Evaluations of CALPUFF, HPAC, and VLSTRACK with Two Mesoscale Field Datasets

JOSEPH C. CHANG AND PASQUALE FRANZESE

George Mason University, Fairfax, Virginia

KITTISAK CHAYANTRAKOM

Old Dominion University, Richmond, Virginia

STEVEN R. HANNA

George Mason University, Fairfax, Virginia

(Manuscript received 24 December 2001, in final form 29 July 2002)

ABSTRACT

Results of evaluations of transport and dispersion models with field data are summarized. The California Puff (CALPUFF), Hazard Prediction and Assessment Capability (HPAC), and Chemical/Biological Agent Vapor, Liquid, and Solid Tracking (VLSTRACK) models were compared using two recent mesoscale field datasets—the Dipole Pride 26 (DP26) and the Overland Along-wind Dispersion (OLAD). Both field experiments involved instantaneous releases of sulfur hexafluoride tracer gas in a mesoscale region with desert basins and mountains. DP26 involved point sources, and OLAD involved line sources. Networks of surface wind observations and special radiosonde and pilot balloon soundings were available, and tracer concentrations were observed along lines of whole-air samplers and some fast-response instruments at distances up to 20 km. The models were evaluated using the maximum 3-h dosage (concentration integrated over time) along a sampling line. It was found that the solutions were highly dependent upon the diagnostic wind field model used to interpolate the spatially variable observed wind fields. At the DP26 site, CALPUFF and HPAC had better performance than VLSTRACK. Overall, the three models had mean biases within 35% and random scatters of about a factor of 3–4. About 50%–60% of CALPUFF and HPAC predictions and about 40% of VLSTRACK predictions were within a factor of 2 of observations. At the OLAD site, all three models underpredicted by a factor of 2–3, on average, with random scatters of a factor of 3–7. About 50% of HPAC predictions and only 25%–30% of CALPUFF and VLSTRACK predictions were within a factor of 2 of observations.

1. Introduction

Transport and dispersion models are powerful tools for assessing the consequences resulting from routine industrial emissions, accidental releases of hazardous materials, and dissemination of chemical and biological warfare agents, either in a conventional tactical setting or in a terrorist attack on civilians. This model evaluation exercise concerns three puff models: 1) the California Puff (CALPUFF) model (Scire et al. 2000b); 2) the Defense Threat Reduction Agency's (DTRA) Hazard Prediction and Assessment Capability (HPAC) modeling system (DTRA 1999); and 3) the Naval Surface Warfare Center's Chemical/Biological Agent Vapor, Liquid, and Solid Tracking (VLSTRACK) model (Bauer and Gibbs 1998). HPAC and VLSTRACK are currently used by the U.S. Department of Defense (DoD) for assessing the transport and dispersion of chemical and biological warfare agents. CALPUFF is a model recommended by the U.S. Environmental Protection Agency (EPA) for regulatory applications and is being tested in the current evaluation exercise to determine how its predictions compare with those of the two DoD models. The dispersion model within HPAC is the Second-Order Closure Integrated Puff (SCIPUFF; Sykes et al. 1998) model. Fundamentally, all three models are based on a Gaussian puff dispersion formulation. HPAC (or SCIPUFF) is unique in that it is capable of predicting both the mean and variance of the concentration field. The three models are different in terms of their areas of focus, general level of sophistication, boundary layer parameterizations, treatments of terrain and the transport wind field, handling of surface characteristics, and data ingestion methods and requirements. Therefore, it is necessary to set up a proper framework in order to perform an objective, meaningful evaluation. This mainly involved the use of the same observed meteorological

Corresponding author address: Joseph Chang, School of Computational Sciences, MS 5C3, George Mason University, Fairfax, VA 22030-4444.
E-mail: jchang4@scs.gmu.edu

data and modeling domain. Whenever applicable, default model options were also chosen. Nevertheless, there are still some intermodel differences that cannot be reconciled.

A systematic model evaluation methodology was used to measure model performance. A transport and dispersion model can be evaluated using at least three approaches: statistical, scientific, and operational (e.g., Hanna et al. 1991). In a statistical evaluation, the model can be treated as a ''black box'' in which model outputs are examined to see how well they match observations. It is sometimes possible for a model to give the right answers but as a result of compensating errors. In a scientific evaluation, the model algorithms, physics, and assumptions are examined in detail for their consistency, accuracy, efficiency, and sensitivity. In an operational evaluation, issues related to the model's user-friendliness are considered, such as the user's guide, the user interface, error checking of input data, internal model diagnostics, and output display. Error checking of input data may include different levels of sophistication to validate the range of input data. Internal model diagnostics may include procedures to check the reasonableness of intermediate results. The main focus of this study is on statistical evaluation. Scientific evaluation of the three models can be found in references such as Allwine et al. (1998) for CALPUFF, Nappo et al. (1998) for SCIPUFF (HPAC), and Pendergrass et al. (1996) for VLSTRACK.

The three dispersion models were evaluated using tracer data from two recent mesoscale (~20 km) field experiments: 1) the Dipole Pride 26 (DP26) experiment (Biltoft 1998; Watson et al. 1998) at the Nevada Test Site, Nevada, and 2) the Overland Along-wind Dispersion (OLAD) experiment (Biltoft et al. 1999; Watson et al. 2000) at Dugway Proving Ground, Utah.

DP26 and OLAD are of great value because there are not that many field experiments with similar scales and resolution. These two field datasets have not been previously described in peer-reviewed literature. Moreover, both datasets bear particular relevance to the growing concern on the use of chemical and biological warfare agents by terrorist groups as weapons of mass destruction.

Section 2 of this paper describes the two field experiments. Section 3 briefly summarizes the three models and how they were configured. The evaluation methodology appears in section 4, with the evaluation results given in section 5. Section 6 gives conclusions and discussion. The results are discussed in more detail in separate reports and a conference paper by the authors (Chang et al. 1999, 2000, 2001).

2. Field experiments

Two tracer experiments were used in this model evaluation study: DP26 and OLAD. DP26 involved instantaneous point sources in both the morning and afternoon hours. OLAD involved instantaneous line sources, mainly in the morning hours. There are many similarities between the two experiments in terms of the spatial scale and the type of meteorological and sampling instruments used. The following sections introduce the two experiments in some detail.

a. DP26 field experiments

The DP26 field experiments were conducted in November of 1996 at Yucca Flat (~37°N, 116°W), the Nevada Test Site, Nevada. The experiments were sponsored by the DTRA, with a primary goal to validate transport and dispersion models. Watson et al. (1998) and Biltoft (1998) provide a detailed description of the experiments. Figure 1 shows the test site and instrument layout. Surface roughness length of the valley portion of the site is in the range of 3–6 cm (Biltoft 1998).

The experiments involved instantaneous releases (~10–20 kg) of sulfur hexafluoride (SF_6) at roughly 6 m above the ground, mostly in the early morning or early afternoon hours. Depending on the prevailing wind direction at the test site, the release was either north (locations N2 and N3 in Fig. 1) or south (locations S2 and S3 in Fig. 1) of Yucca Flat. The main sampling array consisted of three lines; each line had 30 whole-air samplers 1.5 m above the ground with a 15-min sampling interval. That is, the samplers measured 15-min-average concentrations. The average spacing between adjacent samplers was about 250 m. The total sampling period was 3 h for each trial. To optimize data collection, the initiation of sampling along the farthest line (~20 km downwind) from the release point was delayed by 30 min. There were also six continuous tracer gas analyzer (TGA) instruments deployed along the middle sampling line that measured SF_6 concentrations at a frequency of 4 Hz. These high-frequency data were not extensively used in this study because the instruments were placed at a distance of about 1.5 km from each other, which gives a spatial resolution too low to represent adequately the crosswind structure of a cloud. Moreover, one of the dispersion models, CALPUFF, produces only hourly average results and does not permit a study of the detailed along-wind structure of a cloud.

There were a total of 21 releases for which sampler data were successfully collected. However, to take advantage of the 3-h sampling period, sometimes two consecutive releases were made about 90 min apart, and the two releases were considered as one trial. The sampler data for these trials included contributions from two separate releases (puffs). Therefore, only 14 separate trials were identified, and the sampler data were arranged and available according to trials.

Surface meteorological conditions were measured by eight meteorological data (MEDA) stations at 15-min intervals at a height of 10 m above the ground. These stations are designated as M1, M2, and so on, in Fig. 1. One radiosonde station (UCC near M6) and two pilot

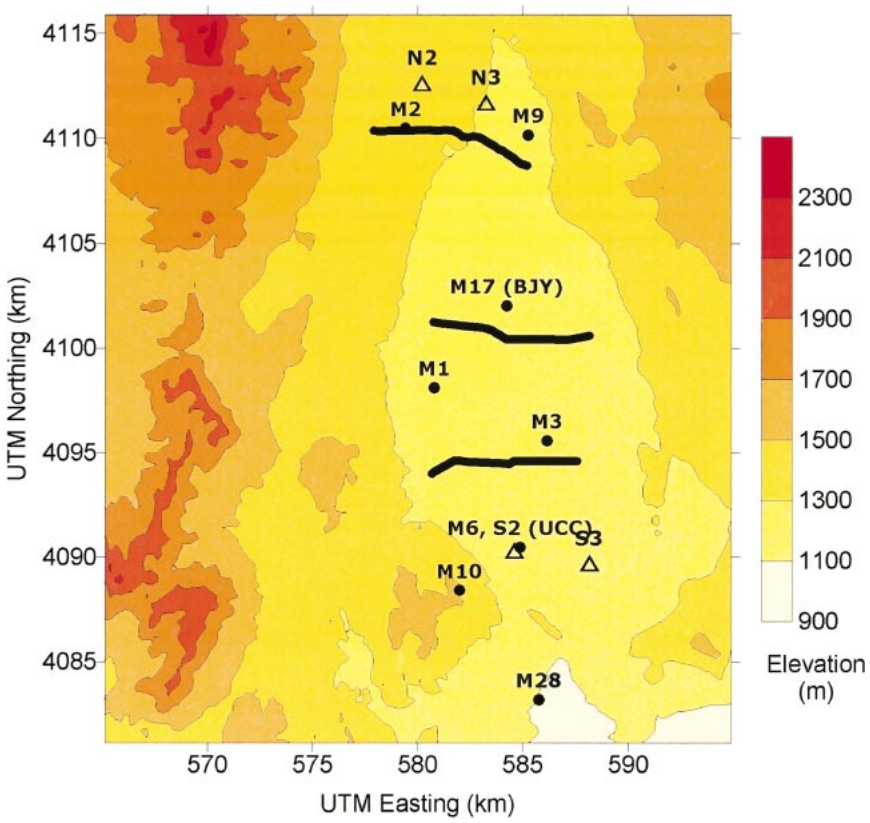

FIG. 1. Terrain elevation at the DP26 test site at Yucca Flat, Nevada Test Site, NV. Also shown are the three SF_6 sampling lines (thick lines, 30 samplers per line), eight MEDA surface meteorological stations (solid circles), and four possible release locations (open triangles). There are also two pibal stations (BJY near M17, and UCC near M6) and one radiosonde station (UCC). The map covers an area of 30 km × 35 km. The southwest corner of the map roughly corresponds to (36.9°N, 116.3°W).

balloon (pibal) stations (UCC, and BJY near M17) provided upper-air measurements. Radiosondes were released every 3–12 h. The 3-h interval applied during the tracer releases. Pibals were typically released every hour in conjunction with the releases. Pibals provided only wind measurements, whereas radiosondes also provided temperature data. Pibal and radiosonde launches were staggered, that is, not simultaneous. Moreover, two sonic anemometers were installed at locations UCC and BJY.

Table 1 summarizes DP26's release information (time, location, and quantity) and typical 10-m wind speeds over the test site. The standard deviations [or root-mean-square (rms) differences] for hourly average wind speeds and directions measured at the eight MEDA stations are roughly 0.5–2 m s^{-1} and 10°–30°, respectively. Note that each standard deviation is from eight 1-h-average observations, and the ranges of standard deviations are based on the 3-h periods for all DP26 trials.

b. OLAD field experiments

The OLAD field experiments were conducted in September of 1997 at West Desert Test Center (WDTC; ~40°N, 113°W), U.S. Army Dugway Proving Ground, Utah. The experiments were jointly sponsored by the Joint Chemical/Biological Contract Point and Test Management Office, WDTC, and the Naval Surface Warfare Center, Dahlgren Division, Dahlgren, Virginia. Watson et al. (2000) and Biltoft et al. (1999) provide detailed descriptions of the experiments. The test domain is mostly a dry mud flat with surrounding mountains (see Fig. 2). The typical value of the surface roughness length in the broad valley is 3 cm, according to the analysis by Biltoft et al. (1999).

The experiments involved releases of SF_6 from a truck or an aircraft in the early morning hours. The predominant wind direction was from the southeast during all experiments. Approximately 12 kg of SF_6 were released by a truck-mounted disseminator traveling at a normal speed of 64 km h^{-1} over a distance of 8 km and at a height of 3 m above the ground. During the aircraft releases, about 100 kg of SF_6 were released by an aircraft traveling at 200 km h^{-1} over a distance of 16 km and at 100 m above the ground. The truck and aircraft release lines are indicated as thin and thick dashed lines, respectively, in Fig. 2. Nine successful

TABLE 1. DP26 trial dissemination times, positions, and mass (Biltoft 1998) and typical wind speeds at 10 m above the ground. Time is Pacific standard time (PST). Trial numbers that end with a letter ''a'' or ''b'' indicate separate releases within a trial. Therefore, the total number of trials is 14. See Fig. 1 for release locations.

Trial No.	Date (Nov 1996)	Yearday	Release time (PST)	Mass released (kg)	Release location	Typical 10-m wind speed (m s^{-1})
1	4	309	1441	8.0	S2	4.3
3	8	313	0400	12.3	N3	3.1
4a	9	314	0400	11.5	N3	3.8
4b	9	314	0538	11.5	N3	3.8
5	11	316	0440	11.5	N2	2.9
6	12	317	0400	11.6	N2	2.3
7a	12	317	1300	19.3	S3	3.2
7b	12	317	1447	10.0	S3	2.7
9	13	318	1400	10.4	S2	3.8
11a	14	319	1430	10.6	N2	3.1
11b	14	319	1551	10.8	N2	2.8
12a	15	320	0900	11.5	N2	3.2
12b	15	320	1030	11.3	N2	5.5
13	15	320	1430	21.6	N2	4.5
14	16	321	1300	21.1	S2	4.2
15a	18	323	1130	10.8	S2	2.9
15b	18	323	1300	20.2	S2	5.0
16a	19	324	1200	20.3	S3	4.0
16b	19	324	1330	20.3	S2	6.0
17a	20	325	1200	20.4	S3	3.7
17b	20	325	1330	20.1	S2	4.1

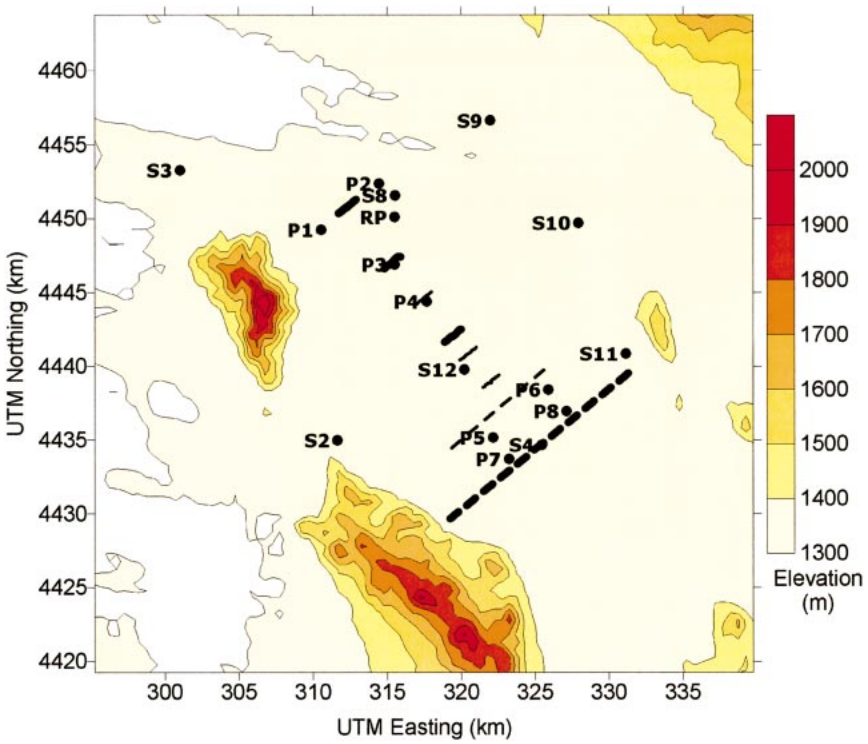

FIG. 2. Terrain elevation at the OLAD test site at West Desert Test Center, U.S. Army Dugway Proving Ground, UT; 2-m PWIDS instrument masts are listed as P1, P2, etc.; 10-m SAMS instrument towers are listed as S2, S3, etc. The radiosonde and pibal measuring site is listed as RP. The thin dashed line is the line source from the truck, and thin solid lines are the corresponding sampling lines. The thick dashed line is the line source from the aircraft, and thick solid lines are the corresponding sampling lines. The map covers an area of 45 km × 45 km. The southwest corner of the map roughly corresponds to (39.9°N, 113.4°W).

TABLE 2. OLAD trial dissemination start and end times, type, and mass (Biltoft et al. 1999) and typical wind speeds at 10 m above the ground. Time is mountain standard time (MST). Trial 12 was interrupted at 6 min by a disseminator failure and resumed 5 min later. The three truck releases under trial 6 were considered as a single trial. Therefore, the total number of trials is 11.

Trial No.	Date (Sep 1997)	Yearday	Start time (MST)	End time (MST)	Mass (kg)	Release type	Typical 10-m wind speed (m s^{-1})
1	8	251	0606	0614	12.4	Truck	2
2	9	252	0545	0553	12.9	Truck	4
3	10	253	0629	0632	100.3	Aircraft	2
4	11	254	0556	0559	100.5	Aircraft	6
5	12	255	0558	0608	12.8	Truck	1.5
6a	15	258	0545	0552	12.5	Truck	9
6b	15	258	0646	0658	11.4	Truck	9
6c	15	258	0730	0743	12.6	Truck	9
7	15	258	0945	0956	12.8	Truck	8
9	17	260	0548	0551	96.1	Aircraft	2
10	18	261	0655	0705	7.1	Truck	3.5
11	24	267	0609	0612	99.1	Aircraft	1.5
12	25	268	0300	0306	12.1	Truck	3
			0311	0319			

truck releases and four successful aircraft releases were made. However, there are only 11 separate trial designations, because one trial (trial 6) included three truck releases. Table 2 shows the OLAD trial dissemination start and end times, release type, and mass released as well as typical wind speeds at 10 m above the ground.

Three lines of whole-air samplers were deployed with 15 samplers on each line to measure SF_6 concentrations. Each whole-air sampler measured 15-min-average concentration and was located roughly 100 m from adjacent samplers. Therefore, the OLAD sampling lines were about 5 times as short as those for DP26, because of smaller sampler spacing (100 vs 250 m) and a smaller number of samplers per line (15 vs 30). The OLAD sampling lines for the truck and aircraft releases are indicated as thin and thick solid lines, respectively, in Fig. 2. The sampling lines were located 2, 5, and 10 km downwind for the truck releases and 10, 15, and 20 km downwind for the aircraft releases. The total sampling period was 3 h for each trial. To account for the travel time from the release location to the farthest sampling line, those samplers along the farthest line were not turned on until 30 min after the release. In addition to the 15-min-average concentrations measured by the whole-air samplers, 4-Hz SF_6 concentrations were measured by the TGA deployed at both ends of each sampling line. One TGA was also used on the aircraft.

Winds were measured by eight 2-m Portable Weather Information and Display System (PWIDS) masts, and eight 10-m Surface Atmospheric Measurement System (SAMS) towers (see Fig. 2). Upper-air winds were measured by pibals and radiosondes. The pibal and radiosonde launching sites were collocated (at location RP in Fig. 2), but the balloons were not simultaneously released.

The surface wind fields measured by the PWIDS and SAMS stations exhibited high spatial variability. The standard deviations for hourly average wind speeds and directions are less than 1 m s^{-1} and 10°–30° for PWIDS and 0.5–2 m s^{-1} and 10°–70° for SAMS. As in DP26, each standard deviation is from eight 1-h-average PWIDS or SAMS observations, and the ranges of standard deviations are based on the 3-h periods for all OLAD trials. Larger variations in the SAMS data are mainly due to the larger area occupied by the SAMS network.

c. Quality assurance

Before actually conducting any model evaluation exercise, it is imperative that the meteorological and concentration data be subject to rigorous quality-assurance (QA) procedures. This step is necessary despite built-in QA checks in dataloggers and other secondary checks before data distribution. Here, the QA procedures included 1) screening data for possible inconsistencies in time zone designation, units, and missing data indicator; 2) examining spatial distributions of surface winds to flag visibly questionable data; 3) comparing concentration and wind data from neighboring instruments; and 4) inspecting the whole-air sampler data according to the quality indicator.

In general, the quality of the two datasets was satisfactory. However, DP26's radiosonde data in the first 500 m or so above the ground were found to be questionable and were not used. This lack of quality is because the radiosonde tracker sometimes had problems tracking a rapidly ascending balloon in a strongly sheared environment, which possibly formed because of the surrounding mountains. In addition, only about 75% of the OLAD whole-air sampler data had satisfactory quality flags (i.e., indicating either ''good data'' or ''below detection limit''). Watson et al. (2000) attribute OLAD's poor sampler recovery rate to a problem with the data collection software. Chang et al. (1999, 2001) describe the results of the quality-assurance analysis in more detail.

3. Dispersion models

The CALPUFF (version 5.0; Scire et al. 2000b), HPAC (version 3.2.1; DTRA 1999), and VLSTRACK (version 3.0; Bauer and Gibbs 1998) transport and dispersion models have been evaluated with data collected during the DP26 and OLAD field experiments. The reader is referred to the respective technical documents for each model's formulations and theoretical background. The similarities and differences among the three models are briefly highlighted below. All three models are based on a Gaussian puff dispersion formulation, although the dispersion model within HPAC is the SCIPUFF (Sykes et al. 1998) model, which is more sophisticated and is capable of predicting both the mean and variance of the concentration field.

The current version of CALPUFF accepts only hourly average meteorological information and predicts only hourly average concentrations. As a result, the model predictions cannot be directly compared with high-frequency concentration data that are typical of puff experiments such as DP26 and OLAD. This limitation of a 1-h averaging time primarily results from the fact that CALPUFF has traditionally been used in EPA's regulatory applications, for which environmental impacts from routine industrial releases are modeled, the hourly average concentration is the predominant variable of interest, and only hourly meteorological data are routinely available. HPAC and VLSTRACK can readily accept and produce higher-frequency data. HPAC was run with an optional output frequency of 20 s. VLSTRACK always generates concentration results every 60 s. Because all models need to be compared on an equal basis, the higher-frequency HPAC and VLSTRACK results and TGA measurements were mostly not used.

All three models treat line sources by approximating them with many volume sources along the length of the line. The approximation is done internally in the codes and does not require user intervention.

All three dispersion models require gridded wind fields for dispersion calculations. CALPUFF has its own diagnostic wind field model called ''CALMET'' (Scire et al. 2000a). HPAC has two optional diagnostic wind field models, the HPAC mass-consistent wind model (MC-SCIPUFF) and the more advanced Stationary Wind Fit and Turbulence (SWIFT) model. SWIFT is adapted from the MINERVE (méthode d'interpolation et de reconstitution tridimensionnelle d'un champ de vent) diagnostic model (Perdriel et al. 1995). SWIFT is the default choice for HPAC and was used in all HPAC runs in this study. VLSTRACK does not have an integrated wind field model but uses the observed wind measurements directly to create the required gridded wind field by means of a three-point interpolation scheme. That is, the wind at a given location is determined by the three closest observations through interpolation.

Chang et al. (1999, 2001) describe in detail the modeling assumptions for DP26 and OLAD. In summary, the modeling domains for all three dispersion models were the same, that is, 30 km $\times$ 35 km for DP26 and 45 km $\times$ 45 km for OLAD (see also Figs. 1 and 2). The grid spacing used for interpolating wind fields was the same for CALPUFF and VLSTRACK—250 m for DP26 and 500 m for OLAD. HPAC used a fixed grid with about 3000 grid cells over the modeling domain, which corresponds to about 590-m grid spacing for DP26 and 820 m for OLAD. CALPUFF used the terrain and land use data from the U.S. Geological Survey. The HPAC package includes its own spatially varying geophysical database. VLSTRACK, on the other hand, does not require such sophisticated geophysical data, other than the base elevation and a simple classification of nine ground surface types for each surface meteorological station.

To carry out model evaluation on the same basis, all three models were run using the same observed meteorological data. Each dispersion model would further process these observational data with its own wind field model, that is, CALMET for CALPUFF, SWIFT for HPAC, and three-point interpolation for VLSTRACK. CALPUFF and HPAC used data from both surface and upper-air stations. VLSTRACK used only surface meteorological data, because the model requires uniform data availability for all stations. For example, if one station's upper-air data were to be included, then the model requires that all other stations' upper-air data are also to be included. As a result, this requirement precludes VLSTRACK from accepting a mixture of data from surface and upper-air stations.

4. Evaluation methodology

Because the total sampling period for a given trial in both field experiments was 3 h, concentrations from the whole-air samplers integrated over the 3-h measuring period (i.e., 3-h dosages) will be the primary model output used in model evaluation. CALPUFF is limited to producing only hourly average concentrations, whereas HPAC and VLSTRACK can produce results at a much shorter time interval (20 and 60 s, respectively). Therefore, the 3-h sampling period corresponds to only 3 CALPUFF predictions, but 540 and 180 predictions for HPAC and VLSTRACK, respectively. CALPUFF's low-resolution concentration data are not sufficient to study issues such as the cloud arrival and departure times and along-wind dispersion. Model evaluation was mainly based on the maximum dosage anywhere along a sampling line in the current paper. Chang et al. (1999, 2001) also discuss the results based on the summation of dosages over all samplers along a sampling line.

The performance of the three models was assessed using two basic methodologies. The first methodology involved the use of scatterplots for direct quantitative comparisons of observed and predicted dosages at the

two field sites. The second methodology involved the application of statistical procedures that quantify several relevant performance measures (Hanna 1989; Hanna et al. 1993).

Hanna et al. (1993) consider the following statistical measures to determine quantitative model performance: the fractional bias (FB), the geometric mean bias (MG), the normalized mean square error (NMSE), the geometric variance (VG), and the fraction of predictions within a factor of 2 of observations (FAC2). They are defined as

$$\text{FB} = \frac{(\overline{C_o} - \overline{C_p})}{0.5(\overline{C_o} + \overline{C_p})}, \quad (1)$$

$$\text{MG} = \exp(\overline{\ln C_o} - \overline{\ln C_p}), \quad (2)$$

$$\text{NMSE} = \frac{\overline{(C_o - C_p)^2}}{\overline{C_o}\,\overline{C_p}}, \quad (3)$$

$$\text{VG} = \exp[\overline{(\ln C_o - \ln C_p)^2}], \quad \text{and} \quad (4)$$

$$\text{FAC2} = \text{fraction of data for which}$$

$$0.5 \le \frac{C_p}{C_o} \le 2.0, \quad (5)$$

where C is the evaluation objective (here, the maximum dosage along a sampling line), C_p is the model predictions, C_o is the observations, and overbar ($\overline{C}$) is the average over all data values.

Bootstrap resampling (Efron 1987) was used to estimate the mean μ and standard deviation σ of the above performance measures. The 95% confidence intervals for the performance measures are defined as

$$\mu \pm t_{95\%}\sigma\left(\frac{n}{n-1}\right)^{1/2}, \quad (6)$$

where n is the number of resamples (e.g., 1000) and $t_{95\%}$ is the Student's t value at the 95% confidence limits with $n - 1$ degrees of freedom.

Both FB and MG deal with mean biases; however, FB uses concentrations directly, whereas MG uses the logarithm of concentrations. A similar relation occurs between NMSE and VG, which both deal with variances or random scatter. A perfect model would have MG, VG, and FAC2 = 1.0 and FB and NMSE = 0.0. Each performance measure has its advantages and shortcomings. The relative advantages of a measure are partly determined by the distribution of the variable of interest. NMSE and FB are more strongly influenced by infrequently occurring high observed and predicted values, whereas MG and VG provide a more balanced treatment of both high and low values. For a dataset in which both predicted and observed values vary by several orders of magnitude, MG and VG would be more appropriate. Therefore, only the results based on MG and VG will be presented below.

However, MG and VG are strongly influenced by extremely low values and are undefined for zero values. These low and zero values are not uncommon in dispersion modeling. Therefore, when calculating MG and VG for observed or predicted values whose magnitude may be very low, it is useful to impose a minimum threshold below which the data values are not allowed to drop. The whole-air samplers' limit of detection (LOD) for SF_6 concentrations was chosen as the basis of this minimum threshold. Watson et al. (1998, 2000) report that the LOD for DP26 and OLAD was around 10 and 3 parts per trillion (ppt), respectively. Therefore, the corresponding minimum thresholds for the 3-h dosage were around 30 and 10 ppt h for DP26 and OLAD, respectively. These lower thresholds were used in the scatterplots and in the calculations of MG and VG, to be presented later.

It can be shown that MG is simply the ratio of the geometric mean of C_o to the geometric mean of C_p. Therefore, a factor-of-N mean bias would mean MG = $1/N$ or N. However, it is more difficult to discern, for example, what value of VG would correspond to a factor-of-2 bias, or what VG = 12 would mean. One way to relate the values of VG to other quantities that are more easily understood is to assume that the ratio of C_p/C_o is equal to a constant A, that is, to ignore the random scatter between C_p and C_o. Then Eq. (4) becomes

$$\text{VG} = \exp[(\ln A)^2], \quad \text{or} \quad A = \exp(\sqrt{\ln \text{VG}}). \quad (7)$$

In other words, a factor-of-2 mean bias (i.e., $A = 2.0$ or 0.5) would mean VG = 1.6, and VG = 12 would indicate a random scatter that is equivalent to roughly a factor-of-5 mean bias ($A = 4.84$). Note that Eq. (7) mainly provides for VG, which includes both systematic bias and relative scatter, an alternative interpretation that is easier to conceptualize. It does not exactly equate VG with systematic bias.

5. Evaluation results

a. Model evaluation with DP26 data

The maximum SF_6 dosage anywhere along a sampling line was chosen for model evaluation at DP26. There were a total of 14 trials (7 of which had two releases) monitored by three sampling lines. Therefore, the sample size is 42 ($=14 \times 3$). Table 3 summarizes the results of statistical performance evaluation for DP26. Overall, the performances for CALPUFF and HPAC were comparable. VLSTRACK tended to overpredict and had a larger scatter. CALPUFF, HPAC, and VLSTRACK yielded values of MG corresponding to 5% underprediction, 25% underprediction, and 35% overprediction, respectively, and values of VG corresponding to random scatter of a factor of 3, 3, and 4, respectively [see Eq. (7)]. FAC2 was about 50%–60% for CALPUFF and HPAC and slightly lower, about 40%, for VLSTRACK. Model performance was similar based

TABLE 3. Summary of performance measures, including geometric mean bias, geometric variance, and fraction of predictions within a factor of 2 of observations, for the maximum dosage (ppt h) anywhere along a sampling line for CALPUFF, HPAC, and VLSTRACK at the DP26 site. The avg, std dev, highest value (HI1), and second highest value (HI2) of the data are shown. A perfect model would have MG, VG, and FAC2 = 1.0. Observed values are based on whole-air samplers. Sample size is 42.

	Observed	CALPUFF	HPAC	VLSTRACK
MG	—	1.069	1.316	0.739
VG	—	3.06	2.87	6.45
FAC2	—	0.524	0.595	0.429
Avg (ppt h)	927	917	838	1461
Std dev (ppt h)	1463	1979	1610	3170
HI1 (ppt h)	7036	9232	7761	19 092
HI2 (ppt h)	5002	9033	5334	7048

on the crosswind summed dosages, which are not shown here but are fully described in Chang et al. (2001).

Figure 3 shows the scatterplots of observed versus predicted maximum dosages for CALPUFF, HPAC, and VLSTRACK, for which a lower threshold of 30 ppt h mentioned above was used. The scatterplots show relatively few low predictions for CALPUFF. This may be a result of the more robust approach adopted by CALMET, CALPUFF's diagnostic wind field model, in which surface winds are extrapolated upward, based on similarity theories, to blend with upper-air wind observations. (However, as explained later, this approach may not be universally valid.) Both HPAC and VLSTRACK showed a few cases with very low simulated dosages, one to two orders of magnitude smaller than observed. This result is mainly due to the predicted puff missing the sampling line.

The scatterplots can also be studied to see how well the models predicted the highest dosages. Of the five highest observed dosage points, three of these points were included in the five highest predicted dosage points by CALPUFF, four of them by HPAC, and three of them by VLSTRACK. The single highest observed dosage (ppt h) was 7036. The single highest predicted dosage on the plots was 9232 for CALPUFF, 7761 for HPAC, and 19 092 for VLSTRACK, or overpredictions of 30%, 10%, and a factor of 2.7, respectively. In no case was the time of the highest observed dosage the same as the time of the highest predicted dosage. However, the time of the second-highest observed dosage, 5002, corresponded to the time of the highest predicted dosage for all three models.

Based on the values of MG, VG, and FAC2, it is concluded that CALPUFF and HPAC had comparable performance, whereas the performance of VLSTRACK was slightly worse. However, only the values of MG for HPAC and VLSTRACK were significantly different at the 95% confidence limits. The values of VG for all three models were not significantly different.

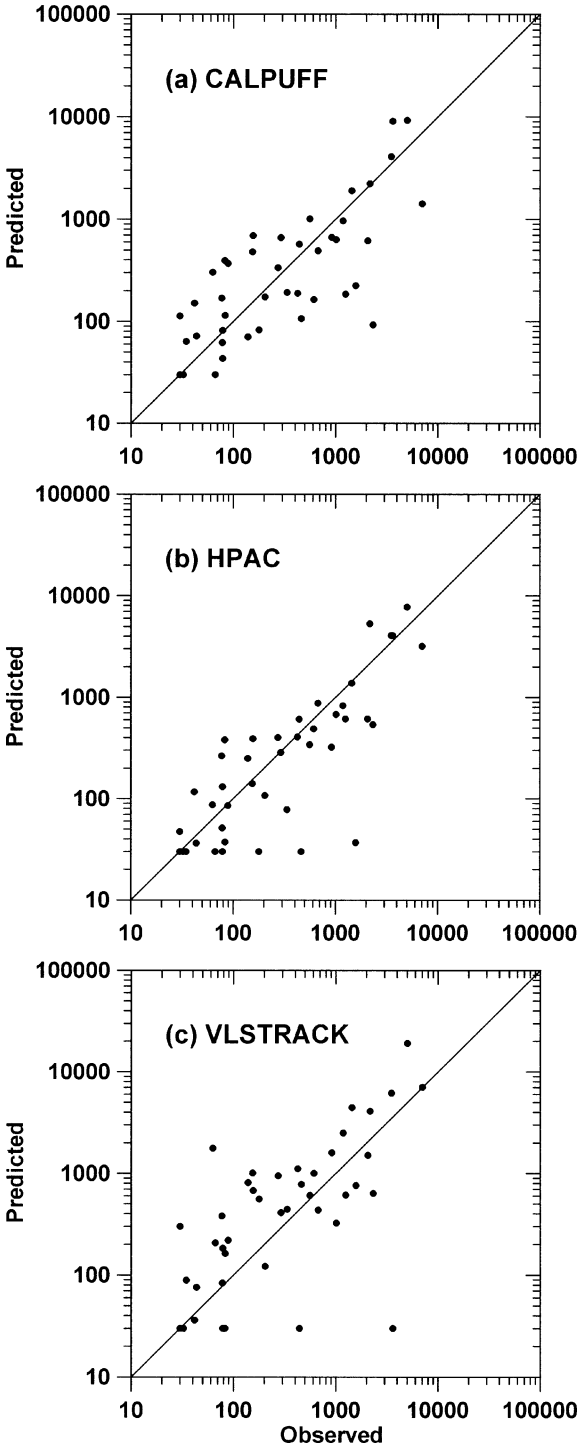

FIG. 3. Scatterplots of the maximum dosage (ppt h) at each sampling line predicted by (a) CALPUFF, (b) HPAC, and (c) VLSTRACK vs observations from the whole-air samplers at the DP26 site. Note that a lower dosage threshold of 30 ppt h, based on the limit of detection, was imposed. Sample size is 42.

TABLE 4. Same as Table 3 but at the OLAD site. Sample size is 21.

	Observed	CALPUFF	HPAC	VLSTRACK
MG	—	1.819	2.057	3.340
VG	—	9.82	3.61	45.68
FAC2	—	0.286	0.476	0.238
Avg (ppt h)	2101	748	854	418
Std dev (ppt h)	2935	743	1044	338
HI1 (ppt h)	10 210	2988	3993	1004
HI2 (ppt h)	8603	2067	2426	955

b. Model evaluation with OLAD data

As in DP26, the maximum SF_6 dosage anywhere along a sampling line was also chosen for model evaluation. Although there were 11 trials with three sampling lines each for OLAD, the sample size was only 21 instead of 33 ($=11 \times 3$), because of the poor data recovery rate (75%) mentioned above. Many sampling lines did not have sufficient data to determine the maximum values. Table 4 summarizes the results of statistical performance evaluation for OLAD. In general, all three models underpredicted at the OLAD site, with CALPUFF having the smallest mean underprediction. CALPUFF, HPAC, and VLSTRACK yielded predictions that correspond to a factor of 2–3 mean underprediction based on MG and a factor of 3–7 random scatter based on VG. The values of MG and VG for the three models are not significantly different at the 95% confidence limits. It seems counterintuitive that the VG values for the three models, although very different, are not significantly different in a statistical sense. This result is mainly due to a smaller sample size (21). FAC2 was about 30% for CALPUFF, 50% for HPAC, and 25% for VLSTRACK.

Figure 4 shows the scatterplots for the maximum dosage, for which a threshold dosage of 10 ppt h mentioned above was used. The agreement is clearly worse than that for DP26 (Fig. 3). This result is probably because many of the OLAD trials were conducted in the morning transition periods (Table 2). There are again a few cases of large model errors (up to two orders of magnitude). Further investigation shows that model performance was appreciably better for the high-wind (≥ 6 m s^{-1}) cases. Chang et al. (2001) suggested that the causes for a systematic underprediction of dosage for an instantaneous line source could be overprediction of either the vertical dispersion coefficient or the cloud advective speed. Figure 4 also shows that VLSTRACK had three ''zeros,'' as indicated by the threshold dosage of 10 ppt h. This may seem questionable, because the release line was comparable to, if not longer than, the travel distance for OLAD. Two factors may have contributed to this anomaly. First, VLSTRACK has an internal concentration threshold of 0.0001 mg m^{-3}, below which no results will be printed. This internal threshold equals about 17 ppt for SF_6, which is almost 6 times as high as the LOD for OLAD (3 ppt). Therefore, there were some low con-

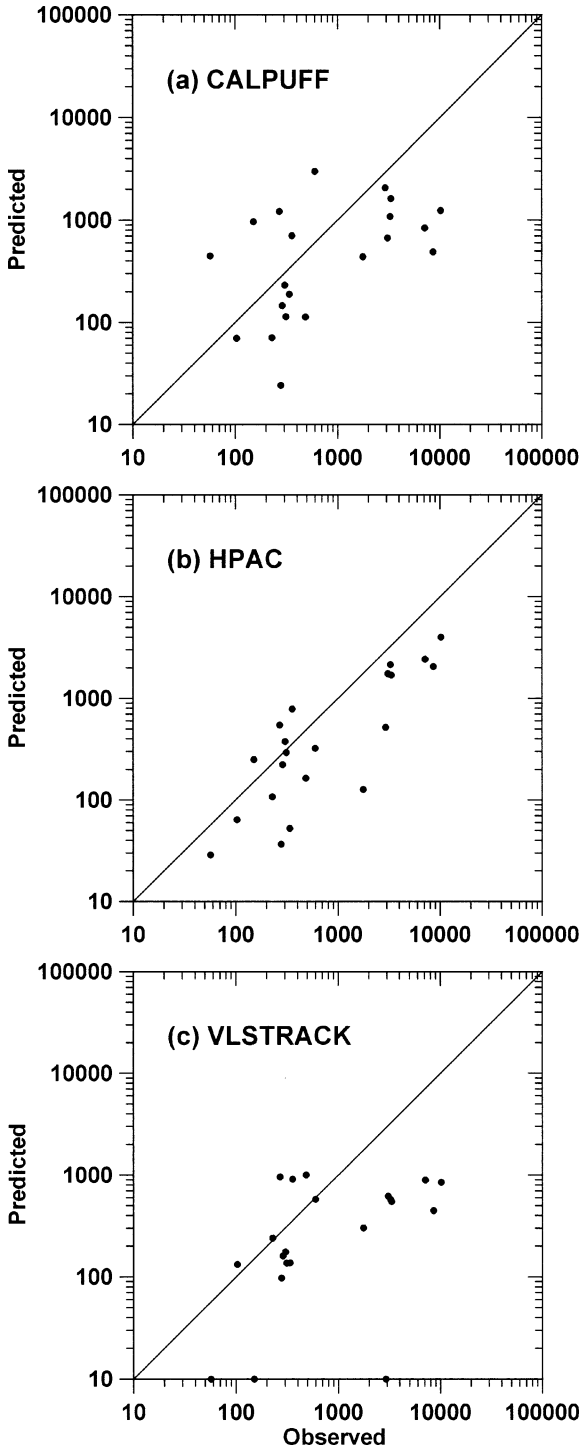

FIG. 4. Scatterplots of the maximum dosage (ppt h) at each sampling line predicted by (a) CALPUFF, (b) HPAC, and (c) VLSTRACK vs observations from the whole-air samplers at the OLAD site. Note that a lower dosage threshold of 10 ppt h, based on the limit of detection, was imposed. Sample size is 21.

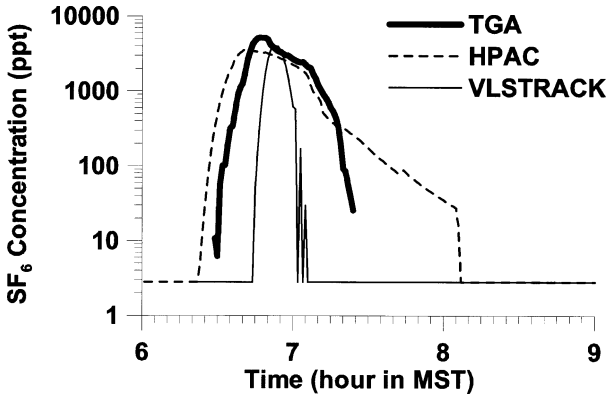

FIG. 5. Time series of 60-s-average SF_6 concentrations observed by the TGA (thick solid) and predicted by HPAC (dashed) and VLSTRACK (thin solid) at the east end of the middle sampling line (~5 km from the source) for OLAD trial 5. A lower concentration threshold of 3 ppt was used.

centration values that were not reported by VLSTRACK and were then treated as zero. Second, the unsophisticated three-point method used by VLSTRACK to interpolate the wind measurements may have failed to capture the high spatial variability exhibited in the observed surface wind fields.

As was done with the DP26 data, the scatterplots in Fig. 4 can be studied to see how well the models predicted the highest dosages. Two of the five highest CALPUFF predictions were also included in the five highest observed dosage values. The same count for HPAC and VLSTRACK was four and two, respectively. The overall maximum observed dosage (ppt h) was 10 210. The overall maximum predicted dosages by CALPUFF, HPAC, and VLSTRACK were 2988, 3993, and 1004, respectively. These correspond to underpredictions by a factor of 3.4, 2.6, and 10.2, respectively, which is contrary to the case for DP26 in which all three models overpredicted the overall maximum. HPAC was the only model for which the time of the highest observed dosage was the same as the time of the highest predicted dosage.

High-frequency (4 Hz) TGA data were also taken at both ends of the three sampling lines during the OLAD experiments. The TGA data have not been extensively used thus far because of inadequate spatial coverage and a relatively short total sampling period (~1 h). The quantitative performance measures and scatterplots presented above were all based on the whole-air sampler data. Nevertheless, it would be of interest to investigate how model-predicted concentrations qualitatively compare with TGA measurements. The CALPUFF results are not included in the analysis, because the model gives only hourly predictions and the TGA sampling period was at best slightly longer than 1 h. On the other hand, HPAC and VLSTRACK gave concentration results every 20 and 60 s, respectively, for the current study. As a result, time series of 60-s-averaged TGA concentrations, together with HPAC and VLSTRACK predicted concentrations, were plotted and investigated. HPAC generally showed better agreement with the observed concentration time series. VLSTRACK consistently calculated shorter puff passage times, the difference between the puff departure and arrival times. This can be seen in Fig. 5, a sample time series comparison for the TGA located at the east end of the middle sampling line for OLAD trial 5. A closer inspection of all available time series revealed that both models predicted earlier puff arrival times (or higher cloud advective speeds) about 65% of the time, consistent with the conjecture made by Chang et al. (2001) regarding the potential overprediction of the cloud advective speed.

TABLE 5. Rms differences in wind speeds and wind directions over the test domain for 1-h-averaged DP26 and OLAD observations during the trial times.

	Wind speed rms diff ($m\ s^{-1}$)	Wind direction rms diff (°)
DP26, 10-m MEDA	0.5–2	10–30
OLAD, 10-m SAMS	0.5–2	10–70
OLAD, 2-m PWIDS	<1	10–30

c. Sensitivity to surface wind fields

As mentioned above, even though the DP26 and OLAD field trials were conducted over test domains that were mostly flat valleys with little vegetation, the observed surface wind fields exhibited large spatial variability, possibly because of the influence of surrounding mountains. The rms differences in wind speeds and wind directions over the test domain for 1-h-averaged DP26 and OLAD observations during the trial times were mentioned earlier and are now summarized in Table 5.

Because diagnostic wind field models essentially extrapolate randomly spaced observations to create the necessary gridded wind fields for dispersion calculations, simulated wind fields are very sensitive to the spatial variability in observed wind fields. Figure 6 shows the observed surface wind fields for DP26 trial 3 (0400–0700 LST 8 November 1996), for which there was considerable spatial variability among the eight MEDA stations. The data-withholding technique (e.g., Bergin et al. 1999) was used to investigate the sensitivity of dispersion model results to input meteorological data, in which data from one of the surface stations were withheld at a time from diagnostic wind modeling. In other words, if data are available from, say, *N* surface stations, then, in addition to the base run in which data from all stations were used, *N* additional sensitivity runs were made, each with data from one of the stations withheld. Figures 7a–i show the 3-h dosage contours for the base run and the eight additional data-withholding runs generated by HPAC. The dosage contours look qualitatively similar overall, but there are appreciable differences in cloud widths and in dosage predictions at fixed locations. Figure 6 shows that the wind vector at MEDA station 9 [M9, approxi-

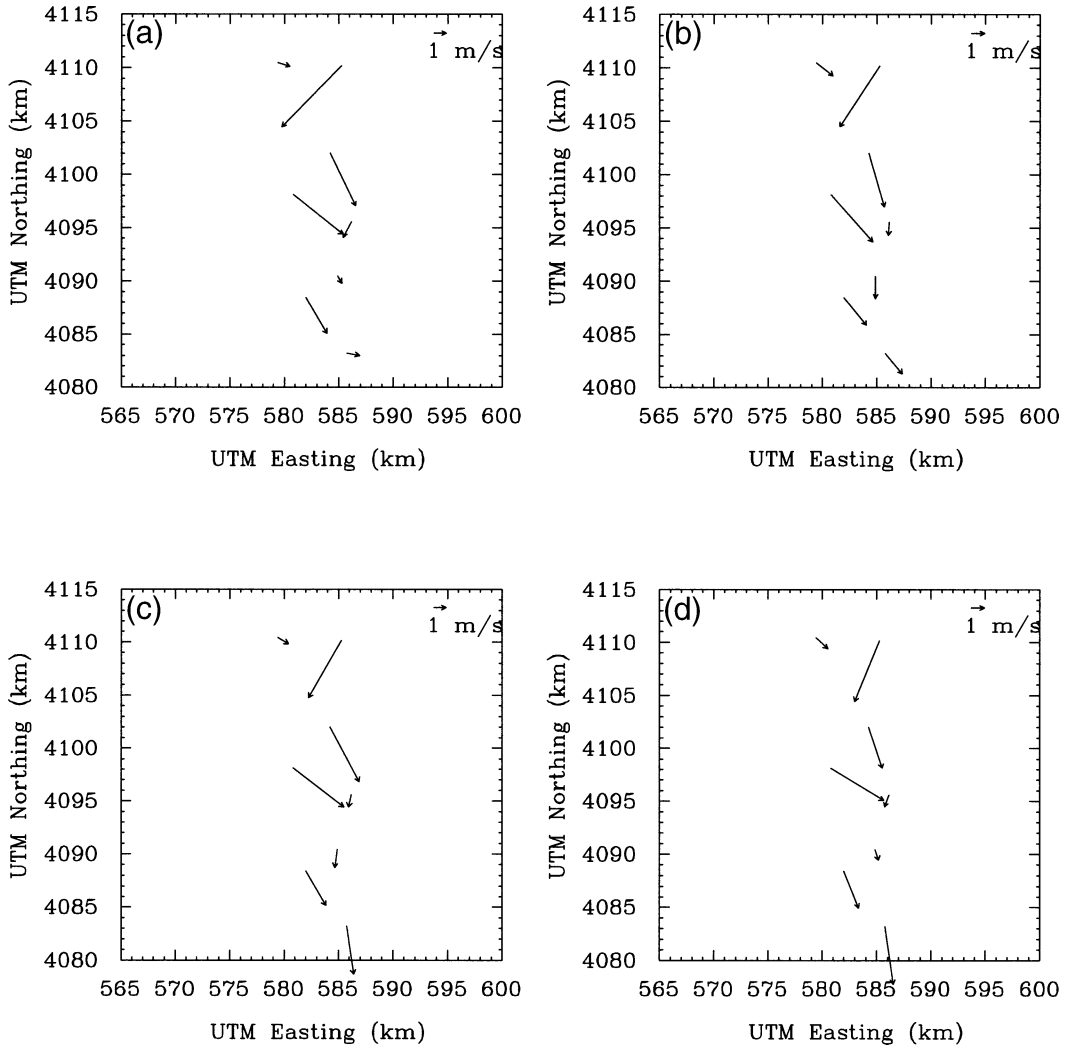

FIG. 6. Observed surface wind fields for DP26 trial 3, 8 Nov 1996: (a) 0400, (b) 0500, (c) 0600, and (d) 0700 LST. See Fig. 1 for detailed terrain elevation of the meteorological stations.

mately located at 585 km Universal Transverse Mercator (UTM) east and 4110 km UTM north] deviates more from the remaining stations. Station M9 was the station closest to the release (Fig. 1 and Table 1). Figure 7f shows the predicted dosage contour with the data from station M9 withheld. The initial southwestern cloud trajectory common to all other sensitivity runs was absent in this sensitivity run. As a result, the overall dosage contours shifted more to the east, to such an extent that the cloud was barely captured by the farthest sampling line. The results also demonstrate that diagnostic wind models may not be capable of preserving the spatial variability present in the observed wind field even with a relatively dense network of surface wind monitors. This shortcoming is because modeled wind fields are sensitive to the number of stations considered in the simulation and may be significantly affected when the data from just one station were withheld.

The sensitivity of DP26 dispersion model results to input wind fields was further investigated by driving HPAC with 1) its own alternate MC-SCIPUFF wind field model (the default SWIFT model has been used in all of the analyses thus far) and 2) the CALMET wind fields (Chang 2002). The dispersion model results were often significantly different, despite the fact that the same set of surface meteorological measurements was always used. The sensitivity study also revealed some subtle intermodel differences that led to large differences in the results. For example, even though the robust CALMET wind fields led to relatively few CALPUFF underpredictions (Fig. 3), the same CALMET wind fields coupled with HPAC often resulted in underpredictions at the surface. This is because CALPUFF uses only the horizontal east–west (u) and north–south (v) wind components created by CALMET and assumes the pollutant cloud centerline follows a constant height above terrain, whereas HPAC uses all three wind components produced by CALMET. The CALMET wind

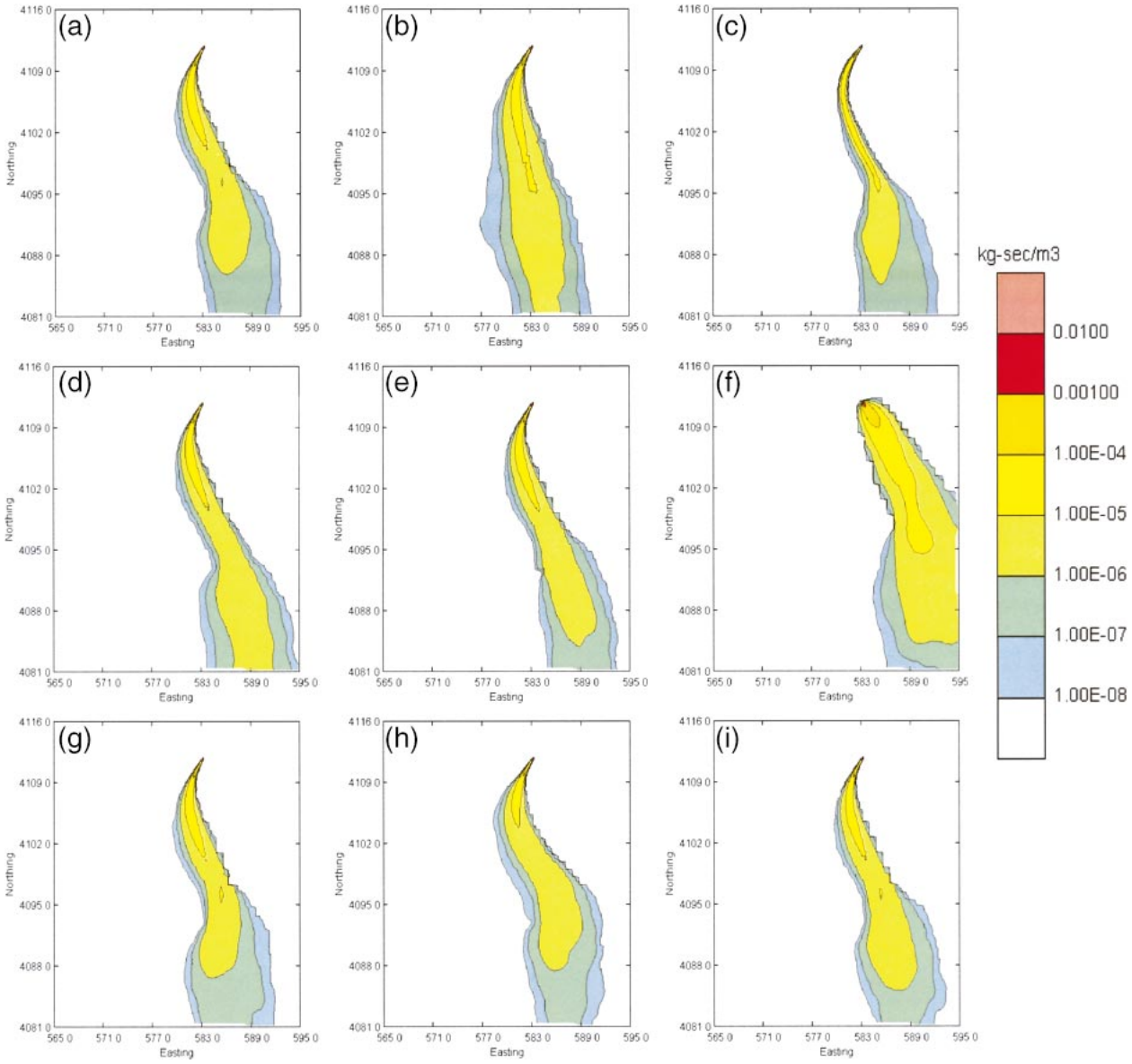

FIG. 7. Surface dosage (concentration integrated over 3 h after the release) contours predicted by HPAC for DP26 trial 3, 8 Nov 1996: (a) data from all eight MEDA stations used (see Fig. 1), (b) data from station M1 withheld, (c) data from station M2 withheld, (d) data from station M3 withheld, (e) data from station M6 withheld, (f) data from station M9 withheld, (g) data from station M10 withheld, (h) data from station M17 withheld, and (i) data from station M28 withheld.

fields often contained strong vertical motions (i.e., a few meters per second), which resulted from vertically extrapolating surface observations to blend with upper-air data on these mesoscale domains. As a result, the puff is likely to be transported aloft by HPAC instead of remaining close to the ground. These findings underline the importance of wind fields in influencing the dispersion model results.

6. Conclusions and discussion

The CALPUFF, HPAC, and VLSTRACK model evaluation exercises using the DP26 (point sources) and OLAD (line sources) field data demonstrate the difficulties in simulating transport and dispersion at mesoscales in flat areas with surrounding mountains. The networks of surface wind monitors at both sites suggest much spatial variability; for example, with root-mean-square differences in 10-m, 1-h-averaged wind speeds and directions of up to 2 m s^{-1} and 70°, respectively, for OLAD. The model comparisons were found to be strongly influenced by the diagnostic wind model that was used to generate gridded wind fields from the observed winds. Limited sensitivity studies also showed that, despite the relatively dense network of surface wind instruments, diagnostic wind models are not al-

ways able to reproduce the spatial variability originally present in all observations when the data from just one station were withheld.

Fortunately, there was an extensive set of concentration observations made at both field sites, and the instantaneous tracer gas releases were straightforward and well defined. Extensive comparisons of model simulations for many interesting characteristics of the data, such as peak 1-min-average concentrations and puff arrival times at sampling lines, could not be performed because CALPUFF does not allow averaging times of less than 1 h. To compare the three models on an equitable basis, the major emphasis had to be on long-term averages or dosages, such as the single maximum dosage at individual monitors on each line in this study.

In general, at the DP26 site, CALPUFF and HPAC were found to have similar performance, with overall mean biases not much different from zero and within a factor of 2 of each other. Occasional large (one–two orders of magnitude) errors occurred and were due to the predicted puff missing the sampling lines because of problems with the derived wind fields. The VLSTRACK performance at the DP26 site was slightly worse than the CALPUFF and HPAC performances. For example, the fraction of predictions within a factor of 2 of observations was about 50%–60% for CALPUFF and HPAC but was about 40% for VLSTRACK.

At the OLAD site, all three models underpredicted by a factor of 2–3, on average. FAC2 was about 50% for HPAC and about 25% for CALPUFF and VLSTRACK. HPAC was better able to match the absolute observed maximum, although the underprediction was still a factor of approximately 2.6.

The major difference between the two sites was that the models did not show as much of a mean bias at the DP26 site as at the OLAD site. There was a consistent tendency toward underpredictions, by a factor of 2, 3, or more, at the OLAD site. The reason for this difference at the OLAD site is unclear, although it is most likely related to the models' simulations of vertical dispersion and cloud advective speed. Moreover, many of the OLAD trials were conducted in the morning transition periods, which are more difficult to simulate.

Acknowledgments. This research has been sponsored by the Defense Threat Reduction Agency, with Major Thomas Smith, Mr. Ronald Meris, and Major Brian Beitler as technical contract representatives. The data files and experiment reports were provided by Christopher Biltoft of the U.S. Army's Dugway Proving Ground, and Thomas Watson of the National Oceanic and Atmospheric Administration's Idaho National Engineering Laboratory.

REFERENCES

Allwine, K. J., W. F. Dabberdt, and L. L. Simmons, 1998: Peer review of the CALMET/CALPUFF modeling system. Tech. Rep., KEVRIC Company, Inc., Durham, NC, EPA Contract No. 68-D-98-092, 40 pp. [Available online at http://www.epa.gov/ttn/scram/7thconf/calpuff/calpeer.pdf.]

Bauer, T. J., and R. L. Gibbs, 1998: Software user's manual for the Chemical/Biological Agent Vapor, Liquid, and Solid Tracking (VLSTRACK) computer model, version 3.0. NSWC Doc. NSWCDD/TR-98/62, Systems Research and Technology Department, Dahlgren Division, Naval Surface Warfare Center, Dahlgren, VA, 170 pp.

Bergin, M. S., G. S. Noblet, K. Petrini, J. R. Dhieux, J. B. Milford, and R. A. Harley, 1999: Formal uncertainty analysis of a Lagrangian photochemical air pollution model. *Environ. Sci. Technol.,* **33,** 1116–1126.

Biltoft, C. A., 1998: Dipole Pride 26: Phase II of Defense Special Weapons Agency transport and dispersion model validation. DPG Doc. DPG-FR-98-001, prepared for Defense Threat Reduction Agency by Meteorology and Obscurants Divisions, West Desert Test Center, U.S. Army Dugway Proving Ground, Dugway, UT, 77 pp.

——, S. D. Turley, T. B. Watson, G. H. Crescenti, and R. G. Carter, 1999: Over-Land Along-Wind Dispersion (OLAD) test summary and analysis. WDTC Doc. WDTC/JCP-99/048, Meteorology and Obscurants Divisions, West Desert Test Center, U.S. Army Dugway Proving Ground, Dugway, UT, 51 pp.

Chang, J. C., 2002: Uncertainty and sensitivity of dispersion model results to meteorological inputs: Two case studies. *Quantitative Methods for Current Environmental Issues*, C. W. Anderson et al., Eds., Springer-Verlag, 167–203.

——, P. Franzese, and S. R. Hanna, 1999: Evaluation of CALPUFF, HPAC, and VLSTRACK with the Dipole Pride 26 field data. Tech. Rep., School of Computational Sciences, George Mason University, Fairfax, VA, 46 pp. [Available online at http://squall.scs.gmu.edu/camp/publica.html.]

——, ——, and ——, 2000: Evaluation of CALPUFF, HPAC, and VLSTRACK with the Dipole Pride 26 field data. Preprints, *11th Joint Conf. on the Applications of Air Pollution Meteorology with the A&WMA,* Long Beach, CA, Amer. Meteor. Soc., 340–345.

——, K. Chayantrakom, and S. R. Hanna, 2001: Evaluation of CALPUFF, HPAC, and VLSTRACK with the Over-Land Along-Wind Dispersion (OLAD) field data. Tech. Rep., School of Computational Sciences, George Mason University, Fairfax, VA, 82 pp. [Available online at http://squall.scs.gmu.edu/camp/publica.html.]

DTRA, 1999: HPAC hazard prediction and assessment capability, version 3.2. Defense Threat Reduction Agency, Alexandria, VA, 406 pp.

Efron, B., 1987: Better bootstrap confidence intervals. *J. Amer. Stat. Assoc.,* **82,** 171–185.

Hanna, S. R., 1989: Confidence limits for air quality model evaluations, as estimated by bootstrap and jackknife resampling methods. *Atmos. Environ.,* **23,** 1385–1398.

——, D. G. Strimaitis, and J. C. Chang, 1991: Hazard response modeling uncertainty (a quantitative method). Vol. 1, User's guide for software for evaluating hazardous gas dispersion models. Tech. Rep. prepared for Engineering and Services Laboratory, Air Force Engineering and Services Center, Tyndall Air Force Base, and American Petroleum Institute, by Earth Tech, 58 pp.

——, J. C. Chang, and D. G. Strimaitis, 1993: Hazardous gas model evaluation with field observations. *Atmos. Environ.,* **27A,** 2265–2285.

Nappo, C. J., R. M. Eckman, K. S. Rao, J. A. Herwehe, and R. L. Gunter, 1998: Second Order Closure Integrated Puff (SCIPUFF) model verification and evaluation study. NOAA Air Resources Laboratory Tech. Memo. ERL ARL-227, 63 pp.

Pendergrass, W. R., C. J. Nappo, K. S. Rao, and R. M. Echmann, 1996: Technical review of the VLSTRACK dispersion model. NOAA Air Resources Laboratory, Tech. Memo. ERL/ARL-218, 39 pp.

Perdriel, S., J. Moussafir, and B. Carissimo, 1995: Note de principe

du code MINERVE, version 4.0 (MINERVE, version 4.0, theory manual). Tech. Rep. HE/33/95/008, Electricité de France (EDF), Chatou, France, 67 pp.

Scire, J. S., F. R. Robe, M. E. Fernau, and R. J. Yamartino, 2000a: A user's guide for the CALMET Meteorological Model (version 5.0). Tech. Rep., Earth Tech, Inc., Concord, MA, 332 pp. [Available online at http://www.src.com/calpuff/calpuffl.htm.]

——, D. G. Strimaitis, and R. J. Yamartino, 2000b: A user's guide for the CALPUFF Dispersion Model (version 5.0). Tech. Rep., Earth Tech, Inc., Concord, MA, 521 pp. [Available online at http://www.src.com/calpuff/calpuffl.htm.]

Sykes, R. I., S. F. Parker, D. S. Henn, C. P. Cerasoli, and L. P. Santos, 1998: PC-SCIPUFF version 1.2PD, technical documentation. ARAP Rep. 718, Titan Research and Technology Division, Titan Corp., Princeton, NJ, 172 pp. [Available online at http://www.titan.com/appliedtech/Pages/TRT/pages/scipuff/scipuff_files.htm.]

Watson, T. B., R. E. Keislar, B. Reese, D. H. George, and C. A. Biltoft, 1998: The Defense Special Weapons Agency Dipole Pride 26 field experiment. NOAA Air Resources Laboratory Tech. Memo. ERL ARL-225, 90 pp.

——, G. H. Crescenti, R. C. Johnson, B. R. Reese, R. G. Carter, S. D. Turley, B. Grim, and C. A. Biltoft, 2000: The Over-Land Along-Wind Dispersion (OLAD) field experiment. NOAA Air Resources Laboratory Tech. Memo. OAR ARL-235, 141 pp.