Atmospheric Environment 89 (2014) 707–720

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Performance evaluation of AERMOD, CALPUFF, and legacy air dispersion models using the Winter Validation Tracer Study dataset

Arthur S. Rood

K-Spar Inc., 4835 W. Foxtrail Lane, Idaho Falls, ID 83402, USA

HIGHLIGHTS

- 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 unpaired 1- and 9-h average concentrations performance objectives.
- Puff models showed the lower bias and variance and higher correlation.
- Steady-state models were less likely to underpredict concentrations.

ARTICLE INFO

Article history:
Received 28 August 2013
Received in revised form
20 February 2014
Accepted 24 February 2014
Available online 25 February 2014

Keywords:
Air dispersion models
CALPUFF
AERMOD
Validation
Tracer

ABSTRACT

The performance of the steady-state air dispersion models AERMOD and Industrial Source Complex 2 (ISC2), and Lagrangian puff models CALPUFF and RATCHET were evaluated using the Winter Validation Tracer Study dataset. The Winter Validation Tracer Study was performed in February 1991 at the former Rocky Flats Environmental Technology Site near Denver, Colorado. Twelve, 11-h tests were conducted where a conservative tracer was released and measured hourly at 140 samplers in concentric rings 8 km and 16 km from the release point. Performance objectives were unpaired maximum one- and nine-hour average concentration, location of plume maximum, plume impact area, arc-integrated concentration, unpaired nine-hour average concentration, and paired ensemble means. Performance objectives were aimed at addressing regulatory compliance, and dose reconstruction assessment questions. The objective of regulatory compliance is not to underestimate maximum concentrations whereas for dose reconstruction, the objective is an unbiased estimate of concentration in space and time. Performance measures included the fractional bias, normalized mean square error, geometric mean, geometric mean variance, correlation coefficient, and fraction of observations within a factor of two. The Lagrangian puff models tended to exhibit the smallest variance, highest correlation, and highest number of predictions within a factor of two compared to the steady-state models at both the 8-km and 16-km distance. Maximum one- and nine-hour average concentrations were less likely to be under-predicted by the steady-state models compared to the Lagrangian puff models. The characteristic of the steady-state models not to under-predict maximum concentrations make them well suited for regulatory compliance demonstration, whereas the Lagrangian puff models are better suited for dose reconstruction and long range transport.

© 2014 The Author. Published by Elsevier Ltd. Open access under CC BY-NC-ND license.

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.

E-mail address: asr@kspar.com.

http://dx.doi.org/10.1016/j.atmosenv.2014.02.054
1352-2310 © 2014 The Author. Published by Elsevier Ltd. Open access under CC BY-NC-ND license.

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 legacy models, Industrial Source Complex Short Term Version 2 (ISC2) (EPA, 1992) and Regional Atmospheric Transport Code for Hanford Emission Tracking (RATCHET) (Ramsdell et al., 1994), were included in the evaluation because the formulations of these models are currently used in radiological assessment codes (EPA, 2007; Chanin et al., 1998; Ramsdell et al., 2010). Model simulations and performance evaluation of ISC2 using the WVTS was originally reported in Haugen and Fontino (1993). Performance evaluation of RATCHET using the WVTS dataset was originally reported in Rood (1999) and Rood et al. (1999). Model-predicted concentrations for ISC2 and RATCHET were taken from Haugen and Fontino (1993) and Rood (1999), respectively, and were used without modification.

A description of the tracer measurements and meteorological data is provided first, followed by modeling protocol, performance objectives, and performance measures. Finally, the model performance results, in terms of addressing the prospective and retrospective assessment questions, are discussed.

2. Methods

Model performance was evaluated in terms of fundamental plume properties, paired ensemble mean concentrations, and concentrations unpaired in space. The WVTS dataset and meteorological data are presented first, followed by modeling protocol, performance objectives, and performance measures.

2.1. Winter Validation Tracer Study

The WVTS was conducted in February 1991 near the former RFETS located on the Front Range of the Colorado Rocky Mountains about 25 km northwest of Denver (Brown, 1991). The study consisted of 12 separate tests (Table 1). For each test, an inert tracer (sulfur hexafluoride [SF_6]) was emitted continuously for 11 h from a 10-m high stack located on the east side of the main plant complex (Fig. 1). The main plant complex was located about 2.5 km east of the foothills on an alluvial plain ranging in elevation from 1750 m to 1850 m above sea level. The primary purpose of the study was to gather data for validation of emergency response atmospheric transport models. Samplers were arranged in concentric circles 8-km and 16-km from the release point so as to capture any possible transport trajectories. One-hour average air concentrations were then measured for the last nine hours of the release at each of the 140 samplers. Six tests were performed under nighttime conditions, four under daytime conditions, one under day–night transition, and one under night–day transition. A total of 108 h of data were recorded. Seventy-two samplers were distributed at the 8-km distance and 68 samplers at the 16-km distance. Sampler elevations ranged from about 1600 m to 2600 m above sea level. The study domain is considered near field because the maximum distance to the samplers is <50 km.

Previous investigators (Haugen and Fontino, 1993) used this data set in a performance evaluation of the TRAC (Hodgin, 1991) and ISC2 models. The electronic copy of this data set was obtained from Haugen and Fontino (1993) for use in the model performance evaluation for the Historical Public Exposures Studies at Rocky Flats (Rood, 1999). These data included the observed hourly-average concentrations for all 12 tests, the sampler ID numbers and locations, and the TRAC and ISC2 predicted concentrations. The ISC2 results provided by Haugen were used in this paper without modification.

2.2. Meteorological data

Meteorological data were recorded for every tracer test at the 10-m and 61-m level from the RFETS 61-m tower located 790 m west and 87 m south of the release point. Only data from the 10-m level were used in the model simulations. Data were provided as 15-min averages of wind speed and direction, temperature, heat flux, and standard deviations of these parameters. Hourly averages of these data were calculated using EPA protocol (EPA, 2000).

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 h^{-1})	CWL[c] (kg h^{-1})	Average (kg h^{-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).

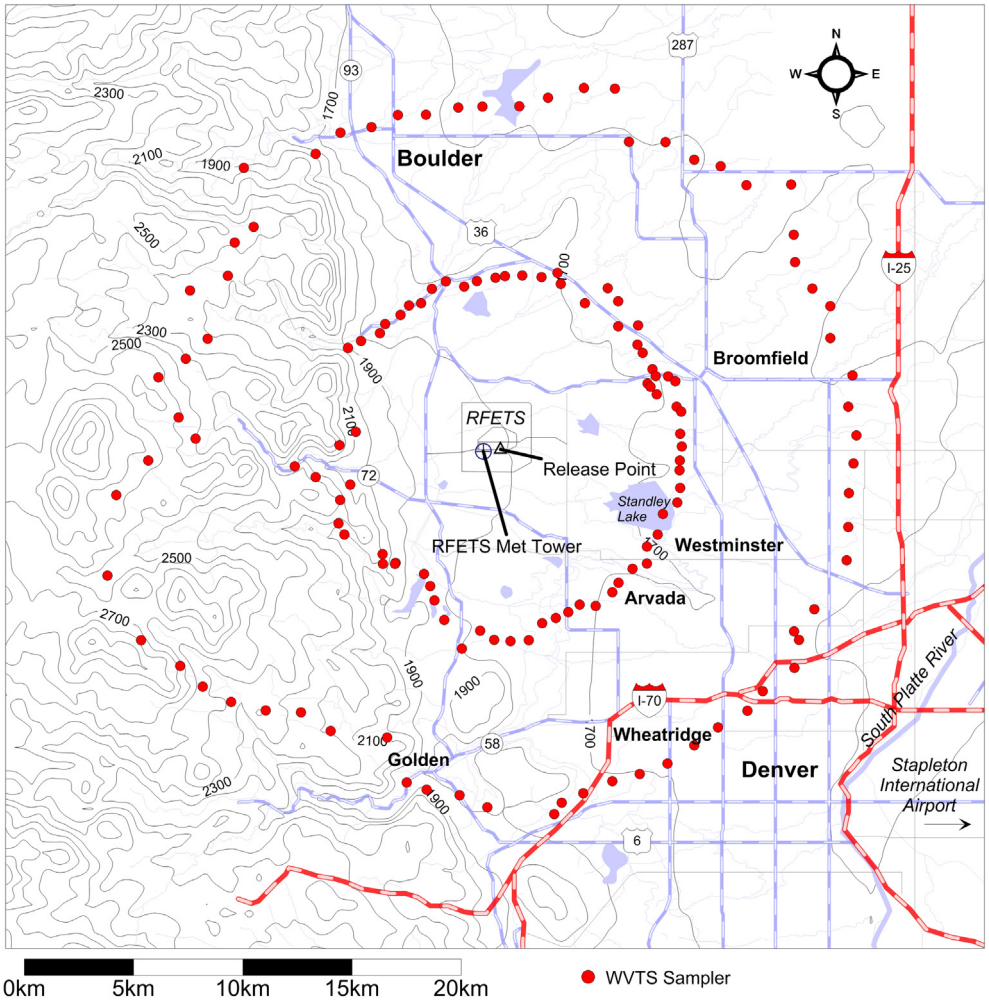

Fig. 1. CALPUFF model domain showing location of the WVTS samplers and terrain features of the Colorado Front Range.

Mixing depth estimates for the ISC2 simulations were derived from linear interpolation for each 15-min period from the rawinsonde data taken every 12 h at Denver Stapleton airport. No precipitation was measured during any of the 12 tests, and there was no snow cover during February 1991.

Additional surface and upper air meteorological data were obtained from Denver Stapleton airport located about 25 km southeast of the RFETS. Surface data included wind speed and direction, cloud cover, and ceiling height and were used with the RFETS data to calculate stability class. The RATCHET and CALPUFF models utilized the wind speed and direction measurements in their wind field interpolation.

Surface observations from Denver Stapleton are strongly influenced by air movement within the Platte River Valley, which flows to the northeast from the city center. By contrast, Rocky Flats is more strongly influenced by its proximity to the foothills. Both locations are influenced by the diurnal pattern of upslope-downslope conditions that characterize the general air movement on the Colorado Front Range environs. Downslope conditions typically occur during the evening hours and are characterized by drainage flow of cooler surface air from the foothills and upper reaches of the Platte River Valley northeastward to the plains. Airflow at Rocky Flats is typically from the west-northwest, and converges with the flow from the south within the Platte River Valley in a broad zone 20 km–30 km east–northeast of the RFETS (Lange, 1992). During daylight hours and after surface heating has eliminated the cooler surface layer, the downslope conditions cease. This is followed by a brief period of relatively calm winds, which in turn is followed by return of air up the valley or upslope conditions. Upslope conditions were weak to non-existent during the WVTS.

2.3. Atmospheric transport models and protocol

A brief description of each of the models included in this study is presented along with the modeling protocol. Because the SF_6 tracer is an inert gas, all model simulations did not include deposition and plume depletion.

2.3.1. AERMOD

The American Meteorological Society and EPA developed the AERMOD modeling system (Cimorelli et al. 2004). Model development began in 1991 with the objective to incorporate current planetary boundary layer concepts into regulatory compliance models. Treatment of both surface and elevated point sources, area sources, and volume sources in a simple or complex terrain model domain are addressed in the model. It was intended as a replacement of the Industrial Source Complex Version 3 (ISC3) model. Currently, AERMOD is the EPA's preferred model for regulatory compliance demonstration for criteria pollutants in the near field (<50 km).

2.3.1.1. AERMOD modeling protocol. The modeling domain consisted of a 46.2-km × 49.8-km region centered on the RFETS. Meteorological data were the same used in the CALPUFF simulation and included surface and upper-air data from Denver Stapleton airport and onsite data from RFETS.

AERMOD and AERMET version 12345 were used in the model simulations. Special processing of AERMOD output was required because minimum model-simulation times are one-day, and nine-hour averaging times are not an option. Nine-hour average concentrations were calculated by inputting the tracer release rate for the last nine hours of the 11-h tracer test and setting the release rate to zero for the remaining hours in the simulation. The maximum one-hour, and 12-h average concentration at each of the samplers for a simulation period that encompassed each test was output. The nine-hour average concentration was calculated by multiplying the 12-h average concentration by the ratio of 12/9.

2.3.2. CALPUFF

The CALPUFF modeling system (Scire et al., 2002) is a non-steady-state Lagrangian puff model that simulates pollutant transport, transformation, and deposition in a three-dimensional spatially and temporally variable wind field. CALPUFF can be applied on local and regional scales. The modeling system is composed of three primary modules: CALMET, CALPUFF, and CALPOST and collectively these are referred to as the CALPUFF modeling system.

The CALMET module is a meteorological model that generates a three-dimensional hourly wind field within a three-dimensional gridded modeling domain. The CALPUFF module uses the CALMET-generated wind field and micrometeorological parameters to advect and disperse "puffs". The CALPOST module reads the CALPUFF concentration and deposition flux files and produces time-averaged concentration and deposition output along with visibility impacts.

CALPUFF is currently the EPA preferred long-range (>50 km) dispersion model for demonstration of compliance with Prevention of Significant Deterioration increment levels and National Ambient Air Quality Standards in Class I areas.

2.3.2.1. CALPUFF modeling protocol. A 45.2-km × 43.6-km model domain having a grid spacing of 400 m (113 × 109 grid cells) was established (Fig. 1). It was centered approximately on the WVTS release point and included the Platte River Valley in the southeast corner. Vertical discretization consisted of eight layers 20 m, 40 m, 60 m, 100 m, 200 m, 400 m, 800 m, 1200 m, and 1800 m above land surface.

For the CALMET simulations, EPA-Federal-Land-Manager-recommended parameter values (Fox, 2009) or CALMET default values were generally used where applicable. The model parameters BIAS, RMAX1, RMAX2, TERRAD, R1, and R2 were chosen on a site-specific basis. The BIAS parameter assigns weights to the surface and upper-air stations data for each vertical layer. Surface data were given 100% of the weight (BIAS = 1) in the first layer with zero weight in the last two vertical layers (BIAS = 1). Equal weight was assigned to the fourth layer (BIAS = 0) with a gradation of weights between the lower and upper for the remaining layers.

The RMAX1 and RMAX2 parameters define the maximum radius of influence for surface and upper data, respectively, over land surfaces. To incorporate the influence of flow in the Platte River Valley as represented by the surface data at Denver Stapleton airport, a value of 32 km for RMAX1 was used. The RMAX2 value serves the same purpose as RMAX1 but is used for upper-air data. A value of 100 km was selected for the RMAX2 parameter.

The TERRAD parameter defines the radius of influence of terrain features. A TERRAD value of 7 km was used so that the terrain influence of the foothills encompassed RFETS. The parameters R1 and R2 are the distance from an observation where the observation and the initial guess field are equally weighted for surface and layers aloft, respectively. A value of 17 km and 56 km were chosen for R1 and R2, respectively.

The CALPUFF runs were performed using dispersion coefficients calculated from micrometeorological variables (MDISP = 2), and the simple CALPUFF-type terrain adjustment algorithm (MCTADJ = 2). The remaining parameters were CALPUFF defaults. CALPUFF Version 5.8, Level 070623, and CALMET Version 5.8, Level 060811, were used in the model simulations.

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

The ISC2 model is an augmented steady-state Gaussian plume model primarily used to demonstrate regulatory compliance with criteria pollutants emitted by industrial facilities. It was replaced by ISC Version 3 and was the EPA preferred model until the promulgation of AERMOD in December of 2005. The model was included in the evaluation because straight-line Gaussian plume models form the basis of the CAP88 (EPA, 2007) model for demonstration of compliance with the Clean Air Act, and of the MACCS2 code (Chanin et al., 1998) for reactor accident consequence analysis.

2.3.3.1. ISC2 modeling protocol. The ISC2 simulations were performed by Haugen and Fontino (1993) using 15-min meteorological data from the 10-m level taken at the RFETS and mixing depth estimated from the rawinsonde data at Denver Stapleton airport. Four ISC2 runs were performed for each hour of simulation using the four, 15-min average data from the Rocky Flats meteorological station. The results from the four simulations were averaged to provide hourly-average concentrations at each of the sampler locations. These hourly concentrations were then averaged across each test by Rood (1999) to provide nine-hour average concentrations at each sampler.

The performance evaluation of this model was originally reported in Rood (1999) and Rood et al. (1999). The results presented here are based on the original data using slightly different performance 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 of the maximum one-hour average concentration unpaired in space and time was also included to provide insight into model performance for short-term maximum concentrations. Descriptions of each modeling objective follows.

2.4.1. Maximum hourly and nine-hour average concentration and plume maximum location

This modeling objective compared the predicted and observed maximum one-hour and nine-hour average concentration measured at a sampler during the nine-hour test period at either the 8-km or 16-km distance from the release point. The predicted maximum concentration was not paired in space, and also unpaired in time for the maximum-hourly average concentration. The nine-hour average concentration was determined by a simple arithmetic average of the nine, one-hour average concentrations. Sampler data that were missing were not included when computing the predicted or observed average concentration.

The plume maximum location was only computed for the nine-hour average concentration and was quantified in terms of the absolute value of angular difference between the predicted and observed location of the plume maximum.

2.4.2. Plume width

The plume width objective evaluated the predicted impact area of the plume. Each sampler was assigned an arc length equal to the arc length between the midpoints of the sampler and each of its adjacent samplers. The plume width was sum of the arc lengths of samplers that had a concentration greater than zero, or in the case of the observed values, a concentration greater than the minimum detectable concentration.

2.4.3. Arc-integrated concentration

The arc-integrated concentration evaluated the plume mass at the 8-km and 16-km distance. The arc-integrated concentration is the sum of the product of the sampler arc lengths as defined in Section 2.4.2 and the nine-hour average predicted or observed concentration.

2.4.4. Unpaired time-averaged concentration

This modeling objective compared the ranked predicted and observed time-averaged (nine-hour) concentrations. Only predicted and observed concentrations that met the selection criteria stated in Section 2.6 were included. Samples were blocked into those performed at night (Tests 1, 2, 3, 4, 10, and 11), those performed during the day (Tests 6, 7, 9, and 12), and those performed during transition periods (Tests 5 and 8). Sample blocking is used in bootstrap resampling to avoid block-to-block variance (Chang and Hanna, 2005).

2.4.5. Paired ensemble means (ASTM procedure)

The American Society for Testing and Materials (ASTM) proposed a procedure for evaluation of models, recognizing that model predictions are ensemble-mean predictions, while observations correspond to realizations of ensembles (ASTM, 2000). An ensemble is defined as a set of experiments having fixed external conditions, such as meteorological conditions and downwind distance. In the WVTS, fixed external conditions were the distance to the sampling arc and meteorological conditions. In general, repeatable diurnal flow and stability regimes are established during nighttime, daytime, and day–night transitional periods along the Colorado Front Range. Thus, averages across the tests representing these similar conditions would approximate ensemble means to compare with model predictions for the same period.

Predictions and observations were grouped into three blocks consisting of nighttime (six tests), daytime (four tests), and transition period (two tests). Average concentrations were calculated across all tests in the block for each sampler, and performance statistics were calculated separately for each block.

2.5. Performance measures

Several simplified measures were used to evaluate model performance (Cox and Tikvart, 1990; Weil et al., 1992). These measures were the fractional bias (*FB*) and normalized mean square error (*NMSE*). Fractional bias is given by

$$FB = \frac{2\left(\overline{C}_o - \overline{C}_p\right)}{\overline{C}_o + \overline{C}_p} \tag{1}$$

where C_p and C_o are the predicted and observed concentrations, respectively. Overbars indicate averages over the sample. The *NMSE* given by

$$NMSE = \frac{\overline{\left(C_o - C_p\right)^2}}{\overline{C}_o \overline{C}_p} \tag{2}$$

The *FB* is a measure of mean bias. A *FB* of 0.6 is equivalent to model under-prediction by about a factor of two. A negative value indicates model over-prediction. The *NMSE* is a measure of variance, and a value of 1.0 indicates that a typical difference between predictions and observations is approximately equal to the mean. The *NMSE* and *FB* are appropriate when the typical difference between the predictions and observations are approximately a factor of two (Hanna et al., 1991) and the range of predictions and observations in the dataset is small (i.e., less than a factor of two). This was not the case in this study where ratios of model predictions to observations often ranged from 0.01 to 100, and within a data set, the predicted and observed concentrations ranged from the zero to ~10,000 ng m^{-3}. In these cases a log-transformed measure of model bias and variance is more appropriate because it provides a more balanced approach (Hanna et al., 1991). The log-transformed measures described in Hanna et al. (1991) are the geometric mean bias (*MG*) and the geometric mean variance (*VG*) and are defined by

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

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

Geometric mean bias values of 0.5 and 2.0 indicate a factor of two over-prediction and under-prediction, respectively. A *VG* value

of 1.6 indicates about a factor of two difference between predicted and observed data pairs.

A more easily understood log-transformed quantity that is related to the *MG* and *VG* is the geometric mean (*GM*) and geometric standard deviation (*GSD*) of the predicted-to-observed ratio (C_p/C_o). The *GM* and *GSD* are given by

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

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

where n = the sample size. Because the *MG* is simply the inverse of the *GM*, only the *GM* is reported. A perfect model would have *FB* and *NMSE* values of 0, and *GM, GSD,* and *VG* values of 1.0. With the exception of the plume width and location of plume maximum performance objectives, the log-transformed performance measures are considered more appropriate than the *FB* and *NMSE*, and thus only the log-transformed measures are reported. The location of plume maximum does not lend itself to the above performance measures, mainly because the objective considered the absolute angular difference between the predicted and observed location of maximum. For this performance objective, the mean difference, standard deviation of the mean (i.e., standard error), and the minimum and maximum differences are reported. Because differences between predicted and observed values and the range of predictions and observations were less than about a factor of two, the *FB* and *NMSE* were considered more appropriate for the plume width objective.

In addition to the above measures, the correlation coefficient (r) between predicted and observed values and the number of predictions within a factor of two of the observations were also reported. The correlation coefficient was determined using least-squares linear regression and log-transformed data except for the plume width performance objective. Scatter plots were also included as visual measures of performance for the paired ensemble means and unpaired time-averaged concentration modeling objectives.

Confidence intervals were estimated for each of the performance measures using the bootstrap methodology described in BOOT software (Hanna et al., 1991; Chang and Hanna, 2005). Confidence intervals were used to determine if the estimated performance measure was significantly different than its optimum value and whether a statistically significant difference existed between the performance measures for each model. Confidence interval estimates were based on the cumulative density function generated from 1000 bootstrap samples.

2.6. Selection criteria

The observed data set only reported nonzero hourly average concentrations greater than the minimum detectable sampler concentration (*mdc*) of 33 ng m^{-3}. Measured concentrations below this value were reported as zero. A sampler that had only one hour of data (in the nine-hour measurement period) greater than the *mdc* would have a nine-hour average concentration of 3.7 ng m^{-3} (33 ng m^{-3}/9). This value represents the nine-hour time-averaged *mdc* for a sampler.

For the paired ensemble means performance objective, the dataset was based on the union of the predicted and observed concentrations. The *mdc* was substituted for predicted concentrations that were less than the *mdc* if the paired observed concentration was greater than zero. Likewise, the *mdc* was substituted for observed concentrations less than the *mdc* if the paired predicted concentration was greater than zero. Predicted and observed pairs that were both zero were omitted from the analysis.

For the unpaired analysis only predicted *and* observed concentration pairs greater than the *mdc* were considered. Samplers missing all nine hours of data were eliminated from the data set.

3. Results

The paired ensemble means and unpaired scatter plots are perhaps the most illustrative in terms of summarizing model performance qualitatively (Figs. 2–5). In general, the highest predicted and observed concentrations were during nighttime and transition period tests and the lowest during daytime tests.

The paired ensemble mean scatter plot at the 8-km distance (Fig. 2) showed the transition period tests as having the highest observed concentrations, some exceeding 8000 ng m^{-3}. Nighttime tests had maximum observed concentrations between 3000 and 4000 ng m^{-3}. As expected, daytime tests had the lowest observed maximum concentrations, the maximum being slightly less than 600 ng m^{-3}. All models performed poorly for the transition period tests, underestimating observed concentrations that were >1000 ng m^{-3}. In general, the puff models exhibited better correlation to the observations for daytime and nighttime tests and concentrations that were >100 ng m^{-3} compared to the steady-state models.

At the 16-km distance (Fig. 3) the nighttime period tests had the highest observed concentrations (~3500 ng m^{-3}), followed by transition period tests (~1800 ng m^{-3}). Daytime tests had maximum observed concentrations that were ~70 ng m^{-3}. The puff models exhibited better correlation, less variability, and a greater number of points within a factor of two of the observations compared to the steady state models for nighttime tests and concentrations >100 ng m^{-3}.

Scatter plots of the unpaired data at the 8-km distance (Fig. 4) showed that all models underestimated transition period observed concentrations that were greater than 1000 ng m^{-3}. Predicted concentrations from RATCHET were within a factor of two of the observations for almost all the daytime tests and most of the nighttime tests for the entire concentration range. Most of the ISC2 concentrations for daytime and nighttime tests were within a factor of two of the observations for concentrations that were >100 ng m^{-3}.

At the 16-km distance (Fig. 5), scatter plots of the unpaired data were similar to those at the 8-km distance, although CALPUFF underestimated almost all the concentrations for transition period tests by more than a factor of two. The three highest observed concentrations were within a factor of two of the corresponding AERMOD predicted concentrations. A similar result was found for ISC2, except the highest observed nighttime concentration was underestimated by more than a factor of two. Observed nighttime concentrations that were <100 ng m^{-3} were overestimated by more than a factor of two by RATCHET.

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

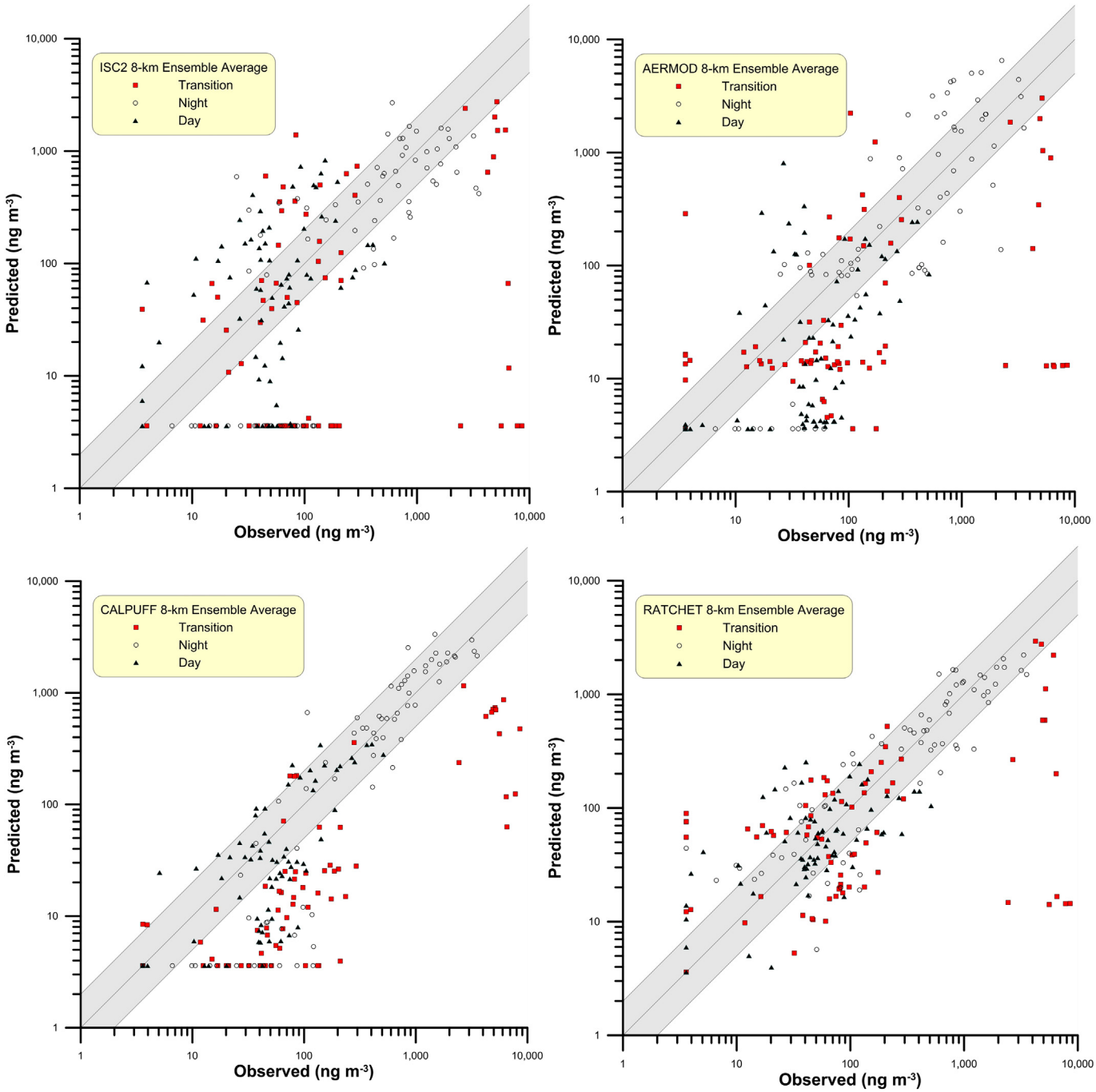

Fig. 2. Scatter plots of paired nine-hour average ensemble mean concentrations at the 8-km distance. Points that lie within the shaded region are within a factor of 2 of the observations.

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°) and greatest for CALPUFF and ISC2 (26 and 24° respectively). Based on a *t*-test difference of the means, the

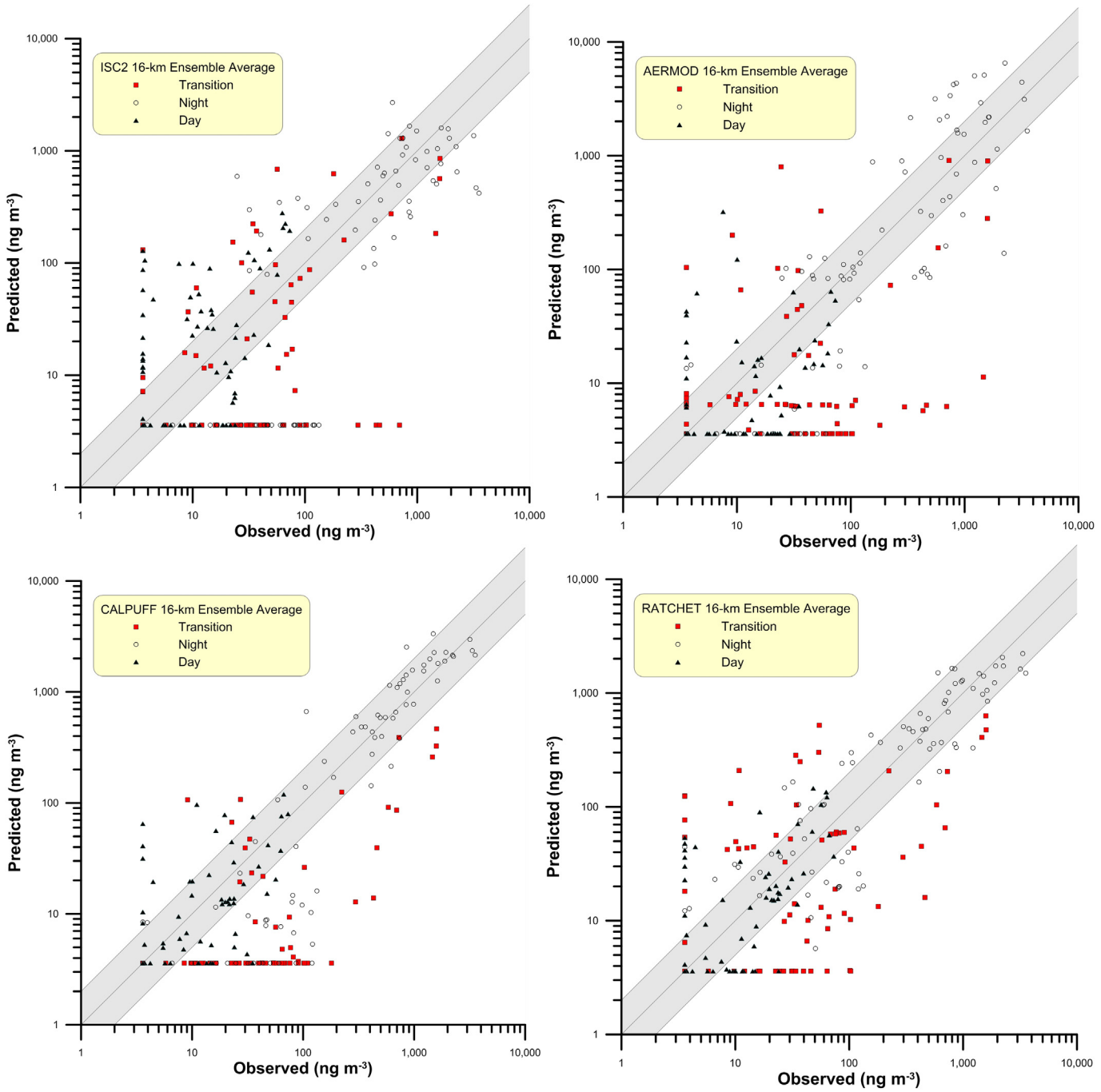

Fig. 3. Scatter plots of paired nine-hour average ensemble mean concentrations at the 16-km distance. Points that lie within the shaded region are within a factor of 2 of the observations.

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.

Arc-integrated concentration at the 8-km distance showed little bias for the steady state models (GM confidence interval included 1.0) and a negative bias for the puff models. However, only 50% of the predictions were within a factor of two for the steady-state models while over 90% of the predictions were within a factor of two for the puff models.

Plume maximum location at the 16-km distance (Table 6) showed that the mean deviation was smallest for CALPUFF and RATCHET (26 and 24 degrees respectively) and greatest for ISC2 and AERMOD (34 and 36 degrees respectively). Based on a *t*-test difference of the means, the differences between the steady state and puff models were significant at the 99% level ($0.01 < P < 0.005$).

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

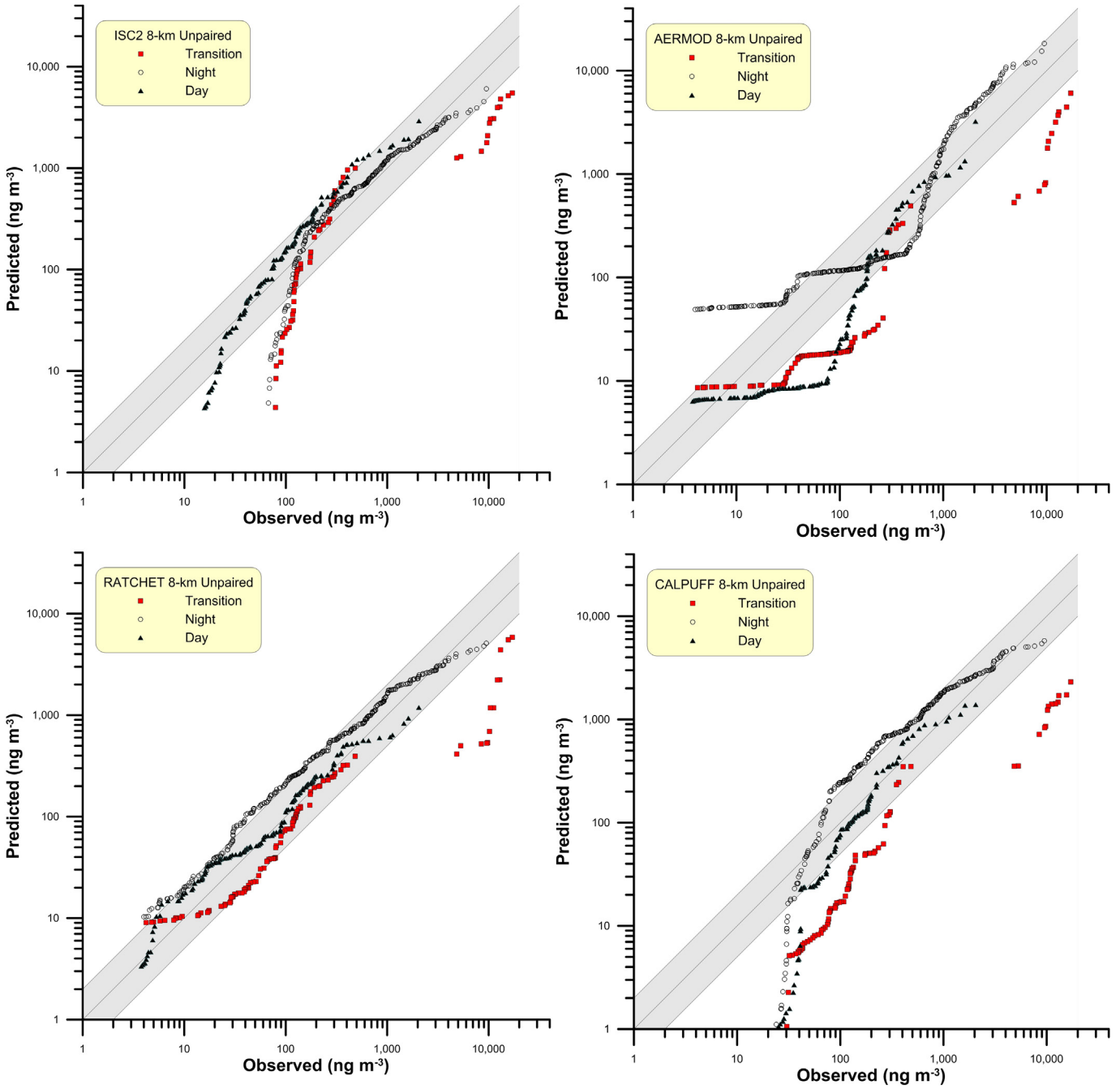

Fig. 4. Scatter plots of unpaired nine-hour average concentrations at the 8-km distance. Points that lie within the shaded region are within a factor of 2 of the observations.

RATCHET and AERMOD overestimated plume width. The CALPUFF *FB* was not significantly different from zero.

Arc-integrated concentration at the 16-km distance showed positive bias for ISC2. The GM confidence interval for the other models included the optimum value of 1.0. Puff models showed a greater percentage of predictions within a factor of two of the observations.

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

3.3. Unpaired nine-hour average concentration

The performance measure results at the 8-km distance for the unpaired nine-hour average concentration (Table 8) indicated a negative bias in predicted concentrations for AERMOD and CALPUFF, slight negative bias for ISC2, and positive bias for RATCHET. ISC2 and RATCHET also had the smallest variance, highest correlation coefficient, and highest percentage of predictions within a factor of two of the observations.

At the 16-km distance, ISC2 and RATCHET exhibited positive bias, AERMOD exhibited negative bias, and CALPUFF nearly no bias. RATCHET and ISC2 had the highest correlation coefficients and CALPUFF and RATCHET had the highest percentage of predictions within a factor of two of the observations.

Except for AERMOD and CALPUFF at the 8-km distance, and ISC2 and RATCHET at the 16-km distance, all bias performance measures among the models were significantly different from one another (Table 9).

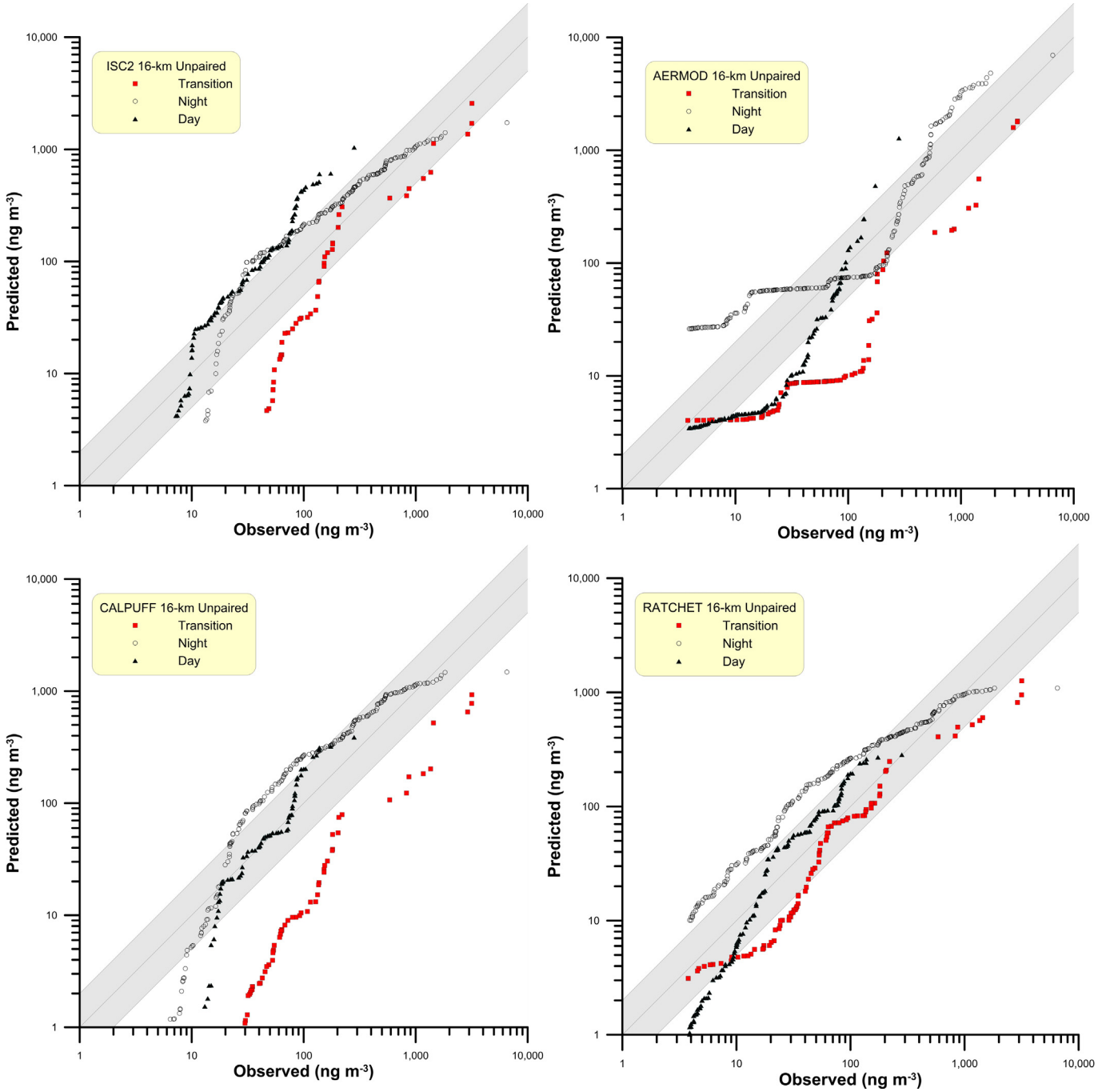

Fig. 5. Scatter plots of unpaired nine-hour average concentrations at the 16-km distance. Points that lie within the shaded region are within a factor of 2 of the observations.

3.4. Paired ensemble means

No one model showed overall better performance across all ensemble groups and all models performed poorly for the transition period ensemble means (Tables 10 and 11). However, excluding the transition ensemble means, RATCHET and CALPUFF had the highest percentage of predictions within a factor of two of the observations, the highest correlation coefficients, and generally the lowest variance compared to the steady-state models. For daytime and nighttime ensemble means, the GM confidence interval for RATCHET encompassed 1.0 at both the 8- and 16-km distances. AERMOD was biased low for daytime tests and showed little bias for nighttime tests (GM confidence interval encompassed 1.0). ISC2 exhibited no bias for daytime tests (GM = 1.0) at the 8-km distance but was biased high at the 16 km distance (GM confidence interval excluded 1.0). For nighttime tests, ISC2 was biased low at the 8-km distance but exhibited nearly zero bias at the 16 km distance. CALPUFF was biased low for daytime and nighttime tests at the 8-km distance (GM confidence interval excluded 1.0), but the GM confidence interval encompassed 1.0 at the 16-km distance.

All models were biased low for the transition period ensemble means at both the 8 and 16-km distances, and exhibited large variances, although the variance measures for CALPUFF and RATCHET were considerable smaller than those for ISC2 and AERMOD. Puff model *VG* and *r* values were significantly different than steady-state models for nighttime tests at the 16-km distance (Table 12).

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

	ISC2	AERMOD	CALPUFF	RATCHET
8-km results				
GM C_p/C_o	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 C_p/C_o	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				
GM C_p/C_o	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 C_p/C_o	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%

4. Discussion

Model performance is judged in terms of the assessment question that the model is intended to address. As stated in the introduction, the assessment questions are different for a prospective regulatory compliance calculation compared to a retrospective dose reconstruction.

In terms of the prospective assessment where it is important that regulatory limits are not exceeded (i.e., highest concentrations are not underestimated), the steady-state models were less likely to underestimate maximum one- and nine-hour average concentrations compared to the Lagrangian puff models. However, this result is not only due to differences in model formulation, but also the model parameters such as diffusion coefficients. The unpaired scatter plots (Figs. 4 and 5) showed that the *maximum* observed concentration across all tests was not underestimated by AERMOD at both the 8 and 16-km distance, although the time and place of the observed maximum was not the same as the predicted maximum.

In terms of the retrospective assessment where the objective is an unbiased estimate of the concentration in space and time, the Lagrangian puffs models showed overall better performance, especially at the 16-km distance. In most cases, the Lagrangian puff models for the paired ensemble means exhibited lower variance higher correlation to observed values and a higher percentage of observations within a factor of two of the observations compared to steady-state models.

In terms of the four fundamental plume properties, the steady-state models tended to overestimate maximum concentrations but provide unbiased estimates of the plume mass at the 8-km distance and the 16-km distance for ISC2. Puff models tended to slightly underestimate plume maximums, but were better at locating the

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

	ISC2	AERMOD	CALPUFF	RATCHET
8-km results				
GM C_p/C_o	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 C_p/C_o	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				
GM C_p/C_o	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 C_p/C_o	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.

Model	8-km data			16-km data		
	GM	VG	r	GM	VG	r
Maximum one-hour						
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						
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.

	ISC2	AERMOD	CALPUFF	RATCHET
Plume maximum location				
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				
FB	0.42	−0.43	0.14	−0.34
FB confidence interval	0.31–0.54	−0.52 to −0.35	−0.06 to 0.37	−0.51 to −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 to 0.50	0.44–0.88	−0.79 to 1.00
% within a factor of 2	83.3%	91.7%	75.0%	91.7%
Arc-integrated concentration				
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.

	ISC2	AERMOD	CALPUFF	RATCHET
Plume maximum location				
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				
FB	0.30	−0.57	0.081	−0.47
FB confidence interval	0.10–0.49	−0.71 to −0.44	−0.18 to 0.37	−0.66 to −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 to 0.51	−0.63 to 0.36	0.20–0.79	−0.57 to 0.54
% within a factor of 2	83.3%	50.0%	66.7%	58.3%
Arc-integrated concentration				
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%

plume maximum at the 16-km distance. CALPUFF appeared to more accurately estimate the plume impact region, whereas AERMOD and RATCHET tended to overestimate it and ISC2 underestimated it.

The WVTS consists of only 108 h of measurements taken during February 1991 and are not representative of annual average concentrations. However, the high sampler density resulted in the likelihood that the maximum concentration was detected at either the 8-km or 16-km sampling distance. Moreover, the tests were conducted during the wintertime when stable dispersion conditions would likely result in the maximum one- or eight-hour average concentration over the course of a year. Achieving compliance with National Ambient Air Quality Standards typically is limited by the short-term average concentration limits. Therefore, these results have relevance in terms of model performance for short-term averages over the period of a year.

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

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

	ISC2	AERMOD	CALPUFF	RATCHET
8-km results				
GM C_p/C_o	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 C_p/C_o	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				
GM C_p/C_o	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 C_p/C_o	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%

5. Conclusions

No one single model consistently out-performed the others in all performance objectives or measures and the state-of the art models (CALPUFF and AERMOD) did not exhibit superior performance in all performance objectives to the legacy models (ISC2 and RATCHET). Lagrangian puff models generally exhibited smaller variances, higher correlation, and higher percentage of predictions within a factor of two compared to the steady-state models at these distances. The conceptual framework of a Lagrangian puff model is better suited for long range transport where winds vary spatially across the model domain. Hence, Lagrangian puff models may be preferable for dose reconstruction where model domains can be large and where the assessment question is an unbiased estimate of concentration in time and space. However, model choice depends on site-specific considerations and the assessment questions to be addressed, and therefore no categorical statement can be made

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.

	ISC2	AERMOD	CALPUFF	RATCHET
Daytime tests				
GM C_p/C_o	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 C_p/C_o	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				
GM C_p/C_o	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 C_p/C_o	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				
GM C_p/C_o	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 C_p/C_o	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.

	ISC2	AERMOD	CALPUFF	RATCHET
Daytime tests				
GM C_p/C_o	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 C_p/C_o	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				
GM C_p/C_o	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 C_p/C_o	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 to 0.49	0.35–0.76	0.07–0.54
% within a factor of 2	37%	30%	30%	27%
Nighttime period tests				
GM C_p/C_o	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 C_p/C_o	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.

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

about the performance of one type of model over the other for a specific application.

The steady-state models generally did not underestimate the high-end concentrations at the distances studied, and therefore provide a sound basis for regulatory compliance modeling. Based on the overall performance of ISC2, assessment models that rely on the Gaussian plume model are not necessarily inferior to the current state-of-the-art models in terms of meeting regulatory performance objectives.

There was a general tendency for the steady-state models to predict relatively higher concentrations at the 16-km distance compared to the 8-km distance. This effect is important because it manifests itself at substantially shorter distances (16 km) than what is defined by the EPA as the near-field environment ($\leq$50 km). The EPA requires AERMOD for the near-field environment for demonstration of regulatory compliance, unless compelling reasons are provided to justify the use of an alternative model. Thus, estimated maximum hourly-average concentrations at distances >16 km are likely to be overestimated based on these results. Other investigators (Dresser and Huizer, 2011) found AERMOD to underestimate near-field maximum one- and three-hour average SO_2 concentrations based on data from the Martins Creek Power Plant. However, only eight samplers were used in the study and it is possible (if not likely) that the true maximum concentration in the model domain was not captured. In the WVTS, it was less likely that the maximum concentration within a sampling arc went undetected.

Finally, a compelling reason to use steady-state models for regulatory compliance demonstration is the fact that they are simpler to run, require less user judgment, and are less prone to error than Lagrangian puff models. The CALMET/CALPUFF model simulation in this paper required numerous iterations using different values of RMAX1, RMAX2, and other parameters so that the wind field matched what was expected. In a prospective analysis, it is unlikely tracer or other validation data would be available to test model performance and adjust model parameters accordingly to improve model performance. The need for consistency and assurance that estimated concentrations are not underestimated

are legitimate reasons for using steady-state models for regulatory compliance determination.

Acknowledgments

This work was partially funded by Risk Assessment Corporation, Neeses South Carolina.

References

ASTM (American Society for Testing and Materials), 2000. Standard Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance. American Society for Testing and Materials, West Conshohocken, Pennsylvania. Designation D 6589-00.

Brown, K.J., 1991. Rocky Flats 1990–91 Winter Validation Tracer Study. North American Weather Consultants, Salt Lake City, Utah. Report AG91-19.

Chang, J.C., Hanna, S.R., 2005. Technical Descriptions and User's Guide for the BOOT Statistical Model Evaluation Software Package, Version 2.0.

Chanin, D., Young, M.L., Randall, J., 1998. Code Manual for MACCS2. NUREG/CR-6613. U.S. Nuclear Regulatory Commission, Washington, D.C.

Cimorelli, A.J., Perry, S.G., Venkatram, A., Weil, J.C., Paine, R.J., Wilson, R.B., Lee, R.F., Peters, W.D., Brode, R.W., Paumier, J.O., 2004. AERMOD: Description of Model Formulation. EPA-454/R-03–004. U.S. Environmental Protection Agency, Research Triangle Park, North Carolina.

Cox, W.M., Tikvart, J.A., 1990. A statistical procedure for determining the best performing air quality simulation model. Atmospheric Environment 24A, 2387–2395.

Dresser, A.L., Huizer, R.D., 2011. CALPUFF and AERMOD model validation study in the near field: Martins Creek revisited. Journal of the Air and Waste Management Assoc. 61, 647–659.

EPA (U.S. Environmental Protection Agency), 1992. User's Instructions. User's Guide for the Industrial Source Complex (ISC) Dispersion Models, vol. 1. EPA 450/4 92-008a, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina.

EPA, 2000. Meteorological Monitoring Guidance for Regulatory Modeling Applications. EPA-454/R-99-005. U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park, North Carolina.

EPA, 2007. Updated User's Guide for CAP88-PC Version 3.0. EPA 402-R-00-004. U.S. Environmental Protection Agency, Office of Radiation and Indoor Air, Washington D.C.

Farris, W.T., Napier, B.A., Eslinger, P.W., Ikenberry, T.A., Shipler, D.B., Simpson, J.C., 1994. Atmospheric Pathway Dosimetry Report, 1944–1992. PNWD-2228 HEDR. Pacific Northwest Laboratories, Richland, Washington.

Fox, T.J., 2009. Clarification on EPA-FLM Recommended Settings for CALMET. Memorandum from T.J. Fox to Regional, Modeling Contacts, August 31, 2009. U.S. Environmental Protection Agency, Research Triangle Park, NC.

Grogan, H.A., Aanenson, J.W., McGavran, P.D., Meyer, K.R., Mohler, H.J., Mohler, S.S., Rocco, J.R., Rood, A.S., Till, J.E., Wilson, L.H., 2007. Modeling of the Cerro Grande fire at Los Alamos: an independent analysis of exposure, heath risk, and communication with the public. In: Applied Modeling and Computations in Nuclear Science. ACS Symposium Series, vol. 945. American Chemical Society, Washington, D.C.

Hanna, S.R., Strimaitis, D.G., Chang, J.C., 1991. Hazard response modeling uncertainty (a quantitative method). User's Guide for Software for Evaluating Hazardous Gas Dispersion Models, vol. 1. Air Force Engineering and Service Center, Tyndall Air Force Base, Florida.

Haugen, D.A., Fontino, I.P., 1993. Performance Evaluation of the Terrain-responsive Atmospheric Code (TRAC) Model. Colorado School of Mines, Golden, Colorado.

Hodgin, C.R., 1991. Terrain Responsive Atmospheric Code (TRAC) Transport and Diffusion: Features and Software Overview. RFP-4516. EG&G Rocky Flats, Golden, Colorado.

Lange, R., 1992. Modeling the Dispersion of Tracer Plumes in the Colorado Front Range Boundary Layer During Night- and Day-time Conditions. American Meteorological Society Tenth Symposium on Turbulence and Diffusion, Portland Oregon.

Napier, B.A., 2009. GENII2 Users' Guide, Rev 3. PNNL-14583. Pacific Northwest National Laboratories, Richland, Washington.

Ramsdell Jr., J.V., Simonen, C.A., Burk, K.W., 1994. Regional Atmospheric Transport Code for Hanford Emission Tracking (RATCHET). PNWD-2224-HEDR. Pacific Northwest Laboratories, Richland, Washington.

Ramsdell Jr., J.V., Athey, G.F., McGuire, S.A., Brandon, L.K., 2010. RASCAL 4: Description of Models and Methods. NUREG 1940. U.S. Nuclear Regulatory Commission, Office of Nuclear Security and Incident Response, Washington, D.C.

Rood, A.S., 1999. Performance Evaluation of Atmospheric Transport Models, Revision 1. 3-CDPHE-RFP-1996-FINAL (Rev 1). Risk Assessment Corporation, Neeses, South Carolina.

Rood, A.S., Killough, G.G., Till, J.E., 1999. Evaluation of atmospheric transport models for use in phase II of the historical public exposures studies at the Rocky Flats plant. Risk Analysis 19 (4), 559–576.

Rood, A.S., Grogan, H.A., Till, J.E., 2002. A model for a comprehensive assessment of exposure and lifetime cancer incidence risk from plutonium released from the Rocky Flats plant, 1953–1989. Health Physics 82 (2), 182–212.

Rood, A.S., Voillequé, P.G., Rope, S.K., Grogan, H.A., Till, J.E., 2008. Reconstruction of atmospheric concentrations and deposition of uranium and decay products released from the former uranium mill at Uravan, Colorado USA. Journal of Environmental Radioactivity 99, 1258–1278.

Scire, J.S., Strimatis, D.J., Yamartino, R.J., 2002. A User's Guide to the CALPUFF Dispersion Model Version 5.7. Earth Tech Inc, Concord, Massachusetts.

Till, J.E., Killough, G.G., Meyer, K.R., Sinclair, W.S., Voillequé, P.G., Rope, S.K., Case, M.J., 2000. The Fernald dosimetry reconstruction project. Technology 7, 270–295.

Till, J.E., Rood, A.S., Voilleque, P.G., McGavran, P.D., Meyer, K.R., Grogan, H.A., Sinclair, W.K., Aanenson, J.W., Meyer, H.R., Mohler, H.J., Rope, S.K., Case, M.J., 2002. Risks to the public from historical releases of radionuclides and chemicals at the Rocky Flats Environmental Technology Site. Journal of Exposure Analysis and Environmental Epidemiology 12 (5), 355–372.

Weil, J.C., Sykes, R.I., Venkatram, A., 1992. Evaluating air quality models: review and outlook. Journal of Applied Meteorology 31, 1121–1145.