Uncertainty in Lagrangian pollutant transport simulations due to meteorological uncertainty at mesoscale

Wayne M. Angevine[1,2], Jerome Brioude[1,2,3], Stuart McKeen[1,2], John S. Holloway[1,2]

[1]Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado, Boulder, Colorado, USA

[2]NOAA Earth System Research Laboratory, Boulder, Colorado USA

[3]Laboratoire de l'Atmosphere et des Cyclones, UMR8105, CNRS-Meteo France-Universite La Reunion, La Reunion, France

Revised for final submission to GMD 6 October 2014

Corresponding author: Wayne M. Angevine, NOAA ESRL R/CSD4, 325 Broadway, Boulder, CO 80305

Email: Wayne.M.Angevine@noaa.gov

Phone: 303-497-3747

Key Points:

Ensemble spread of tracer concentrations from a Lagrangian particle dispersion model (FLEXPART-WRF) is presented.

Uncertainty of tracer concentrations at grid scale due only to meteorological uncertainty is 30-40%.

Uncertainty of tracer age due only to meteorological uncertainty is 15-20%.

No simple relationships are found between tracer spread and local physical parameters.

Abstract

Lagrangian particle dispersion models require meteorological fields as input. Uncertainty in the driving meteorology is one of the major uncertainties in the results. The propagation of uncertainty through the system is not simple, and has not been thoroughly explored. Here, we take an ensemble approach. Six different configurations of the Weather Research and Forecast (WRF) model drive otherwise identical simulations with FLEXPART-WRF for 49 days over eastern North America. The ensemble spreads of wind speed, mixing height, and tracer concentration are presented. Uncertainty of tracer concentrations due solely to meteorological uncertainty is 30-40%. Spatial and temporal averaging reduces the uncertainty marginally. Tracer age uncertainty due solely to meteorological uncertainty is 15-20%. These are lower bounds on the uncertainty, because a number of processes are not accounted for in the analysis.

Index Terms:

0345 Pollution: urban and regional

0368 Troposphere: constituent transport and chemistry

3307 Boundary layer processes

3355 Regional modeling

Keywords:

Lagrangian particle dispersion model (LPDM)

Uncertainty

FLEXPART

WRF

Tracer transport

Southeast U.S.

1. Introduction

Lagrangian particle dispersion models (LPDMs) are commonly used to simulate transport of trace gasses and aerosols for air pollution studies, greenhouse gas tracking, determination of sources of radiative releases [*Stohl et al.*, 2012], and forecasting of volcanic impacts. Lagrangian models are efficient, flexible, and self-adjoint. The latter property means that simulations can be run backward in time to find the sources of species observed at a particular time and place, which can provide a very large gain in efficiency. Backward runs are used, among other uses, to invert measurements to find source emission strengths and locations [*Brioude et al.*, 2011; *Brioude et al.*, 2013b; *Locatelli et al.*, 2013] (and many others). LPDMs are used at scales ranging from global to mesoscale.

Uncertainty in LPDM results is difficult to assess. Many sources of uncertainty exist, among the most important being uncertainty in emissions and uncertainty in the driving meteorology. Lagrangian models require meteorological fields as input. These are usually provided by operational output or reanalysis from a numerical weather prediction model.. For global or large-scale simulations, output from global operational models or associated reanalyses is commonly used. Mesoscale simulations require more finely resolved input data (here we define mesoscale as intended to resolve features 10-100 kilometers in size). Many groups run their own mesoscale meteorological simulations. Assessing the uncertainties and biases in those simulations is itself difficult, since observations are sparse and themselves uncertain. Further, the propagation of errors from the meteorological fields through the LPDM is not trivial. Some aspects are obvious. For example, random errors in wind direction will broaden the plume from a small source. However, the limits of this kind of thinking become clear quite quickly when one considers a plume propagating in a spatially inhomogeneous and temporally changing atmosphere, in which the errors also change in space and time. This is precisely the situation for which mesoscale simulations are needed.

In this paper we present LPDM (FLEXPART-WRF) [*Brioude et al.*, 2013a] simulations driven by a six-member ensemble of meteorological model runs. Other than the driving meteorology, the FLEXPART-WRF runs are identical. FLEXPART-WRF is run forward in time, transporting specified tracer emissions. We postulate that the ensemble spread of wind speed and of mixing height represent the uncertainty of the meteorological simulation. The spread of the tracer concentrations then represents the meteorological uncertainty as propagated through FLEXPART-WRF. However, such a small ensemble probably does not represent the full range of uncertainty. Biases due to errors in parts of the model common to all configurations will produce biases in the ensemble output that cannot be detected. We therefore attempt to interpret the results with suitable modesty. Many results are presented with one significant figure, or as ranges, to avoid unwarranted precision. We also note that the generality of the results is unknown. The region we cover is in the middle of a continent, with only modest terrain, and we only consider six weeks of one season. We use spatially distributed emissions; point sources might produce rather different results.

Hegarty et al. [2013] showed that differences between LPDMs are much smaller than differences between meteorological models, pointing out the fact that uncertainties most likely arise from the meteorological models when Lagrangian models are used. The propagation of uncertainty from meteorological fields through an LPDM was addressed by *Lin and Gerbig* [2005] for horizontal wind uncertainty, and by *Gerbig et al.* [2008] for uncertainty in vertical mixing. In both cases, they found that failing to account for meteorological uncertainty produced backward simulations with insufficient dispersion. They pointed out the importance of spatial correlation in the random errors.

All errors were assumed to be random, that is, biases were not addressed. Numerical uncertainties, especially those due to terrain, were addressed by *Brioude et al.* [2012]. Meteorological performance of a group of regional air quality models was evaluated by *Vautard et al.* [2012]. Ensemble forecasts were used to evaluate ozone predictability in Texas by *Zhang et al.* [2007]. *Locatelli et al.* [2013] used several different global meteorological and transport model pairs to evaluate uncertainty in methane inversions, finding large uncertainties at regional and smaller scales. Several recent studies [*Chevallier et al.*, 2010; *Houweling et al.*, 2010; *Kretschmer et al.*, 2012; *Lauvaux and Davis*, 2014] used small numbers of models or configurations of one model to evaluate uncertainties in carbon dioxide (CO_2) simulations. Of these, *Kretschmer et al.* [2012] and *Lauvaux and Davis* [2014] worked at mesoscale with WRF meteorology. They explored only the differences due to parameterization of vertical mixing.

The Southeast Nexus (SENEX) campaign (http://www.esrl.noaa.gov/csd/projects/senex/) was conducted in June and July 2013. The NOAA WP3 aircraft made 19 science flights (figure 1) from its base in Smyrna, TN (near Nashville). The aircraft carried a comprehensive package of gas-phase and aerosol chemistry instruments, as well as standard meteorological instruments.

After presenting the model configurations (section 2), we evaluate the ensemble and its members against specifically relevant observations (section 3). Then we present the ensemble spreads (section 4) followed by discussion and conclusions.

2. Model configurations

Six WRF configurations are used, as shown in table 1. They cover three axes of the configuration space, including two different initial and boundary condition datasets, two different planetary boundary layer parameterizations, and two different treatments of the soil variables. All are run on a single 12 km horizontal grid covering most of the eastern half of North America (figure 1). The vertical grid has 60 levels with 19 below 1 km AGL and the lowest level at 16 m. We note that the goal is to produce several reasonable solutions, not to establish a single "best" configuration. All configurations use WRF version 3.5, RRTMG shortwave and longwave radiation, Eta microphysics, and the Noah land surface model with single-level urban canopy. The Grell 3D cumulus scheme was used, with shallow cumulus option on for runs with the MYNN PBL scheme and off for runs with the TEMF PBL. The model was initialized at 0000 UTC each day and run for 30 hours. Except for the runs with cycled soil moisture and temperature, all initial and land boundary conditions were taken from the global analysis (GFS or ERA-Interim). To make a continuous output dataset, the first six hours of each daily run were discarded as spinup. Sea surface temperature was provided by the U.S. Navy GODAE high-resolution SST, (see http://www.usgodae.org/ftp/outgoing/fnmoc/models/ghrsst/docs/ghrsst_doc.txt) updated every six hours and interpolated between updates. No observed data was directly assimilated into WRF, nor were the WRF runs nudged toward any analysis. Most of these configuration choices were the same as used for California in [*Angevine et al.*, 2012]. References for all WRF options can be found in [*Skamarock et al.*, 2008].

We used a version of the FLEXPART Lagrangian particle dispersion model [*Stohl et al.*, 2005] modified to use WRF output [*Brioude et al.*, 2013a]. FLEXPART-WRF uses the same grid spacing as in WRF. FLEXPART-WRF solves turbulent motion in a Lagrangian framework using first-order Langevin equations. The turbulent motion is stochastic and parameterized using the Hanna scheme. That scheme uses PBL height, Monin-Obukhov length, convective velocity scale, roughness length and friction velocity. The PBL height and friction velocity are read from the WRF output. The PBL height in WRF with the MYNN PBL scheme is calculated based on a TKE threshold. With the

TEMF PBL scheme, the PBL height is the level reached by an entraining thermal from the surface [*Angevine et al.*, 2010]. FLEXPART-WRF prescribes a turbulent profile based on the Hanna scheme [*Stohl et al.*, 2005], depending on convective, neutral or stable conditions. Horizontal and vertical turbulence are both calculated from the Hanna scheme.We used the WRF output with an output time interval of 30 minutes. The number of particles emitted per unit time in each grid square is proportional to the tracer emissions at that time and place in the inventory (described below). Runs begin at 0000 UTC 4 May 2010 and run until 0000 UTC 26 June 2010. Particles are retained until they leave the domain. Each particle carries a fixed quantity of tracer. The time of emission is carried with each particle. We used time-average wind out of WRF to reduce trajectory uncertainties [*Brioude et al.*, 2012] as time-average wind is more representative of the wind variability than instantaneous wind out of WRF. *Brioude et al.* [2012] have shown that this setup conserves the well mixed criterion in the PBL in FLEXPART-WRF. Above the PBL, a simple coefficient of diffusivity is used to simulate the horizontal turbulent motion in the free troposphere. Particles are not exchanged directly by turbulence between the PBL and the free troposphere but by horizontal displacement or by the resolved vertical displacement in the WRF wind.

We defined the FLEXPART-WRF output grid (which is independent of the transport calculation) with a 12 km grid spacing in both horizontal dimensions and 28 vertical layers, each 100 m thick. The horizontal grid corresponds to that used for the driving WRF simulations. Particles are grouped into six age classes on output, with maximum ages of 3, 6, 12, 24, 48, and 120 hours since emission.

Approximately 1.8 million particles were emitted each day of the simulation. No chemical transformation or deposition was simulated. The spatial and temporal pattern of emissions is that of carbon monoxide (CO) specified according the U.S. EPA 2011 National Emission Inventory, version 1, available as of November 8, 2013 (http://www.epa.gov/ttn/chief/net/2011inventory.html#inventorydoc). Gridded (4-km resolution), hourly emissions for a July average weekday in 2011 have been derived from this inventory, and are publically available at the WRF/Chem data site: ftp://aftp.fsl.noaa.gov/divisions/taq/emissions_data_2011/. Specific details on the files and data-sets used for spatial and temporal partitioning are supplied in the readme.txt file at the data site. Because the map projection and domain used in the WRF and FLEXPART-WRF simulations is chosen to overlap with the U.S. EPA emissions grid, hourly emissions from the 4-km NEI inventory are simply combined together within the 12km grid resolution used here. Details of the emissions are not directly relevant here, since all runs use the same emissions and results are normalized. When comparing with observed CO, it must be kept in mind that there are a number of CO sources not accounted for in these simulations. These include biomass burning, class-3 commercial marine vessels, and oxidation of methane and volatile organic compounds.

3. Meteorological evaluation

Here we present some evaluation of the performance of each of the WRF configurations. Our goal is to establish that each of the runs has reasonable and comparable performance and therefore that each is a suitable ensemble member. We do not intend to comprehensively evaluate each run in this context. Evaluation of specific processes such as vertical transport by clouds is reserved for future analyses.

Table 2 presents a statistical comparison of each model run to data from all 19 flights of the NOAA WP3 aircraft during SENEX. All data below 1000 m ASL are used, that is, data in the daytime boundary layer and the nighttime residual layer. All the runs produce statistics in the range usually considered in the literature to be “good agreement.” While small differences may be statistically significant with such a large dataset, we do not consider the differences to be of practical significance. These data, and all WP3 data presented herein, are averaged to 120 s (approximately 12 km) to match the model output grid. Calculations with 10 s data (not shown) produce very similar results.

Soil moisture is a key control on meteorological model performance [*Chen et al.*, 2007; *Koster et al.*, 2010; *Kumar et al.*, 2006; *LeMone et al.*, 2008] because it governs the partitioning of incoming solar radiation into sensible heat flux (heating the boundary layer) and latent heat flux (moistening the boundary layer). The six WRF runs use three different strategies to initialize soil moisture and temperature. The runs with GFS initial and boundary conditions (“G” runs) use the soil moisture directly from the GFS analysis at 0000 UTC each day, interpolated to the WRF grid. Runs with ERA-Interim (“E” runs) do the same with the ERA-Interim soil moisture. Cycled runs (“ExC”) start with the soil moisture from ERA-Interim at 0000 UTC on 28 May, and then run open loop. That is, the soil moisture for each day’s run is taken from the 24-hour forecast initialized the previous day. This approach was shown by *Angevine et al.* [2014] and *Di Giuseppe et al.* [2011] to improve results under some conditions, although the differences in these runs are small.

The Climate Reference Network (CRN) [*Diamond et al.*, 2013] provides measurements of soil moisture at multiple levels at 28 sites within our model domain. The time series of modeled and observed soil moisture is shown in figure 2. The runs using GFS soil moisture directly are clearly too moist, and a strong tendency to dry down in the course of each day is visible. Runs with ERA-Interim start and stay close to the observations. Without cycling, these runs (EM and ET) are too moist after day 170, and a diurnal cycle is visible, but smaller than with GFS. Run EMC stays closest to the observations through the period. Around day 160 run ETC falls below the observations and remains there until late in the period. In figure 3, the observations of daily maximum and minimum near-surface air temperature at the CRN sites are shown along with the simulations from each WRF run. All runs have a larger diurnal cycle than the observations. Some of the differences between runs can be traced to the soil moisture and shallow cloud treatment, but the details are outside the scope of this paper.

Cycling soil moisture is vulnerable to errors in modeled precipitation. Figure 4 shows the observed precipitation for the whole period from the NOAA Stage IV analysis (http://data.eol.ucar.edu/codiac/dss/id=21.093), a blend of gauge and radar measurements. The corresponding modeled precipitation is shown in figure 5, and the totals are in table 4. All of the WRF runs miss an area of precipitation in the north-central part of the domain (roughly 38-40N, 87-89W) that occurs in late June, but otherwise the spatial patterns are similar. All runs underestimate the total precipitation except GM, which comes quite close despite the previously mentioned missing area.

4. Ensemble spreads and their relationships

The ensemble spread of wind speed is shown in figure 6. The averages are taken over all 50 days and hours 1000-1200 UTC (denoted AM) and 1800-2000 UTC (denoted PM). Throughout the text, we discuss the “2/3” spread, that is, the difference between the fourth and second ranking values of the six models at each point. This corresponds to the common idea of uncertainty as a standard deviation [*Taylor*, 1997]. The choice is discussed further in the Discussion section below. Some tables also show the “full” spread (maximum minus minimum value). If the spread is not explicitly qualified as “2/3” or “full”, the 2/3 spread is intended. In the figures, spreads are normalized by the mean value at that point from the six models, so a plotted value of 1 means that the spread is equal to the mean value. The level of approximately 200 m AGL is chosen to be relevant to both daytime and nighttime transport. Mean and median spreads are approximately 20%. This includes the narrow band at the domain edges where the spread is small, but the results are only slightly reduced thereby. Some geographic features are apparent, for example the Appalachian Mountains have larger spreads than surrounding lowlands both in the morning and especially at midday. The largest spreads are found in northern Florida, probably due to differences in thunderstorms between the WRF runs.

Mixing height is a key parameter in Lagrangian models. The ensemble spread of mixing height (also called PBL height here) is shown in figure 7. The mixing height as used within FLEXPART-WRF is shown, which is somewhat modified from the direct WRF output. In particular, a minimum height of 100 m is imposed upon input to FLEXPART-WRF. The early morning PBL heights (1000-1200 UTC) have large spreads in the eastern part of the domain and even larger in the western part. This is largely because the TEMF PBL scheme allows very low PBL heights as designed, while the MYNN PBL scheme diagnoses higher heights. Near the western edge of the domain, the three runs with TEMF PBL differ on the location and extent of high PBLs, which are not present in the MYNN runs at all. In the afternoon (1800-2000 UTC), PBL height spreads are moderate except over water. Most land areas have spreads around 20%. The large spreads over water arise from differences in the temperature and wind speed and direction. Overwater PBLs can be stable and therefore shallow in the afternoon, but not at the exact same times and places in the different runs. Mean PBL height spreads over the whole domain are 50% in the early morning and 25% at midday.

The effects of mixing height and wind speed can be combined into a single quantity called "ventilation", which roughly expresses the tendency of emissions to be diluted horizontally and vertically. The ventilation is simply the product of mixing height and wind speed, in this case at 200 m AGL (figure 8). The ventilation spread maps inherit primary features from the wind speed (figure 6) and PBL height (figure 7) maps. In the early morning, the ventilation spread is moderate in the east and large in the west. At midday, the Appalachian Mountains stand out as areas of moderately large spread, with quite large values over the Great Lakes, Florida, and the Atlantic and Gulf Coasts. Mean ventilation spreads for the whole domain are 60% in the early morning and 35% at midday.

Figure 9 shows the mean ensemble spread of tracer mixing ratio in the lowest FLEXPART-WRF level (0-100 m AGL). Points with small mean values (<10 ppbv) are masked out. In the afternoon (lower panel) moderate spreads (roughly 30%) are present over most of the central part of the domain. Spreads are large near the Gulf Coast, Great Lakes, and offshore. Mean spread for the whole domain is 35% (Table 5). In the morning (upper panel), the area of moderate spreads is smaller but the spatial distribution of values is similar. Mean spreads are larger, roughly 40%. Some areas with large emissions, for example Atlanta, Georgia (approximate coordinates -84, 34), have relatively small spreads. Table 5 gives the means for several threshold values of mean mixing ratio, showing that areas with larger concentrations have slightly smaller spreads. Note that the tracer values do not include any background CO, so areas unaffected by emissions within the domain have zero mixing ratio. Absolute values of mean tracer concentration and spread are shown in the Supplemental Material. These are useful for checking the reasonableness of the results, but difficult to interpret in terms of uncertainty.

The near-surface layer is perhaps the most difficult layer for the models, so in figure 10 we show the tracer spread in the 400-500 m AGL layer. The afternoon pattern and mean values are similar to the 0-100 m layer, which makes sense because boundary layer turbulence couples these levels strongly during the day. In the early morning, normalized spreads are larger in the upper layer than near the surface, because the upper layer is decoupled from surface emissions.

The WP3 aircraft flights provide another perspective on the ensemble behavior of the CO tracer. Table 6 displays correlations between the measured CO and the tracer from each member and the ensemble mean. Biases and standard deviations are not shown because computing them requires strong assumptions about the emissions and background. In figure 11, a two-dimensional histogram shows the frequency of occurrence of tracer mixing ratio spread and mean age along the flight tracks for all points with CO measurements below 1000 m AGL. The peak of the spread histogram is at about 20% and 30 hours age, and the mean spread is 30% (median 21%). Although the diagram suggests a correlation between age and spread, its value is only 0.12 (Spearman). There are a number of points with short ages and large spreads, and a wide distribution of spread at any age. Fresh plumes near sources explain the large spreads at short ages. These plumes can be rather narrow and small differences in wind direction move them to slightly different locations. At longer ages, the spread distribution narrows because the air being sampled has circulated through the domain for several days, and differences in transport and mixing in specific locations have been smoothed out. The spread may be asymptotic to a value of 50-60% at long ages.

We might have expected that spread and mixing ratio would correlate inversely, plumes measured near sources having little time to be transported differently, but the lower panel of figure 11 shows no such correlation. Larger mixing ratios occur near sources, but different source strengths place those occurrences at different places on the X-axis. In fact, the peak of the histogram occurs at small to moderate spread (10-20%) and small mixing ratio (~15 ppb).

Tracer age is another important product from the FLEXPART-WRF simulations, and its uncertainty should also be evaluated. Figure 12 shows two-dimensional histograms of age spread. The peak of the histogram is at moderate ages (25-35 hours) and spreads of 15-20%. Overall mean spread is 17% and its median is 13%. Age spread is not correlated with age or mixing ratio.

5. Discussion

A key question in working with an ensemble is whether it is reliable, that is, does the probability with which an event occurs in the ensemble correspond to the probability of that event in reality? For our application, we are interested in a simpler but related criterion, whether the spread of the ensemble is a good estimate of the uncertainty of the CO mixing ratio (above background) at a particular time and place. Uncertainty is often expressed by a standard deviation. One standard deviation each side of the mean covers 66% of a Gaussian distribution. For those times, places, and variables for which we have observations, we can compare the error (simulation-obs) with the ensemble spread. These relationships are tabulated in Table 7. Of the meteorological variables, potential temperature and water vapor from the aircraft show spreads somewhat larger than the standard deviation of the errors. Wind speed has approximately equal spread and error. Temperature at 2 m from the Climate Reference Network sites has errors twice the spread. The CO tracer error is sensitive to the choice of mean for normalization, since the observed mean (minus its minimum) is twice as large as the simulated mean. This is due largely to the neglect of non-anthropogenic sources in the simulations. The spread-error relationship is therefore not useful in this situation.

Rank histograms [*Hamill*, 2001] are a method to visualize the relationship between spread and error. Each measurement is ranked among the values from the ensemble members and the ranks are counted. The expectation is that an observation should fall with equal probability into each bin of a ranked ensemble if the ensemble is reliable. Therefore the histogram should be approximately flat, although caveats apply. In figure 13, the rank histograms for meteorological variables measured by the P3 are shown. The potential temperature histogram is fairly flat, indicating reasonable reliability. An excess of points in the leftmost bin indicates a small bias consistent with the values in table 2. A more significant bias to the right is found for water vapor. The wind speed spread may be somewhat too small as indicated by the U shape of the histogram. Figure 14 shows the rank histogram for 2m T at the CRN sites, for which the ensemble clearly has too little spread.

For our six-member ensemble, the standard deviation can be approximated as the range of the four inner members (leaving out the minimum and maximum). This quantity is tabulated as “2/3” spread in table 7, and shown in the preceding figures. It agrees better with the error (also defined as a standard deviation) than the full spread for potential temperature and water vapor. This is the reason we have used the 2/3 spread above and in our conclusions below. The 2/3 spread is clearly too small for 2m T at the CRN sites, for reasons we have not explored.

The ensemble spreads presented above represent, by our postulate, the uncertainty at a single point of a 12-km grid in a single realization. For the maps in figures 9 and 10, the spreads were computed with 3-hour averaging. The comparisons with WP3 data (table 6) include no temporal averaging. The uncertainty can be reduced by further averaging in space or time. The effect of averaging depends on the degree of independence of the samples. Figure 15 shows the behavior of the ensemble spread (uncertainty) with respect to spatial and temporal averaging. Results are shown for individual hours (1100 UTC and 1900 UTC) and for 3-h averages, each spatially averaged over 1, 3, 5, 7, 9, and 19 grid points in each direction (1, 9, 25, 49, 81, and 361 points total). Averaging is done to each mixing ratio field before the spread is calculated. Points are also shown on the right axis for averaging over the entire spatial domain. Removing the three-hour averaging increases the spread by about 5%. Averaging over 3 points in each direction reduces the spread by about 5%. Further reductions come with increased averaging, but the gain is rather slow. Even averaging over 9 points in each direction only reduces the spread by 5-10%. The reduction is much slower than would be expected if we naively assumed that all points in the average or all points in each direction were independent, in which case averaging would reduce uncertainty by the inverse square root of the number of samples (green and red lines respectively). The spreads for 1-h and 3-h averaging converge as spatial averaging increases. The pattern of improvement with averaging is similar at the surface and in the 400-500 m layer. Averaging over the entire domain, a rather extreme procedure, reduces the spread to roughly 5%. This remnant spread is due to the fact that the tracer can leave the finite domain at different rates with different wind patterns.

The results we have presented (figures 6-10) show that patterns of ensemble spread of CO tracer are not simply related to patterns of wind speed, PBL height, or ventilation (their product). This result may appear surprising at first glance. However, we are dealing with a large area with moderately complex terrain, distributed sources, and complex meteorology. The LPDM simulates all of the complex patterns, including medium-range transport between regions and partial recirculation or stagnation of the tracer. There is some tendency toward larger spread of all variables in mountainous areas, at night, and over coastal waters (see for example [*Ngan et al.*, 2012]).

Previous work of *Gerbig et al.* [2008] and *Lin and Gerbig* [2005] addressed uncertainty in meteorology driving an LPDM by adding a correlated random error, effectively increasing the diffusion terms in the transport equations. Our work shows that the uncertainty is highly variable in space and time, and it is not clear how one would account for this in an approach like theirs. Most likely, uncertainties from meteorological model runs cannot be fully addressed by correlated random errors, and an ensemble approach should be used instead.

6. Conclusions

We have presented ensemble spreads of tracer mixing ratio from the FLEXPART-WRF Lagrangian particle dispersion model driven by meteorological fields from six different configurations of WRF. The FLEXPART-WRF model and WRF model source codes are publicly available online. Interested parties can contact us to access the (large) amount of WRF and FLEXPART-WRF output used in this study.

The spreads of a passive tracer emitted according to all inventoried CO sources are 30-40% , for transport time of 5 days or less, whether they are taken over the whole domain at the surface or in the daytime boundary layer (table 5), or sampled by the aircraft (table 7). Excluding points with small tracer mixing ratios keeps the spreads near the smaller end of those ranges (table 7). Spatial or temporal averaging reduces the spreads, but rather slowly (figure 15).

We postulated that the tracer spread is a measure of uncertainty in the LPDM simulation due to meteorological uncertainty. This is verified by comparing spreads to errors in meteorological variables. Among meteorological variables compared with measurements on the aircraft, the ensemble is roughly reliable for potential temperature and water vapor, but has too little spread for wind speed. For near-surface temperature at the CRN sites, the ensemble has significantly too little spread.

No member of a valid ensemble should be obviously bad or obviously superior. The direct comparisons with observations in tables 2 and 3 verify this. The best and worst performing members for one variable or platform are not the same as for others. It is also interesting to note that the ensemble mean is not obviously better than the best member for any particular variable.

We examined wind speed, boundary layer height, and ventilation looking for relationships between the spreads of these parameters and the tracer spread. No obvious relationships were found. Spreads of meteorological variables are largest where we would expect, in complex terrain, at night, and over coastal waters. Simple relationships among the uncertainties of meteorological parameters and the tracer uncertainty are missing because of terrain, partial recirculation, medium-range (order 100 km) transport, and long tracer lifetime. These are the reasons why an LPDM is needed in this and similar real mesoscale situations. We do not think that tracer spreads can be predicted from known error characteristics of the meteorological variables. We recommend that an ensemble approach like this one, or even more sophisticated, be used to assess the uncertainty of Lagrangian simulations.

Uncertainty in single LPDM simulations of passive tracers at mesoscale due solely to uncertainty in the meteorological forcing is 30-40% of the tracer mixing ratio. The uncertainty is somewhat less, perhaps as little as 20%, under particularly favorable conditions (strong, broad plumes sampled in daytime at moderate distance/time downwind of their sources). It is greater, as much as 60%, under less favorable conditions (weak or narrow plumes, undifferentiated background, or sampling at night). Spatial averaging can reduce the uncertainty with loss of resolution. Uncertainty of simulated tracer age is 15-20%.

References

Angevine, W. M., H. Jiang, and T. Mauritsen (2010), Performance of an eddy diffusivity - mass flux scheme for shallow cumulus boundary layers, *Monthly Weather Review*, *138*, 2895-2912.

Angevine, W. M., E. Bazile, D. Legain, and D. Pino (2014), Land surface spinup for episodic modeling, *Atmos. Chem. Phys. Discuss.*, *14*(4), 4723-4744.

Angevine, W. M., L. Eddington, K. Durkee, C. Fairall, L. Bianco, and J. Brioude (2012), Meteorological model evaulation for CalNex 2010, *Monthly Weather Review*, *140*, 3885-3906.

Brioude, J., W. M. Angevine, S. A. McKeen, and E. Y. Hsie (2012), Numerical uncertainty at mesoscale in a Lagrangian model in complex terrain, *Geosci. Model Dev.*, *5*(5), 1127-1136.

Brioude, J., et al. (2011), Top-down estimate of anthropogenic emission inventories and their interannual variability in Houston using a mesoscale inverse modeling technique, *Journal of Geophysical Research: Atmospheres*, *116*(D20), D20305.

Brioude, J., et al. (2013a), The Lagrangian particle dispersion model FLEXPART-WRF version 3.1, *Geosci. Model Dev.*, *6*(6), 1889-1904.

Brioude, J., et al. (2013b), Top-down estimate of surface flux in the Los Angeles Basin using a mesoscale inverse modeling technique: assessing anthropogenic emissions of CO, NOx and CO2 and their impacts, *Atmos. Chem. Phys.*, *13*(7), 3661-3677.

Chen, F., et al. (2007), Description and Evaluation of the Characteristics of the NCAR High-Resolution Land Data Assimilation System, *Journal of Applied Meteorology and Climatology*, *46*(6), 694-713.

Chevallier, F., L. Feng, H. Bösch, P. I. Palmer, and P. J. Rayner (2010), On the impact of transport model errors for the estimation of CO2 surface fluxes from GOSAT observations, *Geophysical Research Letters*, *37*(21), L21803.

Di Giuseppe, F., D. Cesari, and G. Bonafe (2011), Soil initialization strategy for use in limited-area weather prediction systems, *Monthly Weather Review*, *139*, 1844-1860.

Diamond, H. J., et al. (2013), U.S. Climate Reference Network after One Decade of Operations: Status and Assessment, *Bulletin of the American Meteorological Society*, *94*(4), 485-498.

Gerbig, C., S. Koerner, and J. C. Lin (2008), Vertical mixing in atmospheric tracer transport models: Error characterization and propagation, *Atmos. Chem. Phys.*, *8*, 591-602.

Hamill, T. M. (2001), Interpretation of rank histograms for verifying ensemble forecasts, *Monthly Weather Review*, *129*, 550-560.

Hegarty, J., R. R. Draxler, A. F. Stein, J. Brioude, M. Mountain, J. Eluszkiewicz, T. Nehrkorn, F. Ngan, and A. Andrews (2013), Evaluation of Lagrangian Particle Dispersion Models with Measurements from Controlled Tracer Releases, *Journal of Applied Meteorology and Climatology*, *52*(12), 2623-2637.

Houweling, S., et al. (2010), The importance of transport model uncertainties for the estimation of CO2 sources and sinks using satellite measurements, *Atmos. Chem. Phys.*, *10*(20), 9981-9992.

Koster, R. D., et al. (2010), Contribution of land surface initialization to subseasonal forecast skill: First results from a multi-model experiment, *Geophysical Research Letters*, *37*(2), L02402.

Kretschmer, R., C. Gerbig, U. Karstens, and F. T. Koch (2012), Error characterization of CO2 vertical mixing in the atmospheric transport model WRF-VPRM, *Atmos. Chem. Phys.*, *12*(5), 2441-2458.

Kumar, S. V., et al. (2006), Land information system: An interoperable framework for high resolution land surface modeling, *Environmental Modelling & Software*, *21*(10), 1402-1415.

Lauvaux, T., and K. J. Davis (2014), Planetary boundary layer errors in mesoscale inversions of column-integrated CO2 measurements, *Journal of Geophysical Research: Atmospheres*, *119*(2), 490-508.

LeMone, M. A., M. Tewari, F. Chen, J. G. Alfieri, and D. Niyogi (2008), Evaluation of the Noah land surface model using data from a fair-weather IHOP_2002 day with heterogeneous surface fluxes, *Monthly Weather Review*, *136*, 4915-4941.

Lin, J. C., and C. Gerbig (2005), Accounting for the effect of transport errors on tracer inversions, *Geophysical Research Letters*, *32*, L01802.

Locatelli, R., et al. (2013), Impact of transport model errors on the global and regional methane emissions estimated by inverse modelling, *Atmos. Chem. Phys.*, *13*(19), 9917-9937.

Ngan, F., D. Byun, H. Kim, D. Lee, B. Rappenglueck, and A. Pour-Biazar (2012), Performance assessment of retrospective meteorological inputs for use in air quality modeling during TexAQS 2006, *Atmospheric Environment*, *54*, 86-96.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, M. G. Duda, X.-Y. Huang, W. Wang, and J. G. Powers (2008), A description of the Advanced Research WRF version 3, NCAR Technical Note *TN-475*, 113 pp.

Stohl, A., C. Forster, A. Frank, P. Seibert, and G. Wotawa (2005), Technical note: The Lagrangian particle dispersion model FLEXPART version 6.2, *Atmospheric Chemistry and Physics*, *5*, 2461-2474.

Stohl, A., P. Seibert, G. Wotawa, D. Arnold, J. F. Burkhart, S. Eckhardt, C. Tapia, A. Vargas, and T. J. Yasunari (2012), Xenon-133 and caesium-137 releases into the atmosphere from the Fukushima Dai-ichi nuclear power plant: determination of the source term, atmospheric dispersion, and deposition, *Atmos. Chem. Phys.*, *12*(5), 2313-2343.

Taylor, J. R. (1997), *An Introduction to Error Analysis*, 2nd ed., 327 pp., University Science Books, Sausalito, CA.

Vautard, R., et al. (2012), Evaluation of the meteorological forcing used for the Air Quality Model Evaluation International Initiative (AQMEII) air quality simulations, *Atmospheric Environment*, *53*(0), 15-37.

Zhang, F., N. Bei, J. W. Nielsen-Gammon, G. Li, R. Zhang, A. Stuart, and A. Aksoy (2007), Impacts of meteorological uncertainties on ozone pollution predictability estimated through meteorological and photochemical ensemble forecasts, *Journal of Geophysical Research*, *112*, D04304,.

Tables

Table 1: Names and primary definitions of the six WRF configurations to be discussed.

Name	Initialization	PBL scheme	Soil treatment	Cumulus
GM	GFS	MYNN2	Direct	Grell 3D with shallow
EM	ERA	MYNN2	Direct	Grell 3D with shallow
EMC	ERA	MYNN2	Cycled	Grell 3D with shallow
GT	GFS	TEMF	Direct	Grell 3D NO shallow
ET	ERA	TEMF	Direct	Grell 3D NO shallow
ETC	ERA	TEMF	Cycled	Grell 3D NO shallow

Table 2: Comparison statistics for all WP3 aircraft flights below 1000 m ASL. Model points are extracted along the flight track every 10 s, linearly interpolated in space and time, and then averaged to 120 s. Std.Dev. is the standard deviation of the differences, and r is the Spearman rank correlation coefficient. Units are m s^{-1} for wind speed, K for potential temperature, and g/kg for water vapor mixing ratio. Sign of bias is (model – measurement). Number of points is 2026.

WP3	GM	EM	EMC	GT	ET	ETC	Ensemble mean
Wind speed Mean bias	0.26	-0.14	-0.16	0.48	0.15	0.14	0.12
Std.Dev.	1.7	1.7	1.7	1.8	1.8	1.8	1.5
r	0.64	0.72	0.72	0.66	0.67	0.68	0.72
Potential temperature Mean bias	-0.30	0.07	0.16	-0.16	0.30	0.58	0.11
Std.Dev.	0.94	1.1	1.1	1.1	1.2	1.2	0.96
r	0.93	0.90	0.90	0.92	0.90	0.90	0.92
Water vapor mixing ratio Mean bias	-0.20	-0.73	-0.88	-0.76	-1.3	-1.6	-0.91
Std.Dev.	1.6	1.5	1.5	1.5	1.5	1.5	1.3
r	0.74	0.79	0.79	0.76	0.78	0.78	0.82

Table 3: Comparison statistics of near-surface (2 m) temperature for 28 Climate Reference Network sites. Model results are from the nearest grid point to each site. Sign of biases is model-measurement.

	GM	EM	EMC	GT	ET	ETC	Ensemble mean
Daily maximum bias	1.4	2.2	2.4	1.8	2.8	3.6	2.4
Daily maximum std. dev.	2.2	1.9	2.0	2.3	2.4	2.9	2.1
Daily maximum r	0.35	0.43	0.42	0.43	0.40	0.34	0.44
Daily maximum 2/3 spread							1.5
Daily minimum bias	-1.6	-0.86	-1.4	-2.0	-1.3	-2.1	-1.5
Daily minimum std. dev.	2.9	2.8	3.0	2.7	2.6	3.0	2.8
Daily	0.46	0.48	0.44	0.47	0.49	0.45	0.47

minimum r							
Daily minimum 2/3 spread							1.0
Daily mean bias	-0.13	-0.54	-0.41	-0.13	0.48	0.47	0.27
Daily mean std. dev.	1.7	1.8	1.9	1.7	1.8	1.9	1.7
Daily mean r	0.54	0.46	0.44	0.52	0.43	0.39	0.48
Daily mean 2/3 spread							1.4

Table 4: Precipitation totals in the portion of the domain shown in figures 3 and 4.

Stage IV observed	GM	EM	EMC	GT	ET	ETC
237	245	189	185	199	154	147

Table 5: Mean normalized CO tracer spreads at two levels of the whole domain with varying mixing ratio thresholds. Number of points is also shown. Grid size is 216*236 so maximum possible N = 50976.

Threshold (mean mixing ratio >)	10 ppb	20 ppb	30 ppb	40 ppb	50 ppb
0-100 m AGL AM 2/3	0.39	0.36	0.34	0.32	0.32
AM full	0.70	0.65	0.61	0.58	0.57
N	38354	29169	19822	13299	8568
PM 2/3	0.35	0.32	0.30	0.29	0.29
PM full	0.62	0.57	0.54	0.52	0.51
N	35856	22698	14238	7739	3487
400-500 m AGL AM 2/3	0.43	0.40	0.38	0.38	N too small
AM full	0.78	0.71	0.69	0.68	N too small
N	34248	18127	10383	1963	20
PM 2/3	0.35	0.32	0.30	0.30	0.29
PM full	0.61	0.56	0.53	0.52	0.51
N	34525	19240	10993	4045	1069

Table 6: Comparison statistics of CO and CO tracer for all WP3 aircraft flights below 1000 m ASL. Model points are extracted along the flight track every 10 s, linearly interpolated in space and time, and then further averaged to 120 s. r is the Spearman rank correlation coefficient. Number of samples is 1597.

	GM	EM	EMC	GT	ET	ETC	Ensemble mean
CO tracer mixing ratio r	0.62	0.61	0.61	0.59	0.59	0.59	0.62

Table 7: Spread and standard deviation statistics for all WP3 aircraft flights below 1000 m ASL and for CRN 2m temperature. Model points are extracted along the flight track every 10 s, linearly interpolated in space and time, and then averaged over 120 s. CRN 2m temperature statistics are for all available hourly observations (N=33569). N=2026 for P3 meteorology, N=1597 for P3 CO. CO spreads and simulated CO standard deviation are normalized by the simulated ensemble mean. Observed CO standard deviation is normalized by the observed mean with minimum value subtracted to account for background. For simulated-observed standard deviation of CO, two values are shown, the smaller is normalized by the observed mean with minimum subtracted (71 ppb) and the larger is normalized by the simulated mean (32 ppb).

	Standard deviation of difference (simulated-observed)	Ensemble spread (full)	Ensemble spread (2/3)	Standard deviation observed	Standard deviation simulated ensemble mean
Potential temperature (P3) (K)	0.96	1.5	0.89	9.1	9.1
Water vapor mixing ratio (P3) (g/kg)	1.3	2.0	1.3	2.3	2.1
Wind speed (P3) (m/s)	1.5	1.9	1.3	2.2	2.3

2m T (CRN) (K)	4.7	2.4	1.5	4.5	4.9
CO tracer mixing ratio (normalized)	0.39 (0.87)	0.54	0.31	0.46	0.58

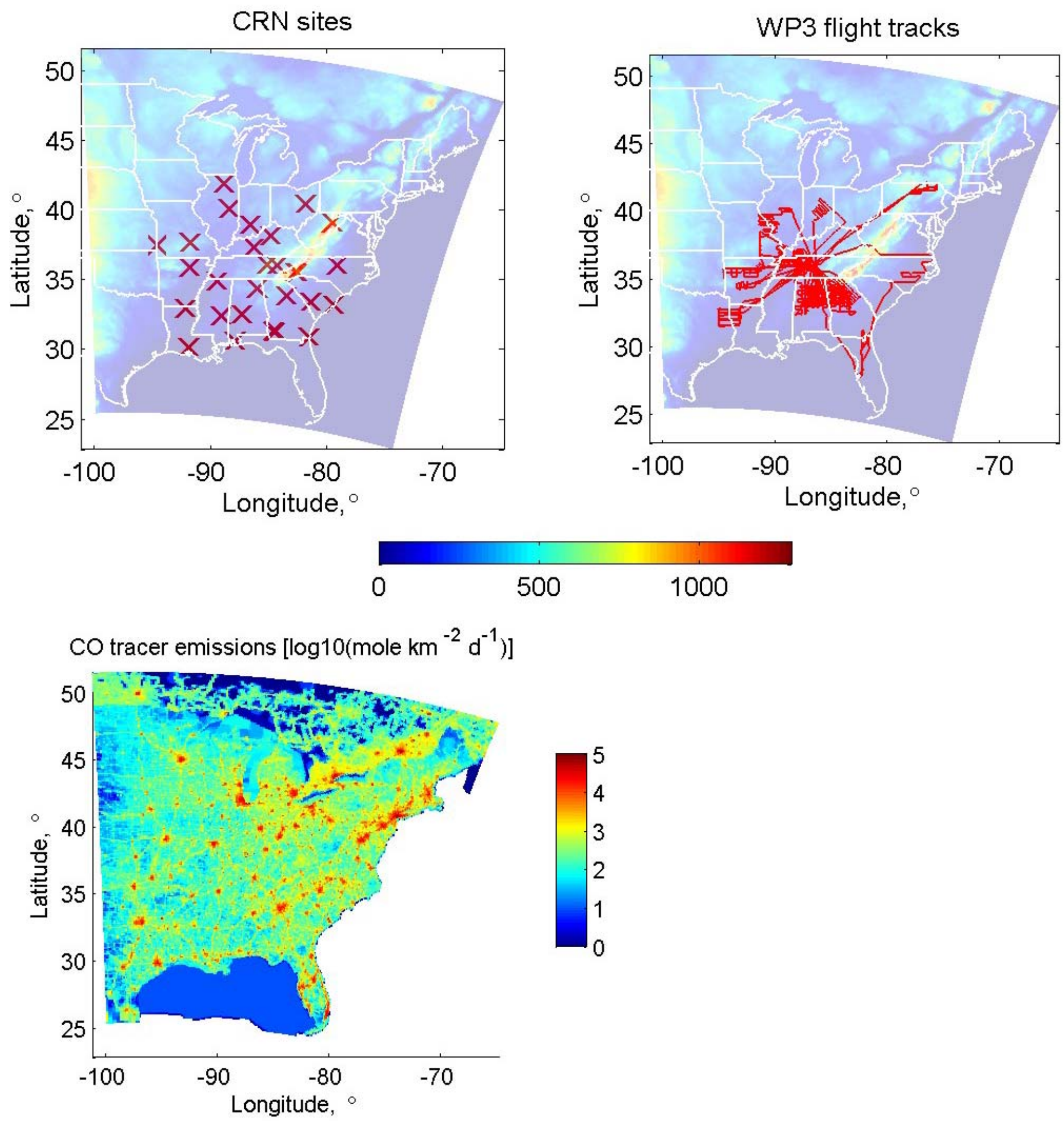


Figure 1: Maps of the WRF domain with terrain height (m ASL) colored as background and showing Climate Reference Network sites (upper left) and flight tracks of the NOAA WP3 (upper right). Lower panel shows CO tracer emissions used in the FLEXPART-WRF runs.

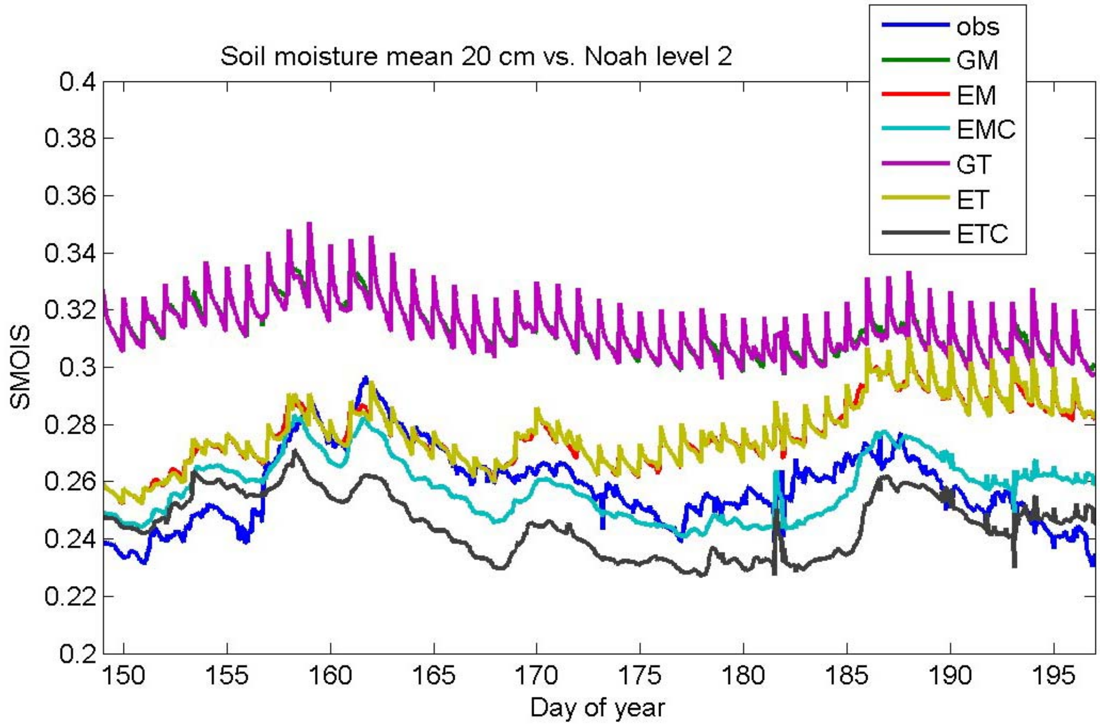


Figure 2: Soil moisture mean of 28 Climate Reference Network stations. Measurement at 20 cm depth is compared to second model level (10-40 cm). Legend refers to table 1. Run GM is often obscured by GT.

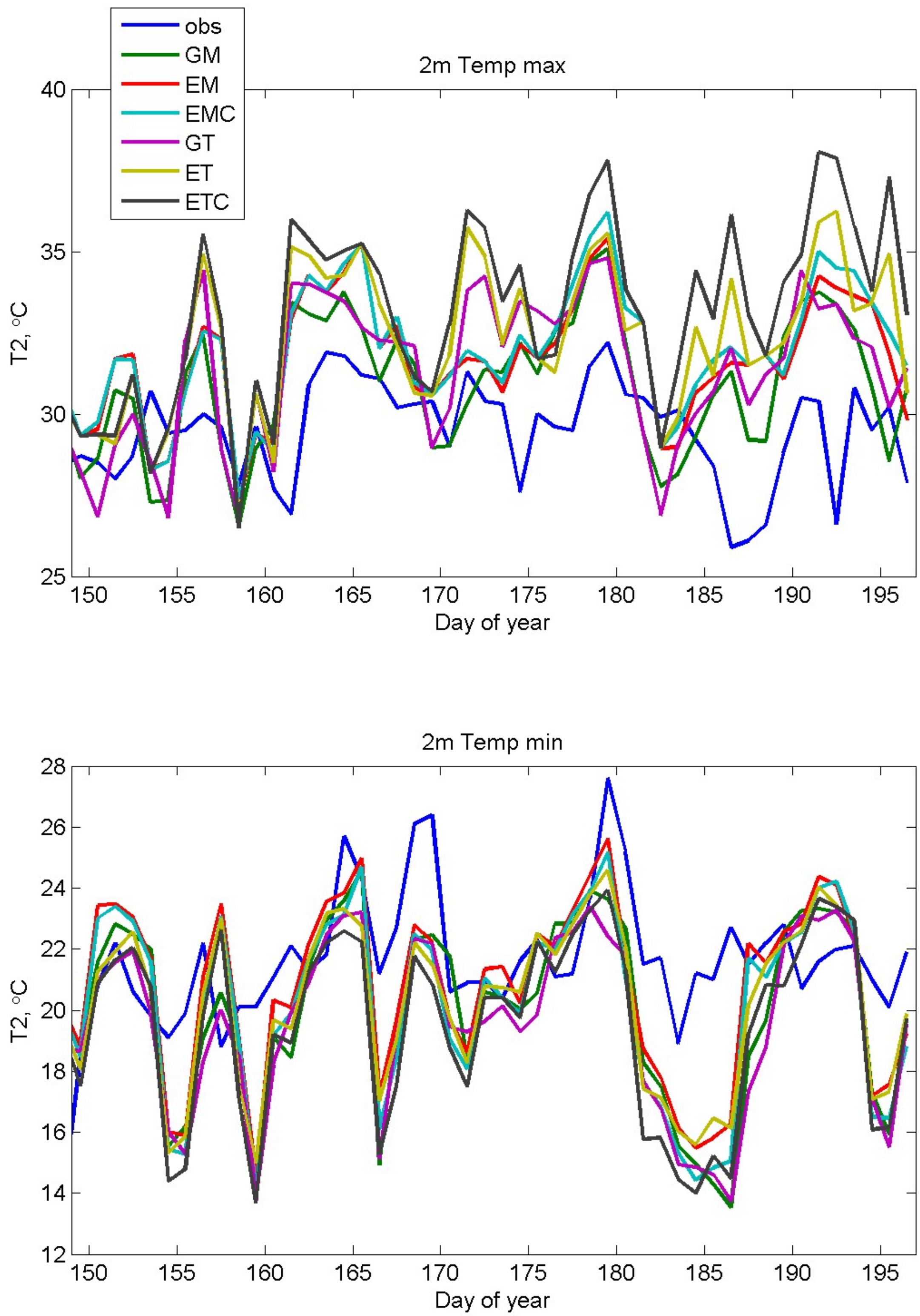


Figure 3: Daily maximum and minimum near-surface temperature averaged over 28 Climate Reference Network sites.

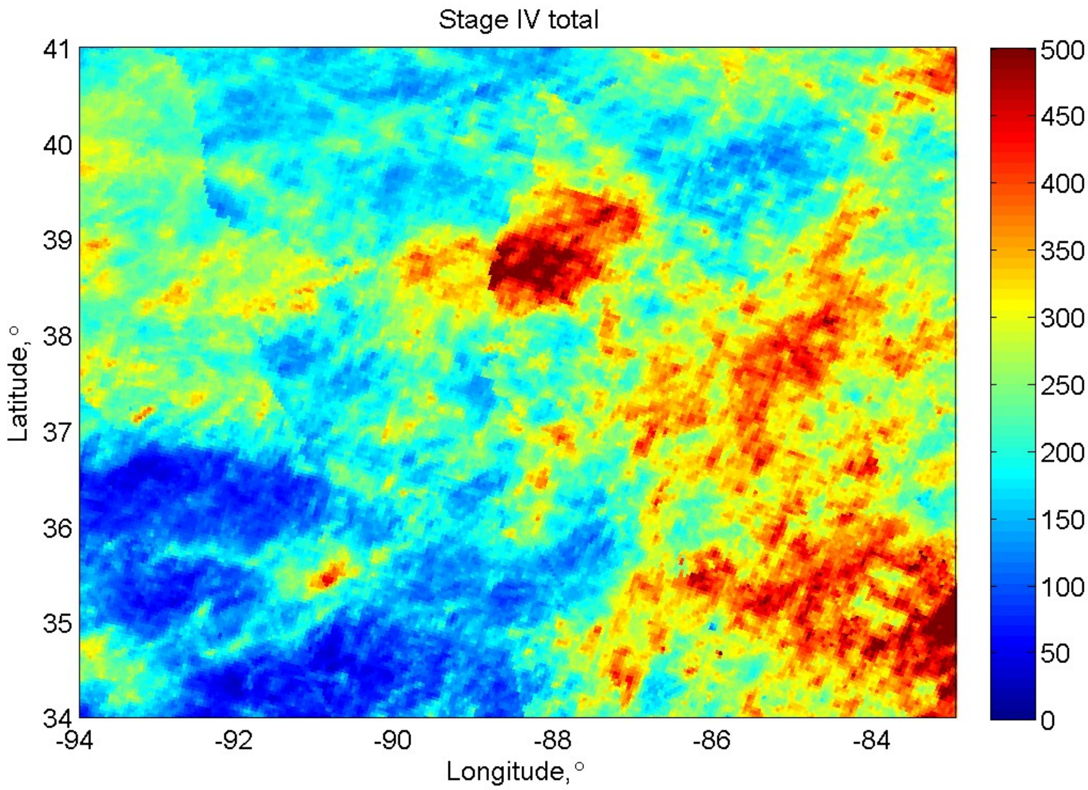


Figure 4: Observed precipitation from the NOAA Stage IV product for 28 May – 15 July 2013 (mm). Edges of the domain are excluded for clarity.

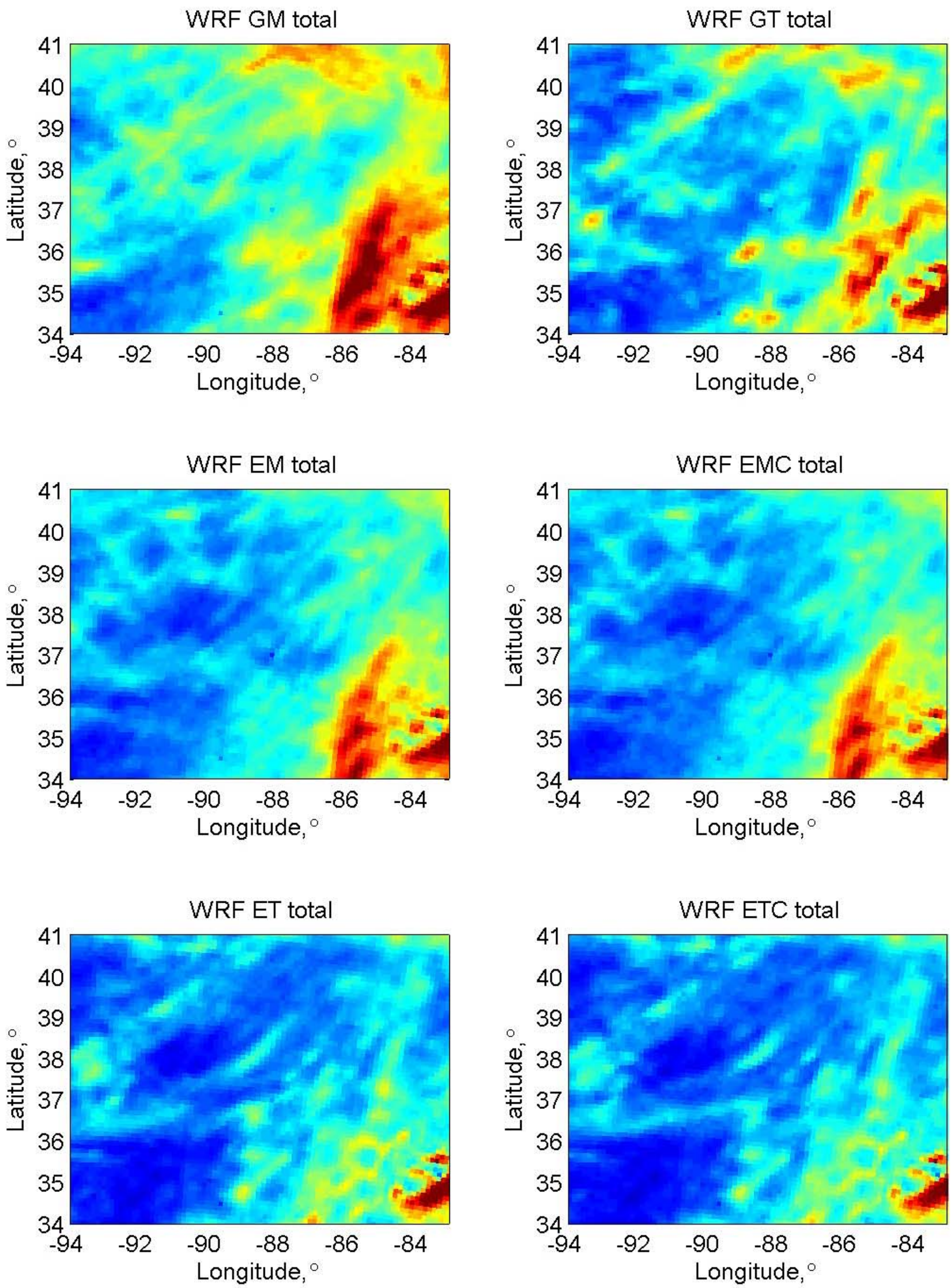


Figure 5: Total precipitation from each WRF run for 28 May – 15 July 2013. Color scale same as figure 3.

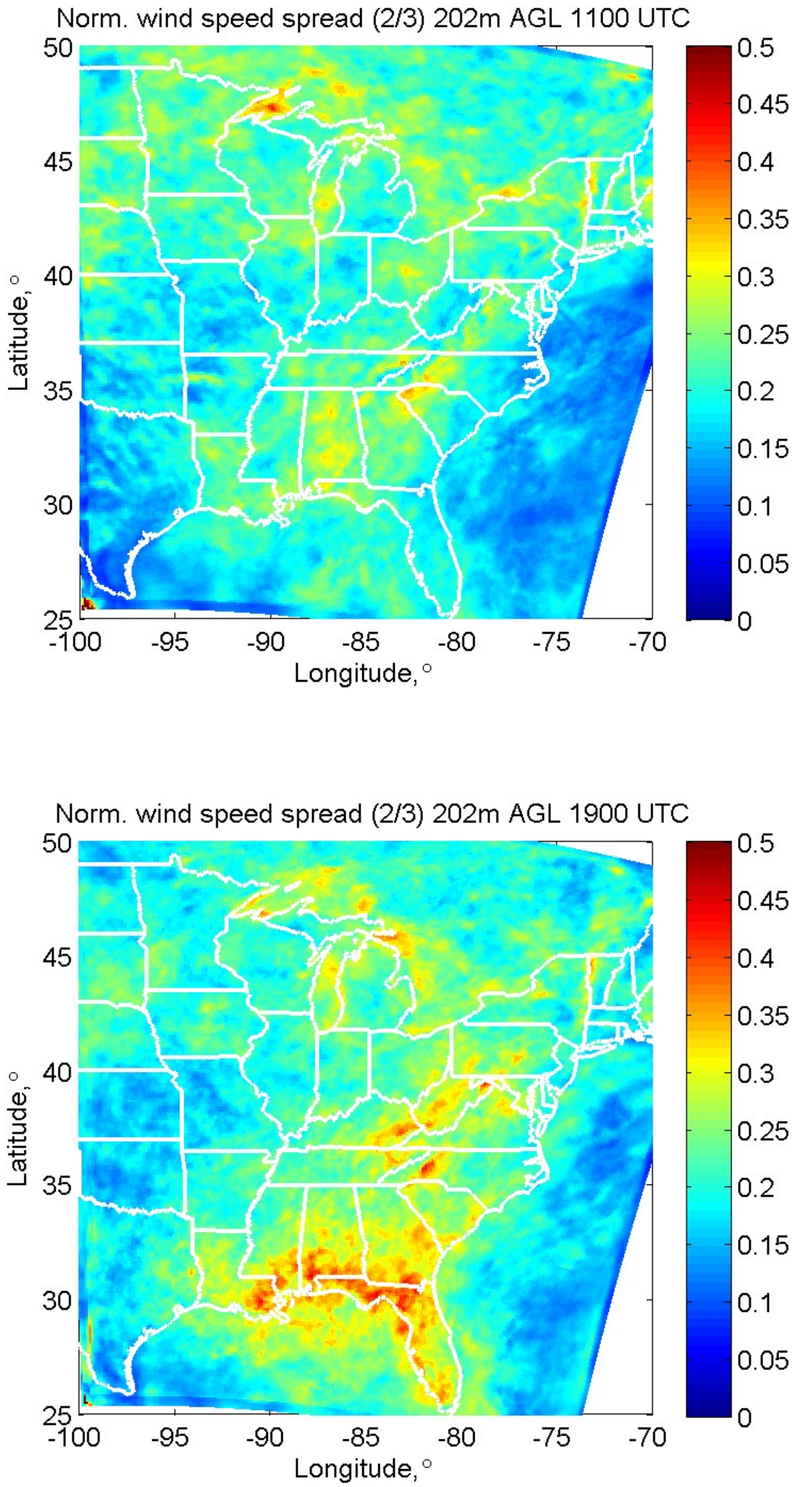


Figure 6: Wind speed spread in early morning and midday from the WRF ensemble. Spread is normalized by mean speed (therefore unitless) and averaged over all 49 days.

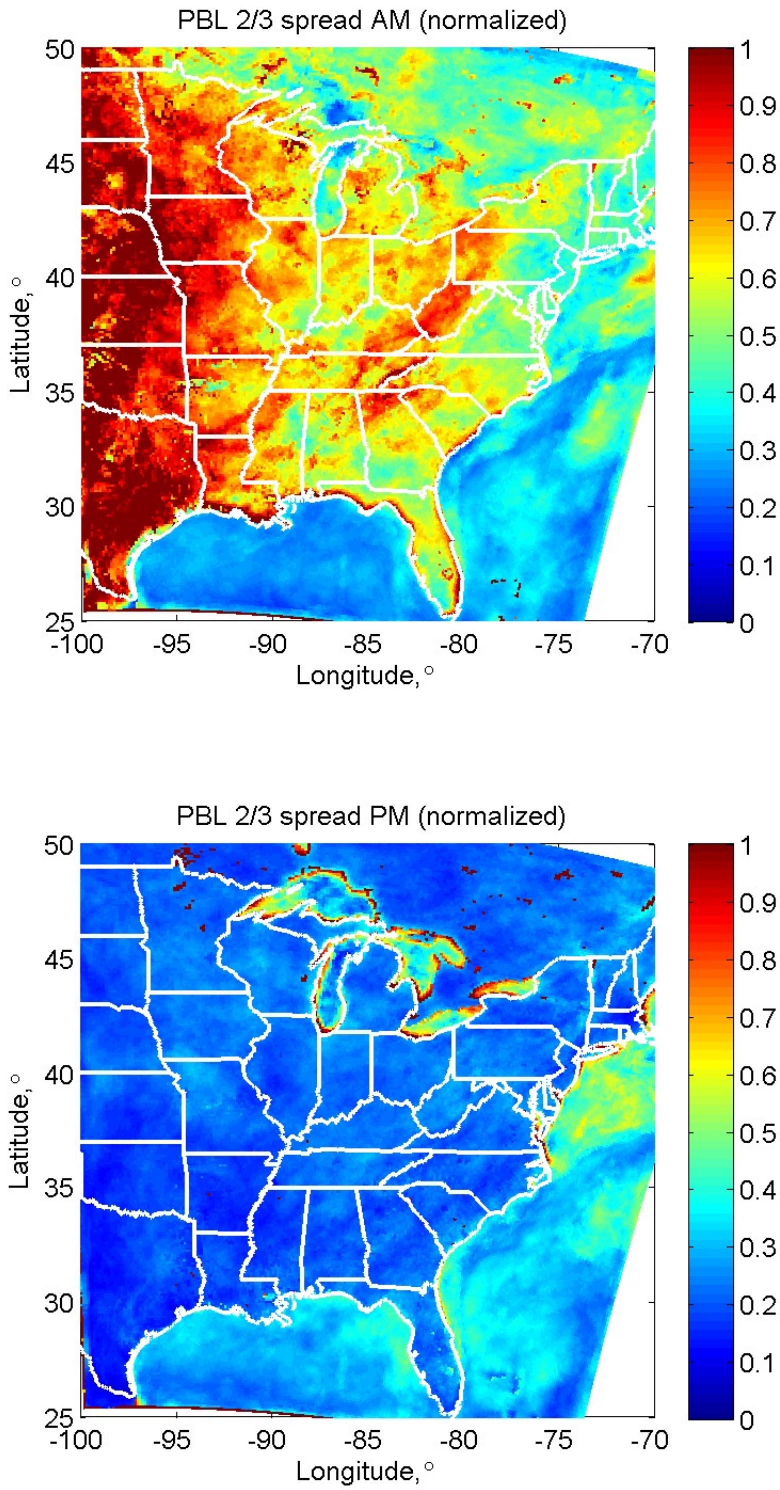


Figure 7: Spread of boundary layer height (mixing height) in the early morning and midday as interpreted by FLEXPART-WRF from the WRF ensemble input. Spread is normalized by the mean value.

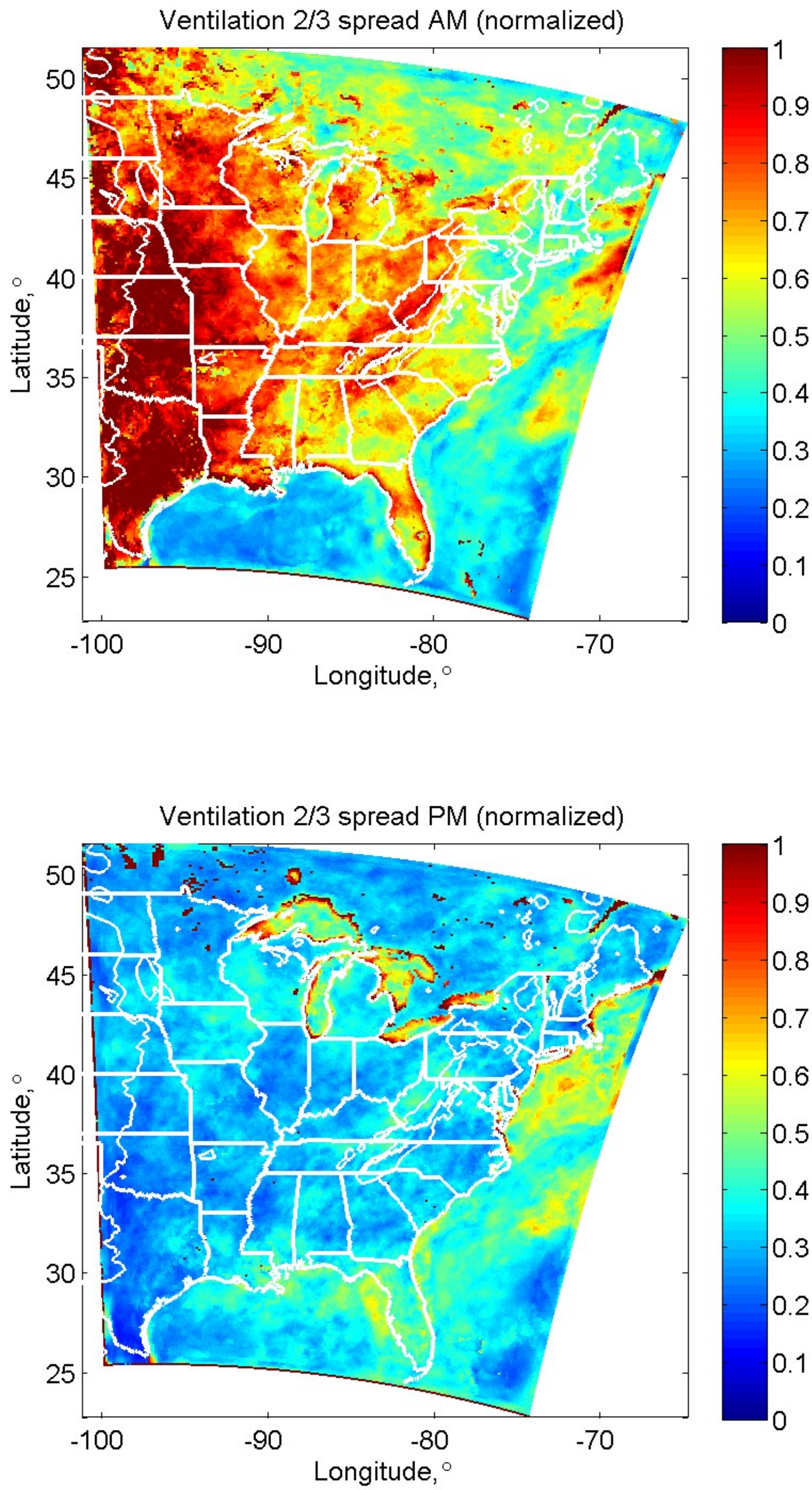


Figure 8: Spread of ventilation (PBL height * wind speed) in the early morning and midday. Spread is normalized by the mean value.

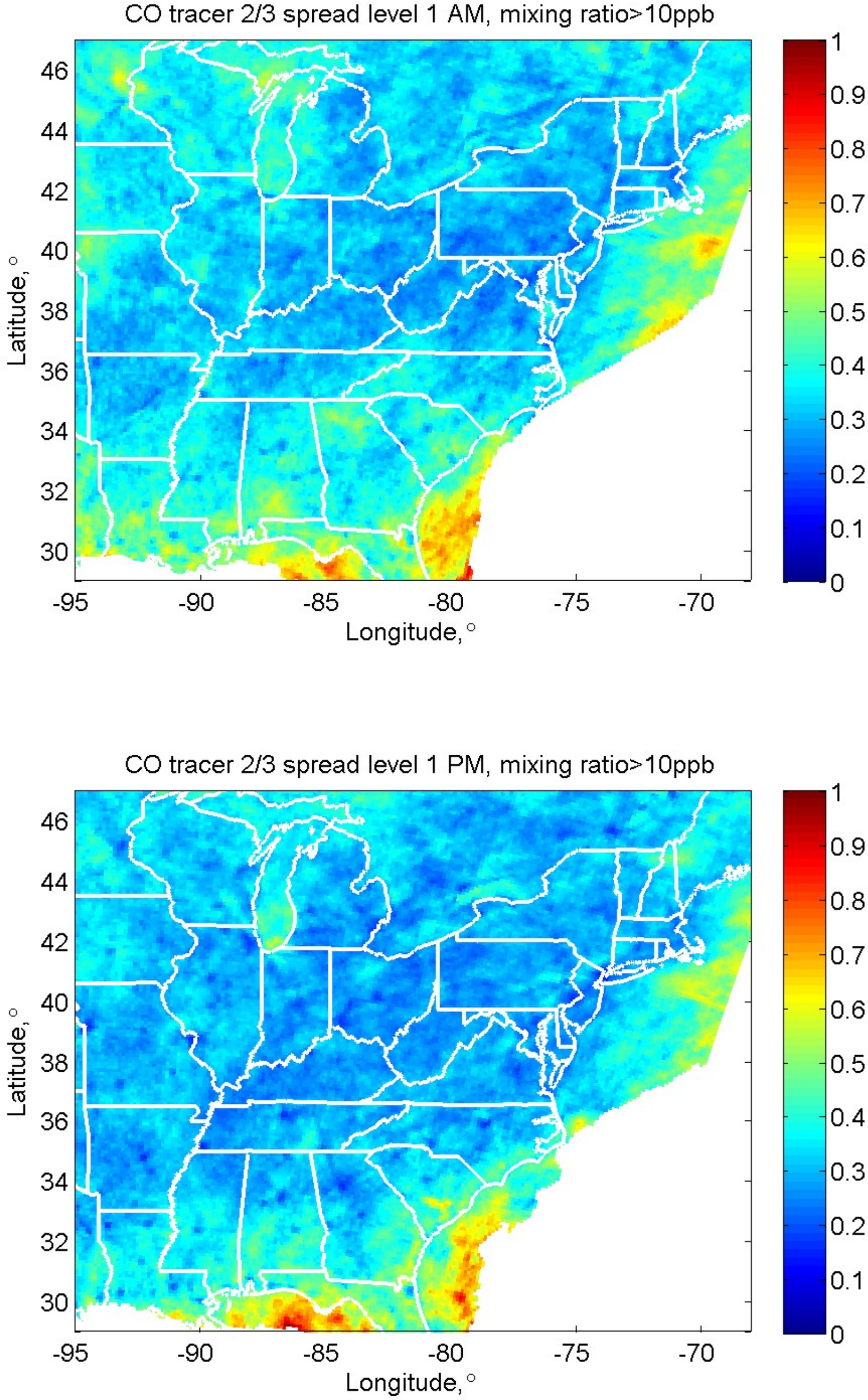


Figure 9: Mean ensemble spread of tracer mixing ratio at level 1 (0-100 m AGL). The averages are taken over all 49 days and hours 0400-0600 LST (AM, top) and 1300-1500 LST (PM, bottom). The spread is normalized by the mean mixing ratio at each point. Points with small mean values (<10 ppbv) are masked out.

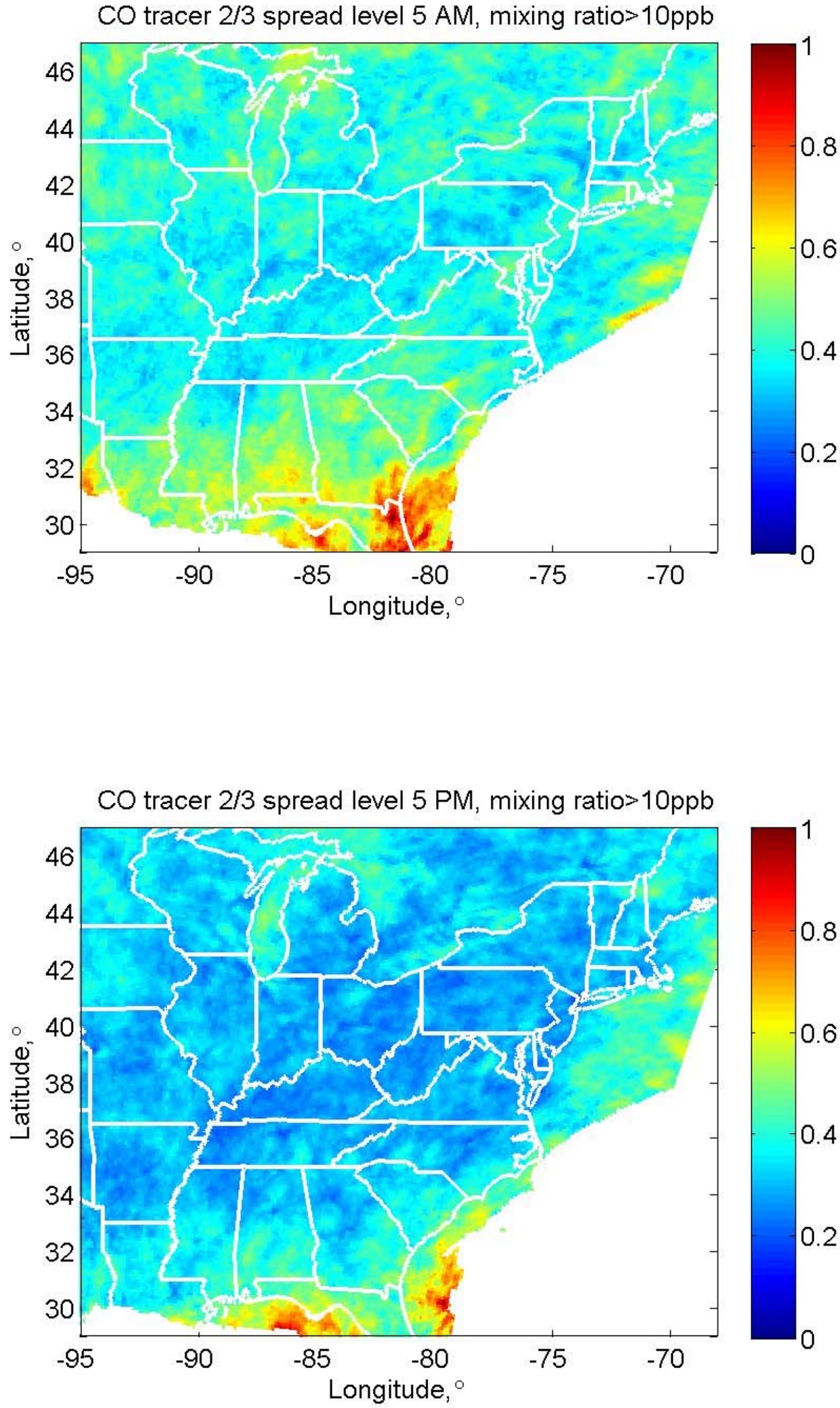
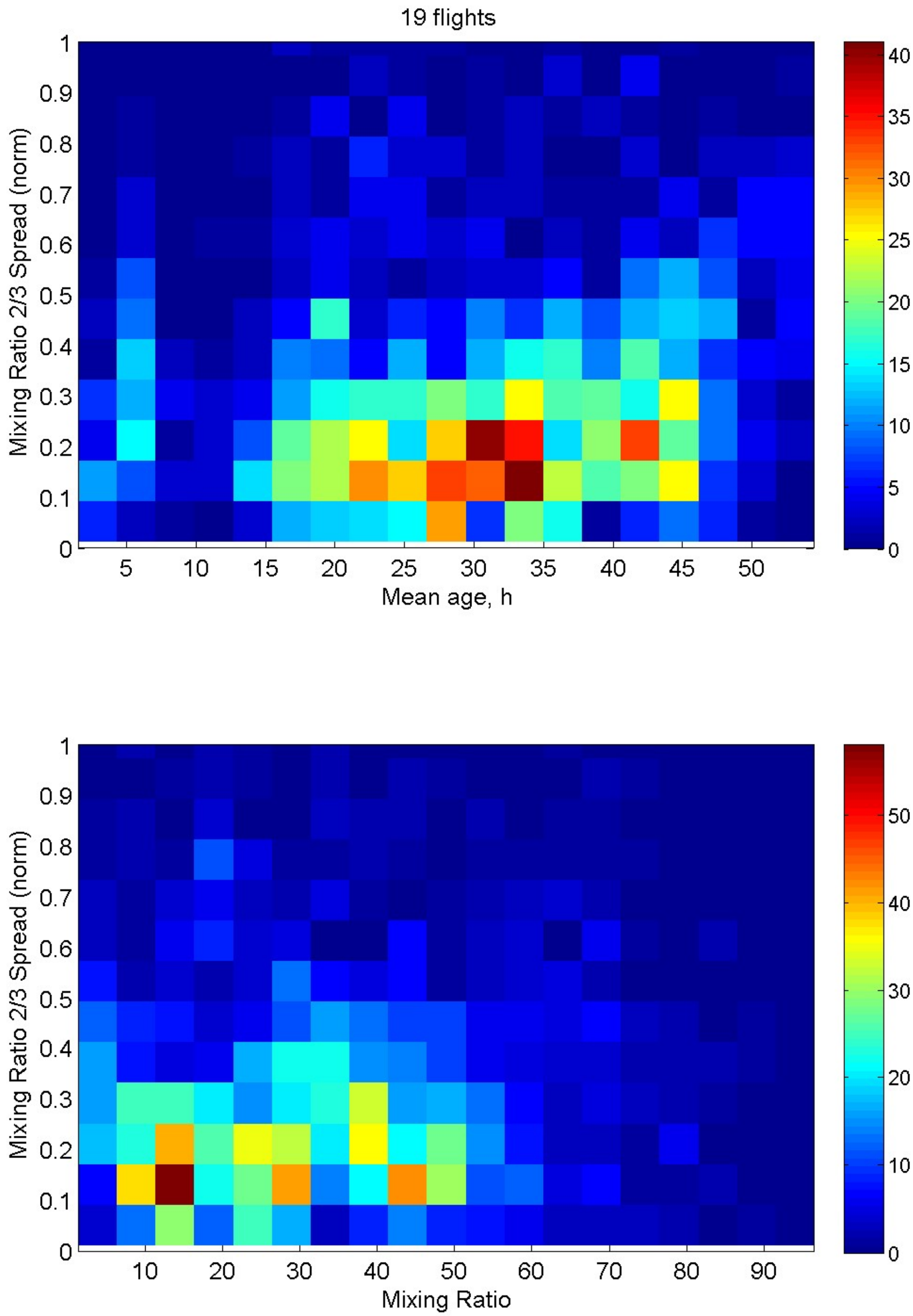


Figure 10: Mean ensemble spread of tracer mixing ratio at level 5 (400-500 m AGL). The averages are taken over all 49 days and hours 0400-0600 LST (AM, top) and 1300-1500 LST (PM, bottom). The spread is normalized by the mean mixing ratio at each point. Points with small mean values (<10 ppbv) are masked out.



557 Figure 11: Frequency of occurrence of CO tracer spread along the P3 flight tracks vs. simulated
558 mean tracer age (top) and simulated mixing ratio (bottom) for all points with valid CO
559 measurements below 1000 m AGL.

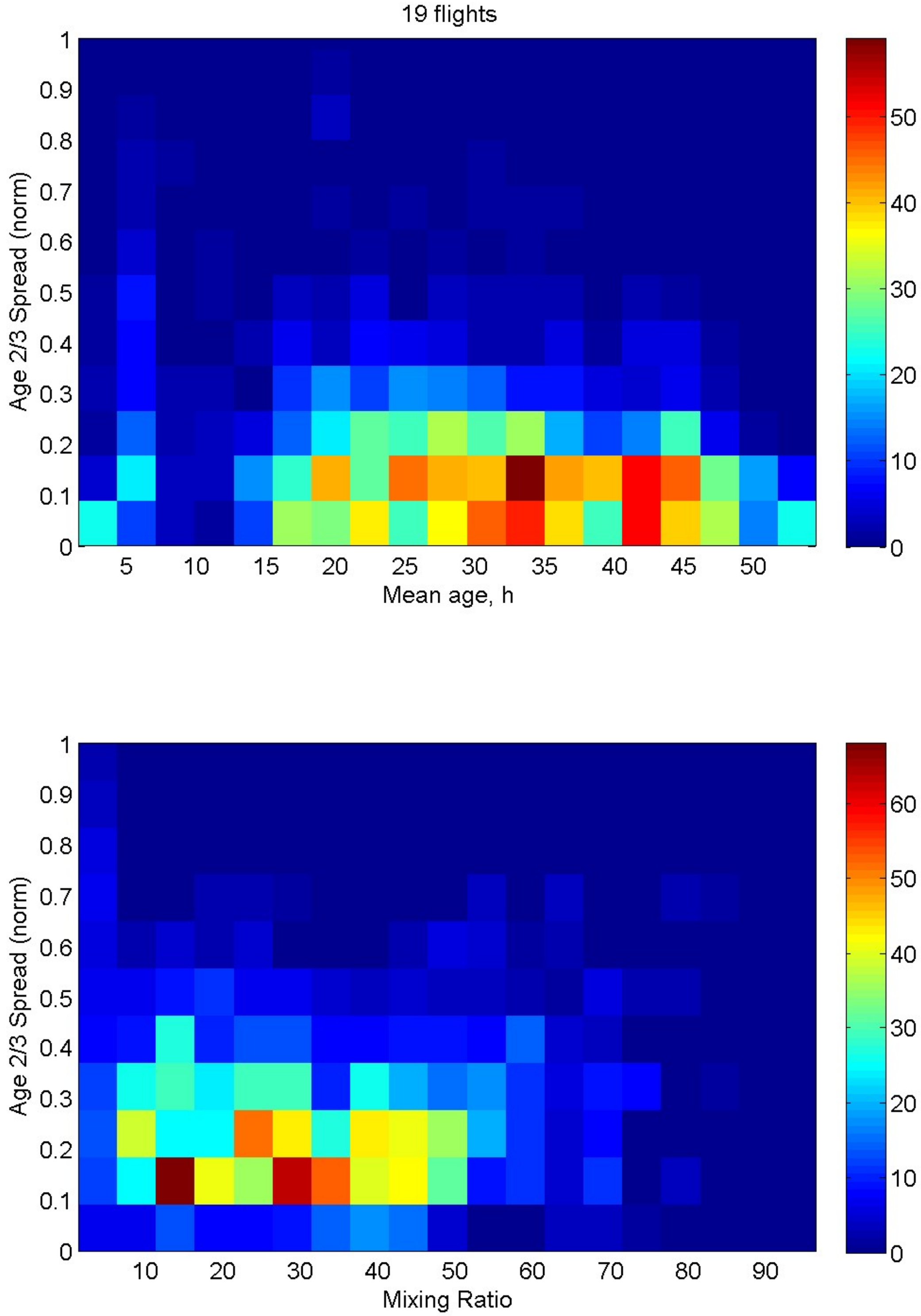


Figure 12: Frequency of occurrence of CO tracer age spread along the P3 flight tracks vs. simulated mean tracer age (top) and simulated mixing ratio (bottom) for all points with valid CO measurements below 1000 m AGL.

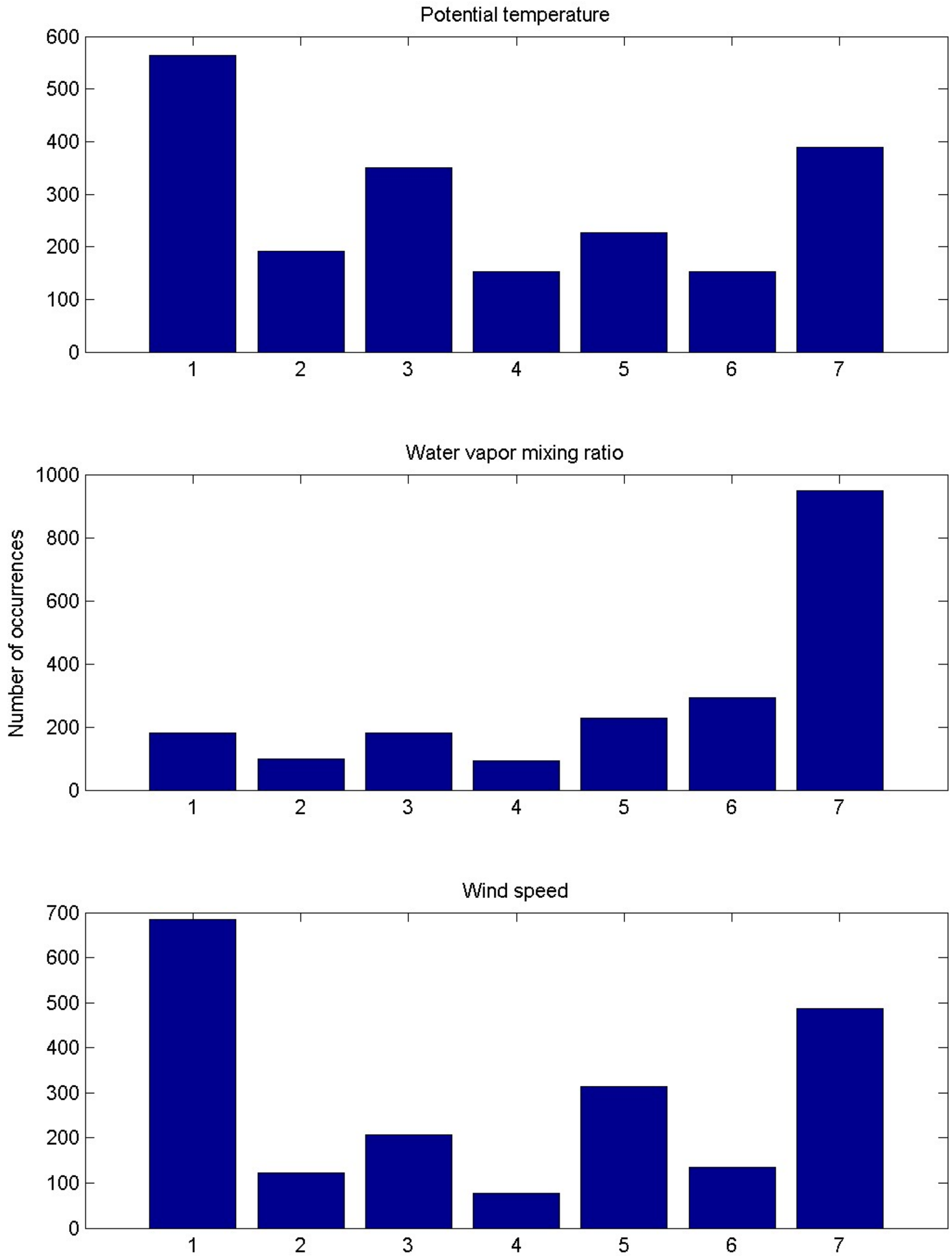


Figure 13: Rank histograms for all P3 flight data below 1000 m AGL for potential temperature, water vapor mixing ratio, and wind speed (as labeled).

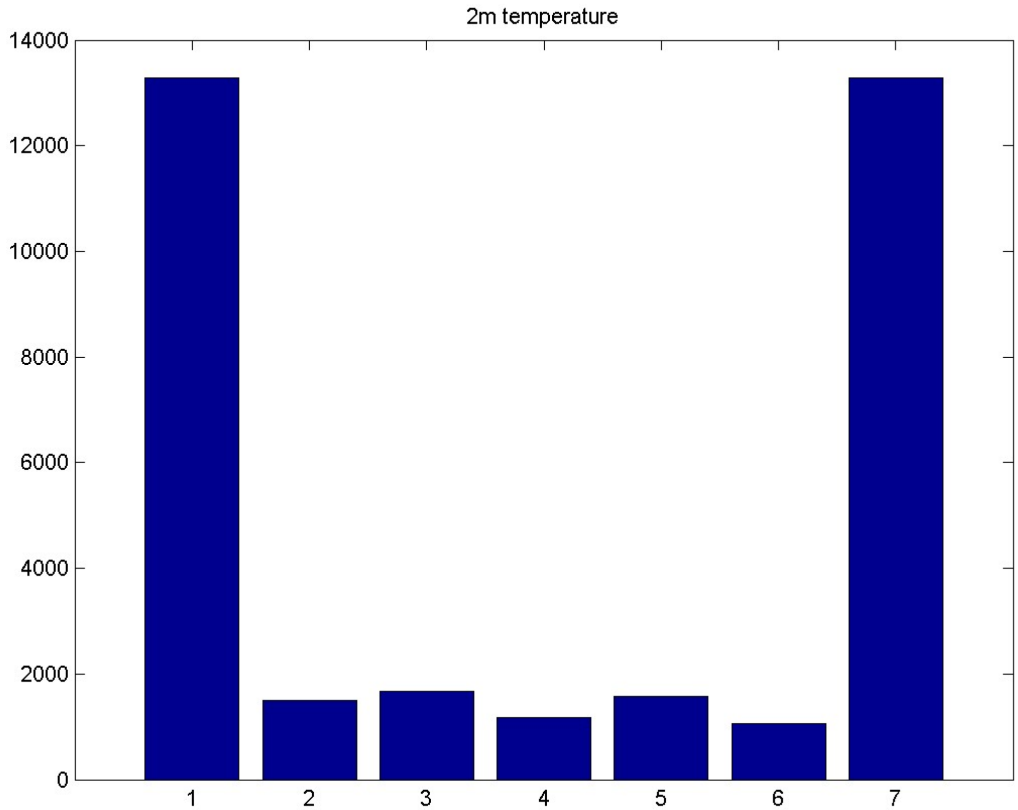


Figure 14: Rank histogram for all hourly near-surface temperature observations at 28 Climate Reference Network sites.

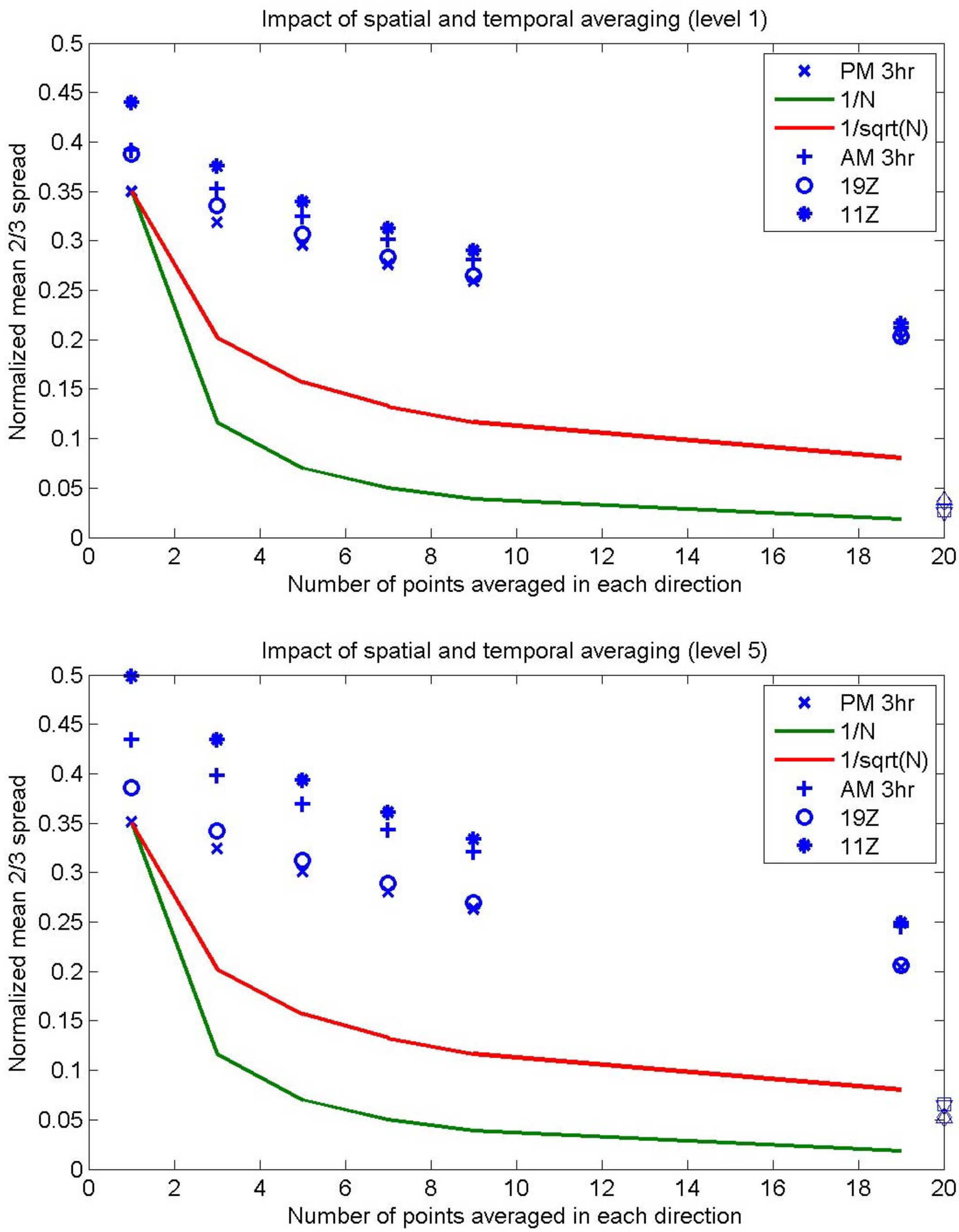


Figure 15: CO tracer spread as a function of averaging for surface (top) and 400-500 m AGL (bottom). The points (+ and x) for AM and PM 3h averaging without spatial averaging are the means shown in the figures and in the second column of table 5. Points on the right axis are for averages over the entire domain (216x236 points).