Annual sums of carbon dioxide exchange over a heterogeneous urban landscape through machine learning based gap-filling

Olaf Menzer, Wendy Meiring, Phaedon C. Kyriakidis, Joseph P. McFadden
Department of Geography, University of California, Santa Barbara, CA 93106-4060, USA
Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106-3110, USA

HIGHLIGHTS
- Estimates of three years of carbon dioxide exchange in an urban environment.
- Gap-filling models explained between 64% and 88% of the variability in the data.
- Traffic is an important driving variable for gap-filling models at a suburban site.
- Modeled annual carbon budgets varied by a factor of two between wind sectors.
- Machine learning based gap-filling can be applied at other heterogeneous sites.

ABSTRACT
A small, but growing, number of flux towers in urban environments measure surface-atmospheric exchanges of carbon dioxide by the eddy covariance method. As in all eddy covariance studies, obtaining annual sums of urban CO2 exchange requires imputation of data gaps due to low turbulence and non-stationary conditions, adverse weather, and instrument failures. Gap-filling approaches that are widely used for measurements from towers in natural vegetation are based on light and temperature response models. However, they do not account for key features of the urban environment including tower footprint heterogeneity and localized CO2 sources. Here, we present a novel gap-filling modeling framework that uses machine learning to select explanatory variables, such as continuous traffic counts and temporal variables, and then constrains models separately for spatially classified subsets of the data. We applied the modeling framework to a three year time series of measurements from a tall broadcast tower in a suburban neighborhood of Minneapolis-Saint Paul, Minnesota, USA. The gap-filling performance was similar to that reported for natural measurement sites, explaining 64% to 88% of the variability in the fluxes. Simulated carbon budgets were in good agreement with an ecophysiological bottom-up study at the same site. Total annual carbon dioxide flux sums for the tower site ranged from 1064 to 1382 g C m⁻² yr⁻¹, across different years and different gap-filling methods. Bias errors of annual sums resulting from gap-filling did not exceed 18 g C m⁻² yr⁻¹ and random uncertainties did not exceed ±44 g C m⁻² yr⁻¹ (or ±3.8% of the annual flux). Regardless of the gap-filling method used, the year-to-year differences in carbon exchange at this site were small. In contrast, the modeled annual sums of CO2 exchange differed by a factor of two depending on wind direction. This indicated that the modeled time series captured the spatial variability in both the biogenic and anthropogenic CO2 sources and sinks in a reproducible way. The gap-filling approach developed here may also be useful for inhomogeneous sites other than urban areas, such as logged forests or ecosystems under disturbance from fire or pests.

1. Introduction

Urban areas account for at least 70% of anthropogenic carbon dioxide emissions (Canadell et al., 2009), mainly due to the burning of fossil fuels for transportation, space heating, industrial uses, and electric power generation (which may take place outside cities, Satterthwaite (2008)). This will become only more important as an
increasing proportion of the world’s population lives in cities (Seto et al., 2012). Cities provide opportunities for emission reduction through urban design choices. Despite the dominant role of anthropogenic emissions in the urban environment, vegetated areas within cities have been shown to influence the net CO₂ emissions. Cities with a higher percentage of vegetation cover tend to have lower net CO₂ exchange per unit area (Velasco and Roth, 2010). CO₂ uptake by vegetation reduces daytime emissions, but does not transform sites that are net carbon sources into sinks (Grimmond et al., 2002). However, urban vegetation and soils have a significant effect on the seasonality of CO₂ exchange in temperate cities (Peters and McFadden, 2012). Yet photosynthetic uptake and respiratory release of CO₂ from urban green spaces are seldom included in carbon emission inventories of cities (Nordbo et al., 2012), which calculate carbon budgets by bottom-up modeling. Recently, Christen et al. (2011) modeled ecosystem uptake and emissions of CO₂ by combining chamber measurements of soil and above ground biomass respiration as well as irradiance measurements to infer photosynthetic activity and up-scale it to the neighborhood level using airborne light detection and ranging (LIDAR) and satellite data. Nevertheless, uncertainties remain among the different methods to quantify CO₂ fluxes from urban vegetation, and studies need to be extended to cities in different climate zones and the southern hemisphere (Weissert et al., 2014).

Seasonal and annual sums of CO₂ emissions in cities are important for planning and design at fine spatial scales such as at the neighborhood, block, or building level (Kellett et al., 2013). The eddy covariance (EC) method (Baldocchi, 2008) has been used to validate local scale emission inventories (Christen et al., 2011), study the impact of urbanization on CO₂ fluxes (Ramment and Pardyjak, 2011), and improve the understanding of urban ecosystems (Velasco and Roth, 2010). EC provides a means to directly measure the net CO₂ flux, thereby accounting for both the dominant anthropogenic sources of CO₂ emissions and the contributions of vegetation and soils (Feigenwinter et al., 2012). The limitations of EC measurements include fragmentation of data sets due to system failures (caused by e.g., snow, lightning, or birds) and low turbulence atmospheric conditions (e.g., during night), the latter resulting in rejection of some observations. On an annual basis, these so-called gaps typically account for 20–60% of a flux data set (Moffat et al., 2007). Gap-filling of the flux time series is essential to obtain annual sums of net CO₂ flux which can be used to calculate carbon budgets and to evaluate process-based model predictions. If gaps are filled according to the respective wind direction, spatially unbiased estimates of CO₂ flux that do not depend on the frequency distribution of the wind rose at the site can be provided. This could potentially help to quantify the location bias of individual tower sites. In addition, implementing meaningful gap-filling offers insights into the controls of urban carbon fluxes and a subsequent partitioning of the net flux into different components.

Eddy covariance time series of CO₂ exchange are routinely gap-filled using methods such as look-up tables (e.g., Falge et al. (2001); Reichstein et al. (2005)), nonlinear regressions (e.g., Falge et al. (2001); Hollinger et al. (2004)) and artificial neural networks (ANN, e.g., Papale and Valentini (2003); Braswell et al. (2005); Moffat (2012)). All of these are based on well understood relationships of the fluxes with environmental drivers such as temperature and light. Typically, for sites in natural ecosystems, these gap-filling methods can explain 60–90% of the variability of daytime CO₂ fluxes and 25–66% of the variability of nighttime CO₂ fluxes (Moffat et al., 2007).

In the urban environment, however, gap-filling of EC measurements is more complex compared to natural and managed ecosystems, mainly due to two reasons. First, there are additional explanatory variables related to anthropogenic CO₂ emissions, such as vehicular traffic or the diurnal pattern of building energy use. Second, urban surfaces are more heterogeneous in terms of the spatial pattern of vegetation and localized sources of CO₂ emission such as buildings and motor vehicles (Vesala et al., 2008; Pickett et al., 2008; Kotthaus and Grimmond, 2012). This means that urban CO₂ flux measurements have a higher dependence on the prevailing wind direction (Vesala et al., 2008; Kordowski and Kuttler, 2010; Crawford et al., 2011; Jarvis et al., 2012). The underlying patterns of anthropogenic and ecological sources and sinks of CO₂ differ among urban sites (Velasco and Roth, 2010; Crawford et al., 2011; Nordbo et al., 2012), making it a challenge to identify and rank the importance of explanatory variables needed to model the net CO₂ exchanges in cities.

Most of the urban flux studies in the literature have circumvented the problem of gap-filling by working with diurnal fluxes averaged over monthly time spans (Grimmond et al., 2002; Soegaard and Moller-Jensen, 2003; Christen et al., 2011; Peters and McFadden, 2012), which is sufficient for evaluation of models or for comparison to satellite data with a coarser temporal resolution. Some studies have taken advantage of site-specific conditions such as the fraction of vegetation in the footprint of a tower in Vancouver, Canada (Crawford et al., 2011), and Crawford and Christen (2014) modeled the flux footprint of a tower in Essen, Germany. Others have considered systematic and random uncertainties. Some studies have taken advantage of site-specific conditions such as the fraction of vegetation in the footprint of a tower in Vancouver, Canada (Crawford et al., 2011), and Crawford and Christen (2014) modeled the flux footprint of a tower in Essen, Germany. Others have considered systematic and random uncertainties. Some studies have taken advantage of site-specific conditions such as the fraction of vegetation in the footprint of a tower in Essen, Germany. Others have considered systematic and random uncertainties.
residential land use (single-family, detached housing) to the northwest and northeast, and recreational land use (a golf course) to the south-west as well as a two-lane county road carrying primarily commuting traffic (≈ 10,000 vehicles day\(^{-1}\)) to the north and east of the tower. The KUOM tower site was representative of an open low-rise local climate zone (LCZ 6) in the urban landscape classification system of Stewart and Oke (2012).

### 2.2. Net CO\(_2\) flux measurements, filtering and data processing

Total net CO\(_2\) fluxes (\(F_C\)) over the suburban landscape were measured from June 27, 2006 to June 26, 2009 at 40 m on the tall tower using an eddy covariance system consisting of a 3-D sonic anemometer (CSAT3, Campbell Scientific, Inc., Logan, Utah) and a closed-path infrared gas analyzer (LI-7000, LI-Cor, Lincoln, Nebraska). Further details of the measurement system are described in Peters et al. (2011). Briefly, 30-min turbulent \(F_C\) fluxes were computed using eth-flux (Mauder et al., 2008). Data processing steps, including corrections (Webb et al., 1980; Schontanu et al., 1983; Moore, 1986), are described in detail in Hiller et al. (2011). The CO\(_2\) flux was corrected for changes in storage in the air column below the sensor by calculating the change in the CO\(_2\) concentration measured by the LI-7000 over each 30-min period, which is a reasonable assumption for an urban site (Crawford and Christen, 2014). Positive \(F_C\) values stand for release of CO\(_2\) into the atmosphere, and negative \(F_C\) values stand for removal of CO\(_2\) from the atmosphere (i.e., CO\(_2\) uptake by photosynthesis). Of the 52,608 half hours at 40 m level during the 3-year period, about 32% (or 17,051 values) were missing due to power outages and system maintenance. Another 6854 values (or 13%) with wind directions from the southeast were removed to avoid interference from the tower structure. This affected observations for wind directions between 105\(^\circ\) and 165\(^\circ\), as determined by binning the fluxes into 10\(^\circ\) containers and computing average percent deviations from the internal turbulence criterion (Foken and Wichura, 1996), which peaked at the 120\(^\circ\) bin. The land use associated with those wind directions was also atypical of the predominantly residential area because it included experimental crop plots cultivated by the University of Minnesota.

We applied filtering for quality assurance and quality control following Foken and Wichura (1996) and Mauder and Foken (2011). In this scheme, each half hour is classified into three quality classes according to tests of stationarity and well-developed turbulence. Preliminary modeling work showed that selecting high quality values only (corresponding to a zero flag following Foken and Wichura (1996)) resulted in highest gap-filling performance. These high quality observations were not biased toward a specific season or time of day. By filtering for only high quality data, we removed 11,207 half hours (or 21%). Finally, we calculated an initial \(u^*\) (friction velocity) threshold of 0.2 m s\(^{-1}\) for both day and night, as the point where increases in \(u^*\) had little effect on the measured CO\(_2\) flux. This was needed because very low \(u^*\) values indicate periods of low turbulence and stable atmosphere, which results in an underestimation of instantaneous fluxes by eddy covariance measurements (Goulden et al., 1996; Moncrieff et al., 1996; Auliet et al., 2000). Overall, 2640 observations (or 5%) made at \(u^*\) < 0.2 were removed (threshold value determined following Papale et al. (2006) using the most conservative value). A total of \(N = 14,858\) (or 28%) high quality observations then remained available for modeling.

Along with the \(F_C\) observations, we selected a total of 14 variables as potential explanatory variables for the gap-filling models (Table 1). This selection was the result of a preliminary sensitivity analysis, which confirmed our hypothesis that fluxes in the urban environment depend on factors such as wind direction, season, time of day, and day of week. The average diurnal course of the fluxes is depicted in Fig. 2, showing differences between growing and non-growing season, such as weekday rush hour peak emissions in morning and afternoon of winter months and periods of CO\(_2\) uptake during daytime in summer months. Flux patterns also
with anthropogenic and temporal variables (Table 1). Fluxes are seven meteorological variables within our set of predictors along and temperature do pertain at our urban site, thus we included models. Figs. 1 and 2 illustrate that ecological drivers such as light and correlation to turf tower gross primary production (GPP, see Hiller et al. (2011)) and by comparison to the onset and offset dates of the growing season.

2.3. Data resampling and model setup

Land–atmosphere interactions in urban environments are complex and only partially described in existing process-based models. Figs. 1 and 2 illustrate that ecological drivers such as light and temperature do pertain at our urban site, thus we included seven meteorological variables within our set of predictors along with anthropogenic and temporal variables (Table 1). Fluxes are known to be highly variable depending on wind direction, thus, we designed gap-filling models so that they explicitly account for spatial differences in the tower footprint. Therefore, we trained models independently on data subsets binned into ten 30° wind sectors to account for the complexity of the urban landscape (Fig. 4). We selected 30° bin sizes after a preliminary analysis in which we tested sizes between 10° and 45° and found no further improvement of models with smaller bins. Larger bins provided larger samples of flux observations in each bin, but at the cost of higher spatial heterogeneity. Observations from wind sectors centered at 120° and 150° were not included due to interference from the tall tower structure, resulting in a total of ten 30° bins. All hourly F<sub>c</sub> observations that fell into a given wind sector bin were further divided into an 80% subset for model training and another 20% subset that was used as an independent test data set for model validation (later termed the validation data set). This partitioning of training and test observations was realized by a random draw that was repeated 1000 times (Fig. 4, M = 1000). Next, we repeatedly (1000 times) randomized the set of predictors used in each model to a subset of five variables. This reduced the dimensionality of the training data set based on the notion that many of the predictors in Table 1 were highly collinear. For example, consider the temporal variable hour<sub>N</sub>, which was highly correlated with PPFD and traffic variables such as traf<sub>SN</sub>. A combination of these three variables may not be beneficial to use in a model, although this can also depend on the wind sector for which the model is trained. By repeatedly sampling a random subset of all explanatory variables, we were able to explore the space of explanatory variables, rank models depending on their performance (Fig. 4, Model Selection), and understand better the inherent dependence of fluxes on their controlling factors. The model selection step can be based on models either ranked by R² or by RMSE, however, when comparing both rankings in terms of their Taylor diagrams (Taylor, 2001), we did not find differences between the

### Table 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Name</th>
<th>Filled?</th>
<th>Pre-processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daytime</td>
<td>hour&lt;sub&gt;N&lt;/sub&gt;</td>
<td>No</td>
<td>Cosine transformed time of day variable (indicating night/day).</td>
</tr>
<tr>
<td>Daytime</td>
<td>hour&lt;sub&gt;D&lt;/sub&gt;</td>
<td>No</td>
<td>Sine transformed time of day variable (indicating morning/evening).</td>
</tr>
<tr>
<td>Binary weekend flag</td>
<td>we</td>
<td>No</td>
<td>Binary variable, 1 for Saturdays and Sundays, 0 for weekdays.</td>
</tr>
<tr>
<td>Air temperature</td>
<td>T&lt;sub&gt;air&lt;/sub&gt;</td>
<td>Yes</td>
<td>Air temperature measured by the sonic anemometer at the 40 m level.</td>
</tr>
<tr>
<td>Soil temperature</td>
<td>T&lt;sub&gt;soil&lt;/sub&gt;</td>
<td>Yes</td>
<td>Measured with a REBS model STP-1 platinum resistance temperature at 5 cm below ground at the nearby turfgrass site.</td>
</tr>
<tr>
<td>Heating degree days</td>
<td>hdd</td>
<td>Yes</td>
<td>Gap-filled using local weather station data.</td>
</tr>
<tr>
<td>Photosynthetic photon flux density</td>
<td>PPFD</td>
<td>Yes</td>
<td>Gap-filled using solar radiation data recorded at the same time.</td>
</tr>
<tr>
<td>Golf course enhanced vegetation index</td>
<td>EVI</td>
<td>Yes</td>
<td>Measured with a Li-Cor model LI-190 quantum sensor mounted at 2 m above our turfgrass flux site adjacent to the tall tower.</td>
</tr>
<tr>
<td>Precipitation</td>
<td>P</td>
<td>Yes</td>
<td>Measured with a tipping bucket rain gauge, at the University of Minnesota climate station located &lt;1 km from the KUOM tower.</td>
</tr>
<tr>
<td>Wind speed</td>
<td>u</td>
<td>Yes</td>
<td>Measured at 40 m, gap-filled using data from the turf tower and two other local stations.</td>
</tr>
<tr>
<td>Wind direction</td>
<td>wd</td>
<td>Yes</td>
<td>Gap-filled using circular regression on data from the turf tower and two other local stations.</td>
</tr>
<tr>
<td>Local traffic east—west</td>
<td>traf&lt;sub&gt;EW&lt;/sub&gt;</td>
<td>Yes</td>
<td>Sum of eastbound and westbound traffic at the nearest intersection ~ 70 m away from the tower. Based on 15 min resolution automated traffic recording data. Source: Ramsey County, Minnesota.</td>
</tr>
<tr>
<td>Local traffic south—north</td>
<td>traf&lt;sub&gt;SN&lt;/sub&gt;</td>
<td>Yes</td>
<td>Sum of southbound and northbound traffic at the nearest intersection. Same location, resolution, and source as for traf&lt;sub&gt;EW&lt;/sub&gt;.</td>
</tr>
<tr>
<td>Highway traffic</td>
<td>traf&lt;sub&gt;SN&lt;/sub&gt;</td>
<td>No</td>
<td>Highways traffic counts recorded by Minnesota Department of Transportation at station ATR 389 in the northwest vicinity of the tower. Interpolated from hourly to half-hourly resolution.</td>
</tr>
</tbody>
</table>

Set of 13 potential driving variables plus wind direction (used for wind sector binning only) and descriptions. Third column indicates if gap-filled. Note that wd has only been implicitly used as a driving variable by constraining models separately for each 30° sector. The weekend flag, we, was included in every model together with a random selection of five of the remaining twelve variables. A total of M = 1000 random permutations (Fig. 4) were saved and fixed for all methods to ensure comparability.
models ranked highest by each statistic. Every randomized subset of five variables was then extended to also include a binary weekend indicator flag to discriminate between weekends and weekdays, based on the dependence of urban fluxes on vehicle commuting patterns. Note that through preliminary analyses, in which we trained models with all 13 available driving variables (Table 1) and removed them one by one, we chose to include six variables in our final models because the decrease in performance was significant when models had fewer variables. We caution that this is an empirical result from a single study site and should be examined independently given local conditions. All data in this study were normalized by subtracting the mean and dividing by the standard deviation of the sample.

2.4. Gap-filling techniques

One of our main objectives was to apply gap-filling techniques that had been shown to perform well at natural sites and extend them to be applicable in urban environments. We focused on machine learning methods for nonlinear regression that can extract functional relationships directly by generalizing from sample data, and do not rely on an explicit, process-based formulation. Schmidt et al. (2008) first used artificial neural networks (ANNs) for gap-filling in urban environments. ANNs previously have been widely used for gap-filling of fluxes from forest or grassland sites (Papale and Valentini, 2003; Braswell et al., 2005; Moffat et al., 2007; Moffat, 2012). Appendix A1 provides further ANN details. Here,

![Fig. 2. Mean diurnal courses of half-hourly $F_C$ measurements averaged over two month periods for each year (2006 in red, 2007 in green, 2008 in blue and 2009 in magenta) in the study period. Error bars represent one standard deviation centered around the mean. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)](image)
we used the same parameterization of the ANN network and learning scheme as in Schmidt et al. (2008), which is also comparable to Moffat (2012) and Järvi et al. (2012). The specific ANN is a multilayer perceptron containing one single hidden layer with five neurons, based on the results in Schmidt et al. (2008) and Järvi et al. (2012).

Radial basis function networks (RBFs, Broomhead and Lowe (1988)) are another form of feed-forward neural networks that have been used for gap-filling of flux data in urban environments (Schmidt et al., 2008; Kordowski and Kuttler, 2010). They have an alternative node structure using Gaussian exponential functions as the activation function at each neuron. Here, we used RBFs as another gap-filling method in our modeling framework, but reduced them to a simpler version that used parameter initializations based on the results in Schmidt et al. (2008) instead of optimizing them before network training. Hence, the number of...
neurons for this type of network was set to 75, while the initial spread parameter for the Gaussian exponential functions was set to 1.2. Both ANN and RBF models were implemented through the MATLAB R2013a Neural Network Toolbox.

In addition, we used another machine learning technique, Gaussian Process regression (GP). GPs are a nonparametric framework for regression and classification, applicable to multidimensional, noisy data sets with underlying nonlinear relationships (Rasmussen, 1996; Neal, 1997). They directly relate to ANNs as they can be considered an ANN (or RBF) with infinitely many hidden units (MacKay, 1998). GPs are widely implemented as Kriging in geostatistics, following Matheron (1963), and have more recently been applied for random uncertainty estimation in EC flux time series (Menzer et al., 2013). A GP is a stochastic process that can be defined by a mean function and a covariance function. The key idea of the learning process with GPs is optimizing the hyperparameters in the covariance function given the observations. For a more detailed description of the GP method, also see the appendix (A1). Here, we used the MATLAB based gpml toolbox 3.0 by Rasmussen and Nickisch (2010).

Finally, we evaluated Marginal Distribution Sampling (MDS, Reichstein et al. (2005)), as a reference method because it is widely used in the eddy flux community for gap-filling at natural sites. MDS-based gap-filling uses a combined moving window and mean diurnal variation approach that considers both the covariation of fluxes with meteorological variables and the temporal autocorrelation of the fluxes. MDS uses only the five following meteorological variables: \( T_{an} \), global radiation, relative humidity, vapor pressure deficit, and \( w \) (for filtering), along with information on time of day and day of the year. We performed MDS gap-filling on the high quality data record of \( N = 14,858 \) observations, randomly dividing it into 80% of the data for training and 20% for model validation, and repeating this process ten times, i.e., for ten different splits of the data into training and test set across observations from all wind sectors (MDS currently does not use wind direction information).

2.5. Sensitivity analysis and hypothesis tests

We used estimates of model parameters to compare the importance of our input variables, such as the characteristic length-scale hyperparameters of the GP’s covariance function (Eq. (A2)). In addition, we tested the sensitivity of our models by removing variables from the model. Specifically, we calculated the change in performance of the best model when excluding a specific variable (i.e., setting the variable to its mean and then re-training the model). By doing so, the explainable variability in terms of adjusted \( R^2_{test} \) (referring to performance on the independent test data set, accounting for number of explanatory variables) or \( R^2_{train} \) (training data set) can be compared between a model that contains an explanatory variable of interest compared to a model lacking that same variable. If the performance measure does not change significantly we can conclude that a variable is not important for the regression, whereas a performance drop after the knock out indicates a significant contribution of that variable to the model.

2.6. Carbon dioxide flux sums and uncertainties

To obtain annual \( F_C \) sums we needed to gap-fill a total of 31,289 half hours that were missing due to no measurement being taken or filtered out according to violated model assumptions. We filled short gaps (<2 h) by linear interpolation (3324 half hours), but did not use those for model input. A small set of 552 missing half hours in May and June 2009 was filled using mean diurnal averages of the two months because meteorological data were missing and data from nearby stations were also unavailable. All remaining gaps were filled using an ensemble average of the ten best machine learning model runs for each of the ten wind sectors individually (Fig. 4). For gaps resulting from tower interference when the prevailing wind direction was southeast, we used average predictions of models fitted to observations from each of the other ten sectors. This strategy was chosen as the best compromise to obtain “toward sums”, i.e., sums over the observed footprint within a given time period.

Following Moffat et al. (2007) and Jarvi et al. (2012), we computed the annual bias error \( BE_{ann} \) as

\[
BE_{ann} = \sum_{w=1}^{10} BE_{wd}(w) \cdot N_{gap,w},
\]

where \( N_{gap,w} \) is the number of gaps in the w-th wind sector and \( BE_{wd}(w) \) is the bias error for \( w \) (in \( \mu mol \ CO_2 \ m^{-2} \ s^{-1} \)), computed as the average bias error over the ten best models in a sector, given by

\[
BE_{wd}(w) = 1/10 \sum BE_{mj}(w),
\]

where \( BE_{mj}(w) \) is the bias error for a model \( j \) in the \( w \)-th sector defined as

\[
BE_{mj}(w) = 1/N_{w test} (\sum (y_i - \hat{y}_i)),
\]

where \( y_i \) are the observations (i.e., \( F_C \) in the test data set, \( \hat{y}_i \) the corresponding model predictions (i.e., \( \hat{F}_C \)), and \( N_{w test} \) the number of observations withheld for model testing in wind sector \( w \). Annual sums were converted to \( g \ C m^{-2} yr^{-1} \) by multiplying the half hourly values by 1800 s \( \cdot \) 12 g mol\(^{-1} \).

In addition, we computed monthly and annual CO\(_2\) flux sums for each of the ten different wind sectors. For this, we used input variables observed at the tower (irrespective of wind direction and thus identical input for each wind sector) for the entire time span as an input to the gap-filling models. In this way, time series were computed for each wind sector separately and then labeled by the land use found within that wind sector. Subsequently, those predictions were summed to obtain monthly, seasonal, or annual sums that were spatially variable. For random uncertainty assessment of the data we used bootstrapping (details in the appendix A2), and for assessment of systematic uncertainties we used propagation of the \( \psi \) threshold distribution following Papale et al. (2006). Details are given in the appendix A3.

3. Results

3.1. Gap-filling model performance

The gap-filling models in this study were evaluated by computing statistics at several levels of aggregation. First, we calculated performance statistics by each 30° wind sector \( W = 10 \) wind sectors) for the three machine learning techniques and for each of the \( M = 1000 \) models that were trained (totaling 30,000 performances). We ranked model performance by the \( R^2 \) of predicted values against observed values in the independent 20% of the data set reserved for model testing \( (R^2_{test}) \). The statistics of the best model for each wind sector are summarized in Table 2. \( R^2_{test} \) scores ranged from 0.60 to 0.86 for ANNs (across the 10 wind sectors), from 0.53 to 0.86 for RBFs, and from 0.64 to 0.88 for GPs. The amount of variability explained by the models was dependent on the wind sector, with lower performance \( (R^2_{test} > 0.83) \) found in the south-west sectors \( (w_{210}, w_{260}) \) and the north-west sector \( (w_{330}) \), all associated with recreational land use types (i.e., golf courses). The \( R^2 \) statistic on the training data set \( (R^2_{train}) \) was in a similar range as on the test data set \( (R^2_{test}) \), i.e., between 0.41 and 0.79 for ANNs (along the 10 wind
increasing the training data set size to 20,850), the
and Foken (2011), but also moderate quality (QA/QC scores
sectors obtained from independent models into one data set
by combining the test data predictions from each of the ten wind
sectors), between 0.63 and 0.82 for RBFS, and between 0.51 and
0.83 for GPs. This statistic is useful to detect when models overfit,
as indicated by training performances that are much larger than
respective test performances. In the presented models, $R^2_{\text{train}}$ was
lower than $R^2_{\text{test}}$ in most cases (28 out of 30 paired performances).
Note that the RMSE value should not be compared across wind
sectors due to its dependence on the absolute values of
fluxes, as shown in the
time scales, because it included model outputs from all the spatially
variable wind sectors. In addition, predictions were included from
all of the ten best models for each wind sector, resulting in a total of
29710 data points, on which performance statistics were then
calculated. Results of the different machine learning methods are
shown in Fig. 5(a)–(c), whereas results of MDS (marginal distribution
sampling) algorithm on ten training samples are shown in
Fig. 5(d). The test performance $R^2_{\text{test}}$ on the aggregated data sets was
as high as 0.74 for ANN models, 0.75 for RBF models and 0.78 for GP
models, compared to 0.60 for MDS. Positive bias errors ($BE$) on the
scale of the entire data set in the range of [0.035, 0.059] mol CO₂ m⁻² s⁻¹ indicated overestimation of half-hourly predicted
values, whereas fluxes of large magnitude were underestimated (in absolute terms) by all methods. When subsetting the test data using
a large flux magnitude threshold of $>|5| \text{ mol CO₂ m}^{-2} \text{ s}^{-1}$, ANN
had a $BE$ of $-0.759$, RBF had a $BE$ of $-0.664 \text{ mol CO₂ m}^{-2} \text{ s}^{-1}$ and GP
had a $BE$ of $-0.634 \text{ mol CO₂ m}^{-2} \text{ s}^{-1}$. In comparison, MDS had a
more negative $BE$ value of $-1.048 \text{ mol CO₂ m}^{-2} \text{ s}^{-1}$, when only
considering large magnitude fluxes. In fact, MDS also had a negative
bias error ($-0.030 \text{ mol CO₂ m}^{-2} \text{ s}^{-1}$) on the scale of the entire data set,
mainly resulting from the underestimation of large positive
fluxes.

Differences in $R^2_{\text{test}}$ performance between the three machine
learning methods ANN, RBF, and GP ranged from 1–11% at the wind
sector level (Table 2). Overall, the test data performance was
marginally but consistently higher for the GP method compared to
ANN and RBF models both in terms of $R^2_{\text{test}}$ and RMSE (except in
sectors w60 and w180). When all wind sectors were considered
together (Fig. 5), differences in performance on the test data be-
 tween the three machine learning methods were about 4% in $R^2_{\text{test}}$
and similarly small in terms of RMSE, $BE$, and regression param-
eters. Notably, the regression slopes and offsets for the RBF and GP
models were almost identical. Performances of the three machine
learning methods ANN, RBF, and GP were consistently higher than
for MDS in terms of all computed statistics. This reflects in the
better ability of the machine learning methods to reproduce the
diurnal cycle of fluxes, as shown in the fingerprint plots of pre-
predicted data for the entire 3 year time series (Fig. 6). Predictions by
the gap-filling models using machine learning methods (e.g., ANN,
Fig. 5(b)) followed clearer diurnal and seasonal time courses and
were smoother as compared to the more scattered MDS predictions
(Fig. 5(a)).

---

**Table 2**

<table>
<thead>
<tr>
<th>Sector [-]</th>
<th>$N_{\text{train}}$ [-]</th>
<th>$N_{\text{test}}$ [-]</th>
<th>Method [-]</th>
<th>$R^2_{\text{test}}$ [-]</th>
<th>RMSE [μmol CO₂ m⁻² s⁻¹]</th>
<th>$BE_{\text{MDS}}$ [μmol CO₂ m⁻² s⁻¹]</th>
<th>$R^2_{\text{train}}$ [-]</th>
<th>$R^2_{\text{MDM}}$ [-]</th>
</tr>
</thead>
<tbody>
<tr>
<td>w₀</td>
<td>319</td>
<td>1277</td>
<td>ANN</td>
<td>0.71</td>
<td>2.40</td>
<td>0.172</td>
<td>0.61</td>
<td>0.52</td>
</tr>
<tr>
<td>w₃₀</td>
<td>184</td>
<td>738</td>
<td>GP</td>
<td>0.76</td>
<td>2.22</td>
<td>0.108</td>
<td>0.67</td>
<td>0.51</td>
</tr>
<tr>
<td>w₆₀</td>
<td>129</td>
<td>515</td>
<td>GP</td>
<td>0.64</td>
<td>2.47</td>
<td>0.002</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>w₉₀</td>
<td>147</td>
<td>588</td>
<td>ANN</td>
<td>0.78</td>
<td>2.01</td>
<td>0.149</td>
<td>0.62</td>
<td>0.46</td>
</tr>
<tr>
<td>w₁₂₀</td>
<td>396</td>
<td>1586</td>
<td>RBF</td>
<td>0.79</td>
<td>1.92</td>
<td>0.104</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>w₂₀</td>
<td>287</td>
<td>1146</td>
<td>RBF</td>
<td>0.85</td>
<td>2.87</td>
<td>-0.032</td>
<td>0.73</td>
<td>0.66</td>
</tr>
<tr>
<td>w₂₁₀</td>
<td>140</td>
<td>561</td>
<td>GP</td>
<td>0.86</td>
<td>2.98</td>
<td>-0.305</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>w₂₇₀</td>
<td>342</td>
<td>1368</td>
<td>ANN</td>
<td>0.67</td>
<td>3.64</td>
<td>-0.202</td>
<td>0.54</td>
<td>0.42</td>
</tr>
<tr>
<td>w₃₀</td>
<td>622</td>
<td>2488</td>
<td>RBF</td>
<td>0.68</td>
<td>3.43</td>
<td>-0.158</td>
<td>0.63</td>
<td></td>
</tr>
<tr>
<td>w₃₉₀</td>
<td>405</td>
<td>1620</td>
<td>RBF</td>
<td>0.71</td>
<td>2.98</td>
<td>-0.254</td>
<td>0.58</td>
<td></td>
</tr>
</tbody>
</table>

**Notes:**
- ANN (Artificial Neural Network); RBF (Radial Basis Function Network) and GP (Gaussian Process).
- $N_{\text{train}}$ and $N_{\text{test}}$ indicate the number of half-hours for validation and training, respectively.
- $R^2_{\text{test}}$, RMSE (root mean squared error) and $BE_{\text{MDS}}$ (bias error) are all computed for the model performance of the best model ($j = 1$) on the independent test data set (20% of the data reserved for model testing), while $R^2_{\text{train}}$ is the fit statistic on the training data set. The last column reports the $R^2_{\text{MDM}}$ of the reference method MDS (Marginal Distribution Sampling) on the respective wind sectors, here we show the best $R^2_{\text{MDM}}$ of ten MDS iterations for each wind direction.

---
3.2. Model sensitivity to driving variables

We analyzed the sensitivity of the gap-filling models to their respective driving (or input) variables, primarily by comparing the log hyperparameters of the GP (Gaussian Processes) method, because GPs showed high gap-filling performance throughout. In short, each of the six variables \(d_1\)…\(d_6\) in a given GP gap-filling model was associated with a characteristic length-scale hyperparameter \(s_d\) that indicated the relative importance of that variable for the regression. That is, the shorter the estimated length-scale hyperparameter, the more important that variable was for the model’s ability to capture variability in the data. This allowed us to compare the importance of any variable relative to the other variables within a given GP gap-filling model. We examined the impact of traffic on the CO\(_2\) flux observations using sector \(w_{30}\) (Table A1), which included the main road, and we compared \(w_{30}\) estimated log hyperparameters of different input variables to those from \(w_{240}\), which contained mainly vegetation (Table A2). The variable \(trafSN\) appeared frequently and was highly ranked in models for \(w_{30}\), a wind sector that included the main road. In contrast, \(trafSN\) was lower ranked or was absent in models for \(w_{240}\), a wind sector dominated by the golf course.

The results of this sensitivity analysis were supported by carrying out model variable knockout experiments. For wind sector \(w_{30}\), we manipulated the best performing GP model, which included \(trafSN\) as a predictor, by setting \(trafSN\) to its mean value and re-training the GP. This caused the test data performance, as measured by adjusted \(R^2\) test, to drop from 0.63 to 0.24. Similarly, the test data performance dropped down from 0.63 to 0.25 when knocking out only \(PPFD\). In contrast, when knocking out e.g., the less important wind speed variable \(u\) (ranked 5th in the best model, with a log length-scale parameter \(s_5 = 3.50\), three times as high as for \(trafSN\) and \(PPFD\)), the model performance only dropped from 0.63 to 0.52. In addition, we also found large decreases in adjusted \(R^2\) test after removing \(trafSN\) variables in the adjacent wind sectors \(w_{0}\) and \(w_{60}\) of about 21% and 40%, respectively; and also in one of the north-west sectors by 14% (\(w_{300}\)). To further investigate the impact of traffic on the flux measurements, we also re-trained models on a summer subset of the data (June, July and August), when non-vehicular anthropogenic CO\(_2\) sources (such as space heating) could be expected to be minimal. In this analysis, we consistently found a decrease in adjusted \(R^2\) test performance between 26% and 74% when knocking out traffic variables in models for sectors \(w_{0}\), \(w_{30}\) and \(w_{60}\) (not shown).
3.3. Carbon dioxide flux sums and uncertainties

Because the measurement period at the tower site started in June of 2006 and ended in June 2009, we estimated annual sums for the three one-year periods of July 1 through June 30 of the following year (Table 3). Across all models and years, estimated tower sums of $F_C$ ranged from 1064 to 1382 g C m$^{-2}$ yr$^{-1}$, signifying that the urban ecosystem was a net carbon source on an annual basis (Table 3).

Gap-filled annual $F_C$ sums are estimates and are therefore subject to uncertainty. We found BEann to be 18 g C m$^{-2}$ yr$^{-1}$ (or 1.7%) at maximum, across the three years of study and across all methods (Table 3). Uncertainty resulting from the selection of the $u^*$ threshold was between 3 g C m$^{-2}$ yr$^{-1}$ (or 0.2% of the annual sum) and 64 g C m$^{-2}$ yr$^{-1}$ (or 4.6% of the annual sum), across all methods. However, compared to the other methods, GP had the lowest $u^*$ uncertainties, at most 2% of the annual sum. We also re-trained the GP gap-filling models including the data points that were below the $u^*$ threshold. This was done for the lower and upper bounds as well as the median of the estimated $u^*$ threshold distribution in order to test how sensitive the gap-filling model results were to the $u^*$ threshold setting. Through this approach, we found sums (not shown) to be in a similar range as the ones reported in Table 3 with uncertainties not higher than 24 g C m$^{-2}$ yr$^{-1}$ (or 2.1%). Random uncertainties inferred from bootstrapping were between 18 g C m$^{-2}$ yr$^{-1}$ (or 1.6% of the annual sum) and 44 g C m$^{-2}$ yr$^{-1}$ (or 3.8% of the annual sum) across years and methods (one standard deviation).

We also calculated wind sector sums of $F_C$ by predicting artificial time series with models that had been trained for each wind sector...
separately, by forcing models with the entire set of meteorological and traffic observations made over a given time span. This allowed us to compare \( F_C \) budgets across wind sectors and associate them with underlying land use types (Fig. 7). There was high variability (by a factor of two) of the modeled wind sector \( F_C \) budgets, ranging from 890–1990 g C m\(^{-2}\) yr\(^{-1}\) for ANNs, 670–1820 g C m\(^{-2}\) yr\(^{-1}\) for RBFs, and 670–1780 g C m\(^{-2}\) yr\(^{-1}\) for GPs. The lowest annual \( F_C \) corresponded to wind sectors over the golf course (e.g., \( w_{210} \)) and highest annual \( F_C \) corresponded to residential land use types and nearby roads (e.g., \( w_{30} \)), with the budgets fairly consistent between the years 2007 and 2008. Uncertainties derived from bootstrap simulations (error bars in Fig. 7) also showed variability with wind direction. The uncertainties attached to wind sectors over recreational land use types were consistently lower than those for wind sectors over residential land use and traffic, by a factor of two (between 57 g C m\(^{-2}\) yr\(^{-1}\) and 106 g C m\(^{-2}\) yr\(^{-1}\)).

Modeled \( F_C \) budgets for the month of June 2007 showed net CO\(_2\) uptake in sectors \( w_{210} \) (\(-14.3 \pm 8.0\) g C m\(^{-2}\)), \( w_{240} \) (\(-29.2 \pm 10.1\) g C m\(^{-2}\)) and \( w_{270} \) (\(-27.4 \pm 11.1\) g C m\(^{-2}\)). These monthly sums, representing the total net CO\(_2\) exchange, were expected to be smaller than the value of \(-36.0\) g C m\(^{-2}\) that was reported from a bottom-up estimate of only the vegetation and soil CO\(_2\) exchange of this site (Peters and McFadden, 2012). This is because our gap-filled tall tower sums captured some anthropogenic emission sources, which were not part of the bottom-up estimates of vegetation and soil CO\(_2\) flux by Peters and McFadden (2012). Note that Fig. 7 only shows the budgets from ANN simulations, but annual sums simulated by GP and RBF models also showed very similar patterns.

4. Discussion

4.1. Annual urban CO\(_2\) flux sums and uncertainties

The annual \( F_C \) sums at the KUOM tower fell at the lower end of those reported from other urban sites, which range from 0 to

---

Table 3

<table>
<thead>
<tr>
<th>Year [–]</th>
<th>Method [–]</th>
<th>( F_C ) (±1 SD) [g C m(^{-2}) yr(^{-1})]</th>
<th>( BE_{ann} ) [g C m(^{-2}) yr(^{-1})]</th>
<th>( \mu)-range (5%-95%) [g C m(^{-2}) yr(^{-1})]</th>
<th>( R_{bias}^2 ) [–]</th>
<th>RMSE [g C m(^{-2}) yr(^{-1})]</th>
<th>( R^2_{ann} ) [–]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006–2007</td>
<td>ANN</td>
<td>1341 (±37)</td>
<td>14</td>
<td>(1329, 1342)</td>
<td>0.72</td>
<td>3.27</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>RBF</td>
<td>1064 (±21)</td>
<td>18</td>
<td>(1058, 1064)</td>
<td>0.70</td>
<td>3.23</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>GP</td>
<td>1147 (±44)</td>
<td>16</td>
<td>(1150, 1147)</td>
<td>0.78</td>
<td>3.06</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>MDS</td>
<td>1396</td>
<td>–</td>
<td>(1375, 1397)</td>
<td>0.55</td>
<td>4.29</td>
<td>0.61</td>
</tr>
<tr>
<td>2007–2008</td>
<td>ANN</td>
<td>1382 (±35)</td>
<td>9</td>
<td>(1321, 1384)</td>
<td>0.74</td>
<td>3.16</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>RBF</td>
<td>1177 (±19)</td>
<td>12</td>
<td>(1138, 1179)</td>
<td>0.75</td>
<td>3.13</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>GP</td>
<td>1172 (±30)</td>
<td>10</td>
<td>(1153, 1172)</td>
<td>0.77</td>
<td>2.97</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>MDS</td>
<td>1372</td>
<td>–</td>
<td>(1303, 1375)</td>
<td>0.62</td>
<td>3.74</td>
<td>0.66</td>
</tr>
<tr>
<td>2008–2009</td>
<td>ANN</td>
<td>1366 (±26)</td>
<td>13</td>
<td>(1317, 1368)</td>
<td>0.75</td>
<td>2.87</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>RBF</td>
<td>1130 (±19)</td>
<td>16</td>
<td>(1091, 1113)</td>
<td>0.76</td>
<td>2.83</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>GP</td>
<td>1148 (±25)</td>
<td>15</td>
<td>(1126, 1149)</td>
<td>0.80</td>
<td>2.62</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>MDS</td>
<td>1446</td>
<td>–</td>
<td>(1384, 1448)</td>
<td>0.65</td>
<td>3.44</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Annual, gap-filled \( F_C \) sums and uncertainties, computed by different gap-filling methods for the three annual periods of July 1 of one year through June 30 of the following year. One standard deviation (SD) of the bootstrapping distribution of sums is given in parentheses behind the \( F_C \) sum, the annual bias error \( BE_{ann} \) resulting from gap-filling the time series was computed as in Eq. (1), systematic uncertainties from \( \mu\)-filtering are given as a range at the 5% (lower quantile of distribution of \( \mu\)-values) and 95% (upper quantile of distribution of \( \mu\)-values) confidence level. Also given are the performance statistics of the methods for each year, however, note that gap-filling models were trained on a high quality data set across all years and not by year. All values are in \([\text{g C m}^{-2}\text{ yr}^{-1}]\).
10,000 g C m$^{-2}$ yr$^{-1}$ (Nordbo et al., 2012). This was reasonable given the KUOM location in a suburban neighborhood, whereas the other urban sites were situated in more densely built and populated environments. The vegetation cover at our site was 82% (Peters et al., 2011), which is higher than the other urban flux tower sites summarized in Velasco and Roth (2010) and Nordbo et al. (2012). The amount of traffic in the area was relatively low, compared to more central urban sites in the literature, which was reasonable given that the primary intent of the KUOM flux study was to quantify the contribution of vegetation and soils to the net CO$_2$ exchange. In contrast, at an urban tower site in Vancouver (Christen et al., 2011) as many as 58,000 vehicles day$^{-1}$ were counted on a nearby commuter route. At a site in Helsinki (Jarvi et al., 2012) the traffic count was 44,000 vehicles day$^{-1}$ whereas at our site it was only 10,000 vehicles day$^{-1}$.

Bias errors of the annual sums were within the same range as what has been reported at other urban sites (Jarvi et al., 2012), resulting from the canceling of positive and negative errors. However, it must be highlighted that there was an underestimation of large magnitude fluxes combined with overestimation of small fluxes. This is a somewhat fortunate constellation for the robustness of annual sums, but something that is important to bear in mind when working with gap-filled half hourly values or smaller subsets of the data. This phenomenon also results in the particular, slightly compressed (along the y-axis) scatterplot shapes (Fig. 5) for all gap-filling methods. Nevertheless, the machine learning methods did not underestimate large magnitude fluxes as severely as the MDS method.

We estimated systematic uncertainties resulting from setting a u$^*$ threshold to filter out periods of low turbulence. We found the resulting uncertainties in annual sums due to the u$^*$-threshold to be relatively low, not exceeding 64 g C m$^{-2}$ yr$^{-1}$ (or 4.6% of the annual sum). Natural sites such as a Pacific Northwest Douglas-fir forest (Morgenstern et al., 2004) found systematic uncertainties due to u$^*$ filtering of 90 g C m$^{-2}$ yr$^{-1}$ (or 33 % of the annual flux), with an average from eleven different sites of approximately 65 g C m$^{-2}$ yr$^{-1}$ (Falge et al., 2001). We are not aware of any previous attempts to estimate such systematic uncertainties at urban sites. The general applicability of methods to determine u$^*$-thresholds at urban sites needs further study, across multiple sites, since considerable variation and uncertainty in different methods remains (Papale et al., 2006; Barr et al., 2013). At urban sites, there are more different processes affecting the nighttime efflux (such as natural gas emissions and vehicle emissions) compared to forest sites (only soil respiration). In addition, the urban boundary layer does not always become stable at night (Nakamura and Oke, 1988). After a partitioning of fluxes into biogenic and anthropogenic contributions, one could potentially find a more robust estimate of the u$^*$-threshold.

The estimated random uncertainties of the gap-filled annual sums at our urban site ranged from 18 g C m$^{-2}$ yr$^{-1}$ (or 1.6%) to 44 g C m$^{-2}$ yr$^{-1}$ (or 3.8%). This was only slightly higher than the range reported in the literature for forest sites (Hollinger and Richardson, 2005; Liu et al., 2009; Menzer et al., 2013), which increases our confidence in eddy covariance CO$_2$ sums from urban sites. We also found the half-hourly uncertainties inferred from our models to increase with the flux magnitude (i.e., error variances were heteroscedastic) and approximately Laplacian distributed, which are typical statistical properties of eddy covariance measurement error (Hollinger and Richardson, 2005; Richardson et al., 2006).

4.2. Spatial variability of fluxes

In addition to imputing missing data, our wind direction dependent modeling approach provided us with information about the spatial representativeness of the eddy covariance measurements. We found that simulated annual sums of F$_C$ differed by a factor of two depending on wind direction (Fig. 7). This indicated that the modeled time series captured the spatial variability in both the biogenic and anthropogenic CO$_2$ sources and sinks in a reproducible way. The monthly sums for June in sectors w240 and w270 were in close agreement with a bottom-up study by Peters and McFadden (2012), which suggests that fluxes from the south-west of the tower (associated with the golf course) were primarily representative of CO$_2$ exchange by urban vegetation and soils. Furthermore, fluxes over the vegetated area had less uncertainty than fluxes over wind sectors that had more vehicle emissions (Fig. 7), also by a factor of about two. This could imply that biogenic fluxes are less uncertain than anthropogenic fluxes in terms of the random error component, possibly due to larger spatiotemporal inhomogeneity in half-hourly aggregated flux observations associated with anthropogenic activity. Future work on partitioning of fluxes into their components will be needed to further investigate this hypothesis.

We found that using wind direction as a proxy for the spatial variability of fluxes was sufficient to gap-fill the time series at our site. This is consistent with the wind sector approach used for most diagnostic analyses of urban eddy covariance measurements (e.g., Sorensen and Moller-Jensen (2003); Grimmmond et al., 2004; Christen et al., 2011). However, Crawford and Christen (2014) found that using wind direction as a proxy for the turbulent flux source area may not account for all observed variations in plan-area land cover and CO$_2$ fluxes, and suggest to include information about the atmospheric stability into models. We did not find a significant relationship between gap-filling model performance and atmospheric stability; however, this would be site specific depending on the surface cover heterogeneity and flux footprint domain, and should be evaluated at each site.

4.3. Role of anthropogenic emissions

Initially, we found our models insensitive to traffic variables when we performed preliminary analyses using the entire record of data, without explicitly accounting for wind sector differences. However, when we binned the observations and trained gap-filling models by wind sector, we found significantly different sensitivities to the explanatory variables. For example, for observations in w30, which were influenced by local vehicle emissions, the best gap-filling models favored trafSN as a predictor. In addition, sensitivity to traffic variables was also confirmed in knockout experiments for subsets of the data. The gap-filling models (for wind sectors w0, w30, w60 and w90) performed better if traffic was included as a variable. This was in contrast to a study in Helsinki, Finland (Jarvi et al., 2012), which found that traffic observations did not increase gap-filling performance significantly, despite a much larger traffic value (44,000 vehicles day$^{-1}$ compared to 10,000 vehicles day$^{-1}$ at our site). However the measurement height in Helsinki was lower (31 m versus 40 m at our site) resulting in smaller source areas, and the road was also further away from the tower (at 150 m distance versus 70 m at our site). Other studies, such as Christen et al. (2011), found emissions from vehicular traffic to be a significant emission source and showed that traffic influenced the flux measurements’ temporal signatures (e.g., through weekday-weekend differences). Additional work is required at other urban sites, combined with synthesis studies to further improve understanding of traffic impacts on urban eddy covariance measurements and implement gap-filling models accordingly. Traffic variables improved gap-filling performance at our site, thus, we recommend that traffic counts be considered for gap-filling at other urban sites with vehicular emissions in the footprint area.
Measurements that quantify other anthropogenic emissions, such as natural gas combustion for space heating or cooking, follow diurnal cycles of human behavior and may be correlated with temporal variables (e.g., time of day). However, we did not find that such variables were more important in wind sectors with more homes. These anthropogenic drivers may be better represented by directly incorporating additional information into gap-filling models, such as from household energy consumption statistics or by output from models that attempt flux partitioning through intercomparing seasonal differences.

4.4. Gap-filling methods for urban sites

The machine learning based, wind direction dependent gap-filling models developed here had an overall performance of 74%–78% (Fig. 5) on unseen test data ($R_{\text{test}}^2$). These results were significantly better than MDS (by 14%–18% in explainable variability in the data), which has been widely used for gap-filling natural and managed vegetation sites (Falge et al., 2001; Reichstein et al., 2005; Moffat et al., 2007). More importantly, the performance of the machine learning based gap-filling for this urban site was similar to those commonly obtained for forest or grassland sites (Moffat et al., 2007). The overall model performance compared favorably with previous studies that have gap-filled urban CO$_2$ flux time series, although there are relatively few (Schmidt et al., 2008; Jarvi et al., 2012). The use of ANN and RBF gap-filling for CO$_2$ flux observations by Schmidt et al. (2008) provided a $R_{\text{test}}^2$ performance of 0.67 for six weeks of measurements at 65 m over an urban neighborhood in Münster, Germany. The study of Jarvi et al. (2012) reported annual sums for a five year record in Helsinki, Finland, using ANN based gap-filling with an $R_{\text{test}}^2$ performance of up to 0.4. Note that statistical performances can vary depending on the complexity in the source area of the tower and the length of the data record. Nevertheless, our results showed that gap-filling in urban environments can successfully provide similar performance as in natural environments, if the spatial heterogeneity in the source area is taken into account (e.g., through training models by wind sector) and the available predictors can explain the majority of the variability in the data. Our suburban tower site had higher vegetation cover compared to many urban sites; therefore, we believe the modeling framework should be tested at other urban sites. Not limited to CO$_2$ measurements, but also carbon monoxide or methane, our gap-filling models could potentially contribute to estimating total carbon emissions of urban environments.

The objective of this study was not to determine which of the three methods is the best gap-filling method but rather to evaluate the potential of each method within the modeling framework that we have presented (Fig. 4). Estimates of annual $F_C$ sums among the different gap-filling methods varied by up to about 275 g C m$^{-2}$ yr$^{-1}$ (Table 3), which may be attributed to larger estimates of small nighttime fluxes by the ANNs compared to RBF and GP (as inferred from fingerprint plots for all methods, not shown). At the same time, there were differences in estimates of large magnitude fluxes (positive or negative), both underestimation of CO$_2$ uptake by the ANNs and underestimation of CO$_2$ release by RBF and GP. RBF and GP predictions were more similar to each other than to ANNs, but $GP$ gap-filling had a slightly higher performance on the test data sets than both ANN and RBF (better by 4% and 3%, respectively). All three machine learning gap-filling methods performed substantially better in reproducing diurnal cycles of the flux observations than MDS, as shown in the fingerprint plots (Fig. 6). The machine learning predicted time series showed a more distinct and regular diurnal cycle, with recurring events such as rush hour emission peaks more clearly discernable compared to MDS. There were also fewer “stripe patterns” (indicating a nearly constant diurnal cycle), in the fingerprint plots of the machine learning methods (Fig. 6). RBF did best in reproducing large (positive or negative) fluxes (Fig. 4(b)), although tending to overfit the data. At the same time, random uncertainties inferred from model runs on the bootstrap samples were consistently lower for RBF than for GP and ANN. This, together with the RBF showing slightly higher gap-filling performance than ANN, suggests that RBF is superior to ANN for gap-filling in urban environments, consistent with the findings of Schmidt et al. (2008).

We tested our models using 13 explanatory variables including meteorological observations, traffic counts, and satellite derived greenness indices (Table 1), but our analyses showed that not all of them were needed to do gap-filling with high performance. Models that used a subset of only six variables performed as well as the gap-filling models commonly used in natural vegetation. It is likely that any of the machine learning regression techniques described here could give satisfactory gap-filling results at other sites in heterogeneous terrain. Site-specific (spatial and/or temporal) subsetting of data, as well as applying re-sampling methods when observations are scarce, will aid in this endeavor. However, the results will also depend on the quality of the flux measurements and the availability of explanatory variables. Regarding the required length of flux observations, we advise that the presented methods should comprise a measurement period of at least several months to constrain representative models.

5. Conclusions

We presented an approach to calculate robust annual sums of CO$_2$ flux for three years of EC observations in a suburban landscape, via wind direction dependent, machine learning gap-filling approaches. Such a model was needed because currently available gap-filling models had been designed for homogeneous terrain and for sites in which biological processes predominate. At the KUOM urban flux tower site, we obtained total annual CO$_2$ flux sums ranging from 1064 to 1382 g C m$^{-2}$ yr$^{-1}$, across different years and different gap-filling methods. Bias errors of annual sums resulting from gap-filling did not exceed 18 g C m$^{-2}$ yr$^{-1}$ (or 1.8 % of the annual flux), and random uncertainties did not exceed $\pm 44$ g C m$^{-2}$ yr$^{-1}$ (or $\pm 3.8\%$). Regardless of the gap-filling method used, the year-to-year differences in carbon exchange at this site were small. In contrast, the modeled annual sums of $F_C$ differed by a factor of two depending on wind direction. This indicated that the models captured the spatial variability in both the biogenic and anthropogenic CO$_2$ sources and sinks in a reproducible way.

Our results suggest that gap-filling for urban eddy covariance sites may be improved using new models that explicitly incorporate wind direction. The resulting gap-filling performance was comparable to that typically obtained at natural sites, explaining between 64% and 88% of the variability in the fluxes. Monthly carbon budgets simulated by the gap-filling models were also in good agreement with an ecophysiological bottom-up study at the same site. Our gap-filling models considered numerous meteorological, anthropogenic and temporal explanatory variables. However, the machine learning regression methods are not limited to incorporate the variables used in this study, but rather are flexible enough to use other site-specific variables, such as data from footprint models, or information about anthropogenic emission sources other than traffic. In addition, the method presented here may be useful at other sites in complex terrain, such as in logged forests or ecosystems under disturbance from fire or pests.
Acknowledgments

The KUOM flux study was funded by a grant to J.P.M. from NASA Earth Science Division (NNG04GN80C), as a component of the North American Carbon Program (NACP). The authors would like to thank Andres Schmidt and Antjie Maria Moffat for insightful and valuable discussions about gap-filling methods. Also, the paper was improved by the highly constructive comments by the two anonymous reviewers and the editor. We thank Dario Papale and Tristan Quaife for sharing their code.

Appendix

A1. Machine learning methods specification

Three different machine learning techniques have been used in this work for gap-filling. Two of them, artificial neural networks (ANNs) and Gaussian Processes (GPs) are specified in more detail in the following.

ANNs are a set of interconnected neurons (or units), which are organized in several layers ranging from an input layer (containing one node for each predictor or input variable, i.e., six out of 14 variables in Table 1), one or several hidden layer(s) and one output layer (containing one node for each target variable, in this study only one for FC). If there are no cycles in the network architecture and information flows in only one direction, the network is called a feed-forward network. The concept of processing information in parallel and learning independently has been inspired by the biological nervous system. An artificial neural network learns functional relationships through a training process with sample data in which the weights (or connections) between neurons are optimized by minimizing the error of a training function. Mathematically, it has been shown by Cybenko (1989) that a feed-forward network with a single hidden layer and sufficient number of neurons can approximate any continuous, multivariate function. Hence, in the case of our specific gap-filling application, it follows that ANNs have the capability to model continuous FC time series by relating them to meteorological, temporal and anthropogenic input variables. In the parameterization we used here, each neuron in the hidden layer uses a logistic sigmoid function to determine if it is active or not based on the weighted inputs (transfer function) and the output neuron uses a linear function. Typically, the performance of a specific network is measured by mean squared error (MSE), which is the squared difference between the observations in the training data set and ANN model predictions for that same data set. During the iterative learning process, the network weights are adjusted using a conjugate gradient descent approach (Hestenes, 1980; Bishop, 1995) towards the minimal MSE. At every step, weights in the network are adjusted by the error backpropagation algorithm (Rumelhart et al., 1986), which directly uses the partial derivative of the MSE function w.r.t. to every neuron to determine the change in weights. For a more detailed description of this specific ANN parameterization refer to Schmidt et al. (2008). For more theoretical details about ANNs in general, we recommend Bishop (1995) and Gurney (1997) as a starting point.

A Gaussian Process is a generalization of a multivariate Gaussian distribution to infinitely many variables (Rasmussen and Williams, 2006), giving a distribution over functions (Bishop, 2006). Thus, in this probabilistic GP framework, random functions are sampled instead of random variables. We can fully describe a GP by a mean function \( m(x) \) (instead of a simple, one dimensional mean) and a covariance function \( k(x, x') \) (instead of a simple, one dimensional variance):

\[
f(x) \sim GP(m(x), k(x, x')) . \tag{A1}
\]

The random variable \( f(x) \) represents the value of the function for a data point \( x \). In our application, \( f(x) \) is the modeled carbon dioxide flux \( F_C \), and we set the mean function \( m(x) \) equal to zero since our data were standard normalized. The covariance function characterizes the correlation between each pair of points \( (f(x), f(x')) \) in the function space. In this study, we assume that this covariance is a function of the Euclidean distance between the predictors \( x \) and \( x' \). This is based on the notion that points in the function space that are close to each other are highly correlated whereas points more distant from each other are uncorrelated. For example, fluxes measured during similar meteorological conditions \( x \) and \( x' \) would be expected to be more similar to each other than fluxes that were measured when \( x \) and \( x' \) differ significantly. We chose the squared exponential covariance function (also known as a Gaussian kernel) with automatic relevance determination (ARD, cf. MacKay (1994); Bishop (2006))

\[
\text{cov}(f(x), f(x')) = k(x, x') = \sigma_f^2 \exp \left[ -\frac{1}{2 \sigma_d^2} \| x - x' \|^2 \right] . \tag{A2}
\]

This covariance function has a hyperparameter \( \sigma_f^2 \) for the signal variance (or amplitude) and length-scale hyperparameters \( \sigma_d \) (sometimes also called width or range parameters). There is one individual \( \sigma_d \) length-scale hyperparameter for each input variable \( x_d \) in this study, six variables per model, \( D = 6 \). These parameters are called hyperparameters to stress that they are parameters of a nonparametric model (or stochastic process) as opposed to parameters of a closed mathematical model, such as slope and intercept parameters in a simple linear regression.

The key idea of the learning process with GPs is optimizing the hyperparameters in the covariance function given the observations. This was implemented as an iterative, conjugate gradient-based maximization (Golub and Van Loan, 1989) of the likelihood of the data given the hyperparameters. Following the automatic relevance determination principle, after this optimization, every hyperparameter \( \sigma_d \) in Eq. (A2) corresponds to the relevance (or importance) of an input variable \( x_d \) for the regression. Small \( \sigma_d \) parameters increase the exponential term in the covariance function in Eq. (A2) through scaling (or weighting) the differences between the data points \( x \) and \( x' \). If \( \sigma_d \) is large then there will be only a small impact on the covariance between \( x \) and \( x' \) and therefore the GP model will be relatively insensitive to that variable. In our study, this allows us to detect variables that have little effect on the predictions (Bishop, 2006), and to compare the relative importance of variables within one model.

A2. Random uncertainties computed by bootstrapping

To estimate random uncertainties of \( F_C \), we used bootstrapping to resample (re-sampling with repetition, Efron (1979)) the entire available high quality data set 1000 times (Fig. 4, Predictions), based on the predictors of the best model for each wind sector and using all available high quality data without dividing into training and test data sets. We re-trained the gap-filling models for each of the bootstrap samples and calculated the standard deviations of those 1000 repetitions as an estimate of random error for each half hour as well as for seasonal and annual sums of fluxes (by calculating standard deviations of the respective sums). Note that the standard deviations that correspond to the ensemble averages of the ten (or even one hundred) initial best models for each method would be difficult to interpret, because they would likely be overestimating errors due to various explanations of the data resulting from
different predictor sets (permutations of variables and data within the 100 training samples). Therefore, we did not use them as random uncertainty estimates but instead used bootstrapping on one single set of predictors (i.e., the one corresponding to the best model) as described above.

A3. Systematic uncertainties associated with \( \hat{u} \)-filtering

Finally, we assessed systematic uncertainties resulting from \( \hat{u} \)-filtering in the following way. We estimated \( \hat{u} \)-thresholds following Papale et al. (2006), splitting the nighttime data by air temperature into seven classes with equal number of values. For each temperature class, the data set was then further divided into 20 classes according to \( \hat{u} \), and the threshold was selected as the class where the nighttime flux reached more than 95% of the average flux of the 10 following larger \( \hat{u} \) classes. Because the estimation of the \( \hat{u} \)-threshold can be made only at a given confidence level, the distribution of \( \hat{u} \)-thresholds was estimated using 100 bootstrap samples. Both the threshold values at the 5% (\( \hat{u}_{0.05} = 0.075 \), lower quantiles of distribution of \( \hat{u} \)-values) and 95% (\( \hat{u}_{0.95} = 0.205 \), upper quantile of distribution of \( \hat{u} \)-values) confidence levels were then used to filter the data one by one, and recompute annual sums using model predictions for any missing values that resulted from \( \hat{u} \)-filtering. This resulted in a range of \( \hat{F}_C \) sums with lower and upper bounds that reflect systematic uncertainties. In another step of the analysis, we also re-trained gap-filling algorithms entirely, based on a priori filtered data sets using quantiles of the \( \hat{u} \)-threshold distribution to test model robustness to \( \hat{u} \)-filtering and to compare the resulting uncertainties in seasonal and annual sums.

A4. Additional results

### Table A1

<table>
<thead>
<tr>
<th>( R^2 )</th>
<th>RMSE</th>
<th>BEM_{(240)}</th>
<th>( \sigma_1 )</th>
<th>( \sigma_2 )</th>
<th>( \sigma_3 )</th>
<th>( \sigma_4 )</th>
<th>( \sigma_5 )</th>
<th>( \sigma_6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.88</td>
<td>2.68</td>
<td>0.387</td>
<td>0.45</td>
<td>0.48</td>
<td>0.74</td>
<td>1.19</td>
<td>2.36</td>
</tr>
<tr>
<td>2</td>
<td>0.86</td>
<td>2.91</td>
<td>0.285</td>
<td>0.41</td>
<td>0.47</td>
<td>0.67</td>
<td>0.74</td>
<td>0.64</td>
</tr>
<tr>
<td>3</td>
<td>0.86</td>
<td>3.14</td>
<td>0.333</td>
<td>0.50</td>
<td>0.58</td>
<td>0.95</td>
<td>1.10</td>
<td>2.15</td>
</tr>
<tr>
<td>4</td>
<td>0.86</td>
<td>2.78</td>
<td>0.251</td>
<td>0.17</td>
<td>0.67</td>
<td>0.93</td>
<td>1.77</td>
<td>2.47</td>
</tr>
<tr>
<td>5</td>
<td>0.85</td>
<td>2.91</td>
<td>0.136</td>
<td>0.20</td>
<td>0.84</td>
<td>0.97</td>
<td>2.09</td>
<td>2.12</td>
</tr>
<tr>
<td>6</td>
<td>0.85</td>
<td>3.21</td>
<td>0.271</td>
<td>0.18</td>
<td>0.62</td>
<td>0.70</td>
<td>1.39</td>
<td>2.01</td>
</tr>
<tr>
<td>7</td>
<td>0.85</td>
<td>2.93</td>
<td>0.165</td>
<td>0.20</td>
<td>0.61</td>
<td>0.99</td>
<td>1.93</td>
<td>1.97</td>
</tr>
<tr>
<td>8</td>
<td>0.85</td>
<td>3.28</td>
<td>0.317</td>
<td>0.18</td>
<td>0.62</td>
<td>0.70</td>
<td>1.39</td>
<td>2.01</td>
</tr>
<tr>
<td>9</td>
<td>0.85</td>
<td>3.29</td>
<td>0.169</td>
<td>0.17</td>
<td>0.65</td>
<td>0.99</td>
<td>1.93</td>
<td>1.97</td>
</tr>
<tr>
<td>10</td>
<td>0.85</td>
<td>2.79</td>
<td>0.374</td>
<td>0.04</td>
<td>0.78</td>
<td>0.84</td>
<td>1.53</td>
<td>2.76</td>
</tr>
</tbody>
</table>

Gap-filling performance and log length-scale hyperparameters \( \sigma_j \) (ordered by their importance for the regression) of the 10 best GPs for wind sector \( w_{40} \). Smaller length-scale parameters are less relevant for the regression than large length-scale parameters. Note that \( \sigma_6 \) can be compared within only a single model, not between models. For acronyms please refer to Table A1. Units of RMSE, BEM_{(240)} and \( \sigma_j \) are all in [$\mu$mol CO$_2$ m$^{-2}$ s$^{-1}$].

### References


