AERMOD Model Evaluation

EPA-454/B-23-011
October 2023

AERMOD Model Evaluation

U.S. Environmental Protection Agency
Office of Air Quality Planning and Standards
Air Quality Assessment Division
Research Triangle Park, NC

Table of Contents

Section	Page
Table of Contents	iv
Figures	vi
Tables	vii
1. Introduction	1
2. Database descriptions	1
2.1. Martin's Creek	3
2.2. Tracy Power Plant	5
2.3. Lovett Power Plant	7
2.4. Westvaco Mill	9
2.5. Duane Arnold Energy Center	11
2.6. Experimental Organically Cooled Reactor	13
2.7. Alaska North Slope	14
2.8. Prairie Grass	16
2.9. Indianapolis	18
2.10. Kincaid	21
2.11. AGA	23
2.12. Millstone Nuclear Power Plant	24
2.13. Bowline	26
2.14. Baldwin Power Plant	27
2.15. Clifty Creek Power Plant	29
3. Evaluation methodology	31
3.1. AERMET/AERMOD comparisons	31
3.2. Evaluation procedures	32
3.2.1. Robust highest concentrations	32
3.2.2. EPA Protocol for determining best performing model	32
3.3. Results	36

3.3.1. Turbulence cases 36
3.3.2. Non-turbulence cases 39
3.3.3. Statistical evaluations 40
4. Summary/Conclusions 49
5. References 50

Figures

Figure **Page**

Figure 1. Martin's Creek study area. 4

Figure 2. Tracy power plant study area. 6

Figure 3. Lovett study area. 8

Figure 4. Westvaco study area. 10

Figure 5. DAEC study area (SF6 releases). 12

Figure 6. Terrain map featuring the entire EOCR grid with the source at the grid center (SF6 releases). Arcs are at distances of about 40, 80, 200, 400, 800, 1200, and 1600 m. 13

Figure 7. Depiction of Alaska North Slope Oil Gathering Center turbine stack, meteorological tower (X), and camera locations used to visualize plume rise. 15

Figure 8. Prairie Grass study area 17

Figure 9. Map showing the location of the Perry-K Station (A), the Hoosier Dome (B), and the central Indianapolis business district (C). The downtown surface meteorological site is located at (D) and the "bank tower" site was on top of the building at (E). 19

Figure 10. Indianapolis meteorological sites and emissions site (Perry K Station). 20

Figure 11. Kincaid study area. 22

Figure 12. Plan view of the locations of tracer samplers at Site 1, AGA field study (SF6 releases). 23

Figure 13. Millstone study area (SF6 and freon releases). 25

Figure 14. Bowline Point study area (SO2 releases). 26

Figure 15. Baldwin study area. 28

Figure 16. Clifty Creek study area. 30

Tables

Table **Page**

Table 1. AERMOD evaluation databases used for comparisons of AERMOD 22112 and AERMOD 23132. Databases in gray are also subject to the EPA's protocol for determining best performing model. 2

Table 2. Hourly, 3-hour, and 24-hour RHC for turbulence cases. 38

Table 3. Hourly, 3-hour, and 24-hour RHC for non-turbulence cases. 39

Table 4. Composite Performance Measure (CPM) for turbulence cases. 40

Table 5. Composite Performance Measure (CPM) for non-turbulence databases. 41

Table 6. Martins Creek Model Comparison Measure (MCM) results. 43

Table 7. Lovett Model Comparison Measure (MCM) results. 44

Table 8. Westvaco Model Comparison Measure (MCM) results. 45

Table 9. Kincaid Model Comparison Measure (MCM) results. 46

Table 10. Bowline Model Comparison Measure (MCM) results. 47

Table 11. Baldwin Model Comparison Measure (MCM) results. 47

Table 12. Clifty Creek Model Comparison Measure (MCM) results. 48

1. Introduction

This evaluation presents a benchmark of model performance based on the original field studies presented in Cimorelli, et al, 2005 and Perry, et al, 2005. The evaluation focused on the performance of the 23132 version of the AERMOD modeling system compared to the previous version, 22112. The statistical analysis determines the best performing version of the model for 15 of the original 17 databases, including the adjust u* option[1] formally adopted as a regulatory option in the version 16216r of AERMOD.

2. Database descriptions

The 15 databases used in this evaluation are briefly described in this section and summarized in Table 1. The stack heights, terrain complexity, urban/rural status, importance of downwash, inclusion of turbulence parameters and meteorological data included for the database are listed for each area. A more complete description of these databases can be found in U.S. EPA, 2003. The databases are arranged by the following hierarchy: Two categories of turbulence inclusion (inclusion of turbulence or no turbulence). Within each of those categories, databases were ordered by complexity of terrain (complex or flat), and within those two categories, databases were ordered by increasing height.

[1] The adjust u* option accounts for low wind speeds when calculating u* in AERMET.

Table 1. AERMOD evaluation databases used for comparisons of AERMOD 22112 and AERMOD 23132. Databases in gray are also subject to the EPA's protocol for determining best performing model.

Location	Stack heights	Urban/ rural	Terrain	Downwash	Turbulence parameters	Site specific AERMET inputs
Martins Creek	59, 76, 183 m	Rural	Complex	Yes	10 m σ_v, σ_w	10m wind, temperature; 90-420 m wind (every 30 m).
Tracy	91 m	Rural	Complex	No	σ_v, σ_w	10 and 50-400 m (every 25 m) wind, temperature
Lovett	145 m	Rural	Complex	No	σ_v, σ_w	10, 50, and 100 m wind, temperature
Westvaco	190 m	Rural	Complex	No	σ_v, σ_w	30, 210, 326, 366, and 416 m wind, temperature[2]
DAEC	1 m, 24 m, 46 m	Rural	Flat	Yes	σ_v	Insolation, 10, 23.5 and 50 m wind, temperature
EOCR	1, 25, 30 m	Rural	Flat	Yes	σ_v	4, 10, and 30 m wind, temperature
Alaska	39.2 m	Rural	Flat	Yes	σ_v, σ_w	33 m wind, temperature
Prairie Grass	0.46 m	Rural	Flat	No	2 n σ_v, σ_w	1, 2, 4, 8, and 16 m temperature, 1 m wind, u*, mixing height, sky cover
Indianapolis	84 m	Urban	Flat	No	σ_v, σ_w	Station pressure, net radiation, 10 m wind, temperature
Kincaid	187 m	Rural	Flat	No	σ_v, σ_w	Net radiation, insolation, 10, 30, and 50 m wind, temperature
AGA	9.8, 14.5, 24.4 m	Rural	Flat	Yes	None	10 m wind and temperature
Millstone	3 stacks 29 m (freon) 48 m (SF6)	Rural	Flat	Yes	None	10 m wind speed; 43.3 m wind and temperature
Bowline	2 stacks 86.87 m	Rural	Flat	Yes	None	100 m winds and temperature
Baldwin	3 stacks 184.4 m	Rural	Flat	Yes	None[2]	10 and 100 m wind, temperature
Clifty Creek	3 stacks 207.9 m	Rural	Flat/Elev	No	None	10 m temperature; 60 m wind

[2] 30 m observations removed from AERMOD profile before running AERMOD.

2.1. Martin's Creek

The Martins Creek Steam Electric Station is located in a rural area along the Delaware River on the Pennsylvania/New Jersey border, approximately 30 km northeast of Allentown, PA and 95 km north of Philadelphia, PA (Figure 1). The area is characterized by complex terrain rising above the stacks. Sources include multiple tall stacks ranging from 59 to 183 m in height, including Martins Creek and three background sources located between 5 and 10 km from Martins Creek. The seven SO_2 monitors were located on Scotts Mountain, which is about 2.5 - 8 km southeast of the Martins Creek facility. On-site meteorological data covered the period from May 1, 1992 through May 19, 1993. Hourly temperature, wind speed, wind direction, and sigma-theta (standard deviation of the horizontal wind direction) at 10 m were recorded from an instrumented tower located in a flat area approximately 2.5 km west of the plant. In addition, hourly multi-level wind measurements were taken by sound detection and ranging (SODAR) located approximately three kilometers southwest of the Martins Creek station.

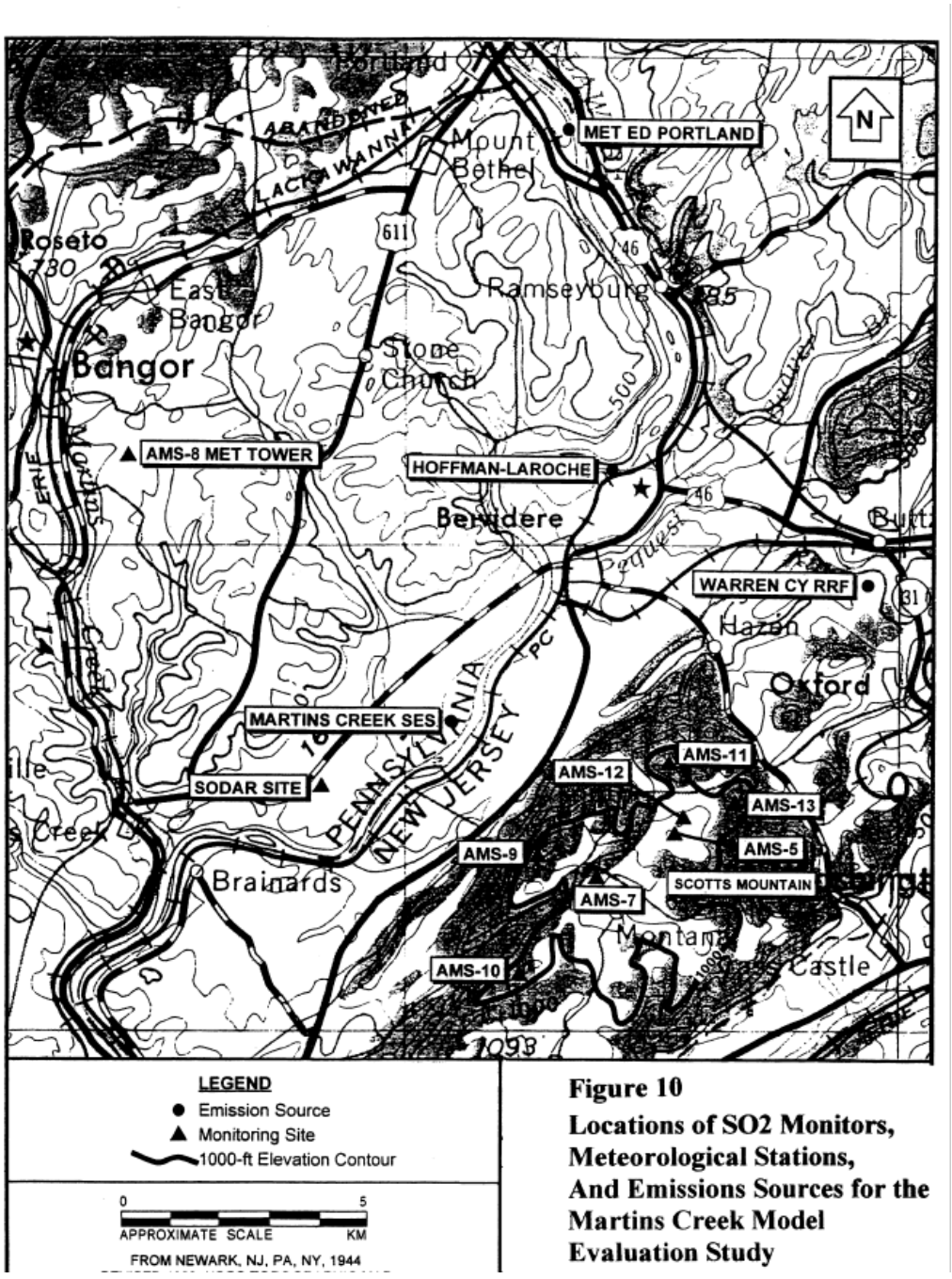


Figure 10
Locations of SO2 Monitors, Meteorological Stations, And Emissions Sources for the Martins Creek Model Evaluation Study

Figure 1. Martin's Creek study area.

2.2. Tracy Power Plant

The Tracy Power Plant is located 27 km east of Reno, Nevada in the rural Truckee River valley completely surrounded by mountainous terrain (Figure 2). A field tracer study was conducted at the power plant in August 1984 with SF_6 being released with the moderately buoyant plume from a 91-m stack. A total of 128 hours of data were collected over 14 experimental periods. Stable atmospheric conditions were dominant for this study. Site-specific meteorological data (wind, temperature, and turbulence) for Tracy were collected from an instrumented 150-m tower located 1.2 km east of the power plant. The wind measurements from the tower were extended above 150 meters using a Doppler acoustic sounder and temperature measurements were extended with a tethersonde.

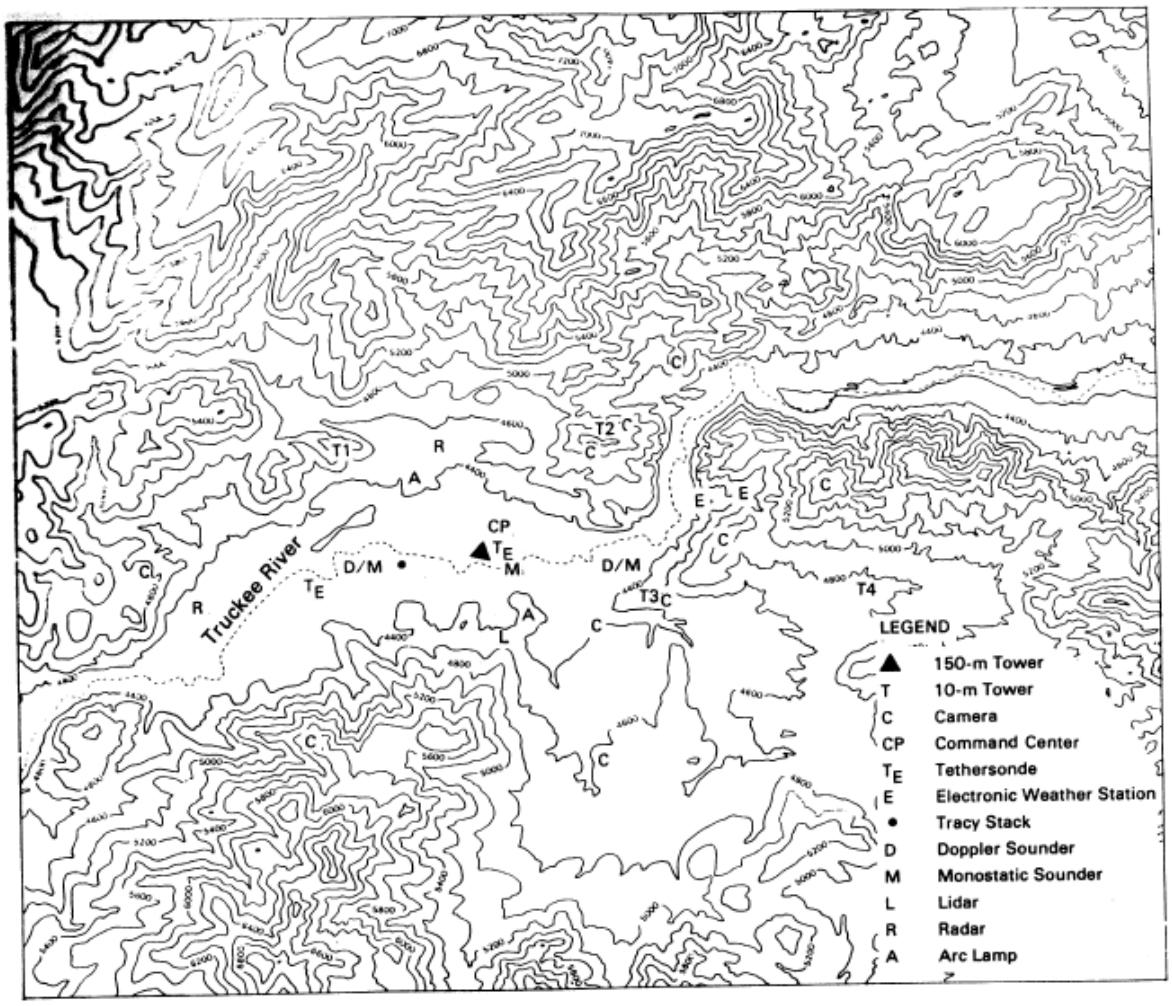


Figure 2. Tracy power plant study area.

2.3. Lovett Power Plant

The Lovett Power Plant study consisted of a buoyant, continuous release of SO2 from a 145 m tall stack located in complex terrain, rural area in New York State (Figure 3). The data spanned one year from December 1987 through December 1988. Data were collected from 12 monitoring sites (ten on elevated terrain and two near stack-base elevation) that were located about 2 to 3 km from the plant. The monitors provided hourly-averaged concentrations. The important terrain features rise approximately 250 m to 330 m above stack base at about 2 to 3 km downwind from the stack. Meteorological data include winds, turbulence, and ΔT from a tower instrumented at 10 m, 50 m, and 100 m. National Weather Service surface data were available from a station 45 km away.

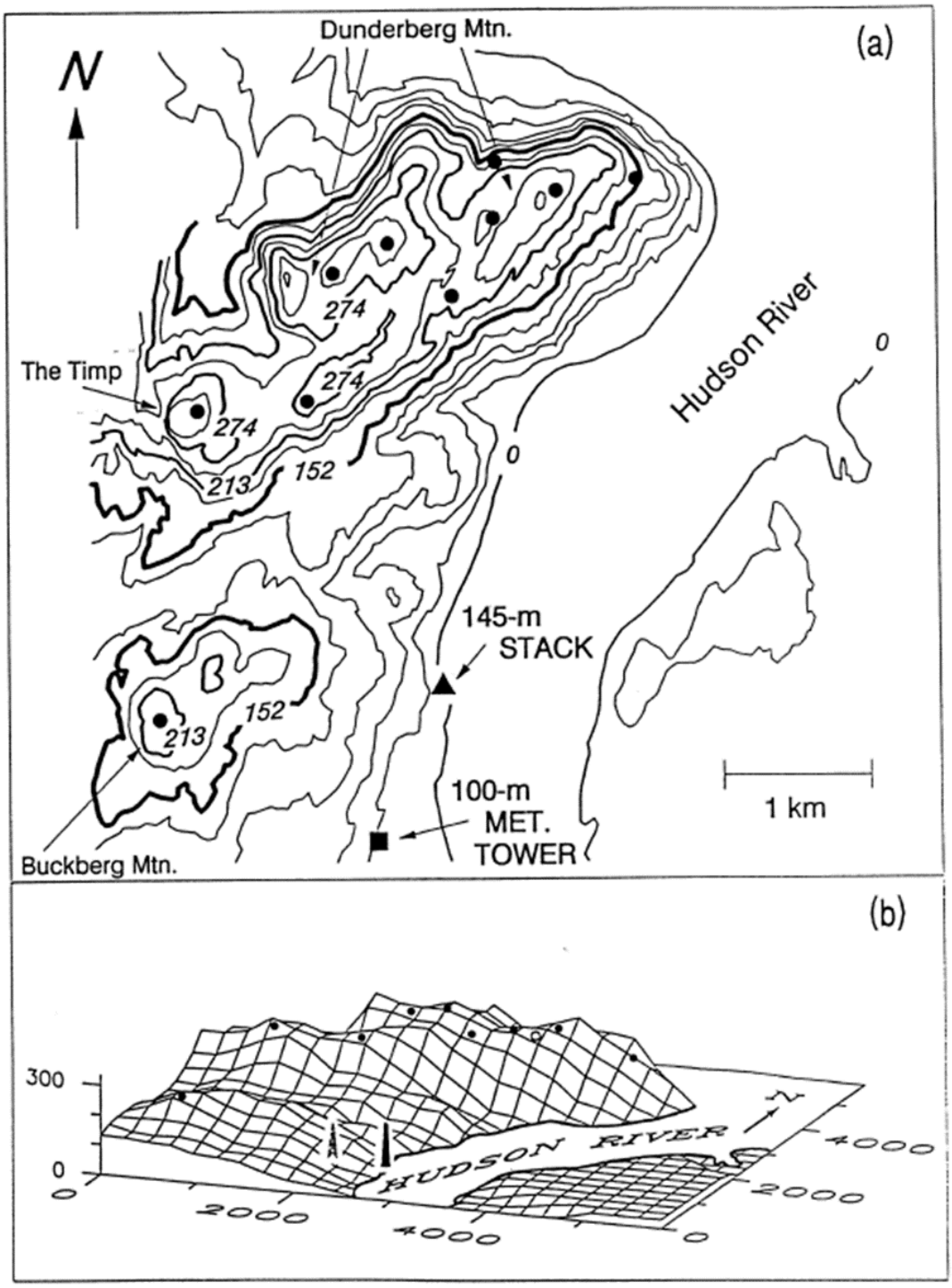


Figure 3. Lovett study area.

2.4. Westvaco Mill

The Westvaco Corporation's pulp and paper mill in rural Luke, Maryland is located in a complex terrain setting in the Potomac River valley (Figure 4). A single 183-m buoyant source was modeled for this evaluation. There were 11 SO_2 monitors surrounding the facility, with eight monitors well above stack top on the high terrain east and south of the mill at a distance of 800 - 1500 m. Hourly meteorological data (wind, temperature, and turbulence) were collected between December 1980 and November 1991 at three instrumented towers: the 100-m Beryl tower in the river valley about 400 m southwest of the facility; the 30-m Luke Hill tower on a ridge 900 meters north-northwest of the facility; and the 100-m Met tower located 900 m east southeast of the facility on a ridge across the river.

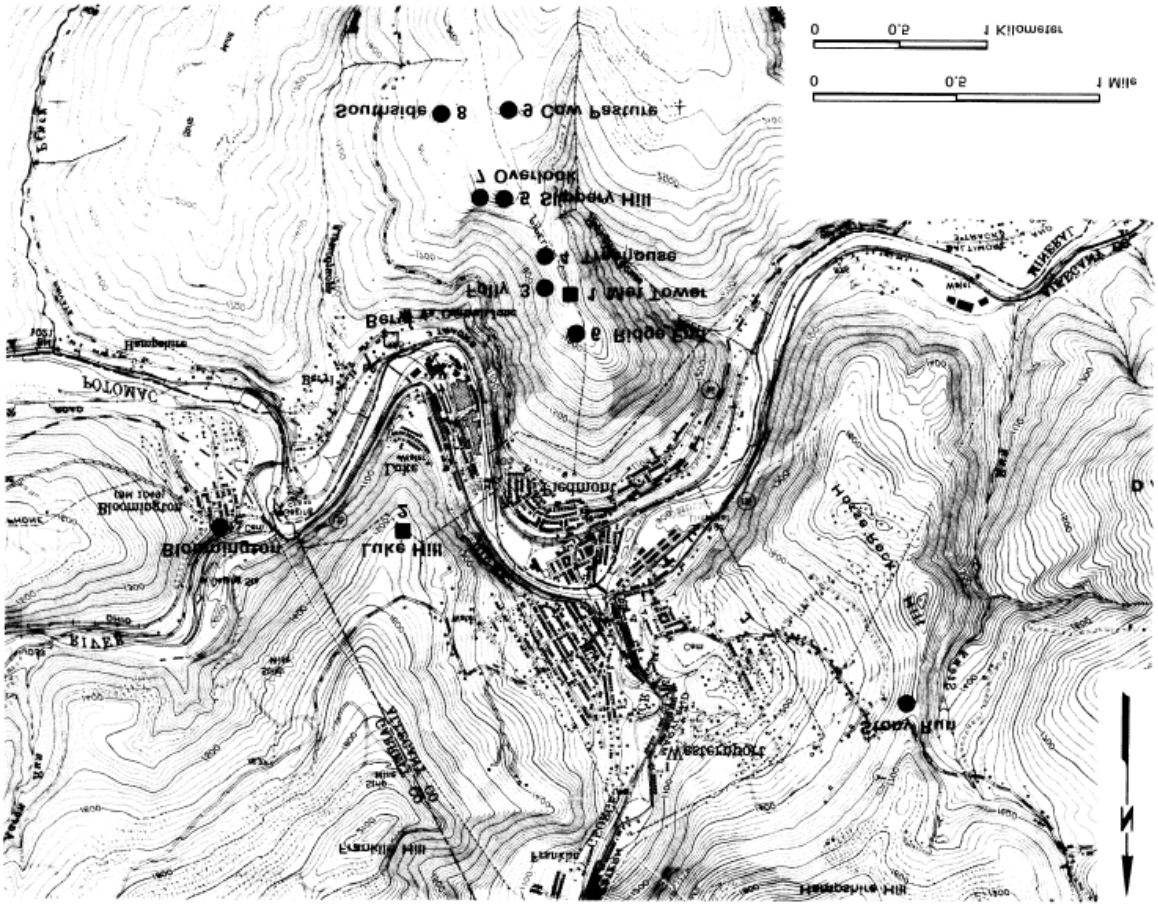


Figure 4. Westvaco study area.

2.5. Duane Arnold Energy Center

The Duane Arnold Energy Center (DAEC) is located in rural Iowa, located about 16 km northwest of Cedar Rapids. It is located in a river valley with some bluffs on the east side. Terrain varies by about 30 m across the receptor network with the eastern half of the semicircular receptor arcs being flat and the western half elevated. The tracer study[35] involved SF_6 releases from two rooftops (46-m and 24-m levels) and the ground (1-m level). Building tiers for the rooftop releases were 43 and 24 m high, respectively. The 1-m and 24-m releases were non-buoyant, non-momentum, while the 46-m release was close to ambient but had about a 10 m/s exit velocity. The number of tracer release hours was 12, 16 and 11 from the release heights of 46 m, 24 m, and 1 m, respectively. There were two arcs of monitors at downwind distances of 300 and 1000 m (see Figure 5). Meteorological data consisted of winds at 10, 24, and 50 m. The meteorological conditions were mostly convective (30 out of 39 hours), with fairly light wind speeds. Only one hour had a wind speed above 4 m/s (4.6), and almost half of the hours were less than 2 m/s.

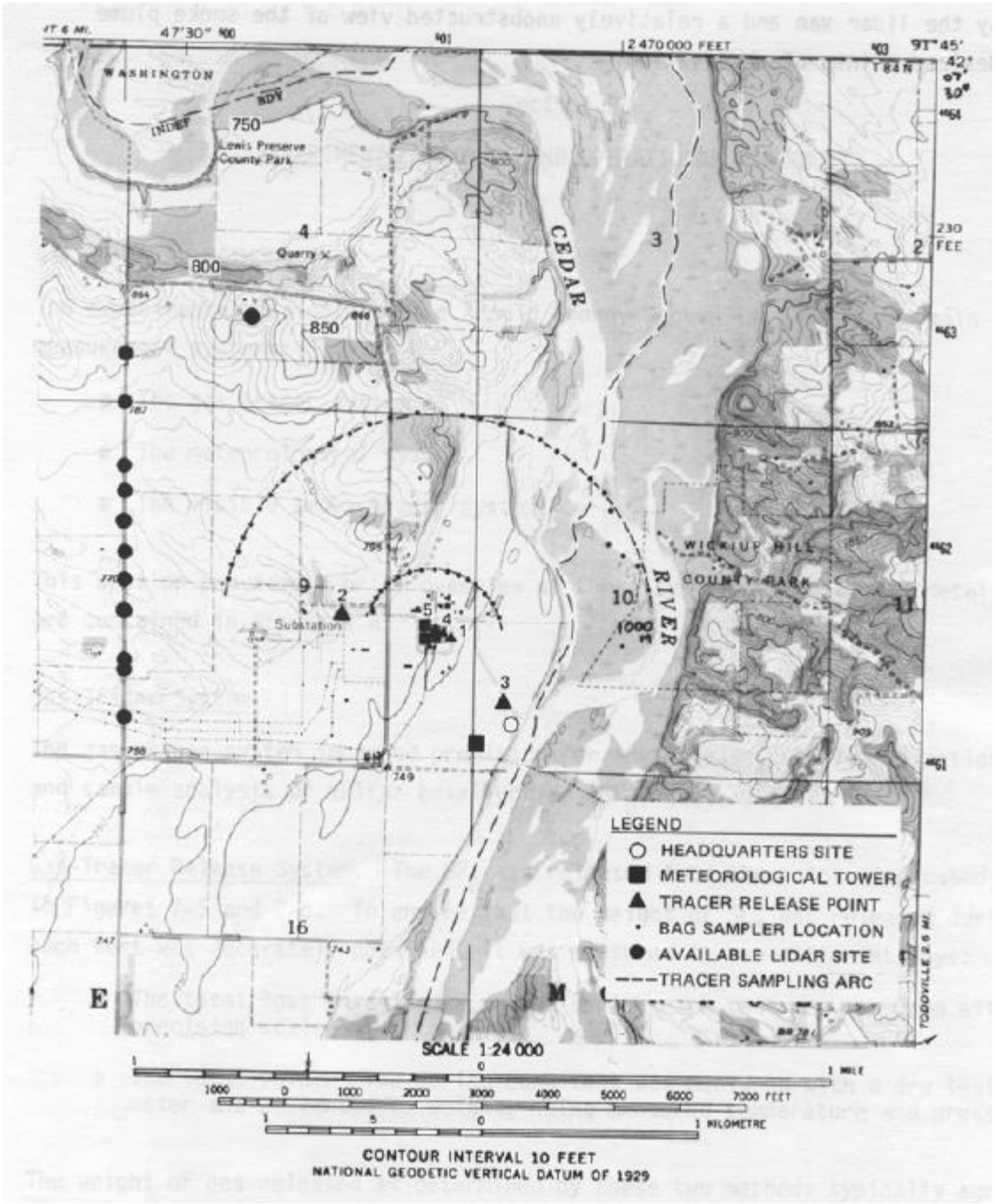


Figure 5. DAEC study area (SF6 releases).

2.6. Experimental Organically Cooled Reactor

The Experimental Organically Cooled Reactor (EOCR) study involved the simultaneous release of three tracer gases (SF_6, F_{12}, and Freon-$12B_2$) at three levels around the EOCR test reactor building at the Idaho National Engineering Laboratory in Southeast Idaho. The terrain was flat with low-lying shrubs. The main building was 25 m high with an effective width of 25 m. The tracer releases typically occurred simultaneously and were conducted during 22 separate time periods. Tracer sampler coverage was provided at eight concentric rings at distances of about 50, 100, 200, 400, 800, 1200, and 1600 m from the release points (see Figure 6). The stability classes ranged from stable to unstable. The 10 m wind speeds for the cases selected ranged from 3 to 8 m/s.

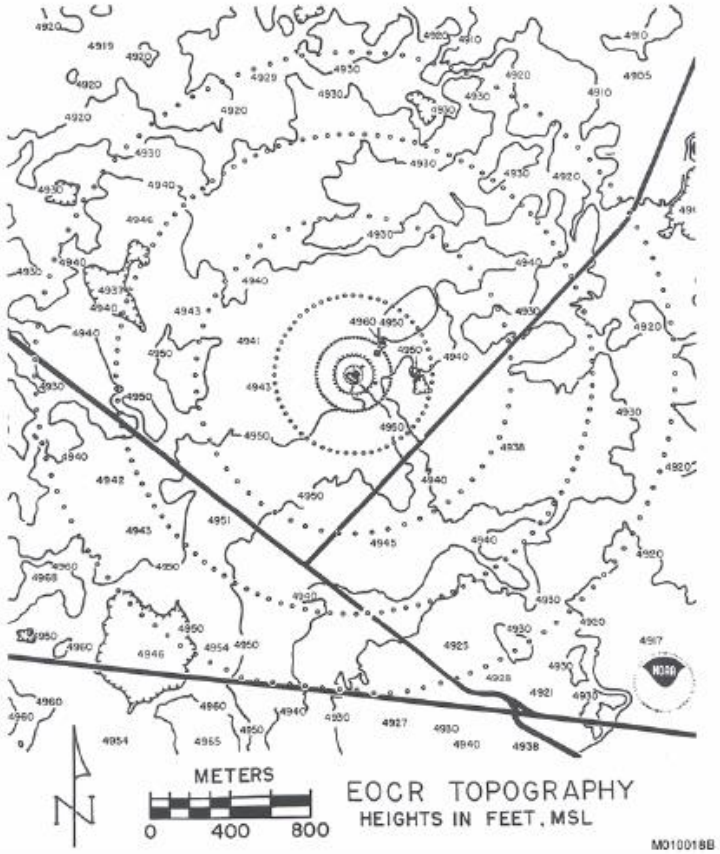


Figure 6. Terrain map featuring the entire EOCR grid with the source at the grid center (SF6 releases). Arcs are at distances of about 40, 80, 200, 400, 800, 1200, and 1600 m.

2.7. Alaska North Slope

The Alaska North Slope tracer study (see Figure 7) involved 44 hours of buoyant SF_6 releases from a 39 m high turbine stack. Tracer sampler coverage ranged over seven arcs from 50 to 3,000 m downwind. Meteorological data, including wind speed, wind direction, temperature, sigma-theta, and sigma-w, were available from an on-site tower at the 33 m level. Atmospheric stability and wind speed profiles were influenced by the smooth snow-covered tundra surface with negligible levels of solar radiation in the autumn months. All experiments (44 usable hours) were conducted during the abbreviated day light hours (0900 – 1600). Wind speeds taken at the 33-m level during the tests were less than 6 m/s during one and part of another test, between 6 and 15 m/s during four tests, and in excess of 15 m/s during three tests. Stability conditions were generally neutral or slightly stable.

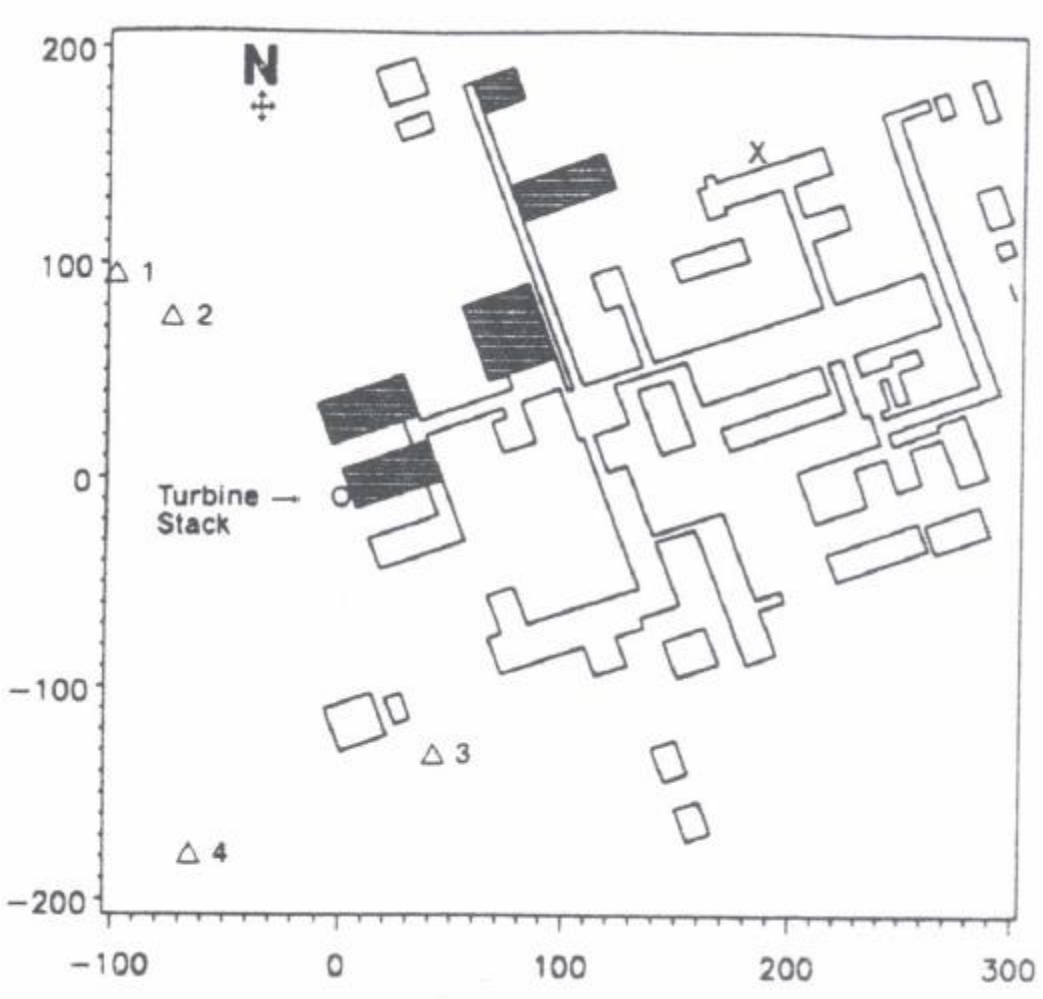


Figure 7. Depiction of Alaska North Slope Oil Gathering Center turbine stack, meteorological tower (X), and camera locations used to visualize plume rise.

2.8. Prairie Grass

The Prairie Grass study used a near-surface, non-buoyant tracer release in a flat rural area in Nebraska. This study involved a tracer of SO2 released at 0.46 m above the surface. Surface sampling arrays (arcs) were positioned from 50 m to 800 m downwind. Meteorological data included the 2 -m level wind direction and speed, the root-mean-square wind direction fluctuation, and the temperature difference (ΔT) between 2 m and 16 m. Other surface parameters, including friction velocity, Monin-Obukhov length, and lateral plume spread were estimated. Wind, turbulence, and temperature were obtained from a multi-leveled instrumented 16 m meteorological tower. A total of 44 ten-minute sampling periods were used, including both convective and stable conditions.

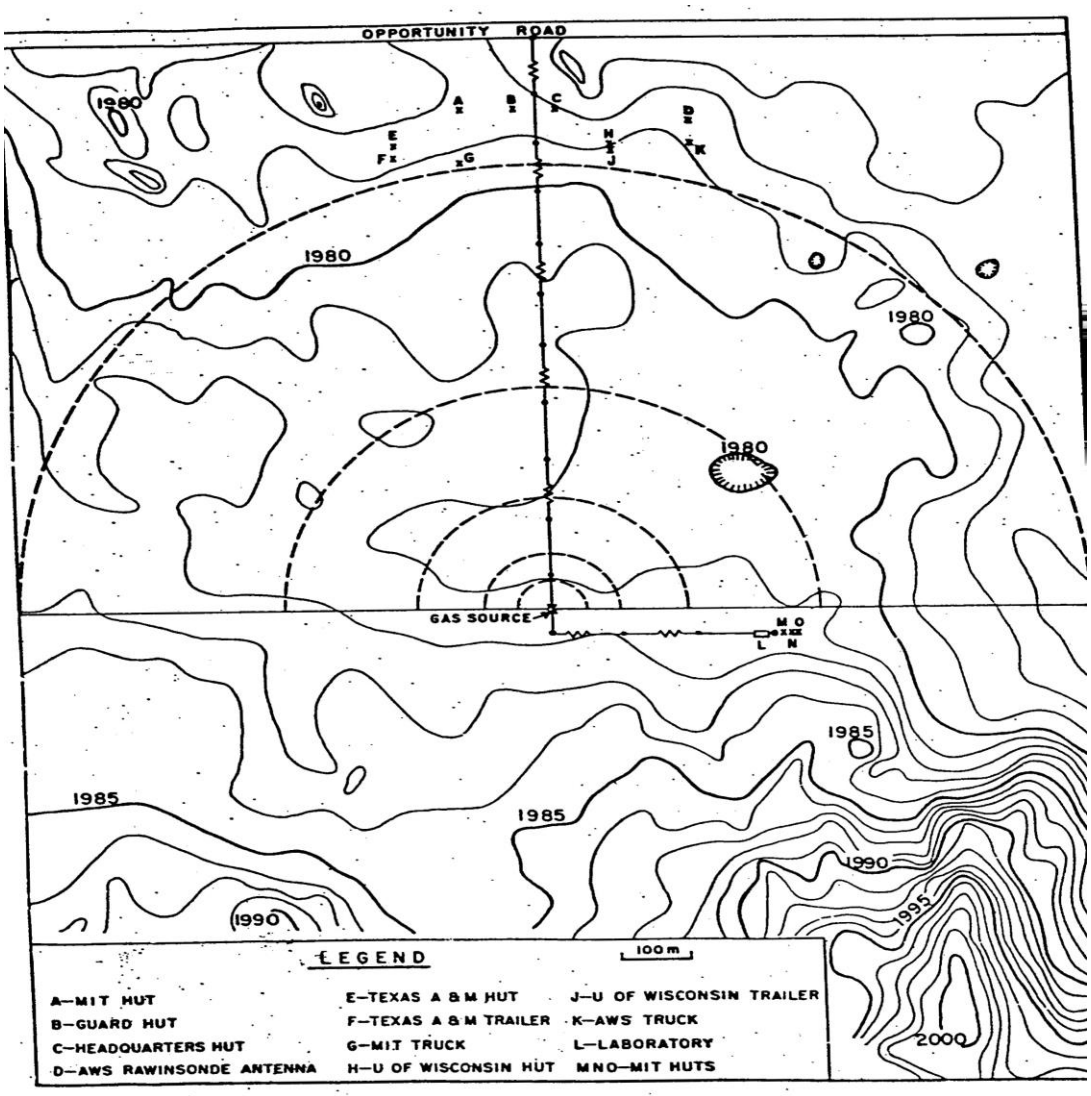


Figure 8. Prairie Grass study area

2.9. Indianapolis

The Indianapolis study consisted of an elevated, buoyant tracer (SF6) released in a flat-terrain urban to suburban area from a single 84-m stack (Figure 9). Data are available for approximately a four- to five-week period with 177 monitors providing 1-hour averaged samples along arcs from 250 m to 12 km downwind for a total of 1,297 arc-hours. Meteorological data included wind speed and direction, sigma-theta on a 94-meter tower; and wind speed, ΔT (2m - 10m) and other supporting surface data at three other 10-m towers (Figure 10). Observed plume rise and estimates of plume sigma-y are also available from the database.

Figure 9. Map showing the location of the Perry-K Station (A), the Hoosier Dome (B), and the central Indianapolis business district (C). The downtown surface meteorological site is located at (D) and the "bank tower" site was on top of the building at (E).

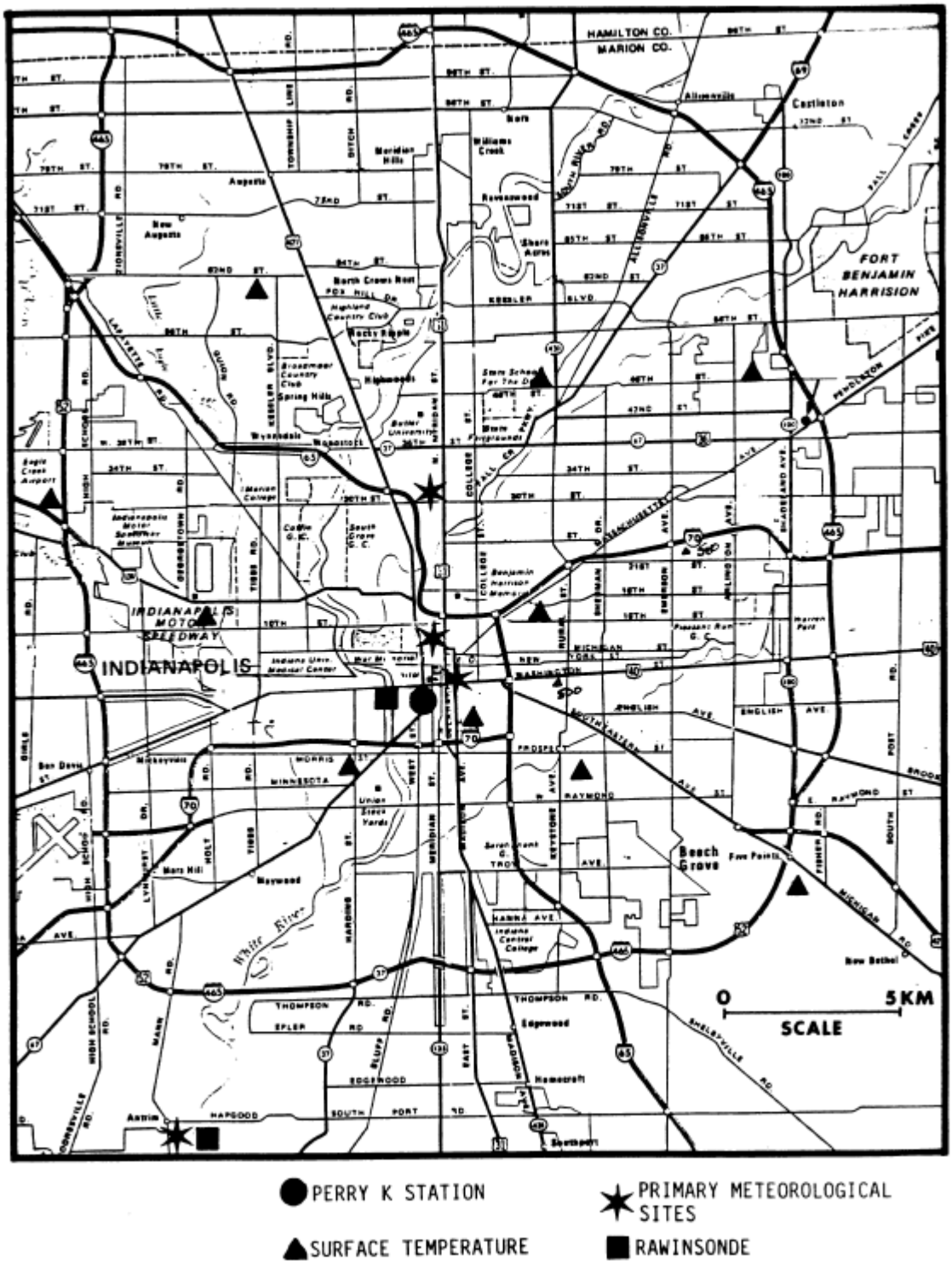


Figure 10. Indianapolis meteorological sites and emissions site (Perry K Station).

2.10. Kincaid

The Kincaid SO2 study was conducted in a flat rural area of Illinois (Figure 11). It involved a buoyant, continuous release of SO2 from a 187-m stack in rural flat terrain. The study included about six months of data between April 1980 and June 1981 (a total of 4,614 hours of samples). There were 30 SO2 monitoring stations providing 1-hour averaged samples from about 2 km to 20 km downwind of the stack. Meteorological data included wind speed, direction, and temperature from a tower instrumented at 2, 10, 50, and 100 m levels, and nearby National Weather Service (NWS) data.

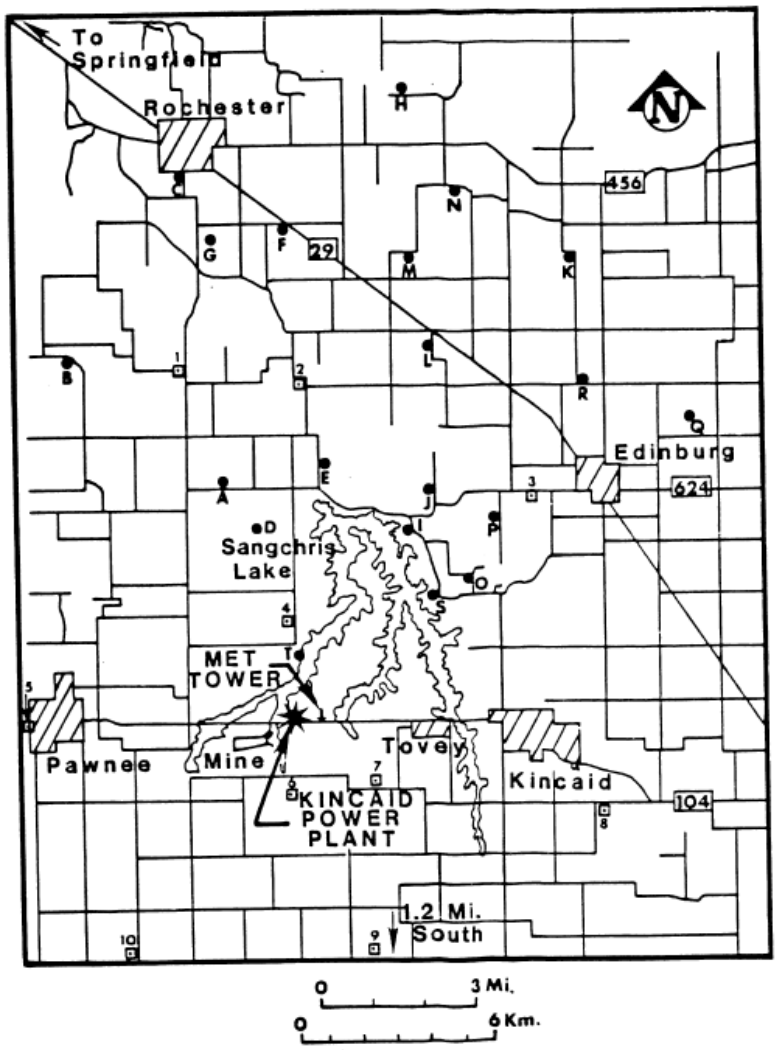


Figure 11. Kincaid study area.

2.11. AGA

The AGA experiments occurred during spring and summer 1980 at gas compressor stations in Texas and Kansas (Figure 12). At each test facility, one of the gas compressor stacks was retrofitted to accommodate SF_6 tracer gas emissions. In addition, stack height extensions were provided for some of the experiments (with the normal stack height close to 10 m). The stack height to building height ratios for the tests ranged from 0.95 to 2.52. There were a total of 63 tracer releases over the course of the tests, and the tracer samplers were located between 50 and 200 m away from the release point (see Figure 12). An instrumented 10-m tower was operated at both experimental sites. The tracer releases were generally restricted to daytime hours. Stability classes range from neutral to extremely unstable, except for three hours that were slightly stable. Wind speeds range from 2 to 11 m/s over the 63 hours.

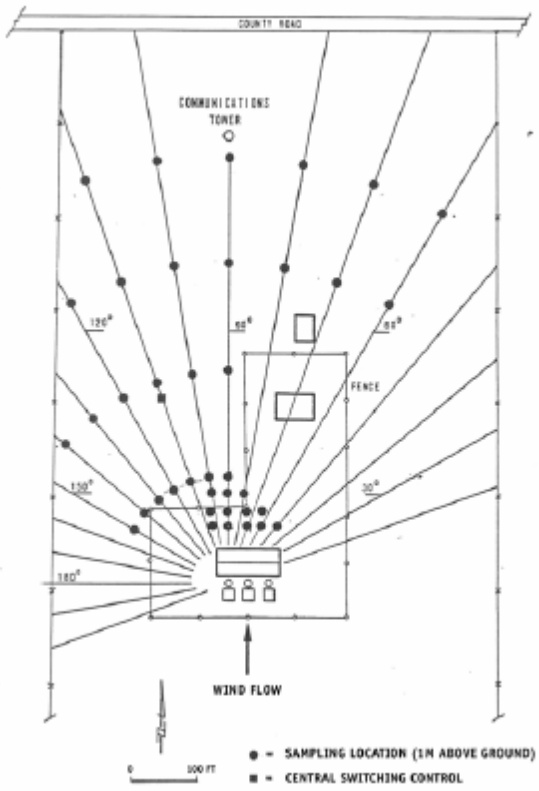


Figure 12. Plan view of the locations of tracer samplers at Site 1, AGA field study (SF6 releases).

2.12. Millstone Nuclear Power Plant

The Millstone nuclear power plant is located on the Connecticut coast, near Niantic. The model evaluation database features 36 hours of SF_6 emissions from a 48-m reactor stack and 26 hours of Freon emissions from a 29-m turbine stack. Exit temperatures were close to ambient (about 295K) with exit velocities of about 10 m/s for both the reactor stack (48.3 m) and the three turbine stacks (29.1 m). These stacks were associated with 45-m and 28-m building tiers, respectively. The monitoring data consisted of three arcs at 350, 800 and 1,500 m. Meteorological data were available from an on-site tower at the 10-m and 43-m levels. There was about an even split between stable and unstable hours, with mostly onshore winds and fairly high wind speeds. There were only 3 stable hours with wind speed less than 4 m/s, and the majority was above about 7 m/s and several above 10 m/s. Figure 13 shows the layout of the study area.

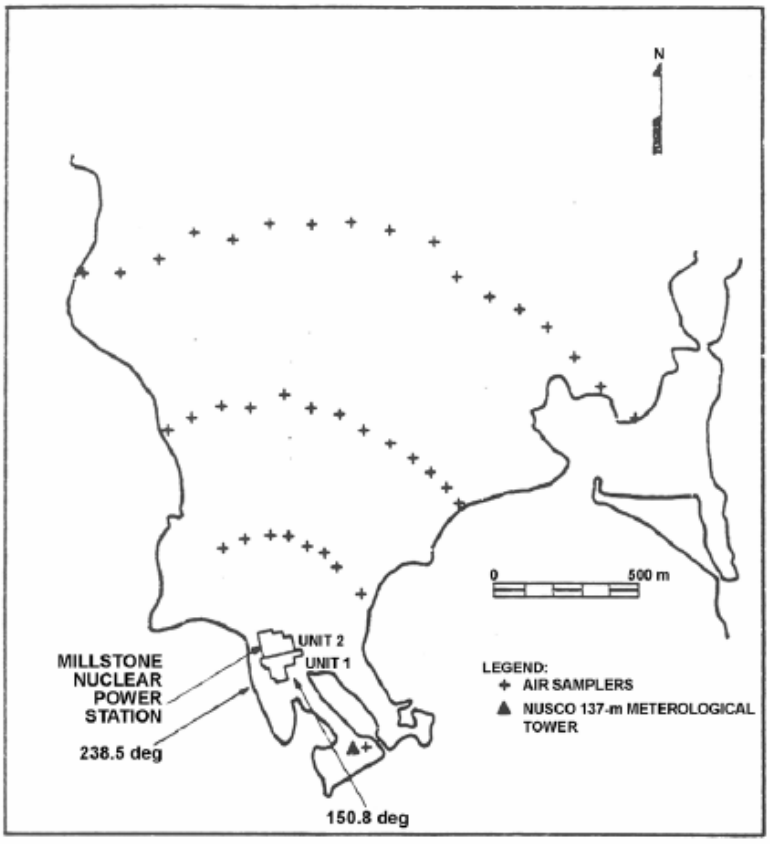


Figure 13. Millstone study area (SF6 and freon releases).

2.13. Bowline

The Bowline Point site[33], located in the Hudson River valley in New York State, is shown in Figure 14 (topographic map). The electric utility site included two 600-MW units, each with an 86.9-m stack and a dominant roof tier with a height of 65.2 m high in a rural area. There were four monitoring sites as shown in Figure 14 that ranged from about 250 to 850 m from the stacks. Hourly emissions data was determined from load data, coal analyses, and site-specific relationships between loads and fuel consumption. Meteorological data was obtained from a 100-m tower at the site. This site was also used as an independent evaluation database with the entire year included.

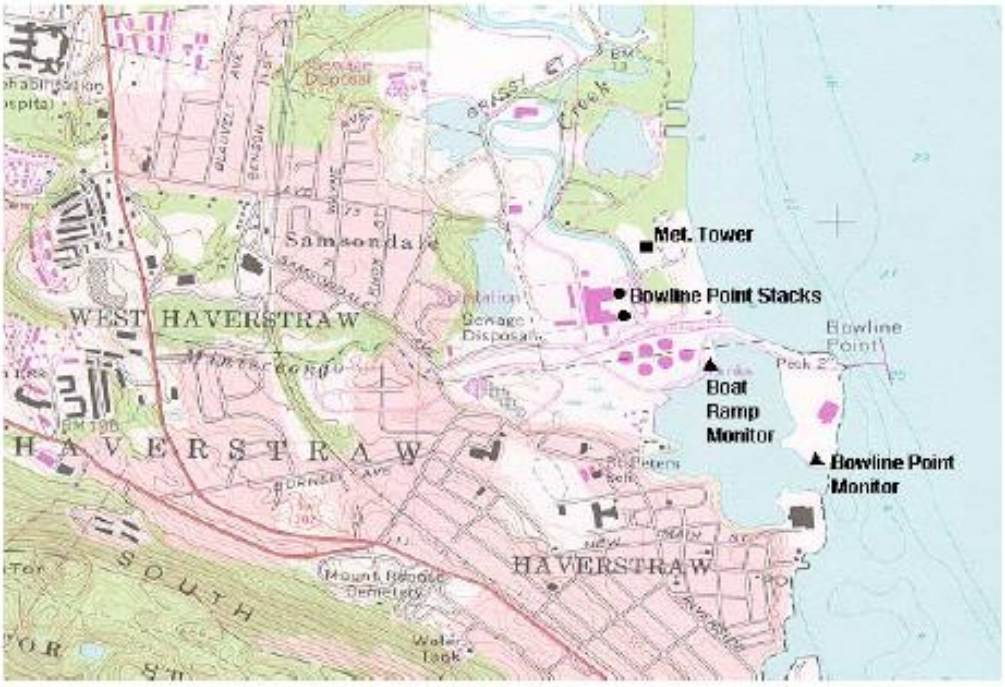


Figure 14. Bowline Point study area (SO2 releases).

2.14. Baldwin Power Plant

The Baldwin Power Plant is located in a rural, flat terrain setting of southwestern Illinois and has three identical 184-m stacks aligned approximately north-south with a horizontal spacing of about 100 m (Figure 15). There were 10 SO2 monitors that surrounded the facility, ranging in distance from two to ten km. On-site meteorological data was available during the study period of April 1, 1982 through March 31, 1983 and consisted of hourly averaged wind speed, wind direction, and temperature measurements taken at 10 m and wind speed and wind direction at 100 m.

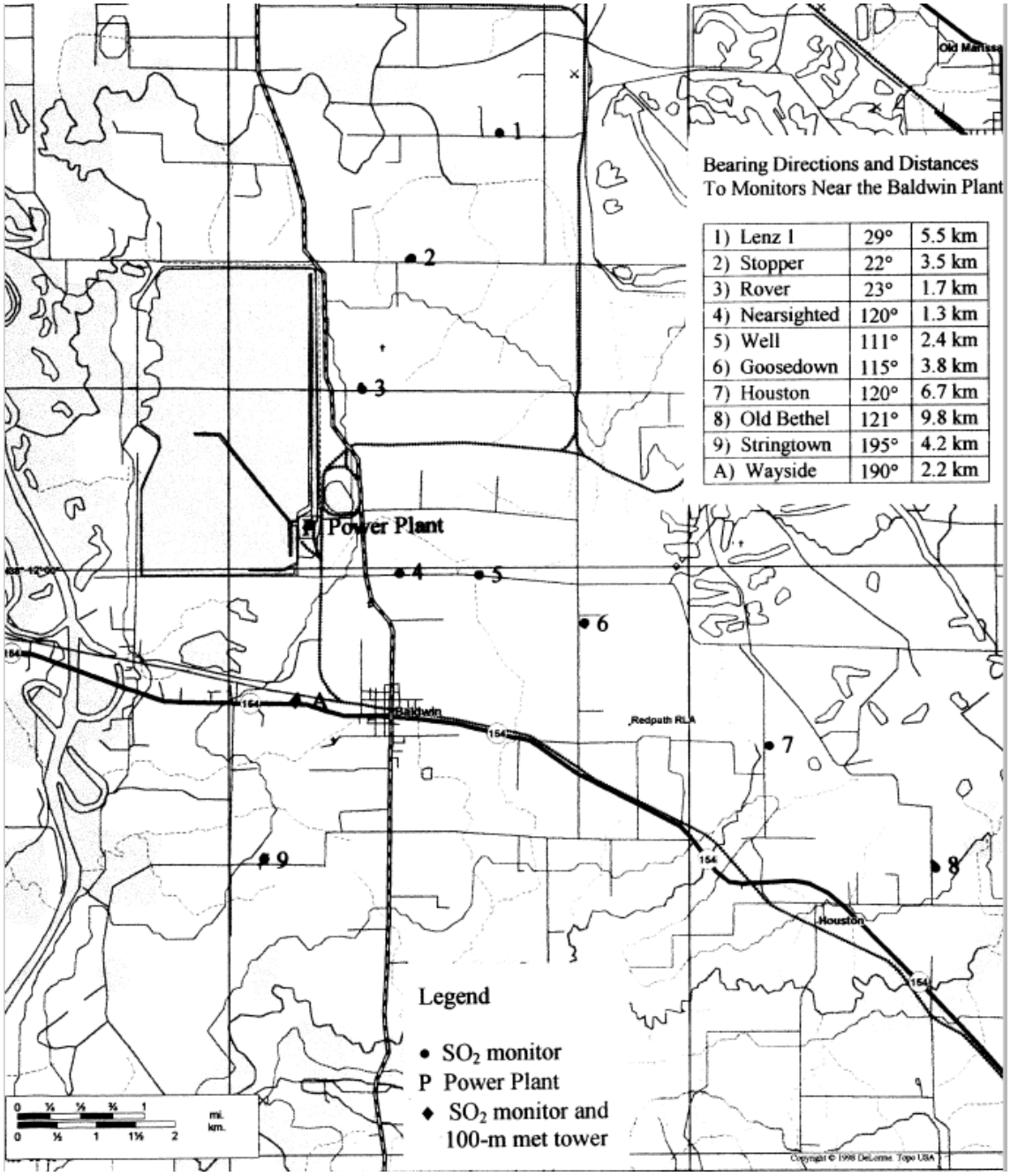
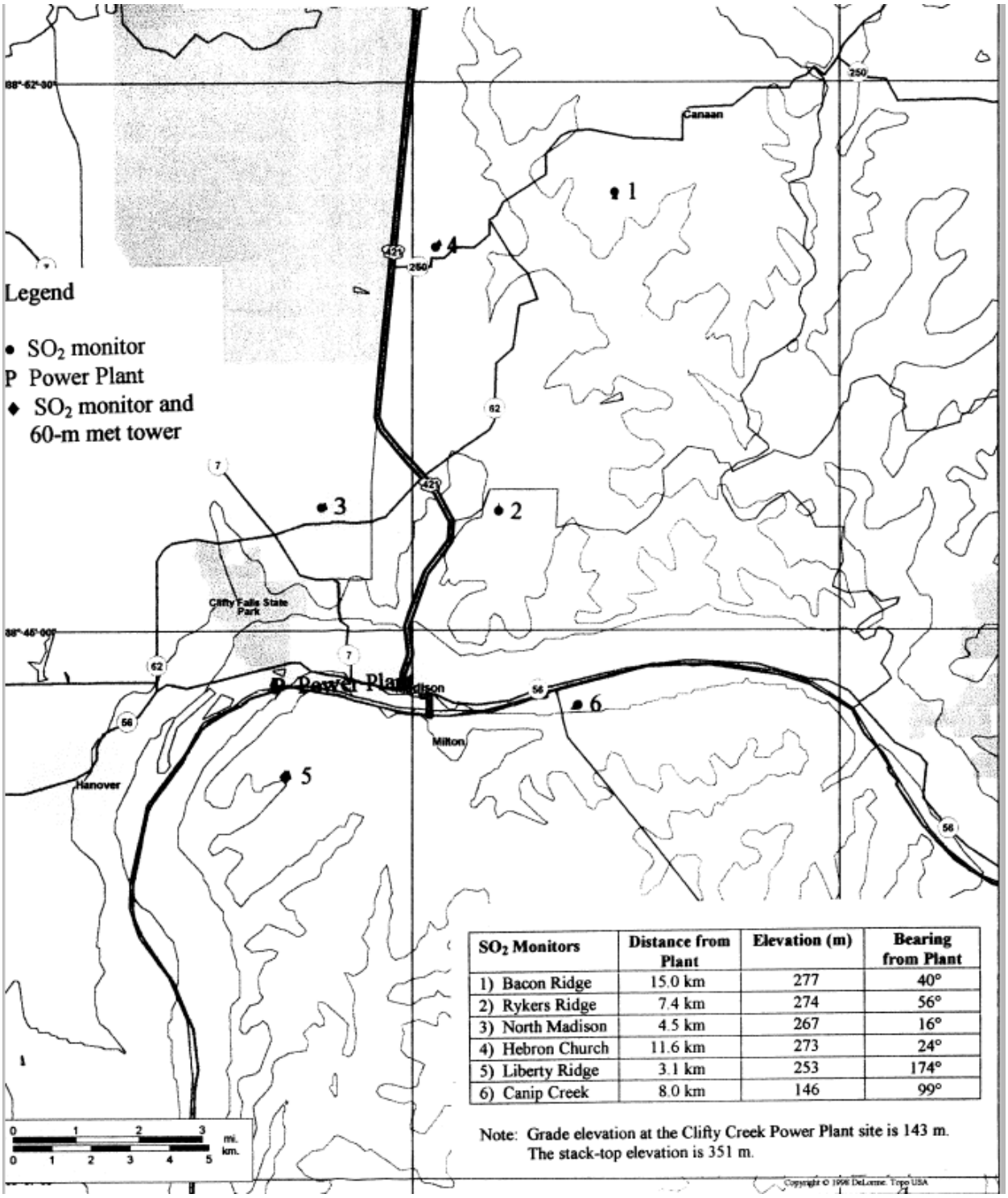


Figure 15. Baldwin study area.

2.15. Clifty Creek Power Plant

The Clifty Creek Power Plant is located in rural southern Indiana along the Ohio River with emissions from three 208-m stacks during this study (Figure 16). The area immediately north of the facility is characterized by cliffs rising about 115 m above the river and intersected by creek valleys. Six nearby SO2 monitors (out to 16 km from the stacks) provided hourly averaged concentration data. Meteorological data from a nearby 60-m tower covered the two-year period from January 1, 1975 through December 31, 1976, although only the data from 1975 were used in this evaluation. This database was also used in a major EPA-funded evaluation of rural air quality dispersion models in the early 1980s.



SO_2 Monitors	Distance from Plant	Elevation (m)	Bearing from Plant
1) Bacon Ridge	15.0 km	277	40°
2) Rykers Ridge	7.4 km	274	56°
3) North Madison	4.5 km	267	16°
4) Hebron Church	11.6 km	273	24°
5) Liberty Ridge	3.1 km	253	174°
6) Canip Creek	8.0 km	146	99°

Figure 16. Clifty Creek study area.

3. Evaluation methodology

3.1. AERMET/AERMOD comparisons

Two versions of AERMET/AERMOD will be compared using Robust highest concentrations and the EPA Protocol for determining best performing model. AERMET 22112/AERMOD 22112 will be compared against AERMET 23132/AERMOD 23132 with various combinations of adjusted or non-adjusted surface friction velocity (u*) and inclusion/exclusion of turbulence parameters (s_v and s_w). The modeled scenarios are:

- 22112_no_u*_with_turb: AERMET/AERMOD 22112 with no u* adjustment and turbulence included in the meteorological data
- 22112_with_u*_no_turb: AERMET/AERMOD 22112 with u* adjustment and no turbulence included in the meteorological data.
- 22112_no_u*_no_turb: AERMET/AERMOD 22112 with no u* adjustment and no turbuluence included in the meteorological data
- 23132_no_u*_with turb: AERMET/AERMOD 23132 with no u* adjustment and turbulence included in the meteorological data
- 23132_with_u*_no_turb: AERMET/AERMOD 23132 with u* adjustment and no turbulence included in the meteorological data.
- 23132_no_u*_no_turb:AERMET/AERMOD 23132 with no u* adjustment and no turbulence included in the meteorological data.

3.2. Evaluation procedures

3.2.1. Robust highest concentrations

Robust highest concentrations (RHC) were calculated for each averaging period of each database. The RHC statistic is calculated as:

$$RHC = \mathrm{X}(N) + [\bar{\mathrm{X}} - \mathrm{X}(N)] \times \ln\left[\frac{3N-1}{2}\right] \quad \textbf{(1)}$$

where X(N) is the Nth largest value, $\overline{\mathrm{X}}$ is the average of N-1 values, and N is the number of values exceeding the threshold value, usually 26.

For the 1-hour RHC, the RHC is calculated based on N=26 across all modeled and monitored values (i.e., not paired in time or space). For the 3-hour and 24-hour the RHC is calculated separately for each monitor within the network for observations and modeled values. The highest observed RHC is then compared to the highest modeled RHC.

3.2.2. EPA Protocol for determining best performing model

AERMOD output, among the different meteorological datasets, was evaluated using the EPA's Protocol for Determining the Best Performing Model, or Cox-Tikvart method (U.S. EPA, 1992; Cox and Tikvart, 1990). The protocol uses a two-step process for determining the better performing model when comparing models. The first step is a screening test that fails to perform at a minimal operational level. The second test applies to those models that pass the screening test that uses bootstrapping to generate a probability distribution of feasible outcomes (U.S. EPA, 1992). This section will discuss the methodology using the evaluation cases as examples.

The first step is to perform a screening test based on fractional bias:

$$FB = 2\left[\frac{OB - PR}{OB + PR}\right] \quad \textbf{(2)}$$

where FB is the fractional bias, OB is the average of the highest 25 observed concentrations and PR is the average of the highest 25 predicted averages. The fractional bias is also calculated for the standard deviation where OB and PR refer to the standard deviation of the highest 25 observed and predicted concentrations respectively. This is done across all monitors and modeled receptors, unpaired in time and space for the 3-hour and 24-hour averaging periods. The fractional bias of the means is plotted against the fractional bias of the standard deviation. Biases that exceed a factor-of-two under-prediction or over-prediction are considered grounds for excluding a model for further evaluation (U.S. EPA, 1992).

Models that pass the screening test are subjected to a more comprehensive statistical comparison that involves both an operational and scientific component using the RHC (Eq. 1). For the evaluations presented here, the screening step was skipped. The operational component is to measure the model's ability to estimate concentration statistics most directly used for regulatory purposes and the scientific component evaluates the model's ability to perform accurately throughout the range of meteorological conditions and the geographic area of concern (U.S. EPA, 1992).

The operational component of the evaluation compares performance in terms of the largest network-wide RHC test statistic. The RHC is calculated separately for each monitor within the network for observations and modeled values. The highest observed RHC is then compared to the highest modeled RHC using Equation 2, where RHC now replaces the means of the top 25 values of observed or modeled concentrations. Absolute fractional bias (the absolute value of fractional bias), AFB is calculated for 3 and 24-hour averages.

The scientific component of the evaluation is also based on absolute fractional bias, but the bias is calculated using the RHC for each meteorological condition and monitor. The meteorological conditions are a function of atmospheric stability and wind speed. For the purposes of these studies, six unique conditions were defined based on two wind speed categories (below and above 2.0 m/s) and three stability categories: unstable, neutral, and

stable.[3] In this evaluation, only 1-hour concentrations are used, and the AFB is based on RHC values paired in space and stability/wind speed combination.

A composite performance measure (CPM) is calculated from the 1-hour, 3-hour, and 24-hour AFB's:

$$CPM = \frac{1}{3} \times \left(\overline{AFB_{i,j}}\right) + \frac{2}{3} \times \left[\frac{AFB_3 - AFB_{24}}{2}\right] \quad \textbf{(3)}$$

where $AFB_{i,j}$ is the absolute fractional bias for monitor i and meteorological condition j, $\overline{AFB_{i,j}}$ is the average absolute fractional bias across all monitors and meteorological conditions, AFB_3 is the absolute fractional bias for the 3-hour average, and AFB_{24} is the absolute fractional bias for the 24-hour average. Once CPM values have been calculated for each model, a model comparison measure is calculated to compare the models:

$$MCM_{A,B} = CPM_A - CPM_B \quad \textbf{(4)}$$

where CPM_A is the CPM for model A and CPM_B is the CPM for model B. When more than two models are being compared simultaneously, the number of MCM values is equal to the total of the number of unique combinations of two models. For Martins Creek, Lovett, Westvaco, and Kincaid, there are four scenarios each, so there were six MCM comparisons for each location. For Bowline, Baldwin, and Clifty Creek, there are three scenarios each, resulting in three MCM comparisons for each location.

In order to determine if the difference between models was statistically significant, the standard error was calculated. A bootstrapping technique was used to create 1000 sample years based on methodology outlined in U.S. EPA (1992). The original data is divided into 3-day

[3] In U.S. EPA (1992), the three stability categories are related to the Pasquill-Gifford categories, unstable being A, B, and C, neutral being D, and stable being E and F. Since AERMOD does not use the stability categories, the stability class was determined using Monin-Obukhov length and surface roughness using methodology from AERMOD subroutine LTOPG.

blocks. Within each season, the 3-day blocks are sampled with replacement until a total season is created. The process is repeated until 1000 boot-strap years are created[4]. The standard error is calculated as the standard deviation of the bootstrap generated outcomes for the MCM.

The magnitude and sign of the MCM are indicative of relative performance of each pair of models. The smaller the CPM the better the overall performance of the model. This means that for two models, A and B, a negative difference between the CPM for A and CPM for B implies that model A is performing better (Model A has a smaller CPM) while a positive difference indicates that Model B is performing better.

Since more than two scenarios are being evaluated in these studies, simultaneous confidence intervals of 90 and 95 percent were calculated. These were calculated by finding the 90^{th} and 95^{th} percentiles of the distribution across all MCM values from the bootstrapping procedure for all model comparisons. The confidence intervals were then found by:

$$CI_{X,A,B} = MCM_{A,B} \pm c_X s_{A,B} \quad \textbf{(5)}$$

where $CI_{X,A,B}$ is the confidence interval for X percent (90 or 95^{th}) for models A and B, $MCM_{A,B}$ is as defined in Equation 4, c_X is the X percentile of the MCM values from the bootstrap results and $s_{A,B}$ is the standard deviation of the bootstrap MCM results for models A and B. Note that in Equation 5, $MCM_{A,B}$ is the MCM value from the original data, not the bootstrap results.

For each pair of model comparisons, the significance of the model comparison measure depended on whether the confidence interval overlapped zero. If the confidence interval overlapped zero, then the two models were not performing at a level which was considered

[4] The bootstrapping was completed using the SAS© SURVEYSELECT procedure with resampling for 1000 replicates.

statistically different. Otherwise, if they did not overlap zero, then there was a statistically significant difference between the two models.

3.3. Results

3.3.1. Turbulence cases

Table 2 lists the hourly observed and modeled RHC, as well as 3-hour and 24-hour RHC for applicable databases, for the databases that initially included turbulence. Table 3 lists the RHC values for those databases initially without turbulence. The modeled scenario(s) closest to the observed RHC are highlighted in gray for each database.

Results in Table 2 indicate that the 22112 and 23132 modeled RHC's are generally identical with the exception of Martins Creek 24-hour values, Westvaco 24-hour values (no u* with turbulence and no u* and no turbulence), EOCR, and Kincaid 24-hour values (no u* with turbulence). Results in Table 2 also indicate that for the most part for the databases with turbulence data, the 22112 or 23132 cases without the u* adjustment and with turbulence data were the better performers against observations. For a few instances, depending on the averaging period, the cases with the u* adjustment and no turbulence, or the cases with no u* adjustment and no turbulence were the better performers.

Table 3 indicates that for the non-turbulence databases, the use of adjusted u* increased modeled performance in some cases depending on the averaging period or stack height. While decreasing or not changing model performance in other cases, depending on averaging period or stack height. For the databases that had multiple averaging periods (Martins Creek, Lovett, Westvaco, and Kincaid), there was not a consistent better performing model across the averaging periods. For example, for Martins Creek, 22112_with_u*_no_turb and 23132_with_u*_no_turb performed better for the 24-hour averaging period, while 22112_no_u*_with_turb and 23132_no_u*_with turb performed better for the 1 and 3-hour period. For DAEC, which had observed concentrations for emissions from different stack heights, the better performing modeling appeared to be dependent on stack height. Overall, it appears that the use of adjusted

u* did not increase model performance for most of the cases and that the inclusion of turbulence is more important to model performance than the u* adjustment.

Table 2. Hourly, 3-hour, and 24-hour RHC for turbulence cases.
Best performing model compared to observed RHC are highlighted in gray.

Database	Avg. period (hr)	RHC						
		Observed	AERMOD version					
			22112			23132		
			No_u*_with turb	With_u*_no_ turb	No_u*_ no_turb	No_u*_ with turb	With_u* _no_tur b	No_u*_ no_turb
Martins Creek	1	1216	1133	1034	1427	1133	1034	1427
	3	461	497	505	655	497	505	655
	24	79	143	132	158	143	132	158
Tracy	1	15	13	18	25	13	18	25
Lovett	1	426	374	538	622	374	538	622
	3	187	169	239	254	169	239	254
	24	52	48	63	68	48	63	68
Westvaco	1	2757	2460	1252	2091	2460	1252	2091
	3	1575	1731	783	1654	1731	783	1654
	24	480	522	457	613	522	457	613
DAEC (h=1m)	1	346	240	188	222	240	188	222
DAEC (h=24m)	1	253	84	71	75	84	71	75
DAEC (h=46m)	1	140	91	59	99	91	59	99
EOCR	1	3763	5822	5731	8250	5822	5731	8250
Alaska	1	6	5	8	8	5	8	8
Prairie Grass	1	925087	989003	873817	989003	987307	867946	883444
Indianapolis	1	6	4	4	5	4	4	5
Kincaid	1	1611	1312	717	717	1312	717	717
	3	618	615	470	470	615	470	470
	24	113	101	167	167	101	167	167

3.3.2. Non-turbulence cases

Table 3 lists the RHC values for the non-turbulence databases for 22112 and 23132. In these databases, because of the lack of turbulence in the meteorological data, the effect of the u* adjustment has more impact in improving model performance. Also, the results indicate the changes made to AERMOD between 22112 and 23132 did not impact these findings.

Table 3. Hourly, 3-hour, and 24-hour RHC for non-turbulence cases.
Best performing model compared to observed RHC are highlighted in gray.

Database	Avg. period (hr)	RHC				
		Observed	AERMOD version			
			22112		23132	
			With_u*_ no_turb	No_u*_ no_turb	With_u* _no_turb	No_u*_ no_turb
AGA	1	296	262	281	262	281
Millstone (Freon)	1	76	96	101	96	101
Millstone (SF6)	1	79	33	35	33	35
Bowline	1	763	552	547	552	547
	3	469	514	523	514	523
	24	204	307	290	307	290
Baldwin	1	2348	3531	3531	3531	3531
	3	920	1183	1184	1183	1184
	24	209	230	230	230	230
Clifty Creek	1	1451	1360	1360	1360	1360
	3	796	871	870	871	870
	24	243	170	165	170	165

3.3.3. Statistical evaluations

While the review of RHC can indicate general model performance, the use of the EPA Protocol for Determining Best Performing Model (U.S. EPA, 1992) provides a statistical basis of determining the best performing model. Tables 4 and 5 show the composite performance measure (CPM) for the turbulence databases and non-turbulence databases respectively. For the databases with turbulence (Table 4), the best performing models for Martins Creek were the cases with adjusted u* and no turbulence but for the remaining areas, the better performing models were the adjusted u* and no turbulence scenarios. This means the use of adjusted u* did not increase model performance and the use of turbulence was important to model performance. For the non-turbulence databases (Table 5), the use of adjusted u* increased model performance for Baldwin and Clifty Creek, while for Bowline, the use of adjusted u* slightly decreased model performance. For all cases, the CPM values were identical for the 22112 and 23132 model versions, suggesting the changes between 22112 and 23132 had minimal to no impact on model performance, which was expected based on the changes made to AERMET and AERMOD and no changes to the adjusted u* equations.

Table 4. Composite Performance Measure (CPM) for turbulence cases. Scenarios with lowest CPM's for each study location are highlighted in gray.

	Database			
Scenario	Martins Creek	Lovett	Westvaco	Kincaid
22112_no_u*_with_turb	0.35	0.40	0.41	0.37
22112_with_u*_no_turb	0.31	0.52	0.60	0.56
22112_no_u*_no_turb	0.49	0.58	0.44	0.56
23132_no_u*_with_turb	0.35	0.40	0.41	0.37
23132_with_u*_no_turb	0.31	0.52	0.60	0.56
23132_no_u*_no_turb	0.49	0.58	0.44	0.56

Table 5. Composite Performance Measure (CPM) for non-turbulence databases. Scenarios with lowest CPM's for each study location are highlighted in gray.

Scenario	Database		
	Bowline	Baldwin	Clifty Creek
22112_no_u*_no_turb	0.47	0.46	0.51
22112_with_u*_no_turb	0.50	0.45	0.49
23132_no_u*_no_turb	0.47	0.46	0.51
23132_with_u*_no_turb	0.50	0.45	0.49

Tables 6 through 9 show the model comparison measure (MCM) for the turbulence databases while Tables 10 through 12 show the MCM for the non-turbulence databases. Also shown are the 90 and 95% confidence intervals of the MCM based on the bootstrapping results. Confidence intervals highlighted in gray indicated statistical significance in the specific MCM cases. The original pairings of 22112 scenarios to other 22112 scenarios are shown for comparison to the analogous 23132 pairings. MCM pairings for the same u*/turbulence pairings between 23132 and 22112 are also shown to show if model changes made differences to results. For all such cases, such comparisons are zero.

Martins Creek (Table 6): The better performing models were 22112 and 23132 with u* and no turbulence. Also, the MCM results indicate that the use of adjusted u* with no turbulence is not statistically significant when compared to no adjusted u* with turbulence for both 22112 and 23132. There were three statistically significant MCM pairings that were statistically significant at the 90% confidence interval, and these were the difference between no u* adjustment and no turbulence and the cases (no adjusted u* with turbulence or adjusted u* with no turbulence) for both 22112 and 23132, indicating that not using adjusted u* and not using turbulence noticeably decreases model performance. At the 95% confidence interval, the two statistically significant differences were between 23132 no adjusted u*/ no turbulence and adjusted u*/ with turbulence for 23132 and for 23132 no adjusted u*/ no turbulence and adjusted u*/ no turbulence for 23132.

Lovett (Table 7): All cases of AERMET/AERMOD 22112 are statistically insignificant when compared AERMET/AERMOD 22112 at both the 90% and 95% CI with the exception of the no u* and no turbulence case compared to the no u* with turbulence case. For 23132 all

cases are statistically insignificant compared to each other at the 90% CI, with the exception of the 23132 no u* and no turbulence case compared to the 23132 no u* with turbulence case. However, the lower bound of the 90% CI is close to zero.

Westvaco (Table 8): The use of adjusted u* decreases model performance significantly at both the 90% and 95% CI for both 22112 and 23132. The use of no adjusted u* and no turbulence also decreases model performance at a statistically significant level for both 22112 and 23132.

Kincaid (Table 9): None of the MCM differences were statistically significant at 90% or 95% CI. The better performers were 22112 or 23132 with no u* adjustment and inclusion of turbulence, but as previously stated, were not statistically different from the adjusted u* case or the case with no adjusted u* and no turbulence.

For the non-turbulence databases (Tables 10-12), the use of adjusted u* was statistically insignificant compared to not using adjusted u* and as expected, the MCM values indicated no difference between 22112 and 23132.

Table 6. Martins Creek Model Comparison Measure (MCM) results.
Confidence intervals highlighted in gray are significant at that percent.

MCM Comparison	MCM	Confidence Intervals			
		90%		95%	
		Lower bound	Upper bound	Lower bound	Upper bound
22112_with_u*_no_turb - 22112_no_u*_with_turb	-0.04	-0.14	0.07	-0.16	0.09
22112_no_u*_no_turb - 22112_no_u*_with_turb	0.14	0.03	0.26	-0.003	0.29
22112_no_u*_no_turb - 22112_with_u*_no_turb	0.18	0.07	0.29	0.04	0.31
23132_no_u*_no_turb - 22112_no_u*_no_turb	0	-0.13	0.13	-0.16	0.16
23132_no_u*_with_turb - 22112_no_u*_with_turb	0	-0.10	0.10	-0.12	0.12
23132_with_u*_no_turb – 22112_with_u*_no_turb	0	-0.12	0.12	-0.14	0.14
23132_with_u*_no_turb - 23112_no_u*_with_turb	-0.03	-0.14	0.06	-0.15	0.09
23132_no_u*_no_turb - 23132_no_u*_with_turb	0.14	0.03	0.26	0.007	0.28
23132_no_u*_no_turb - 23132_with_u*_no_turb	0.18	0.07	0.29	0.05	0.31

Table 7. Lovett Model Comparison Measure (MCM) results.
Confidence intervals highlighted in gray are significant at that percent.

MCM Comparison	MCM	Confidence Intervals			
		90%		95%	
		Lower bound	Upper bound	Upper bound	Lower bound
22112_with_u*_no_turb - 22112_no_u*_with_turb	0.13	-0.05	0.30	-0.08	0.34
22112_no_u*_no_turb - 22112_no_u*_with_turb	0.18	0.01	0.35	-0.0	0.39
22112_no_u*_no_turb - 22112_with_u*_no_turb	0.05	-0.05	0.14	-0.06	0.17
23132_no_u*_no_turb - 22112_no_u*_no_turb	0	-0.12	0.12	-0.14	0.14
23132_no_u*_with_turb - 22112_no_u*_with_turb	0	-0.13	0.12	-0.15	0.15
23132_with_u*_no_turb – 22112_with_u*_no_turb	0	-0.11	0.11	-0.13	0.13
23132_with_u*_no_turb - 23132_no_u*_with_turb	0.12	-0.04	0.30	-0.08	0.33
23132_no_u*_no_turb - 23132_no_u*_with_turb	0.18	0.001	0.36	-0.03	0.39
23132_no_u*_no_turb - 23132_with_u*_no_turb	0.05	-0.04	0.15	-0.06	0.16

Table 8. Westvaco Model Comparison Measure (MCM) results.
Confidence intervals highlighted in gray are significant at that percent.

MCM Comparison	MCM	Confidence Intervals			
		90%		95%	
		Lower bound	Upper bound	Lower bound	Upper bound
22112_with_u*_no_turb - 22112_no_u*_with_turb	0.19	0.05	0.33	0.02	0.36
22112_no_u*_no_turb - 22112_no_u*_with_turb	0.03	-0.05	0.12	-0.07	0.13
22112_no_u*_no_turb - 22112_with_u*_no_turb	-0.16	-0.31	-0.01	-0.34	0.02
23132_no_u*_no_turb - 22112_no_u*_no_turb	0	-0.09	0.09	-0.11	0.11
23132_no_u*_with_turb - 22112_no_u*_with_turb	0	-0.08	0.08	-0.09	0.09
23132_with_u*_no_turb – 22112_with_u*_no_turb	0	-0.07	0.07	-0.09	0.09
23132_with_u*_no_turb - 23132_no_u*_with_turb	0.19	0.04	0.34	0.01	0.37
23132_no_u*_no_turb - 23132_no_u*_with_turb	0.03	-0.05	0.11	-0.07	0.13
23132_no_u*_no_turb - 23132_with_u*_no_turb	-0.16	-0.31	-0.01	-0.34	0.02

Table 9. Kincaid Model Comparison Measure (MCM) results.
Confidence intervals highlighted in gray are significant at that percent.

MCM Comparison	MCM	Confidence Intervals			
		90%		95%	
		Lower bound	Upper bound	Lower bound	Upper bound
22112_with_u*_no_turb - 22112_no_u*_with_turb	0.19	-0.27	0.66	-0.32	0.70
22112_no_u*_no_turb - 22112_no_u*_with_turb	0.19	-0.29	0.67	-0.34	0.72
22112_no_u*_no_turb - 22112_with_u*_no_turb	-5.1×10^{-4}	-0.13	0.13	-0.15	0.15
23132_no_u*_no_turb - 22112_no_u*_no_turb	2.0×10^{-5}	-0.14	0.14	-0.16	0.16
23132_no_u*_with_turb - 22112_no_u*_with_turb	6.0×10^{-5}	-0.56	0.51	-0.61	0.61
23132_with_u*_no_turb – 22112_with_u*_no_turb	2.0×10^{-5}	-0.14	0.14	-0.15	0.15
23132_with_u*_no_turb - 23132_no_u*_with_turb	0.19	-0.27	0.65	-0.32	0.70
23132_no_u*_no_turb - 23132_no_u*_with_turb	0.19	-0.28	0.66	-0.33	0.71
23132_no_u*_no_turb - 23132_with_u*_no_turb	-5.1×10^{-4}	-0.13	0.13	-0.14	0.14

Table 10. Bowline Model Comparison Measure (MCM) results.
Confidence intervals highlighted in gray are significant at that percent.

MCM Comparison	MCM	Confidence Intervals			
		90%		95%	
		Lower bound	Upper bound	Lower bound	Upper bound
22112_no_u*_no_turb - 22112_with_u*_no_turb	-0.03	-0.11	0.05	-0.12	0.06
23132_no_u*_no_turb - 22112_no_u*_no_turb	0.0	-0.10	0.10	-0.12	0.12
23132_with_u*_no_turb - 22112_with_u*_no_turb	0.0	-0.09	0.09	-0.12	0.12
23132_no_u*_no_turb - 23132_with_u*_no_turb	-0.03	-0.10	0.04	-0.12	0.06

Table 11. Baldwin Model Comparison Measure (MCM) results.
Confidence intervals highlighted in gray are significant at that percent.

MCM Comparison	MCM	Confidence Intervals			
		90%		95%	
		Lower bound	Upper bound	Lower bound	Upper bound
22112_no_u*_no_turb - 22112_with_u*_no_turb	0.002	-0.07	0.08	-0.09	0.09
23132_no_u*_no_turb - 22112_no_u*_no_turb	2.0×10^{-5}	-0.10	0.10	-0. 12	0.12
23132_with_u*_no_turb - 22112_with_u*_no_turb	2.0×10^{-5}	-0.10	0.10	-0. 12	0.12
23132_no_u*_no_turb - 23132_with_u*_no_turb	0.002	-0.07	0.08	-0.09	0.09

Table 12. Clifty Creek Model Comparison Measure (MCM) results.
Confidence intervals highlighted in gray are significant at that percent.

MCM Comparison	MCM	Confidence Intervals			
		90%		95%	
		Lower bound	Upper bound	Lower bound	Upper bound
22112_no_u*_no_turb - 22112_with_u*_no_turb	0.02	-0.04	0.07	-0.05	0.08
23132_no_u*_no_turb - 22112_no_u*_no_turb	$3x10^{-5}$	-0.07	0.07	-0.08	0.08
23132_with_u*_no_turb - 22112_with_u*_no_turb	$3x10^{-5}$	-0.06	0.06	-0.08	0.08
23132_no_u*_no_turb - 23132_with_u*_no_turb	0.02	-0.04	0.07	-0.05	0.08

4. Summary/Conclusions

Based on the results the RHC comparisons and the EPA protocol for determining best performing model, in situations involving turbulence, the use of turbulence without adjusting u* usually led to better performance than using adjusted u* without turbulence, especially in areas of complex terrain. In some instances, the differences between the adjusted u* cases were statistically worse than non-adjusted u* cases. For situations where turbulence is not in the meteorological data, the use of adjusted u* often resulted in little change or some increase in model performance. However, the databases without turbulence were in flat terrain and had talk stacks, so model performance for non-turbulence cases with complex terrain cannot be determined from these results. The results of the RHC and EPA protocol also indicate that changes made to AERMOD 23132 had no unexpected changes from AERMOD 22112.

5. References

Cimorelli, A. J., S. G. Perry, A. Venkatram, J. C. Weil, R. J. Paine, R. B. Wilson, R. F. Lee, W. D. Peters, and R. W. Brode, 2005: AERMOD: A dispersion model for industrial source applications Part I: General model formulation and boundary layer characterization. *J.Appl.Meteor.* **44**, 682-693

Cox, W. M. and J. A. Tikvart, 1990. A statistical procedure for determining the best performing air quality simulation model. Atmos. Environ., 24A(9): 2387-2395.

Perry, S. G., A. J. Cimorelli, R. J. Paine, R. W. Brode, J. C. Weil, A. Venkatram, R. B. Wilson, R. F. Lee, and W. D. Peters, 2005: AERMOD: A dispersion model for industrial source applications Part II: Model performance against seventeen field-study databases. *J.Appl.Meteor*. **44**, 694-708. U.S. Environmental Protection Agency, 1992: Protocol for Determining Best Performing Model. EPA-454/R-92-025, U.S. Environmental Protection Agency, RTP, NC.

U.S. Environmental Protection Agency, 2003: AERMOD: Latest Features and Evaluation Results. EPA-454/R-03-003, U.S. Environmental Protection Agency, RTP, NC.

United States Environmental Protection Agency	Office of Air Quality Planning and Standards Air Quality Assessment Division Research Triangle Park, NC	Publication No. EPA-454/B-23-011 October 2023