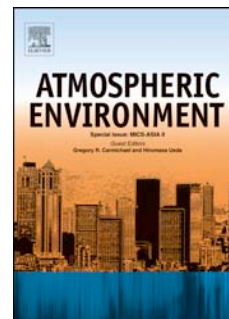# Accepted Manuscript

Performance evaluation of AERMOD, CALPUFF, and legacy air dispersion models using the winter validation tracer study dataset

Arthur S. Rood

- Steady-state and Lagrangian puff models compared.
- SF_6 tracer data at 140 samplers in concentric rings 8- and 16-km from release point.
- Paired and unparied 1- and 9-hour average concentrations performance objectives.
- Puff models showed the lower bias and variance and higher correlation
- Steady-state models were less likely to underpredict concentrations

1

PERFORMANCE EVALUATION OF AERMOD, CALPUFF, AND LEGACY AIR DISPERSION MODELS USING THE WINTER VALIDATION TRACER STUDY DATASET

Arthur S. Rood[*]

[*] Corresponding author, K-Spar Inc.,4835 W. Foxtrail Lane, Idaho Falls, Idaho 83402, Phone: 208 528-0670, asr@kspar.com

ABSTRACT

11

12   The performance of the steady-state air dispersion models AERMOD and Industrial Source Complex 2

13   (ISC2), and Lagrangian puff models CALPUFF and RATCHET were evaluated using the Winter

14   Validation Tracer Study dataset. The Winter Validation Tracer Study was performed in February 1991 at

15   the former Rocky Flats Environmental Technology Site near Denver, Colorado. Twelve, 11-hour tests

16   were conducted where a conservative tracer was released and measured hourly at 140 samplers in

17   concentric rings 8 km and 16 km from the release point. Performance objectives were unpaired maximum

18   one- and nine-hour average concentration, location of plume maximum, plume impact area, arc-integrated

19   concentration, unpaired nine-hour average concentration, and paired ensemble means. Performance

20   objectives were aimed at addressing regulatory compliance, and dose reconstruction assessment

21   questions. The objective of regulatory compliance is not to underestimate maximum concentrations

22   whereas for dose reconstruction, the objective is an unbiased estimate of concentration in space and time.

23   Performance measures included the fractional bias, normalized mean square error, geometric mean,

24   geometric mean variance, correlation coefficient, and fraction of observations within a factor of two. The

25   Lagrangian puff models tended to exhibit the smallest variance, highest correlation, and highest number

26   of predictions within a factor of two compared to the steady-state models at both the 8-km and 16-km

27   distance. Maximum one- and nine-hour average concentrations were less likely to be under-predicted by

28   the steady-state models compared to the Lagrangian puff models. The characteristic of the steady-state

29   models not to under-predict maximum concentrations make them well suited for regulatory compliance

30   demonstration, whereas the Lagrangian puff models are better suited for dose reconstruction and long

31   range transport.

32   **Keywords:** air dispersion models, CALPUFF, AERMOD, validation, tracer

33

34

1. INTRODUCTION

The steady-state model AERMOD and Lagrangian puff model CALPUFF are the U.S. Environmental Protection Agency (EPA) preferred models for demonstrating regulatory compliance in the near field (< 50 km) and far field (>50 km), respectively. The CALPUFF model has also been used in non-regulatory retrospective studies of radiation dose in the near field (Rood et al. 2008) and far field environments (Grogan et al. 2007). Demonstration of regulatory compliance and accident consequence analysis are generally prospective assessments, whereas dose reconstruction and epidemiological studies are generally retrospective in nature. The assessment questions for the prospective and retrospective analyses are fundamentally different and require different model performance objectives.

For the prospective analysis, the assessment question is whether air emissions will exceed ambient air quality standards, or result in impacts that are unacceptable. This assessment question can initially be addressed using conservative assumptions and simple models. It may not be critical to accurately estimate temporal and spatial variations in concentration, as long as the estimated impacts do not exceed the standards within a safety margin of error. More detailed model applications may be required if simple models cannot demonstrate that regulatory standards are achieved.

For a retrospective assessment, the assessment question is an unbiased estimate of the temporal and spatial distribution of concentration and deposition. Examples of a retrospective analysis include the dose reconstructions performed at U.S. Department of Energy Facilities (Farris et al. 1994; Till et al. 2000, 2002; Rood et al. 2002) and other special studies (Rood et al. 2008; Grogan et al. 2007). Simple models may be used in initial scoping calculations. However, ultimately an unbiased estimate of the temporal and spatial distribution of air concentration and deposition with estimated uncertainty is desired.

The purpose of this paper is to examine the performance of AERMOD, CALPUFF, and two legacy models using the Winter Validation Tracer Study (WVTS) dataset conducted in February 1991 at the former Rocky Flats Environmental Technology Site (RFETS). Performance objectives were tailored toward addressing the assessment questions posed by the prospective and retrospective analysis. Two

60 legacy models, Industrial Source Complex Short Term Version 2 (ISC2) (EPA 1992) and Regional

61 Atmospheric Transport Code for Hanford Emission Tracking (RATCHET) (Ramsdell et al. 1994), were

62 included in the evaluation because the formulations of these models are currently used in radiological

63 assessment codes (EPA 2007; Chanin et al. 1998; Ramsdell et al. 2010). Model simulations and

64 performance evaluation of ISC2 using the WVTS was originally reported in Haugen and Fontino (1993).

65 Performance evaluation of RATCHET using the WVTS dataset was originally reported in Rood (1999)

66 and Rood et al. (1999). Model-predicted concentrations for ISC2 and RATCHET were taken from

67 Haugen and Fontino (1993) and Rood (1999), respectively, and were used without modification.

68   A description of the tracer measurements and meteorological data is provided first, followed by

69 modeling protocol, performance objectives, and performance measures. Finally, the model performance

70 results, in terms of addressing the prospective and retrospective assessment questions, are discussed.

71

72   **2. METHODS**

73   Model performance was evaluated in terms of fundamental plume properties, paired ensemble mean

74 concentrations, and concentrations unpaired in space. The WVTS dataset and meteorological data are

75 presented first, followed by modeling protocol, performance objectives, and performance measures.

76 **2.1. Winter Validation Tracer Study**

77   The WVTS was conducted in February 1991 near the former RFETS located on the Front Range of

78 the Colorado Rocky Mountains about 25 km northwest of Denver (Brown 1991). The study consisted of

79 12 separate tests (Table 1). For each test, an inert tracer (sulfur hexafluoride [SF_6]) was emitted

80 continuously for 11 hours from a 10-m high stack located on the east side of the main plant complex (Fig.

81 1). The main plant complex was located about 2.5 km east of the foothills on an alluvial plain ranging in

82 elevation from 1,750 m to 1,850 m above sea level. The primary purpose of the study was to gather data

83 for validation of emergency response atmospheric transport models. Samplers were arranged in

84 concentric circles 8-km and 16-km from the release point so as to capture any possible transport

85 trajectories. One-hour average air concentrations were then measured for the last nine hours of the release

86   at each of the 140 samplers. Six tests were performed under nighttime conditions, four under daytime

87   conditions, one under day-night transition, and one under night-day transition. A total of 108 hours of

88   data were recorded. Seventy-two samplers were distributed at the 8-km distance and 68 samplers at the

89   16-km distance. Sampler elevations ranged from about 1,600 m to 2,600 m above sea level. The study

90   domain is considered near field because the maximum distance to the samplers is <50 km.

91   Previous investigators (Haugen and Fontino 1993) used this data set in a performance evaluation of

92   the TRAC (Hodgin 1991) and ISC2 models. The electronic copy of this data set was obtained from

93   Haugen and Fontino (1993) for use in the model performance evaluation for the Historical Public

94   Exposures Studies at Rocky Flats (Rood 1999). These data included the observed hourly-average

95   concentrations for all 12 tests, the sampler ID numbers and locations, and the TRAC and ISC2 predicted

96   concentrations. The ISC2 results provided by Haugen were used in this paper without modification.

97   **2.2. Meteorological Data**

98   Meteorological data were recorded for every tracer test at the 10-m and 61-m level from the RFETS

99   61-m tower located 790 m west and 87 m south of the release point. Only data from the 10-m level were

100   used in the model simulations. Data were provided as 15-minute averages of wind speed and direction,

101   temperature, heat flux, and standard deviations of these parameters. Hourly averages of these data were

102   calculated using EPA protocol (EPA 2000). Mixing depth estimates for the ISC2 simulations were

103   derived from linear interpolation for each 15-minute period from the rawinsonde data taken every 12

104   hours at Denver Stapleton airport. No precipitation was measured during any of the 12 tests, and there

105   was no snow cover during February 1991.

106   Additional surface and upper air meteorological data were obtained from Denver Stapleton airport

107   located about 25 km southeast of the RFETS. Surface data included wind speed and direction, cloud

108   cover, and ceiling height and were used with the RFETS data to calculate stability class. The RATCHET

109   and CALPUFF models utilized the wind speed and direction measurements in their wind field

110   interpolation.

5

111   Surface observations from Denver Stapleton are strongly influenced by air movement within the

112   Platte River Valley, which flows to the northeast from the city center. By contrast, Rocky Flats is more

113   strongly influenced by its proximity to the foothills. Both locations are influenced by the diurnal pattern

114   of upslope-downslope conditions that characterize the general air movement on the Colorado Front Range

115   environs. Downslope conditions typically occur during the evening hours and are characterized by

116   drainage flow of cooler surface air from the foothills and upper reaches of the Platte River Valley

117   northeastward to the plains. Airflow at Rocky Flats is typically from the west-northwest, and converges

118   with the flow from the south within the Platte River Valley in a broad zone 20 km to 30 km east-northeast

119   of the RFETS (Lange, 1992). During daylight hours and after surface heating has eliminated the cooler

120   surface layer, the downslope conditions cease. This is followed by a brief period of relatively calm winds,

121   which in turn is followed by return of air up the valley or upslope conditions. Upslope conditions were

122   weak to non-existent during the WVTS.

123   2.3. Atmospheric Transport Models and Protocol

124   A brief description of each of the models included in this study is presented along with the modeling

125   protocol. Because the SF_6 tracer is an inert gas, all model simulations did not include deposition and

126   plume depletion.

127   2.3.1. AERMOD

128   The American Meteorological Society and EPA developed the AERMOD modeling system

129   (Cimorelli et al. 2004). Model development began in 1991 with the objective to incorporate current

130   planetary boundary layer concepts into regulatory compliance models. Treatment of both surface and

131   elevated point sources, area sources, and volume sources in a simple or complex terrain model domain are

132   addressed in the model. It was intended as a replacement of the Industrial Source Complex Version 3

133   (ISC3) model. Currently, AERMOD is the EPA's preferred model for regulatory compliance

134   demonstration for criteria pollutants in the near field (<50 km).

135 **2.3.1.1. AERMOD Modeling Protocol**

136   The modeling domain consisted of a A 46.2-km × 49.8-km region centered on the RFETS.

137 Meteorological data were the same used in the CALPUFF simulation and included surface and upper-air

138 data from Denver Stapleton airport and onsite data from RFETS.

139   AERMOD and AERMET version 12345 were used in the model simulations. Special processing of

140 AERMOD output was required because minimum model-simulation times are one-day, and nine-hour

141 averaging times are not an option. Nine-hour average concentrations were calculated by inputting the

142 tracer release rate for the last nine hours of the 11-hour tracer test and setting the release rate to zero for

143 the remaining hours in the simulation. The maximum one-hour, and 12-hour average concentration at

144 each of the samplers for a simulation period that encompassed each test was output. The nine-hour

145 average concentration was calculated by multiplying the 12-hour average concentration by the ratio of

146 12/9.

147 **2.3.2. CALPUFF**

148   The CALPUFF modeling system (Scire et al. 2002) is a non-steady-state Lagrangian puff model that

149 simulates pollutant transport, transformation, and deposition in a three-dimensional spatially and

150 temporally variable wind field. CALPUFF can be applied on local and regional scales. The modeling

151 system is composed of three primary modules: CALMET, CALPUFF, and CALPOST and collectively

152 these are referred to as the CALPUFF modeling system.

153   The CALMET module is a meteorological model that generates a three-dimensional hourly wind

154 field within a three-dimensional gridded modeling domain. The CALPUFF module uses the CALMET-

155 generated wind field and micrometeorological parameters to advect and disperse "puffs". The CALPOST

156 module reads the CALPUFF concentration and deposition flux files and produces time-averaged

157 concentration and deposition output along with visibility impacts.

7

158   CALPUFF is currently the EPA preferred long-range (>50 km) dispersion model for demonstration

159   of compliance with Prevention of Significant Deterioration increment levels and National Ambient Air

160   Quality Standards in Class I areas.

161   **2.3.2.1. CALPUFF Modeling Protocol**

162   A 45.2-km × 43.6-km model domain having a grid spacing of 400 m (113 × 109 grid cells) was

163   established (Fig. 1). It was centered approximately on the WVTS release point and included the Platte

164   River Valley in the southeast corner. Vertical discretization consisted of eight layers 20 m, 40 m, 60 m,

165   100 m, 200 m, 400 m, 800 m, 1,200 m, and 1,800 m above land surface.

166   For the CALMET simulations, EPA-Federal-Land-Manager-recommended parameter values (Fox

167   2009) or CALMET default values were generally used where applicable. The model parameters BIAS,

168   RMAX1, RMAX2, TERRAD, R1, and R2 were chosen on a site-specific basis. The BIAS parameter

169   assigns weights to the surface and upper-air stations data for each vertical layer. Surface data were given

170   100% of the weight (BIAS = 1) in the first layer with zero weight in the last two vertical layers (BIAS =

171   1). Equal weight was assigned to the fourth layer (BIAS = 0) with a gradation of weights between the

172   lower and upper for the remaining layers.

173   The RMAX1 and RMAX2 parameters define the maximum radius of influence for surface and upper

174   data, respectively, over land surfaces. To incorporate the influence of flow in the Platte River Valley as

175   represented by the surface data at Denver Stapleton airport, a value of 32 km for RMAX1 was used. The

176   RMAX2 value serves the same purpose as RMAX1 but is used for upper-air data. A value of 100 km was

177   selected for the RMAX2 parameter

178   The TERRAD parameter defines the radius of influence of terrain features. A TERRAD value of 7

179   km was used so that the terrain influence of the foothills encompassed RFETS. The parameters R1 and R2

180   are the distance from an observation where the observation and the initial guess field are equally weighted

181   for surface and layers aloft, respectively. A value of 17 km and 56 km were chosen for R1 and R2,

182   respectively.

183   The CALPUFF runs were performed using dispersion coefficients calculated from

184 micrometeorological variables (MDISP = 2), and the simple CALPUFF-type terrain adjustment algorithm

185 (MCTADJ = 2). The remaining parameters were CALPUFF defaults. CALPUFF Version 5.8, Level

186 070623, and CALMET Version 5.8, Level 060811, were used in the model simulations.

187 **2.3.3. Industrial Source Complex Short Term Version 2 (ISC2)**

188   The ISC2 model is an augmented steady-state Gaussian plume model primarily used to demonstrate

189 regulatory compliance with criteria pollutants emitted by industrial facilities. It was replaced by ISC

190 Version 3 and was the EPA preferred model until the promulgation of AERMOD in December of 2005.

191 The model was included in the evaluation because straight-line Gaussian plume models form the basis of

192 the CAP88 (EPA 2007) model for demonstration of compliance with the Clean Air Act, and of the

193 MACCS2 code (Chanin et al. 1998) for reactor accident consequence analysis.

194 **2.3.3.1. ISC2 Modeling Protocol**

195   The ISC2 simulations were performed by Haugen and Fontino (1993) using 15-minute

196 meteorological data from the 10-m level taken at the RFETS and mixing depth estimated from the

197 rawinsonde data at Denver Stapleton airport. Four ISC2 runs were performed for each hour of simulation

198 using the four, 15-minute average data from the Rocky Flats meteorological station. The results from the

199 four simulations were averaged to provide hourly-average concentrations at each of the sampler locations.

200 These hourly concentrations were then averaged across each test by Rood (1999) to provide nine-hour

201 average concentrations at each sampler.

202   The performance evaluation of this model was originally reported in Rood (1999) and Rood et al.

203 (1999). The results presented here are based on the original data using slightly different performance

204 measures.

2.3.4. RATCHET

The Regional Atmospheric Transport Code for Hanford Emission Tracking (RATCHET) (Ramsdell et al. 1994) is a Lagrangian puff model developed by Pacific Northwest Laboratories for the Hanford Dose Reconstruction Project (Farris et al. 1994). Its primary purpose was to estimate transport and deposition of ^{131}I released from the Hanford facility across a 194,250 km^2 model domain located in eastern Washington State. The model includes a surface wind field interpolator that allows incorporation of multiple surface meteorological stations into a model simulation. Upper-air data were not considered in the model. Terrain complexities were not explicitly treated, but are implicitly represented by using multiple meteorological surface stations that reflect major topographical features. Surface roughness features are spatially variable across the model domain. Diffusion coefficients are estimated from statistics of atmospheric turbulence that are inferred from estimates of atmospheric stability, surface roughness length, and the Monin Obukhov length. The current radiological assessment models, RASCAL (Ramsdell et al. 2010) and GENII (Napier, 2009), employ the RATCHET air dispersion model.

2.3.4.1. RATCHET Modeling Protocol

Model simulations with RATCHET were performed by the author (Rood et al. 1999) as part of Phase II of the Historical Public Exposures at Rocky Flats (Till et al. 2002). Hourly-average meteorological data at the 10-m level from the Rocky Flats plant and Denver Stapleton airport were used in the model simulations. A 37-km × 37-km model domain centered on the RFETS with 500-m grid spacing was established. RATCHET does not allow discrete receptors, and therefore, calculated concentrations were extracted from the grid node nearest the sampler. Surface roughness lengths (z_0) ranged from 2-m in the foothills to 0.05-m in the farmland east of the RFETS.

2.4. Performance Objectives

Performance objectives consisted of four fundamental plume properties and a paired and unpaired comparison of individual samplers. The four fundamental plume properties were maximum concentration, location of the plume maximum, plume width, and arc-integrated concentration. An additional objective

230   of the maximum one-hour average concentration unpaired in space and time was also included to provide

231   insight into model performance for short-term maximum concentrations. Descriptions of each modeling

232   objective follows.

233   **2.4.1. Maximum Hourly and Nine-Hour Average Concentration and Plume Maximum Location**

234   This modeling objective compared the predicted and observed maximum one-hour and nine-hour

235   average concentration measured at a sampler during the nine-hour test period at either the 8-km or 16-km

236   distance from the release point. The predicted maximum concentration was not paired in space, and also

237   unpaired in time for the maximum-hourly average concentration. The nine-hour average concentration

238   was determined by a simple arithmetic average of the nine, one-hour average concentrations. Sampler

239   data that were missing were not included when computing the predicted or observed average

240   concentration.

241   The plume maximum location was only computed for the nine-hour average concentration and was

242   quantified in terms of the absolute value of angular difference between the predicted and observed

243   location of the plume maximum.

244   **2.4.2. Plume Width**

245   The plume width objective evaluated the predicted impact area of the plume. Each sampler was

246   assigned an arc length equal to the arc length between the midpoints of the sampler and each of its

247   adjacent samplers. The plume width was sum of the arc lengths of samplers that had a concentration

248   greater than zero, or in the case of the observed values, a concentration greater the the minimum

249   detectable concentration.

250   **2.4.3. Arc-Integrated Concentration**

251   The arc-integrated concentration evaluated the plume mass at the 8-km and 16-km distance. The arc-

252   integrated concentration is the sum of the product of the sampler arc lengths as defined in Section 2.4.2

253   and the nine-hour average predicted or observed concentration.

254 **2.4.4. Unpaired Time-Averaged Concentration**

255   This modeling objective compared the ranked predicted and observed time-averaged (nine-hour)

256 concentrations. Only predicted and observed concentrations that met the selection criteria stated in

257 Section 2.6 were included. Samples were blocked into those performed at night (Tests 1, 2, 3, 4, 10, and

258 11), those performed during the day (Tests 6, 7, 9, and 12), and those performed during transition periods

259 (Tests 5 and 8). Sample blocking is used in bootstrap resampling to avoid block-to-block variance (Chang

260 and Hanna 2005).

261 **2.4.5. Paired Ensemble Means (ASTM Procedure)**

262   The American Society for Testing and Materials (ASTM) proposed a procedure for evaluation of

263 models, recognizing that model predictions are ensemble-mean predictions, while observations

264 correspond to realizations of ensembles (ASTM 2000). An ensemble is defined as a set of experiments

265 having fixed external conditions, such as meteorological conditions and downwind distance. In the

266 WVTS, fixed external conditions were the distance to the sampling arc and meteorological conditions. In

267 general, repeatable diurnal flow and stability regimes are established during nighttime, daytime, and day-

268 night transitional periods along the Colorado Front Range. Thus, averages across the tests representing

269 these similar conditions would approximate ensemble means to compare with model predictions for the

270 same period.

271   Predictions and observations were grouped into three blocks consisting of nighttime (six tests),

272 daytime (four tests), and transition period (two tests). Average concentrations were calculated across all

273 tests in the block for each sampler, and performance statistics were calculated separately for each block.

274 ## 2.5. Performance Measures

275   Several simplified measures were used to evaluate model performance (Cox and Tikvart 1990; Weil

276 et al. 1992). These measures were the fractional bias (*FB*) and normalized mean square error (*NMSE*).

277 Fractional bias is given by

278
$$FB = \frac{2(\overline{C}_o - \overline{C}_p)}{\overline{C}_o + \overline{C}_p} \qquad (1)$$

279 where C_p and C_o are the predicted and observed concentrations, respectively. Overbars indicate averages

280 over the sample. The *NMSE* given by

281
$$NMSE = \frac{\overline{(C_o - C_p)^2}}{\overline{C}_o\,\overline{C}_p} \qquad (2)$$

282 The *FB* is a measure of mean bias. A *FB* of 0.6 is equivalent to model under-prediction by about a factor

283 of two. A negative value indicates model over-prediction. The *NMSE* is a measure of variance, and a

284 value of 1.0 indicates that a typical difference between predictions and observations is approximately

285 equal to the mean. The *NMSE* and *FB* are appropriate when the typical difference between the predictions

286 and observations are approximately a factor of two (Hanna et al. 1991) and the range of predictions and

287 observations in the dataset is small (i.e., less than a factor of two). This was not the case in this study

288 where ratios of model predictions to observations often ranged from 0.01 to 100, and within a data set, the

289 predicted and observed concentrations ranged from the zero to ~10,000 ng m^{-3}. In these cases a log-

290 transformed measure of model bias and variance is more appropriate because it provides a more balanced

291 approach (Hanna et al. 1991). The log-transformed measures described in Hanna et al. (1991) are the

292 geometric mean bias (*MG*) and the geometric mean variance (*VG*) and are defined by

293
$$MG = \exp\left(\overline{\ln C_o} - \overline{\ln C_p}\right) = \exp\left(\overline{\ln \frac{C_o}{C_p}}\right) \qquad (3)$$

294
$$VG = \exp\left[\overline{\left(\ln C_o - \ln C_p\right)^2}\right] \qquad (4)$$

295 Geometric mean bias values of 0.5 and 2.0 indicate a factor of two over-prediction and under-prediction,

296 respectively. A *VG* value of 1.6 indicates about a factor of two difference between predicted and observed

297 data pairs.

298   A more easily understood log-transformed quantity that is related to the *MG* and *VG* is the geometric

299   mean (*GM*) and geometric standard deviation (*GSD*) of the predicted-to-observed ratio (C_p/C_o). The *GM*

300   and *GSD* are given by

301
$$GM \; = \; \exp\!\left(\overline{\ln \frac{C_p}{C_o}}\right) \tag{5}$$

302
$$GSD \; = \; \exp\!\left[\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\ln \frac{C_{p_i}}{C_{o_i}} \; - \; \overline{\ln \frac{C_p}{C_o}}\right)^2}\right] \tag{6}$$

303   where *n* = the sample size. Because the *MG* is simply the inverse of the *GM*, only the *GM* is reported. A

304   perfect model would have *FB* and *NMSE* values of 0, and *GM, GSD*, and *VG* values of 1.0. With the

305   exception of the plume width and location of plume maximum performance objectives, the log-

306   transformed performance measures are considered more appropriate than the *FB* and *NMSE*, and thus

307   only the log-transformed measures are reported. The location of plume maximum does not lend itself to

308   the above performance measures, mainly because the objective considered the absolute angular difference

309   between the predicted and observed location of maximum. For this performance objective, the mean

310   difference, standard deviation of the mean (i.e., standard error), and the minimum and maximum

311   differences are reported. Because differences between predicted and observed values and the range of

312   predictions and observations were less than about a factor of two, the *FB* and *NMSE* were considered

313   more appropriate for the plume width objective.

314   In addition to the above measures, the correlation coefficient (*r*) between predicted and observed

315   values and the number of predictions within a factor of two of the observations were also reported. The

316   correlation coefficient was determined using least-squares linear regression and log-transformed data

317   except for the plume width performance objective. Scatter plots were also included as visual measures of

318   performance for the paired ensemble means and unpaired time-averaged concentration modeling

319   objectives.

14

320   Confidence intervals were estimated for each of the performance measures using the bootstrap

321   methodology described in BOOT software (Hanna et al. 1991; Chang and Hanna 2005). Confidence

322   intervals were used to determine if the estimated performance measure was significantly different than its

323   optimum value and whether a statistically significant difference existed between the performance

324   measures for each model. Confidence interval estimates were based on the cumulative density function

325   generated from 1000 bootstrap samples.

2.6. Selection Criteria

327   The observed data set only reported nonzero hourly average concentrations greater than the minimum

328   detectable sampler concentration (*mdc*) of 33 ng m^{-3}. Measured concentrations below this value were

329   reported as zero. A sampler that had only one hour of data (in the nine-hour measurement period) greater

330   than the *mdc* would have a nine-hour average concentration of 3.7 ng m^{-3} (33 ng m^{-3} /9). This value

331   represents the nine-hour time-averaged *mdc* for a sampler.

332   For the paired ensemble means performance objective, the dataset was based on the union of the

333   predicted and observed concentrations. The *mdc* was substituted for predicted concentrations that were

334   less than the *mdc* if the paired observed concentration was greater than zero. Likewise, the *mdc* was

335   substituted for observed concentrations less than the *mdc* if the paired predicted concentration was greater

336   than zero. Predicted and observed pairs that were both zero were omitted from the analysis.

337   For the unpaired analysis only predicted *and* observed concentration pairs greater than the *mdc* were

338   considered. Samplers missing all nine hours of data were eliminated from the data set.

3. RESULTS

340   The paired ensemble means and unpaired scatter plots are perhaps the most illustrative in terms of

341   summarizing model performance qualitatively (Figs. 2, 3, 4 and 5). In general, the highest predicted and

342   observed concentrations were during nighttime and transition period tests and the lowest during daytime

343   tests.

344   The paired ensemble mean scatter plot at the 8-km distance (Fig. 2) showed the transition period tests

345   as having the highest observed concentrations, some exceeding 8000 ng m^{-3}. Nighttime tests had

346   maximum observed concentrations between 3000 and 4000 ng m^{-3}. As expected, daytime tests had the

347   lowest observed maximum concentrations, the maximum being slightly less than 600 ng m^{-3}. All models

348   performed poorly for the transition period tests, underestimating observed concentrations that were >1000

349   ng m^{-3}. In general, the puff models exhibited better correlation to the observations for daytime and

350   nighttime tests and concentrations that were >100 ng m^{-3} compared to the steady-state models.

351   At the 16-km distance (Fig. 3) the nighttime period tests had the highest observed concentrations, (~

352   3500 ng m^{-3}), followed by transition period tests (~ 1800 ng m^{-3}). Daytime tests had maximum observed

353   concentrations that were ~70 ng m^{-3}. The puff models exhibited better correlation, less variability, and a

354   greater number of points within a factor of two of the observations compared to the steady state models

355   for nighttime tests and concentrations >100 ng m^{-3}.

356   Scatter plots of the unpaired data at the 8-km distance (Fig. 4) showed that all models underestimated

357   transition period observed concentrations that were greater than 1000 ng m^{-3}. Predicted concentrations

358   from RATCHET were within a factor of two of the observations for almost all the daytime tests and most

359   of the nighttime tests for the entire concentration range. Most of the ISC2 concentrations for daytime and

360   nighttime tests were within a factor of two of the observations for concentrations that were >100 ng m^{-3}.

361   At the 16-km distance (Fig.5), scatter plots of the unpaired data were similar to those at the 8-km

362   distance, although CALPUFF underestimated almost all the concentrations for transition period tests by

363   more than a factor of two. The three highest observed concentrations were within a factor of two of the

364   corresponding AERMOD predicted concentrations. A similar result was found for ISC2, except the

365   highest observed nighttime concentration was underestimated by more than a factor of two. Observed

366   nighttime concentrations that were <100 ng m^{-3} were overestimated by more than a factor of two by

367   RATCHET.

368

16

3.1. Maximum One-Hour and Nine-Hour Average Concentration

Performance measure results for the maximum one-hour average concentration modeling objective (Table 2) indicate a strong positive bias for the steady-state models, especially AERMOD, and nearly no bias for puff models RATCHET and CALPUFF (GM confidence interval included 1.0). The positive bias for the steady-state models was greater at the 16-km distance. Ninety-two percent of the ISC2-estimated maximum one-hour average concentrations and 83% of the AERMOD values had predicted-to-observed ratios of 0.95 or higher. In contrast, only 50% of the CALPUFF- and RATCHET-estimated maximum one-hour average concentrations had a predicted- to-observed ratio greater than 0.95.

Predicted maximum nine-hour average concentrations (Table 3) showed a similar trend to those of the maximum hourly-average concentrations. That is, the steady-state models exhibited positive bias, while the Lagrangian puff models exhibited negative bias. However, the GM confidence interval included 1.0 for all models. Measures of variance were generally lower for the puff models and correlation coefficients were higher compared to the steady state models.

Measures of bias among the steady state models were significantly different from those of the puff models (Table 4). None of the model performance measures for CALPUFF and RATCHET were significantly different from one another.

3.2. Plume Maximum Location, Plume Width, and Arc-Integrated Concentration

Plume maximum location at the 8-km distance (Table 5) showed that the mean deviation was smallest for AERMOD and RATCHET (14 degrees) and greatest for CALPUFF and ISC2 (26 and 24 degrees respectively). Based on a t-test difference of the means, the mean deviation for CALPUFF and ISC2 was significantly different from the mean deviation for AERMOD and RATCHET ($P>0.005$).

Plume width performance measures at the 8-km distance (Table 5) showed that ISC2 and CALPUFF underestimated plume width while RATCHET and AERMOD overestimated plume width. Although the FB confidence interval for CALPUFF included the optimum value of zero and CALPUFF had the smallest NMSE value.

17

395   Arc-integrated concentration at the 8-km distance showed little bias for the steady state models (GM

396   confidence interval included 1.0) and a negative bias for the puff models. However, only 50% of the

397   predictions were within a factor of two for the steady-state models while over 90% of the predictions

398   were within a factor of two for the puff models.

399   Plume maximum location at the 16-km distance (Table 6) showed that the mean deviation was

400   smallest for CALPUFF and RATCHET (26 and 24 degrees respectively) and greatest for ISC2 and

401   AERMOD (34 and 36 degrees respectively). Based on a t-test difference of the means, the differences

402   between the steady state and puff models were significant at the 99% level ($0.01 < P < 0.005$).

403   Plume width performance measures at the 16-km distance (Table 6) showed that ISC2

404   underestimated plume width while RATCHET and AERMOD overestimated plume width. The

405   CALPUFF *FB* was not significantly different from zero.

406   Arc-integrated concentration at the 16-km distance showed positive bias for ISC2. The GM

407   confidence interval for the other models included the optimum value of 1.0. Puff models showed a greater

408   percentage of predictions within a factor of two of the observations.

409   Significant differences among models (Table 7) were noted for the bias performance measures and

410   the correlation coefficients.

3.3. Unpaired Nine-Hour Average Concentration

412   The performance measure results at the 8-km distance for the unpaired nine-hour average

413   concentration (Table 8) indicated a negative bias in predicted concentrations for AERMOD and

414   CALPUFF, slight negative bias for ISC2, and positive bias for RATCHET. ISC2 and RATCHET also

415   had the smallest variance, highest correlation coefficient, and highest percentage of predictions within a

416   factor of two of the observations.

417   At the 16-km distance, ISC2 and RATCHET exhibited positive bias, AERMOD exhibited negative

418   bias, and CALPUFF nearly no bias. RATCHET and ISC2 had the highest correlation coefficients and

419   CALPUFF and RATCHET had the highest percentage of predictions within a factor of two of the

420   observations.

421   Except for AERMOD and CALPUFF at the 8-km distance, and ISC2 and RATCHET at the 16-km

422   distance, all bias performance measures among the models were significantly different from one another

423   (Table 9).

3.4. Paired Ensemble Means

425   No one model showed overall better performance across all ensemble groups and all models

426   performed poorly for the transition period ensemble means (Tables 10 and 11). However, excluding the

427   transition ensemble means, RATCHET and CALPUFF had the highest percentage of predictions within a

428   factor of two of the observations, the highest correlation coefficients, and generally the lowest variance

429   compared to the steady-state models. For daytime and nighttime ensemble means, the GM confidence

430   interval for RATCHET encompassed 1.0 at both the 8- and 16-km distances. AERMOD was biased low

431   for daytime tests and showed little bias for nighttime tests (GM confidence interval encompassed 1.0).

432   ISC2 exhibited no bias for daytime tests (GM=1.0) at the 8-km distance but was biased high at the 16 km

433   distance (GM confidence interval excluded 1.0). For nighttime tests, ISC2 was biased low at the 8-km

434   distance but exhibited nearly zero bias at the 16 km distance. CALPUFF was biased low for daytime and

435   nighttime tests at the 8-km distance (GM confidence interval excluded 1.0), but the GM confidence

436   interval encompassed 1.0 at the 16-km distance.

437   All models were biased low for the transition period ensemble means at both the 8 and 16-km

438   distances, and exhibited large variances, although the variance measures for CALPUFF and RATCHET

439   were considerable smaller than those for ISC2 and AERMOD. Puff model *VG* and *r* values were

440   significantly different than steady-state models for nighttime tests at the 16-km distance (Table 12).

4. DISCUSSION

442   Model performance is judged in terms of the assessment question that the model is intended to

443   address. As stated in the introduction, the assessment questions are different for a prospective regulatory

444   compliance calculation compared to a retrospective dose reconstruction.

445   In terms of the prospective assessment where it is important that regulatory limits are not exceeded

446   (i.e., highest concentrations are not underestimated), the steady-state models were less likely to

447   underestimate maximum one- and nine-hour average concentrations compared to the Lagrangian puff

448   models. However, this result is not only due to differences in model formulation, but also the model

449   parameters such as diffusion coefficients. The unpaired scatter plots (Fig. 4 and 5) showed that the

450   *maximum* observed concentration across all tests was not underestimated by AERMOD at both the 8 and

451   and 16-km distance, although the time and place of the observed maximum was not the same as the

452   predicted maximum.

453   In terms of the retrospective assessment where the objective is an unbiased estimate of the

454   concentration in space and time, the Lagrangian puffs models showed overall better performance,

455   especially at the 16-km distance. In most cases, the Lagrangian puff models for the paired ensemble

456   means exhibited lower variance higher correlation to observed values and a higher percentage of

457   observations within a factor of two of the observations compared to steady-state models.

458   In terms of the four fundamental plume properties, the steady-state models tended to overestimate

459   maximum concentrations but provide unbiased estimates of the plume mass at the 8-km distance and the

460   16-km distance for ISC2. Puff models tended to slightly underestimate plume maximums, but were better

461   at locating the plume maximum at the 16-km distance. CALPUFF appeared to more accurately estimate

462   the plume impact region, whereas AERMOD and RATCHET tended to overestimate it and ISC2

463   underestimated it.

464   The WVTS consists of only 108 hours of measurements taken during February 1991 and are not

465   representative of annual average concentrations. However, the high sampler density resulted in the

466   likelihood that the maximum concentration was detected at either the 8-km or 16-km sampling distance.

467   Moreover, the tests were conducted during the wintertime when stable dispersion conditions would likely

468   result in the maximum one- or eight-hour average concentration over the course of a year. Achieving

469   compliance with National Ambient Air Quality Standards typically is limited by the short-term average

470 concentration limits. Therefore, these results have relevance in terms of model performance for short-term

471 averages over the period of a year.

5. CONCLUSIONS

473   No one single model consistently out-performed the others in all performance objectives or measures

474 and the state-of the art models (CALPUFF and AERMOD) did not exhibit superior performance in all

475 performance objectives to the legacy models (ISC2 and RATCHET). Lagrangian puff models generally

476 exhibited smaller variances, higher correlation, and higher percentage of predictions within a factor of

477 two compared to the steady-state models at these distances. The conceptual framework of a Lagrangian

478 puff model is better suited for long range transport where winds vary spatially across the model domain.

479 Hence, Lagrangian puff models may be preferable for dose reconstruction where model domains can be

480 large and where the assessment question is an unbiased estimate of concentration in time and space.

481 However, model choice depends on site-specific considerations and the assessment questions to be

482 addressed, and therefore no categorical statement can be made about the performance of one type of

483 model over the other for a specific application.

484   The steady-state models generally did not underestimate the high-end concentrations at the distances

485 studied, and therefore provide a sound basis for regulatory compliance modeling. Based on the overall

486 performance of ISC2, assessment models that rely on the Gaussian plume model are not necessarily

487 inferior to the current state-of-the-art models in terms of meeting regulatory performance objectives.

488   There was a general tendency for the steady-state models to predict relatively higher concentrations at

489 the 16-km distance compared to the 8-km distance. This effect is important because it manifests itself at

490 substantially shorter distances (16 km) than what is defined by the EPA as the near-field environment (≤

491 50 km). The EPA requires AERMOD for the near-field environment for demonstration of regulatory

492 compliance, unless compelling reasons are provided to justify the use of an alternative model. Thus,

493 estimated maximum hourly-average concentrations at distances >16 km are likely to be overestimated

494 based on these results. Other investigators (Dresser and Huizer 2011) found AERMOD to underestimate

21

495 near-field maximum one- and three-hour average SO_2 concentrations based on data from the Martins

496 Creek Power Plant. However, only eight samplers were used in the study and it is possible (if not likely)

497 that the true maximum concentration in the model domain was not captured. In the WVTS, it was less

498 likely that the maximum concentration within a sampling arc went undetected.

499   Finally, a compelling reason to use steady-state models for regulatory compliance demonstration is

500 the fact that they are simpler to run, require less user judgment, and are less prone to error than

501 Lagrangian puff models. The CALMET/CALPUFF model simulation in this paper required numerous

502 iterations using different values of RMAX1, RMAX2, and other parameters so that the wind field

503 matched what was expected. In a prospective analysis, it is unlikely tracer or other validation data would

504 be available to test model performance and adjust model parameters accordingly to improve model

505 performance. The need for consistency and assurance that estimated concentrations are not

506 underestimated are legitimate reasons for using steady-state models for regulatory compliance

507 determination.

Acknowledgments

510

6. REFERENCES

511 ASTM (American Society for Testing and Materials), 2000. Standard guide for statistical evaluation of

513 atmospheric dispersion model performance. West Conshohocken, Pennsylvania: American Society

514 for Testing and Materials; Designation D 6589-00.

515 Brown, K.J., 1991. Rocky Flats 1990–91 winter validation tracer study. North American Weather

516 Consultants, Salt Lake City, Utah, Report AG91-19.

517 Chang J.C. and Hanna S.R, 2005. Technical descriptions and user's guide for the BOOT statistical model

518 evaluation software package, version 2.0.

519 Chanin D., Young M. L., and Randall J., 1998. Code manual for MACCS2. NUREG/CR-6613, U.S.

520 Nuclear Regulatory Commission, Washington, D.C.

521 Cimorelli, A.J., Perry, S.G., Venkatram, A., Weil, J.C., Paine, R.J., Wilson, R.B., Lee, R.F., Peters, W.D.,

522 Brode, R.W., and Paumier, J.O., 2005. AERMOD: Description of model formulation. EPA-454/R-

523 03-004, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina.

524 Cox W.M., and Tikvart, J.A., 1990. A statistical procedure for determining the best performing air quality

525 simulation model. Atmospheric Environment 24A: 2387–2395.

526 Dresser, A.L. and Huizer, R.D., 2011. CALPUFF and AERMOD model validation study in the near field:

527 Martins Creek revisited. Journal of the Air and Waste Management Assoc. 61: 647–659.

528 EPA (U.S. Environmental protection Agency), 1992. User's guide for the industrial source complex (ISC)

529 dispersion models vol. 1, user's instructions. EPA-450/4-92-008a, U.S. Environmental Protection

530 Agency, Research Triangle Park, North Carolina.

531 EPA, 2000. Meteorological monitoring guidance for regulatory modeling applications. EPA-454/R-99-

532 005, U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research

533 Triangle Park, North Carolina.

534 EPA, 2007. Updated user's guide for CAP88-PC version 3.0. EPA 402-R-00-004, U.S. Environmental

535 Protection Agency, Office of Radiation and Indoor Air, Washington D.C.

536   Farris, W.T., Napier. B.A., Eslinger, P.W., Ikenberry, T.A., Shipler, D.B., and Simpson, J.C., 1994.

537   Atmospheric pathway dosimetry report, 1944–1992. PNWD-2228 HEDR, Pacific Northwest

538   Laboratories , Richland, Washington.

539   Fox, T,J., 2009. Clarification on EPA-FLM recommended settings for CALMET. Memorandum from T.J.

540   Fox to Regional, Modeling Contacts, August 31, 2009. U.S. Environmental Protection Agency,

541   Research Triangle Park, NC.

542   Grogan H.A., Aanenson J.W., McGavran P.D., Meyer K.R., Mohler H.J., Mohler S.S., Rocco J.R., Rood

543   A.S., Till J.E., and Wilson L.H., 2007. Modeling of the Cerro Grande fire at Los Alamos: An

544   independent analysis of exposure, heath risk, and communication with the public. In: Applied

545   modeling and computations in nuclear science. ACS Symposium Series 945, American Chemical

546   Society, Washington, D.C.

547   Hanna S.R., Strimaitis D.G., and Chang J.C., 1991. Hazard response modeling uncertainty (a quantitative

548   method), volume 1: User's guide for software for evaluating hazardous gas dispersion models.

549   Florida: Air Force Engineering and Service Center, Tyndall Air Force Base.

550   Haugen D.A. and Fontino I. P., 1993. Performance evaluation of the terrain-responsive atmospheric code

551   (TRAC) model. Colorado School of Mines, Golden, Colorado.

552   Hodgin C.R., 1991. Terrain responsive atmospheric code (TRAC) transport and diffusion: Features and

553   software overview. RFP-4516, EG&G Rocky Flats, Golden, Colorado.

554   Lange R., 1992, Modeling the dispersion of tracer plumes in the Colorado Front Range boundary layer

555   during night- and day-time conditions. American Meteorological Society Tenth Symposium on

556   Turbulence and Diffusion, Portland Oregon.

557   Napier, B.A., 2009. GENII2 users' guide, rev 3. PNNL-14583, Pacific Northwest National Laboratories ,

558   Richland, Washington.

559 Ramsdell Jr J.V., Simonen C.A., and Burk K.W., 1994. Regional atmospheric transport code for Hanford

560   emission tracking (RATCHET). PNWD-2224-HEDR, Pacific Northwest Laboratories, Richland,

561   Washington.

562 Ramsdell Jr J.V., Athey G.F., McGuire S.A., and Brandon L.K., 2010. RASCAL 4: description of models

563   and methods. NUREG 1940, U.S. Nuclear Regulatory Commission, Office of Nuclear Security and

564   Incident Response, Washington, D.C..

565 Rood, A.S., 1999. Performance evaluation of atmospheric transport models, revision 1. 3-CDPHE-RFP-

566   1996-FINAL (Rev 1), Risk Assessment Corporation, Neeses, South Carolina.

567 Rood, A.S, Killough, G.G., and Till, J.E., 1999. Evaluation of atmospheric transport models for use in

568   phase II of the historical public exposures studies at the Rocky Flats plant. Risk Analysis 19 (4):

569   559–576.

570 Rood, A.S., Grogan, H.A., and Till, J.E., 2002. A model for a comprehensive assessment of exposure and

571   lifetime cancer incidence risk from plutonium released from the Rocky Flats plant, 1953–1989.

572   Health Physics 82 (2); 182–212.

573 Rood, A.S., Voillequé, P.G., Rope, S.K., Grogan, H.A., and Till, J.E., 2008. Reconstruction of

574   atmospheric concentrations and deposition of uranium and decay products released from the former

575   uranium mill at Uravan, Colorado USA. Journal of Environmental Radioactivity 99: 1258–1278.

576 Scire J.S., Strimatis D.J., and Yamartino R.J., 2002. A user's guide to the CALPUFF dispersion model

577   version 5.7. Earth Tech Inc, Concord, Massachusetts.

578 Till J.E., Killough G.G., Meyer K.R., Sinclair W.S., Voillequé P.G., Rope S.K., and Case M.J., 2000. The

579   Fernald dosimetry reconstruction project. Technology 7: 270–295.

580 Till J.E., Rood A.S., Voilleque P.G., McGavran P.D., Meyer K.R., Grogan H.A., Sinclair W.K.,

581   Aanenson J.W., Meyer H.R., Mohler H.J., Rope S.K., and Case M.J., 2002. Risks to the public from

582   historical releases of radionuclides and chemicals at the Rocky Flats Environmental Technology Site.

583   Journal of Exposure Analysis and Environmental Epidemiology 12 (5): 355–372.

584   Weil J.C., Sykes R.I., and Venkatram A., 1992. Evaluating air quality models: Review and outlook.

585   Journal of Applied Meteorology 31: 1121–1145.

Figure Captions

Fig 1. CALPUFF model domain showing location of the WVTS samplers and terrain features of the Colorado Front Range.

Fig 2. Scatter plots of paired nine-hour average ensemble mean concentrations at the 8-km distance. Points that lie within the shaded region are within a factor of 2 of the observations.

Fig 3. Scatter plots of paired nine-hour average ensemble mean concentrations at the 16-km distance. Points that lie within the shaded region are within a factor of 2 of the observations.

Fig 4. Scatter plots of unpaired nine-hour average concentrations at the 8-km distance. Points that lie within the shaded region are within a factor of 2 of the observations.

Fig 5. Scatter plots of unpaired nine-hour average concentrations at the 16-km distance. Points that lie within the shaded region are within a factor of 2 of the observations.

Table Captions

Table 1. Winter Validation Tracer Study start and end times and source strength

Table 2. Performance measures for the maximum one-hour average concentration unpaired in time and space modeling objective

Table 3. Performance measures for the maximum nine-hour average concentration unpaired in time and space modeling objective

Table 4. Significant differences in model performance measures for the maximum one- and nine-hour-average concentration unpaired in space and time modeling objective. An "X" indicates a significant difference.

Table 5. Performance measures for plume maximum location, plume width, and the arc integrated concentration at the 8-km distance.

Table 6. Performance measures for plume maximum location, plume width, and the arc integrated concentration at the 16-km distance.

Table 7. Significant differences in model performance measures for the plume width and arc integrated concentration modeling objective. An "X" indicates a significant difference.

Table 8. Performance measures for the unpaired nine-hour averaged concentration modeling objective.

Table 9. Significant differences in model performance measures for the unpaired nine-hour average concentration modeling objective. An "X" indicates a significant difference.

Table 10. Performance measures for the daytime, transition, and nighttime period ensemble means modeling objective at the 8-km distance.

Table 11. Performance measures for the daytime, transition, and nighttime period ensemble means modeling objective at the 16-km distance.

Table 12. Significant differences in model performance measures for the day, transition, and nighttime ensemble means modeling objective. An "X" indicates a significant difference.

Table 1. Winter Validation Tracer Study start and end times and source strength.

Test	Start date	Start time (MST)[a]	End date	End time (MST)[a]	MFC[b] (kg hr^{-1})	CWL[c] (kg hr^{-1})	Average (kg hr^{-1})
1	02/03/91	20:00:00	02/04/91	07:00:00	13.71	13.24	13.48
2	02/04/91	20:00:00	02/05/91	07:00:00	13.05	12.16	12.61
3	02/06/91	20:00:00	02/07/91	07:00:00	13.71	13.33	13.52
4	02/07/91	20:00:00	02/08/91	07:00:00	16.53	16.84	16.69
5	02/09/91	13:00:00	02/09/91	00:00:00	23.61	22.63	23.12
6	02/11/91	07:00:00	02/11/91	18:00:00	23.61	22.94	23.28
7	02/12/91	07:00:00	02/12/91	18:00:00	23.61	23.99	23.80
8	02/14/91	01:00:00	02/14/91	12:00:00	23.61	23.44	23.53
9	02/15/91	07:00:00	02/15/91	18:00:00	23.61	23.29	23.45
10	02/16/91	20:00:00	02/17/91	07:00:00	23.61	23.47	23.54
11	02/17/91	20:00:00	02/18/91	07:00:00	23.61	23.04	23.33
12	02/19/91	07:00:00	02/19/91	18:00:00	23.21	22.97	23.09

[a] Mountain standard time.

[b] Release rate calculated from mass flow controllers (MFCs) that were calibrated at 760 mm Hg, 21.11°C.

[c] Release rate determined from cylinder weight loss (CWL).

Table 2. Performance measures for the maximum one-hour average concentration unpaired in time and space modeling objective.

8-km results	ISC2	AERMOD	CALPUFF	RATCHET
GM Cp/Co	1.9	3.4	0.99	0.92
GM Confidence Interval	1.24 — 2.87	2.15 — 5.94	0.65 — 1.56	0.55 — 1.65
GSD Cp/Co	2.2	3.4	2.2	2.8
VG	2.7	17.4	1.8	2.6
log(VG) Confidence Interval	0.40 — 1.59	1.78 — 4.15	0.26 — 0.92	0.48 — 1.35
r	0.748	0.742	0.76	0.612
r Confidence Interval	0.55 — 0.92	0.46 — 0.90	0.58 — 0.96	0.27 — 0.89
% within a factor of 2	33.3%	25.0%	58.3%	58.3%
16-km results	ISC2	AERMOD	CALPUFF	RATCHET
GM Cp/Co	2.7	4.9	0.98	0.93
GM Confidence Interval	1.71 — 3.94	2.40 — 9.34	0.59 — 1.62	0.57 — 1.47
GSD Cp/Co	2.2	3.5	2.8	2.5
VG	4.7	54.8	2.7	2.2
log(VG) Confidence Interval	0.57 — 2.42	2.25 — 5.85	0.41 — 1.47	0.55 — 1.07
r	0.885	0.727	0.742	0.828
r Confidence Interval	0.78 — 0.97	0.55 — 0.92	0.58 — 0.93	0.69 — 0.95
% within a factor of 2	50.0%	16.7%	50.0%	58.3%

Table 3. Performance measures for the maximum nine-hour average concentration unpaired in time and space modeling objective.

8-km results	ISC2	AERMOD	CALPUFF	RATCHET
GM Cp/Co	1.2	1.5	0.79	0.75
GM Confidence Interval	0.70 — 2.02	0.94 — 2.52	0.59 — 1.03	0.55 — 1.00
GSD Cp/Co	2.9	3.0	1.9	1.8
VG	2.9	3.6	1.6	1.5
log(VG) Confidence Interval	0.78 — 1.39	0.69 — 1.88	0.08 — 0.82	0.13 — 0.68
r	0.56	0.73	0.86	0.90
r Confidence Interval	0.35 — 0.88	0.49 — 0.89	0.67 — 0.98	0.78 — 0.98
% within a factor of 2	33.3%	16.7%	91.7%	75.0%
16-km results	ISC2	AERMOD	CALPUFF	RATCHET
GM Cp/Co	1.4	1.8	0.90	0.85
GM Confidence Interval	0.96 — 1.99	0.94 — 3.14	0.59 — 1.30	0.56 — 1.23
GSD Cp/Co	2.7	3.1	2.4	2.6
VG	2.9	4.4	2.0	2.4
log(VG) Confidence Interval	0.41 — 1.76	0.86 — 2.17	0.27 — 1.31	0.40 — 1.44
r	0.80	0.76	0.87	0.84
r Confidence Interval	0.66 — 0.96	0.63 — 0.90	0.81 — 0.94	0.68 — 0.97
% within a factor of 2	58.3%	16.7%	66.7%	41.7%

Table 4. Significant differences in model performance measures for the maximum one-hour and nine-hour-average concentration unpaired in space and time modeling objective. An "X" indicates a significant difference.

Maximum one-hour	8-km data			16-km data		
Model	GM	VG	r	GM	VG	r
ISC2-AERMOD	X	X		X	X	X
ISC2-CALPUFF	X			X		
ISC2-RATCHET	X			X		
AERMOD-CALPUFF	X	X		X	X	
AERMOD-RATCHET	X	X		X	X	
CALPUFF-RATCHET						
Maximum nine-hour	8-km data			16-km data		
Model	GM	VG	r	GM	VG	r
ISC2-AERMOD						
ISC2-CALPUFF	X			X		
ISC2-RATCHET	X	X	X	X		
AERMOD-CALPUFF	X			X		
AERMOD-RATCHET	X	X	X	X		
CALPUFF-RATCHET						

Table 5. Performance measures for plume maximum location, plume width, and the arc integrated concentration at the 8-km distance.

Plume maximum location	ISC2	AERMOD	CALPUFF	RATCHET
Mean deviation (degrees)	26	14	24	14
Standard Error (degrees)	11	3.0	10	3.2
Minimum (degrees)	4.2	0.0	3.3	0.0
Maximum (degrees)	133	31	129	29
Plume width	ISC2	AERMOD	CALPUFF	RATCHET
FB	0.42	-0.43	0.14	-0.34
FB Confidence Interval	0.31 — 0.54	-0.52 — -0.35	-0.06 — 0.37	-0.51 — -0.15
NMSE	0.25	0.22	0.15	0.25
NMSE Confidence Interval	0.13 — 0.38	0.15 — 0.32	0.07 — 0.31	0.15 — 0.36
r	0.57	0.19	0.71	-0.22
r Confidence Interval	0.00 — 0.88	-0.24 — 0.50	0.44 — 0.88	-0.79 — 1.00
% within a factor of 2	83.3%	91.7%	75.0%	91.7%
Arc-integrated concentration	ISC2	AERMOD	CALPUFF	RATCHET
GM C_p/C_o	1.1	0.94	0.76	0.79
GM Confidence Interval	0.71 — 1.66	0.68 — 1.29	0.61 — 0.94	0.60 — 1.03
GSD C_p/C_o	2.4	2.3	2.0	1.8
VG	2.1	1.9	1.7	1.5
log(VG) Confidence Interval	0.52 — 0.93	0.29 — 0.94	0.07 — 0.98	0.05 — 0.70
r	0.81	0.86	0.86	0.89
r Confidence Interval	0.64 — 0.96	0.67 — 0.96	0.60 — 1.00	0.75 — 0.98
% within a factor of 2	50.0%	50.0%	91.7%	91.7%

Table 6. Performance measures for plume maximum location, plume width, and the arc integrated concentration at the 16-km distance.

Plume maximum location	ISC2	AERMOD	CALPUFF	RATCHET
Mean deviation (degrees)	34	36	26	23
Standard Error (degrees)	9.4	9.0	8.0	7.2
Minimum (degrees)	0.0	0.0	3.8	0.0
Maximum (degrees)	95	98	86	74
Plume width	ISC2	AERMOD	CALPUFF	RATCHET
FB	0.30	-0.57	0.081	-0.47
FB Confidence Interval	0.10 — 0.49	-0.71 — -0.44	-0.18 — 0.37	-0.66 — -0.27
NMSE	0.26	0.44	0.26	0.39
NMSE Confidence Interval	0.09 — 0.42	0.28 — 0.63	0.11 — 0.48	0.20 — 0.60
r	0.074	-0.022	0.49	0.065
r Confidence Interval	-0.36 — 0.51	-0.63 — 0.36	0.20 — 0.79	-0.57 — 0.54
% within a factor of 2	83.3%	50.0%	66.7%	58.3%
Arc-integrated concentration	ISC2	AERMOD	CALPUFF	RATCHET
GM C_p/C_o	1.4	1.2	0.88	1.1
GM Confidence Interval	1.11 — 1.87	0.77 — 1.70	0.66 — 1.16	0.77 — 1.60
GSD C_p/C_o	2.0	2.4	1.9	2.1
VG	1.8	2.1	1.5	1.6
log(VG) Confidence Interval	0.28 — 0.85	0.36 — 1.15	0.13 — 0.71	0.40 — 0.58
r	0.92	0.85	0.87	0.82
r Confidence Interval	0.88 — 0.97	0.75 — 0.96	0.70 — 0.98	0.76 — 0.89
% within a factor of 2	66.7%	50.0%	75.0%	83.3%

Table 7. Significant differences in model performance measures for the plume width and arc integrated concentration modeling objective. An "X" indicates a significant difference.

Plume width	8-km data			16-km data		
Model	FB	NMSE	r	FB	NMSE	r
ISC2-AERMOD	X			X		
ISC2-CALPUFF						
ISC2-RATCHET	X			X		
AERMOD-CALPUFF	X		X	X		
AERMOD-RATCHET						
CALPUFF-RATCHET	X		X	X		
Arc-integrated concentration	8-km data			16-km data		
Model	GM	VG	r	GM	VG	r
ISC2-AERMOD						
ISC2-CALPUFF				X		
ISC2-RATCHET	X					X
AERMOD-CALPUFF						
AERMOD-RATCHET						
CALPUFF-RATCHET						

Table 8. Performance measures for the unpaired nine-hour averaged concentration modeling objective.

8-km results	ISC2	AERMOD	CALPUFF	RATCHET
GM Cp/Co	0.93	0.82	0.89	1.3
GM confidence interval	0.90 — 1.00	0.77 — 0.87	0.84 — 0.92	1.22 — 1.31
GSD Cp/Co	1.9	2.8	2.7	1.8
GSD confidence interval	1.77 — 1.96	2.62 — 2.88	2.56 — 2.76	1.66 — 1.90
n	392	564	430	560
VG	1.5	2.9	2.5	1.5
log(VG) confidence interval	0.33 — 0.46	0.97 — 1.16	0.89 — 1.04	0.32 — 0.46
r	0.92	0.87	0.87	0.95
r confidence interval	0.91 — 0.94	0.85 — 0.89	0.85 — 0.89	0.94 — 0.96
% within a factor of 2	85%	56%	58%	81%
16-km results	ISC2	AERMOD	CALPUFF	RATCHET
GM Cp/Co	1.4	0.86	1.1	1.4
GM confidence interval	1.35 — 1.50	0.81 — 0.92	1.08 — 1.19	1.32 — 1.45
GSD Cp/Co	2.1	2.9	2.5	2.0
GSD confidence interval	1.93 — 2.17	2.75 — 3.04	2.43 — 2.62	1.94 — 2.08
n	316	410	299	399
VG	1.9	3.2	2.4	1.8
log(VG) confidence interval	0.57 — 0.71	1.04 — 1.25	0.80 — 0.94	0.54 — 0.63
r	0.88	0.82	0.82	0.90
r confidence interval	0.86 — 0.90	0.79 — 0.85	0.80 — 0.85	0.89 — 0.91
% within a factor of 2	51%	49%	62%	62%

Table 9. Significant differences in model performance measures for the unpaired nine-hour average concentration modeling objective. An "X" indicates a significant difference.

Model	8-km data			16-km data		
	GM	VG	r	GM	VG	r
ISC2-AERMOD	X	X	X	X	X	X
ISC2-CALPUFF	X	X	X	X	X	X
ISC2-RATCHET	X		X			
AERMOD-CALPUFF		X		X	X	
AERMOD-RATCHET	X	X	X	X	X	X
CALPUFF-RATCHET	X	X	X	X	X	X

Table 10. Performance measures for the daytime, transition, and nighttime period ensemble means modeling objective at the 8-km distance.

Daytime tests	ISC2	AERMOD	CALPUFF	RATCHET
GM Cp/Co	1.0	0.43	0.61	1.05
GM confidence interval	0.71 — 1.46	0.31 — 0.59	0.49 — 0.76	0.87 — 1.30
GSD Cp/Co	4.7	4.3	2.6	2.5
GSD confidence interval	3.87 — 5.60	3.25 — 5.36	2.29 — 2.92	2.16 — 2.87
VG	11	16	3.2	2.3
log(VG) confidence interval	1.83 — 2.97	2.12 — 3.45	0.81 — 1.52	0.59 — 1.11
r	0.45	0.46	0.74	0.63
r confidence interval	0.26 — 0.58	0.28 — 0.62	0.62 — 0.82	0.46 — 0.76
% within a factor of 2	32%	38%	61%	69%
Transition period tests	ISC2	AERMOD	CALPUFF	RATCHET
GM Cp/Co	0.3	0.31	0.19	0.56
GM confidence interval	0.15 — 0.48	0.19 — 0.51	0.15 — 0.25	0.36 — 0.90
GSD Cp/Co	12.1	8.3	3.4	7.4
GSD confidence interval	8.06 — 19.6	5.30 — 12.5	2.70 — 4.07	4.44 — 11.0
VG	2024	326	68.2	71.2
log(VG) confidence interval	5.19 — 11.50	3.49 — 8.31	3.32 — 5.17	2.31 — 6.56
r	0.36	0.41	0.82	0.43
r confidence interval	0.01 — 0.51	0.17 — 0.61	0.73 — 0.88	0.18 — 0.66
% within a factor of 2	35%	38%	18%	40%
Nighttime period tests	ISC2	AERMOD	CALPUFF	RATCHET
GM Cp/Co	0.5	0.84	0.56	1.0
GM confidence interval	0.36 — 0.71	0.64 — 1.10	0.43 — 0.75	0.86 — 1.25
GSD Cp/Co	4.3	3.4	3.3	2.4
GSD confidence interval	3.45 — 5.47	2.89 — 3.95	2.67 — 3.84	2.00 — 2.72
VG	12.9	4.6	5.6	2.1
log(VG) confidence interval	1.81 — 3.46	1.13 — 1.96	1.08 — 2.41	0.48 — 1.00
r	0.79	0.86	0.92	0.87
r confidence interval	0.68 — 0.86	0.79 — 0.90	0.88 — 0.95	0.82 — 0.91
% within a factor of 2	53%	47%	68%	72%

Table 11. Performance measures for the daytime, transition, and nighttime period ensemble means modeling objective at the 16-km distance.

Daytime tests	ISC2	AERMOD	CALPUFF	RATCHET
GM Cp/Co	1.6	0.72	0.94	1.10
GM confidence interval	1.18 — 2.27	0.54 — 0.96	0.75 — 1.19	0.90 — 1.41
GSD Cp/Co	3.7	3.4	2.6	2.7
GSD confidence interval	3.11 — 4.54	2.60 — 4.39	2.15 — 3.21	2.21 — 3.24
VG	6.81	5.04	2.54	2.73
log(VG) confidence interval	1.44 — 2.68	1.09 — 2.23	0.60 — 1.35	0.63 — 1.42
r	0.41	0.27	0.55	0.57
r confidence interval	0.11 — 0.57	0.02 — 0.53	0.33 — 0.72	0.34 — 0.74
% within a factor of 2	40%	56%	69%	72%
Transition period tests	ISC2	AERMOD	CALPUFF	RATCHET
GM Cp/Co	0.4	0.29	0.20	0.61
GM confidence interval	0.24 — 0.59	0.18 — 0.47	0.15 — 0.28	0.39 — 0.94
GSD Cp/Co	6.3	6.8	3.8	6.2
GSD confidence interval	4.67 — 8.31	4.91 — 9.00	2.99 — 4.59	4.70 — 7.72
VG	72.31	169.10	69.42	33.82
log(VG) confidence interval	3.02 — 5.85	3.79 — 6.46	3.41 — 5.12	2.76 — 4.38
r	0.40	0.22	0.60	0.32
r confidence interval	0.11 — 0.60	-0.13 — 0.49	0.35 — 0.76	0.07 — 0.54
% within a factor of 2	37%	30%	30%	27%
Nighttime period tests	ISC2	AERMOD	CALPUFF	RATCHET
GM Cp/Co	1.2	1.2	0.90	1.1
GM confidence interval	0.89 — 1.72	0.86 — 1.51	0.73 — 1.09	0.93 — 1.33
GSD Cp/Co	3.8	3.5	2.5	2.2
GSD confidence interval	2.82 — 5.07	2.84 — 4.08	2.11 — 2.83	1.80 — 2.62
VG	6.0	4.6	2.3	1.8
log(VG) confidence interval	1.07 — 2.79	1.09 — 2.02	0.57 — 1.10	0.35 — 0.95
r	0.74	0.82	0.91	0.92
r confidence interval	0.60 — 0.85	0.73 — 0.89	0.87 — 0.94	0.87 — 0.95
% within a factor of 2	54%	53%	68%	76%

Table 12. Significant differences in model performance measures for the day, transition, and nighttime ensemble means modeling objective. An "X" indicates a significant difference.

Daytime	8-km data			16-km data		
Model	GM	VG	r	GM	VG	r
ISC2-AERMOD	X			X		
ISC2-CALPUFF	X	X	X	X	X	
ISC2-RATCHET		X			X	
AERMOD-CALPUFF		X	X		X	
AERMOD-RATCHET	X	X		X		
CALPUFF-RATCHET	X			X		
Transition	8-km data			16-km data		
Model	GM	VG	r	GM	VG	r
ISC2-AERMOD						
ISC2-CALPUFF		X	X	X		
ISC2-RATCHET		X				
AERMOD-CALPUFF			X			X
AERMOD-RATCHET				X		
CALPUFF-RATCHET	X		X	X		
Nighttime	8-km data			16-km data		
Model	GM	VG	r	GM	VG	r
ISC2-AERMOD	X	X				
ISC2-CALPUFF			X		X	X
ISC2-RATCHET	X	X			X	X
AERMOD-CALPUFF			X		X	X
AERMOD-RATCHET		X			X	X
CALPUFF-RATCHET	X	X				