Evaluation of Lagrangian Particle Dispersion Models with Measurements from Controlled Tracer Releases

JENNIFER HEGARTY,* ROLAND R. DRAXLER,+ ARIEL F. STEIN,* JEROME BRIOULE,#,@ MARIKATE MOUNTAIN,* JANusz ELUSZKIEWICZ,* THOMAS NEHRKORN,* FONG NGAN,+,& AND ARLYN ANDREWS**

* Atmospheric and Environmental Research, Lexington, Massachusetts
+ NOAA/Air Resources Laboratory, College Park, Maryland
# Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado
@ Chemical Science Division, NOAA/Earth System Research Laboratory, Boulder, Colorado
& Cooperative Institute for Climate and Satellites, University of Maryland, College Park, College Park, Maryland
** Global Monitoring Division, NOAA/Earth System Research Laboratory, Boulder, Colorado

(Manuscript received 9 April 2013, in final form 26 July 2013)

ABSTRACT

Three widely used Lagrangian particle dispersion models (LPDMs)—the Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT), Stochastic Time-Inverted Lagrangian Transport (STILT), and Flexible Particle (FLEXPART) models—are evaluated with measurements from the controlled tracer-release experiments Cross-Appalachian Tracer Experiment (CAPTEX) and Across North America Tracer Experiment (ANATEX). The LPDMs are run forward in time driven by identical meteorological inputs from the North American Regional Reanalysis (NARR) and several configurations of the Weather Research and Forecasting (WRF) model, and the simulations of tracer concentrations are evaluated against the measurements with a ranking procedure derived from the combination of four statistical parameters. The statistical evaluation reveals that all three LPDMs have comparable skill in simulating the tracer plumes when driven by the same meteorological inputs, indicating that the differences in their formulations play a secondary role. Simulations with HYSPLIT and STILT demonstrate the benefit of using customized hourly WRF fields with 10-km horizontal grid spacing over the use of 3-hourly NARR fields with 32-km horizontal grid spacing. All three LPDMs perform better when the WRF wind fields in the planetary boundary layer are nudged to NARR, with FLEXPART benefiting the most. Case studies indicate that the nudging corrects an overestimate in plume transport speed possibly caused by a positive bias in WRF wind speeds near the surface. All three LPDMs also benefit from the use of time-averaged velocity and convective mass flux fields generated by WRF, but the impact on HYSPLIT and STILT is much greater than on FLEXPART. STILT backward runs perform as well as their forward counterparts, demonstrating this model’s reversibility and its suitability for application to inverse flux estimates.

1. Introduction

Lagrangian particle dispersion models (LPDMs) are a powerful tool for modeling atmospheric transport (Lin et al. 2012). An LPDM tracks a set of tracer particles either forward in time from a source region or backward in time from a measurement location (receptor). Each particle is transported by advective wind fields (obtained from a meteorological model) plus an unresolved turbulent (subgrid) velocity component. Lagrangian models have minimal numerical diffusion (e.g., Eluszkiewicz et al. 2000) and can preserve tracer gradients at smaller spatial scales than can Eulerian models because the latter smooth tracer concentrations to the resolution of the meteorological model grid. The inclusion of both the mean (resolved) and stochastic (unresolved) wind components (Uliasz 1994) sets the LPDMs apart from conventional trajectory models that employ only mean winds and thus cannot properly simulate dispersion or surface interactions (Stohl 1998). Through careful utilization of outputs from numerical weather prediction...
models (Uliasz 1993), meteorological realism and mass conservation can be achieved (Nehrkorn et al. 2010; Brioude et al. 2012a).

Forward-in-time LPDM computations are a natural choice for examining the dispersion of tracers from known source regions. An example of forward computations is a release of radioactive materials (e.g., Wotawa et al. 2006) such as that resulting from the accident at Japan’s Fukushima nuclear reactor. The spread of volcanic ash from Iceland’s Eyjafjallajökull eruption (e.g., Stohl et al. 2011), which caused massive disruptions of European air traffic, is another example with direct societal impacts.

When LPDMs are run in the backward-in-time mode from receptors, they can provide the sensitivity of the modeled concentration to upwind sources (Seibert and Frank 2004). These source–receptor sensitivities represent the adjoint of the mean and turbulent transport fields. When applied to surface sources they are often called “footprints” (e.g., Kljun et al. 2002) and are calculated by counting the number of particles in a surface-influenced region, defined as a fraction of the estimated planetary boundary layer (PBL) height, and the time spent in the region (e.g., Lin et al. 2003; Brioude et al. 2011). When multiplied by an assumed field of surface flux, the footprint gives the associated contribution to the mixing ratio measured at the receptor. In Eulerian chemical transport models (CTMs), computation of the corresponding sensitivities requires either a massive number of perturbation forward-transport computations or the coding and maintenance of the adjoint of the CTM, which is a difficult process, even with modern tools (Errico 1997; Henze et al. 2007)—in particular, for the rapidly evolving research codes that are employed in the Earth sciences. For this reason, LPDM-based footprint computations are an enabling tool for top-down (inverse) estimates of greenhouse gas (GHG) fluxes, particularly at policy-relevant regional and local scales (e.g., Kort et al. 2008; Zhao et al. 2009; Schuh et al. 2010; Lauvaux et al. 2012; Gourdji et al. 2012; Miller et al. 2012; McKain et al. 2012; Brioude et al. 2012b).

In recent decades a number of LPDMs have been developed that use a variety of meteorological inputs and employ different methods for calculating atmospheric transport and dispersion. In this study, we compare three widely used LPDMs—the Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) model (Draxler and Hess 1997, 1998; Draxler 1999), the Stochastic Time-Inverted Lagrangian Transport (STILT) model (Lin et al. 2003), and the Flexible Particle (FLEXPART) model (Stohl et al. 1998, 2005)—with controlled tracer measurements from the Cross-Appalachian Tracer Experiment (CAPTEX; Ferber et al. 1986) and the Across North America Tracer Experiment (ANATEX; Draxler and Heffter 1989). Although other tracers have been used for validation of transport models (e.g., Jacob et al. 1997), controlled releases for which the emissions and resulting concentrations are accurately known arguably constitute the “gold standard” for evaluating transport models. Data from a single CAPTEX release have been used in several model-evaluation studies that used high-resolution (12 and 4 km) meteorological data (Deng et al. 2004; Lee et al. 2009; Deng and Stauffer 2006), and both ANATEX and CAPTEX data for multiple releases were evaluated with FLEXPART by Stohl et al. (1998) with 1° global meteorological fields. In the evaluation that is presented here, each LPDM is driven by hourly meteorological inputs from the Weather Research and Forecasting (WRF) model (Skamarock and Klemp 2008) with 10-km horizontal grid spacing. The evaluation includes statistical assessments of each LPDM to quantify its ability to simulate the observed tracer concentrations and its sensitivity to different WRF inputs, with particular regard to the PBL variables and the use of time-averaged versus instantaneous velocity fields. The overall objectives of this study are thus to identify any substantial differences among the LPDMs and how the specific meteorological inputs from WRF influence the performance of each LPDM. The applicability of the statistical assessment to inverse methods is also addressed through a reversibility analysis of the STILT model.

2. Experimental data

CAPTEX consisted of six 3-h perfluoromonomethylocyclohexane (PMCH) tracer releases from 18 September to 29 October 1983. Four releases, hereinafter referred to as CAPTEX-1 through CAPTEX-4, were from Dayton, Ohio, and two (CAPTEX-5 and CAPTEX-7) were from Sudbury, Ontario, Canada. One additional short (30 min) release from Dayton was not included in the sampling data. Samples of the released tracers were collected from 19 September through 30 October at 84 sites, 300–800 km from the source, as 3- and 6-h averages for periods of 48–60 h after each release. The releases from Dayton were during the afternoons to allow for good vertical mixing, and the releases from Sudbury occurred at nights after frontal passages when strong wind conditions also caused significant vertical mixing. All releases were from ground level. Concentrations were reported in volume units above background and were converted to mass units according to the molecular weight of the tracer to simplify comparison with model results.

ANATEX consisted of 66 perfluorocarbon tracer releases, 33 each from two different locations, from 5 January to 26 March 1987. The releases occurred at 2.5-day intervals so that release times alternated...
between afternoon and nighttime. Air samples were collected from 5 January through 29 March and were averaged over 24-h periods at 75 sites located in the eastern United States and southeastern Canada. Perfluorotrimethylcyclohexane (PTCH) was released from Glasgow, Montana (GGW), and perfluorodimethylcyclohexane (PDCH) and PMCH were released from St. Cloud, Minnesota (STC). Because the PMCH releases from STC are coincident with the PDCH releases and hence do not provide any “meteorologically independent” data, we only include the PDCH data in our analysis. Releases at both locations were from the rooftops of one-story buildings through pipes that extended 2 m above the rooftops. The measured concentrations were reported as above background and were converted to mass units. For our analysis, we only consider data collected during 5–16 January, which includes tracers from the first 10 releases (five from each site). We included only the first five ANATEX release periods to be of comparable duration to the CAPTEX period and to be representative of winter conditions in contrast with the more summer-like conditions that prevailed during the first four CAPTEX releases.

3. WRF model configuration

The goal of this study is to establish whether differences in plume transport among the models are due to differences in the LPDM formulation or to the way the meteorological driver fields are generated and applied. When the initial model-evaluation studies using CAPTEX and ANATEX data were conducted (e.g., Stohl et al. 1998), only global meteorological data were available. Mesoscale models have been more recently used to generate the driver fields (e.g., Deng et al. 2004) for dispersion simulations of controlled tracer-release experiments.

Mesoscale models have many options in how the meteorological fields can be calculated, and these selections influence the subsequent dispersion calculations. In this work, the three LPDMs were driven by meteorological fields created by the Advanced Research version of the WRF model (ARW, version 3.2.1). The initial and boundary conditions for WRF were provided by the North American Regional Reanalysis (NARR; Mesinger et al. 2006) available at 32-km resolution. WRF was configured with two nest levels of 30- and 10-km resolution and one-way boundary conditions between the two nest levels (Fig. 1). The vertical grid had 43 layers, with the lowest layer being approximately 33 m thick. The WRF physics options included Lin et al. (1983) microphysics, Rapid Radiative Transfer Model–Global Climate Model Applications version (RRTM-G) longwave and shortwave radiation (Iacono et al. 2008), the National Centers for Environmental Prediction–Oregon State University–U.S. Air Force–National Weather Service Hydrologic Development Office (“Noah”) land surface model (Mitchell et al. 2005), the Yonsei University PBL scheme (Hong et al. 2006), and the cumulus parameterization of Grell and Devenyi (2002).

The WRF simulations were restarted 0000 UTC for each day, and 3D grid nudging of both mass (temperature and moisture) and wind fields to NARR was applied in both domains every 3 h. To eliminate spinup errors associated with the daily restarts, WRF was run for 30 h and the first 6 h of each run were replaced with the last six overlap hours of the previous day’s run. Nudging of winds in the free troposphere was done for all simulations, but nudging of winds within the PBL was optionally included. In the discussion of results, such runs with PBL wind nudging are denoted pbl = 1 whereas runs without PBL wind nudging are denoted pbl = 0.

Fig. 1. The WRF domains for (a) CAPTEX and (b) ANATEX. The horizontal grid spacing is 30 km for d01 and 10 km for d02.
Transport models may use WRF standard outputs as is; WRF can also produce additional outputs implemented specifically for the benefit of transport simulations (Nehrkorn et al. 2010). These include 1) time-averaged, mass-coupled horizontal and vertical velocities (using the WRF terrain-following vertical coordinate $\eta$) to improve mass conservation and the representation of wind variability and 2) time-averaged mass fluxes from the WRF moist convective parameterization scheme. In our experiments, we drove the LPDMs with both the time-averaged fields (denoted as avg in results) and the instantaneous fields (denoted as inst) in separate sensitivity tests. The instantaneous fields were not mass coupled, used the geometric vertical velocity $w$, and did not make use of the convective mass fluxes.

4. Lagrangian particle dispersion models

For the evaluations, the three LPDMs were run forward in time, driven with identical meteorological fields provided by WRF (in several computational and/or output configurations). In addition, baseline calculations with HYSPLIT and STILT were performed for each experiment using only the NARR meteorological data.\(^1\) When an LPDM is run in the forward mode, a number of pollutant particles are released at the source location and passively follow the wind. The particle trajectory is the integration of the particle’s position vector in space and time using a velocity field composed of a mean component from the meteorological model and a turbulent component. The turbulent component of the motion defines the dispersion of the pollutant cloud and is computed by adding a random component to the mean advection velocity in each of the three-dimensional wind-component directions. The vertical and horizontal turbulence is computed from the various meteorological inputs by using parameterization schemes that differ in each LPDM, as will be discussed in the sections that follow. Models employ different methods to compute air concentrations, from simply summing the mass of particles as they pass over a concentration grid cell and dividing the result by the cell’s volume to applying a technique to distribute the mass of each particle between adjacent cells following a predefined functional distribution kernel. For our study, the calculated concentrations were averaged over 6 and 24 h for CAPTEX and ANATEX, respectively, to match the tracer-sampling periods. A more detailed description of each LPDM that focuses on their differences is provided in sections 4a–c.

\(^1\)The FLEXPART version used here could not be driven with NARR.

a. HYSPLIT

In HYSPLIT, the calculation of particle dispersion is carried out on the native WRF horizontal grid (Lambert conformal in this study). Given that HYSPLIT uses an internal terrain-following vertical coordinate system, however, the WRF meteorological profiles at each horizontal grid point are linearly interpolated to this internal vertical grid. The vertical grid is defined with the first level at ~10 m above ground level (AGL) and with decreasing vertical resolution with height.

Random horizontal and vertical turbulence components with a Gaussian distribution are added to the mean trajectory. The random components are scaled by the turbulent velocity standard deviation computed from the Lagrangian time scale of turbulence and the diffusion coefficients estimated on the basis of inputs from the driving meteorological model (Hanna 1982; Draxler and Hess 1998; Kanthar and Clayson 2000). Whereas Draxler and Hess (1998) used a single constant value for the vertical Lagrangian time scale (usually 200 s), in this application we implemented a two-value vertical Lagrangian time scale to account for differences in atmospheric stability conditions (Hanna 1982): 200 s for unstable conditions and 5 s for stable conditions. The drift-correction scheme of Wilson et al. (1983) is used to preserve well-mixed particle distributions.

The overall integration time step was 1 min (although the integration time step for the dispersion can be as low as 1 s), and the 3-h tracer releases were represented by 50 000 particles in CAPTEX and 25 000 particles in ANATEX. The fewer particles released for ANATEX were compensated for by the longer averaging time. Furthermore, because two ANATEX tracer releases could be on the computational domain at the same time, the same number of total particles was followed as for each CAPTEX release. A limited number of tests indicated little sensitivity of the results to the number of particles released beyond 25 000 particles. The resolution of the output concentration grid was set to 0.25° in both latitude and longitude and 100 m in the vertical direction.

b. STILT

STILT is built upon the HYSPLIT software and uses the same mean advection scheme but a different turbulence module (Lin et al. 2003). The STILT configuration for the tracer simulations is identical to HYSPLIT in terms of the concentration grid, number of particles released in the forward runs, and many of the remaining parameters. STILT contains several distinct features not available in HYSPLIT, however. One of them is a reflection/transmission scheme for Gaussian turbulence
(Thomson et al. 1997) that preserves well-mixed distributions of particles moving across interfaces between step changes in turbulence parameters. Lin et al. (2003) demonstrated that such a scheme is necessary to prevent accumulation of particles in low-turbulence regions of the strongly inhomogeneous environment of the PBL where a simple drift correction may not work. Another feature is the ability of STILT to directly use the convective mass fluxes generated in WRF by the Grell–Devenyi cumulus scheme in the dispersion of particles, incorporating the vertical profiles of up- and downdrafts and entrainment and detrainment fluxes between the environment and convective cells (Nehrkorn et al. 2010).

c. **FLEXPART**

We use a version of FLEXPART modified for use with WRF (Fast and Easter 2006; Brioude et al. 2012a). It uses the native horizontal grid of WRF and interpolates the WRF vertical levels onto an internal terrain-following vertical coordinate for the computation of particle transport and dispersion. The concentrations were output on a 25 km × 25 km horizontal grid using the same projection as WRF. The WRF-compatible version of FLEXPART has the option of outputting regular latitude–longitude concentration grids similar to those in HYSPLIT and STILT, but we found that using the WRF projection produced better results, and only those results are used in the statistical evaluation reported below.2 The horizontal grid spacing for the FLEXPART concentration output was selected to match closely the 0.25° latitude–longitude grid of HYSPLIT and STILT. As with HYSPLIT and STILT, the concentrations were calculated over the lowest 100 m AGL. The scheme of Hanna (1982) is used to represent turbulent mixing in the PBL, with PBL height and frictional velocity parameters retrieved directly from the WRF output. The integration time step is calculated dynamically depending on the values of the Lagrangian time scale, vertical velocity, and mixing height, with upper limits of 90 and 18 s for horizontal and vertical mixing, respectively. The 3-h tracer releases were represented by 100 000 particles.

5. **Statistical evaluation**

Procedures for evaluating dispersion models have a long history (Fox 1984; Hanna 1989, 1993; Chang and Hanna 2004). The problem eludes simple solutions because of the variability in sensitivity of statistical parameters to small mismatches between predicted and measured concentrations paired in space and time. The dispersion-model evaluation protocol used here follows Mosca et al. (1998) and Stohl et al. (1998), but only four statistical parameters were selected from their broad list to represent well-defined evaluation categories: the correlation coefficient $R$, the fractional (normalized) bias (FB), the figure-of-merit in space (FMS) to measure the modeled–observed tracer-plume overlap (defined here with a threshold value of 0 pg m$^{-3}$), and the Kolmogorov–Smirnov parameter (KSP) to represent the similarity between modeled and observed concentration distributions. See Mosca et al. (1998) for the computational details.

Both Mosca et al. (1998) and Stohl et al. (1998) recognized the problem presented by uncertainties in “near background” measurements that affect statistical parameters that are sensitive to small variations in the measurement values, such as ratios between measured and calculated concentrations. For a quick evaluation, it is desirable to have a single parameter that represents the overall model performance. Stohl et al. (1998) found that the ratio-based statistics are indeed very sensitive to measurement errors whereas the correlation coefficient is more robust. Chang and Hanna (2004) are more critical of the correlation coefficient because of its sensitivity to high concentrations. This sensitivity can be partially mitigated by employing a bootstrapping technique (Efron 1987) to calculate the average correlation. This was not found to be an issue with the experimental data used in our analysis, however, because the large downwind distances and longer averaging times reduced the range of concentration values. Chang and Hanna (2004) also summarized attempts to define a single model evaluation parameter, such as ranking models by each statistic and then ordering by the total rank. In our study, the ranking method of Draxler (2006) was implemented by giving equal weight to the normalized (0–1) sum of $R^2$ using the bootstrap technique (hereinafter referred to as just $R^2$), FB, FMS, and KSP, such that the total model rank [Eq. (1)] would range from 0 to 4, worst to best:

$$\text{Rank} = R^2 + 1 - |\text{FB}/2| + \text{FMS}/100 + (1 - \text{KSP}/100)$$ (1)

6. **Results**

This section summarizes the statistical evaluation of the LPDM runs for six different CAPTEX experiments and two ANATEX experiments (GGW and STC), each ANATEX “experiment” consisting of five releases. The combination of multiple ANATEX releases into a single
experiment for this evaluation resulted in a comparable number of measured–calculated data pairs for analysis between the 6-h sample duration during CAPTEX and the 24-h sample duration during ANATEX (354, 365, 346, 299, 306, 183, 562, and 556 for the six CAPTEX experiments and two ANATEX experiments, respectively).

The normalized statistical parameters composing the ranks for each of the experiments are represented graphically in Fig. 2. From this figure it is clear that a single statistic such as the correlation coefficient, although generally helpful in ranking models, is not always useful on its own. For example, for CAPTEX-4 the correlation coefficient is near zero for all models because of the simulated trajectory of the plume center being too far south and provides little information on the differences in model skill that are indicated by the other statistical parameters. Therefore, while we recognize that there are limitations in any one parameter, we select the rank as a convenient summary of the modeling results.

The HYSPLIT, STILT, and FLEXPART statistical ranks are presented in Tables 1, 2, and 3, respectively. Except for the column labeled “NARR only,” the ranks in each table were generated with identical meteorological inputs provided by WRF in the configurations discussed in section 3. The NARR-only column shows ranks for HYSPLIT and STILT driven directly by the NARR fields, providing a baseline performance metric to evaluate the impact of the higher temporal and spatial resolution afforded by WRF.

The statistical ranks indicate that in all eight experiments there is at least one and usually several WRF-based simulations that perform better than their NARR counterparts, and this result is true for both HYSPLIT and STILT. This result confirms that the higher temporal (1 vs 3 h) and spatial (10 vs 32 km) resolution available from WRF provides substantial benefit to the LPDM simulations. For the WRF-based runs, it is clear that the best configuration for all three LPDMs included nudging of the PBL winds (pbl = 1) and time-averaged fields (avrg). For this WRF configuration there is little difference between the LPDMs, as shown by the fact that the average ranks over all experiments (shown in the last row of the tables) differ by less than 0.1. This fact indicates that each of the LPDMs has the about the same skill level if provided with identical meteorological inputs. There are some notable differences among the LPDMs for other WRF configurations, however, and even for the same LPDM using different WRF configurations that indicate various model sensitivities to errors in meteorological inputs.

With one exception, the nudging of winds within the PBL has either a substantial positive impact or a small negative impact on the ranks (the negative impact is...
limited to HYSPLIT and STILT and occurs mostly when instantaneous fields are employed, which, as discussed below, result in inferior results when compared with the time-averaged fields). The one exception is ANATEX-STC, when the PBL wind nudging lowers the ranks by about 0.3 for all three LPDMs. This case is also unique in that the NARR-based runs for HYSPLIT and STILT are generally better than the WRF-driven runs. The cause of this outlier behavior is likely related to the spacing of the sampling network (this is further discussed in section 7). We consequently believe that this experiment should be given less weight in our evaluation and does not invalidate the overall conclusions.

The FLEXPART model shows the greatest positive impact of PBL wind nudging, with average ranks improving by 0.21 and seven (avrg) and six (inst) of eight experiments improving when it is enabled (Table 3). For HYSPLIT and STILT, the impact is smaller but still consistent, with average ranks improving by 0.08 and 0.13, respectively, and five and six of eight experiments improving, respectively, when time-averaged fields are used. The impact of PBL wind nudging for HYSPLIT and STILT is not as consistently positive when instantaneous fields are used, as average ranks only improve by 0.05 for HYSPLIT and actually decrease by 0.01 for STILT, with four HYSPLIT and six STILT experiments showing improvement. Despite this overall lesser impact, several experiments still show a substantial positive impact of PBL nudging even when instantaneous fields are used. For example, for CAPTEX-2, the experiment simulated best by all LPDMs, PBL wind nudging improves the avrg (inst) ranks by 0.55 (0.29) for HYSPLIT and by 0.50 (0.15) for STILT; while these improvements are smaller than the corresponding values of 0.70 (0.63) for FLEXPART, they are still substantial. Both HYSPLIT and STILT have higher average ranks than FLEXPART, however, suggesting there may be a limit to the improvement that can be achieved by PBL wind nudging and time averaging.

The inclusion of nudging of the PBL winds in WRF seems to have had the most noticeable impact on the LPDM simulations. In our initial design phase of the experiments, we tested several other configurations, including continuous runs versus daily restarts, one-way versus two-way nesting, observational nudging, and grid nudging above the PBL in the outer domain versus in both domains for HYSPLIT and STILT simulations of the CAPTEX cases. The results of all of these runs were similar, however, and it was only with the inclusion of PBL wind nudging that substantial, generally positive, impacts were achieved. Although grid nudging of winds at all vertical levels was recommended for previous generations of mesoscale models at coarser resolution (Stauffer and Seaman 1990; Stauffer et al. 1991), our initial hypothesis was that, given the relatively high spatial resolution of our inner domain, grid

### Table 1. Rank results from the HYSPLIT model evaluation.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>NARR only Avrg</th>
<th>Pbl = 1 Avrg</th>
<th>Pbl = 0 Avrg</th>
<th>Inst Avrg</th>
<th>Pbl = 1 Inst</th>
<th>Pbl = 0 Inst</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAPTEX-1</td>
<td>2.04 2.40 2.43</td>
<td>2.48 2.62</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPTEX-2</td>
<td>2.53 3.47 2.92</td>
<td>3.25 2.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPTEX-3</td>
<td>1.74 1.83 1.96</td>
<td>1.89 1.86</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPTEX-4</td>
<td>1.99 2.23 2.14</td>
<td>2.29 2.23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPTEX-5</td>
<td>2.30 2.66 2.37</td>
<td>2.56 2.31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPTEX-7</td>
<td>2.62 2.94 2.75</td>
<td>2.40 2.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANATEX-GGW</td>
<td>2.01 2.43 2.36</td>
<td>2.06 2.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANATEX-STC</td>
<td>1.80 1.40 1.76</td>
<td>1.40 1.70</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg</td>
<td>2.13 2.42 2.34</td>
<td>2.29 2.29</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 2. As in Table 1, but for the STILT model.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>NARR only Avrg</th>
<th>Pbl = 1 Avrg</th>
<th>Pbl = 0 Avrg</th>
<th>Inst Avrg</th>
<th>Pbl = 1 Inst</th>
<th>Pbl = 0 Inst</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAPTEX-1</td>
<td>2.24 2.43 2.48</td>
<td>2.42 2.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPTEX-2</td>
<td>2.59 3.27 2.77</td>
<td>2.99 2.84</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPTEX-3</td>
<td>1.72 1.86 1.81</td>
<td>1.80 1.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPTEX-4</td>
<td>1.98 2.33 2.04</td>
<td>2.19 2.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPTEX-5</td>
<td>1.93 2.73 2.43</td>
<td>2.54 2.31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPTEX-7</td>
<td>2.58 2.61 2.52</td>
<td>2.29 2.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANATEX-GGW</td>
<td>2.14 2.34 2.13</td>
<td>2.35 2.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANATEX-STC</td>
<td>1.77 1.57 1.90</td>
<td>1.54 1.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg</td>
<td>2.12 2.39 2.26</td>
<td>2.26 2.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 3. As in Table 1, but for the FLEXPART model. Note that NARR-only results are not available for the WRF version of the FLEXPART model.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Avrg</th>
<th>Pbl = 1 Avrg</th>
<th>Pbl = 0 Avrg</th>
<th>Inst</th>
<th>Pbl = 1 Inst</th>
<th>Pbl = 0 Inst</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAPTEX-1</td>
<td>2.42</td>
<td>2.12</td>
<td>2.44</td>
<td>2.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPTEX-2</td>
<td>3.17</td>
<td>2.47</td>
<td>3.13</td>
<td>2.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPTEX-3</td>
<td>1.91</td>
<td>1.80</td>
<td>1.90</td>
<td>1.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPTEX-4</td>
<td>1.95</td>
<td>1.93</td>
<td>1.95</td>
<td>1.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPTEX-5</td>
<td>2.54</td>
<td>2.09</td>
<td>2.41</td>
<td>2.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPTEX-7</td>
<td>3.14</td>
<td>2.85</td>
<td>3.17</td>
<td>2.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANATEX-GGW</td>
<td>2.24</td>
<td>2.07</td>
<td>2.23</td>
<td>2.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANATEX-STC</td>
<td>1.30</td>
<td>1.66</td>
<td>1.31</td>
<td>1.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg</td>
<td>2.33</td>
<td>2.12</td>
<td>2.32</td>
<td>2.11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
nudging should only be applied above the PBL to allow the presumably more advanced WRF PBL schemes to develop mesoscale features in response to the influences of the resolved finescale terrain and land-use features without the artificial damping effects of nudging to a comparatively coarse analysis. From our results it appears that the advantages of nudging to control error growth in WRF outweigh the possible damping effects, however. As discussed in section 7, nudging helps to reduce larger-scale plume transport errors that may result from a speed bias in the WRF PBL winds.

Time-averaged fields are clearly beneficial for HYSPLIT and STILT, producing an average rank improvement over the instantaneous velocities of 0.13 for both models and improvements in four and six of eight experiments, respectively in the pbl = 1 configuration. For FLEXPART, time-averaged fields are of lesser benefit, producing an average improvement in rank of only 0.01 and the same or improved ranks in five of eight experiments with or without PBL wind nudging. This result is consistent with Brioude et al. (2012a), who found that using time-averaged velocities reduced the bias and uncertainty by less than 5% for FLEXPART simulations over complex terrain. For some cases, however, the impact of time-averaged fields is substantial for all three LPDMs—for example, improving the ranks for the CAPTEX-5 pbl = 1 simulations by 0.10, 0.19, and 0.13 for HYSPLIT, STILT, and FLEXPART, respectively. As discussed in more detail in section 7, the use of time-averaged fields produces a more uniform simulation of the plume dispersion.

7. Case studies

In section 6, we presented several statistical conclusions regarding the overall LPDM performance. In this section, the results will be examined in more detail by analyzing model combinations that showed the largest differences in rank between the simulations for both regional and continental scales. We use concentration plots together with the statistics presented earlier to examine the impact of PBL wind nudging on regional-scale transport with CAPTEX-2 (Fig. 3) and on continental-scale transport using ANATEX-GGW (Fig. 4) and the contrasting case of ANATEX-STC (Fig. 5). For the impact of time-average flux fields, we examine CAPTEX-5 (Fig. 6). All of the concentration plots that are presented were generated using the same HYSPLIT plotting software to facilitate comparisons among the LPDMs. Since FLEXPART produces concentrations on the WRF projection grid, these outputs required interpolation to the regular latitude–longitude grids of HYSPLIT and STILT. A visual comparison of the FLEXPART interpolated concentration plots and noninterpolated concentration plots (generated by the FLEXPART plotting software) indicated that the interpolation did not substantially alter the dominant features of the tracer plumes.

The impact of PBL wind nudging is most striking for CAPTEX-2, with rank improvements of greater than 0.5 for all three LPDMs when time-averaged fields are used (Tables 1–3). The concentration plots for these runs for the 6-h sampling period that began 28 h after the start of the tracer release indicate that the plumes simulated with PBL wind nudging (Fig. 3) tend to move more slowly, intersecting samplers in northwestern Pennsylvania and western New York with very high tracer concentrations. In particular, the 1000 pg m$^{-3}$ contour extends southwest from the center of the plume and encompasses these high concentration measurements, whereas in pbl = 0 simulations it is noticeably farther north and east. Although there is an indication that the faster movement in the simulations without PBL wind nudging (Figs. 3e,f) provides slightly better results at the northeastern leading edge of the plume, there are more obvious improvements seen along the trailing western edge in the pbl = 1 simulation and, furthermore, the slightly better results at the northeastern edge are also offset by large overestimates of the concentrations near Albany, New York, and in southwestern Vermont. All three LPDMs show the same general movement of the plume centers, but the FLEXPART plumes are narrower with higher concentration values near the center. In addition, FLEXPART seems to be most affected by the lack of PBL wind nudging, as this simulation spreads high concentrations (>500 pg m$^{-3}$) too far east all the way to southeastern Vermont while almost completely missing two high observations in western Pennsylvania. PBL wind nudging appears to correct the faster eastward movement in all LPDM simulations and brings the FLEXPART simulation more in line with those of HYSPLIT and STILT.

The nudging of PBL winds has a similar impact on the continental-scale simulations, as illustrated by the concentration plots for the 24-h period beginning 1400 UTC 13 January 1987 for ANATEX-GGW (Fig. 4). It is clear from these plots that the eastern plume resulting from the release at 1700 UTC on 10 January 1987 moved more slowly eastward for all pbl = 1 simulations and produced a better match with observations near the center of the plume located at the extreme western tip of the Upper Peninsula of Michigan and in northeastern Wisconsin. In contrast, the western edge of the plume in the pbl = 0 simulations had already moved east of this location by this time period. The impact of PBL wind nudging on the western plume is less clear because the
sampling period begins only 9 h after the start of the related release at 0500 UTC 13 January 1987. Although the slightly greater eastern spread of the pbl = 1 plumes seems to contradict the findings from the eastern plume analysis, the greater southern spread along the main plume axis in the pbl = 0 simulations supports the conclusion that the simulated plumes generally move away faster from their source when PBL nudging is not applied.

Figure 4 also illustrates some substantial differences among the LPDM simulations. The eastern plume is more cohesive in the HYSPLIT and STILT simulations, whereas in the FLEXPART pbl = 1 simulation it splits into northern and southern branches. The split plume is hinted at in the observations, with lower values of 2–50 pg m$^{-3}$ in Ohio with higher values of 66 and 88 pg m$^{-3}$ in North Carolina, although the FLEXPART southern branch does not exactly match the location of these higher observations. As with CAPTEX-2, the FLEXPART plumes are narrower, with higher values concentrated toward the centers. For the eastern plume during this time period, this feature in the FLEXPART pbl = 1 simulation causes some concentration overestimates in south-central Canada and some underestimates in Wisconsin and the Upper Peninsula of Michigan (Fig. 4c). (The apparent abrupt end to the plume in south-central Canada in Fig. 4c is an artifact of fitting a square latitude–longitude interpolation grid within the WRF inner domain.) In contrast to the eastern plume, the more concentrated FLEXPART plumes better match the observations of the western plume. Thus, while there are noticeable differences between the LPDM simulations, it does not appear that one is consistently better than the others in capturing the plume dispersion for these cases of continental-scale transport.

As noted in the statistical evaluation presented in section 6, the ANATEX-STC simulations are degraded when PBL wind nudging is employed, in a clear contrast to other cases. A visual inspection of the concentration plots for ANATEX-STC indicates that the plumes in the pbl = 1 simulations agree with observations about as well
as those with pbl = 0, however. A possible explanation for this inconsistency and for why the ANATEX-STC ranks are so low may be the sparseness of the observations downwind of the release. As illustrated in Fig. 5 for the HYSPLIT ANATEX-STC simulation of the 24-h period beginning 1400 UTC 8 January 1987, the main observing arc stretching from Oklahoma to just north of Lake Superior was located a relatively short distance (400 km) downwind of St. Cloud. This proximity to the release point means that the plumes may have been relatively narrow and thus undetected by measurements (spaced roughly 170 km apart) when they intersected the arc. The narrow simulated plume shown in Fig. 5 has maximum concentrations of greater than 500 pg m\(^{-3}\) but passes between the two observation locations in northeastern Wisconsin and the Upper Peninsula of Michigan and does not spread out until after passing the arc. If the real plume followed the same path and had the same dimensions, it would be relatively undetected as it passed by the arc, which seems likely to have occurred given the relatively low values of the measured concentrations. We believe that, for this reason, the plumes were inadequately sampled for the ANATEX-STC releases and that their statistics are less reliable than for the other experiments.

Another key result regarding the use of meteorological data is that the calculations using time-averaged fields are generally better than those using instantaneous fields. This finding is not as universally applicable as the impact of PBL wind nudging, however, being limited to HYSPLIT and STILT. The only case for which it has a substantial positive effect for all three LPDMs is CAPTEX-5, illustrated in Fig. 6 for the 6-h sampling period that began 17 h after the release. In this experiment, the winds were strong, the plume was very narrow, and all simulated plumes essentially covered the same sampling stations. Calculations using time-averaged fields (Figs. 6a–c) generally provided smoother concentration patterns, more consistent with measurements than those simulated using instantaneous fields (Figs. 6d–f).

Although we have only shown selected cases, the graphical differences in these examples represent a

---

**Fig. 4.** As in Fig. 3, but for ANATEX-GGW average PTCH concentrations (pg m\(^{-3}\)) for the 24-h period starting 1400 UTC 13 Jan 1987 resulting from tracer releases at Glasgow (shown as an open star) from 1700 to 2000 UTC 10 Jan 1987 and from 0500 to 0800 UTC 13 Jan 1987.
consistent pattern across all experiments and the three LPDMs, as evidenced in Tables 1–3. In particular, the tendency for the plumes to be transported too quickly downwind in all simulations without PBL wind nudging suggests a bias in the near-surface wind speeds generated by WRF. Such a bias in WRF is noted by Jiménez and Dudhia (2012), who attribute it to an inadequate representation of the frictional drag imposed by subgrid-scale topographic features. It is interesting that nudging to NARR acts to correct this problem, suggesting that the NARR analysis does not have this bias.

8. STILT reversibility

LPDMs are often used in the backward mode (e.g., to support top-down estimates of GHG fluxes), and this fact raises the question of how well the forward results presented so far reflect their accuracy in this common mode of application. Since resources did not permit a comparison of all three models in forward and backward mode and because STILT is primarily used in backward mode, we selected this model for these comparisons. A similar forward–backward comparison for FLEXPART was recently performed by Brioude et al. (2012a), however, and the similarities between HYSPLIT and STILT reflected in the forward runs suggest that the forward–backward comparison for STILT should be representative of HYSPLIT.

For the backward-mode tests, we simulated CAPTEX and ANATEX measurements by convolving STILT-generated footprints with the known tracer fluxes and then averaging over the same time periods as the measurements (either 6 or 24 h). Fluxes were approximated on a 1° grid centered on the tracer release location, at hourly temporal resolution, over the 3-h tracer release time periods. The fluxes convolved with the STILT particle footprints yield a “delta” mixing ratio for each point in time along the particle trajectory, and these are summed over the entire particle trajectory to yield the total delta mixing ratio for that particle. The total delta mixing ratio for a single receptor is the average of the total delta mixing ratios of all particles emanating from that receptor. The simulated 6- or 24-h averaged measurement at each location was calculated as the average total delta mixing ratios of 6 or 24 receptors, defined at hourly increments throughout the measurement period. Note that the delta mixing ratios represent the actual mixing ratios because the background concentrations of the artificial tracers are assumed to be zero. They were converted to concentrations at the measurement sites, and the validation statistics were calculated in the same manner as in the forward calculations.

Only time-averaged fields were used in these backward runs, with particles released from heights of 50 and 10 m AGL. The 50-m height is the midpoint of the averaging layer for which the concentrations were calculated in the forward runs, allowing a direct comparison.

3 In the most recent WRF version (version 3.5), the surface-layer parameterization has been upgraded to correct this bias by adding a momentum sink.
and the 10-m height is closer to ground level at which the measurements were taken. As in most STILT-based footprint calculations (e.g., McKain et al. 2012), we employed 500 particles in the backward runs. For selected experiments, sensitivity runs using 5000 particles produced minimal changes in the results. Because of the computational expense of the ANATEX runs, we only simulated the Glasgow release, using 50-m receptor heights and the PBL wind nudging.

The ranks for the backward STILT runs shown in Table 4 indicate that the backward runs were as accurate as their forward counterparts. For the 50-m-AGL pbl = 1 case, the average backward rank for the CAPTEX runs (2.65) is actually slightly greater (by 0.11) than its forward counterpart (2.54), whereas for ANATEX-GGW it is slightly lower (by 0.11). For the best overall case, CAPTEX-2, the backward and forward ranks are within 0.01 of each other. Furthermore, PBL wind nudging exerts similar positive influence on the simulations, increasing the average ranks by 0.21 and 0.20 for the backward and forward CAPTEX runs, respectively. The ranks for the 10-m-AGL runs are only slightly lower than for 50 m (by an average of 0.04 when PBL wind nudging is employed), reflecting the ability of the LPDM to work well near the surface. Overall, these backward STILT runs demonstrate the reversibility and good performance of this LPDM in the receptor-oriented mode.

9. Summary

We evaluated three widely used LPDMs (HYSPLIT, STILT, and FLEXPART) with controlled tracer release data from the CAPTEX and ANATEX experiments. The LPDMs were driven by identical meteorological fields, enabling the differences attributable to the LPDM formulation and to their sensitivities to the various configurations of input data to be separated. With the exception of one case in which sampling artifacts may have degraded the reliability of our statistical analysis, several robust conclusions have emerged from this study. In particular, all three LPDMs had comparable skill in simulating the tracer plumes when driven with modern
meteorological inputs (including NARR and several configurations of WRF), indicating that differences in their formulations play a secondary role. Our simulations demonstrate the benefit of employing 10-km customized WRF runs over NARR at 32-km resolution. The LPDMs exhibited significant sensitivity to the WRF configurations. Perhaps most striking is that all three LPDMs performed substantially better when the WRF wind fields within the PBL were nudged to NARR, with FLEXPART benefitting most from this nudging. The PBL wind nudging appears to correct an overestimate of the plume transport speed, possibly caused by a positive wind speed bias near the surface. On the basis of only this study we cannot generalize that nudging of PBL winds is necessary for all LPDMs, but it should be a consideration for future LPDM studies using WRF. Another consistent finding from our study is that all three LPDMs benefited from the use of time-averaged velocity and convective mass flux fields generated by WRF, but in this case the impact on HYSPLIT and STILT was greater than on FLEXPART. The STILT backward runs performed as well as their forward counterparts, demonstrating this model’s reversibility and good performance in its most common application to inverse flux estimates.

Transport simulations offer an opportunity to evaluate meteorological models that extends beyond conventional verification that involves comparisons of the model-predicted meteorological variables with observational data. These typical comparisons are fixed in space and time, whereas the comparison with tracer data represents a time and space integration of the meteorological field in which small differences accumulate to provide a more integrated comparison. Our study provides a detailed evaluation of the current state of LPDMs using historical tracer-release data and thus provides a limited benchmark of their performance as they are used in a growing range of applications, including inverse GHG flux estimates, air quality, and the dispersion of toxic airborne contaminants.

Acknowledgments. We gratefully acknowledge support for this study from the National Science Foundation (Grants ATM-0836153 and ATM-1207983) and NOAA (Grants WE133R11SE2312 and RA-133R-12-SE-2198). The WRF-STILT development at AER has been supported by NSF, NASA, NOAA, and the U.S. intelligence community. The field data and the original reports are available through the Data Archive of Tracer Experiments and Meteorology (DATEM; http://www.arl.noaa.gov/DATEM.php), statistical and graphical outputs from the LPDM simulations are available at ftp://arlftp.arlhq.noaa.gov/117/results, and the WRF meteorological fields are available at http://www.arl.noaa.gov/DATEM_WRF-ARW.php. Please contact Roland Draxler for more information about the DATEM database.

### REFERENCES


