Analyzing "Gray Zone" Turbulent Kinetic Energy Predictions in the Boundary Layer from Three WRF PBL Schemes over New York City and Comparison with Aircraft Measurements

AUSTIN P. HOPE,[a,b] ISRAEL LOPEZ-COTO,[b,a] KRIS HAJNY,[a] JAY M. TOMLIN,[c] ROBERT KAESER,[c] BRIAN STIRM,[c] ANNA KARION,[b] AND PAUL B. SHEPSON[a,c]

[a] *Stony Brook University, State University of New York, Stony Brook, New York*
[b] *National Institute of Standards and Technology, Gaithersburg, Maryland*
[c] *Purdue University, West Lafayette, Indiana*

(Manuscript received 6 November 2022, in final form 6 November 2023, accepted 13 November 2023)

ABSTRACT: We investigated the ability of three planetary boundary layer (PBL) schemes in the Weather Research and Forecasting (WRF) Model to simulate boundary layer turbulence in the "gray zone" (i.e., resolutions from 100 m to 1 km). The three schemes chosen are the well-established MYNN PBL scheme and the two newest PBL schemes added to WRF: the three-dimensional scale-adaptive turbulent kinetic energy scheme (SMS-3DTKE) and the E–ε parameterization scheme (EEPS). The SMS-3DTKE scheme is designed to be scale aware and avoid the double counting of TKE in simulations within the gray zone. We evaluated their performance using aircraft measurements obtained during three research flights immediately downwind of Manhattan, New York City, New York. The MYNN PBL scheme simulates TKE best, despite not being scale aware and slightly underestimating TKE from observations, whereas the SMS-3DTKE scheme appears to be overly scale aware for the three flights examined, in particular, when combined with the MM5 surface layer scheme. The EEPS scheme significantly underestimates TKE, mostly in the elevated layers of the boundary layer. In addition, we examined the impact of flow over tall buildings on observed TKE and found that only the windiest day showed a significant increase in TKE directly downwind of Manhattan. This impact was not reproduced by any of the model configurations, regardless of the land-use data selected, although the better resolved National Land Cover Database (NLCD) land use led to a slight improvement of the spatial distribution of TKE, implying that more explicit representation of the impact of tall buildings may be needed to fully capture their impact on boundary layer turbulence.

SIGNIFICANCE STATEMENT: Because the majority of the world's population lives in cities, it is important to accurately simulate the atmosphere above and around these cities including the turbulence caused by tall buildings. This turbulence can significantly impact the mixing and dilution of air pollutants and other toxins in highly populated urban environments. The scale of cities often falls into what is known as the "gray zone" for turbulence modeling, which has been analyzed theoretically before but rarely in varied real-world conditions. Our analysis around New York City, New York, suggests that model turbulence schemes can match observations relatively well even at gray zone scales, although newer schemes require refinement, and all schemes tend to underestimate turbulence downwind of tall buildings.

KEYWORDS: Turbulence; Aircraft observations; Model evaluation/performance; Numerical weather prediction/forecasting

1. Introduction

Continued advances in computing power have allowed weather prediction models to steadily increase the resolution at which they run weather and air quality simulations. These advances have allowed mesoscale weather models and some climate models to run with resolutions nearing the scales relevant to cities and pollutant exposure spatial variability. There is increasing interest in urban greenhouse gas emissions estimation (Lauvaux et al. 2020; Lopez-Coto et al. 2020a; Yadav et al. 2021; Pitt et al. 2022), as well as environmental justice issues related to differential exposures to various pollutants (Banzhaf et al. 2019; Gardner-Frolick et al. 2022). Understanding such exposures requires high-resolution models for urban environments (Demetillo et al. 2021). Unfortunately, as model resolution shrinks to scales more appropriate for major urban centers, the physics for modeling turbulent eddies enters what is known as the "terra incognita" or "gray zone." In this gray zone, neither traditional mesoscale models nor large-eddy simulations (LES), the two traditional means of modeling turbulence, are fully appropriate (de Roode et al. 2019; Senel et al. 2020; Honnert et al. 2020, 2021; Juliano et al. 2022). In traditional mesoscale modeling, with model grid spacing greater than 1 km, PBL schemes parameterize the turbulence from eddies when calculating vertical mixing, using the assumption that most or all energy-containing turbulence is too small to be resolved given the grid spacing. Conversely, models with grid spacing below 100 m often use LES, working under the assumption that most or all eddies can be resolved by that smaller spacing (Giometto et al. 2016; Li et al. 2017; Yoshida et al. 2018; Kim et al.

Denotes content that is immediately available upon publication as open access.

Corresponding author: Austin P. Hope, austin.hope@stonybrook.edu

DOI: 10.1175/JAMC-D-22-0181.1

2023). As such, mesoscale PBL parameterizations may start double counting some turbulence as models enter the gray zone from the larger scales, while LES models will undercount turbulence at grid spacings above the typical LES resolutions of 100 m. As most turbulent kinetic energy (TKE) results from eddies ranging in size from 200 m to 5 km (Honnert et al. 2020), a lack of an accurate means for calculating TKE between 100 m and 1 km represents a significant potential shortcoming in modeling of TKE and related quantities.

The Weather Research and Forecasting (WRF) Model is designed to run at a large range of model grid resolutions and includes many options for modeling the PBL ranging from mesoscale parameterizations to LES schemes (Skamarock et al. 2019). This work aims to evaluate two of the newest PBL schemes in WRF, one of which is explicitly designed to bridge the gray zone issue. To evaluate these schemes, we utilized TKE measurements from three research flights over New York City (NYC), New York, performed with the Airborne Laboratory for Atmospheric Research (ALAR; Garman et al. 2006, 2008). We then simulated the data from those flight days with WRF using the three different PBL schemes: Mellor–Yamada–Nakanishi–Niino (MYNN; Nakanishi and Niino 2004, 2006; Olson and Brown 2009; Olson et al. 2019), a well-established and commonly used PBL scheme; the three-dimensional scale-adaptive turbulent kinetic energy scheme (SMS-3DTKE; Zhang et al. 2018; Wyngaard 2004), new to WRF 4.2 and designed to be scale aware, that is, reducing the amount of parameterized TKE as the scale decreases and more TKE is resolved; and the E–ε parameterization scheme (EEPS; Zhang et al. 2020), the newest PBL scheme, introduced in WRF 4.3. Because the focus here is to better understand the performance of these PBL schemes in the gray zone, we purposely did not add an urban canopy model (UCM) that could alter the behavior of the schemes (Lee et al. 2011; Teixeira et al. 2019; Lopez-Coto et al. 2020b). Instead, we experimented with different surface layer options and land-use representation. We then compared simulated TKE with that calculated from our flights' high-frequency wind data. This comparison was made for the full flight durations, as well as for the transects portion just immediately downwind of the (borough of) Manhattan skyline. Because these flights are low-altitude and through the middle of the greater NYC area, they are very well suited to capture the impacts of dense urban environments on turbulence and vertical mixing and yield better understanding of the performance of the modeling tools currently available.

Projects such as this that aim to compare modeled TKE to observations with a focus on gray zone resolutions in urban-influenced turbulence is relatively uncommon. To our knowledge, the use of aircraft-based TKE observations is unique. Section 2 of this paper details the aircraft instrumentation used to obtain high-frequency wind data and summarizes the flights used in this experiment. Section 3 outlines the WRF v4.3 setups that we use for our simulations. Section 4 then presents the results of our simulations in four subsections: section 4a, "simulation of meteorology," provides meteorological context; section 4b, "TKE as a function of horizontal resolution," to understand the specifics of the gray zone impacts in each model looking at locations along the flight path and as domain average vertical profiles of TKE and eddy diffusivity; section 4c, "TKE partition," to quantify the simulated TKE partitions; and section 4d, "simulating Manhattan TKE" to understand the spatial distribution and the impacts of the building-induced turbulence. Last, section 5 summarizes the main results and conclusions obtained.

2. TKE from observations

For this study, we calculated TKE data based on high-frequency three-dimensional wind velocity observations obtained using a turbulence probe from Purdue University's ALAR. ALAR is a Beechcraft Duchess research aircraft, flown and maintained by Purdue University in the experimental category, allowing for streamlined modifications of the class of instruments contained in ALAR at any time. The nose-mounted "best air turbulence" (BAT) probe uses nine pressure ports situated around a carbon-fiber dome to measure differential pressures during flight operations. These measurements, combined with the aircraft's inertial navigation and global positioning systems (INS/GPS), enable calculation of 3D wind velocity data relative to the ground and adjusted for the aircraft's sideslip and attack angle in all three component directions at roughly 50 Hz (Hacker and Crawford 1999; Garman et al. 2006, 2008). Specifically, the BAT probe combines measurements from nine pressure inlets and one sensor positioned around the dome end of the probe with data from the INS/GPS measurements using a potential flow model and a calibrated lookup table to then calculate the three wind speed components. This process removes any dependency on the heading of the aircraft. The probe is positioned at the nose of the aircraft to both limit the effect of the aircraft body flexing during flight and the upwash effect from the wings, though an upwash correction model is still necessary as nonnegligible upwash still occurs when lift is changing (Garman et al. 2006, 2008). The bulk of the data used in this study are taken from relatively level transects over the Hudson River, minimizing the effect of changing lift and upwash on our analysis.

We combine the wind data into TKE using Eq. (1), combining the variances [i.e., average (overbar) of the squares of the velocity component perturbations] of each directional component of the winds (u, v, w):

$$\mathrm{TKE} = \frac{1}{2}(\overline{u'^2} + \overline{v'^2} + \overline{w'^2}). \tag{1}$$

Because we wish to capture turbulence through gray zone scales (from 100 m to 1 km) and slightly above, we aggregate our 50-Hz wind data in 60-s intervals. This time interval allows our calculated TKE to capture eddies up to 4 km in size based on the typical speed of the aircraft, which is usually 60–70 m s^{-1}. This 4-km size maximum allows us to capture TKE from our observed wind data well past the spectral energy density maximum around 1 km, meaning we should be capturing all size scales of turbulent eddies. Using shorter time intervals (and thus a smaller sampling length scale) would mean we fail to capture the full turbulence energy spectrum, resulting in underestimated TKE. Conversely, using longer time intervals would sample both turbulent motion and spatial variability induced by differences in the larger-scale winds, and the derived TKE would be overestimated. For thoroughness, however, we tested the sensitivity of

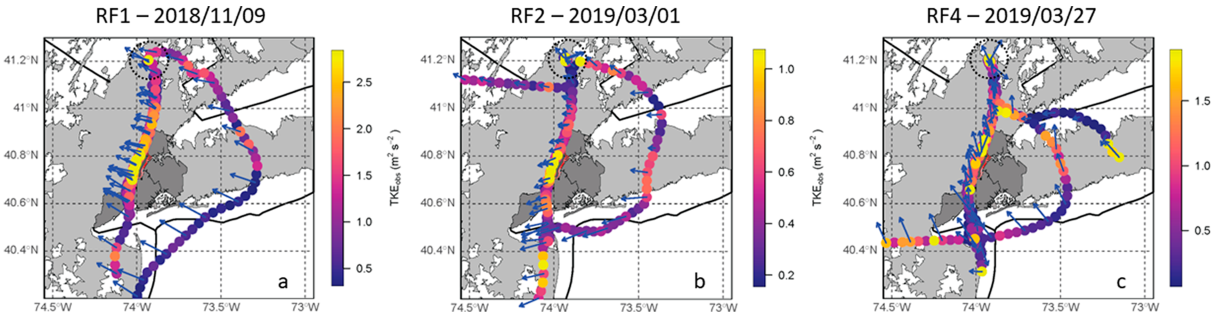

FIG. 1. Flight paths for the three ALAR flights analyzed in this paper. Color corresponds to the calculated 1-min TKE values along each flight path. Wind vectors (blue) show wind speed and direction, with wind vector length proportional across all three panels—the maximum 1-min wind speed among the three flights is 13.6 m s^{-1} in RF1; average wind speeds per flight are approximately 9 m s^{-1} in RF1 and approximately 4 m s^{-1} in both RF2 and RF4. The urban area is shown in gray, with the five boroughs of New York City shown in dark gray. The large, dashed circle at the top of each map at the north end of the Hudson River transects represents the location of the spiral vertical profile of ALAR for each flight.

our TKE calculations to the time interval. While there is a noticeable increase in TKE going from 10-s intervals to 30-s intervals due to the undersampling of the energy spectra at 10 s, the increase in TKE from 30 to 60 s is comparatively small, indicating that the full energy turbulence spectrum is likely well captured with our choice of 60-s intervals.

For this study, we examine TKE from three flights selected from a collection of nine ALAR flights conducted over NYC (Pitt et al. 2022; Hajny et al. 2022). These three flights were chosen because they are the only flights that contain multiple transits directly over the Hudson River with easterly winds, effectively capturing the impact of the tall buildings of Manhattan on TKE. The three flights took place on 9 November 2018 (RF1), 1 March 2019 (RF2), and 27 March 2019 (RF4), with each occurring in local afternoon. RF1 and RF2 were both on cloudy days with precipitation outside the flight windows, because NYC was between a low pressure system over the southern Appalachians and a high pressure offshore of Massachusetts each day. RF4 was on a moderately cloudy day, with a high pressure ridge over much of the mid-Atlantic region. Figure 1 shows each of the three flight paths, with the color of each point showing the 1-min TKE values. Wind vectors are also attached to every third point to display the direction and relative speed of the winds each day.

3. WRF simulations

We use the WRF Model, version 4.3 (Skamarock et al. 2019), to simulate the atmosphere on each of our flight days. Our four nested domains (Fig. 2) have resolutions ranging from 9 km in the outermost domain to 333 m in the innermost domain, decreasing by a factor of 3 in each domain, and our simulations represent 72 h of real time. The two innermost domains thus operate on resolutions within the gray zone, that is, below 1 km but above 100 m, with the innermost domain focused on the five boroughs of NYC. Our initial and boundary conditions for each WRF simulation come from the North American Mesoscale (NAM) 12-km analysis (NCEP/NWS/NOAA/U.S. Department of Commerce 2015) at 6-h intervals.

To streamline the calculation of resolved TKE from the model simulations, we also apply the velocity variance (VELVAR) diagnostic module (Lopez-Coto et al. 2020c) to our two innermost gray-zone-scale domains. This module computes the 3D velocity variances and mean winds at each grid cell during the model run between fixed time intervals. We chose a 15-min interval for the calculation to allow for turbulent eddies up to several kilometers in size to fully pass over any given grid point, at the mean simulated mesoscale wind speeds for our three flight days. This interval ensures that both our BAT probe data and WRF output are processed to include turbulence from eddies up to the same size.

Our model simulations vary in choice of land-use datasets, the surface layer physics schemes, and the PBL schemes. The

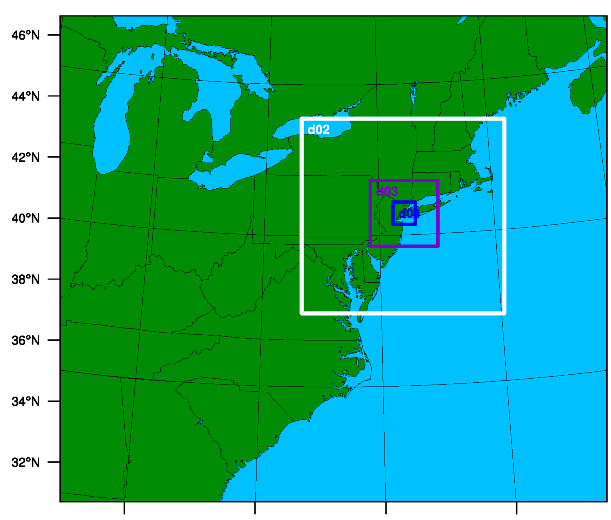

FIG. 2. Map of the four domains used in our WRF simulations, with the innermost domain covering the area under study, New York City. The outermost domain (full map) has a horizontal resolution of 9 km, the second domain (d02; white) is at 3 km, the third domain (d03; purple) is at 1 km, and the innermost domain (d04; blue) is at 333 m.

two land-use datasets we consider are the MODIS dataset (Friedl et al. 2002, 2010), which is the current WRF default, and the National Land Cover Database (NLCD) 2016 (Yang et al. 2018; Jin et al. 2019). The two surface layer physics schemes we consider are the Revised Fifth-Generation Mesoscale Model (MM5) surface layer scheme [with updates from Jiménez et al. (2012) and Fairall et al. (2003)] and the MYNN surface layer scheme (Nakanishi and Niino 2006). Last, the three PBL schemes we consider are MYNN (Nakanishi and Niino 2006; Olson et al. 2019), SMS-3DTKE (Zhang et al. 2018; Wyngaard 2004), and EEPS (Zhang et al. 2020).

The first PBL scheme we examine is the MYNN–eddy diffusivity/mass flux (EDMF) level-2.5 scheme. The foundation of the scheme comes from a second-order turbulence closure model (Mellor and Yamada 1982) with major improvements being made based on analysis with LES (Nakanish 2001; Nakanishi and Niino 2004, 2006). Specifically, the LES analyses informed the improvement of buoyancy-pressure effects in the PBL and the definition of the master length scale, as well as the slight altering of closure coefficients for better performance. The implementation of MYNN in the most recent versions of WRF also includes further length scale refinements, including the addition of a scale-aware mixing length (Ito et al. 2015) and the addition of an eddy mass-flux option to this scheme that provides some nonlocal characteristics (Olson et al. 2019; Angevine et al. 2020). The scale-aware mixing length makes the eddy-diffusivity component of the MYNN partially scale adaptive with respect to the model resolution (Olson et al. 2019). While there are other versions of the MYNN scheme available in WRF, we apply the commonly used level-2.5 MYNN scheme for our simulations.

The second PBL scheme we examine is the SMS-3DTKE scheme (Zhang et al. 2018). This scheme combines the horizontal and vertical subgrid turbulent mixing into a single energetically consistent framework and uses a partition function based on a dynamically calculated vertical master mixing length to smoothly transition between explicit LES scales and traditional mesoscale PBL scales. This combination eliminates the separation of vertical 1D PBL scheme equations and horizontal 2D mixing, creating a unified 3D set of equations, and provides scale awareness to the scheme. The partition function is based on previous LES analysis (Honnert et al. 2011) determining the ratio of resolved versus parameterized TKE that should exist over a range of grid scales spanning the gray zone. This scheme thus produces scale-aware TKE output by reducing the amount of subgrid TKE for simulations with small grid resolutions, as these also produce resolved TKE, ideally keeping total TKE relatively constant across a range of domains.

The third PBL scheme we examine is EEPS (Zhang et al. 2020). Unlike the SMS-3DTKE scheme, EEPS was not designed to bridge the gray zone [the smallest grid resolution used in the introductory Zhang et al. (2020) paper was 3 km]. We include EEPS in our analysis because of its recent addition to WRF in version 4.3 and its novelty regarding the turbulence dissipation rate formulation. EEPS differs from MYNN and many other E-l PBL schemes by using a prognostic equation for the dissipation term ε instead of a diagnostic equation based on the length scale l. This difference allows the scheme to more directly calculate many PBL processes and features, with EEPS performing particularly well in unstable atmospheric conditions (Zhang et al. 2020). We have tested both the version of EEPS included in the WRF v4.3 package and an updated version of the EEPS code we received from the developers of that PBL scheme. Our results are insensitive to the version of the EEPS scheme used.

Each parameterization provides a diagnostic of the PBL height, which we use when comparing modeled PBL height to that derived from the aircraft vertical profiles. MYNN uses a hybrid method that blends a θ_v-based definition in the neutral/convective boundary layer and a TKE-based definition in stable conditions. EEPS uses a method based on the bulk Richardson number, and the 3DTKE scheme uses a θ-based definition.

We consider two different land surface datasets to address concerns about the representation of NYC urban surface characteristics and their effect on turbulence. The default land-use classification used in WRF comes from MODIS. In MODIS, only one of the 21 land surface categories represents urban land cover, creating a relatively homogeneous surface for most of the NYC metropolitan area. Conversely, NLCD includes 4 different urban categories (of 40 available in the WRF implementation) each with different values for properties such as areal fractional coverage of green vegetation, albedo, and effective roughness length. The maximum surface roughness length value for the highest-urbanized category in NLCD (z_0 = 2.0 m) is 4 times the single urban surface roughness value in MODIS (z_0 = 0.5 m). The native resolution of NLCD is a much finer resolution than MODIS, that is, 30 m versus 500 m. Figure 3 shows how the land surface data appear in our innermost model domain, both gridded at 333-m resolution. In theory, NLCD thus allows for a more accurate representation of NYC, particularly with so many very tall buildings in Manhattan. We considered both MODIS and NLCD in this analysis to examine how the finer land surface representation affects TKE calculations.

In summary, this work includes 36 different simulations, or 12 for each of the three flight days (one in November 2018 and two in March 2019). These 12 represent three PBL schemes (MYNN, 3DTKE, and EEPS), two land-use datasets (MODIS and NLCD), and two surface layer physics schemes (MM5 and MYNN). For each flight day, we will label these configurations as 01–12, as defined in Table 1 below.

4. Results

a. Simulation of meteorological conditions

We first examine how well each of the 36 model simulations recreates the basic meteorology of each flight day. ALAR includes instruments that measure temperature and humidity in addition to the wind data from the BAT probe. We can also infer the height of the planetary boundary layer as the bottom of an elevated temperature inversion (Seidel et al. 2010) though these inferences could only be made once or twice per flight when ALAR performs vertical profiles.

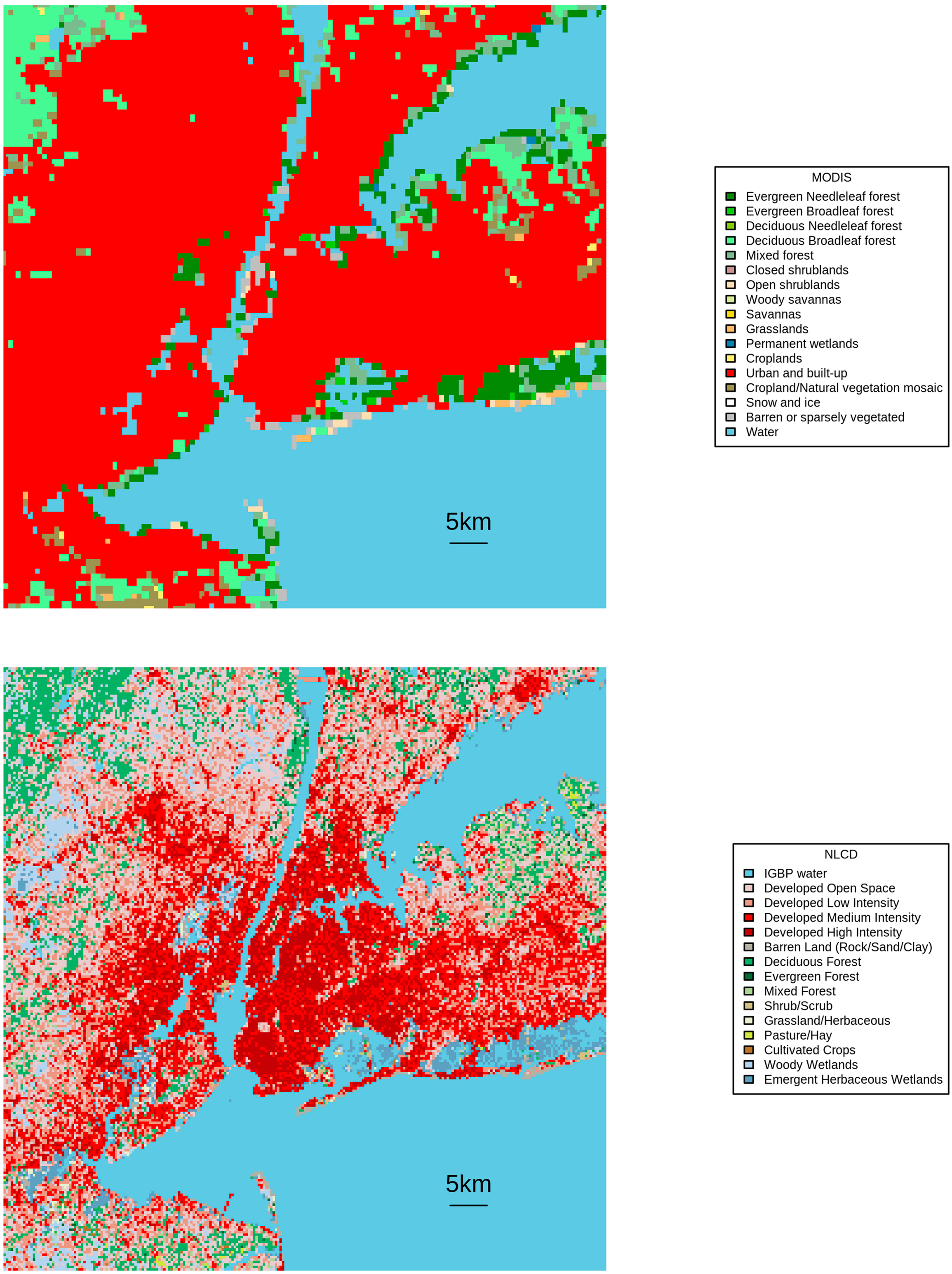

FIG. 3. Maps of the land surface categories in (top) MODIS (native 500-m resolution) and (bottom) NLCD (native 30-m resolution) for the innermost WRF domain in our analysis (both gridded and plotted at 333-m resolution). Light blue is water; dark blue is wetlands; greens are forested area; and reds are urban area, with the dark red being highly developed urban area in NLCD.

TABLE 1. Description of model configurations.

Configuration No.	PBL scheme	Land-use dataset	Surface layer scheme
01	MYNN-EDMF	MODIS	MYNN
02	SMS-3DTKE	MODIS	MYNN
03	EEPS	MODIS	MYNN
04	MYNN-EDMF	MODIS	MM5
05	SMS-3DTKE	MODIS	MM5
06	EEPS	MODIS	MM5
07	MYNN-EDMF	NLCD	MM5
08	SMS-3DTKE	NLCD	MM5
09	EEPS	NLCD	MM5
10	MYNN-EDMF	NLCD	MYNN
11	SMS-3DTKE	NLCD	MYNN
12	EEPS	NLCD	MYNN

Figure 4 shows these four meteorological comparisons—temperature, humidity, and wind speed at the aircraft's exact position, as well as PBL height—using boxplots to summarize each WRF simulation in the innermost domain along the ALAR flight track for each respective flight time. Overall, all WRF simulations produce results that largely match the observations from ALAR for each flight, although there are some notable exceptions. Most differences between model configurations occur with wind speed, often with the SMS-3DTKE scheme for RF1 and RF4 but also with the EEPS scheme for RF2. There is also some variability in water vapor for the 3DTKE scheme. The 3DTKE scheme produces lower wind speeds than observed for RF1 when paired with the MM5 surface layer scheme (configurations RF1-05 and RF1-08; Figs. 4f,j) and produces faster winds speeds in three RF4 cases (all except RF4-02; Fig. 4b) as well as for RF2-05. The 3DTKE scheme also occasionally produces drier air than observed, again when paired with the MM5 surface layer scheme, though this dryness is not consistent across the simulated flights (RF1-05, RF2-05, and RF1-08; Figs. 4g,k). It is also important to note that the observed potential temperature data here have been corrected to account for a known bias of roughly 3.5 K in the fast ultrasensitive temperature probe installed on the underside of the BAT probe on ALAR (Salmon et al. 2017).

A moderate match exists between observed and simulated PBL height (Fig. 4, fourth column), especially for RF1 and RF2. In WRF, the PBL height is calculated within each PBL scheme as described in the previous section. The main reason for the larger discrepancies observed in RF4 seems to be because these ALAR-based vertical profiles occurred at the northernmost points of each flight path (see Fig. 1), making these vertical profiles not necessarily representative of the PBL height for the entirety of each flight. The potentially nonrepresentative effect of this placement is made particularly clear in RF4, where the flight path through the WRF simulation often takes the virtual aircraft through two noticeably distinct air masses (not the case for RF1 and RF2)—one with a lower PBL height of 400–600 m and one with a greater PBL height of 1000–1200 m. The latter air mass exists over the inland area of our domain while the former air mass exists over water and coastal land. The split in PBL heights from these two air masses is reproduced by MYNN and EEPS for all configurations but fails to appear for 3 of 4 configurations using the 3DTKE scheme (RF4-05, RF4-08, and RF4-11). In addition, when the 3DTKE PBL scheme produces a single, noticeably larger PBL height than the other schemes that show the split, it does so in the simulations in which it also overpredicts wind speeds. With that said, only the 3DTKE scheme simulates the observed RF4 PBL height within the range of its data, which it does in all RF4 simulations, while the other two PBL schemes produce PBL-height values lower than the observed values, even as they simulate two air masses with noticeably different PBL thicknesses.

b. *TKE as a function of horizontal resolution*

Figure 5 shows boxplots of the TKE calculated from ALAR's BAT probe wind data compared to the total TKE from our WRF simulations along each flight track for all four domains from each simulation, limited to the extent of the innermost domain (i.e., to see the impact of coarser resolution on the results). Total TKE from WRF is calculated as the subgrid TKE determined by the PBL schemes (Fig. 6) plus the resolved TKE based on WRF-simulated winds for the domains where the VELVAR module is active. The MYNN PBL scheme simulated TKE values generally agree the best with the observed TKE, while the EEPS PBL scheme consistently underestimates total TKE. The 3DTKE PBL scheme varies, often producing similar results to the MYNN PBL scheme, but also significantly overestimating TKE in some simulations where it also overestimates wind speed and PBL height. It similarly produces too high TKE in the outermost domain in the RF4 simulation when using MODIS land-use and the MYNN surface scheme (Fig. 5a).

Figure 5 also illustrates the scale awareness (or lack thereof) for each of the three PBL schemes. Neither the MYNN scheme nor the EEPS scheme is designed to be scale aware for TKE calculations, although the modified mixing length scale in the recent version of MYNN should give it some scale awareness, as mentioned in section 3. A lack of programmed scale awareness means that we expect that the calculated subgrid TKE should remain roughly constant across all four resolutions while resolved TKE from simulated winds will grow as grid resolution shrinks due to the smaller grid size being able to explicitly resolve smaller turbulent eddies. We see this clearly in Fig. 5, with TKE from the 333-m domain often having the highest mean, median, and interquartile range (IQR) within each MYNN and EEPS grouping, usually having increased consistently from the 3-km domain. In

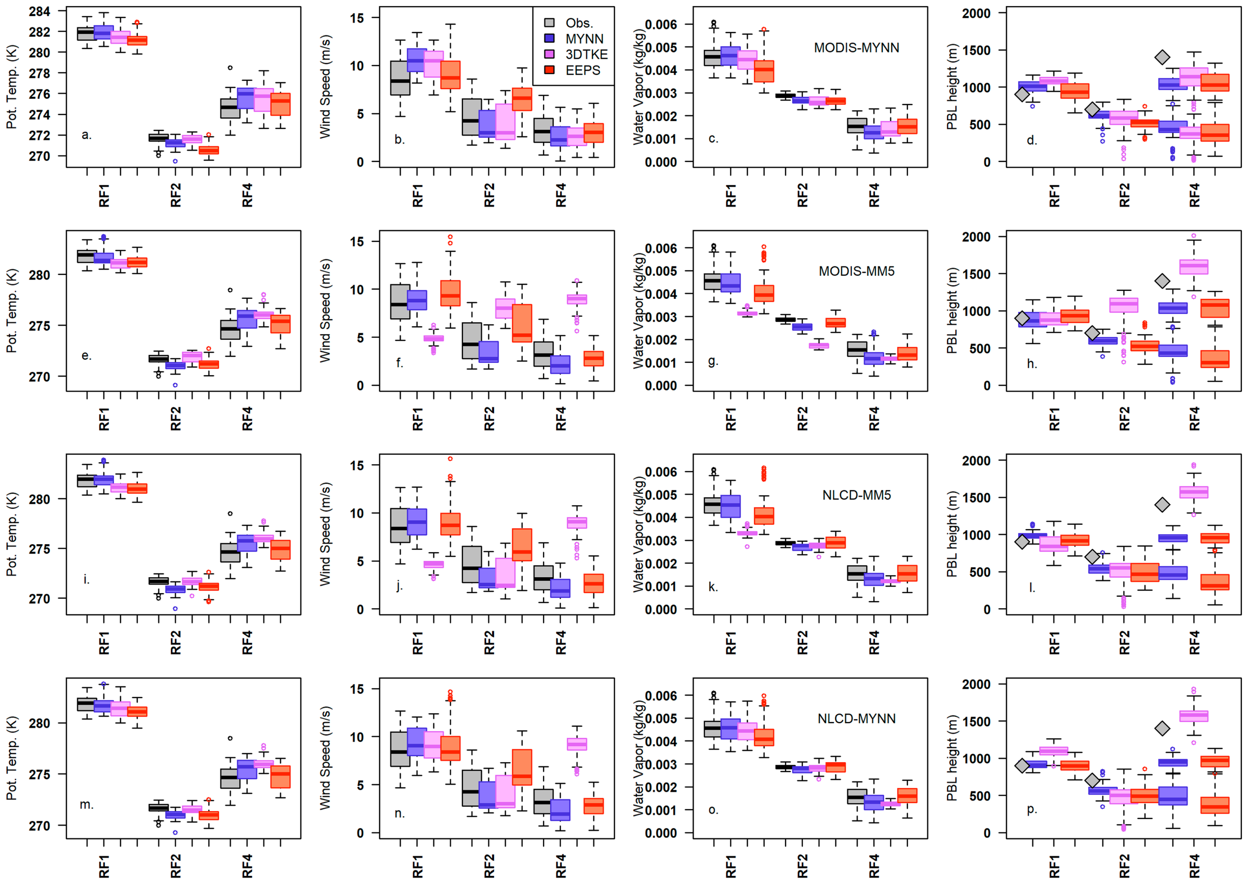

FIG. 4. Box-and-whisker plots of various meteorological comparisons between ALAR flights and WRF simulations (333-m-resolution domain): (left) potential temperature relative to standard at ground level, (left center) wind speed, (right center) humidity, and (right) PBL height. In RF4 model simulations aside from three of the four 3DTKE runs, PBL-height data are divided for overland (thicker PBL) and overwater (thinner PBL) air masses. PBL heights inferred from ALAR flights are single data points obtained from vertical profiles. Data from ALAR are shown in gray; WRF simulations using the MYNN, 3DTKE, and EEPS schemes are shown in blue, light purple, and red, respectively. Each row represents one of our four land-use-data+surface-layer-scheme combinations, which are named in the right-center column. Each box shows the first quartile, median, and third quartile of the data from each flight, and the whiskers and circles show the remaining data. Any data points more than 1.5 times the IQR outside the IQR are shown as circles beyond the extent of the whiskers. Otherwise, the whiskers span the entire range from each respective flight or WRF simulation. Versions of this figure using data from the larger WRF domains are nearly identical.

some cases, however, especially with EEPS, the highest or second highest TKE occurs for the coarsest 9-km resolution.

The SMS-3DTKE scheme is designed to be scale aware, meaning it should reduce subgrid TKE as the resolved TKE grows in smaller-resolution domains, leaving the total to be roughly constant. While this goal of constant TKE is largely realized for the MODIS-MYNN surface configuration (Fig. 5a), there is a large exception for configuration 02 in the outermost domain of RF4 producing oddly high TKE. Then, for all other surface configurations, we clearly see that the d01–d04 means, medians, and IQRs among the three sets of 3DTKE results are trending downward for eight of the other nine model simulations (configurations 05, 08, and 11 for the three flight days). In four of those eight simulations, the erroneous trend derives from the overestimated TKE in the coarser domains as the simulated TKE from the 333-m domain generally best matches that from observed winds. In the other four simulations, the 9-km domain has TKE closest to observed, although the decreasing TKE with increasing resolution remains within the range of observations. This trend suggests that the equations governing the scale-aware representation in the 3DKTE scheme might be tuned slightly too aggressively for the conditions of these three flight days, with subgrid TKE (Fig. 6) also being too high in the mesoscale domains for half of the 3DTKE simulations. That is, while the 3DTKE scheme directly includes equations that account for grid resolution by explicitly reducing subgrid TKE while resolved TKE increases, these equations are reducing subgrid TKE far faster than resolved TKE is increasing as grid size gets smaller. The total TKE in our 333-m domains for several 3DTKE simulations approximately matches observed TKE only because the 3DTKE scheme is providing far too much subgrid TKE in the coarser domains and then scaling that subgrid TKE down very strongly. This result might indicate that the resolution dependency used in the partition function is being overemphasized in the current implementation of the function.

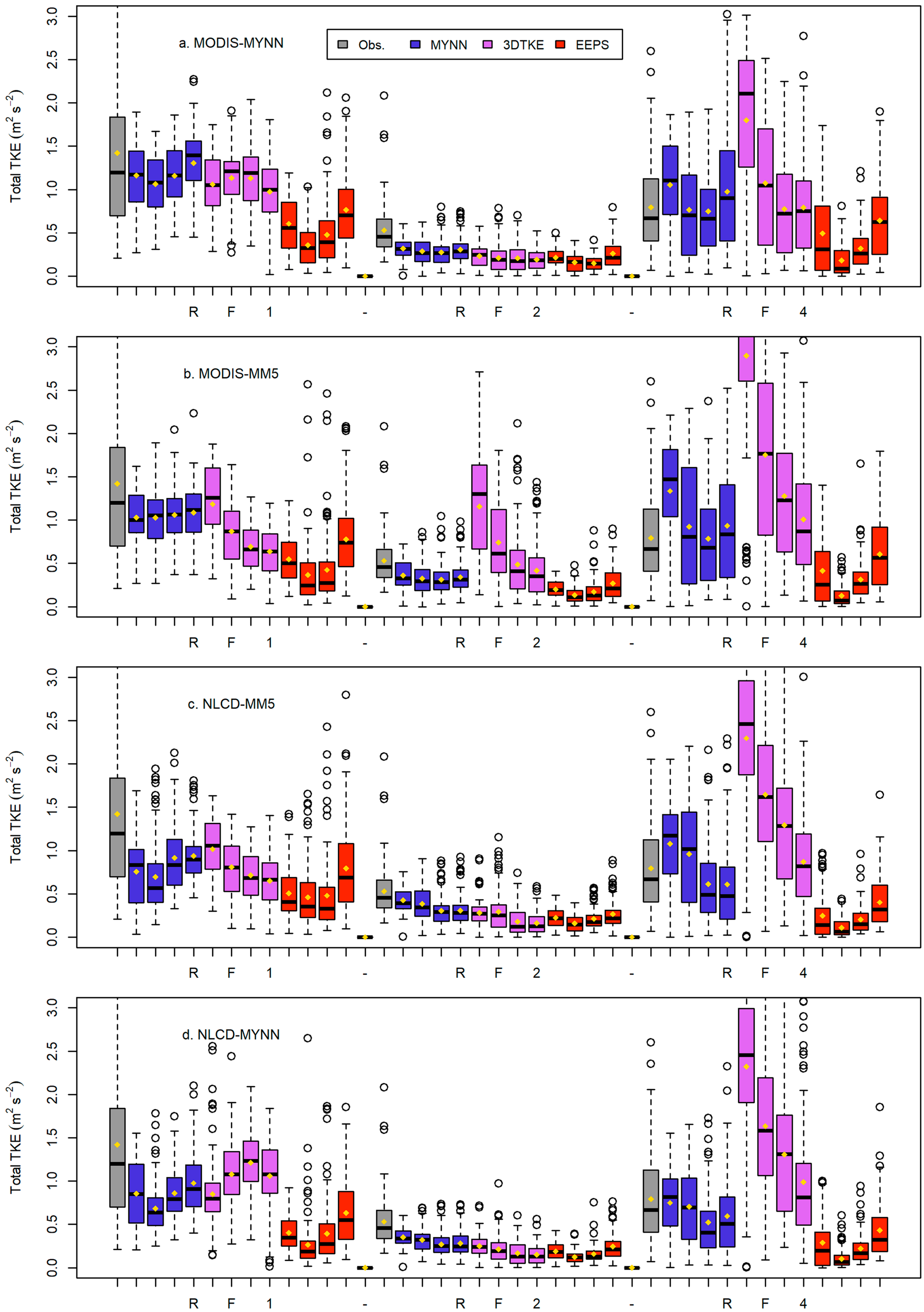

FIG. 5. Box-and-whisker plots showing TKE from our flights (gray) and from the model simulations (blue, light purple, and red, similar to Fig. 4). Gold diamonds represent the mean TKE along the flight track. The four box-and-whisker plots for each color of WRF data represent the TKE calculated from the four nested domains of each simulation, from our outermost domain (d01; 9-km grid resolution) to our innermost domain (d04; 333 m) from left to right for each of the flight days. Measurements for which the aircraft was above the boundary layer have been excluded, as have measurements for which the plane was outside the extent of the innermost domain.

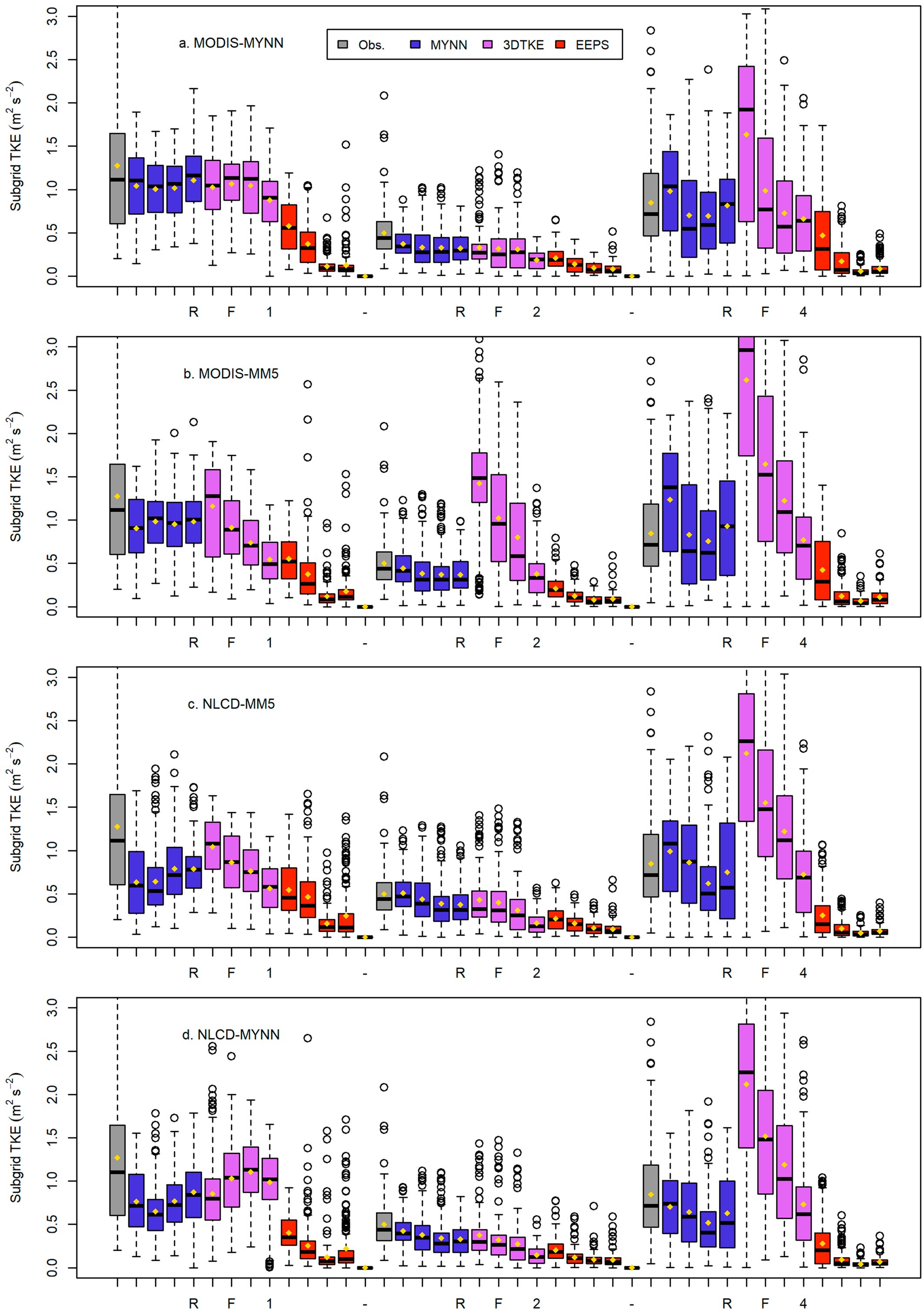

FIG. 6. As in Fig. 5, but focusing on subgrid TKE instead of total TKE, i.e., removing the resolved TKE from Fig. 5.

The effects of changing the land-use and surface layer schemes are less noticeable with the other two PBL schemes, though it does appear that the MYNN surface layer scheme might produce less subgrid TKE with the NLCD land use than with the MODIS land use, in particular, for coarser domains. One possibility for this modeled decrease in simulated TKE could be that while there are larger urban roughness values in NLCD, the finer resolution also allows for more representation

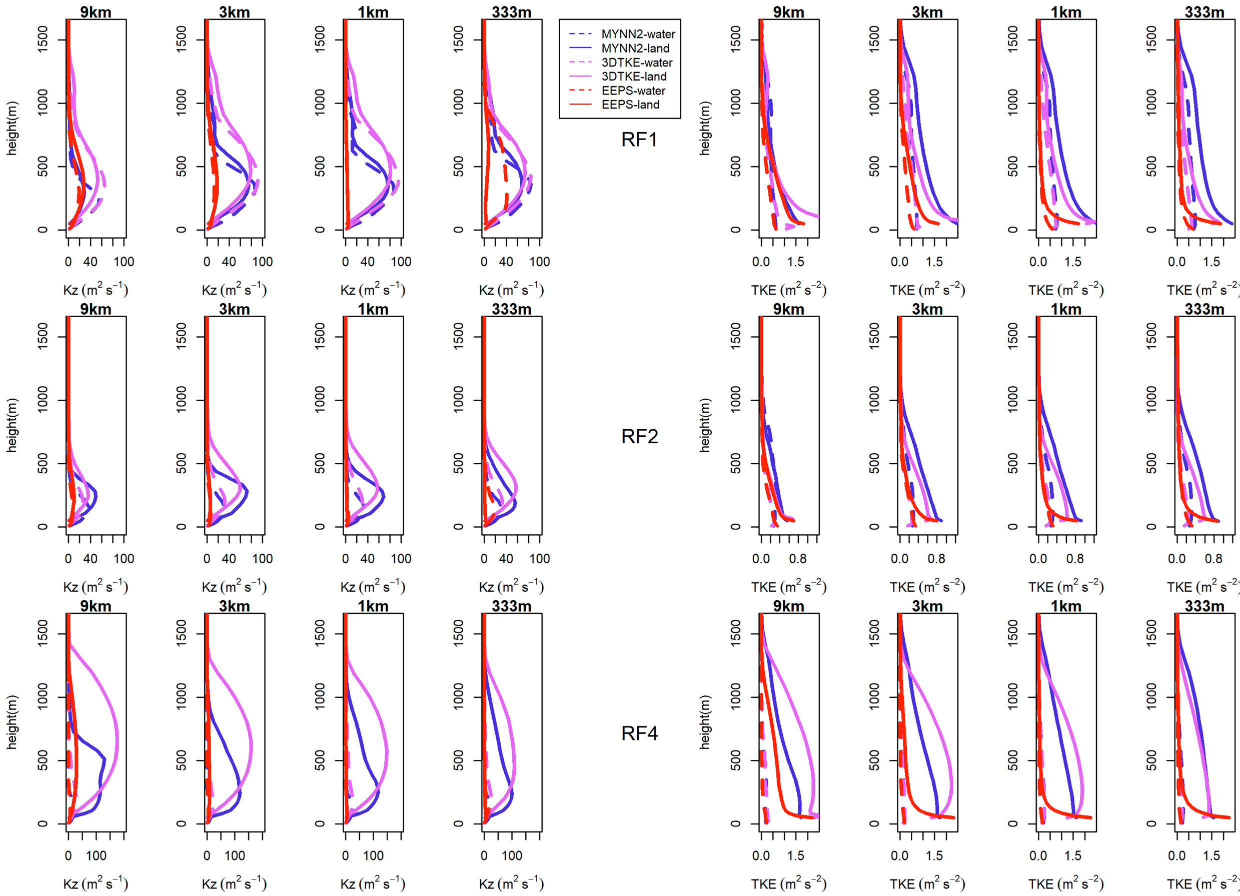

FIG. 7. Area-average profiles from WRF at 1830 UTC (1330 local time for RF1 and RF2; 1430 LT for RF4) for the (left) vertical exchange coefficient and (right) subgrid TKE for all RFs (separated by row) using MODIS land-use data and the MYNN surface layer scheme (model configurations 01–03). The color scheme matches that of previous figures. Dashed lines represent area averages over water-surface grid points in the extent of our innermost domain, and solid lines represent the same but over land surface grid points.

of the lower-roughness suburban and rural areas upwind, lowering the overall TKE. Another possibility is that the varied surfaces in NLCD allow for different meteorology above the city—such as a greater extent of cloud cover than with MODIS—decreasing total energy in the system.

In addition to analyzing WRF TKE along the simulated flight tracks of ALAR, we also compute vertical profiles of subgrid TKE and the vertical exchange coefficient (Kz) as area averages. These area averages are calculated at each model height for all domains over the horizontal extent of the innermost domain. Kz is obtained directly from WRF simulation output, specifically the variable "EXCH_H" for the MYNN and EEPS schemes and "XKHV" for the 3DTKE scheme. For the TKE schemes without explicit scale awareness, it takes the form of $\mathrm{Kz} = sL\sqrt{\mathrm{TKE}}$, where s is a stability function and L is a length scale. For the 3DTKE scheme, it takes a similar form but includes grid resolution considerations in the blending of the mesoscale and LES limits of the length scale (Zhang et al. 2018). Figure 7 shows these area averages from model configurations 01–03 (all MODIS-land-MYNN-surface simulations) as profiles, where the value at each pressure level is shown for the average height for that pressure level. We also separate the area averages into two gridcell types, computing the averages over land and over water distinctly.

In the subgrid TKE profiles of Fig. 7, we see the scale awareness of the 3DTKE scheme with lower values of subgrid TKE in the smaller-scale domains, particularly at the lower altitudes in each profile. We can also see a slight scale-aware aspect to the exchange coefficients in the MYNN simulations.

These profiles also give some insight as to why the EEPS results generally fail to match observations from ALAR. Within the EEPS simulations of Fig. 7, the surface-level subgrid TKE appears mostly in line with surface-level TKE from the simulations using the other PBL schemes, but the subgrid TKE from EEPS drops very quickly with increasing altitude, in contrast to the profiles from the other two schemes. This drop-off is particularly pronounced in the gray-zone-resolution domains, approaching near zero just a couple hundred meters above the model surface, both over land and over water. The larger amounts of resolved TKE do not sufficiently compensate for the near-zero subgrid TKE values in these domains. The EEPS scheme still performs moderately well in our outer two domains (3- and 9-km resolutions), providing a partial match to previous evidence of strong performance with model simulations using a single 3-km grid domain (Zhang et al. 2020).

The exchange coefficients within the EEPS simulation also tend to remain near zero through the entire height of the model atmosphere instead of reaching a maximum in the mid- to lower levels of the PBL as it does with the other schemes. The EEPS scheme does produce noticeably nonzero coefficients, but only a moderate maximum in our outermost domain and a relatively strong maximum specifically over water in our innermost domain. These qualities of the EEPS PBL scheme are consistent across all flight days and all surface configurations.

As shown in Fig. 7, the 3DTKE results are also consistent across all flight days and all surface configurations. This consistency in the 3DTKE profiles is in slight contrast to the greater variability from TKE along the flight track itself (Figs. 5 and 6) comparing across surface schemes for a given flight. Two clear atypical cases are RF4-02, the MODIS-MYNN surface configuration, and RF2-05, the MODIS-MM5 configuration, with the third possible exception being for RF1-11, the NLCD-MYNN configuration. For all three cases, the modeled PBL height (Fig. 4, right column) is noticeably different for these exception surface combinations than for the other three combinations on that day. Consequently, between RF2 and RF4, the simulations where 3DTKE significantly overestimates subgrid TKE in the coarser domains also overestimate PBL heights. While there is no clear explanation for why RF4-02 has more consistent TKE across domains for RF4 than the other three surface configurations or why RF1-11 has increasing TKE for finer domains, the RF2-05 case is readily explainable based on the spatial distribution of high TKE in the RF2 3DTKE simulations (regardless of surface and land use). In those WRF runs, there is an area of high TKE immediately north of New York City (not shown) in contrast to much lower TKE around most of the city itself. This highlights the importance of the spatial distribution of the turbulence and shows that differences in the location of the measurements can significantly change the sampled TKE.

c. TKE partition

Figure 8 shows the fraction of total TKE that is directly resolved by WRF in the 1- and 333-m-resolution domains, taken along the ALAR flight tracks within the extent of our innermost domain. As expected, this ratio is largest for the domain with the highest resolution, due to the model's ability to resolve more and smaller turbulent eddies. This expected pattern holds true whether the subgrid TKE remains roughly constant across resolutions or decreases in a scale-aware system. The decrease in subgrid TKE in the scale-aware 3DTKE PBL scheme should cause this partition ratio to rise more than in non-scale-aware PBL schemes, and we see this in the comparison with the MYNN and 3DTKE results. Notably, resolved TKE accounts for less than 20% of the total TKE along at least 80% of each flight path, even in our 333-m grid-resolution domain, again with the exceptions of the three RF4 3DTKE simulations that also had strong high bias for the wind speed.

For RF1 and RF2, we can also see a general increase in the TKE ratio when using the MM5 surface layer scheme when compared to the MYNN surface layer scheme. The 3DTKE PBL scheme seems to show the pattern the most consistently, with one exception in the EEPS results and two with MYNN. For MYNN and 3DTKE, this pattern also affects our 333-m domain more than the 1-km domain, providing a greater difference in ratios between the two resolutions, and showing the impact of the choice of surface layer scheme in the TKE partition. The RF4 ratios do not show a consistent difference between the two surface layer schemes.

The partition ratios for our EEPS simulations are significantly greater than those from the other two PBL schemes. These large values are driven by both more resolved TKE and near-zero subgrid TKE as discussed in the previous section. As before, this pattern holds true regardless of which surface type data and surface layer scheme are used.

d. Simulating Manhattan TKE

As we have discussed, the overall performance of the MYNN PBL scheme in reproducing the observed TKE along the flight path is mostly satisfactory in aggregate, with the mean, median, and range of MYNN-simulated TKE closely matching the observations (e.g., Fig. 5). Indeed, Fig. 9 (scatterplots of simulated versus observed TKE) clearly shows that for MYNN, the predicted total TKE falls mostly around the 1:1 line when compared with the observations along the flight tracks. For all MYNN simulations, the TKE from our innermost domain has a bias (model minus observations) of -0.18 m^2 s^{-2} and an RMSE of 0.68 m^2 s^{-2}; the next two domains have similar if slightly larger values for both statistics. In addition, it shows that this is the case for different surface types as the model can clearly differentiate measurements taken over water (blue), over low-rise urban and suburban (green), and over the Hudson River directly downwind of Manhattan (red), which means that the general spatial distribution of the TKE is also well reproduced. We can also directly see the increase in TKE from the 3DTKE scheme (mean bias from -0.32 to $+0.03$ m^2 s^{-2}) and decrease in TKE from the EEPS scheme (mean bias from -0.44 to -0.67 m^2 s^{-2}) as we go from the 333-m domain to the 3-km domain, as noted earlier. Importantly, we can see that not all measurements downwind of Manhattan are equally well reproduced and that a few points with very high observed TKE stand as outliers that the model cannot simulate. The mean bias per panel for the Hudson transect points in Fig. 9 range from -0.41 to -1.03 m^2 s^{-2} with RMSE values from 0.94 to 1.37 m^2 s^{-2}. Contrast this mismatch in TKE immediately downwind of Manhattan to the corresponding statistics for TKE values over water, with mean biases from $+0.06$ to -0.51 m^2 s^{-2} and RSME values from 0.51 to 0.80 m^2 s^{-2}.

To better understand the origin of these high TKE observations, Fig. 10 shows observed and modeled TKE as a function of height and latitude along the flight track immediately downwind of Manhattan for RF1, that is, the multiple passes

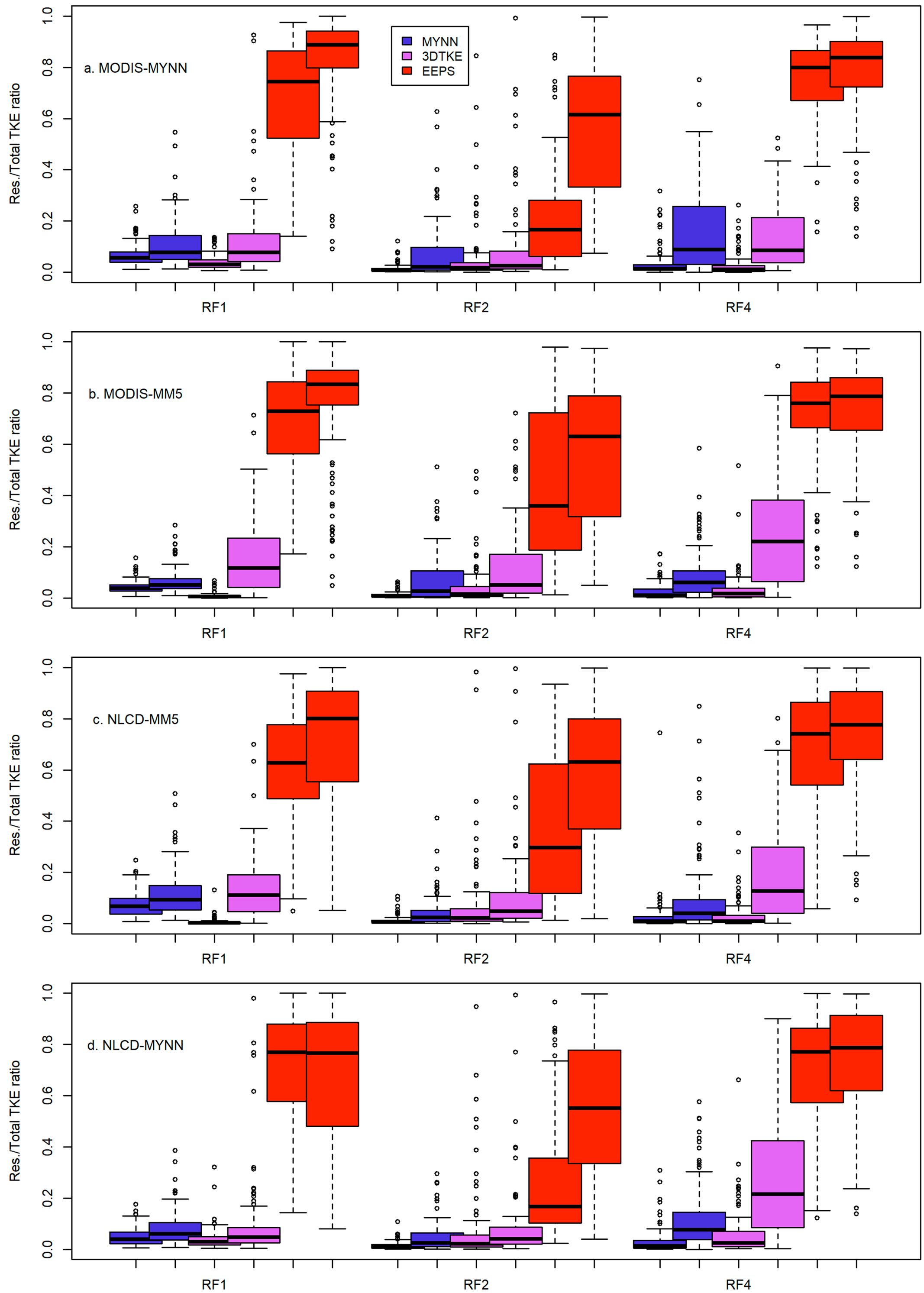

FIG. 8. Box-and-whisker plots showing the partition ratio within our model simulations (blue, magenta, and red, as in Fig. 4) for our two innermost domains (1 km on the left and 333 m on the right in each pairing) taken along the ALAR flight tracks within the extent of our innermost domain. Partition ratio is defined here as the amount of resolved TKE divided by the total (resolved + subgrid) TKE.

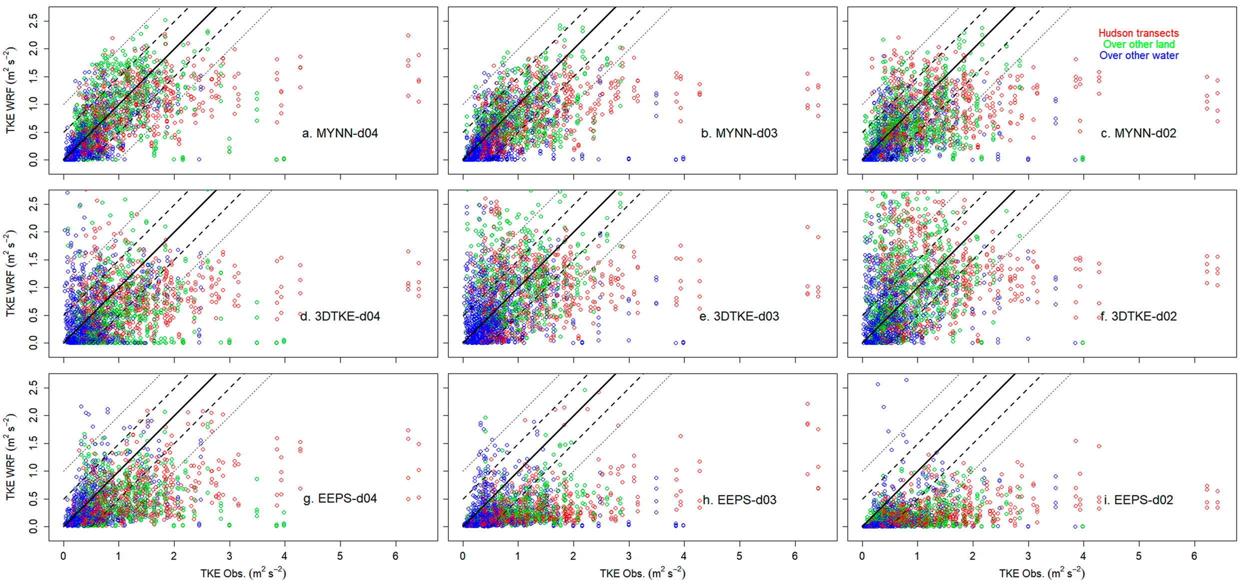

FIG. 9. Scatterplots of total TKE from the three innermost domains (333-m grid size in d04 to 3 km in d02) of all WRF simulations in this study, including the three days and all configurations, separated by PBL scheme, as a function of the observed TKE along the flight paths. Red points represent TKE from our transects over the Hudson River, i.e., from 40.6° to 40.9°N; green represents all other points over land, and blue represents all other points over open water. The one-to-one line is shown (thick solid) with ±0.5 and ±1 m^2 s^{-2} as dashed and dotted lines, respectively.

over the Hudson River shown in Fig. 1a and the red points in Fig. 9. Figures 11 and 12 show the same as Fig. 10, but from RF2 and RF4, respectively. The discussion here will focus on RF1, as RF2 and RF4 respectively show partially or significantly weaker TKE increases from Manhattan based on the observed winds. For all RFs, the maximum values of TKE generally occur at lower altitudes between 40.7° and 40.8°N.

The maximum observed TKE values correspond directly to downtown Manhattan, which has roughly 100 buildings that are at least 200 m tall with several over 500 m tall, meaning we are observing the small-scale turbulent eddies behind these very tall buildings. The much larger building impacts on the observations for RF1 relative to the other two RFs are likely caused by the stronger winds as compared to the other two days (Fig. 4). Also, due to the lack of subgrid TKE simulated within the EEPS scheme in our gray-zone-resolution domains, we will focus on the other two schemes for this analysis.

Neither the MYNN PBL scheme nor the 3DTKE PBL scheme produces TKE values immediately downwind of Manhattan as high as the observed values from ALAR, in particular, for RF1 when the building-induced turbulence is significant in our measurements. This local mismatch implies that the generally good matches across the entire flight path (e.g., Fig. 5) are more influenced by TKE above low-rise urban, suburban, and rural areas with lower surface roughness, as the simulated TKE downwind of the very tall buildings was considerably underestimated for both WRF schemes. The MYNN PBL scheme tends to produce more TKE in the Manhattan downwind region (especially for the lowest altitudes) than the 3DTKE scheme does, with the MODIS-MYNN surface configuration (Fig. 10b) producing TKE values very similar to those from observations through most of the curtain, except for the lower altitudes immediately downwind of Manhattan. Even if the MYNN scheme does not produce TKE values above 2 m^2 s^{-2} in any of these simulations, it mostly places its highest TKE values in the lower altitudes between 40.7° and 40.8°N. The lowest values in the MYNN PBL simulations also tend to occur at higher altitudes near the northern and southern edges of the curtains, which is a good match to TKE from observations. The 3DTKE PBL simulations show a weaker vertical gradient in TKE values than the MYNN simulations do, and only with the NLCD-MYNN surface configuration (Fig. 10i) does the 3DTKE PBL scheme seem to produce a noticeable maximum at lower altitudes downwind of Manhattan. It appears that the MYNN scheme is responding to the general turbulence created by the larger New York metropolitan area better than the 3DTKE scheme is, even though neither PBL scheme is fully capturing the impacts from Manhattan's very tall buildings.

Looking at the effect of our surface configurations on the TKE results also for RF1, we again see that the combination of the 3DTKE PBL scheme with the MM5 surface layer scheme (Figs. 10e,g) produces results significantly below the observed TKE values downwind of NYC for the innermost domain (333-m grid resolution). The choice of the surface layer scheme has much less effect on MYNN PBL simulations, though we can see some decrease in the resulting TKE between the MODIS-MYNN simulation (Fig. 10b) and the MODIS-MM5 simulation (Fig. 10d).

Comparing results from MODIS land-use simulations to NLCD land use, there is a negligible change of TKE in some cases and an overall decrease from NLCD in others. As the

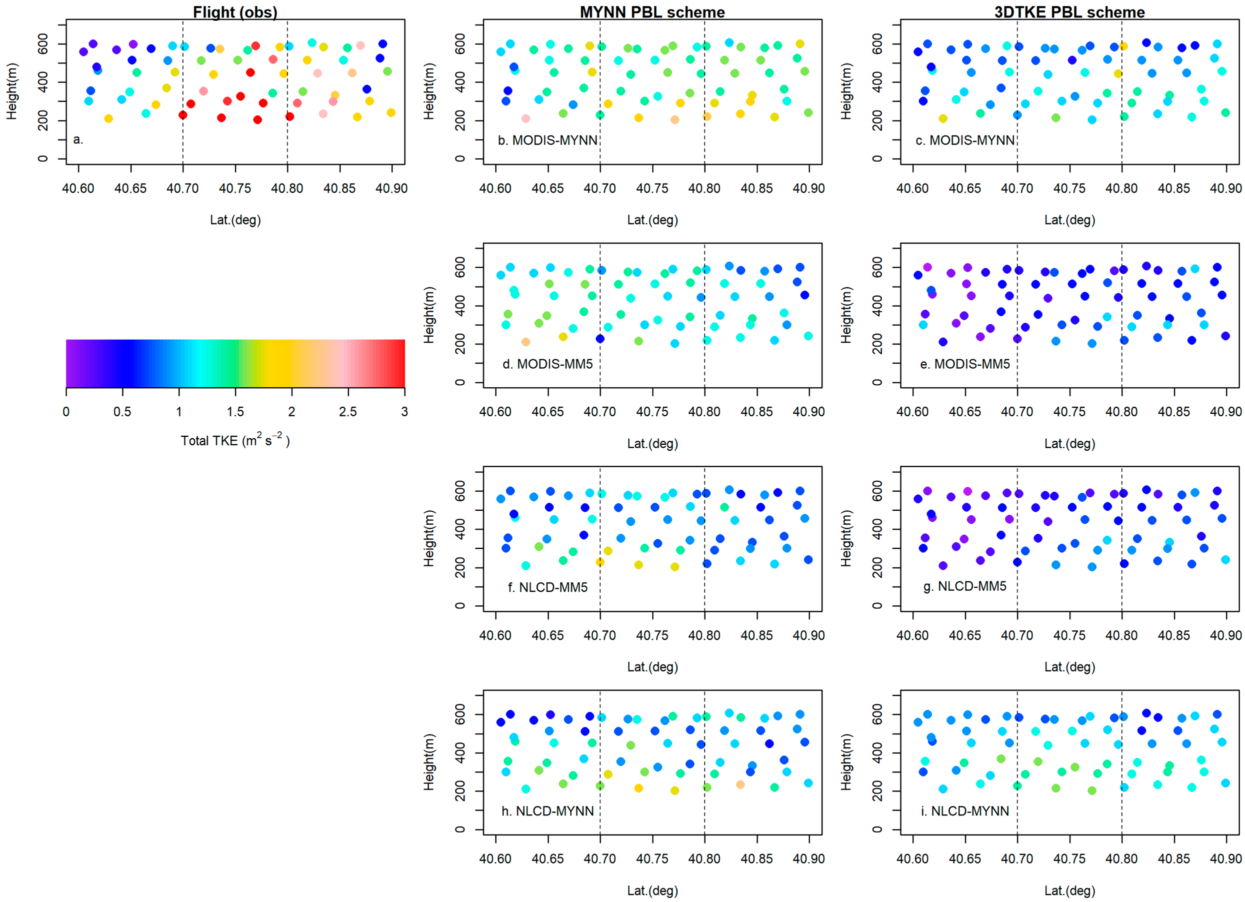

FIG. 10. Curtain plots of total TKE from RF1 in our innermost domain (333-m grid resolution) from the (center) MYNN and (right) 3DTK WRF simulations with all surface combinations. Each circle represents one of the 1 min data points from transects flown over the Hudson River. For reference, the borough of Manhattan spans roughly 40.7°–40.8°N, i.e., the middle one-third of each plot, marked here with dashed vertical lines. Several points in the (left) observations have TKE well above 3 m^2 s^{-2}, with the color bar here capped to show as many features as possible among the various curtains.

more heterogeneous nature of the NLCD surface categorization and the larger values for surface roughness in the urban categories both would suggest an increase in turbulence in the atmosphere above and downwind of NYC, this negligible to negative change was unexpected. This issue was discussed previously in section 4b (e.g., Figs. 5 and 6) where we saw this decrease in TKE from MODIS to NLCD for MYNN PBL scheme simulations. For the RF1 MYNN PBL cases with NLCD shown here (Figs. 9f,h), the decrease in overall TKE could be due to an increase in cloud cover in the simulations, reducing the energy flux into the system and thus depressing TKE, as well as the presence of less urban areas upwind of the city.

Notably, while these NLCD simulations have an overall decrease in TKE across the curtains, the TKE immediately downwind of Manhattan (i.e., center region lower levels of each curtain) is similar to or greater than the TKE in the same area for the MODIS simulations. Similarly, across these lower altitudes, the relative increase in TKE due to Manhattan is generally more pronounced in the NLCD simulations. This local difference suggests that the increased heterogeneity and larger roughness values are producing the expected greater TKE locally, even while the TKE values at higher altitudes and farther from the city are decreasing because of the increase in cloud cover and the more heterogeneous upwind areas.

We have also conducted several WRF simulations using a modification of the NLCD land-use data where we increased the surface roughness of the urban categories by roughly 25% to test if this change had any noticeable effect on the TKE produced (not shown). While some of these simulations showed a small increase in TKE, others showed a small decrease. Even in the cases with increases, the change was insignificant compared to the difference between the WRF-simulated TKE and the TKE from observations, particularly immediately downwind of Manhattan. As such, the WRF land-use categorization alone is not able to account for the observed high TKE immediately downwind of Manhattan, and likely a more explicit representation of the buildings in the model would be needed to account for the building-induced turbulence.

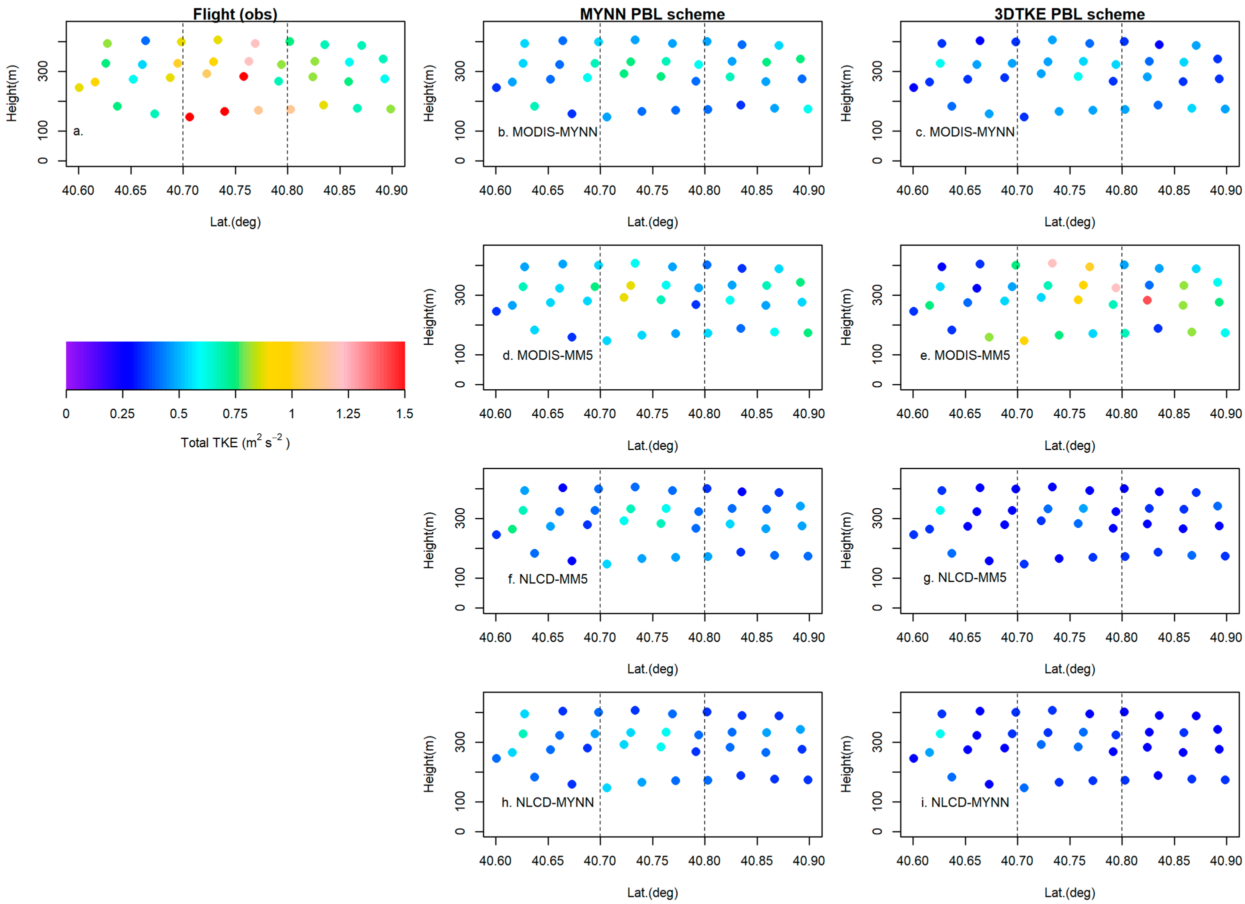

FIG. 11. As in Fig. 10, but for RF2, with adjusted color scale.

The above discussion comparing schemes and land-use classification also generally holds true for RF2 and RF4 (Figs. 11 and 12), but with generally lower values of observed TKE, indicating that the building-induced turbulence is much less significant, and so the WRF simulations more successfully match the observations. The WRF-simulated TKE from RF4 is quite similar to observations, even in the low altitudes between 40.7° and 40.8°N. The overall lower TKE values that allow for this match, however, are likely because of the wind direction, which in RF4 is southeasterly instead of more directly easterly in the other two flights (Fig. 1), as well as generally lower wind speeds. Taken together, these two facts about the RF4 winds result in less production turbulence due to Manhattan's buildings and slower advection of the turbulence to the flight positions over the Hudson River, relative to the other two flights.

5. Summary and conclusions

We investigated the ability to simulate boundary layer turbulence in the "gray zone" in a complex urban–water interface, as well as the impacts of the building-induced turbulence. We tested three different WRF PBL schemes combined with several model surface configurations by comparing with observations from three separate research aircraft flights over New York City.

We found that the commonly used MYNN PBL scheme simulates TKE values that generally agree the best with the observed values, including at the gray zone resolutions we tested (333 m and 1 km). On the other hand, the EEPS PBL scheme consistently underestimates total TKE and the SMS-3DTKE PBL scheme varies, often producing similar results to the MYNN PBL scheme but also significantly overestimating TKE in the simulations where it also overestimates PBL height. For all schemes tested, as we increase model resolution from mesoscale to gray zone scale, we observed that the resolved TKE is larger as the resolution increases due to the model being able to resolve smaller eddies, as expected. Consequently, the total TKE (subgrid plus resolved) tends to be larger at higher resolution for the more traditional mesoscale PBL scheme MYNN. To mitigate this double-counting effect, the scale-aware SMS-3DTKE scheme attempts to compensate the production of resolved TKE by reducing the subgrid TKE using a partition function that depends on the grid resolution. However, its internal partition function appears to be too severe, as it reduces subgrid TKE faster than resolved TKE increases at higher resolution. In addition, the resolved TKE accounts for less than 20% of the total TKE along at least

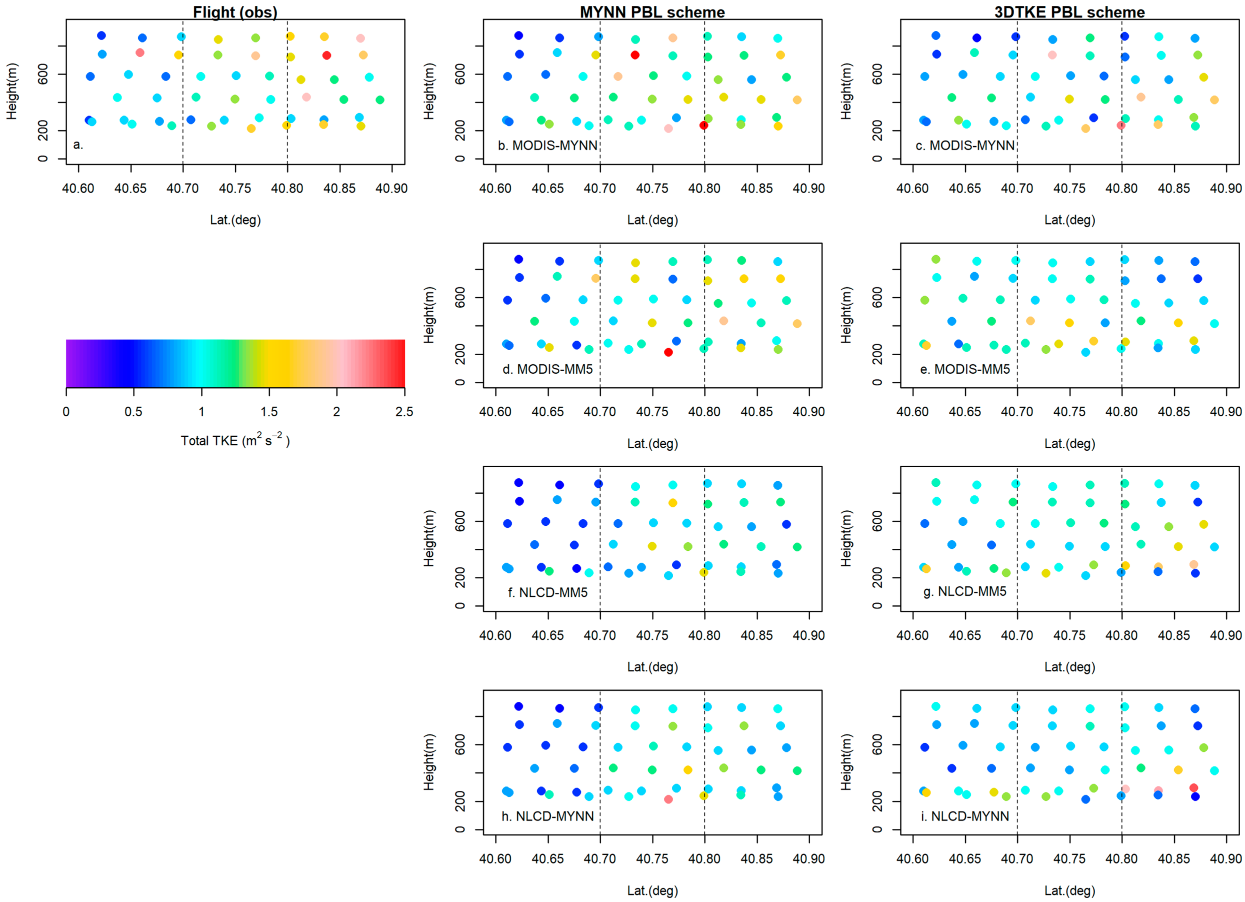

FIG. 12. As in Fig. 10, but for RF4, with adjusted color scale.

80% of each flight path in our 333-m grid resolution and much less at 1-km resolution for the MYNN and SMS-3DTKE boundary layer schemes, demonstrating that the gray zone issue can play a significant role at 333-m resolution and should not be overlooked.

Overall, the choice of the surface layer scheme does not show a very noticeable impact on the total TKE produced by the PBL schemes. The partition ratios tend to be marginally larger at the finest resolution when using the MM5 surface layer scheme when compared with the MYNN surface layer scheme. More noticeably, the choice of land-use classification impacts the results by producing less subgrid TKE with NLCD than with the MODIS land use, in particular, for coarser domains. This is readily explainable by the fact that the more resolved, heterogeneous, finer-resolution NLCD dataset also includes more lower-roughness suburban and rural areas upwind. However, while the NLCD land use decreases the total TKE, thus worsening overall agreement with the observations, we argue that the improved spatial distribution in NLCD also improves spatial features observed in the TKE along the flight path, indicating that better resolved land-use classifications might be a step forward in improving TKE predictions in complex surface scenarios.

In contrast, the EEPS PBL scheme exhibits a significant underestimation of total TKE and vertical exchange coefficient. The parameterized subgrid TKE along the flight paths falls to near zero in model domains with grid resolution well within gray zone scales, partially due to the vertical distribution of TKE. The scheme also produces much greater resolved TKE, showing relatively high ratios of about 80% on average. These results may seem in contrast to the originally published EEPS scheme results (Zhang et al. 2020). However, we note that the scheme was originally tested at 3-km spatial resolution, outside of the gray zone, and that our results support better performance of this scheme at coarser resolutions. The fact that the EEPS scheme does not perform as expected at the gray zone resolutions acts as a cautionary tale about using models outside the range for which they were developed and were known to perform well. As such, the burden of awareness of the model performance characteristics lies with the user.

Although WRF generally produces a good match to TKE from observations over most of the flight path (e.g., over rural and suburban areas, with particularly good matching over water), it produces a low bias in TKE immediately downwind (~1–2 km) of the very tall buildings of Manhattan, mostly in the model layers closest to the rooftops and when the wind speed is strong enough to cause significant building-induced turbulence. This result implies that building-resolving LES simulations (Muñoz-Esparza et al. 2020, 2021; Wiersema et al. 2022) may be needed to better capture building effects. However, urban canopy models, in combination with high-resolution land-use and urban morphology data (Brousse et al. 2016; Nadimpalli et al. 2022), might also be an alternative to a more explicit building-induced turbulence representation. A valuable follow-up to this paper would be to examine how the differences between PBL schemes shown here translate to pollutant dispersion and mixing.

Acknowledgments. We thank NIST for support of this work, through NIST Awards 70NANB19H167 and 70NANB21H021. We also thank the developers of the EEPS scheme, Chunxi Zhang (NOAA/NWS/NCEP/EMC) and Yuqing Wang (Department of Atmospheric Sciences and International Pacific Research Center, University of Hawaiʻi at Mānoa), for their communication and code. Certain commercial equipment, instruments, or materials are identified in this paper to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

Data availability statement. The data that support the findings of this study are available from corresponding author A. Hope upon reasonable request.

REFERENCES

Angevine, W. M., J. Olson, J. J. Gristey, I. Glenn, G. Feingold, and D. D. Turner, 2020: Scale awareness, resolved circulations, and practical limits in the MYNN–EDMF boundary layer and shallow cumulus scheme. *Mon. Wea. Rev.*, **148**, 4629–4639, https://doi.org/10.1175/MWR-D-20-0066.1.

Banzhaf, S., L. Ma, and C. Timmins, 2019: Environmental justice: The economics of race, place, and pollution. *J. Econ. Perspect.*, **33**, 185–208, https://doi.org/10.1257/jep.33.1.185.

Brousse, O., A. Martilli, M. Foley, G. Mills, and B. Bechtel, 2016: WUDAPT, an efficient land use producing data tool for mesoscale models? Integration of urban LCZ in WRF over Madrid. *Urban Climate*, **17**, 116–134, https://doi.org/10.1016/j.uclim.2016.04.001.

Demetillo, M. A. G., C. Harkins, B. C. McDonald, P. S. Chodrow, K. Sun, and S. E. Pusede, 2021: Space-based observational constraints on NO_2 air pollution inequality from diesel traffic in major US cities. *Geophys. Res. Lett.*, **48**, e2021GL094333, https://doi.org/10.1029/2021GL094333.

de Roode, S. R., and Coauthors, 2019: Turbulent transport in the gray zone: A large eddy model intercomparison study of the CONSTRAIN cold air outbreak case. *J. Adv. Model. Earth Syst.*, **11**, 597–623, https://doi.org/10.1029/2018MS001443.

Fairall, C. W., E. F. Bradley, J. E. Hare, A. A. Grachev, and J. B. Edson, 2003: Bulk parameterization of air–sea fluxes: Updates and verification for the COARE algorithm. *J. Climate*, **16**, 571–591, https://doi.org/10.1175/1520-0442(2003)016<0571:BPOASF>2.0.CO;2.

Friedl, M. A., and Coauthors, 2002: Global land cover mapping from MODIS: Algorithms and early results. *Remote Sens. Environ.*, **83**, 287–302, https://doi.org/10.1016/S0034-4257(02)00078-0.

——, D. Sulla-Menashe, B. Tan, A. Schneider, N. Ramankutty, A. Sibley, and X. Huang, 2010: MODIS Collection 5 Global Land Cover: Algorithm refinements and characterization of new datasets. *Remote Sens. Environ.*, **114**, 168–182, https://doi.org/10.1016/j.rse.2009.08.016.

Gardner-Frolick, R., D. Boyd, and A. Giang, 2022: Selecting data analytic and modeling methods to support air pollution and environmental justice investigations: A critical review and guidance framework. *Environ. Sci. Technol.*, **56**, 2843–2860, https://doi.org/10.1021/acs.est.1c01739.

Garman, K. E., and Coauthors, 2006: An airborne and wind tunnel evaluation of a wind turbulence measurement system for aircraft-based flux measurements. *J. Atmos. Oceanic Technol.*, **23**, 1696–1708, https://doi.org/10.1175/JTECH1940.1.

——, P. Wyss, M. Carlsen, J. R. Zimmerman, B. H. Stirm, T. Q. Carney, R. Santini, and P. B. Shepson, 2008: The contribution of variability of lift-induced upwash to the uncertainty in vertical winds determined from an aircraft platform. *Bound.-Layer Meteor.*, **126**, 461–476, https://doi.org/10.1007/s10546-007-9237-y.

Giometto, M. G., A. Christen, C. Meneveau, J. Fang, M. Krafczyk, and M. B. Parlange, 2016: Spatial characteristics of roughness sublayer mean flow and turbulence over a realistic urban surface. *Bound.-Layer Meteor.*, **160**, 425–452, https://doi.org/10.1007/s10546-016-0157-6.

Hacker, J., and T. Crawford, 1999: The BAT-probe: The ultimate tool to measure turbulence from any kind of aircraft (or sailplane). *Tech. Soaring*, **23**, 43–46.

Hajny, K. D., and Coauthors, 2022: A spatially explicit inventory scaling approach to estimate urban CO_2 emissions. *Elem. Sci. Anthropocene*, **10**, 00121, https://doi.org/10.1525/elementa.2021.00121.

Honnert, R., V. Masson, and F. Couvreux, 2011: A diagnostic for evaluating the representation of turbulence in atmospheric models at the kilometric scale. *J. Atmos. Sci.*, **68**, 3112–3131, https://doi.org/10.1175/JAS-D-11-061.1.

——, and Coauthors, 2020: The atmospheric boundary layer and the "gray zone" of turbulence: A critical review. *J. Geophys. Res. Atmos.*, **125**, e2019JD030317, https://doi.org/10.1029/2019JD030317.

——, V. Masson, C. Lac, and T. Nagel, 2021: A theoretical analysis of mixing length for atmospheric models from micro to large scales. *Front. Earth Sci.*, **8**, 582056, https://doi.org/10.3389/feart.2020.582056.

Ito, J., H. Niino, M. Nakanishi, and C.-H. Moeng, 2015: An extension of the Mellor–Yamada model to the terra incognita zone for dry convective mixed layers in the free convection regime. *Bound.-Layer Meteor.*, **157**, 23–43, https://doi.org/10.1007/s10546-015-0045-5.

Jiménez, P. A., J. Dudhia, J. F. González-Rouco, J. Navarro, J. P. Montávez, and E. García-Bustamante, 2012: A revised scheme for the WRF surface layer formulation. *Mon. Wea. Rev.*, **140**, 898–918, https://doi.org/10.1175/MWR-D-11-00056.1.

Jin, S., and Coauthors, 2019: Overall methodology design for the United States National Land Cover Database 2016 products. *Remote Sens.*, **11**, 2971, https://doi.org/10.3390/rs11242971.

Juliano, T. W., B. Kosović, P. A. Jiménez, M. Eghdami, S. E. Haupt, and A. Martilli, 2022: "Gray zone" simulations using a three-dimensional planetary boundary layer parameterization in the Weather Research and Forecasting Model. *Mon. Wea. Rev.*, **150**, 1585–1619, https://doi.org/10.1175/MWR-D-21-0164.1.

Kim, E., and Coauthors, 2023: Turbulent characteristics in complex coastal areas assessed using BSWO observations and

WRF-LES simulation results. *Atmos. Res.*, **289**, 106756, https://doi.org/10.1016/j.atmosres.2023.106756.

Lauvaux, T., and Coauthors, 2020: Policy-relevant assessment of urban CO_2 emissions. *Environ. Sci. Technol.*, **54**, 10 237–10 245, https://doi.org/10.1021/acs.est.0c00343.

Lee, S.-H., and Coauthors, 2011: Evaluation of urban surface parameterizations in the WRF model using measurements during the Texas Air Quality Study 2006 field campaign. *Atmos. Chem. Phys.*, **11**, 2127–2143, https://doi.org/10.5194/acp-11-2127-2011.

Li, H., G. Cui, and Z. Zhang, 2017: A new scheme for the simulation of microscale flow and dispersion in urban areas by coupling large-eddy simulation with mesoscale models. *Bound.-Layer Meteor.*, **167**, 145–170, https://doi.org/10.1007/s10546-017-0323-5.

Lopez-Coto, I., and Coauthors, 2020a: Wintertime CO_2, CH_4, and CO emissions estimation for the Washington, DC–Baltimore metropolitan area using an inverse modeling technique. *Environ. Sci. Technol.*, **54**, 2606–2614, https://doi.org/10.1021/acs.est.9b06619.

——, M. Hicks, A. Karion, R. K. Sakai, B. Demoz, K. Prasad, and J. Whetstone, 2020b: Assessment of planetary boundary layer parameterizations and urban heat island comparison: Impacts and implications for tracer transport. *J. Appl. Meteor. Climatol.*, **59**, 1637–1653, https://doi.org/10.1175/JAMC-D-19-0168.1.

——, P. B. Shepson, A. Karion, C. Gerbig, K. Prasad, and J. R. Whetstone, 2020c: Coupling an off-line Lagrangian dispersion model with large eddy simulations as a tool for vertical mixing parameterization development in mesoscale applications. *21st Joint Conf. on the Applications of Air Pollution Meteorology*, Boston, MA, Amer. Meteor. Soc., 10.3, https://ams.confex.com/ams/2020Annual/meetingapp.cgi/Paper/368360.

Mellor, G. L., and T. Yamada, 1982: Development of a turbulence closure model for geophysical fluid problems. *Rev. Geophys.*, **20**, 851–875, https://doi.org/10.1029/RG020i004p00851.

Muñoz-Esparza, D., and Coauthors, 2020: Inclusion of building-resolving capabilities into the FastEddy GPU-LES model using an immersed body force method. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002141, https://doi.org/10.1029/2020MS002141.

——, and Coauthors, 2021: Efficient graphics processing unit modeling of street-scale weather effects in support of aerial operations in the urban environment. *AGU Adv.*, **2**, e2021AV000432, https://doi.org/10.1029/2021AV000432.

Nadimpalli, R., P. Patel, U. C. Mohanty, S. D. Attri, and D. Niyogi, 2022: Impact of urban parameterization and integration of WUDAPT on the severe convection. *Comput. Urban Soc.*, **2**, 41, https://doi.org/10.1007/s43762-022-00071-w.

Nakanish, M., 2001: Improvement of the Mellor–Yamada turbulence closure model based on large-eddy simulation data. *Bound.-Layer Meteor.*, **99**, 349–378, https://doi.org/10.1023/A:1018915827400.

Nakanishi, M., and H. Niino, 2004: An improved Mellor–Yamada level-3 model with condensation physics: Its design and verification. *Bound.-Layer Meteor.*, **112** (1), 1–31, https://doi.org/10.1023/B:BOUN.0000020164.04146.98.

——, and ——, 2006: An improved Mellor–Yamada level-3 model: Its numerical stability and application to a regional prediction of advection fog. *Bound.-Layer Meteor.*, **119**, 397–407, https://doi.org/10.1007/s10546-005-9030-8.

NCEP/NWS/NOAA/U.S. Department of Commerce, 2015: NCEP North American Mesoscale (NAM) 12 km analysis (updated daily). Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, accessed 2 March 2021, https://doi.org/10.5065/G4RC-1N91.

Olson, J. B., and J. M. Brown, 2009: A comparison of two Mellor–Yamada-based PBL schemes in simulating a hybrid barrier jet. *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., JP1.13, https://ams.confex.com/ams/pdfpapers/154321.pdf.

——, J. S. Kenyon, W. A. Angevine, J. M. Brown, M. Pagowski, and K. Sušelj, 2019: A description of the MYNN-EDMF scheme and the coupling to other components in WRF–ARW. NOAA Tech. Memo. OAR GSD 61, 42 pp., https://doi.org/10.25923/n9wm-be49.

Pitt, J. R., and Coauthors, 2022: New York City greenhouse gas emissions estimated with inverse modeling of aircraft measurements. *Elem. Sci. Anthropocene*, **10**, 00082, https://doi.org/10.1525/elementa.2021.00082.

Salmon, O. E., and Coauthors, 2017: Urban emissions of water vapor in winter. *J. Geophys. Res. Atmos.*, **122**, 9467–9484, https://doi.org/10.1002/2016JD026074.

Seidel, D. J., C. O. Ao, and K. Li, 2010: Estimating climatological planetary boundary layer heights from radiosonde observations: Comparison of methods and uncertainty analysis. *J. Geophys. Res.*, **115**, D16113, https://doi.org/10.1029/2009JD013680.

Senel, C. B., O. Temel, D. Muñoz-Esparza, A. Parente, and J. van Beeck, 2020: Gray zone partitioning functions and parameterization of turbulence fluxes in the convective atmospheric boundary layer. *J. Geophys. Res. Atmos.*, **125**, e2020JD033581, https://doi.org/10.1029/2020JD033581.

Skamarock, W. C., and Coauthors, 2019: A description of the Advanced Research WRF Model version 4. NCAR Tech. Note NCAR/TN-556+STR, 145 pp., https://doi.org/10.5065/1dfh-6p97.

Teixeira, J. C., J. Fallmann, A. C. Carvalho, and A. Rocha, 2019: Surface to boundary layer coupling in the urban area of Lisbon comparing different urban canopy models in WRF. *Urban Climate*, **28**, 100454, https://doi.org/10.1016/j.uclim.2019.100454.

Wiersema, D. J., K. A. Lundquist, J. D. Mirocha, and F. Katopodes Chow, 2022: Evaluation of turbulence and dispersion in multiscale atmospheric simulations over complex urban terrain during the joint urban 2003 field campaign. *Mon. Wea. Rev.*, **150**, 3195–3209, https://doi.org/10.1175/MWR-D-22-0056.1.

Wyngaard, J. C., 2004: Changing the face of small-scale meteorology. *Atmospheric Turbulence and Mesoscale Meteorology*, Cambridge University Press, 17–34.

Yadav, V., and Coauthors, 2021: The impact of COVID-19 on CO_2 emissions in the Los Angeles and Washington DC/Baltimore metropolitan areas. *Geophys. Res. Lett.*, **48**, e2021GL092744, https://doi.org/10.1029/2021GL092744.

Yang, L., and Coauthors, 2018: A new generation of the United States National Land Cover Database: Requirements, research priorities, design, and implementation strategies. *ISPRS J. Photogramm. Remote Sens.*, **146**, 108–123, https://doi.org/10.1016/j.isprsjprs.2018.09.006.

Yoshida, T., T. Takemi, and M. Horiguchi, 2018: Large-eddy-simulation study of the effects of building-height variability on turbulent flows over an actual urban area. *Bound.-Layer Meteor.*, **168**, 127–153, https://doi.org/10.1007/s10546-018-0344-8.

Zhang, C., Y. Wang, and M. Xue, 2020: Evaluation of an E–ε and three other boundary layer parameterization schemes in the WRF Model over the southeast Pacific and the southern Great Plains. *Mon. Wea. Rev.*, **148**, 1121–1145, https://doi.org/10.1175/MWR-D-19-0084.1.

Zhang, X., J. W. Bao, B. Chen, and E. D. Grell, 2018: A three-dimensional scale-adaptive turbulent kinetic energy scheme in the WRF-ARW model. *Mon. Wea. Rev.*, **146**, 2023–2045, https://doi.org/10.1175/MWR-D-17-0356.1.