Impacts of different plant functional types on ambient ozone predictions in the Seoul Metropolitan Areas (SMAs), Korea

H.-K. Kim[1]**, J.-H. Woo**[1]**, R. S. Park**[2]**, C. H. Song**[2]**, J.-H. Kim**[3]**, S.-J. Ban**[4]**, and J.-H. Park**[4]

[1]Konkuk University, Department of Advanced Technology Fusion, Seoul, Korea
[2]Gwangju Institute of Science and Technology, School of Environmental Science and Engineering, Gwangju, Korea
[3]Green-ECOS, Seoul, Korea
[4]National Institute of Environmental Research, Air Quality Research Division, Incheon, Korea

Correspondence to: J.-H. Woo (woojh21@gmail.com)

Abstract. Plant functional type (PFT) distributions affect the results of biogenic emission modeling as well as O_3 and particulate matter (PM) simulations using chemistry-transport models (CTMs). This paper analyzes the variations of both surface biogenic volatile organic compound (BVOC) emissions and O_3 concentrations due to changes in the PFT distributions in the Seoul Metropolitan Areas, Korea. The Fifth-Generation NCAR/Pennsylvania State Mesoscale Model (MM5)/the Model of Emissions of Gases and Aerosols from Nature (MEGAN)/the Sparse Matrix Operator Kernel Emissions (SMOKE)/the Community Multiscale Air Quality (CMAQ) model simulations were implemented over the Seoul Metropolitan Areas in Korea to predict surface O_3 concentrations for the period of 1 May to 31 June 2008. Starting from a performance check of CTM predictions, we consecutively assessed the effects of PFT area deviations on the MEGAN BVOC and CTM O_3 predictions, and we further considered the basis of geospatial and statistical analyses. The three PFT data sets considered were (1) the Korean PFT, developed with Korea-specific vegetation database; (2) the CDP PFT, adopted from the community data portal (CDP) of US National Center for Atmospheric Research in the United States (NCAR); (3) MODIS PFT, reclassified from the NASA Terra and Aqua combined land cover products. Although the CMAQ performance check reveals that all of the three different PFT data sets are applicable choices for regulatory modeling practice, noticeable primary data (i.e., PFT and Leaf Area Index (LAI)) was observed to be missing in many geographic locations. Based on the assessed effect of such missing data on CMAQ O_3 predictions, we found that this missing data can cause spatially increased bias in CMAQ O_3. Thus, it must be resolved in the near future to obtain more accurate biogenic emission and chemistry transport modeling results.

Comparisons of MEGAN biogenic emission results with the three different PFT data showed that broadleaf trees (BTs) are the most significant contributor, followed by needleleaf trees (NTs), shrub (SB), and herbaceous plants (HBs) to the total BVOCs. In addition, isoprene from BTs and terpene from NTs were recognized as significant primary and secondary BVOC species in terms of BVOC emissions distributions and O_3-forming potentials in the study domain. A geographically weighted regression analysis with locally compensated ridge (LCR-GWR) with the different PFT data (δO_3 vs. δPFTs) suggests that addition of BT, SB, and NT areas can contribute to O_3 increase, whereas addition of an HB area contributes to O_3 decrease in the domain.

Assessment results of the simulated spatial and temporal changes of O_3 distributions with the different PFT scenarios reveal that hourly and local impacts from the different PFT distributions on occasional inter-deviations of O_3 are quite noticeable, reaching up to 13 ppb. The simulated maximum 1 h O_3 inter-deviations between different PFT scenarios have an asymmetric diurnal distribution pattern (low in the early morning, rising during the day, peaking at 05:00 p.m., and decreasing during the night) in the study domain. Exponentially diverging hourly BVOC emissions and O_3 concentrations with increasing ambient temperature suggest that the use of different PFT distribution data requires much caution when modeling (or forecasting) O_3 air quality in complicated

urban atmospheric conditions in terms of whether uncertainties in O_3 prediction results are expected to be mild or severe.

1 Introduction

Biogenic volatile organic compounds (BVOCs) emitted from vegetated areas play an important role in the chemistry of the lower troposphere and atmospheric boundary layer via a series of oxidation reactions with OH and NO_3 radicals and O_3 (Finlayson-Pitts and Fitts, 2000; Atkinson and Arey, 2003). It is known that BVOC emissions can enhance O_3 formation in the areas with high NO_x concentrations because BVOC oxidation increases the concentrations of hydroperoxy and organic peroxy radicals (HO_2 and RO_2) that can convert NO into NO_2 without depleting O_3. In addition, the BVOC emissions can reduce O_3 concentrations in the areas with low levels of NO_x because the reaction of O_3 and BVOC reduces hydroxyl radicals and leads to decreased O_3 formation (Finlayson-Pitts and Fitts, 2000; Hogrefe et al., 2011). For example, a regional air quality modeling study reported that biogenic emissions are associated with at least 20 % of surface O_3 concentrations in most areas of the continental United States (Tao et al., 2003). A modeling study with MOZART-4 (Model for Ozone and Related Chemical Species, version 4) showed that biogenic isoprene emissions cause -5 ppb to 10 ppb changes in surface O_3 concentrations over the Amazon region, Indonesia, and parts of South Africa during the spring season (Pfister et al., 2008).

The flux of biogenic emission is a strong function of vegetation area, biomass density, and other environmental factors (Guenther et al., 1995, 2006). In biogenic emission modeling, vegetation area has been commonly considered to be one of the most important driving variables because it reflects the phenological emission capacity of the area of interest. In order to more efficiently prepare vegetation area information as an input to biogenic emission modeling, plant functional type (PFT) data have been used (Guenther et al., 2006; Pfister et al., 2008; Arneth et al., 2011). Conceptually, PFTs are classes of vegetation species that share similar responses to environmental factors, similar functioning at the organismic level, and/or similar effects on ecosystems (Smith et al., 1997; Sun et al., 2008). In a biogenic emissions modeling approach, the PFT data sets specify the type and composition of vegetation classes in a grid cell to determine the capacity of biogenic emissions. Thus, careful consideration of PFT distributions is needed for the estimation of biogenic emissions and subsequent predictive work such as the O_3 prediction with chemistry-transport model (CTM).

Several recent modeling studies have reported sensitivities of PFT to the biogenic emissions and O_3 concentrations. For example, Guenther et al. (2006) reported about -13 to 24 % changes in global annual isoprene emissions from standard Model of Emissions of Gases and Aerosols from Nature (MEGAN) modeling using the 11 different PFT data sets.

Pfister et al. (2008) reported a factor of two or more differences in monthly isoprene emissions on global to regional scales through the examination of the MEGAN sensitivity to three different sets of satellite-derived leaf area index (LAI) and PFT data sets. Pfister et al. (2008) also reported a difference of up to 5 ppb in monthly mean O_3 concentrations from the global-scale simulations of MOZART /MEGAN with three different sets of satellite-derived LAI and PFT input data.

Although the previously summarized modeling studies produced valuable findings, none of the studies represents a local scale and a more detailed situation. For example, Pfister et al. (2008) adopted a coarse-grid resolution (i.e., $2.8° \times 2.8°$) in their study and use only satellite-based PFT distributions to carry out their global modeling study. In addition, they also showed that the sensitivity or uncertainty (e.g., a factor of 2 or more differences in isoprene and up to 5 ppb of surface O_3 inter-deviation) is not only related to changes in LAI data but also to changes in PFT data.

The purpose of our study is to investigate the impacts of different PFT distribution data on O_3 concentrations. Carrying out biogenic emission model and CTM simulations, we investigated changes in surface biogenic emissions and ambient O_3 concentrations with different PFT distribution data sets. Our approach has several distinctive features. First, in this study, we included a region-specific high-resolution PFT distribution data set: the Korean PFT database (KORPFT). The KORPFT was derived from three data sets: (1) Korean land cover classification maps; (2) tree stock maps; and (3) Korean vegetation survey data. Secondly, we adopted a 3 km by 3 km fine grid system in order to more closely investigate spatial patterns of biogenic emissions and O_3 behavior in accordance with finer PFT distribution patterns. Thirdly, we changed only PFT data sets to isolate the impacts of the different PFTs on atmospheric chemistry (or O_3 concentrations).

The main idea behind our approach is that because PFT distributions affect the magnitudes of biogenic emission capacity in the domain of interest, the use of different PFT data can affect biogenic emission estimations and consequently O_3 prediction results. The proposed idea and method was tested over the Seoul Metropolitan Areas (SMAs) in Korea, in which both tremendously developed urban area and densely vegetated areas actually coexist.

Section 2 of this paper describes the PFT distribution data sets developed for or used in this study together with the simulation framework used for biogenic emission estimation and O_3 predictions and some statistical measures used for the evaluation of the modeling results. Section 3 presents and discusses evaluation and comparison results for biogenic emissions and O_3. Finally, Sect. 4 presents some conclusions of this study.

2 PFT data sets and simulation experiment

2.1 Development of the PFT data set

2.1.1 Korean PFT database

A highly resolved PFT distribution data set (resolution of $150 \times 150\,\mathrm{m}^2$) for the Korean Peninsula (KORPFT) was established in this study by compiling various local sources of vegetation data, such as a Korean land cover map (produced in 2007 and renewed in 2009) and vegetation survey (collected for the consecutive years 1994–2004) from the Korea Ministry of Environment, and tree stock maps (produced and renewed for the consecutive years 1996–2005) from the Korea Forest Service.

A four-PFT scheme in which multiple vegetation species were classified into four types, broadleaf (BT), needleleaf (NT), shrub (SB), and herbaceous (HB), was applied. The fraction of individual PFT groups (PFTF_k) was calculated by applying the canopy density-weighted focal average as follows:

$$\mathrm{PFTF}_k(\%) = \sum_{i=1}^{n} \mathrm{fm}_{\mathrm{PFT}_{k\_i}} / \mathrm{SA} \times \delta_{\mathrm{PFT}_{k\_i}} \times 100, \tag{1}$$

where SA is a selected grid region (m^2) with n-grid cells; PFTF_k denotes the percentage area of an individual PFT in the SA (%); k denotes the individual PFT type (e.g., BT, NT, SB, and HB); n denotes the number of unit grid cells in the SA; $\mathrm{fm}_{\mathrm{PFT}_{k\_i}}$ is the spatial moving averaged value (m^2) of the ith unit grid cell for a PFT in the SA calculated by the focal average method; and $\delta_{\mathrm{PFT}_{k\_i}}$ is the vegetation canopy density factor for an individual PFT at ith grid cell ($0 \leq \delta_{\mathrm{PFT}_{k\_i}} \leq 1$).

The focal average of individual PFTs at every focal grid cell (the ith grid cell) can be defined as follows:

$$\mathrm{fm}_{\mathrm{PFT}_{k\_i}} = \frac{1}{N_i} \sum_{\{j : j \sim i\}}^{a_{j_{\mathrm{wrt}}\mathrm{PFT}_k}}, \tag{2}$$

where N_i denotes the number of all neighbors of the ith unit grid cell, the set $\{j : j \sim i\}$ contains all the neighboring locations j of unit i, and $a_{j_{\mathrm{wrt}}\mathrm{PFT}_k}$ is the area sum (m^2) of all neighboring grids of a focal grid cell with respect to an individual PFT group.

2.1.2 CDP and MODIS PFT

Two other PFT distribution data sets used in this study are based on satellite observations. The former PFT distribution data set (resolution of $1 \times 1\,\mathrm{km}^2$) was developed for the global isoprene emission study (Guenther et al., 2006) and derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) land cover product in 2001. As this PFT data set is downloadable from the community data portal (http://cdp.ucar.edu) maintained by the US National Center for Atmospheric Research (NCAR), we refer to this data

set as CDP for brevity in this study. The latter PFT distribution data set (resolution of $0.5 \times 0.5\,\mathrm{km}^2$) was reclassified for this study from the MODIS land cover type 5 products of the Terra and Aqua satellite sensors in 2008. Hereafter, MODIS will only refer to this reclassified 2008 data set. The MODIS land cover type 5 data, with a PFT scheme including eight vegetation and four non-vegetation classes (Bonan et al., 2002; Friedl et al., 2002; Strahler et al., 1999), were processed to generate the MEGAN PFT. The MODIS vegetation classes were converted into MEGAN PFT classes by straightforward mapping of the eight MODIS vegetation classes into the four MEGAN PFTs (i.e., BT, NT, SB, and HB). For example, the PFT distribution was calculated by adopting land cover class descriptions provided by the International-Geosphere–Biosphere Programme (IGBP) (Friedl et al., 2002; Strahler et al., 1999). As an example, geographically distributed pixels (pixel size $= 500\,\mathrm{m} \times 500\,\mathrm{m}$) for MODIS broadleaf evergreen and broadleaf deciduous trees were reclassified as BT. Based on the IGBP description of the vegetation-covered fraction for the broadleaf evergreen and broadleaf deciduous trees (Strahler et al., 1999) (i.e., "Lands dominated by woody vegetation with a percentage cover $> 60\,\%$"), we assumed that the BT-covered fractions for every reclassified pixel range from 60 to $100\,\%$ (mean BT-covered fraction $= 80\,\%$). Therefore, we simply assigned a mean % BT-covered fraction of $80\,\%$ to every BT pixel when we developed a gridded $\mathrm{PFTF}_{\mathrm{BT}}$ database ($500\,\mathrm{m} \times 500\,\mathrm{m}$ spatial resolution) for MEGAN.

2.2 Framework of the simulation experiment

A modeling framework, the Fifth-Generation NCAR/Pennsylvania State Mesoscale Model (MM5)–the Sparse Matrix Operator Kernel Emissions (SMOKE)–the Model of Emissions of Gases and Aerosols from Nature (MEGAN)–the Community Multiscale Air Quality (CMAQ), was established for meteorological modeling, anthropogenic and biogenic emissions processing and modeling, and chemistry-transport model simulations for May–June 2008. In this modeling study, three-step nested gridding was conducted. Horizontal resolutions of $27 \times 27\,\mathrm{km}^2$ for the largest domain, $9 \times 9\,\mathrm{km}^2$ for the second largest domain, and $3 \times 3\,\mathrm{km}^2$ for the smallest domain were used (refer to Fig. 1a). In the framework, the model settings for three-step simulations were the same, except for the biogenic emissions input. The three sets of hourly, gridded, and speciated biogenic emissions (based on KORPFT, CDP, and MODIS data sets) were merged with anthropogenic emissions (described in Sect. 2.2.2).

Because we are focusing on the impact of using different PFT data sets, the specific simulation scenarios are hereafter referred to as KORPFT, CDP, and MODIS (Table 1).
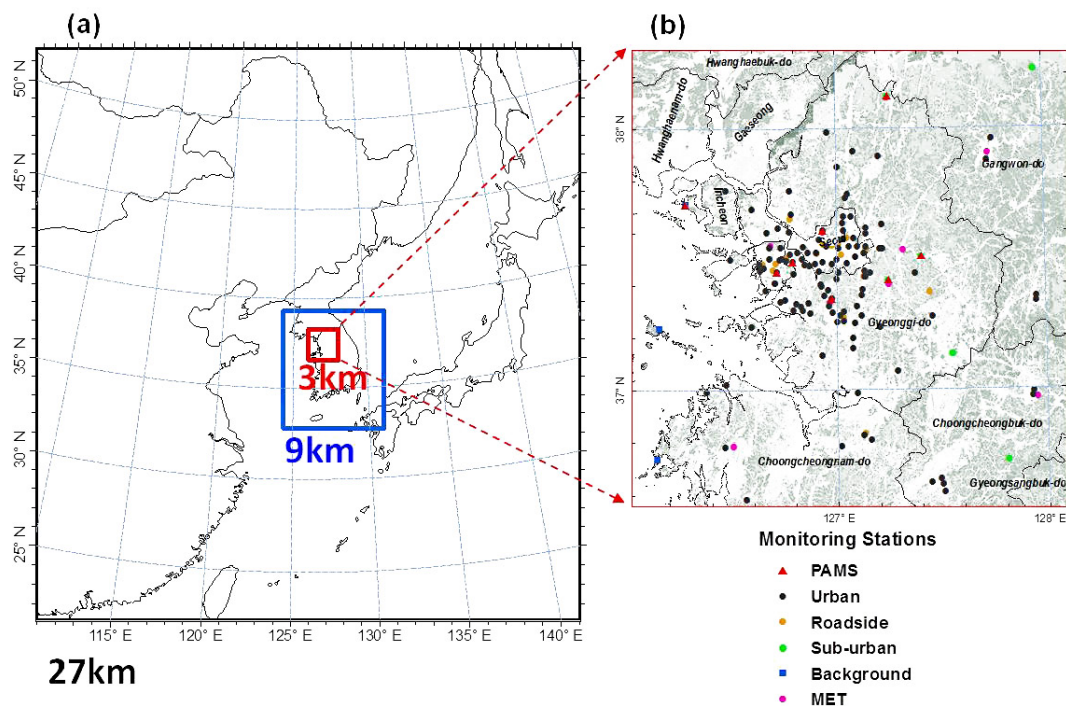
Figure 1. The modeling domain. The target domain is a $3\,km \times 3\,km$ grid system domain covering the Seoul, Gyeonggi, and Incheon Metropolitan Areas. Note that "27 km", "9 km", and "3 km" refer to the model grid resolutions.

Table 1. List of model simulation scenarios and the PFT data sets used for biogenic emissions modeling.

Simulation name	D1 (27 km)	D2 (9 km)	D3 (3 km)
KORPFT	MODIS	KORPFT	KORPFT
CDP	MODIS	CDP	CDP
MODIS	MODIS	MODIS	MODIS

2.2.1 Meteorological and chemical transport modeling

For meteorological inputs to the CMAQ model, MM5 modeling was carried out for a period of May–June 2008. Each of the three domains (i.e., 27 km, 9 km, and 3 km) consisted of 20 vertical layers resolving the atmosphere between the surface and 100 hPa in sigma coordinate. Applying a two-way nesting technique, the meteorological outputs from coarse-grid to fine-grid domains (i.e., 27 to 9 to 3 km) were derived. NCEP/DOE AMIP-II (National Centers for Environmental Prediction/Department of Energy Atmospheric Model Intercomparison Project II) Reanalysis data (Reanalysis-2) (Kanamitsu et al., 2002) were used for the initial and boundary conditions (IC/BC). The Grell scheme (Grell et al., 1994), based on the rate of destabilization or quasi-equilibrium, was employed for cumulus parameterization. The Medium-Range Forecast (MRF) Planetary Boundary Layer (PBL) scheme was applied to obtain high resolution in the PBL (Hong and Pan, 1996). For an explicit

moisture scheme, a mixed-phase option was used (Reisner et al., 1998). In order to reduce meteorological uncertainties, four-dimensional data assimilation (FDDA) was employed with 10 m Advanced Scatterometer (ASCAT) wind data. The MM5 outputs were then processed with the Meteorology–Chemistry Interface Processor utility (Byun and Ching, 1999) to derive the meteorological input variables for CMAQ simulations.

For the O_3 air quality simulation using CMAQv4.6 (Byun and Ching, 1999; Byun and Schere, 2006), we used the State Air Pollution Research Center mechanism, version 1999 (SAPRC-99) chemical mechanism for gas-phase chemistry (Carter, 2000a), AERO3 module for aerosol formation (Binkowski and Roselle, 2003), the piecewise parabolic method (PPM) for advection (Collela and Woodward, 1984), the multi-scale method for horizontal diffusion (CMAQ v4.6 Operational Guidance Document, 2006), the Asymmetric Convective Method (ACM) for cloud (Pleim and Chang, 1992), and an updated version of ACM (ACM2) for vertical diffusion (CMAQ v4.6 Operational Guidance Document, 2006). Based on these emissions and meteorological inputs, one-way nested model simulations were performed at the three domains (Fig. 1a) with a four-day spin-up period. Here, we focus on the fine and detailed domain ($3 \times 3\,km^2$) in which developed urban areas and densely vegetated areas coexist. The BCs for the 3 km CMAQ modeling domain were obtained from the simulation outputs of the coarse domains (27 km down to 9 km) using the CMAQ boundary condition

(BCON) processor to generate hourly concentrations along the outer lateral edges of the 3 km domain. The ICs for the 3 km domain were also obtained from the CTM results at coarse domains.

2.2.2 Anthropogenic emissions modeling

For processing anthropogenic emission, SMOKE-Asia (Woo et al., 2012) was applied to generate CMAQ-ready anthropogenic emissions for our study domain. SMOKE-Asia adopts the Sparse Matrix Operator Kernel Emissions (SMOKE) processing system of the US Environmental Protection Agency (EPA) as the base frame, but with some improved and upgraded contents. For example, SMOKE-Asia includes spatial and temporal surrogate databases, such as 38 source classification codes (SCCs), 2752 administrative division codes, and some regionalized temporal profiles. Details of SMOKE-Asia are described in Woo et al. (2012). A merged version of the INTEX 2006 (Zhang et al., 2009) and TRACE-P 2000 inventories (Streets et al., 2003) was used as a base inventory and was allocated into our study domain using spatial surrogates in SMOKE-Asia. Through further processing with temporal and chemical speciation (SAPRC-99 chemical mechanism) profiles, we generated hourly gridded CMAQ-ready emissions data for this study.

The anthropogenic volatile organic compound (VOC) and NO_x emissions processed by SMOKE-Asia for the study period are shown in Table 2, and their spatial distributions are displayed in Fig. 2.

2.2.3 Biogenic emissions modeling

We used MEGANv2.04 (Guenther et al., 2006) to produce BVOC emission inputs to CMAQ. The BVOC flux is a function of emission factor, vegetation area, and various environmental factors:

$$ER = EF \cdot A \cdot \gamma_F \cdot \gamma_S \cdot \gamma_W \cdot \phi, \qquad (3)$$

where ER is the net emission rate (μg of compound h^{-1}); EF is the an emission factor that represents the net in-canopy emission rate expected at standard conditions ($\mu g\,m^{-2}\,h^{-1}$ at 303 K); A is the vegetation cover area represented by the PFT (A = grid area × PFTF); γ includes environmental activity factors that account for the emission changes due to activity deviations from standard conditions, such as changes in leaf area and age (γ_F), stress due to soil moisture content (γ_S), and environmental effects (i.e., temperature and solar radiation) within the canopy (γ_W); and ϕ is a factor that defines chemical production and loss within plant canopies (Guenther et al., 2006).

In this study, detailed soil-moisture and plant-canopy information was not considered, so γ_S and ϕ were set to 1. For EF, we applied PFT-specific emission factors tabulated in MEGAN module. For γ_W, MM5-derived solar radiation and temperature data sets were used. For the leaf

area and age (γ_F) input calculation, we utilized monthly averaged the LAI of vegetation covered surfaces (LAIv) data using 2008 MODIS LAI (eight-day coverage and 1 km resolution) and each source of PFT distribution data. MEGAN uses an approach that divides the surface of each grid cell into different PFTs and non-vegetated surfaces (Guenther et al., 2006). MEGAN uses LAIv to simulate the seasonal variations in leaf biomass and age distribution rather than LAI values (Guenther and Sakulyanontvittaya, 2011; Guenther et al., 2006). MEGAN assumes that plant leaves cover only that part of the grid cell containing vegetation (Guenther et al., 2006). Thus, the LAIv calculations can only be performed with LAI at the grid cells with PFT values. LAIv is different from LAI in that LAIv is estimated by dividing the grid-average LAI by the vegetation-covered fraction. The upper limit of LAIv is set to 6 to eliminate the very high values that can be estimated for grids with very little vegetation (Guenther et al., 2006). Guenther and Sakulyanontvittaya (2011) suggested two main reasons for using LAIv in MEGAN. First, LAIv is the actual LAI of the canopy and is, thus, a more appropriate input for a canopy environment model. Second, lower and upper bounds can be placed on LAIv because vegetation-covered areas rarely have a maximum LAI of less than 0.1 or more than 10. In this study, we adopted three different sources of PFT data (i.e., KORPFT, CDP, and MODIS) and a single source of LAI data (i.e., MODIS LAI at 1 km × 1 km resolution). Raw MODIS LAI values range from 0 to 10. Following the definition of LAIv in MEGAN literature (e.g., Guenther and Sakulyanontvittaya, 2011; Guenther et al., 2006), the MODIS LAI values were divided by the PFT-covered fraction at each grid cell (3 km × 3 km size) and converted to LAIv values (e.g., 3 km × 3 km zonal averaged LAI values/3 km × 3 km zonal averaged PFT fractional values). The LAIv calculations were performed for three different vegetation data sets (i.e., MODIS LAI-KORPFT, MODIS LAI-CDP, and MODIS LAI-MODIS PFT) independently. It is important to note that the LAIv values were not computed unless a grid cell includes both PFT and LAI values concurrently.

The estimated biogenic VOC and NO emissions from the MEGAN model with the three different PFT scenarios for the study period are shown in Table 2. Overall, the biogenic VOC emissions (BVOC) contribute 44.5 % of the total VOC emissions (anthropogenic + biogenic). The differences of BVOC emissions between the three PFT scenarios ranged from 1.8 Gg (MODIS-KORPFT) to 4.2 Gg (MODIS-CDP) for the study period (May–June 2008). Biogenic NO_x (specifically NO) showed marginal contributions (~ 0.4 %) to the total NO_x emissions and differences (e.g., 0.02 Gg between CDP and KORPFT) between the three different PFT data sets. Investigating the distribution of the VOC / NO_x ratio across the domain revealed that consistently low values of the VOC / NO_x ratio were distributed across the domain for the case of anthropogenic emission only (i.e.,

Table 2. Anthropogenic and biogenic nonmathane VOC (NMVOC) and NO_x emissions and deviations between different PFT-input cases (unit: $Gg\, period^{-1}$).

	Anthropogenic emissions	Biogenic emissions						
		Mean	KORPFT	CDP	MODIS	δKORPFT	δCDP	δMODIS
NMVOC	37.24	29.83	29.98	27.68	31.83	0.15	−2.15	2.00
NO_x	43.37	0.18	0.17	0.19	0.18	−0.01	0.01	0.00
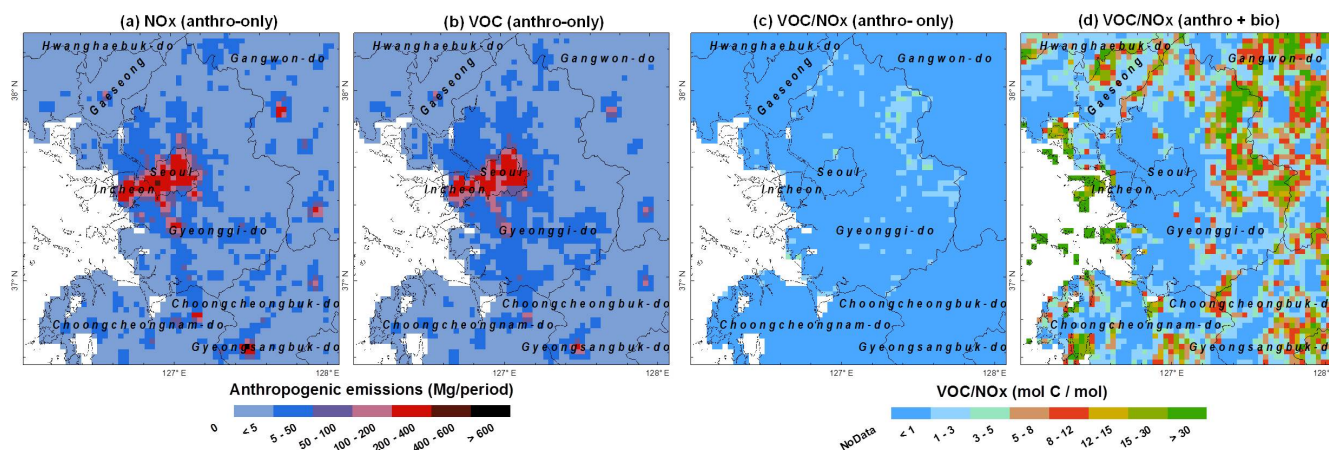
Figure 2. Geospatial distributions of estimated anthropogenic emissions for NO_x and VOCs and their ratios for the study period (May–June 2008). The VOC/NO_x ratio for the "anthro-only" case includes only the anthropogenic VOCs and NO_x emissions; the "anthro + bio" case includes both the anthropogenic and biogenic VOCs and NO_x emissions.

$VOC/NO_x < 3$), whereas noticeably increased VOC/NO_x ratios were distributed over some suburban and border areas for the case of combined anthropogenic and biogenic emissions (Fig. 2).

2.3 Statistical measures for quantitative evaluation

We investigated the impacts of different sources of PFT distributions on CTM O_3 predictions by examining the deviation of each data set (i.e., PFT, BVOC emissions, and O_3) from the norm (here, their mean values). The deviation of an individual data set at each grid cell (δGV) was calculated as follows:

$$\delta GV_{X|(i,j)} = GV_{X|(i,j)} - \overline{GV_{(i,j)}}, \tag{4}$$

where $GV_{X|(i,j)}$ is the grid cell value of an individual source of PFT X at a given grid cell location (i,j), and $\overline{GV_{(i,j)}}$ is the mean value for every source of PFT data at a given grid cell location (i, j). We consider the mean value as the best guess of the true values of each variable (PFT areas, biogenic emissions, or O_3 concentrations) because the mean value for each set of variable data from the given scenarios describes each central tendency of those variable data sets.

The effect of using different PFT data sets on CTM performances was assessed based on some statistical measures:

mean bias (MB), normalized mean bias (NMB), and normalized mean error (NME):

$$\text{mean bias (MB)} = \frac{1}{n} \sum_{i=1}^{n} (P_i - O_i), \tag{5}$$

$$\text{normalized mean bias (NMB)} = \frac{\sum_{i=1}^{n} (P_i - O_i)}{\sum_{i=1}^{n} O_i} \times 100, \tag{6}$$

$$\text{normalized mean error (NME)} = \frac{\sum_{i=1}^{n} |P_i - O_i|}{\sum_{i=1}^{n} O_i} \times 100, \tag{7}$$

where P is model prediction and O is observation.

We used hourly O_3, NO_x, and isoprene data gathered from the ambient air quality monitoring network (Fig. 1b), 148 ambient monitoring stations (AMS), and 8 photochemical air monitoring stations (PAMS) of the National Institute of Environmental Research to assess the CMAQ O_3 predictions over the Seoul-Gyeonggi Metropolitan Area for the study period. The US EPA suggested the informal performance standards for regulatory modeling practices of ±5 to ±15 % for NMB and ±30 to ±35 % for NME (Russell and Dennis, 2000).

Beginning with a performance check of CTM predictions, we consecutively investigated the effects of PFT area deviations on the results of MEGAN BVOC and CTM O_3 predictions.

3 Results and discussion

3.1 Performance evaluation of CMAQ simulations

The CMAQ-simulated O_3, NO_x, and isoprene concentrations were compared with the observed concentrations at 148 monitoring sites in the domain. Although the number of comparisons is limited, we assessed whether the CMAQ simulations with different PFT scenarios can provide reasonable O_3 concentrations in our study domain. Note that the main objective of the performance evaluation is not to determine which one of the three different PFT databases should be preferred.

Figure 3 shows a comparison between the observed and modeled O_3, NO_x, and isoprene concentrations for the period of 1 May–30 June 2008. The CMAQ predictions follow the observations reasonably well with a tendency to underpredict O_3 and isoprene and a tendency to overpredict NO_x. It should be noted that we displayed the averaged value of the CMAQ simulations with the different PFT scenarios in Fig. 3 because the time variations of the individual concentrations were not substantially different in the graph.

In general, the underpredictions in O_3 concentration could primarily be caused by the combined effect of multiple sources of uncertainty, such as in O_3 precursor (e.g., NO_x and VOC) emissions, meteorological fields, and so forth. For example, an overestimation of NO_x emissions and an underestimation of VOC emissions may have contributed to the overall underprediction of O_3 concentration under VOC-limited conditions. The overprediction of surface wind speeds and underprediction of ambient temperature for the simulation period may also have contributed to the O_3 underprediction. The overprediction of NO_x concentrations is primarily due to the overestimation of anthropogenic NO_x emissions, and the underprediction of isoprene concentrations is due to the combined effects of the overestimations of NO_x and underestimations of VOC and ambient temperature (i.e., underpredictions).

Table 3 shows a summary of the statistical measures indicating the CMAQ performance for O_3, NO_x, and isoprene for each PFT scenario at 148 monitoring sites in the domain. Although the effects of using different PFT distributions on CMAQ performance were not readily recognized across the 148 monitoring sites, the mean values from the CMAQ with KORPFT are between those from two other sites. This is because the O_3 and isoprene concentrations from CMAQ with MODIS-based PFT are higher than those with KORPFT, and the concentrations from CDP-based CMAQ are lower than those with KORPFT (Table 3).

For O_3, the overall evaluation statistics fall within the US EPA performance standards (i.e., ± 5 to ± 15 % for NMB and ± 30 to ± 35 % for NME), although CMAQ underpredicts when compared to the hourly O_3 observations. Because urban sites account for about 82 % of the total data sets (121 urban sites in 148 sites) in our comparison, the underpredictions at the urban sites contribute significantly to the overall CMAQ performance, resulting in underpredictions for the 1 h average O_3 concentrations. CMAQ shows the best performance at suburban sites, followed by background, urban, and roadside. For example, the respective values of MB, NMB, and NME for 1 h average O_3 are about -0.86 ppb, -2.14 %, and 21.64 at the suburban sites, about -2.76 ppb, -5.48 %, and 22.86 % at the background sites, about -8.10 ppb, -23.49 %, and 34.60 % at the urban sites, and about -5.37 ppb, -23.17 %, and 41.02 % at the roadside sites.

Comparing the CMAQ O_3 for the three different PFT scenarios (i.e., KORPFT, CDP, and MODIS), the MODIS case provides slightly better agreement compared to the observations than the other cases. The mean error and correlation coefficient are almost the same for three simulations. The resultant maximum O_3 differences between different PFT scenarios were only 0.4 % across the monitoring sites.

For NO_x, an important O_3 precursor, CMAQ overpredicts compared to the average value of 1 h observations for all sites, although it underpredicts at the roadside, suburban, and background sites, because of the significant contribution of the overprediction at the urban sites to the overall performance. This inaccuracy in NO_x predictions may occur due to the effect of uncertain estimations of anthropogenic NO_x emissions, such as overestimations at the urban sites and underestimations at the other sites. The respective values of MB, NMB, and NME for 1 h average NO_x are about 14.47 ppb, 40.55 %, and 71.39 % for all sites, about 19.68 ppb, 66.22 %, and 85.34 % at the urban sites, about -8.20 ppb, -9.78 %, and 59.01 % at the roadside sites, about -4.16 ppb, -31.43 %, and 63.23 % at the suburban sites, and about -12.85 ppb, -75.16 %, and 85.12 % at the background sites.

In contrast, for isoprene, another important O_3 precursor, CMAQ underpredicts compared to the average value of 1 h observations at PAMS distributed over urban (four sites), suburban (three sites), and background (one site) areas in the domain. The respective values of MB, NMB, and NME for 1 h average isoprene are about 0.05 ppb, -21.07 %, and 63.15 % throughout all the sites. Among the three CMAQ isoprene results, the CMAQ provided values closer to the observations with MODIS (MB $= -0.02$ ppb and NMB $= -7.78$ %) than with the others (MB $= -0.05$ and NMB $= -22.62$ % with KORPFT; MB $= -0.08$ and NMB $= -32.82$ % with CDP). For correlation, the CMAQ provided slightly higher r values with KORPFT ($r = 0.622$) than with the others ($r = 0.598$ with MODIS and $r = 0.591$ with CDP).
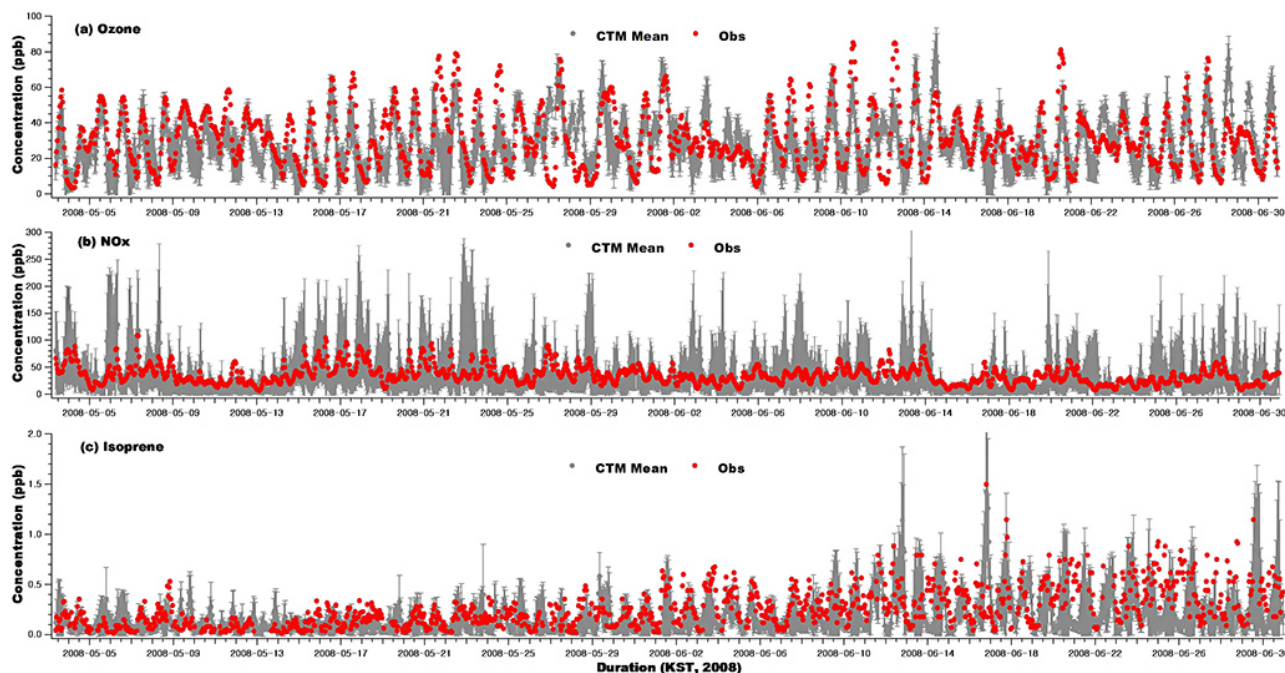
Figure 3. Evaluation of CMAQ results for hourly O_3 concentrations. Note that "CTM Mean" represents the average of the CMAQ predicted O_3 concentrations from the three different BVOC scenarios, and "Obs" represents the average of measured O_3 concentrations at monitoring stations across the domain.

Table 3. Performance statistics for CMAQ O_3 with the three different BVOC emission scenarios against the measured O_3 concentrations.

		Mean (ppb)			MB (ppb)			NMB (%)			NME (%)			Correlation coeff.	
	Obs.	KORPFT	CDP	MODIS	KORPFT	CDP	MODIS	KORPFT	CDP	MODIS	KORPFT	CDP	MODIS	KORPFT	CDP	MODIS
O_3 (overall: 148 sites; urban: 121 sites; roadside: 18 sites; suburban: 6 sites; background: 3 sites)																
Overall	30.16	25.69	25.66	25.77	−4.47	−4.50	−4.39	−14.82	−14.92	−14.56	29.54	29.54	29.53	0.740	0.740	0.739
Urban	34.49	26.37	26.32	26.47	−8.12	−8.17	−8.02	−23.54	−23.68	−23.25	34.61	34.58	34.62	0.671	0.672	0.670
Roadside	24.75	19.00	18.97	19.07	−5.75	−5.77	−5.68	−23.23	−23.33	−22.96	41.00	40.99	41.07	0.658	0.659	0.657
Suburban	40.24	39.35	39.31	39.46	−0.88	−0.92	−0.78	−2.19	−2.29	−1.94	21.63	21.67	21.63	0.790	0.789	0.788
Background	50.43	47.67	47.59	47.73	−2.76	−2.83	−2.69	−5.47	−5.62	−5.34	22.86	22.84	22.87	0.494	0.495	0.493
NO_x (overall: 148 sites; urban: 121 sites; roadside: 18 sites; suburban: 6 sites; background: 3 sites)																
Overall	35.69	50.15	50.16	50.15	14.47	14.48	14.46	40.54	40.57	40.53	71.38	71.39	71.38	0.539	0.539	0.539
Urban	29.71	49.39	49.40	49.38	19.67	19.68	19.67	66.21	66.25	66.20	85.33	85.34	85.33	0.567	0.567	0.567
Roadside	83.82	75.62	75.63	75.61	−8.20	−8.19	−8.21	−9.79	−9.77	−9.79	59.01	59.01	59.01	0.264	0.264	0.264
Suburban	13.25	9.08	9.10	9.08	−4.17	−4.15	−4.17	−31.46	−31.35	−31.49	63.26	63.20	63.23	0.320	0.320	0.320
Background	17.10	4.25	4.25	4.25	−12.85	−12.86	−12.85	−75.15	−75.17	−75.17	85.10	85.14	85.11	N.C.	N.C.	N.C.
Isoprene (PAMS 8 sites)																
PAMS	0.23	0.18	0.15	0.21	−0.05	−0.08	−0.02	−22.62	−32.82	−7.78	60.98	63.69	64.78	0.622	0.591	0.598

N.C.: not correlated.

These results suggest that all three different PFT data sets are acceptable options for regulatory modeling practices when evaluated in terms of hourly O_3 concentrations following the US EPA performance standards.

3.2 Comparison of PFTs and biogenic emissions distributions

In our modeling domain, CDP (25 333 km^2) delivered the largest PFT area followed by KORPFT (25 128 km^2) and MODIS (24 101 km^2). Although the three sources of the PFT data delivered similar cover areas, clear differences were identified in the composition of the PFT classes (i.e., BT, NT, SB, and HB). For example, for BTs (the major source of biogenic isoprene), MODIS (10 115 km^2) delivered the largest area followed by KORPFT (5318 km^2) and CDP (4539 km^2). For NTs (the major source of biogenic terpene), KORPFT (8019 km^2) delivered the largest area followed by CDP (4967 km^2) and MODIS (1258 km^2). For SB (the second most major source of biogenic isoprene), CDP (3920 km^2) delivered the largest area followed by KORPFT (2569 km^2) and MODIS (769 km^2). For HBs (the major

source of biogenic, or soil, NO), MODIS (11 959 km²) delivered the largest area followed by CDP (11 907 km²) and KORPFT (9223 km²).

A comparison of the spatial patterns of the PFT area deviations allowed us to understand qualitative discrepancies among the sources of PFT data. Figure 4 shows the spatial distributions of the δPFT areas for each PFT scenario for the total vegetation and PFT classes (i.e., BT, NT, SB, and HB). The δPFT area of each PFT scenario in the study domain was produced based on Eq. (4). The resulting maps show several noticeable features in the δPFT area distributions:

1. KORPFT delivered larger BT covers over the Seoul and Incheon Metropolitan Areas and other regions (Fig. 4b1).

2. KORPFT delivered larger NT covers across the domain. There are hot spots over Gangwon-do and the border areas of Gaeseong and Hwanghaebuk-do (Fig. 4c1).

3. CDP lacked PFT distribution information for some islands and costal city areas off the Incheon Metropolitan Area (the brightest yellow color area in Fig. 4a2).

4. CDP delivered comparatively larger NT- and HB-type vegetation covers that concentrated in the Seoul and Incheon Metropolitan Areas and were widespread across the domain (Fig. 4a2, c2, d2, and e2) along with larger BT covers over the Gaeseong and Hwanghaebuk-do areas (Fig. 4b2).

5. MODIS delivered larger BT vegetation covers across the domain except for the Seoul and Incheon Metropolitan Areas and Gaeseong and some of its neighborhood areas. There are hot spots over Gangwon-do, Choongcheongbuk-do, and Choongcheongnam-do) (Fig. 4b3).

From the subsequent investigation of the δBVOC emission distributions, we found that the patterns from biogenic emission distributions closely resembled those from the PFT distributions in the study domain, except for the missing biogenic emission zones detected over the Seoul and Incheon Metropolitan Areas and some of their neighborhood areas (Fig. 5). A brief summary of the features that were similar to those from the δPFT distributions follows:

1. KORPFT estimated comparatively higher isoprene emissions from some border areas (between the Gyeonggi-do and Gaeseong areas and between the Gyeonggi-do and Choongcheongbuk-do areas)) as well as some islands areas (Fig. 5b1) due to the influences of BT cover over these areas.

2. KORPFT estimated larger and widespread terpene emissions across the domain except for Seoul and some border areas, (Fig. 5c1) due to the influences of NT cover over these areas.

3. CDP omitted biogenic emissions from some island and costal city areas off the Incheon Metropolitan Area (Fig. 5a2, b2, c2, and d2) due to omitted PFT areas.

4. CDP estimated comparatively higher and widespread terpene and NO emissions, except for Seoul and some border areas, across the domain (Fig. 5c2 and d2) due to the influence of NT and HB cover, and higher isoprene emissions from the Gaeseong and Hwanghaebuk-do areas (Fig. 5b2) due to the influence of BT cover.

5. MODIS estimated larger isoprene emissions across the domain except for the Seoul and Incheon Metropolitan Areas and Gaeseong and some of its neighborhood areas (Fig. 5b3) due to the influence of BT cover over these areas. There are hot spots over Gangwon-do, Choongcheongbuk-do, and Choongcheongnam-do.

We observed zones of missing biogenic emissions in many locations. These zones (the red zones in Fig. 5a–d) occurred because of the problem of primary data missing. In this study, there were two causes of missing data: PFT and LAI data omissions. Either PFT or LAI data omission (Fig. S1 in the Supplement) can inhibit the computation of LAIv and then inhibit the computation of biogenic emissions. The cause of the LAIv computation inhibition (i.e., missing LAIv data) was different according to the sources of PFT data. For the KORPFT, most (approximately 96 %) of the missing LAIv data was due to the omitted MODIS LAI data (Fig. S1b in the Supplement). For the CDP and MODIS, approximately 60 % of the missing LAIv data was due to the omitted MODIS LAI data, and approximately 40 % of the missing LAIv data was due to the omitted PFT data (Fig. S1a2 and a3 in the Supplement). In other words, the MODIS LAI data omission issue caused approximately 96 % of the zones with missing biogenic emissions for the KORPFT scenario and approximately 60 % of the zones with missing biogenic emissions for the CDP and MODIS scenarios, whereas the PFT data omission issue caused approximately 4 % of the zones with missing biogenic emissions for the KORPFT scenario and approximately 40 % of the zones with missing biogenic emissions for CDP and MODIS scenarios.

This problem of lacking biogenic emission can affect the surface-level O_3 simulations of CTM. This issue will be discussed in Sect. 3.3.

3.3 Effect of missing data on CMAQ O_3 simulations

It is worthwhile to investigate the effect of missing data, mentioned in Sect. 3.2, on biogenic emissions and O_3 predictions in the modeling domain.

Because the MEGAN model assumes that plant leaves cover only that part of the grid cell containing vegetation (Guenther et al., 2006), we supplemented the LAIv values into the grid cells that lacked LAIv but contained PFT values. We made up for each LAIv missing data location by

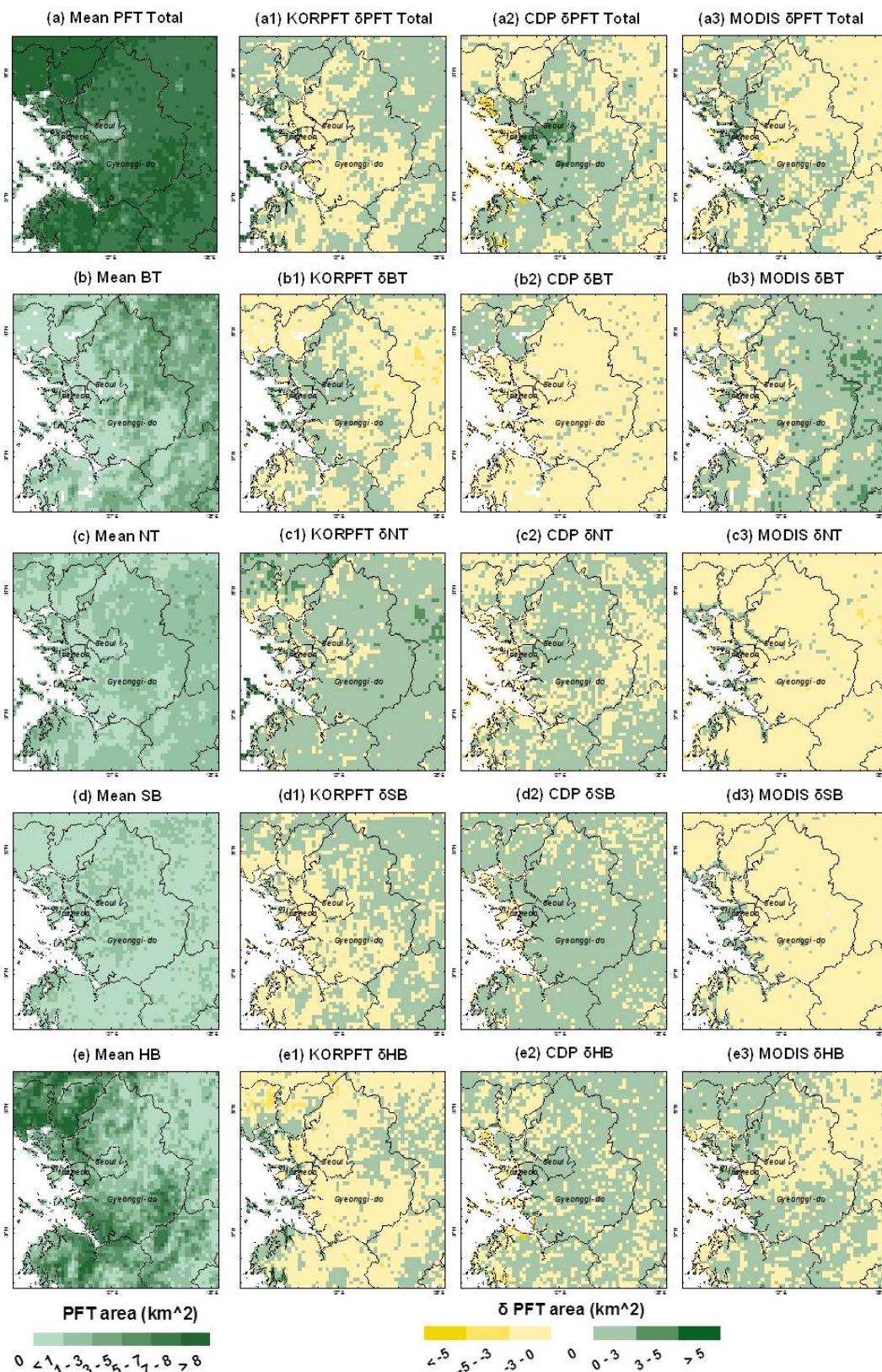**Figure 4.** Spatial distribution of the PFT area deviations for each PFT scenario. The mean spatial distributions for PFT total and the PFT classes (BT, NT, SB, and HB) were derived by averaging the three different distribution data sources.
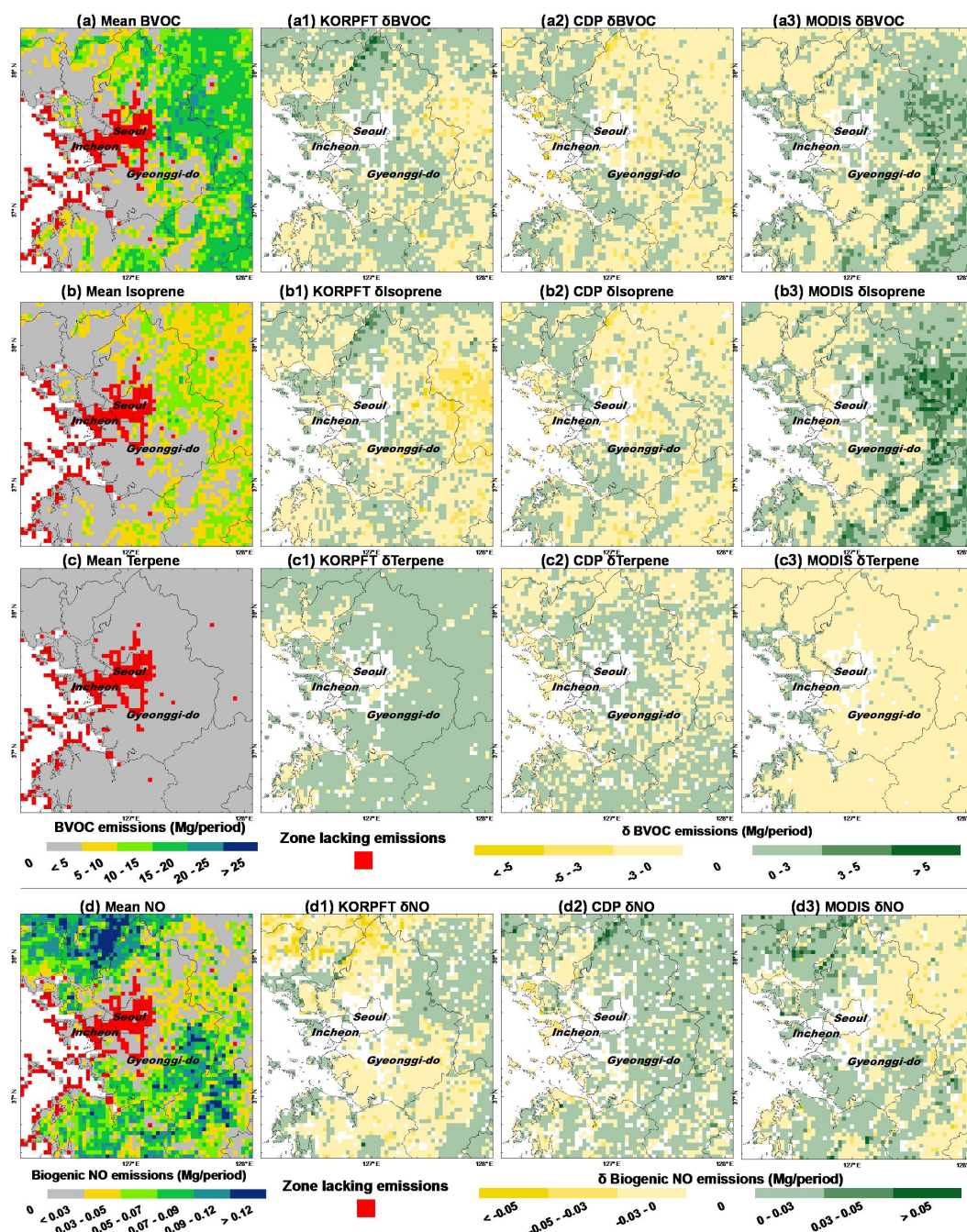
Figure 5. Spatial distribution of biogenic emission deviations for each PFT scenario. The red zones denote the zones lacking biogenic emissions during the MEGAN emission modeling.

using each nearest LAIv value as a supplement. As a result, 274 locations (approximately 96 % of the total 285 locations with missing LAIv) for the KORPFT, 165 locations (approximately 58 % of the total locations with missing LAIv) for CDP, and 165 locations (approximately 59 % of the total locations with missing LAIv) for MODIS were supplemented by their nearest LAIv values. Afterwards, we performed MEGAN biogenic emission modeling and CMAQ

simulations and then assessed the effect of additional missing BVOC emissions data on CMAQ O_3 predictions.

The simulated results show the longitudinally widespread increase in O_3 concentrations (Fig. 6) at the LAIv supplemented grid cells and usual grid cells. The KORPFT scenario shows the most widely spread O_3 increment pattern caused by the wider influence of the additional BVOC emissions compared with those of other PFT scenarios. The highest
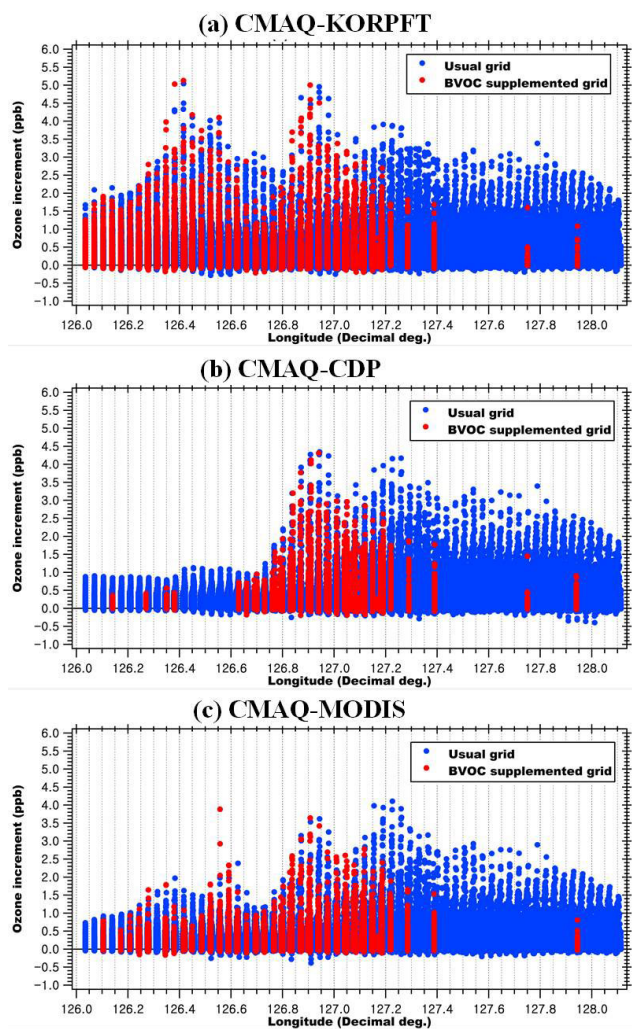
Figure 6. Longitudinal distribution of the incremented ozone concentrations due to the effect of additional BVOC emissions for each PFT scenario via LAIv input supplementation for MEGAN modeling. A filled red circle denotes an O$_3$ increment in an LAIv supplemented grid cell. A filled blue circle indicates an O$_3$ increment in a normal grid cell.

O$_3$ increment values for each PFT scenario occurred at an LAIv-supplemented grid cell (KORPFT and CDP scenarios) or in the neighborhood of the LAIv-supplemented grid cells (MODIS scenario). The simulated increment in hourly O$_3$ concentration was up to 5, 4.5, and 4 ppb for KORPFT, CDP, and MODIS scenarios, respectively. This result suggests that the problem of lacking input data for MEGAN biogenic emission modeling can cause spatially far different 1 h O$_3$ prediction results.

Although PFT area loss might be inevitable at this point in time, it must be resolved in the near future to obtain more accurate biogenic emission and chemistry transport modeling results.

3.4   PFT class-dependent BVOC emissions and their O$_3$-forming potentials

The links between BVOC emissions and their O$_3$-forming potentials (OFPs) were further investigated by focusing on the compositional differences and the OFPs of the three BVOC emissions groups from the different PFT scenarios (Fig. 7). From these investigations, we found that isoprene was the most important BVOC compound showing the largest contributions to the total BVOC emission amounts and potential O$_3$ formation.

In the analyses, we reconstructed the 22 SAPRC-99 VOC species of MEGAN into 10 BVOC compound groups for a clearer presentation of the BVOC emission distributions. The 10 reconstructed compound groups were ALKs (sum of alkanes, i.e., ALK1-ALK5), AROs (sum of aromatics, i.e., ARO1 and ARO2), CCHO (CCHO), ETHENE (ethene), HCHO (HCHO), ISOPRENE (isoprene), MEOH (MEOH), OLEs (sum of olefins, i.e., OLE1 and OLE2), Other Ox-Orgs (sum of ACET, BALD, CCO_OH, HCOOH, MEK, RCHO, and RCO_OH), and TERPENE (terpene). To examine the OFP of each PFT scenario, we employed the maximum incremental reactivity (MIR) (Carter, 2000b). The MIR is a useful quantitative measure of the impact of a VOC on O$_3$ under high NO$_x$ conditions (i.e., grams of O$_3$ generated/grams of VOC added) under which O$_3$ is most sensitive to VOCs and which represent near-source or urban areas (Carter, 2000b). The MIR was derived from several box model scenarios representing various urban areas with NO$_x$ inputs adjusted to yield maximum sensitivities of ozone to changes in VOC levels (Carter, 2000b). The MIR has been adopted in the state of California for the purpose of implementing reactivity-based regulations (CARB, 1993) and often used as a general reactivity scale to study the impact of VOC on O$_3$ formation (Xie et al., 2008; Zheng et al., 2009; Carter and Seinfeld, 2012).

Among the 10 BVOC compound groups, isoprene is shown to have the largest contribution to the total BVOC emissions (about 52 %) followed by MEOH, terpene, and other VOCs (Fig. 5a). Among the different PFT scenarios applied to the MEGAN biogenic emission modeling, MODIS derived the highest isoprene emissions because of the highest BT areas, followed by KORPFT and CDP (Fig. 7b). A distinct result with the KORPFT scenario is the highest level of terpene emissions, shown in Fig. 7b, due to the larger NT areas of KORPFT (Fig. 4c1) compared to the other PFT scenarios.

By simply multiplying BVOC emissions with MIR values (Fig. 7c), we calculated domain-wide total OFPs of BVOC emissions. The calculated distributions of OFPs for each PFT scenario (Fig. 7d) were mainly affected by the spatial distributions of isoprene emissions. For example, most locations with higher OFPs overlapped with the BT and the isoprene emission hot spots in the study domain (Figs. 4b, 5b, and 7d). From the ozone-forming potential computation, we derived the following proportion
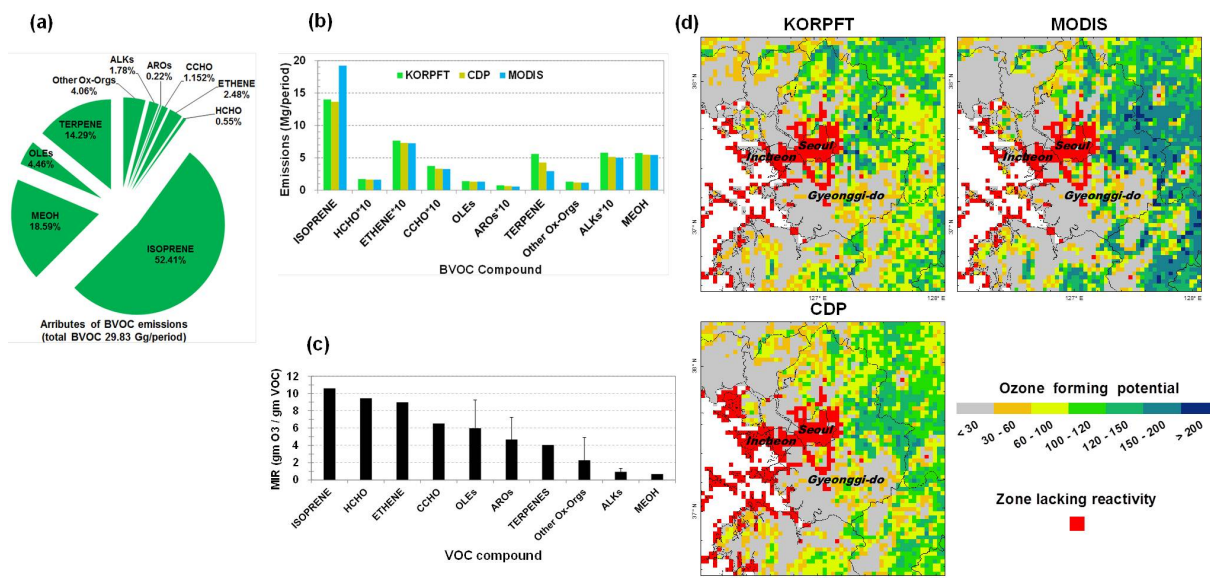
Figure 7. Composition and distribution of the estimated BVOC emissions, along with spatial distribution of the estimated net ozone-forming potentials of biogenic VOC emissions. The total BVOC in **(a)** was derived by averaging the total BVOC emissions of the three sources of data. The 10 BVOC compound groups in **(b)** are arranged in descending order following their O_3 reactivity, i.e., MIR shown in **(c)**. The distributed net ozone-forming potentials in **(d)** were calculated by aggregating the O_3-forming potentials (OFPs) for 10 BVOC compound groups calculated by multiplying the emissions by their corresponding MIR values at each grid cells. The red zones denote the zones missing BVOC reactivity due to the lacking biogenic emissions during the OFP calculations.

for the maximum O_3-forming potentials in the study domain: $MODIS : KORPFT : CDP = 1 : 0.82 : 0.78$. Assuming that under high NO_x level our study domain experiences optimal balance between VOC and NO_x to generate O_3, this rough estimation implies that about 22 % of the maximum difference in the O_3 concentrations can occur between the different PFT scenarios. For example, when the MODIS scenario estimates 60 ppb of domain-average O_3 concentration, the probable O_3 concentrations from the KORPFT and CDP scenarios would be 49.2 and 46.8 ppb, respectively. There are two points to be noted here because the MIR values used for OFP calculations are from the chamber experiments (Carter, 2000b). Firstly, the proportional expression (i.e., $MODIS : KORPFT : CDP = 1 : 0.82 : 0.78$) we obtained was simply a rough one and it is not applicable to the relevant atmospheric conditions for O_3 formation over some areas. From Fig. 7d, we can easily recognize that the estimated proportional expression is not applicable to some areas, e.g., areas with missing BVOC reactivity zones due to the lack of BVOC emissions and areas with small OFP gaps between the different PFT scenarios. Secondly, the MIR-scale-based OFP calculation may not duplicate the relevant atmospheric conditions and chemistry of VOC associated with O_3 formation for the period and domain of this study. Thus, we can only expect that the areas with higher OFP levels have a greater possibility of experiencing higher levels of O_3 formation in the presence of higher NO_x conditions. This MIR-OFP application issue will be discussed further in Sect. 3.6.

Despite the mentioned limitations of MIR-OFP, this approach helps to evaluate the relative importance of biogenic VOC compounds in the production of ground-level ozone in the current study domain. In our OFP calculation, both the isoprene and terpene emission amounts were the first and second largest contributors to the O_3 concentrations (about 79 % contribution by isoprene and 9 % contribution by terpene). The result suggests that the primary and secondary BVOC species of concern are isoprene and terpene because of their high reactivities and large emissions across the domain.

3.5 Spatial relationship between δPFT and δO_3

The results in Sects. 3.2–3.4 suggest some qualitative and quantitative evidence of a causal link between the δPFT areas, δBVOC emission estimations, and δO_3 predictions. To analyze this link more quantitatively, we performed a spatial regression analysis with each gridded data set in a vegetated region by applying the geographically weighted regression technique with a locally compensated ridge (LCR-GWR) (Brunsdon et al., 1996; Fotheringham et al., 2002). The LCR-GWR is a useful method to address spatially varying relationships between the dependent and independent variables and to reduce the problem of collinearity among the explanatory (independent) variables (Brunsdon et al., 2012; Lu et al., 2013; Gollini et al., 2013). The regression models that include δBT, δNT, δSB, and δHB as independent

variables were fitted by LCR-GWR techniques (refer to Appendix A). As the counterpart, the parameters of the ordinary least square (OLS) were also estimated. The main objective of this spatial regression analysis was to show that changes in O_3 concentrations in the modeling grid cells are influenced by the changes in PFT distributions in the modeling grid cells. Because of the short photochemical lifetimes of BVOC (e.g., isoprene ~ 2 h) (Atkinson and Arey, 2003), many studies used the assumption that the emitted BVOCs from local biogenic emission sources (i.e., PFT area distributions) immediately affect the levels of local surface O_3 concentrations rather than move over a long distance. However, this assumption may not be appropriate for our spatial regression analysis. Because of the relatively high wind speed (mean wind speed at 10 m altitude $= 3.7$ m s^{-1}) over our finer modeling grid system (3 km $\times$ 3 km), isoprene emissions generally may not fully react in the grid in which they are emitted. The GWR technique was applied to appropriately address the relationships between those non-quantified O_3 variations with spatial dependency and distributed spatial effects of biogenic sources. Specifications of OLS and LCR-GWR are briefly described in Appendix A, and the regression modeling was conducted in the following three steps:

1. Fit global OLS models to the data sets of a dependent variable (δO_3) and four explanatory variables (i.e., δBT, δNT, δSB, and δHB) for the three PFT scenarios.

2. Fit LCR-GWR models to the data sets of a dependent variable (δO_3) and four explanatory variables (i.e., δBT, δNT, δSB, and δHB) for the three PFT scenarios.

3. Assess the models from Steps 1 and 2.

In LCR-GWR simulation, the GWR model (Eq. A3) was calibrated using dual optimization of both the bandwidth and ridge terms. An optimal bandwidth was selected by leave-one-out cross-validation (Bowman, 1984). Using a Gaussian kernel with the optimal bandwidth, the weighting matrix of GWR was specified. At the locations where collinearity among the independent variables affects the corresponding local coefficient estimates (local condition number > 30), GWR applied a local ridge regression. The functions used in the analysis with OLS and LCR-GWR are included in the GWmodel R package (Lu et al., 2013; Gollini et al., 2013).

The estimated coefficients of δO_3 with respect to δBT, δNT, δSB, and δHB, as well as regression model diagnostic information for OLS and LCR-GWR, are summarized in Table 4.

The OLS provided a global relationship of delta O_3 with δBT, δNT, δSB, and δHB. The OLS for KORPFT demonstrates that gaining the areas of the two major biogenic isoprene sources (i.e., $+0.016 \cdot \delta$BT and $+0.001 \cdot \delta$SB) and the one major biogenic terpene source (i.e., $+0.002 \cdot \delta$NT) contributes to O_3 increase, whereas gaining the area of the major soil NO$_x$ source (i.e., $-0.001 \cdot \delta$HB) contributes to O_3 decrease across the domain.

The OLS for CDP suggests that gaining areas of BT contributes to O_3 increase, whereas gaining NT, SB, and HB areas contributes to O_3 decrease across the domain.

The OLS for MODIS indicates positive contributions from gaining BT, SB, NT, and HB areas to O_3 increase in the domain.

However, these OLS parameter estimates of delta O_3 with respect to δBT, δNT, δSB, and δHB may have the poor applicability in some locations because these models ignored the local spatial effects. The summarized LCR-GWR results reveal the improvements of the model fits and calibrated relationships between model parameters (i.e., δO_3 and δPFTs).

LCR-GWR models provide better fits than the OLS models for the PFT scenarios, where the bias-corrected Akaike information criterion (AICc) values have been reduced by 7666 for KORPFT, 9317 for MODIS, and 9678 for CDP. As expected, the parameter estimates with LCR-GWR show spatially varying impacts of δBT, δNT, δSB, and δHB on δO_3 in the study domain (Fig. S2 in the Supplement). The revealed relationship of δO_3 with δBT, δNT, δSB, and δHB from the summarized coefficient estimates of LCR-GWR for three PFT scenarios is parallel to that of the OLS for KORPFT: gaining BTs, SB, and NTs contributes to O_3 increase, whereas gaining HBs contributes to O_3 decrease in the domain.

To assess the prediction accuracy of the OLS and LCR-GWR models, we compared the δO_3 values predicted by each regression model in turn with those derived from CMAQ outputs in Sect. 3.3. We performed local assessments at the missing data zones (the black squared zones in Fig. S2 in the Supplement). Because the parameter values for the LCR-GWR model were not estimated in the missing data zones, we used the nearest parameter values to given missing data locations for the LCR-GWR model prediction. Results of the local assessment of model prediction accuracy are shown in Fig. 8 and Table 5. The prediction accuracy for the LCR-GWR model was greater than that for the OLS model, which suggests that stationary relationships from OLS do not fit the local situation. Among the three LCR-GWR fits, the most accurate δO_3 predictions were found with KORPFT ($r^2 = 0.57$, SEE $= 0.0081$, and $F = 356.4$), followed by MODIS ($r^2 = 0.4$, SEE $= 0.0087$, and $F = 112.6$) and CDP ($r^2 = 0.33$, SEE $= 0.0086$, and $F = 79.8$).

3.6 Impact of PFT distribution differences on hourly O_3 predictions

From the investigation of the daytime O_3 episodes, we found noticeable deviation patterns of hourly CMAQ O_3 predictions with different PFT scenarios in our modeling domain. Figure 9a and b show the time change of the spatial distributions of δO_3 and mean O_3 concentrations in the surface layer for the daytime of 29 May 2008 and 30 June 2008, respectively. The δO_3 distributions were produced by subtracting the hourly CMAQ O_3 distributions for each scenario (i.e.,

Table 4. The relationship between δO_3 deviations and the corresponding δPFT area across the domain derived from the OLS and LCR-GWR models. The third to seventh columns show the estimated coefficients of the explanatory variables and the intercept of the regression models. The standard errors of the coefficient estimates by OLS and the minimum and maximum values of the coefficient estimates by LCR-GWR are enclosed by square brackets. The 8th column shows the adjusted coefficient of determination representing the explanation power of the regression model. The standard errors of the coefficient estimates by OLS and the minimum and maximum values of the coefficient estimates by LCR-GWR are enclosed by square brackets. The 9th to the 11th columns show test statistics for the goodness of the regression model fits. A higher F statistic, a lower RSS, and a lower AICc value indicates better fit of the regression model and a lower p value indicates greater significance of the regression model. The 12th column shows the discrepancy of AICc between OLS and LCR-GWR fits. A higher reduction (at least over 3) indicates the greater benefits of moving from an OLS to an LCR-GWR. The 13th column shows the number of the fitted data. It should be noted that the data used for this regression analysis are from each of the individual modeling grid locations where both PFT and LAI values are not missing. The significance levels of the parameters and the model fits are (***) for p value < 0.001, (**) for 0.001–0.01, (*) for 0.01–0.05, (●) for 0.05–0.1, and ( ) for > 0.1. RSS: residual sum of square. N.D. stands for "not determined".

Regression model	PFT scenario	Parameters					adj. R^2	F statistic	RSS	AICc	AICc reduction from OLS fit	n
		δBT	δNT	δSB	δHB	constant						
OLS	KORPFT	0.016 [0.001] (***)	0.002 [0.000] (***)	0.001 [0.000] (**)	−0.001 [0.000] (***)	−0.015 [0.000] (***)	0.51	783 (***)	0.580	−16 975.03	0	2977
	CDP	0.007 [0.001] (***)	−0.002 [0.001] (*)	−0.001 [0.001] (●)	−0.00003 [0.000] ( )	−0.05104 [0.001] (***)	0.07	57 (***)	1.050	−15 206.48	0	2977
	MODIS	0.021 [0.001] (***)	0.002 [0.001] (*)	0.006 [0.001] (***)	0.002 [0.001] (**)	0.056 [0.001] (***)	0.35	407 (***)	2.368	−12 785.09	0	2977
LCR-GWR	KORPFT	0.003 [−0.012, 0.015]	0.0004 [−0.015, 0.009]	0.001 [−0.013, 0.011]	−0.002 [−0.015, 0.006]	−0.015 [−0.068, 0.005]	N.D.	N.D.	0.023	−24 641.13	7666.10	2977
	CDP	0.003 [−0.019, 0.035]	0.001 [−0.021, 0.015]	0.001 [−0.018, 0.012]	−0.001 [−0.019, 0.013]	−0.055 [−0.104, −0.016]	N.D.	N.D.	0.024	−24 523.92	9317.44	2977
	MODIS	0.004 [−0.020, 0.026]	0.001 [−0.020, 0.020]	0.002 [−0.031, 0.044]	−0.001 [−0.019, 0.016]	0.066 [0.014, 0.155]	N.D.	N.D.	0.049	−22 463.56	9678.47	2977

Table 5. Evaluation summary for the global OLS and LCR-GWR models of delta O_3 for three PFT scenarios. The significance levels of the parameters based on the t statistic are (***) for p value < 0.001, (**) for 0.001–0.01, (*) for 0.01–0.05, (●) for 0.05–0.1, and ( ) for > 0.1. SEE: standard error of estimate.

Regression model	PFT scenario	R^2	Slope	Constant	SEE	F statistic	P value	n
OLS	KORPFT	0.22	0.447 (***)	0.0039 (***)	0.0109	77.14	< 0.0001	274
	CDP	0.09	1.031 (***)	0.0122	0.01001	15.9	0.0001	165
	MODIS	0.19	0.384 (***)	0.034 (***)	0.01015	38.36	< 0.0001	167
LCR-GWR	KORPFT	0.57	1.074 (***)	0.0086 (***)	0.0081	356.4	< 0.0001	274
	CDP	0.33	1.063 (***)	0.0071	0.0086	79.8	< 0.0001	165
	MODIS	0.4	0.806 (***)	0.006	0.0087	112.6	< 0.0001	167

KORPFT, CDP, and MODIS) from the average of these three CMAQ O_3 distributions (see Eq. 4). The δO_3 distributions show that MODIS usually develops higher positive deviations (Fig. 9a11–a15 and 9b11–b15) while CDP usually develops higher negative O_3 deviations (Fig. 9a6–a10 and 9b6–b10) across the domain. The KORPFT develops both negative and positive O_3 deviations (Fig. 9a1–a5 and 9b1–b5) across the domain. Interestingly, the patterns of these negative or positive O_3 deviations develop in concert with enhancements in ambient O_3 concentration.

The 29 May 2008 episode shows relatively high concentrations of the simulated O_3 over some island and costal city areas off the Incheon Metropolitan Area. At the time, the Incheon coastal and neighborhood areas were located downwind of the Incheon industrial complex and the Gyeonggi-do and Seoul Metropolitan Areas, where anthropogenic NO_x emissions are very strong (emission ratio for VOC $/ NO_x < 3$; see Fig. 2d), and suffered a temperature inversion that limited the vertical mixing of pollutants (see the grey-hatched zones in Fig. 9a). While the released BVOC was trapped within the inversion layer, the easterly winds brought anthropogenic NO_x-rich air from the Incheon industrial complex and some areas further inland (e.g., Gyeonggi-do and Seoul) to the inversion area. With the temperature inversion, mixing between the local BVOC emissions and the intruded anthropogenic NO_x and VOC produced high concentrations of O_3 through photochemical reaction. Eventually, the Incheon coastal and neighborhood areas suffered consistently high O_3 development. The most distinctive feature of the δO_3 pattern on 29 May 2008 is the level of deviations of O_3 (maximum
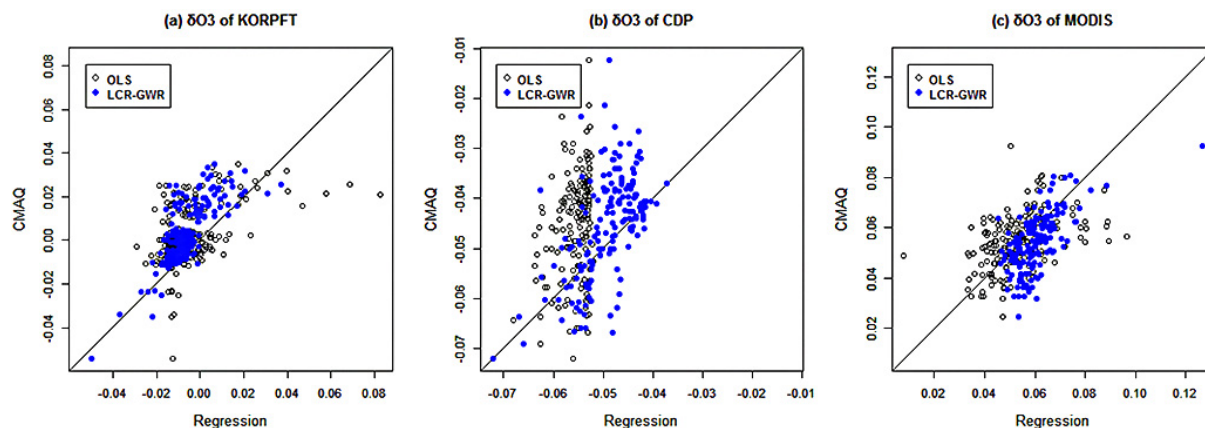
Figure 8. CMAQ-derived δO_3 vs. the fitted δO_3 under each regression model for three PFT scenarios.

difference is up to 7 ppb: 3 ppb of δO_3 for MODIS and -4 ppb of δO_3 for CDP at 03:00 p.m.) that developed and consistently remained in the temperature inversion zone (i.e., the Incheon coastal and neighborhood areas). The location of this deviation is coincident with the location of the PFT deviation in Fig. 4 and the BVOC deviation in Fig. 5. The high negative deviation of CDP O_3 is associated with the influence of the missing PFT areas of CDP (see Figs. 4a2, b2, c2, d2, and e2) on the CMAQ O_3 predictions over these areas. The high positive deviation of O_3 with the MODIS scenario is associated with the impact of the larger PFT areas (e.g., δBT $\sim 1.5 km^2$ and δNT $\sim 2.1 km^2$) of MODIS on the CMAQ predictions over the area (see Fig. 4b3 and c3, Fig. 5b3 and c3, and Fig. 7b).

The 30 June 2008 episode shows more widespread concentrations of the simulated O_3 throughout the domain. At the time, most of the high O_3 concentration regimes (e.g., see the O_3 contour in Fig. 9b3–b4) were located downwind from higher anthropogenic NO_x emission source areas (Fig. 2a). The anthropogenic NO_x emissions transported from urban center areas affected several border areas (e.g., the border areas between Seoul and Gyeonggi-do, between Gyeonggi-do and Gaeseong, between Gyeonggi-do and Gangwon-do, and between Gyeonggi-do and Choongcheongnam-do and Choongcheongbuk-do, etc.) and some suburban areas (e.g., the southern part of Gyeonggi-do and the northern part of Choongcheongnam-do) and some suburban areas (e.g., the southern part of Gyeonggi-do and the northern part of Choongcheongnam-do) where abundant hydroperoxy and organic peroxy radicals are generated through the oxidation of large amounts of locally emitted BVOC, resulting in high O_3 concentrations upon photolysis. Although the Incheon coastal area and the neighborhood areas underwent a temperature inversion similar to the 29 May 2008 episode, these areas were not consistently affected by anthropogenic NO_x transported from inland areas due to a wind direction change (i.e., easterly to westerly). One of the most noticeable features of the δO_3 pattern for 30 June 2008 consists of the

strong deviations of O_3 that develop over Gangwon-do (difference is up to 10 ppb: -4 ppb of δO_3 for KORPFT and 6 ppb for MODIS) and Choongcheongbuk-do (differences of 9–10 ppb: -4 ppb of δO_3 for KORPFT, -3 ppb for CDP, and 6 ppb for MODIS) at 05:00 p.m. Another noticeable feature consists of the consistent deviations, which are not small (difference up to 5 ppb: ~ -2 ppb of δO_3 for CDP and ~ 3 ppb for MODIS), that develop over the border areas between East Seoul and Gyeonggi-do. Figures 4 and 7 show that these higher O_3 deviations are associated with higher deviations in biogenic isoprene emissions that occur due to the higher difference of BT area among the different PFT scenarios (e.g., the positive δBT area of MODIS ($\sim 3.5 km^2$) and the negative δBT areas of other PFTs ($\sim -3 km^2$)).

Interestingly, the 9–10 ppb of difference between CMAQ O_3 with the different scenarios, detected in the 30 June 2008 episode, is about 11 % (i.e., [10 ppb of CMAQ O_3 difference]/[90 ppb of CMAQ O_3] $\times$ 100). This corresponds roughly to half of the maximum difference (i.e., 22 %) estimated based on application of the MIR-OFP approach in Sect. 3.4. As mentioned earlier in Sect. 3.4, this gap could be the result of differences between the assumed CMAQ atmospheric conditions for surface level O_3 prediction and the chamber simulation conditions for MIR estimation: the O_3 simulations were conducted for the period of May–June 2008 under much narrower distributions of NO_x and temperature in the atmosphere (domain-averaged NO_x range: 0–24 ppb; domain-averaged temperature range: 6–26 °C) compared to those of the chamber experiment for MIR developments (chamber environment NO_x range: 150–1000 ppb; chamber experiment temperature: 22–43 °C) (refer to Table B-1 in Carter, 2000b). This indicates that our CMAQ atmospheric conditions for photochemical O_3 formation were not fully developed to derive the maximum O_3 reactivities of BVOC emissions from different PFT scenarios. Thus, it can be said that MIR-OFP approach is crucially dependent on pollution episodes (Carter and Seinfeld, 2012).
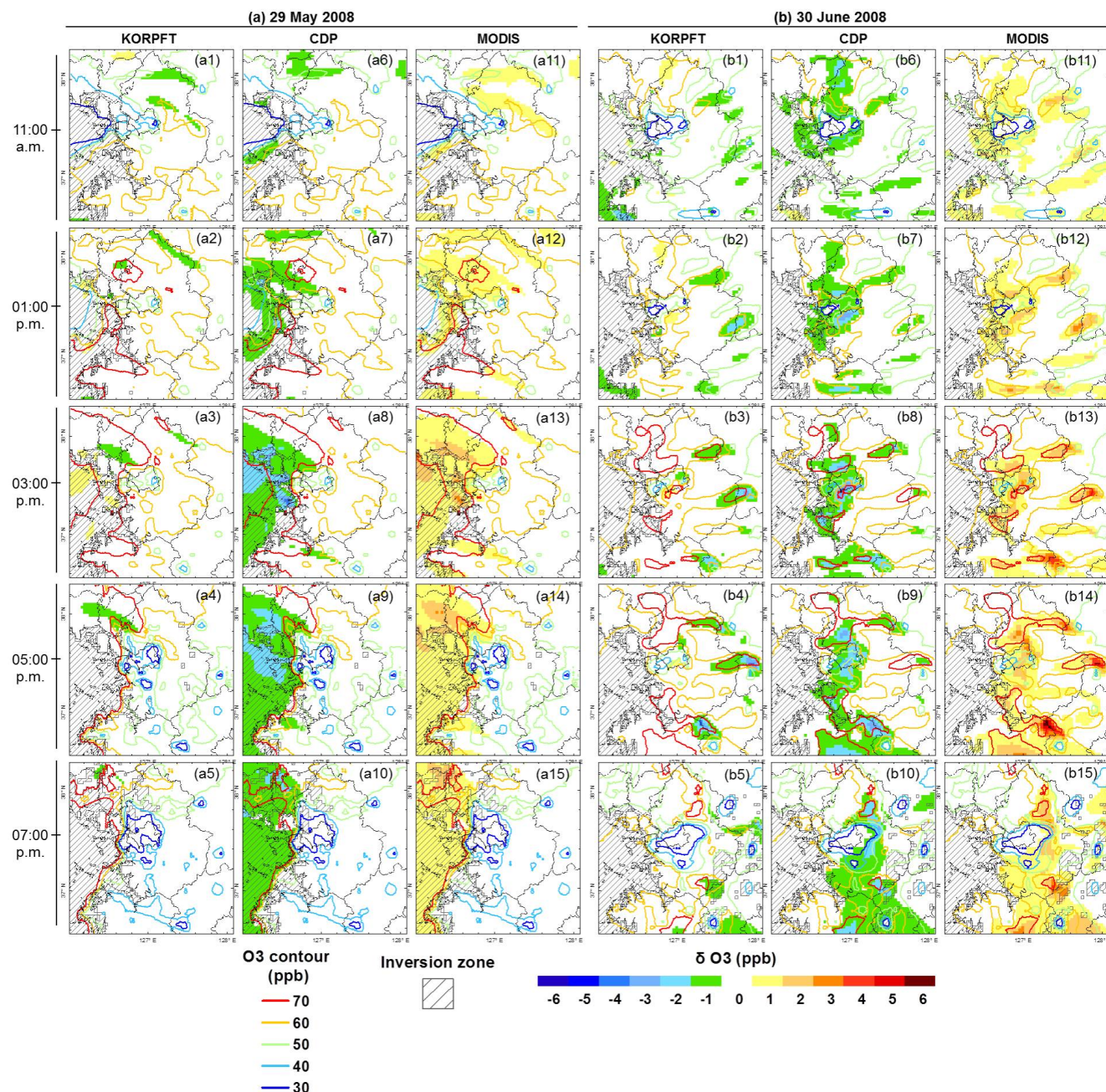
Figure 9. Spatial distributions of CMAQ O_3 hourly deviations for each BVOC emission scenario and CMAQ mean O_3 hourly concentrations. The CMAQ mean O_3 hourly concentrations are the averaged values of CMAQ O_3 concentrations for the three PFT scenarios. These CMAQ O_3 concentrations are illustrated by the contour. The grey-hatched lines represent the temperature inversion zone.

The earlier discussions regarding MIR-OFP and the results in this section could have important implications for designing and implementing biogenic emission estimation and air quality management strategies in this region (the Seoul, Gyeonggi, and Incheon Metropolitan Areas). For example, as described in the literature (Zheng et al., 2009; Carter and Seinfeld, 2012), the distributions of MIRs-OFPs for BVOC compounds can be affected more dynamically by meteoro-

logical factors (e.g., wind direction, temperature, light intensity, etc.) and the locations and magnitudes of source emissions in this region. This discussion points out that developing region-specific reactivity scales for BVOC can be an important tool to better characterize the impact of ozone precursor emissions on regional ozone formation mechanisms and can support the O_3 air quality management strategies in this region. In this sense, an accurate BVOC inventory based

on accurate PFT distribution information would be an essential prerequisite to yield appropriate MIR-OFP information in this region.

3.7 Diurnal variation in the maximum difference between the simulated O_3

To investigate the more accurate impact of the different PFT scenarios on CMAQ O_3 prediction results, we conducted a number of iterative adjustments for various input-variable data sets (e.g., meteorological variables, boundary conditions, anthropogenic NO_x, and isoprene emissions) in CMAQ simulations for June 2008. One of the most important tasks was to fix CMAQ-predicted NO_x concentrations at the level of observed NO_x concentrations. We applied observation data for pollutants and meteorological variables to support these tasks. In this work, we did not consider the spatial variation of the adjustment factors because we assumed that meteorological and chemical variables in the modeling domain are systematically biased. Thus, we applied a common set of temporally varying adjustment factors to every individual modeling grid. Specific adjustment methods are described in Appendix B, and the adjustments were conducted in the following four steps:

1. Using averaged temperature and wind speed data from 11 meteorological monitoring stations (MET in Fig. 1b) maintained by the Korea Meteorological Administration, we corrected meteorological inputs for MEGAN biogenic emission and CMAQ modeling (i.e., temperature and wind speed). Every missing value in observation data was filled by data interpolation.

2. Boundary conditions (BCON) of the simulation domain 3 were adjusted by the hourly varying ratios of observations and corresponding CMAQ concentrations for isoprene, NO_x, and O_3. Here, the CMAQ concentrations represent the averaged values of the CMAQ simulations with the different PFT scenarios. This adjustment was conducted once. Using the adjusted BCON, we performed CMAQ simulations. Every missing value in observation data was filled by interpolation.

3. Using the output from Step 2, we adjusted anthropogenic emission species (isoprene and NO_x) by applying the hourly varying ratios of observations and corresponding CMAQ concentrations for isoprene and NO_x. After emission adjustment, we conducted CMAQ simulation. These emission adjustments and CMAQ simulations were conducted iteratively until the specified criterion was satisfied. Although the optimal criterion was NMB $= 0\%$, we set a relatively moderate criterion because of the limitation of model computing time and storage.

4. When the prespecified convergence criterion in Step 3 was satisfied, the anthropogenic isoprene emissions were adjusted by use of a single ratio of the mean observed isoprene concentration and corresponding CMAQ mean isoprene concentration to make the NO_x / isoprene ratio derived from CMAQ predictions closer to that from observations.

The result of Step 3 is shown in Fig. S3 in the Supplement. After five iterative simulations applying hourly varying adjustment factors for emissions, the specified criterion was satisfied by revealing that NMB of CMAQ NO_x is approximately -7.4%. Furthermore, the corresponding MB (mean bias in Eq. 5) of O_3 for each PFT scenario and the mean CMAQ NO_x / isoprene ratio are shown in Fig. S4a in the Supplement.

The result of Step 4 is shown in Fig. S4 in the Supplement. Through four iterative simulations with averaged adjustment factors for anthropogenic isoprene emissions, we explored the variation of the MB of CMAQ O_3 and the CMAQ NO_x / isoprene ratio (Fig. S4b–d and f in the Supplement). Among four different simulations (Fig. S4b–d and f), we chose the Fig. S4f simulation case for the remainder of the investigation of the impact of the different PFT scenarios on CMAQ O_3 predictions because the CMAQ NO_x / isoprene ratio from the Fig. S4f case ($= 101.29$) is closer to the observed NO_x / isoprene ratio ($= 101.28$) compared with other cases.

After the CMAQ-predicted NO_x values and NO_x / isoprene ratio were adjusted by following the procedure described above, we investigated the impacts of the different PFT scenarios on CMAQ O_3 prediction results. As a result, diurnal distributions of simulated hourly maximum O_3 differences between different PFT scenarios in subregions are shown in Fig. 10. This figure enables the characterization of the periodic impact patterns in everyday CMAQ predictions with different PFT scenarios in the study domain.

In the subregions, the simulated maximum O_3 differences between different PFT scenarios have an asymmetric diurnal distribution pattern (skewed to the left: lower tail is longer than the upper tail). The maximum O_3 difference values in the range (of the whiskers) are low in the early morning (06:00 a.m. to 09:00 a.m.), rise gradually (reaching up to 2 ppb in the Choongcheong regions) during the day with peaks at 05:00 p.m. and then gradually decrease. Although the extreme values (outside the range of the whiskers) of the simulated maximum O_3 differences similarly show an asymmetric diurnal distribution pattern, these distributions have relatively steeper variations, with peaks at 05:00 p.m. at approximately 13 ppb in the Seoul, Incheon, and Gyeonggi regions.

The resultant high differences in the CMAQ 1 h O_3 predictions with the different PFT scenarios could have important implications for air quality decisions and human health studies. For example, the Korea O_3 alert system provides a warning at 120 ppb and an alarm at 300 ppb for 1 h O_3, and the
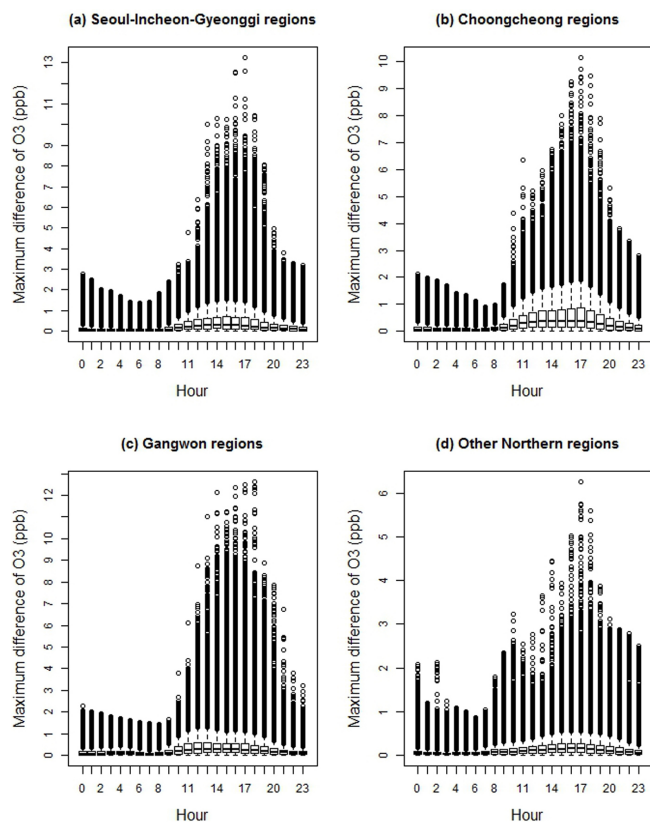
Figure 10. Diurnal distributions of the maximum O_3 differences between different PFT scenarios in subregions derived from CMAQ simulations. The CMAQ simulation was conducted after adjustments of anthropogenic isoprene emission input under fixed CMAQ NO_x at the level of observed NO_x concentrations. Isoprene emission input adjustment was conducted to fix the CMAQ predicted NO_x/isoprene ratio at the level of observed the NO_x/isoprene ratio (see Fig. S4 in the Supplement). For each hour of the day, the boxes represent the range between the 25th and 75th percentile of all maximum difference values for that hour across the simulation period (1–30 June 2008) in the subregions. The line components in boxes indicate the median values. The whiskers indicate the highest ($1.5\times$ IQR (the interquartile range) of the upper quartile) and lowest ($1.5\times$ IQR of the lower quartile) values and connect the extreme points to the box.

Korea NAAQS (national ambient air quality standard) values for 1 h O_3 and 8 h O_3 are 100 ppb and 60 ppb, respectively. In the short term, the highly biased forecasting (e.g., bias up to 13 ppb) of high-O_3 episodes may result in the incorrect issuance of O_3 alerts. Moreover, such inaccurate O_3 forecasting may result in the suggestion of incorrect regulatory design values to O_3 air quality decision supporting authorities. Ji et al. (2011) reported that emergency hospitalizations for total respiratory disease increased by about 3 % per 10 ppb 24 h O_3 among the elderly. This result suggests a critical point at which chemistry transport modeling with highly biased PFT distribution scenarios would predict highly bi-

ased O_3 concentrations and subsequently provide misleading information for studying the relationship between O_3 air quality and human health outcomes.

Another point of concern is the impact from uncertain meteorological variables, especially temperature. Figure 11 shows the deviation tendencies of the simulated BVOC emissions and O_3 concentrations as a function of temperature for the three different PFT scenarios. KORPFT shows a comparatively gentle declining tendency for both δBVOC and δO_3 whereas the other two PFT scenarios show steeper inclining (MODIS) and declining (CDP) tendencies for both δBVOC and δO_3 as temperature increases. These diverging tendencies can be expected to increase (i.e., positive bias for MODIS and negative bias for CDP) as the weather warms due to the seasonal change from early summer to midsummer. Assuming ambient temperature increases due to climate forcing in the future (IPCC, 2007), the use of different PFT distribution data require much caution in terms of whether uncertainties in O_3 prediction results would be mild or severe.

4 Conclusions

The CMAQ performance check suggests that all of three different PFT data sets are acceptable options for regulatory modeling practices when evaluated as to hourly O_3 concentrations based on the US EPA performance standards.

From the investigation of the δPFT areas and δBVOC emission distributions for each PFT scenario (KORPFT, CDP, and MODIS), similar patterns were clearly shown for the PFT and BVOC emission distributions. Three PFT scenarios commonly showed that broadleaf trees (BTs) were the most significant contributor, followed by needleleaf trees (NTs), shrub (SB), and herbaceous plants (HBs), to the total BVOC emissions. Furthermore, isoprene from BTs and terpene from NTs were recognized as significant primary and secondary BVOC species of interest in terms of potential O_3 level increases in the study domain.

The effect of the observed lack of primary data (i.e., PFT and LAI) on CMAQ 1 h O_3 predictions in many geographic locations was noticeable. The lack of this data can cause spatially mistaken CMAQ O_3 prediction results. Thus, we suggest that this issue of missing primary data should be recognized as a current limitation of MEGAN biogenic emission modeling and chemistry transport model O_3 simulations, and it must be resolved in the near future.

An LCR-GWR analysis with different PFT data (δO_3 vs. δPFTs) suggests that addition of BT, SB, and NT areas can contribute to O_3 increase, whereas addition of an HB area contributes to O_3 decrease in the domain.

An assessment of the prediction accuracy of the LCR-GWR model fits with each different PFT scenario showed that the KORPFT provides the best explanations for the
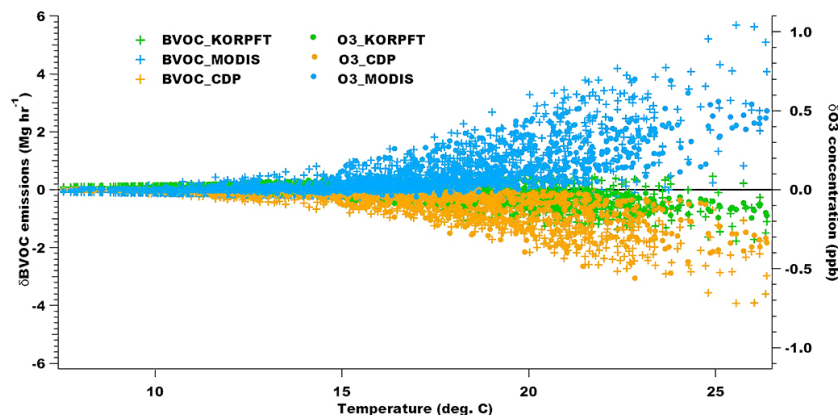
Figure 11. Divergence of the predicted hourly BVOC emissions and O_3 concentrations with temperature change in the modeling domain. Every displayed value of δBVOC and δO_3 in this figure was derived from the hourly domain-averaged values of the MEGAN BVOC emissions and CMAQ O_3 concentrations for the three PFT scenarios (e.g., KORPFT, CDP, and MODIS). The temperature on the horizontal axis in this figure represents averaged values of MM5-predicted ambient temperature.

relationship between PFT, BVOC emissions, and surface-level O_3 changes, followed by MODIS and CDP.

The temporally and spatially averaged effects of the different PFT distributions on CMAQ O_3 simulation results can be regarded as marginal because the usual difference of CMAQ O_3 simulations with different PFT scenarios is less than 0.4 ppb. However, the hourly and local impacts of these are quite noticeable, showing occasional differences of O_3 of up to 13 ppb. The simulated maximum 1 h O_3 differences between different PFT scenarios show an asymmetric diurnal distribution pattern (low by early morning, rising during the daytime, peaking at 05:00 p.m., and decreasing during the nighttime) in the study domain.

Exponentially diverging hourly BVOC emissions and O_3 concentrations were found as a function of temperature change in our modeling domain. Thus, we conclude that the PFT distributions could play the role of a large uncertainty source in hourly O_3 air quality modeling (or forecasting) that supports air quality decision-making and human health studies. The higher the ambient temperature applied to air quality simulation, the larger the likely bias related to PFT distributions.

Appendix A: Specifications of OLS and LCR-GWR

A1 OLS (ordinary least square)

In the OLS approach, the relationship between the dependent and independent (explanatory) variables is modeled as a linear function:

$$y_i = \beta_0 + \sum_{j=1}^{P} \beta_j x_{ij} + \varepsilon_i, \tag{A1}$$

where y_i is the ith value of the dependent variable, x_{ij} is the ith value of the jth independent variable, the ε_is are independent normally distributed error terms with zero means, and β_j must be determined from a sample of n data values. Usually the OLS method is used to estimate the parameters β_js. Using matrix notation this can be stated as

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}^T\mathbf{X}]^{-1}\mathbf{X}^T\boldsymbol{y}, \tag{A2}$$

where $\mathbf{X}$ is an $(n \times (p+1))$ independent variable data matrix, $\boldsymbol{y}$ is an $(n \times 1)$ dependent variable data vector, and $\hat{\boldsymbol{\beta}}$ is the column vector including the coefficient estimates.

A2 LCR-GWR geographically weighted regression with locally compensated ridge

A2.1 GWR (geographically weighted regression)

Brunsdon et al. (1996, 1998) developed the GWR (geographically weighted regression) technique by extending the OLS model (Eq. A1). The GWR approach is preferred when relationships between the dependent and explanatory variables are expected to vary across the geographical domain. The basic form of the GWR model is

$$y_i = \beta_0(u_i, v_i) + \sum_{j=1}^{P} \beta_j(u_i, v_i)x_{ij} + \varepsilon_i, \tag{A3}$$

where (u_i, v_i) is the spatial location of the ith observation, $\beta_0(u_i, v_i)$ is the constant at location i, $\beta_j(u_i, v_i)$ is the local regression coefficient for the jth explanatory variable at location i, P is the number of explanatory variables, and the ε_is are independently normally distributed random error terms with zero mean and σ^2 as common variance at location i.

A local OLS can be calibrated at any location i with data close to location i given greater influence than data farther away by spatial weighting observations according to some kernel functions. The parameter estimation can be formulated as

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = [\mathbf{X}^T\mathbf{W}(u_i, v_i)\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{W}(u_i, v_i)\boldsymbol{y}, \tag{A4}$$

where $\mathbf{W}(u_i, v_i)$ is an $(n \times n)$ spatial weighting diagonal matrix, $\mathbf{X}$ is an $(n \times (p+1))$ explanatory variable data matrix, and $\boldsymbol{y}$ is an $(n \times 1)$ dependent variable data vector.

A2.2 Bandwidth size

An optimum bandwidth can be determined by minimizing some model goodness-of-fit diagnostics (Loader, 1999), such as the cross-validation (CV) score (Cleveland, 1979; Bowman, 1984) or a bias-corrected Akaike information criterion (AICc) statistic (Akaike, 1973; Hurvich et al., 1998). Detailed information about these two methods can be found in Brunsdon et al. (1998, 2000) and Fotheringham et al. (2002). In this study, we calibrate GWR models by the CV approach.

A2.3 Spatial weight

The spatial weighting scheme $\mathbf{W}(u_i, v_i)$ is estimated with a chosen kernel function based on the proximities between regression point i and the N data point around it. A number of kernel functions are available for spatial weighting. In this study, we specified a Gaussian kernel. Its basic continuous form is

$$w_{ij} = \exp\left(-(d_{ij}^2/h^2)/2\right), \tag{A5}$$

where w_{ij} is the jth element of the diagonal of the matrix $\mathbf{W}(u_i, v_i)$, d_{ij} is the distance between observation point j and regression point i. The Euclidean distance was employed for d_{ij} with planar coordinates. h is the kernel bandwidth and controls the degree of distance decay. As the bandwidth gets larger the weights approach unity and the local GWR approaches the OLS.

A2.4 Local ridge regression

The significant collinearity amongst the explanatory variables of a regression model can lead to a loss of precision and power in the parameter estimates. The measure of collinearity includes matrix condition numbers and variance inflation factors (VIFs). Generally condition numbers assess collinearity for all explanatory variables together, whereas VIFs consider each explanatory variable in turn. When condition numbers are above 30 and VIFs above 10, collinearity problems are suspected to be the cause (Belsey et al., 1980; O'Brien, 2007).

The ridge approach (Hoerl, 1962; Hoerl and Kennard, 1970) to the collinearity problem modifies Eq. (A4) by adding a displacement to the leading diagonal of the design matrix:

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = [\mathbf{X}^T\mathbf{W}(u_i, v_i)\mathbf{X} + \lambda\mathbf{I}(u_i, v_i)]^{-1}$$
$$\mathbf{X}^T\mathbf{W}(u_i, v_i)\boldsymbol{y}, \tag{A6}$$

where $\lambda\mathbf{I}(u_i, v_i)$ is the locally compensated value of λ at location (u_i, v_i). Here, it is possible to specify λ_i such that the local condition number is below a chosen threshold (e.g., 30). For a square symmetrical matrix $\mathbf{V}$, the condition number is defined by $\kappa(\mathbf{V}) = \varepsilon_l/\varepsilon_m$, where ε_l is the largest eigenvalue of $\mathbf{V}$ and ε_m is the smallest (Gollini et al., 2013; Brunsdon et al., 2012).

A3   Model assessment diagnostics

A bias-corrected Akaike information criterion (AICc) statistic is usually used (Hurvich and Simonoff, 1998). AIC_c can be determined from

$$AIC_c(h) = 2n \log_e(\hat{\sigma}) + n \log_e(2\pi)$$
$$+ n \left\{ \frac{(n + tr(\mathbf{S}))}{(n - 2 - tr(\mathbf{S}))} \right\}, \qquad (A7)$$

where n is the local sample size according to the kernel bandwidth h, $\hat{\sigma}$ is the maximum likelihood estimate of the standard deviation of the error term (i.e., $\hat{\sigma}^2 = RSS/n$, where RSS is the residual sum of squares of the GWR model), and $tr(\mathbf{S})$ denotes the trace of the hat matrix $\mathbf{S}$.

The hat matrix $\mathbf{S}$ is the projection matrix from the observed $\boldsymbol{y}$ to the fitted values, $\hat{\boldsymbol{y}}$ (Hoaglin and Welsch, 1978), where for GWR each row $\boldsymbol{r}_i$ of $\mathbf{S}$ is

$$\boldsymbol{r}_i = \mathbf{X}_i [\mathbf{X}^T \mathbf{W}_i \mathbf{X}]^{-1} \mathbf{X}^T \mathbf{W}_i, \qquad (A8)$$

where $\mathbf{X}_i$ is its ith row of the matrix of explanatory variables $\mathbf{X}$.

Since the size and sign of the AIC_c values are unrelated, a reduction in AIC_c from the counterpart models can be used as a key diagnostic of the model fit.

Appendix B: Adjustments of CMAQ input variable data

B1   Meteorological variables and biogenic emissions correction

$$MET_{adj(i,t)} = MET_{(i,t)} \times \left(1 + \frac{O_{(i,t)} - P_{(i,t)}}{P_{(i,t)}} \right), \qquad (B1)$$

where MET is short for the meteorological variable data, O is observation, P is prediction (here, MM5-MCIP output), "adj" is short for the adjusted data, and i and t indicate meteorological variable and time ($t = 1, 2, 3, ..., n$ hours), respectively.

B2   BCON adjustment

$$BCON_{adj(i,t)} = BCON_{(i,t)} \times \left(1 + \frac{O_{(i,t)} - P_{(i,t)}}{P_{(i,t)}} \right), \qquad (B2)$$

where BCON represents boundary condition, O is observation, P is prediction (here, CMAQ output), and i and t indicate pollutant species and time ($t = 1, 2, 3, ..., n$ hours), respectively.

B3   Anthropogenic emission adjustment

$$E_{adj(i,t)} = E_{(i,t)} \times \left(1 + \frac{O_{(i,t)} - P_{(i,t)}}{P_{(i,t)}} \right), \qquad (B3)$$

where E is emission (here, SMOKE output), O is observation, P is prediction (here, CMAQ output), i and t are pollutant species and time ($t = 1, 2, 3, ..., n$ hours), respectively.

Criterion: $-10\% < NMB < +10\%$,

where NMB represents the normalized mean bias (see Eq. 6).

B4 NO_x/isoprene ratio adjustment by changing isoprene emission only

$$E_{adj(ISOP)} = E_{(ISOP)} \times \left(1 + \frac{O_{(ISOP)} - P_{(ISOP)}}{P_{(ISOP)}} \right), \qquad (B4)$$

where E_{ISOP} is isoprene emission (here, SMOKE output for isoprene), $O_{(ISOP)}$ is isoprene observation, P_{ISOP} is isoprene prediction (here, CMAQ output for isoprene).

Criterion: $P_{(NO_x)}/P_{(ISOP)} \cong O_{(NO_x)}/O_{(ISOP)}$.

The Supplement related to this article is available online at doi:10.5194/acp-14-7461-2014-supplement.

References

Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: International symposium on information theory, 2nd, Tsahkadsor, Armenian SS, Budapest: Akademiai Kiado, 267–281, 1973.

Arneth, A., Schurgers, G., Lathiere, J., Duhl, T., Beerling, D. J., Hewitt, C. N., Martin, M., and Guenther, A.: Global terrestrial isoprene emission models: sensitivity to variability in climate and vegetation, Atmos. Chem. Phys., 11, 8037–8052, doi:10.5194/acp-11-8037-2011, 2011.

Atkinson, R. and Arey, J.: Gas-phase tropospheric chemistry of biogenic volatile organic compounds: a review, Atmos. Environ., 37 (Suppl. 2), The 1997 Southern California Ozone Study (SCOS97-NARSTO), dedicated to the Memory of Dr. Glen Cass (1947–2001), S197–S219, 2003.

Belsley, D. A., Kuh, E., and Welsch, R. E.: Regression diagnostics: Identifying influential data and sources of collinearity, John Wiley & Sons, Inc., 1980.

Binkowski, F. S. and Roselle, S. J.: Models-3 Community Multi-scale Air Quality (CMAQ) model aerosol components: 1. model description, J. Geophys. Res., 108, 4183, doi:10.1029/2001JD001409, 2003.

Bonan, G. B., Levis, S., Kergoat, L., and Oleson, K. W.: Landscapes as patches of plant functional types: An integrating concept for climate and ecosystem models, Global Biogeochem. Cy., 16, 1021, doi:10.1029/2000GB001360, 2002.

Bowman, A.: An Alternative Method of cross-validation for the smoothing of density estimates, Biometrika, 71, 353–360, 1984.

Brunsdon, C., Fotheringham, A. S., and Charlton, M.: Geographically Weighted Regression:A Method for Exploring Spatial Nonstationarity, Geogr. Anal., 28, 281–289, 1996.

Brunsdon, C., Fotheringham, A. S., and Charlton, M.: Geographically weighted regression-Modeling spatial non-stationarity, The Statistician, 47, 431–443, 1998.

Brunsdon, C., Charlton, M., and Harris, P.: Living with Collinearity in Local Regression Models, in: Proceedings of the 10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Brasil, 2012.

Byun, D. W. and Ching, J. K. S.: Science algorithms of the EPA models-3 Community Multiscale Air Quality (CMAQ) modeling system, EPA/600/R-99/030, US EPA, Research Triangle Park, USA, 1999.

Byun, D. W. and Schere, K. L.: Review of the Governing Equations, Computational Algorithms, and Other Components of the Models-3 Community Multiscale Air Quality (CMAQ) Modeling System, Appl. Mech. Rev., 59, 51–77, 2006.

CARB, Proposed Regulations for Low-Emission Vehicles and Clean Fuels – Staff Report and Technical Support Document, California Air Resources Board, Sacramento, CA, 1993.

Carter, W. P. L.: Implementation of the SAPRC-99 chemical mechanism into the models-3 framework, Report to the United States Environmental Protection Agency, 2000a.

Carter, W. P. L.: Documentation of the SAPRC-99 Chemical Mechanism for VOC Reactivity Assessment," Report to the California Air Resources Board, Contracts 92-329 and 95-308, 8 May, available at: http://cert.ucr.edu/~carter/absts.htm#saprc99 (last access: 21 May 2013), 2000b.

Carter, W. P. L. and Seinfeld, J. H.: Winter ozone formation and VOC incremental reactivities in the Upper Green River Basin of Wyoming, Atmos. Environ., 50, 255–266, 2012.

Cleveland, W. S.: Robust locally weighted regression and smoothing scatterplots, J. Am. Stat. Assoc., 74, 829–836, 1979.

CMAQ v4.6 Operational Guidance Document, http://www.ie.unc.edu/cempd/products/cmaq/op_guidance_4.6/html (last access: 22 August 2006), 2006.

Colella, P. and Woodward, P. L.: The Piecewise Parabolic Method (PPM) for gas dynamical simulation, J. Comput. Phys., 54, 174–201, 1984.

Finlayson-Pitts, B. J. and Pitts Jr., J. N.: Chemistry of the upper and lower atmosphere, Academic Press, San Diego, CA, 2000.

Fotheringham, A. S., Brunsdon, C., and Charlton, M.: Geographically Weighted Regression: the Analysis of Spatially Varying Relationships, Wiley, Chichester, 2002.

Friedl, M. A., McIver, D. K., Hodges, J. C. F., Zhang, X. Y., Muchoney, D., Strahler, A. H., Woodcock, C. E., Gopal, S., Schneider, A., Cooper, A., Baccini, A., Gao, F., and Schaaf, C.: Global Land Cover Mapping from MODIS: Algorithms and Early Results, Remote Sens. Environ., 83, 287–302, 2002.

Gollini, I., Lu, B., Charlton, M., Brunsdon, C., and Harris, P.: GWmodel: an R Package for Exploring Spatial Heterogeneity using Geographically Weighted Models. In arXiv:1306.0413 [stat.AP], available at: http://arxiv.org/pdf/1306.0413v2.pdf (last access: 13 January 2014), 2013.

Grell, G. A., Dudhia, J., and Stauffer, D. R.: A description of the fifth-generation Penn State/NCAR mesoscale model (MM5), NCAR Technical Note, NCAR/TN-398+STR, 117 pp., 1994.

Guenther, A. and Sakulyanontvittaya, T.: Improved Biogenic Emission Inventories across the West, Technical Analysis Report for Western Governors' Association, Project Number: 06-27369, Alex Guenther, Colorado and ENVIRON International Corporation, California, available at: http://www.wrapair2.org/pdf/WGA_Task1_TechnicalAnalysisReport_

ImprovedBiogenicEmissionInventories.pdf (last access: 5 February 2012), 2011.

Guenther, A., Hewitt, C. N., Erickson, D., Fall, R., Geron, C., Graedel, T., Harley, P., Klinger, L., Lerdau, M., Mckay, W. A., Pierce, T., Scholes, B., Steinbrecher, R., Tallamraju, R., Taylor, J., and Zimmerman, P. A.: Global-Model of Natural Volatile Organic-Compound Emissions, J. Geophys. Res.-Atmos., 100, 8873–8892, 1995.

Guenther, A., Karl, T., Harley, P., Wiedinmyer, C., Palmer, P. I., and Geron, C.: Estimates of global terrestrial isoprene emissions using MEGAN (Model of Emissions of Gases and Aerosols from Nature), Atmos. Chem. Phys., 6, 3181–3210, doi:10.5194/acp-6-3181-2006, 2006.

Hoaglin, D. C. and Welsch, R. E.: The hat matrix in regression and ANOVA, The American Statistician, 32, 17–22, 1978.

Hoerl, A. E.: Application of ridge analysis to regression problems, Chemical Engineering Progress, 58, 54–59, 1962.

Hoerl, A. E. and Kennard, R. W.: Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics, 12, 55–67, 1970.

Hogrefe, C., Isukapalli, S., Tang, X., Georgopoulos, P., He, S., Zalewsky, E., Hao, W., Ku, J., Key, T., and Sistla, G.: Impact of biogenic emission uncertainties on the simulated response of ozone and fine Particulate Matter to anthropogenic emission reductions, J. Air Waste Manage., 61, 92–108, 2011.

Hong, S.-Y. and Pan, H.-L.: Nonlocal boundary layer vertical diffusion in a medium-range forecast model, Mon. Weather Rev., 124, 2322–2339, 1996.

Hurvich, C. M., Simonoff, J. S., and Tsai, C. L.: Smoothing parameter selection in non-parametric regression using an improved Akaike information criterion, Journal of Real State Society, Series B (Statistic Methodology), 60, 271–293, 1998.

IPCC Climate Change 2007: The Physical Science Basis, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M., and Miller, H. L., Cambridge University Press, Cambridge, UK and New York, NY, USA, 996 pp., 2007.

Ji, M., Cohan, D. S., and Bell, M. L.: Meta-analysis of the association between short-term exposure to ambient ozone and respiratory hospital admissions, Environ. Res. Lett., 6, 024006, doi:10.1088/1748-9326/6/2/024006, 2011.

Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S. K., Hnilo, J. J., Fiorino, M., and Potter, G. L.: NCEP-DOE AMIP-II reanalysis (R-2), B. Am. Meteor. Soc., 83, 1631–1643, 2002.

Loader, C. R.: Bandwidth selection: classical or plug-in?, The Annals of Statistics, 27, 415–438, 1999.

Lu, B., Harris, P., Charlton, M., Brunsdon, C., Nakaya, T., Gollini, I., and Lu, M. B.: Package "GWmodel" Version 1.2-2, available at: http://cran.r-project.org/src/contrib/Archive/GWmodel/ (last access: 13 January 2014), 2013.

O'Brien, R.: A caution regarding rules of thumb for variance inflation factors, Qual. Quant., 41, 673–690, 2007.

Pfister, G. G., Emmons, L. K., Hess, P. G., Lamarque, J. F., Orlando, J. J., Walters, S., Guenther, A., Palmer, P. I., and Lawrence, P. J.: Contribution of isoprene to chemical budgets: A model tracer study with the NCAR CTM MOZART-4, J. Geophys. Res., 113, D05308, doi:10.1029/2007JD008948, 2008.

Pleim, J. E. and Chang, J. S.: A non-local closure model in the convective boundary layer, Atmos. Environ. A-Gen., 26, 965–981, 1992.

Reisner, J., Rasmussen, R. J., and Bruintjes, R. T.: Explicit forecasting of supercooled liquid water in winter storms using MM5 mesoscale model, Q. J. Roy. Meteor. Soc., 124B, 1071–1107, 1998.

Russell, A. G. and Dennis, R.: NARSTO critical review of photochemical models and modeling. Atmos. Environ., 34, 2284–2324, 2000.

Smith, T. M., Shugart, H. H., and Woodward, F. I. (Eds.): Plant functional types: Their relevance to ecosystem properties and global change, International geosphere-biosphere programme book series 1, Cambridge University Press, Cambridge, 1997.

Strahler, A., Muchoney, D., Borak, J., Friedl, M. A., Gopal, S., Lambin, E., and Moody, A.: MODIS land cover product algorithm theoretical basis document (ATBD) (Version 5.0), available at: http://modis.gsfc.nasa.gov/data/atbd/atbd_mod12.pdf (last access: 3 July 2009), 1999.

Streets, D. G., Bond, T. C., Carmichael, G. R., Fernandes, S. D., Fu, Q., He, D., Klimont, Z., Nelson, S. M., Tsai, N. Y., Wang, M. Q., Woo, J.-H., and Yarber, K. F.: An inventory of gaseous and primary aerosol emissions in Asia in the year 2000, J. Geophys. Res., 108, 8809, doi:10.1029/2002JD003093, 2003.

Sun, W., Liang, S., Xu, G., Fang, H., and Dickinson, R.: Mapping plant functional types from MODIS data using multisource evidential reasoning, Remote Sens. Environ., 112, 1010–1024, 2008.

Tao, Z., Larson, S. M., Wuebbles, D. J., Williams, A., and Caughey, M.: A Summer Simulation of Biogenic Contributions to Ground-Level Ozone over the Continental United States. J. Geophys. Res., 108, 4404, doi:10.1029/2002JD002945, 2003.

Woo, J.-H., Choi, K.-C., Kim, H.-K., Baek, B. H., Jang, M. D., Eum, J.-H., Song, C. H., Ma, Y.-I., Young Sunwoo, Y., Chang, L.-S., and Yoo, Y. H.: Development of an anthropogenic emissions processing system for Asia using SMOKE, Atmos. Environ., 58, 5–13, 2012.

Xie, X., Shao, M., Liu, Y., Lu, S. H., Chang, C. C., and Chen, Z. M.: Estimate of initial isoprene contribution to ozone formation potential in Beijing, China, Atmos. Environ., 42, 6000–6010, 2008.

Zhang, Q., Streets, D. G., Carmichael, G. R., He, K. B., Huo, H., Kannari, A., Klimont, Z., Park, I. S., Reddy, S., Fu, J. S., Chen, D., Duan, L., Lei, Y., Wang, L. T., and Yao, Z. L.: Asian emissions in 2006 for the NASA INTEX-B mission, Atmos. Chem. Phys., 9, 5131–5153, doi:10.5194/acp-9-5131-2009, 2009.

Zheng, J. Y., Shao, M., Che, W. W., Zhang, L. J., Zhong, L. J., Zhang, Y. H., and Streets, D. G.: Speciated VOC emission inventory and spatial patterns of ozone formation potential in the Pearl River Delta, China, Environ. Sci. Technol., 43, 8580–8586, 2009.