US Department of the Interior
National Park Service
National Center for
Preservation Technology and Training
Publication No 1996-21

EVALUATION OF NAPAP AEROMETRIC DATA

Terry J Reedy, PhD

October 1996

EXECUTIVE SUMMARY

The National Acid Precipitation Assessment Program (NAPAP) started in the early 1980s and has continued more or less to the present. At least parts of the program and resulting data have been inherited by the Materials Research Program of the recently established National Center for Preservation Technology and Training (NCPTT) of the National Park Service. One component of this inherited data is the Aerometric Data associated with the NAPAP Briquette Studies.

The problem addressed by this survey is the lack of a list of currently available aerometric data files, along with a list of their variables, formats, and completeness. This report inventories and evaluates the ASCII data files delivered to me either electronically or by MSDOS disks The intention is to help NCPTT managers make decisions on future action with NAPAP projects and data.

Aerometric data for a particular site were intended to be collected hourly for months or years. The files received divide into two groups covering 1984-1986 and 1988-1994. Each group is examined in a separate section.

Data were collected for three years (1984-1986) at four sites in Washington DC, North Carolina, New Jersey, and New York and one year (1986) in Ohio. The first batch of files cover seven of the nine total years for NC, NJ, and NY. After I made some minor adjustments, each file contains data for one year in a uniform format with 4 identifiers (site, year, day, hour) and 14 data variables Two-or three-letter codes are known for all the data variables but some (such as ‘DP’) still need interpretation. Even though there is one line for each hour of the site-years covered, individual variables are missing none to all of their values for a particular year. Section 1.4 report details for each variable.

The second group of later data has three batches of files. The ‘INDE’ batch has monthly .IND’ and .CRX’ files for 1988-1991. While the file formats are easily determined by visual inspection, the meaning of these terms and the names of the variables represented by each data column are not known to me. Files of a second batch, from Washington DC, cover various numbers of days in 1991. Again, variable names are not presently available. The third batch comprises monthly files for WDC 1989 to mid 1993 and NY 1992 to mid 1994 The names and units are known for all variables except for two status variables in the WDC sub-batch.

In all three batches in this group, most files are missing some to most of the hours during the period covered. Though some of the monthly files are complete, others are empty dummy files that merely serve as missing-month markers. In addition, individual variables are occasionally missing even when the rest of the data for an hour are present, but this has not been examined in detail.

Sections of the data with known variables are sufficiently complete to be potentially useful to someone. I believe NCPTT should consider making these

Funding for this report was provided by the National Park Services National Center for Preservation Technology and Training.

available 'as-is' to the outside scientific community. Other sections are empty or close enough to empty so as to be worthless. The batches without variable names can at best be kept on hold until such time as the variable names are obtained. Recommendations for further analyses by NCPTT would require further discussion with NCPTT and an inventory of the briquette studies.

For future projects, I recommend that NCPTT (and NPS also) attend more to the completeness and documentation of the data received Documentation includes methods of data collection and preliminary data reduction and explanations for missing data as well as the file formats, variables, and units.

CONTENTS

1 AEROMETRIC DATA, 1984-1986

1.1 AVAILABILITY

Information is available from both computer files and a blue-cover report entitled “National Park Service Environmental Database: Data Availability Plots and Precipitation Plots’ The latter plots data presence against date for each variable.

SITE		1984	1985	1986
1. DC	(Washington)	D	D	D
2. NC	(Research Triangle Park)	D#P	#	P
3. NJ	(Chester)	D*P	D*P	D P
4. NY	(Newcomb)	D#	D#P	D#P
5. 0H	(Steubenville)			D

D = aerometric data - availability plot
* = aerometric data computer file
P = precipitation data plot

It appears that OH did not start until 1986 NY had precipitation plots for Nov and Dec 1984 The report has explicit notation that precipitation data was not available where not given (except for OH until 86).

1.2 FILE FORMAT

For each site and year combination, there is one 80 character (char) line for each hour of each day. Each line has 4 identifiers and 14 data variables (with a status flag for each) Here is an example (the first line from NC 85) :

285 001 00 -004 0014 -001 0012 0006 0115 0010 0097 0067 0817 0000 0000 0116 9999

Colmn	Sz	Variable
1	1	Site: 2=NC, 3=NJ, OR 4=NY
2- 3	2	Year: 84, 85, or 86
4	1	(blank)
5- 7	3	Day: 001 to 365 (or 366 in 1984
8	1	(blank)
9-10	2	Hour: 00 to 23

11-15	5	Status (1 column) + data (4 (columns)
16-20	5	“ “
21-25	5	“ “
26-30	5	“ “
31-35	5	“ “
36-40	5	“ “
41-45	5	“ “
46-50	5	“ “
51-55	5	“ “
56-60	5	“ “
61-65	5	“ “
66-70	5	“ “
71-75	5	“ “
76-80	5	“ “

Site, year, day, and hour are obvious line identifiers. The names of the 14 aerometric data variables are those used in the data availability charts Comparing the pattern of missing values in the files with the blank spaces on the charts verifies that both display the variables in the same order (with the charts starting with S02 at the bottom). The meaning of WDA to DP and PR to WSV is for someone else to determine.

A letter in the status column apparently indicates that the following value is dubious. In the cases I checked, such values were are treated as missing on the availability charts. Missing values are indicated by 9999 in the data columns. The status column is blank for valid values and, where checked, for missing values.

1.3 FILE PREPARATION

As delivered, the files were almost but not quite ready for computer analysis. After analyzing them for ‘deviations’ with Python language scripts, I modified copies with a text editor as needed.

A. Each file started with a text line identifying the file by site and year. Since these are redundant and would interfere with analysis, I removed them The 2 NJ and 3 NY files each represented one calender year of data. The 4 NC files each represented exactly 1/4 of two years of data Because 1984 was a leap year, the split between the 2nd and 3rd came on Dec 31, 1984 I combined and resplit the data to get two files with one calender year each, as with NJ and NY.

B. 1984 has 366 x 24 8784 hours, while 1985/6 have 365 x 24 = 8760 Each file should have the corresponding number of lines NY84 had about 110 too many. The problem was that day:hour lines 188:12 to 193:23 appeared twice. Since the two chunks had complementary non-overlapping sets of variables present (not missing) I consolidated them into one chunk with all the data available.

C. The day:hour lines should run in a steady sequence from 001:00 to 365:23 (or 366:23) In NC85, 184:05 appeared twice (with identical values), while 184:06 was missing. I deleted the duplicate line and added a 184:06 line with the average of values from 184:05 and 184:07.

D. Each line should be the same length: 82 = 80 data chars + 2 control chars (the carriage return + linefeed at the end of each line). Total file sizes should

then be 8760 x 82 = 718,320 or 8784 x 82 = 720,288. They are now.

1 4 VARIABLE SUMMARIES

To quantify both the presence and range of values, I ran the seven files through a BMDP1d analysis and summarize them below.

			Status: %			Value		
Var	Site	Yr	OK	Dubious	Missing	Min	Mean	Max
1. so2	NC	84	90	3	7	-25	3	41
		85	82	6	12	-76	3	78
	NJ	84	84	1	15	1	9	121
		85	89	1	10	1	6	188
	NY	84	97	0	3	0	3	43
		85	98	0	2	0	2	28
		86	98	0	2	0	2	29
2. no2	NC	84	94	7	5	0	12	120
		85	97	2	1	-42	14	148
	NJ	84	90	0+	10	3	17	108
		85	90	0+	10	1	14	118
	NY	84	90	0	10	0	1	56
		85	73	0	27	0	1	27
		86	65	0	35	0	2	21
3. no	NC	84	94	1	5	-138	10	351
		85	97	2	1	-44	7	302
	NJ	84	88	0+	12	1	6	122
		85	90	0+	10	1	9	279
	NY	84	90	0	10	0	0+	34
		85	73	0	27	0	1	62
		86	66	0	34	0	1	8
4. nox	NC	84	94	1	5	-141	22	372
		85	99	0+	1	-17	21	305
	NJ	84	87	1	12	6	23	205
		85	89	1	10	2	23	397
	NY	84	90	0	10	0	2	80
		85	73	0	27	0	1	119
		86	64	0	36	0	2	26
5. o3	NC	84	90	4	6	-5	25	550
		85	94	4	2	-4	26	119
	NJ	84	87	3	10	1	30	147
		85	90	3	10	1	33	126
	NY	84	96	1	3	0	31	127
		85	96	1	3	0	29	95
		86	96	1	2	0	30	99
6. wda	NC	84	84	8	8	0	174	360
		85	87	7	6	0	175	360
	NJ	84	0	0	100	-	-	-
		85	44	4	52	0	68	360
	NY	84	0	0	100	-	-	-

		85	0	0	100	-	-	-
		86	0	0	100	-	-	-
7. wsa	NC	84	97	0+	1	0	16	81
		85	100	0+	0	0	15	100
	NJ	84	71	29	0+	0	58	825
		85	68	29	3	0	60	825
	NY	84	64	0	36	0	7	80
		85	99	0	1	0	9	69
		86	95	0	5	0	9	97
8. tp	NC	84	97	0+	3	-127	152	350
		85	98	1	1	-211	153	348
	NJ	84	89	1	10	-180	93	349
		85	93	1	6	-238	96	459
	NY	84	66	1	23	-283	46	316
		85	97	2	1	-283	70	339
		86	87	9	4	-278	64	346
9. dp	NC	84	96	1	3	-261	66	231
		85	73	1	26	-335	77	227
	NJ	84	26	0+	74	-182	16	369
		85	89	1	10	-291	13	461
	NY	84	0	0	100	-	-	-
		85	0	0	100	-	-	-
		86	91	7	2	-178	13	258
10.rh	NC	84	97	0+	3	109	618	1000
		85	74	0	26	119	614	1070
	NJ	84	84	0+	16	207	724	1026
		85	0	0	100	-	-	-
	NY	84	56	0+	44	177	790	997
		85	0	0	100	-	-	-
		86	92	0+	8	120	716	1000
11.pr	NC	84	95	0+	5	0	0+	25
		85	99	0	1	0	0+	1
	NJ	84	99	0+	1	0	0+	31
		85	96	0+	4	0	0+	78
	NY	84	14	0+	84	0	1	254
		85	98	0+	2	0	0+	28
		86	97	0+	3	0	0+	25
12.sr	NC	84	96	1	3	0	1021	7600
		85	98	0	2	0	228	1367
	NJ	84	98	1	1	0	1210	8180
		85	95	1	5	0	1150	8390
	NY	84	8	0	92	0	200	3825
		85	95	3	2	0	1130	9618
		86	85	3	15	0	1043	9875
13.wdv	NC	84	65	5	30	0	173	360
		85	87	7	6	0	177	360
	NJ	84	93	6	1	0	231	360

		85	90	6	4	0	229	360
	NY	84	13	2	85	0	212	359
		85	88	10	2	0	217	458
		86	85	9	6	0	177	360
14.wsv	NC	84	0	-	-	-	-	-
		85	0	-	-	-	-	-
	NJ	84	71	28	1	0	58	825
		85	68	28	4	0	59	825
	NY	84	15	0	85	0	9	80
		85	96	0	4	0	9	69
		86	94	0	6	0	8	96

Status ‘dubious’ means flagged with a status code.
Missing means coded ‘9999’.
0+ means rounded down to 0.
NY did not begin to record some variables until Nov 84.

1.5 DATA QUALITY

These seven files now have the form required for any further analysis. They appear to be usable, with care, but are far from ideal.

The percentage of good values ranges from 0 to 100. When missing values are scattered rather than concentrated, some can be eliminated by averaging (if allowable by the type of analysis done). Then, 90% good or even lower should be usable.

A potential problem for multisite analysis is protocol inconsistency. NC recorded negative concentration readings for the first five variables; the other two sites did not. NJ recorded WSA(7) and WSV(l4) about 10 times higher than NY Let us hope that this is simply a shift of the decimal point.

Some analyses of the aerometric variables considered as dependent on time of day, time of year, and place should be possible. However, since this is (I believe) outside the interest of NCPTT, this might best be done by making the data available to appropriate researchers at other institutions. Announcements could be posted on appropriate Internet newsgroups.

It should also be possible to use these variables as independent variables affecting briquette outcomes. This is (I believe) the main interest of NCPTT in these data Any detailed recommendation would require a survey of that data also

I cannot comment on existing analyses since I have not received any.

2 AEROMETRIC DATA, 1988-1994

2.1 INDE, 1988-1991

Self -extracting archive INDEDATA.EXE expands into 91 files labelled YYMM.IND or YYMM.CRX. There is one file of each type for months 8804 to 9110 There are five more CRX files for 9111 (blank), 9112, 9201, 9301 (blank) , and 9302. Both types of files are labelled with day of the year and hour, so both presumably

contain aerometric data I do not know what INDE stands for.

Logical lines in the .IND files are split into two physical lines of 78 and 49 chars for a total length of 78+49+4 = 131 chars There are 13 data fields, including 2 which are either 1 or –1. Example:

```
 8114       0 11 87    11 97     11 86     12.36  21.7   21.5     0.000    0 000
  0.447     0 950 –1.  -1.        21.1     20.3  14.26
```

Lines in the .CRX files have 7 data fields (64+2=66 chars). The midnight reading is labelled 2400 instead of 0, as in all other files Example:

8113, 2400, 12.00, 20.2, 0.0, 3.18, 2.4, 4.43, 14.300

I have no information about the meaning of the 22 data fields. The two which are +-1 are presumably status indicators of some sort. Others might be guessed by comparing their ranges and patterns with those of known variables in other batches of files. However, this would be a dubious basis for any analysis.

If each logical line is exactly the same length, the file size divided by that length is the number of lines in the file. This should be 24 times the number of days in the month. The following table shows how much the files deviate from having exactly the right number.

	.IND	Lines	.CRX	Lines
Date	Bytes	off	Bytes	off
8804	26,069	-521	13,200	-520
8805	51,876	-348	44,022	-77
8806	99,953	43	47,520	0
8807	72,967	-187	49,104	0
8808	76,766	-158	31,416	-268
8809	94,320	0	46,596	-14
8810	97,464	0	12,276	-558
8811	94,320	0	47,520	0
8812	100,608	*24*	49,104	*0*
8901	97,464	0	49,104	0
8902	88,032	0	41,844	-38
8903	100,608	24	49,104	0
8904	*90,914*	*-26*	46,266	*-19*
8905	97,333	-1	49,104	0
8906	94,320	0	47,520	0
8907	64,583	-251	49,104	0
8908	30,523	-511	28,512	-312
8909	74,408	-152	15,048	-492
8910	95,106	-18	46,002	-47
8911	4,847	-683	47,520	0
8912	186,544	680	50,688	24
9001	97,464	0	47,520	-24
9002	88,032	0	44,352	0
9003	91,176	-48	49,104	0
9004	60,653	-257	47,520	0
9005	96,023	-11	12,012	-562
9006	94,320	0	47,520	0
9007	97,464	0	46,596	-38

9008	91,202	-2	49,104	0
9009	89,997	-33	47,322	-3
9010	69,037	-217	49,104	0
9011	91,176	-24	47,190	-5
9012	97,464	0	46,926	-33
9101	97,464	0	38,676	-158
9102	88,032	0	44,286	-1
9103	57,116	-308	48,180	-14
9104	113,315	145	42,504	-76
9105	76,242	-162	49,104	0
9106	87,246	-54	47,520	0
9107	97,464	0	40,392	-132
9108	97,464	0	48,840	-4
9109	70,609	-181	47,455	-1
9110	89,866	-58	8,910	-609
9112			44,682	-67
9201			8,646	-613
9302			8,778	-539

Every file has an even number of lines (no fractions), which suggests that all are the same length, and which accords with visual spot checks. (The alternative is that deficiencies exactly match surpluses.)

Some files have extra data. In most cases, it should be moved to a neighboring month to make up a deficiency. Assuming that this is always the case (exceptfor 8812,) there are an average of 87 lines (3 6 days) missing from each file from 8804 to 9110.

2.2 WDC/NY, 1989-1994

Self-extracting archive WDCLDATA EXE expands into 63 files. The names of 38 are YYMM.WDC where YY and MM are the year and data. They cover 8812 (from Dec 8, noon) to 9201 (up to Jan 6, 4 pm) Each 91 (89÷2) char line begins with a day and hour (X100) identifier like these:

```
 343,  1200,
 344,     0,
```

Each line continues with 8 data fields and 2 1/-1 codes like these:

```
   10.07,  43.77,  498.40,   0.43, 345 6,   0.00     12.100,   12.230,  1,  1
    5.84,  50.68,    0.00,   1.14, 354.7,   0.00      4.980,    5.551,  -1, -1
```

Compressed file DCMET.ZIP contain 68 data files also named YYMM.WDC and with the same format for each line. They cover 8812 to 9407, with many near the end being empty Those from 8812 to 9104 exactly match the previous files in length Those from 9105 to 9201 always differ from the previous files. Those after 9201 are new.

Compressed file NYMET.ZIP contains 30 files named YYMM.NYM running from 9111 to 9404. The format for each 98 (96+2) char line is nearly the same as for the WDC files. Year is added and the two +-1 flags are replaced with regular values. values represent hourly corrected means.

Colmn	Size	Variable
1-2	2	Year: 1 to 4 (last digit, NY only)
3-5	3	Day: 001 to 365 or 366

9-12	4	Hour: 0, 0100 to 2300
15-20	6 2	Temperature, air: deg C
23-28	6 2	Relative humidity: % (-99 = missing)
31-37	7 *2*	Solar radiation: (PAR) w/m2
40-45	6 2	Wind speed: m/s
48-52	5 1	Wind direction: 0 to 359 1 degrees (90 0 = missing?)
55-60	6 2	Precipitation: mm (total, not average)
64-71	8.3	Temperature, stone briquette, upper surface: deg C
74-81	8.3	Temperature, stone briquette, lower surface: deg C
84-89	6 2	Temperature, data logger case: deg C (NY only)
92-96	5 2	Reference supply voltage: vdc (NY only)
84-85	2	Unknown indicator: -1, 1 (WDC only)
88-89	2	Unknown indicator: -1. 1 (WDC only)

The information above is from Don Gatz via Mary Striegel. Columns with commas and blanks are not listed. A size entry such as 6 2 indicates 6 columns with 2 after the decimal point I hypothesize that a ‘mean wind direction’ of 90 0 actually indicates a missing or indeterminate value since a large fraction of entries have exactly this value, usually for several consecutive hours.

As with the INDE date, file sizes divided by line length (89 or 98) come out even, indicating that lines are individually correct. The table below indicates how much each file deviates from completeness.

	.WDCa	Lines	.WDCb	Lines	.NYM	Lines	
Name	Bytes	off	Bytes	off	Bytes	off	
8812	51,233	-1	same	same	(after 343:12 + Dec 8 noon)		
8901	67,704	0	” ”	” ”			
8902	61,152	0	.	.			
8903	64,701	-33	.	.			
8904	65,520	0	.				
8905	69,888	24	.	.	(April 30)		
8906	60,333	-57					
8907	61,425	-69					
8908	64,428	-36					
8909	52,689	-141					
8910	67,704	0					
8911	65,611	1					
8912	63,973	-41					
9001	67,704	0					
9002	61,152	0					
9003	67,704	0					
9004	64,701	-9					
9005	66,339	-15					
9006	63,336	-24					
9007	57,421	-113					
9008	53,690	-154					
9009	36,491	-319					
9010	57,057	-117					
9011	58,422	-78					
9012	63,700	-44					
9101	60,788	-76					

9102	60,060	-12				
9103	42,770	-274				
9104	0	—720	“	“	“	“
9105	0	-744	455	-739		
9106	58,786	-74	65,520	0		
9107	60,333	-81	60,242	-82		
9108	65,520	-24	65,520	0		
9109	89,544	264	87,906	222		
9110	69,888	24	67,704	0		
9111	67,704	24	65,520	0	33,026	-383
9112	65,065	-29	66,612	-12	72,814	-1
9201	8,463	-43	51,051	-183	72, 912	0
9202			63,336	24	68,208	24
9203			64,428	-36	72,912	0
9204			35,126	-334	70,560	0
9205			67,704	0	72,912	0
9206			65,520	0	70,560	0
9207			12,194	-610	27,244	-466
9208			10,920	-624	0	-744
9209			65,793	3	54,586	-163
9210			67,704	0	66,444	-66
9211			65,520	0	69,286	-13
9212			67,704	0	63,896	-92
9301			66,430	-14	59,486	-137
9302			54,236	-76	58,310	-77
9303			14,469	-585	56,546	-167
9304			0	-720	46,158	-249
9305			29,848	-416	0	-744
9306			2,639	-691	5,292	-666
9307			0	-744	64,974	-81
9308			0	-744	38,710	-349
9309			0	-720	63,798	-69
9310			0	-744	67,130	-59
9311			0	-720	70,462	-1
9312			0	-744	59,584	-136
9401			0	-744	43,022	-305
9402			1,729	-653	9,114	-579
9403			0	-744	38,416	-352
9404			0	-720	7,840	-640
9405			0	-744		
9406			10,283	-607		
9407			24,934	-470		

As in the previous section, positive numbers indicate extra lines, which can easily be removed, or possibly transferred to another file to remedy a deficiency. The amount missing in other files ranges from none to all. Counting surplus files as having a deficiency of 0, the 36 months of WDCa files (1989, 90, 91) are missing 3284 lines total. This is about 91 hours or 3 8 days per month WDCb is only slightly different WDCb 9201 to 9302 lack 1877 lines, which is 134 hours or 5 6 days per month. Thereafter, the last 17 WDCb files are either empty place fillers or near so.

2.3 WDC, 1991

Another 12 files extracted from WDCLDATA.EXE, with a different format, have the name YYMMWDC.DAT or YY###***.WDC, where YY and MM are month and year as before and ### and *** are day numbers (A 13th file named 91313322.NYM is present by mistake) With one exception, the year is 91 and the days range from May to Nov. As will be seen below, the day numbers overlap, so we may surmise that the second day number is actually the day after the file stops.

In these files, each logical line of 18 data fields occupies three physical lines with the following format (| indicates the end of line after the trailing spaces).

```
01+0113.   02+0313.   03+1600.   04+1.359    |+on the same line
05+41.59   06+119.9   07+0.283   08+0.141  |
09÷269.1   10+57.45   11+0.000   12+0.000    |+on the same line
13+11.62   14+14.27   15+159.5  16+0.000  |
17+4.506   18+21.13  |
```

Each value is preceded by a field number: 01 to 18 The first, 113 for WDC, 111 for NYM, seems to be a year/site code. The second and third are the day and hour. Sometimes, the field number is replaced by ‘xx’, which likely indicates a dubious or missing value. In such cases, there may be a numerical value, presumably dubious, or there may be a ‘77?????’ missing value indicator.
There are 79+79+20+6 = 184 characters per logical line. Dividing the file size by 184 gives the number of logical lines. Since dividing again by 24 usually gives about the correct number of days, judging from the filename, each line appears to represent one hour, as in previous files.

Filename	Bytes	Hrs	Days	Exp	
91123152 WDC	3,087	17-	7	29	1 reading/day instead of /hour
91152177 WDC	173,144	941	39 2	22	filename is not accurate
91152191 WDC	173,117	941 -	39 2	39	
91191203 WDC	53,360	290	12 1	12	
91203219 WDC	70,656	384	16	16	
91219240 WDC	92,552	503	21 0 -	21	
91240250 WDC	42,872	233	9 7	10	
91250270 WDC	88,136	479	20 0 -	20	
91270289 WDC	85,192	463	19 3	19	
91289320 WDC	135,976	739	30 8	31	
9107WDC DAT	27,600	150	6 2	31?	
9201WDC DAT	17,112	93	3.9	?	

Exp is the number of days expected from the filename (*** - ###).

File 152-177, with 941 logical lines and no missing data, is 39 2 days instead of 22. The next file, 152-191 starts with the same data, but about halfway through has many dubious or missing readings and many physical lines that are not the right length, with a net deficit of 27 bytes compared to what it should be for 941 hours (or a surplus relative to 39 days) If I were to use these data, I would probably delete the 152-191 and rename 152-171 as 152-191.

2 4 DATA QUALITY

The INDE files include many with complete or nearly complete monthly data. They might be usable if NCPTT obtains the identity of the variables. The same would

apply to the nearly complete 6 months of the WDC91 files after some editing to remove duplicates and identify gaps.

Except for some gaps, the longer-term WDC data, from 8812 to 9302, are relatively complete. Since we know most of the variables, they should be usable. However, the unknown +-l indicators are somewhat bothersome. They could indicate data quality, but my uncertain guess, from those I looked at, is that they indicate something else. The last 17 months of this series is worthless.

The corresponding NY series has 7 continuous months of complete data near the beginning and some usable months thereafter. The values I checked are consistent with the information provided on the identity of variables and their measurements units. In this and the other series, data for missing hours, when not too numerous, could be estimated from seasonal and daily patterns and neighboring values.