations of airborne pollutants. Chemometrics and Intelligent Laboratory Systems 30, 69-79.

Owega S., Khan B-U-Z, Evans G., Jervis R.E. and Fila M. 2006: Identification of long-range aerosol transport patterns to Toronto via classification of back trajectories by cluster analysis and neural network techniques. Chemometrics and Intelligent Laboratory Systems, 83 (1), 26-33.

Paatero P. and Tapper U., 1994: Positive Matrix Factorisation: A non-negative factor model with optimal utilisation of error estimates of data values, Environmetrics, Vol 5, 111-126.

Stohl A., 1998: Computation, Accuracy and Applications of Trajectories – A review and Bibliography. Atmospheric Environment Vol 32, No 6, 947-966.

8. Appendix 1 – Format of data file input.txt

The data file input contains a header line (please note that because of the length of the line each line wraps around in the following) followed by any number of data lines. Columns 1 to 3 contain the date of the measurement (day, month, year). Column 4 contains the total mass. The next columns (which can be from 4 to 12) contain the elemental mass or may contain the source fingerprint mass.

Day	Mnth	Year	CMass	Na	SI	S	CL	K	CA	FE	ZN
	PBL	BC	Soil	OMH							
3	1	2001	8129	334	117.8	1021.1	2.7	41	19.8	33.6	3.1
	0.8	1008	490.5	703							
7	1	2001	5011	431.8	44	363.8	351.6	68.3	20.8	16.5	2.6
	0.1	908.6	203.2	2079.8							
10	1	2001	4664	632.2	71.7	356.2	423.6	34.8	27.3	29	2.8
	3.2	652.2	341.2	0							
14	1	2001	8175	642.7	148.2	1173.6	29.1	106.4	32.1	47.5	7.1
	1.1	851.4	666	0							
17	1	2001	6900	480.3	91.2	968.7	47.2	49	26.2	16.2	1.1
	1.4	624	357.2	0							
21	1	2001	7818	290.1	45.4	819.2	66.2	51.6	20.5	20.7	4.4
	6.4	942.8	216.5	1219.8							
24	1	2001	19698	931.7	114.8	2750	0	98.1	42.1	44.3	17.5
	7.7	1015.6	548.7	1406.6							
28	1	2001	8243	770.8	20.1	1107.7	224.6	70.2	32.3	24.4	4.6
	5.4	625.7	178.5	1081.1							
31	1	2001	6676	156.7	13.2	383.9	27.2	29.8	11.4	18.9	7.1
	3.3	596.9	100.8	314.3							
4	2	2001	8110	475.2	35.7	917.2	19.2	32	18.8	9.9	3.4
	3.7	701.9	155.2	387.1							
7	2	2001	5216	736	5.3	355.7	490	32.1	22.8	8.4	3.6
	4.9	750.8	75.8	0							
11	2	2001	5576	187.8	10.6	734.9	36.8	28.5	10.4	11.7	3.4
	2.8	613.9	73.1	0							
14	2	2001	12269	599.5	52.1	1333.6	3.7	63.7	28.4	67.3	23.3
	4.6	1494.2	366.9	1379.7							
18	2	2001	3075	0	26.7	282.1	51	27.6	11.1	13.7	3.7
	3.4	703.9	131.8	0							
21	2	2001	4386	262.4	25	595.4	5.4	29.4	12.5	9.1	1.1
	1.5	604.4	117.9	0							

9. Appendix 2 - format of control file (control.txt)

The control.txt file controls the execution of the program with the following lines of input:

GRID specification

```
Line 1:  90 140 1     - min long, max long, cell size in degrees
Line 2:  10  50 1     - min lat, max lat, cell size in degrees
Line 3:  30  0        - spare value, print_tecplot – a value of 1
                        implies tecplot type output should be
                        produced, 0 results in column labels only.
Line 4:  12  8        - number of factors (4 to 12, matching the input file),
                        time_diff where
                        time_diff is specified such that
                        utc + time_diff = local time.
```

SITE information

```
Line 5: Site          - site identifier – must be 5 characters long
Line 6: 21.033 105.850 300 500  120 – latitude, longitude of site,
        arrival height of back trajectory – two heights are specified and results
        are generated for each height that is greater than 0,
        length in hours of back trajectory (if length is negative) otherwise forward
        trajectories are calculated.
```

DATA of interest

```
Line 7: 3  3 2006     - starting date of data of interest
Line 8: 30 3 2006     - end date of data of interest
```

```
Line 9: X:/2008_IsoTrans/NewPSCF/ - working directory,
Please note that you need a subdirectory "TrajhhhhSite" in the working directory,
where "Site" is the site identifier specified on Line 5 and hhhh is the height of the
trajectory being generated as specified on line6. If this directory doesn't exist the
program creates it.

Line 10: X:\hysplit4\exec\hymodelt – hymodelt executible
Line 11: n                          - 'n' for northern hemisphere, 's'
                                      for southern
Line 12: 2 2006 2007               - number of years for which
                                     fnl-files are specified,
                                     start year, end year – these have to cover the
                                       period of data specified in lines 7 to 8.
```

LOCATION of meteorological data

```
Line 13: X:/2008_IsoTrans/NH_FNL/ - FNL or GDS directory for each
                                     year on line 12, i.e. if line 12 specifies 3 years
                              then 3 lines of FNL/GDAS directories needs to be specified.
Line 14: X:/2008_IsoTrans/GDAS/
```

To generate the back/forward trajectory information the program needs to access the FNL and/or GDAS meteorological data files. If the data to be analysed is prior to 2007 than the FNL files are required, otherwise GDAS files are required. The program automatically works out which files are required and it requests 3 FNL files and 7 GDAS files for each run of the trajectory generation. If backward trajectories are being generated the current months two FNLs are needed as well as the 2nd part of the previous month. For forward trajectories instead of the 2nd part of the previous month the 1st part of the next month needed. Similarly for GDAS, but 2 files either back or forward in time are needed (as GDAS has 5 files per month). It is important that on line 12 you specify all the years for which the meteorological files will be needed (FNLs or GDAS) and follow line 12 by the same number of lines, one for each year which specify the directories where these files will be found). These should cover the periods specified in lines 7 and 8.

10. Appendix 3: Screen dialog – file cutoffs.txt

```
Site RI000
 Location    -33.616   150.748   300   300   -72
Start date    11    3 2005
  End date    30    3 2005
   13    3 2005
   16    3 2005
   20    3 2005
   23    3 2005
   27    3 2005
   30    3 2005
```

(Parameters for entire data set)

```
 Parameters for mass and each factor
Factor      MEAN    MEDIAN       STD       MAX    No     L     M     H
CMass       7365.39   5695.00   6595.91  61471.00   693     0     0     0
Na           190.00    102.50    269.40   1890.30   693     0     0     0
SI            83.81     49.30    140.12   2233.20   693     0     0     0
S            428.59    327.10    342.25   2750.00   693     0     0     0
CL           133.86     39.80    211.80   1650.30   693     0     0     0
K             67.52     43.50     75.61    620.10   693     0     0     0
CA            17.05     13.40     14.06    159.80   693     0     0     0
FE            29.63     22.50     33.46    551.30   693     0     0     0
ZN             6.47      4.30      7.61     85.60   693     0     0     0
PBL            5.27      3.40      9.09    187.30   693     0     0     0

Data larger than mean + m * std will be ignored
```

(M value input by user)

```
 input m: Value of m:      2.000
 Ignoring values for mass and each factor greater than
 CMass       20557.21
 Na            728.79
 SI            364.04
 S            1113.09
 CL            557.46
 K             218.74
 CA             45.16
```

```
 FE             96.55
 ZN             21.69
 PBL            23.45
```

(parameters after removing high data)

```
 New Parameters for mass and each factor
Factor       MEAN    MEDIAN       STD       MAX    No      L     M     H
CMass        6342.10   5495.00   3568.59  19718.00   666     0     0     0
Na            147.07     84.70    182.20    715.50   659     0     0     0
SI             67.95     46.80     62.33    339.10   677     0     0     0
S             388.95    312.30    268.83   1107.70   668     0     0     0
CL             99.18     33.10    132.37    553.80   662     0     0     0
K              53.87     41.60     39.40    207.10   660     0     0     0
CA             15.12     12.90      9.44     45.00   666     0     0     0
FE             25.85     22.00     16.85     92.60   675     0     0     0
ZN              5.46      4.20      4.37     20.80   670     0     0     0
PBL             4.39      3.20      4.37     23.30   676     0     0     0
```

(threshold selected by user)

```
 Threshold std      1.000
 Threshold for mass and each factor greater than
 CMass        9910.70
 Na            329.27
 SI            130.29
 S             657.78
 CL            231.56
 K              93.26
 CA             24.56
 FE             42.70
 ZN              9.83
 PBL             8.77
```

```
 Selected values for mass and each factor
Factor       MEAN    MEDIAN       STD       MAX    No      L     M     H
CMass        3925.00   1970.00   3525.65  11429.00     6     5     1     0
Na             17.88      0.00     39.99    107.30     6     6     0     0
SI             39.67     19.10     29.14     85.30     6     6     0     0
S             379.67    164.60    377.01   1128.80     6     5     0     1
CL             24.82      1.10     40.91    112.20     6     6     0     0
K              27.08     12.20     26.96     84.50     6     6     0     0
CA              6.03      5.70      2.72     10.50     6     6     0     0
FE             21.32     13.60     16.36     55.60     6     5     1     0
ZN              3.63      2.20      2.77      9.40     6     6     0     0
PBL             3.50      2.30      1.69      5.50     6     6     0     0
```

(The last 3 columns give the number of data points in each of the 3 bands. These are out of the 6 data points that were found in the input file (input.txt) that fell between the selected start and end dates in the control.txt file). At this point if you are unhappy with the results you can choose to go repeat the choice of the two values of m and threshold).

```
   Limits at
CMass        9910.70  20557.21
Na            329.27    728.79
SI            130.29    364.04
S             657.78   1113.09
CL            231.56    557.46
K              93.26    218.74
CA             24.56     45.16
FE             42.70     96.55
ZN              9.83     21.69
PBL             8.77     23.45
```

11. Appendix 4 – Content of back/ forward trajectory files

The trajectory files generated (file names of the form “B/F”hhhhyymmddhh and stored in directory Trajhhhhssssss, see appendix 2 for the definition of the fields of the file and directory names) contain header information containing information on the site selection for the particular run. These files are not required to be analysed by the user as they are used internally by the program.

```
     3     1
```

```
FNL      5     2    16     0     0       - FNL or GDAS
FNL      5     3    16     0     0
FNL      5     3     1     0     0
 1 BACKWARD OMEGA                        - backward or forward trajectory
 5     3    12    14  -33.620  150.750   300.0 – date, time, location and height
 1 PRESSURE    - variables prionted
```

The header information if followed by one line of information for each hour (end point) of the trajectory, thus if 240 hours were requested, 241 lines will be printed with the first line being at hour 0, i.e. at the receptor site. The date and time is given with the main information being in the last 5 columns (hours back or forward from the receptor site, latitude, longitude, height and pressure).

```
    1     1     5     3    12    14     0     0     0.0  -33.620  150.750    300.0
960.5
    1     1     5     3    12    13     0     0    -1.0  -33.449  150.897    291.4
964.9
    1     1     5     3    12    12     0     0    -2.0  -33.271  151.067    280.1
966.6
…….
……
```

Convention of naming of the trajectory files is (F|B)hhhhyymmddhh, “F” for forward, “B” for backward followed by 4 characters for the height at the receptor site followed by 2 characters each for year, month, day, hour.

12. Appendix 5 CONTROL file

The CONTROL file is generated and used by the program and is only included here as it may be useful should a problem arise.

```
02 11 27 05            - date tile
1                      - number of sites
-34.0786 149.9173    500 – lat long and height
-240                    - hours back in time if negative else forward
0
50000.0
3                      - number of FNL files
X:/2008_IsoTrans/SH_FNL/  - location of FNL file
fnl.sh.mar05.001          - name of FNL file
X:/2008_IsoTrans/SH_FNL/
fnl.sh.mar05.002
X:/2008_IsoTrans/SH_FNL/
fnl.sh.feb05.002
D:\NewPSCF\TrajRI000/   - Directory where trajectory files are stored
B030005033013           - name of trajectory files, convention of
naming is “F” for forward, “B” for backward followed by 4 characters
for the height at the receptor site followed by 2 characters each for
year, month, day, hour.
```