Prediction of Satellite-Based Column CO$_2$ Concentration by Combining Emission Inventory and LULC Information

Shrutilipi Bhattacharjee, *Member, IEEE,* Jia Chen, *Member, IEEE*

Abstract—In this article, we generate a regional mapping of space-borne carbon dioxide (CO$_2$) concentration through a data fusion approach, including emission estimates and Land Use and Land Cover (LULC) information. NASA's Orbiting Carbon Observatory-2 (OCO-2) satellite measures the column-averaged CO$_2$ dry air mole fraction (XCO$_2$) as contiguous parallelogram footprints. A major hindrance of this data set, specifically with its Level-2 observations, is missing footprints at certain time instants and the sparse sampling density in time. This article aims to generate Level-3 XCO$_2$ maps on a regional scale for different locations worldwide through spatial interpolation of the OCO-2 retrievals. To deal with the sparse OCO-2 sampling, the cokriging-based spatial interpolation methods are suitable, which models auxiliary densely-sampled variables to predict the primary variable. In this article, a cokriging-based approach is applied using auxiliary emission data sets and the principles of the semantic kriging (SemK) method. Two global high-resolution emission data sets, the Open-source Data Inventory for Anthropogenic CO$_2$ (ODIAC) and the Emissions Database for Global Atmospheric Research (EDGAR), are used here. The ontology-based semantic analysis of the SemK method quantifies the interrelationships of LULC classes for analyzing the local XCO$_2$ pattern. Validations have been carried out in different regions worldwide, where the OCO-2 and the Total Carbon Column Observing Network (TCCON) measurements coexist. It is observed that the modeling of auxiliary emission data sets enhances the prediction accuracy of XCO$_2$. This article is one of the initial attempts to generate Level-3 XCO$_2$ mapping of OCO-2 through a data fusion approach using emission data sets.

Index Terms—Emissions Database for Global Atmospheric Research (EDGAR), interpolation, land use and land cover (LULC), Orbiting Carbon Observatory-2 (OCO-2), Open-source Data Inventory for Anthropogenic CO$_2$ (ODIAC), semantic kriging (SemK), column-averaged CO$_2$ dry air mole fractions (XCO$_2$).

I. INTRODUCTION

ATMOSPHERIC carbon dioxide (CO$_2$) is a greenhouse gas (GHG) that affects climate change most significantly. A number of satellites were launched in the last few years that were dedicated to GHG observations. Examples include, NASA's Orbiting Carbon Observatory-2 (OCO-2) [1], [2], Exploratory Satellite for Atmospheric CO$_2$ (TanSat) [3], Japanese Greenhouse gases Observing SATellite (GOSAT) [4], and Environmental Satellite (ENVISAT) [5]. The OCO-2 measures

the column-averaged CO$_2$ dry air mole fractions (XCO$_2$) in the atmosphere as contiguous parallelogram footprints, each having area up to about 3 km^2. The XCO$_2$ measurements are obtained using a number of retrieval algorithms, for example, NASA Atmospheric CO$_2$ Observations from Space (ACOS) algorithm [6]. However, these are Level-2 retrievals and irregular in space as well as in time. Fig. 1 shows the OCO-2 Lite Version 9r XCO$_2$ measurements on October 14, 2017. A portion of a measurement swath near Córdoba, Argentina, is magnified where the parallelogram footprints are visible and the corresponding swath width (approximaely 10.3 km [7]) is shown. One footprint is further magnified to show its maximum dimension. The center of the parallelogram is the representative center point of the sounding footprint. The XCO$_2$ values of these one-day retrievals are categorized into five classes (shown in different colors), and the measurement unit is parts per million (ppm, 10^{-6}). It is evident from the figure that, on a single day, a huge portion of the Earth's surface remains unmeasured, and this problem eventually propagates in its time-series measurements as well.
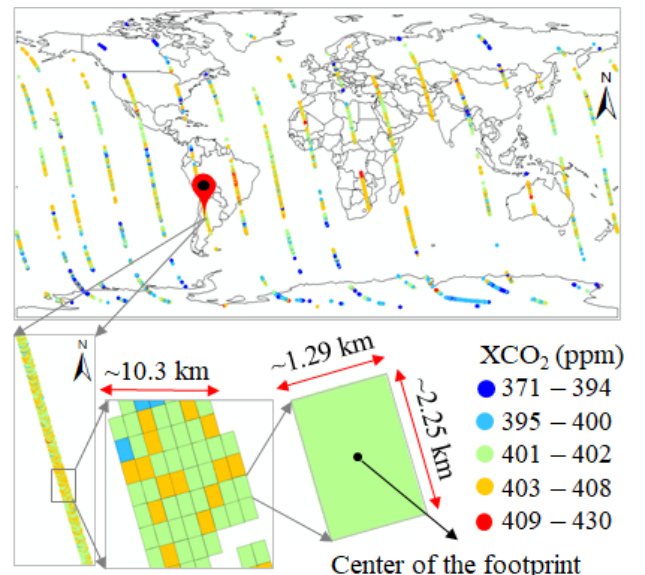

Fig. 1: XCO$_2$ measurements by OCO-2 on October 14, 2017.

This incomplete sampling of atmospheric CO$_2$ prevents us from understanding the global carbon cycle, the mechanisms to control its spatial and temporal variability, the distributions of carbon emission and uptake, and so on. Apart from the satellite remote sensing of CO$_2$, several high-resolution globally

gridded emission inventories are available, which represents the anthropogenic emissions from different activities. For example, the Open-source Data Inventory for Anthropogenic CO_2 (ODIAC) is one such popular emission data product. It is based on the latest country FFCO2 estimates (20002015) made by the Carbon Dioxide Information Analysis Center (CDIAC) at the Oak Ridge National Lab (ORNL), USA, by fuel type (solid, liquid, gas, cement manufacturing, gas flaring, domestic/international shipping, aviation, and marine bunkers) [8]. Currently, it provides monthly emission data from 2000 to 2017. According to Oda *et al.* [8], this data can be considered for the modeling of remotely sensed XCO_2 data from the OCO-2 satellite for understanding the carbon cycle science. Similarly, another comprehensive database of anthropogenic GHG emissions is the Emissions Database for Global Atmospheric Research (EDGAR) [9]. It provides time-series fossil CO_2 emission estimates from 1970 to 2012, including anthropogenic emissions from fossil fuel combustion and production, as well as from industrial processes (cement, steel, liming, urea, and ammonia production, or consumption). These emissions are calculated with a bottomup approach using international statistics for the activity data (such as fuel consumption or crops) and Intergovernmental Panel on Climate Change (IPCC) (2006) values for the emission factors. The uncertainty for the global annual anthropogenic CO_2 emission estimate ranges from 9% to +9% [10]. However, the uncertainties of the GHG inventory vary from nation to nation and are much higher. For example, the estimated uncertainties (2σ) reported for CO_{2eq} in China, India, and Brazil are 11.3%, 17.2%, and 28.3%, respectively, in 2012 [10]. In the subnational scale, the uncertainty is even higher [11]. Considering these emissions to be the major contributing factors to the expansion of CO_2 in the atmosphere, this article aims to predict the missing XCO_2 concentrations of OCO-2 by contemplating the emission estimates of ODIAC or EDGAR as the auxiliary information into the prediction process. The 2017 ODIAC estimates (monthwise) and 2012 EDGAR estimates are used in this article. There is a temporal misalignment of the EDGAR emission data set with the 2017 OCO-2 measurements considered in the empirical study. According to [12], global GHG emissions are dominated by the fossil CO_2 share and steadily increased in the period between 1970 and 2012. Then, the global CO_2 emissions show a slowdown trend and were stalled for the third year in a row with no further increment of the total CO_2 in 2016. Therefore, the EDGAR emission estimate from 2012 can be used as a valid data set for this article. Since, here we try to differentiate between the high and low-emission zones through the emission estimates of a local region, the temporal stability assumption of this auxiliary data should not affect the performance of this approach. The empirical results also justify this assumption in Section V.

To address the drawback of the incomplete OCO-2 measurements stated earlier, spatial prediction, more specifically spatial interpolation, is one of the obvious choices reported in the literature. In contrast to the deterministic spatial prediction methods, the geostatistical interpolation methods or kriging are reported to be more accurate. Among the geostatistical interpolation methods, the variants of kriging, for example,

simple kriging (SK) [13], ordinary kriging (OK) [14], universal kriging (UK) [15], spatial block kriging (BK) [16], fixed rank kriging (FRK) [17], [18], ordinary cokriging (OCK) [19], and spatio-temporal kriging (STK) [20] are popular spatial and spatiotemporal methods to generate a Level-3 mapping from the satellite-based Level-2 retrievals and other related applications. Among other nongeostatistical interpolation methods, inverse distance weighting (IDW), nearest neighbor (NN), thin-plate spline (TPS), and trend surface analysis (TSA) are a few popular methods [21]]. Many studies have reported that the terrestrial land use and land cover (LULC) change have a profound impact on the increase of atmospheric CO_2. For example, Houghton [22] and Houghton and Goodale [23] have stated that the changes in LULC, such as cropland expansion, resulted in the release of 156 Pg of carbon to the atmosphere during the period between 1850 and 1990. They reported this amount to be half of the carbon released from the combustion of fossil fuels over the same period. Hwang and Um [24] have presented an interesting study to find the causal relationship between LULC and carbon emissions using OCO-2's XCO_2 data. They have reported that the LULC classes representing the development activities generally exhibit higher mean XCO_2 compared with the nonanthropogenic LULC types. Therefore, the terrestrial LULC distribution is significant information to be modeled for the mapping of XCO_2 alongside the emission estimates. For this article, the underlying terrestrial distribution of LULC and their influences on the local XCO_2 pattern have to be quantified so that it can be used in any cokriging framework. Here, we have used the semantic modeling component of the semantic-kriging-(SemK)-based interpolation method [25]. This semantic modeling adopts an ontology-based approach to quantify the terrestrial LULC classes for analyzing different environmental variables [26], [27]. This quantified LULC, along with the auxiliary emission data sets, is inserted into a traditional cokriging [19] process. This approach is referred to as cokriging with semantic analysis of LULC (SemCK) [28]. The broader objectives are as follows.

- Spatial interpolation of sparse Level-2 XCO_2 measurements of OCO-2 to create a Level-3 mapping at a particular time instance in a local region.
- Semantically quantifying the auxiliary variable LULC alongside the emission estimates from ODIAC or EDGAR for the primary variable (XCO_2 from OCO-2).
- Validations by comparing the traditional baseline alternative with respect to its extensions with LULC and emission data.
- Comparison of the prediction results with the external Total Carbon Column Observing Network (TCCON) data.

The rest of this article is organized as follows. Section II presents the recent works on creating a Level-3 XCO_2 map from the satellite measurements. Section III describes the SemCK approach focusing on the LULC quantification process of the SemK method. In Section IV, different data sets considered in this article and the specifications of the empirical study have been presented. Section V presents the interpolation results of XCO_2 and other analytical results. Discussions to

understand the XCO_2 mapping application and the SemCK approach are presented in Section VI. Finally, the conclusions are presented in Section VII, including some future prospects.

II. RELATED WORKS

The literature has been reported on different applications of the XCO_2 data from the OCO-2 satellite. Zammit-Mangion et al. [17] have indicated that the gap-filled Level-3 XCO_2 product can be used for validating the OCO-2 retrievals against ground-based measurements or atmospheric transport model output. A few attempts have been made to generate a comprehensive Level-3 mapping of XCO_2 from the partially sampled Level-2 retrievals of the OCO-2 satellite. Article [17] is one such recent work that is reported on generating full coverage satellite Level-3 mapping from the Level-2 product of OCO-2. Zammit-Mangion et al. [17] have developed a spatiotemporal statistical modeling framework of FRK to obtain the global predictions of XCO_2 with Version 7r and Version 8r data of OCO-2. They have validated the prediction framework with TCCON data [29] and found that the Version 8r performs better than the Version 7r. Chevallier et al. [30] have attempted to generate a global continuous map of XCO_2 at daily and monthly temporal resolutions. They have considered a Bayesian Kalman filter (KF) developed on a model of persistence and compared the KF daily-mean XCO_2 maps with TCCON data. Hammerling et al. [31] have adopted a statistical mapping approach for creating a full coverage of the atmospheric CO_2 that is synthetically generated from the PCTM/GEOS-4/CASA-GFED model. They have reported that the overall uncertainty in terms of root-mean-square prediction error (RMSPE) ranges from 0.20 to 0.63 ppm. The prediction accuracy varies with the temporal averaging window (1-day, 4-day, 16-day, and so on) and season (such as January, April, July, and September). The highest and lowest RMPSEs are observed with the combination of one-day averaging window in July and 16-day window in September, respectively.

From the application perspective, similar attempts have been reported to generate a Level-3 XCO_2 mapping with other satellite data as well. Zeng et al. [32] have implemented a spacetime kriging with a moving flexible kriging neighborhood for the ACOS-GOSAT XCO_2 data set. They have validated the prediction results in three ways: through a cross-validation approach, in comparison with TCCON data, and with the model-based simulated data from both CarbonTracker CT2013 [33] and GEOS-Chem [34]. Tadić et al. [16] have proposed a flexible moving window BK method to generate a Level-3 continuous map of satellite data and applied this method for two data sets: the XCO_2 from the GOSAT satellite and the solar-induced fluorescence (SIF) from the Global Ozone Monitoring Experiment-2 (GOME-2) instrument [35]. Jing et al. [14] have proposed a new data fusion technique to generate a global distribution of XCO_2 by fusing GOSAT and SCIAMACHY CO_2 measurements. The OK-based interpolation method applied to this data fusion approach has increased the spatial coverage up to approximately 20% compared with a single data set. Wang et al. [36] have also proposed a method for fusing SCIAMACHY and GOSAT XCO_2 measurements.

This amalgamation approach has increased the global spatial coverage of XCO_2 by 41.3% on a daily and 47.7% on a monthly temporal scale, relative to the measurements of GOSAT. Compared with SCIAMACHY, this article reports even higher spatial coverage. Similarly, Watanabe et al. [37] and Liu et al. [38] have applied kriging interpolation methods to generate a Level-3 spatial distribution of GOSAT XCO_2. Nguyen et al. [39] have proposed a method for fusing the vertical profiles of XCO_2 from OCO-2 and GOSAT. The modified spatial random effects model, applied on the CO_2 profiles, has reduced the mean squared error of data fusion in comparison with the TCCON measurements.

From the methodology perspective, the abovementioned literature on the XCO_2 mapping has used different prediction algorithms. The OK-based interpolation approach is a basic, univariate, minimum variance, linear, unbiased mapping method. The SK assumes the mean of the random field to be known and constant over the study region (SR). The UK approach is a variant of OK, which analyzes the local trend within a predefined search window around the prediction point as a continuous and smoothly varying function of the sampled locations [13]. The BK method can extend any basic kriging method (such as OK and UK) and uses the average expected value over a segment/surface/volume around the prediction point. The point-to-point covariance of the basic kriging is replaced by the point-to-block covariance [13]. The FRK makes the kriging interpolation process efficient to deal with a very large sample size. The reduced complexity in the prediction and error computation process in the FRK is achieved by the choice of covariance function to offer an efficient way to compute the kriging equations [18], [17]. The spacetime kriging [20], [40] method is a variant of the basic spatial interpolation process to include the time component. It decomposes the random function into a deterministic trend component (representing the spatial trend and the temporal trend) and a residual component (representing a stationary spacetime error process). Similarly, the cokriging extension of the basic kriging modifies the baseline method for multivariate analysis [41] to intake multiple variables as input when the prediction variable is sparsely sampled [13], [19].

The assessment of satellite-derived XCO_2 measurements and the CO_2 emission estimates for each other has been found in different applications, for example, to determine the impact of regional fossil fuel emissions on global XCO_2 fields [42], estimate the fire CO_2 emission by using XCO_2 [43], determine CO_2 emissions from megacities by using XCO_2 through inverse modeling approach [44] or from the individual middle to large-sized coal power plants through plume model simulations [45], and observe the anthropogenic CO_2 emission by deriving CO_2 anomalies through deseasonalizing and detrending XCO_2 measurements [46]. This article is one of the initial attempts for modeling the emission inventories as auxiliary variables to generate a mapping surface of spaceborne XCO_2 measurements of OCO-2.

III. METHODOLOGY

From the literature, it is evident that the geostatistical interpolators, mainly different variants of kriging, are the most

widely used techniques for creating a complete mapping from sparsely sampled data. However, variety exists depending on several factors, such as accuracy-complexity tradeoff, smoothing factor of the predictor, available observed/sampled locations, and type of auxiliary information. Here, we have used a cokriging approach that intakes quantified LULC and the emission estimates as the covariates. Our previously proposed SemK-based interpolation method [25], [47] has two components: the quantification of LULC classes for the environmental variables that are inherently influenced by or dependent on the terrestrial LULC dynamics through semantic modeling, and the updating of the traditional Euclidean-distance-based dependences with this semantic modeling. The kriging is often regarded as a "smooth" interpolator that is detrimental to the prediction outcome. The SemK models the LULC to deal with the underlying terrestrial distribution. It aims to reduce the mean absolute error (MAE) and root-mean-square error (RMSE) [25]. Besides LULC, this article attempts to input auxiliary emission estimates to predict the primary variable XCO_2. It uses the first component of SemK, i.e., the semantic modeling of the LULC classes, and then, this quantified LULC is input to a traditional cokriging framework along with the emission estimates. This approach weights all the covariates equally. The basics of the kriging method, semantic modeling of LULC, cokriging method, and the approach to input the quantified LULC to a cokriging framework are presented in Sections III-A–III-D.

A. Overview of Kriging-based Interpolation

The estimation equation of an univariate interpolation method, such as SK and OK, is given in (1) [13]. Here, $\hat{Z}(x_0)$ is the unmeasured primary variable (XCO_2) value at the prediction point x_0, $Z(x_i)$ is the measured XCO_2 value at the ith sampled location x_i, w_i is the impact of/assigned weight to the x_i, $\mathbf{W}$ is the w_i vector, N is the number of sampled locations in the search window, μ is the constant stationary mean, and $\mu(x_0)$ is the local mean of the sample points within the search window. Each x_i is defined on a geographical domain $x \subset \mathbb{R}^2$

$$\hat{Z}(x_0) - \mu = \sum_{i=1}^{N} w_i [Z(x_i) - \mu(x_0)] \qquad (1)$$

Variants of the abovementioned equation are found for different kriging methods based on the weight vector evaluation process. For example, the customized estimation equation for SK is presented in (2), where $\mu(x_0)$ in (1) is replaced by μ, which is the constant mean over the SR [13]

$$\hat{Z}(x_0) = \sum_{i=1}^{N} w_i Z(x_i) + \left[1 - \sum_{i=1}^{N} w_i\right]\mu \qquad (2)$$

The OK is similar to SK but replaces μ in (1) with a local mean $\mu(x_0)$, which leads to the condition $\sum_{i=1}^{N} w_i = 1$.

B. Semantic Modeling of LULC in SemK

The SemK aims to extend the traditional spatial interpolation process, such as SK and OK, with auxiliary terrestrial

LULC information for more accurate prediction. The basic SemK, reported in [25], is presented as an extension of a widely used spatial interpolation method OK [13]. However, the baseline method is not fixed. Depending on the application requirement, any univariate interpolation method can be modified with the LULC quantification process of SemK. In SemK, the terrestrial distribution of LULC is assumed to be the semantic knowledge of the locations such that the equidistant sampled locations, being represented by two different LULC classes, would have different impacts on the prediction point. For quantifying this qualitative property of the terrain, an ontology hierarchy is constructed with all possible LULC classes of the SR, for example, as shown in Fig. 2. It is constructed by following a similar classification proposed in [48]. As the ontology is exhaustive for the whole SR, each of the locations (x_i or x_0) should correspond to one leaf LULC class of the hierarchy. Therefore, proper reasoning of every representative pair of leaf concepts in this hierarchy can quantify the amount of semantic associations between every pair of locations, both sampled and unsampled. Two metrics, proposed by the SemK, to model this association are semantic similarity (SS) and spatial importance (SI). The SS is the measure of the hierarchy-structural association, and the SI is the correlation between the representative measured samples of every pair of LULC classes.
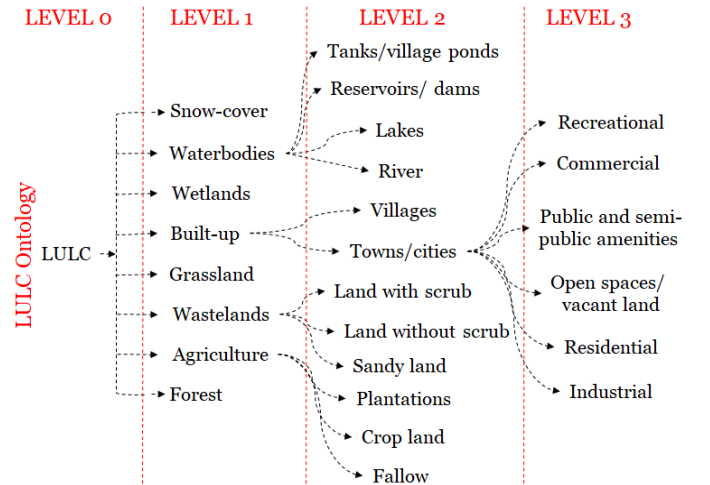

Fig. 2: A LULC ontology.

1) Semantic Similarity: The SS metric considers the ontology hierarchy layout [49], [50] and is modelled by following the modified context resemblance method [25]. The SS score between the pth and qth LULC classes is represented as SS_{pq} and is modeled in (3). Here, $Count_{pq}$ is the count of the common LULCs in the paths of pth and qth LULCs in the hierarchy, starting from level 0 concept LULC. Similarly, the $|LC_p|$ and $|LC_q|$ are the total LULC counts in the respective paths of pth and qth LULCs

$$SS_{pq} = \frac{1}{2}\left(\frac{Count_{pq}}{|LC_p|} + \frac{Count_{pq}}{|LC_q|}\right) \qquad (3)$$

2) Spatial Importance: The SI metric considers the ontology hierarchy layout as well as the measurements from the SR

[51]. As it deals with the measured XCO_2 value of the sampled locations, the SI score for a pair of LULCs is different across the SRs. Mathematically, this metric is evaluated by assessing the Pearson's correlation coefficient between the representative measurements of two LULCs with respect to the primary prediction variable XCO_2. The SI score between two sampled locations or their representative LULC classes, LC_p and LC_q, is denoted as SI_{pq} and is given in (4). Here, $Z(LC_{p_q})$ refers to the XCO_2 value of the qth sample point represented by the LULC class LC_p, and $\overline{Z(LC_p)}$ refers to the average of the XCO_2 value over k sample points taken from the same LULC type LC_p.

$$SI_{pq} = \frac{\sum\limits_{m=1}^{k}(Z(LC_{p_m}) - \overline{Z(LC_p)})(Z(LC_{q_m}) - \overline{Z(LC_q)})}{\sqrt{\sum\limits_{m=1}^{k}(Z(LC_{p_m}) - \overline{Z(LC_p)})^2 \sum\limits_{m=1}^{k}(Z(LC_{q_m}) - \overline{Z(LC_q)})^2}}$$
(4)

For M number of LULC classes, both SS and SI metrics are evaluated for each pair of LULC classes. This results in two $[M \times M]$ symmetric matrices, denoted as $[\mathbf{SS}]_{pq[M \times M]}$ and $[\mathbf{SI}]_{pq[M \times M]}$. These semantic scores together modify the traditional baseline kriging process, followed by modeling the weight vector $\mathbf{W}_{SemK}$ [25]. The uniqueness of SemK lies in the process of modeling LULC through SS and SI metrics, which is the first component of SemK. However, the second component, i.e., the traditional covariance modification process, may vary depending on the application requirements. In this article, the first component is used to quantify the LULC for the primary variable XCO_2, and for the second component, we have used a traditional cokriging framework where the quantified LULC is input as a covariate.

C. Motivating Example

The working principle of the SemK method [25] is described here with a motivating example [52], in comparison to a traditional kriging method OK [53]. For the ease of understanding, the covariance matrix modification in this example is carried out as presented in [25], in contrast to the cokriging principle considered in this article (refer to Section III-D). A SR with nine pixels is considered in Fig. 3(a) where pixel no. 5, represented with a ★, is a missing pixel and its XCO_2 value is required to be predicted from the eight surrounding pixels (pixels: 1 2 3 4 6 7 8 9). All these sampled locations are $\sqrt{2}$ unit apart from the missing location 5. The measured XCO_2s (in ppm) at these surrounding locations are specified in Fig. 3(b), which forms the Z vector as Z = [410 395 404 401 406 403 398 400]. These are assumed XCO_2 values which are not similar to a real life scenario in terms of statistical properties, such as distribution, range, autocorrelation, and others.

We assume that the number of LULC classes in the terrain is five, say, Town/cities: ▨, Waterbodies: ▨, Wastelands: ▨, Agriculture: ▨, and Grassland: ▨. All of these classes belong to the same level-1 of the ontology, except Town/cities in level-2 (refer to Fig. 2). Therefore, the SS score for any pair of LULC classes, among ▨ ▨ ▨ ▨, is the same and that is (1/2+1/2)/2 = 0.5. For class ▨, the SS of the other distinct
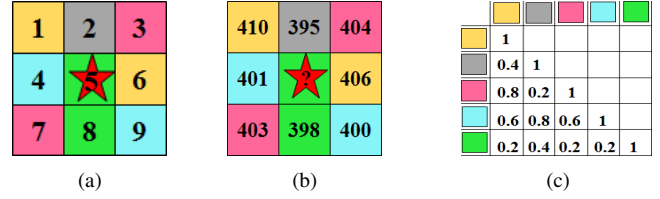

Fig. 3: Motivating example scenario of SemK. (a) SR with 9 pixels. (b) XCO_2 values. (c) Correlation (SI) scores.

LULC classes is: (1/2+1/3)/2 = 0.417. However, their pairwise correlations (SI) are different for every pair and example scores are specified in Fig. 3(c). Ideally it should be calculated from the sampled locations considered, as described in Section III-B2.

There are five different pair-wise Euclidean distances among the sampled locations: 1, $\sqrt{2}$, 2, $\sqrt{5}$, and $\sqrt{8}$. According to the OK method, the semivariances (γ) at these Euclidean distances are evaluated as follows: $\gamma(1) = 28.75$, $\gamma(\sqrt{2}) = 28.75$, $\gamma(2) = 12$, $\gamma(\sqrt{5}) = 19$, and $\gamma(\sqrt{8}) = 25.25$. After fitting the experimental semivariogram, let the sill of the OK is 22.88, and the covariances (cov) are evaluated as follows: $cov(1) = 12.59$, $cov(\sqrt{2}) = 10.65$, $cov(2) = 9.02$, $cov(\sqrt{5}) = 8.61$, and $cov(\sqrt{8}) = 7.92$. Assuming the weight vector of the OK method ($\mathbf{W}_{OK}$) is simply defined as $\mathbf{C}^{-1}\mathbf{D}$ ($\mathbf{C}$ is the traditional covariance matrix given as $[cov(x_i, x_j)]_{N \times N} = [cov(Distance_{ij})]_{N \times N}$ and the $\mathbf{D}$ is the traditional distance matrix given as $[cov(x_0, x_i)]_{N \times 1} = [cov(Distance_{0i})]_{N \times 1}$; i, j $\in$ (1, 2, $\cdots$, N)), the $\mathbf{W}_{OK}$ is evaluated as: $[0.1875\ 0.0625\ 0.1875\ 0.0625\ 0.0625\ 0.1875\ 0.0625\ 0.1875]^T$. The XCO_2 value at ★ is evaluated by the OK as $\sum \mathbf{W}_{OK} \cdot Z = 403.19$.

However, being in the same distance, two pairs of sampled locations can be represented by different LULC pairs. Thus, according to the study region in Fig. 3(a), there are different cases of semantic semivariances [25]. Assuming that the Euclidean-distance-based covariance scores of OK are directly updated by the semantic metrics of the SemK, its spatio-semantic covariances are evaluated as follows: $cov(1, \text{▨▨})$ = $(cov(1) \cdot (SS \cdot SI)_{\text{▨▨}})$ = $(12.59 \cdot (0.417 \cdot 0.4))$ = 2.10, $cov(1, \text{▨▨})$ = 1.26, $cov(\sqrt{2}, \text{▨▨})$ = 4.26, $cov(2, \text{▨▨})$ = 1.80, $cov(\sqrt{5}, \text{▨▨})$ = 0.86, $cov(\sqrt{8}, \text{▨▨})$ = 1.98, and so on. From these covariances, the weight vector of the SemK ($\mathbf{W}_{SemK}$) is evaluated in the same process as the OK method and is given as: $[0.1343\ 0.1158\ 0.1309\ 0.1522\ 0.1336\ 0.1304\ 0.0507\ 0.1521]^T$. Then, the XCO_2 value at ★ is evaluated by the SemK as $\sum \mathbf{W}_{SemK} \cdot Z = 402.53$. In this article, the semantic modeling principle of SemK through SS and SI scores is directly used to quantify the LULC classes and then input to a cokriging framework, as described in Section III-D.

D. Cokriging with Semantic Modeling of LULC (SemCK)

The cokriging-based interpolation method assumes that the primary variable (Z_1) can be estimated with the help of other ($V-1$) auxiliary variables ($Z_2, Z_3, ..., Z_V$) which exhibit some

correlation with the primary variable Z_1. The multivariate cokriging estimator is given as follows [13]:

$$\hat{Z}_1(x_0) - \mu_1 = \sum_{i_1=1}^{N_1} w_{i_1}[Z_1(x_{i_1}) - \mu_1(x_{i_1})] +$$
$$\sum_{j=2}^{V} \sum_{i_j=1}^{N_j} w_{i_j}[Z_j(x_{i_j}) - \mu_j(x_{i_j})] \quad (5)$$

where μ_1 is the stationary mean of the primary variable, $Z_1(x_{i_1})$ is the measured value of the primary variable at the sampled location x_{i_1} and w_{i_1} is the assigned weight to it, N_1 is the total number of sample points, $\mu_1(x_{i_1})$ is the local mean of the sampled locations within the search window, (V-1) is the number of secondary variables, N_j is the number sampled locations (x_{i_j}) of the jth secondary variable within the search window and $\mu_j(x_{i_j})$ local mean of these locations, $Z_j(x_{i_j})$ is the measured value of the jth secondary variable at the sampled location x_{i_j}, and w_{i_j} is the assigned weight considering the jth secondary variable [13].

The weight vector of the cokriging estimator can be evaluated through the covariance and the cross-covariance matrices which should be positive definite. These matrices are modeled through the distance (h)-based semivariogram $(\gamma_{ii}(h))$ and the cross-variogram $(\gamma_{ij}(h))$, defined in (6) and (7) [13], [54]. Here, $Z_i(x_m)$ and $Z_i(x_m + h)$ are the measured values of the variable Z_i at the location x_m and the sample points that are in lag distance h with respect to x_m, that is, $(x_m + h)$, and L is the total pair of samples within the spatial lag h.

$$\gamma_{ii}(h) = \frac{\sum_{m=1}^{L}[Z_i(x_m) - Z_i(x_m + h)]^2}{2L} \quad (6)$$

$$\gamma_{ij}(h) = \frac{\sum_{m=1}^{L}[Z_i(x_m) - Z_i(x_m + h)][Z_j(x_m) - Z_j(x_m + h)]}{2L} \quad (7)$$

Now, the covariance analysis of the SemCK approach also intake the semantic scores of the representative LULCs of each pair of sample points $(x_p, x_p + h)$. According to the SemK method [25], the quantified semantic score between two LULC types, LC_p (for example, urban area) and LC_q (for example, waterbody), is given as $SIS_{pq} = (SS_{pq} \cdot SI_{pq})$. In this article, the SI metric is only used as the SS metric does not give much extra information (as all the LULC classes belong to the same ontology level). The SI metric should be mapped to a positive range before analysis to avoid negative correlation scores. This scaling range may vary across the SRs and the number of available LULC classes. Now, taking the LULC class with the highest primary variable estimate as the reference (urban area in our case), the Z value of the other LULC classes (LC_q) are normalized as $Z^{LC_q} = SIS_{pq}$, where LC_p is the reference LULC class. Therefore, in the SemCK process, the semivariogram considering the LULC dataset $(\gamma_{LC}(h))$ and the cross-variogram $(\gamma_{i\text{-}LC}(h))$ scores that consider LULC as one of the covariates are given in (8) and (9), where $Z^{LC}(x_m)$ and $Z^{LC}(x_m+h)$ are the SIS scores of the representative LULCs of the sampled locations (x_m)

and $(x_m + h)$, respectively.

$$\gamma_{LC}(h) = \frac{1}{2L} \sum_{m=1}^{L}[Z^{LC}(x_m) - Z^{LC}(x_m + h)]^2 \quad (8)$$

$$\gamma_{i\text{-}LC}(h) = \frac{1}{2L} \sum_{m=1}^{L}[Z_i(x_m) - Z_i(x_m + h)]$$
$$[Z^{LC}(x_m) - Z^{LC}(x_m + h)] \quad (9)$$

The semivariances are plotted against distance to derive the covariance and cross-covariance scores in the respective matrices. Finally, these matrices are used by the cokriging process to evaluate the weight vector and predict the missing XCO_2 values.

IV. EMPIRICAL DATA & SPECIFICATIONS

This section presents the details of the empirical study to validate the SemCK interpolation approach for the prediction of missing XCO_2 footprints. The specifications of the datasets and the details of the study regions are described here.

A. Datasets

In this article, we are amalgamating various data sets for a multivariate interpolation scenario. The primary variable is the XCO_2 data from the OCO-2 satellite. Furthermore, three secondary data sets are ESACCI land cover data, the ODIAC data, and the EDGAR data for the emission estimates. Though there is a temporal misalignment of the auxiliary data sets (EDGAR and LULC) with respect to the primary OCO-2 data of the year 2017, the assumption of the temporal stability of the emissions estimates and LULC makes their assessment valid for this article.

1) OCO-2 Lite Level-2 Version 9r Data: The OCO-2 satellite data [1], [55], [56] are obtained through different observation modes, namely, Nadir, Glint, and Target. These modes vary in terms of sensitivity and accuracy of the observations with respect to measurement geometries. The OCO-2 satellite repeats its operation in 16-day cycles. A typical daily Level-2 product is likely to output about 10 000 retrievals worldwide [57], over up to eight different footprints in a swath, observed in 0.333 s, and with a maximum area of 1.29 km $\times$ 2.25 km each [45] (refer to Fig. 1). This Level-2 data are provided in hierarchical data format (HDF or NetCDF), along with latitude, longitude, altitude, and time as the implicit variables of the retrieval. In this article, we have chosen the OCO-2 Lite Level-2 version 9r data [58] of the year 2017 for different regions. Eight SRs are considered for the comparison study. However, for the qualitative performance analysis with the predicted mapping images, two SRs, Lamont, USA, with Nadir and Orléans, France, with Glint XCO_2 samples [59] are chosen to prove the efficacy of the proposed approach for both the observation modes. For the comparison study with TCCON measurements [29], different days are chosen for different SRs, such that OCO-2 and TCCON measurements coexist within a time window and a predefined distance [17]. For example, Fig. 4 depicts the OCO-2 samples in the SR Lamont, USA, which coexists with the corresponding TCCON

measurements, conforming to the aforementioned criteria. The OCO-2's XCO_2 retrievals of a single day are shown in Fig. 1. The OCO-2 science team also provides an interpolated Level-3 product from everyday retrievals, which aims to map XCO_2 with 1 ppm accuracy over the Earth's surface in bins with the resolution of $1° \times 1°$ in latitude and longitude [56]. It is created using the simple averaging method, which is unable to predict outside the OCO-2 swath if the averaging pixel size is relatively small.
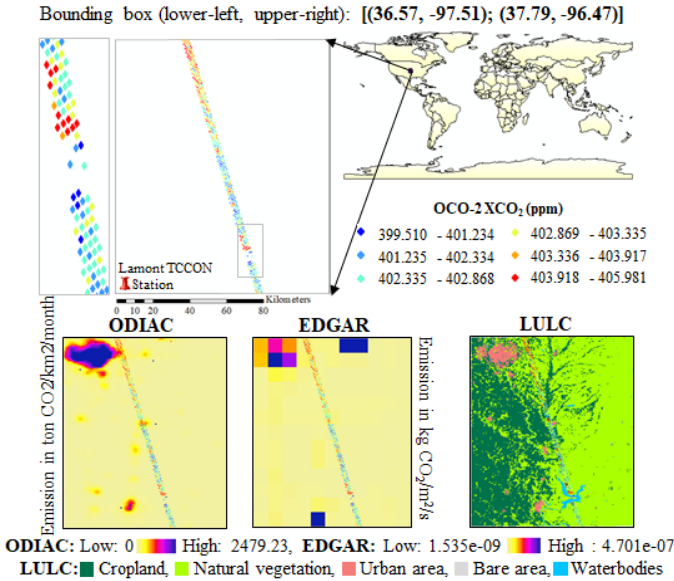

Fig. 4: Details of the SR: Lamont, USA, including the location of the TCCON station, distributions of OCO-2 (Nadir) samples (the center point of each footprint), ODIAC (October 2017), EDGAR (2012), and LULC distributions.

2) ESA CCI LULC Data: We have chosen the 2015 global land cover map from the European Space Agency (ESA) Climate Change Initiative (CCI) project [60]. The resolution of this data is $0.002778°$. Each pixel value (in the range between 0 and 220) corresponds to a land cover class that is defined based on the UN Land Cover Classification System (LCCS). For a simplified analysis, we have generalized the available LULC classes into seven classes, as given in Table I. In the ontology hierarchy, these classes are mapped as follows:cropland as 'agriculture', natural vegetation as 'grassland', urban area as 'built-up', bare area as 'wastelands', waterbodies as 'waterbodies', and Permanent snow and ice as 'snow-cover' (refer to Fig. 2). This article has used a scaled LULC map that is created according to the LULC quantification process (refer to Section III-III-D). For the correlation analysis in SI metric, both LULC and XCO_2 are scaled in the same resolution of 1 km.

3) ODIAC Emission Inventory: The ODIAC inventory [8], [61] is a global monthly emission data set consisting of CO_2 emission estimates from fossil fuel combustion, cement production, gas flaring, domestic/international shipping, aviation, and marine bunkers. It also considers multiple spatial emission proxies by fuel type, such as nighttime light data, specific to gas flaring and ship/aircraft fleet tracks and includes monthly

TABLE I. RECLASSIFICATION OF ESA CCI LULC PRODUCT

Pixel value of ESA CCI LULC data	LULC type
0	No data
10, 11, 12, 20, 30, 40	Crop land
50, 60, 61, 62, 70, 71, 72, 80, 81, 82, 100 ,110, 120, 121, 122, 130, 140, 150, 151, 152, 153, 160, 170, 180	Natural vegetation
190	Urban area
200, 201, 202	Bare area
210	Waterbodies
220	Permanent snow and ice

and interannual emission variations as well. The ODIAC2018 [61] version is used here, which is available for the years from 2000 to 2017. Between the available GeoTIFF and NetCDF formats, we have chosen the former with a resolution of 1 km $\times$ 1 km and in the unit of ton C/cell (monthly total). The weekly/diurnal emissions can also be modeled by multiplying the TIMES temporal scaling factors with monthly emission fields [8], [62]. For this article, we have chosen the data from January 2017 to December 2017 (varies with respect to the SRs) to consider the same month as of OCO-2 data. The global emission data for October 2017 are presented in Fig. 5. The statistical measures of this global data are reported. For better visualization, a region near Mark Twain National Forest, USA, is magnified. The local emission estimates for that region and the corresponding statistical measures are also reported.
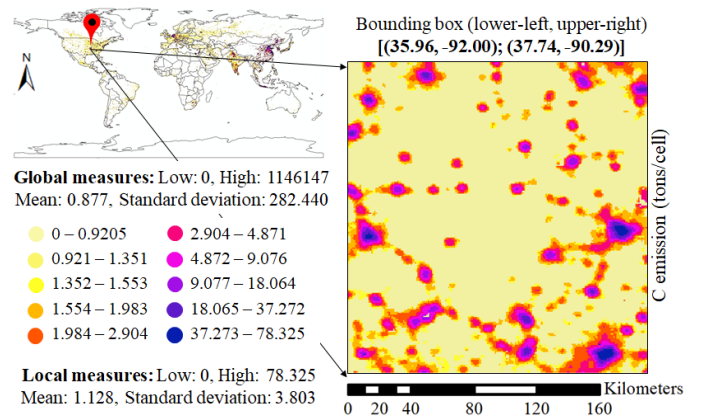

Fig. 5: ODIAC's emission data for October 2017.

4) EDGAR Emission Inventory: The EDGAR emission database [63] provides the annual CO_2 emission data from fossil fuel combustion and industrial processes, excluding short-cycle biomass burning, large-scale biomass burning, and carbon emissions/removals from land-use, land-use change, and forestry. It provides annual emission data for other GHGs as well, such as methane (CH_4), nitrous oxide (N_2O), from 1970 to 2012. We have chosen the latest 2012 data of version EDGAR v4.3.2 (v432_CO2_excl_short-cycle_org_C_201) in NetCDF format with resolution of $0.1° \times 0.1°$ in latitude and

longitude, and the emission estimates given in kg $CO_2/m^2/s$. The EDGAR emission is calculated using a bottomup approach, and it scales the activity data based on international annual statistics with the best-available emission factors. It further considers different proxies, such as road transportation and population, to downscale the data to a finer spatial resolution [10].

5) TCCON Data: For performance analysis of the SemCK approach, the TCCON measurements are chosen as an external validation resource to compare with the interpolated OCO-2 data. The TCCON [64] is a ground-based network of the Fourier transform spectrometers recording the direct solar spectra in the near-infrared spectral region. Currently, there are 26 active TCCON stations (operational sites) and five future sites spread over the world [65]. Apart from the precise column-averaged abundances of CO_2, other GHGs, such as CH_4, N_2O, carbon monoxide (CO), and water (H_2O), are also retrieved from the spectra and reported by TCCON. Here, we have used GGG2014 version of TCCON data [64], available in NetCDF format [66]. Eight TCCON stations [67]–[74] and their surrounding locations in the land region of different continents are chosen for this article (details are given in Section IV-B). In each of the SRs, the OCO-2 and the TCCON measurements should ideally coexist approximately within a distance and time window. For comparison, we have chosen the temporally nearest TCCON measurement with respect to the OCO-2 measurement time, usually in the early afternoon local time. In this article, the maximum spatial and temporal distance considered are 78 km and 11 min, respectively, with an exception in the SR Saga, Japan, having the temporally nearest measurement almost 1 h before.

B. Study Regions

The SemK-based LULC quantification process can be applied everywhere on the Earth's surface. However, for a homogeneous surface, such as ocean or desert, the SemK eventually converges to its baseline kriging method [47], as the semantic scores are same for every pair of locations. Therefore, this semantic modeling process of SemK is only suitable for the land regions with heterogeneous LULC distribution. We have carried out the interpolation process in eight SRs worldwide, in different continents with heterogeneous population densities and LULC distributions. Keeping in mind that the interpolation accuracy of the SemCK approach is supposed to be compared with the TCCON measurements, one of the basic criteria of choosing SRs is to check for the available OCO-2 measurements around the TCCON stations on different days [67]–[75]. The details of the chosen SRs, the corresponding TCCON locations, and the days (YYYYMMDD) of prediction are specified in Table II [76]. Furthermore, the XCO_2 is highly dependent on the wind properties (speed and direction) and the boundary layer height. These atmospheric dynamics are assumed to be constant in this article. Therefore, for comparison, both the OCO-2 and TCCON measurements should coexist almost at the same time.

All the SRs are chosen as the rectangular regions where the OCO-2 measurements are spread as a tilted swath. We have

TABLE II. DETAILS OF THE SRs (TCCON LOCATION AND THE CHOSEN DATE)

SR	Location	Date
Caltech, USA	34.13623N, 118.126897W	20170508
Garmisch, Germany	47.476N, 11.063E	20170311
Jet Propulsion Lab (JPL), NASA, USA*	34.202N, 118.175W	20171218
Karlsruhe, Germany	49.1002N, 8.4385E	20170421
Lamont, USA	36.604N, 97.486W	20171013
Lauder, New Zealand	45.038S, 169.684E	20170123
Saga, Japan**	33.240962N, 130.288239E	20171010
Tsukuba, Japan	36.0513N, 140.1215E	20170322

*non-functioning after May 2018
** unlike the other stations, here the nearest measurement is almost one hour apart

chosen SRs of different sizes, varying numbers of samples, different distances between OCO-2 swath and TCCON location, and so on. For example, the distributions of the OCO-2 measurements in the SR: Lamont, USA, i.e., the approximate location of the corresponding TCCON station (red pushpin, labeled as "Lamont TCCON Station"), are shown in Fig. 4. A magnified view of the XCO_2 swath (showing the center points of the footprint) is shown in the upper-left box where some missing footprints are visible within the swath, alongside the wide unmeasured region outside the swath. The interpolation is carried out for the whole rectangular SR. In the figure, corresponding auxiliary data sets, that is, the emissions of ODIAC, EDGAR, and LULC distribution within the SR are also shown. The considered OCO-2 samples sometimes need preprocessing as the range of the measurement is unrealistic. In this article, the SRs: Lamont, USA, are chosen for reporting the predicted mapping images with Nadir XCO_2 samples, comparing the performances of different approaches within an SR. For the rest of the SRs, the numeric prediction results by different approaches and their comparison studies are reported.

V. RESULTS

Before carrying out the actual interpolation in the chosen SRs, the correlation between the primary variable XCO_2 with the auxiliary emission estimates should be checked. For example, in Lamont, USA, the high-valued XCO_2 samples are mainly located near to the emission hotspots. Otherwise, the auxiliary variables may not be suitable for this article.

Next, the spatial interpolation process is carried out using OCO-2's XCO_2 data with and without emission estimates and LULC for the comparison study. In this article, four different kriging approaches have been applied in each of the SRs: the traditional SK without emission and LULC (the SK is considered as the baseline univariate kriging method for all the cokriging approaches), SemCK with LULC as a covariate (without considering emission as an auxiliary variable), SemCK with ODIAC and LULC, and SemCK with EDGAR and LULC as the auxiliary variables. In addition,

SR	Predicted by **I.** **Natural Neighbor**	Predicted by **II.** **SK** without LULC and emission	Predicted by **III.** **SemCK with** **LULC** as covariate	Predicted by **IV.** **SemCK with** **LULC & ODIAC** as covariates	Predicted by **V.** **SemCK with** **LULC & EDGAR** as covariates	**Legend** (predicted XCO$_2$ in ppm)
Lamont, USA						
Comparison	MAE: 0.464 RMSE: 0.558 PSNR: 57.236	MAE: 0.365 RMSE: 0.443 PSNR: 59.246	MAE: 0.358 RMSE: 0.435 PSNR: 59.400	MAE: 0.312 RMSE: 0.386 PSNR: 60.442	MAE: 0.315 RMSE: 0.397 PSNR: 60.193	

Experimental details: Kriging semivariogram type: stable; samples' base distribution: Student's t; maximum neighboring points: 5; minimum neighboring points: 2; mapping image resolution: 1 km^2 (for SK, *SemCK with LULC*, and *SemCK with ODIAC and LULC*), 0.1° × 0.1° in latitude and longitude (for *SemCK with EDGAR and LULC*)

Fig. 6: Comparison study with the predicted mapping images in the SR: Lamont, USA (Nadir observations).

we have used the natural neighbor method for the comparison study. The quantitative comparisons are carried out considering five different aspects.

- *Intra-SR Comparison Study With OCO-2 Measurements:* We assume that some of the OCO-2 samples are missing from the swath, which are predicted by different approaches and then compared with the measured XCO$_2$, followed by reporting the error measures MAE, RMSE, and peak signal-to-noise ratio (PSNR).
- *Intra-SR Comparison Study With TCCON Measurements:* We report the performance of all the approaches using RMSE and PSNR in the same SR by comparing the prediction with the TCCON measurement (mostly outside the OCO-2 swath).
- *Inter-SR Comparison Study With OCO-2 Measurements:* We show the comparison among all the SRs in terms of RMSE considering the OCO-2 swath and also give an insight on how the RMSEs vary across different SRs with respect to the autocorrelation measure of the XCO$_2$ samples.
- *Inter-SR Comparison Study With TCCON Measurements:* We compare the predicted values reported by different methods with the corresponding TCCON measurements using RMSE in all the SRs.
- *Intra-SR Comparison Study With Glint OCO-2 Measurements:* Since the aforementioned approach A reports the predicted mapping images with Nadir XCO$_2$ samples, another mapping image comparison with Glint XCO$_2$ samples in the SR: Orléans, France is presented. It is reported with the same experimental specifications adopted for approach A.

The abovementioned three error measures, i.e., MAE, RMSE, and PSNR, are popular and standard in the literature of spatial interpolation. They are mathematically expressed in (10) - (12). The $Z Z(x_i)$, $\hat{Z}(x_i)$, MAX, and N represent the measured value of the ith XCO$_2$ sample, predicted value of the same, maximum value of the XCO$_2$ samples in the OCO-2

swath, and the total number of validation samples in the swath, respectively. The MAE is an estimation of the prediction "bias' but does not report the variance of the prediction approach, whereas the RMSE represents the prediction variance using the sum of the squared bias and is commonly referred to as a "gold standard" to analyze the prediction performance [17]. As MAE and RMSE are error measures, the lower their values, the better the model. On the other hand, a higher PSNR indicates a better prediction model and vice versa. It is important to estimate PSNR for the local kriging processes. PSNR is highly used when a subset of observations are used in prediction [17]. Thus, it is required to check whether the SemCK with emission and LULC reports better PSNR compared with others [17]. The three error measures are defined in (10) - (12)

$$\text{MAE}_{XCO_2} = \frac{\sum_{i=1}^{N} |\hat{Z}(x_i) - Z(x_i)|}{N} \text{ ppm} \tag{10}$$

$$\text{RMSE}_{XCO_2} = \sqrt{\frac{\sum_{i=1}^{N} [\hat{Z}(x_i) - Z(x_i)]^2}{N}} \text{ ppm} \tag{11}$$

$$\text{PSNR}_{XCO_2} = 20 \, log_{10}\left(\frac{MAX}{RMSE}\right) \text{ dB} \tag{12}$$

A. Intra-SR Comparison Study with OCO-2 Measurements

In the Lamont, USA, the number of XCO$_2$ samples in the OCO-2 swath is 464. For the performance analysis within the swath, approximately one-tenth, that is, 46 sample points spread all over the swath are first discarded during the interpolation process, and then, their predicted values are compared with the corresponding measured value by each of the interpolation approaches. Here, the five prediction approaches are: I. natural neighbor; II. SK without emission and LULC; III. SemCK with LULC only; IV. SemCK with ODIAC and LULC; and V. SemCK with EDGAR and LULC. The predicted mapping images along with the experimental details of the interpolation processes are presented in Fig. 6. For visual comparison with corresponding input data sets, the resolution of the gridded images is approximately 0.1° × 0.1°

in latitude and longitude for the approach V and e averaged approximately in the 1 km $\times$ 1 km for the rest (refer to Fig. 4). However, for the numeric comparison and error estimation, all the predicted surfaces are averaged approximately in the 1 km $\times$ 1 km resolution. Different semivariogram types are modeled across the SRs based on the distributions of the primary and auxiliary data sets. The maximum and the minimum neighboring points of the interpolation process, the major range, and so on may vary with respect to the SR, prediction approach, and so on.

Fig. 6 presents the predicted mapping images by different interpolation approaches in the SR: Lamont, USA. It helps us to visually analyze the efficacy of the SemCK approach. The extent of the prediction layer is set to the range of the whole SR so that it can predict outside the swath. The first column of the table presents the prediction by the natural neighbor method. However, it must be observed that even if the extent of the SR is defined outside the OCO-2 swath, this method is not suitable to produce the predicted mapping for the whole SR, as the pixel size is 1 km^2. Being variants of the kriging method, the rest of the approaches are capable of producing the whole predicted surface outside the swath. In the predicted surface reported by II and III, average behavioral patterns are observed outside the OCO-2 swath. The approaches IIIV consider the covariates as LULC, ODIAC and LULC, EDGAR, and LULC, respectively.While approach III introduces the LULC-based terrestrial heterogeneity in the predicted surface, the approaches IV and V can identify the high-emission pixels from the corresponding emission data sets. Both IV and V report some high XCO$_2$ pixels outside the OCO-2 swath. These pixels mostly spatially coincide with the high-emission pixels reported by ODIAC or EDGAR (refer to Fig. 4). The main advantages of the SemCK-based approaches are stated as follows.

- it is capable of mapping the XCO$_2$ surface for the whole SR (refer to Fig. 6).
- In addition, it is capable of identifying the high-emission pixels or emission hotspots outside the OCO-2 swath, which are likely to report high XCO$_2$.
- the LULC-based terrestrial heterogeneity introduces the variations of XCO$_2$ among different LULC classes in the predicted surface.

To quantitatively compare the results of different approaches, three standard error measures are reported in Fig. 6 for the SR: Lamont, USA. As described in the experimental details (refer to Section V-A), approximately one-tenth of the sample points that are discarded during the interpolation are considered during this comparison. It is observed from the figure that approach IV, SemCK with ODIAC and LULC, followed by approach V, SemCK with EDGAR and LULC, report the least MAE and RMSE. The approach SemCK with LULC reports the third least error measures. The univariate kriging approach, followed by the natural neighbor method, which has not intake any covariate information, produces the highest error in prediction. For the PSNR, approach IV followed by approach V reports higher measures. The SemCK approach with emission and LULC data introduces

the underlying heterogeneity of the terrain to the prediction process and, consequently, enhances the prediction accuracy.

B. Intra-SR Comparison Study with TCCON Measurements

As mentioned earlier, TCCON is a ground-based network of GHG measurements. For each of the chosen SRs, only one TCCON measurement station/location is available. We have chosen the entire OCO-2 swath for this interpolation approach. Table III reports the error measures RMSE and PSNR (for single location comparison, the MAE is the same as the RMSE) by comparing the TCCON station's measured value with the predicted ones by different approaches. It must be noted that the natural neighbor method is unable to predict values outside the OCO-2 swath for 1 km^2 grid and cannot be compared with the TCCON measurements. For the other approaches, the reported RMSEs show that the SemCK with ODIAC and LULC yields the least RMSE and the highest PSNR, followed by SemCK with EDGAR and LULC. Thus, SemCK approaches also perform better outside the OCO-2 swath in the SR: Lamont, USA.

TABLE III. COMPARISON STUDY WITH TCCON MEASUREMENTS (RMSE AND PSNR) IN THE SR: LAMONT, USA

Error measure	Prediction approach				
	I	II	III	IV	V
RMSE	—	0.290	0.278	0.031	0.200
PSNR	—	62.921	63.301	82.439	66.156

C. Inter-SR Comparison Study with OCO-2 Measurements

In this comparison, the error measure RMSE is reported in Table V using the comparison approach A, i.e., Intra-SR Comparison Study with OCO-2 Measurements, but across the eight SRs. The prediction has been carried out for these regions separately, with suitable kriging assumptions, such as sample's base distribution, semivariogram type, and maximum and minimum neighboring samples. According to Table V, the highest errors reported in all the SRs are either produced by approach I, natural neighbor or by approach II, SK without emission and LULC. The rest of the approaches produce lesser RMSE. The least RMSE is always reported by either approach V or by approach IV, excluding approach I as it has limited possibility to predict outside the swath. Therefore, the inclusion of emission and LULC information in the cokriging process enhances the prediction accuracy in general. Regression analysis is carried out considering the validation samples from all the SRs, and the corresponding coefficient of determination (R^2) values are reported in Table IV. The overall uncertainties in prediction reported by the approaches IV and V are 1.127 and 1.080 ppm, respectively.

However, it must be noted from Table V that the range of RMSEs reported by five prediction approaches varies across different SRs. This occurs due to the heterogeneity present in the underlying terrain, which varies across the SRs. Consequently, the autocorrelation in the OCO-2 swath also varies in different SRs. The spatial autocorrelation among the OCO-2

samples can be calculated using a popular and well-established metric in spatial statistics, Moran's I [77], ranging in [1, 1]. This measure is important to check how the heterogeneous distribution of the samples affects the accuracy of the SemCK results. A positive spatial autocorrelation measure indicates that similar values of XCO_2 occur near one another and vice versa. The Moran's I measure of the OCO-2 samples chosen for the validation is also specified in Table V. An analytical study is reported in Fig. 7 to understand how the autocorrelation of the OCO-2 samples influences the prediction accuracy. In the graph, the X-axis represents the Moran's I value for each SR, and the Y-axis represents the corresponding RMSE in that SR. The trend lines are fit considering the RMSEs of each of the prediction approaches. The generic behavior of the exponential fits shows that the error in prediction decays with the increment of the autocorrelation in the OCO-2 samples. Approach V has generally reported lower average RMSE with respect to its baseline alternative (approach II) across the SRs.

TABLE IV. COMPARISON STUDY WITH OCO-2 MEASUREMENTS (R^2) IN ALL THE EIGHT SRs

R^2	Prediction approach				
	I	II	III	IV	V
	0.861	0.878	0.895	0.898	0.906

TABLE V. COMPARISON STUDY WITH OCO-2 MEASUREMENTS (RMSE) IN EIGHT DIFFERENT SRs AND THEIR MORAN's I MEASURE

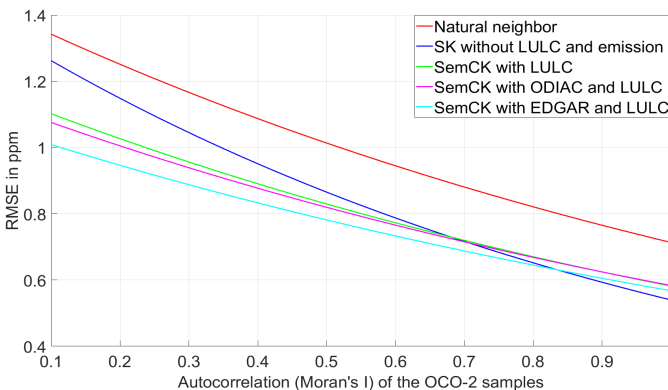| SR	Moran's I	RMSE (ppm)				
		I	II	III	IV	V
Caltech	0.559	0.694	0.548	0.478	0.476	0.454
Garmisch	0.419	2.154	1.454	1.430	1.426	1.422
JPL	0.688	0.669	0.942	0.867	0.863	0.828
Karlsruhe	0.179	1.110	0.960	0.852	0.794	0.614
Lamont	0.200	0.558	0.443	0.435	0.386	0.397
Lauder	0.073	2.148	2.217	2.183	2.172	2.142
Saga	0.163	1.082	0.906	0.885	0.876	0.708
Tsukuba	0.061	1.066	1.140	0.655	0.642	0.644

Fig. 7: Analysis of RMSE with respect to the autocorrelation of the OCO-2 swath (validation samples) across the SRs.

D. Inter-SR Comparison Study with TCCON Measurements

Table VI presents a comparison with different prediction approaches in eight SRs with respect to the TCCON measurements. Similar to Lamont, USA (refer to Table III), the approaches considering auxiliary emission data have produced less error compared with the traditional approaches for other SRs as well. As the SemCK approaches are sensitive to the high-emission pixels or the emission hotspots, the range of the RMSE varies across the SRs. For example, the highest emission pixel around the TCCON station Tsukuba, Japan, spatially coincides with both ODIAC and EDGAR. The approaches IV and V also behave similarly in this SR, producing similar RMSE, that is, $RMSE_{IV}^{Tsukuba} = 1.673$ and $RMSE_{V}^{Tsukuba} = 1.680$. However, for the SR: Saga, Japan, the highest emission pixels reported in ODIAC and EDGAR do not exactly coincide. The RMSE reported by approaches IV and V also vary. Therefore, apart from the methodological differences in ODIAC and EDGAR, the results reported by these two approaches may also differ if the emission pixels do not spatially coincide in these two data sets. From the perspective of the prediction approach, the SemCK with emission and LULC is likely to estimate better in and around the high-emission pixels. It also indicates that this approach performs better and is particularly suitable for the prediction in and around urban areas due to high anthropogenic emission activities compared with other LULC types.

TABLE VI. COMPARISON STUDY WITH THE TCCON MEASUREMENTS (RMSE) IN EIGHT DIFFERENT SRs

SR	RMSE (ppm)				
	I	II	III	IV	V
Caltech	—	0.660	0.529	0.335	0.321
Garmisch	—	0.544	0.329	0.126	0.124
JPL	—	0.850	0.125	0.088	0.028
Karlsruhe	—	0.365	0.142	0.080	0.058
Lamont	—	0.290	0.278	0.031	0.200
Lauder	—	0.471	0.325	0.237	0.013
Saga	—	2.091	1.076	0.956	0.357
Tsukuba	—	1.918	1.710	1.673	1.680

E. Intra-SR Comparison Study with **Glint** OCO-2 Measurements

As mentioned earlier, the qualitative performance evaluation in comparison approach A is carried out with the Nadir XCO_2 samples. However, this approach is applicable to the Glint XCO_2 samples as well. For the performance analysis with the Glint XCO_2 samples [59], the XCO_2 observations are chosen around the Orléans TCCON station on March 16, 2017. Fig. 8 shows the SR details, including the Orléans TCCON station's location, the OCO-2 sample distribution, ODIAC, EDGAR, and LULC distribution. The number of OCO-2 samples in this swath is 547, and we have adopted a similar approach as the SR: Lamont, USA, to split the available OCO-2 samples into the prediction/validation segments. The prediction is carried out with the similar experimental details, as specified in Section V-A, and the predicted mapping images

with these Glint XCO_2 samples are reported for five interpolation approaches. Fig. 9 presents these mapping images with the corresponding numeric error measures. It is observed from the pictorial and the numeric results that the SemCK approaches with emission estimates and LULC generate more realistic prediction surfaces compared with its baseline alternative for the Glint XCO_2 samples as well. In these approaches, the qualitative mapping images can capture the high-emission pixels, and they are assigned higher XCO_2 values compared with the traditional approach.
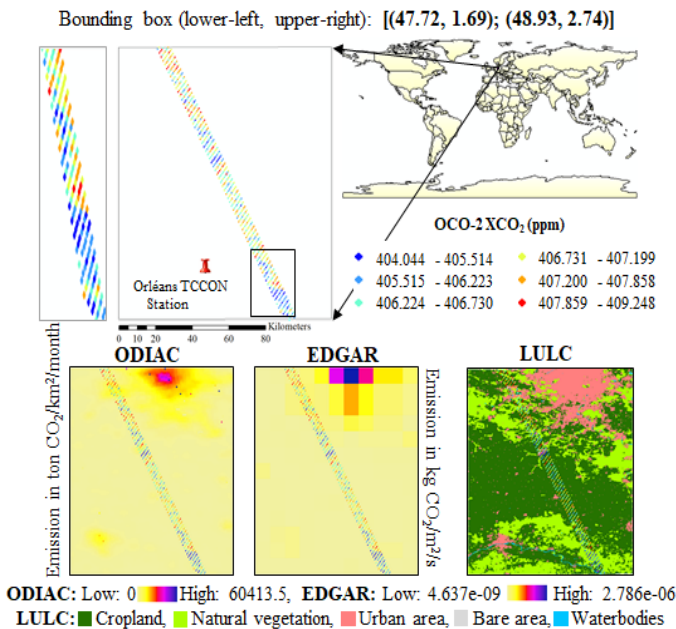

Fig. 8: Details of the SR: Orléans, France, including the location of the TCCON station, distributions of Glint XCO_2 samples, ODIAC (March 2017), EDGAR (2012), and LULC distributions.

VI. DISCUSSION

In this section, few assumptions, restrictions, and the possible solution approaches to be implemented in the future are discussed. This could be helpful to understand the application of the mapping of space-borne XCO_2, the SemCK with covariates approach, for generating a global XCO_2 mapping, etc. As mentioned earlier, a general restriction of the semantic modeling or the LULC quantification process of the SemK method is its applicability over the land region. If it is applied over a homogeneous surface in terms of LULC distribution (for example ocean, desert, permanent snow, and ice) and/or emission inventory, the LULC quantification process will not provide any extra information, and the multivariate approach considering emission and LULC as covariates will be reporting the similar results as that of the baseline kriging method. Therefore, for applying this approach worldwide to create a global Level-3 XCO_2 mapping, the homogeneous surfaces should be taken care of differently. For example, this approach should be extended with a different weighting strategy to differentiate between the waterbodies in the land region and the ocean. Here, the proper scaling of the LULC data using semantic metrics is also important, which may vary across

the SRs. In the future, a strategy should be defined to apply uniformly optimal scaling for the global mapping of XCO_2, which leads to the analysis of scaling-accuracy tradeoff.

Furthermore, the Level-3 XCO_2 mapping reported in this article is representative in a local region for 16 days, as mostly a single OCO-2 swath is available for a particular location in a 16-day cycle. In the SR: Lamont, USA, the local interpolation is carried out considering a region of size 94×135 km^2 approximately (geodesic distance). However, on a single day, the distance between two OCO-2 swaths can be more than 2500 km. Therefore, to create a global map, considering a temporal averaging window is a better choice compared with choosing a single-day measurement. The existing literature has reported the Level-3 maps using different averaging window (for example, 1-day, 4-day, 16-day in [31] and 1-day and 16-day in [17]). The spatiotemporal extension of the SemCK approach can be used for the temporal prediction of XCO_2 [78].

For creating a global XCO_2 mapping, another important factor, besides the accuracy requirement, is the computational overhead of the kriging process. The number of samples associated with the global measurement is very high, which further increases with the chosen averaging window. For example, a 1-day OCO-2 measurement on October 14, 2017, consists of 232265 samples, whereas a 16-day measurement window (October 115, 2017) consists of 1916359 measurement samples. Furthermore, the univariate kriging methods, such as SK and OK, require $O(N^3)$ computations to generate the weight vector for a single point from its covariance matrices, as it involves matrix inversion and multiplication operations (N is the sample size) [79]. However, the multivariate approaches, such as SemCK with emission and LULC, having (M1) auxiliary variables as input, require $O((N \cdot M)^3)$ computations [80]. Therefore, the computational resource requirement is very high for any multivariate kriging approaches. In this scenario, this approach can adopt the computationally intelligent equations of the efficient kriging methods, such as FRK, to speed up the process of global XCO_2 mapping [17].

A major assumption of this article is the static environment in terms of atmospheric transport through the wind. As the static Level-3 XCO_2 mapping, generated by this approach, is representative for 16 days, we have assumed that the wind directions are constantly changing in a 16-day time window, and the pairs of opposite wind directions are annulling the effects of each other in terms of the atmospheric transport of the particles. However, incorporation of this atmospheric dynamics is needed for the mapping of XCO_2. A possible solution is presented here [81]. In this approach, the stochastic time-inverted Lagrangian transport (STILT) model for analyzing the atmospheric transport will be considered [82], [83]. It is a Lagrangian particle dispersion model (LPDM) [84] to derive the upstream influence region for the atmospheric measurement locations. Given a receptor point, this method can model the sensitivity of the atmospheric tracer mixing ratio with respect to upstream surface fluxes. It has wide applications in the modeling of GHGs and other trace gases, and can be coupled with the emission inventories and biospheric flux models to understand their local carbon budgets [82]. To

SR	Predicted by **I.** **Natural Neighbor**	Predicted by **II.** **SK** without LULC and emission	Predicted by **III.** **SemCK** with **LULC** as covariate	Predicted by **IV.** **SemCK** with **LULC & ODIAC** as covariates	Predicted by **V.** **SemCK** with **LULC & EDGAR** as covariates	**Legend** (predicted XCO_2 in ppm)
Orléans, France						404.044 − 405.427 405.427 − 405.837 405.837 − 406.090 406.090 − 406.277 406.277 − 406.461 406.461 − 406.617 406.617 − 406.772 406.772 − 406.896 406.896 − 407.033 407.033 − 407.186 407.186 − 407.337 407.337 − 407.467 407.467 − 409.248
Comparison	MAE: 0.372 RMSE: 0.467 PSNR: 58.846	MAE: 0.313 RMSE: 0.399 PSNR: 60.218	MAE: 0.272 RMSE: 0.355 PSNR: 61.236	MAE: 0.237 RMSE: 0.322 PSNR: 62.091	MAE: 0.210 RMSE: 0.297 PSNR: 62.784	
	Experimental details: Kriging semivariogram type: stable; samples' base distribution: Student's t; (maximum & minimum neighboring points, mapping image resolution are same as specified in Fig. 6)					

Fig. 9: Comparison study with the predicted mapping images in the SR: Orléans, France (Glint observations).

estimate the XCO_2 at the unsampled locations considering the effect of atmospheric transport, emission particles are moved backward in time from the receptor point (prediction location) to simulate possible trajectories from their upstream locations. The simulation will generate a footprint (sensitivity) map, representing the upstream area that influences the air arriving at the receptor point considering the other pixels for the whole SR. According to [82], the footprint maps can be combined with the emission data of the same region to get a prior map of XCO_2. This prior XCO_2 map will be considered as another input to the multivariate cokriging framework that will intake the influence of atmospheric transport. The overall idea is presented as a block diagram in Fig. 10.
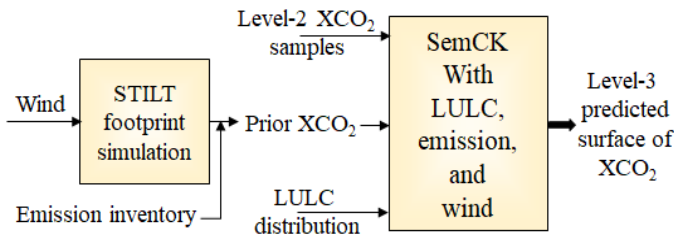

Fig. 10: Modified SemCK approach with STILT model.

Finally, it must be noted that the accuracy of this approach, or any spatial interpolation method in general, depends on the SR, selected validation samples, and their measurement uncertainty. Different validation samples may produce different results for the interpolation approaches. An optimal selection strategy of the validation samples should be implemented in the future. Furthermore, if the measurement is highly uncertain, the prediction error could be high outside the OCO-2 swath. However, if the uncertainty of most of the samples in the whole swath is high, the comparison approach A, that is, the cross-validation within the OCO-2 swath may not give higher prediction error. Therefore, the performance of the interpolation approaches should be validated against more external ground-based measurements outside the OCO-2 swath. In the future, we are considering the validation of

the prediction surfaces with COCCON data [85] alongside the TCCON, as it is a growing network providing accurate column measurements. In addition, using city column networks, we will be able to validate the urban-rural differences in the prediction surface [86], [87], [88].

VII. CONCLUSIONS

This article presents a novel data fusion approach to create a complete mapping of XCO_2 of the OCO-2 satellite. It is one of the initial attempts to map the XCO_2 swath of OCO-2 for the regional scale considering two emission inventories, ODIAC and EDGAR, separately as the auxiliary data. A multivariate spatial interpolation approach is applied, which intakes the semantically quantified LULC and emission estimates as covariates. Following the principle of the SemK method, this approach models the underlying LULC-based heterogeneity in the terrain, and along with the ODIAC or EDGAR emission data, predicts the missing XCO_2 concentrations. Validation has been carried out in different locations worldwide, where the OCO-2 measurements and TCCON measurements coexist within a spatial and temporal window. The performances of these approaches (SemCK with ODIAC and LULC, SemCK with EDGAR and LULC) have been compared with the traditional univariate SK and SemCK with LULC approaches that do not consider the emission estimate as an auxiliary variable. It is observed that the prediction errors reported by the SemCK with emission and LULC together are lower in the considered SRs compared with the two baseline alternatives. The SemCK with ODIAC and LULC, and SemCK with EDGAR and LULC report 12.81% and 18.31% accuracy enhancements (in terms of RMSE), respectively, compared with the univariate SK method across the eight SRs. Furthermore, the SemCK approaches with LULC and emissions can predict the high-emission pixels or the emission hotspots that are outside the OCO-2 swath. The accuracies reported by these approaches vary with different factors, such as heterogeneity of the LULC distribution, the autocorrelation among the OCO-2 samples, and the distance from the nearest high-emission pixels, etc.

Regarding the effectiveness of this approach, the choice of the auxiliary emission data sets is also important. From the empirical studies, the emission data set EDGAR has been found to produce slightly better results (except the SRs: Lamont, USA, and Tsukuba, Japan) for the cross-validation with OCO-2 measurements (refer to Table V). For the validation with external TCCON measurements, which is a more pragmatic approach compared with the basic cross-validation approach, many of the results are comparable between SemCK with ODIAC and EDGAR, and neither of the data sets is shown to be superior across all the SRs. The reason could be the methodological differences between ODIAC and EDGAR. Therefore, for local mapping in a single SR, more case studies are needed to conclude the advantage of one emission data set over the other in terms of accuracy. Another approach could be to combine different emission data sets together. For creating the global Level-3 XCO_2 mapping, it is desirable to consider emission estimates with fine spatial and temporal resolutions, and having a high degree of correlation with the XCO_2 samples.

As future prospects, atmospheric transport should be modeled by this approach to map the XCO_2 because it has a significant influence on the atmospheric CO_2 concentrations. This multivariate approach can be applied worldwide to generate a global mapping of XCO_2. This product would be useful for understanding the global carbon cycle and different climatological studies involving atmospheric CO_2, validating the Level-3 OCO-2 measurements against ground-based measurements or atmospheric transport model output and vice versa.

VIII. Acknowledgement

References

[1] D. Crisp, "Measuring Atmospheric Carbon Dioxide from Space with The Orbiting Carbon Observatory-2 (OCO-2)," in *Earth Observing Systems XX*, vol. 9607. International Society for Optics and Photonics, 2015, p. 960702.

[2] C. Frankenberg, R. Pollock, R. Lee, R. Rosenberg, J. Blavier, D. Crisp, C. O'Dell, G. Osterman, C. Roehl, P. Wennberg *et al.*, "The Orbiting Carbon Observatory (OCO-2): Spectrometer Performance Evaluation Using Pre-launch Direct Sun Measurements," *Atmospheric Measurement Techniques*, vol. 8, no. 1, p. 301, 2015.

[3] Y. Liu, J. Wang, L. Yao, X. Chen, Z. Cai, D. Yang, Z. Yin, S. Gu, L. Tian, N. Lu *et al.*, "The TanSat Mission: Preliminary Global Observations," *Science Bulletin*, vol. 63, no. 18, pp. 1200–1207, 2018.

[4] A. Butz, S. Guerlet, O. Hasekamp, D. Schepers, A. Galli, I. Aben, C. Frankenberg, J.-M. Hartmann, H. Tran, A. Kuze *et al.*, "Toward Accurate CO2 and CH4 Observations from GOSAT," *Geophysical Research Letters*, vol. 38, no. 14, 2011.

[5] C. Frankenberg, J. Meirink, P. Bergamaschi, A. Goede, M. Heimann, S. Körner, U. Platt, M. van Weele, and T. Wagner, "Satellite Chartography of Atmospheric Methane from SCIAMACHY on Board ENVISAT: Analysis of the Years 2003 and 2004," *Journal of Geophysical Research: Atmospheres*, vol. 111, no. D7, 2006.

[6] C. W. O'Dell, B. Connor, H. Bösch, D. O'Brien, C. Frankenberg, R. Castano, M. Christi, D. Eldering, B. Fisher, M. Gunson, J. McDuffie, C. E. Miller, V. Natraj, F. Oyafuso, I. Polonsky, M. Smyth, T. Taylor, G. C. Toon, P. O. Wennberg, and D. Wunch, "The ACOS CO2 Retrieval Algorithm - Part 1: Description and Validation Against Synthetic Observations," *Atmospheric Measurement Techniques*, vol. 5, no. 1, pp. 99–121, 2012. [Online]. Available: http://dx.doi.org/10.5194/amt-5-99-2012

[7] M. Guo, J. Li, L. Wen, and S. Huang, "Estimation of co2 emissions from wildfires using oco-2 data," *Atmosphere*, vol. 10, no. 10, p. 581, 2019.

[8] T. Oda, S. Maksyutov, and R. J. Andres, "The Open-source Data Inventory for Anthropogenic CO2, Version 2016 (ODIAC2016): A Global Monthly Fossil Fuel CO2 Gridded Emissions Data Product for Tracer Transport Simulations and Surface Flux Inversions," *Earth System Science Data*, vol. 10, no. 1, pp. 87–107, 2018.

[9] M. Crippa, D. Guizzardi, M. Muntean, E. Schaaf, F. Dentener, J. A. van Aardenne, S. Monni, U. Doering, J. G. Olivier, V. Pagliari *et al.*, "Gridded Emissions of Air Pollutants for the Period 1970-2012 within EDGAR v4. 3.2," *Earth System Science Data*, vol. 10, pp. 1987–2013, 2018.

[10] G. Janssens-Maenhout, M. Crippa, D. Guizzardi, M. Muntean, E. Schaaf, F. Dentener, P. Bergamaschi, V. Pagliari, J. G. Olivier, J. A. Peters *et al.*, "Edgar v4. 3.2 global atlas of the three major greenhouse gas emissions for the period 1970–2012," *Earth System Science Data*, vol. 11, no. 3, pp. 959–1002, 2019.

[11] L. R. Hutyra, R. Duren, K. R. Gurney, N. Grimm, E. A. Kort, E. Larson, and G. Shrestha, "Urbanization and the Carbon Cycle: Current Capabilities and Research Outlook from the Natural Sciences Perspective," *Earth's Future*, vol. 2, no. 10, pp. 473–495, 2014.

[12] G. Janssens-Maenhout, M. Crippa, D. Guizzardi, M. Muntean, E. Schaaf, J. Olivier, J. Peters, and K. Schure, "Fossil CO2 & GHG Emissions of All World Countries," *document EUR 28766 EN, Publications Office of the European Union, Luxembourg City, Luxembourg*, 2017.

[13] J. Li and A. D. Heap, *A Review of Spatial Interpolation Methods for Environmental Scientists*. Geoscience Australia Canberra, 2008.

[14] Y. Jing, J. Shi, T. Wang, and R. Sussmann, "Mapping Global Atmospheric CO2 Concentration at High Spatiotemporal Resolution," *Atmosphere*, vol. 5, no. 4, pp. 870–888, 2014.

[15] W. Caballero, R. Giraldo, and J. Mateu, "A Universal Kriging Approach for Spatial Functional Data," *Stochastic Environmental Research and Risk Assessment*, vol. 27, no. 7, pp. 1553–1563, 2013.

[16] J. M. Tadić, X. Qiu, V. Yadav, and A. M. Michalak, "Mapping of Satellite Earth Observations Using Moving Window Block Kriging," *Geoscientific Model Development*, vol. 8, no. 10, pp. 3311–3319, 2015.

[17] A. Zammit-Mangion, N. Cressie, and C. Shumack, "On Statistical Approaches to Generate Level-3 Products from Satellite Remote Sensing Retrievals," *Remote Sensing*, vol. 10, no. 1, p. 155, 2018.

[18] N. Cressie and G. Johannesson, "Fixed Rank Kriging for Very Large Spatial Data Sets," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 1, pp. 209–226, 2008.

[19] P. M. Atkinson, E. Pardo-Iguzquiza, and M. Chica-Olmo, "Downscaling Cokriging for Super-Resolution Mapping of Continua in Remotely Sensed Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 2, pp. 573–580, Feb 2008. [Online]. Available: http://dx.doi.org/10.1109/TGRS.2007.909952

[20] J.-M. Montero, J. Mateu *et al.*, *Spatial and Spatio-temporal Geostatistical Modeling and Kriging*. John Wiley & Sons, 2015, vol. 998.

[21] J. Li and A. D. Heap, "A Review of Comparative Studies of Spatial Interpolation Methods in Environmental Sciences: Performance and Impact Factors," *Ecological Informatics*, vol. 6, no. 3, pp. 228–241, 2011.

[22] R. A. Houghton, "Revised Estimates of the Annual Net Flux of Carbon to the Atmosphere from Changes in Land Use and Land Management 1850–2000," *Tellus B*, vol. 55, no. 2, pp. 378–390, 2003.

[23] R. Houghton and C. Goodale, "Effects of Land-Use Change on the Carbon Balance of Terrestrial Ecosystems," *Ecosystems and Land Use Change*, vol. 153, pp. 85–98, 2004.

[24] Y. Hwang and J.-S. Um, "Performance Evaluation of OCO-2 XCO2 Signatures in Exploring Casual Relationship between CO2 emission and Land Cover," *Spatial Information Research*, vol. 24, no. 4, pp. 451–461, 2016.

[25] S. Bhattacharjee, P. Mitra, and S. K. Ghosh, "Spatial Interpolation to Predict Missing Attributes in GIS Using Semantic Kriging," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 8, pp. 4771–4780, August 2014.

[26] S. Bhattacharjee and S. K. Ghosh, "Automatic Resolution of Semantic Heterogeneity in GIS: an Ontology Based Approach," in *Advanced Computing, Networking and Informatics-Volume 1*. Springer, 2014, pp. 585–591.

[27] S. Bhattacharjee, R. R. Prasad, A. Dwivedi, A. Dasgupta, and S. K. Ghosh, "Ontology based Framework for Semantic Resolution of Geospatial Query," in *2012 12th International Conference on Intelligent Systems Design and Applications (ISDA)*. IEEE, 2012, pp. 437–442.

[28] S. Bhattacharjee and J. Chen, "Global Mapping of CO2 Concentration of OCO-2 by Statistical Modeling of Anthropogenic Emission Dataset," in *Geophysical Research Abstracts*, vol. 21, 2019.

[29] D. Wunch, P. O. Wennberg, G. Osterman, B. Fisher, B. Naylor, C. M. Roehl, C. O'Dell, L. Mandrake, C. Viatte, M. Kiel *et al.*, "Comparisons of the Orbiting Carbon Observatory-2 (OCO2) XCO2 measurements with TCCON," *Atmospheric Measurement Techniques*, vol. 10, no. 6, pp. 2209–2238, 2017.

[30] F. Chevallier, G. Broquet, C. Pierangelo, and D. Crisp, "Probabilistic Global Maps of the CO2 Column at Daily and Monthly Scales from Sparse Satellite Measurements," *Journal of Geophysical Research: Atmospheres*, vol. 122, no. 14, pp. 7614–7629, 2017.

[31] D. M. Hammerling, A. M. Michalak, and S. R. Kawa, "Mapping of CO2 at High Spatiotemporal Resolution Using Satellite Observations: Global Distributions from OCO-2," *Journal of Geophysical Research: Atmospheres*, vol. 117, no. D6, 2012.

[32] Z.-C. Zeng, L. Lei, K. Strong, D. B. A. Jones, L. Guo, M. Liu, F. Deng, N. M. Deutscher, M. K. Dubey, D. W. T. Griffith, F. Hase, B. Henderson, R. Kivi, R. Lindenmaier, I. Morino, J. Notholt, H. Ohyama, C. Petri, R. Sussmann, V. A. Velazco, P. O. Wennberg, and H. Lin, "Global Land Mapping of Satellite-observed CO2 Total Columns Using Spatio-temporal Geostatistics," *International Journal of Digital Earth*, vol. 10, pp. 426–456, 2017.

[33] W. Peters, A. R. Jacobson, C. Sweeney, A. E. Andrews, T. J. Conway, K. Masarie, J. B. Miller, L. M. P. Bruhwiler, G. Pétron, A. I. Hirsch, D. E. J. Worthy, G. R. van der Werf, J. T. Randerson, P. O. Wennberg, M. C. Krol, and P. P. Tans, "An Atmospheric Perspective on North American Carbon Dioxide Exchange: CarbonTracker," *Proceedings of the National Academy of Sciences*, vol. 104, no. 48, pp. 18 925–18 930, 2007. [Online]. Available: http://dx.doi.org/10.1073/pnas.0708986104

[34] R. Nassar, D. B. Jones, P. Suntharalingam, J. M. Chen, R. J. Andres, K. J. Wecht, R. M. Yantosca, S. S. Kulawik, K. W. Bowman, J. R. Worden *et al.*, "Modeling Global Atmospheric CO2 with Improved Emission Inventories and CO2 Production from the Oxidation of Other Carbon Species," *Geoscientific Model Development*, vol. 3, no. 2, p. 689, 2010.

[35] J. Joiner, L. Guanter, R. Lindstrot, M. Voigt, A. Vasilkov, E. Middleton, K. Huemmrich, Y. Yoshida, and C. Frankenberg, "Global Monitoring of Terrestrial Chlorophyll Fluorescence from Moderate-Spectral-Resolution Near-Infrared Satellite Measurements: Methodology, Simulations, and Application to GOME-2," *Atmospheric Measurement Techniques*, vol. 6, no. 10, pp. 2803–2823, 2013.

[36] T. Wang, J. Shi, Y. Jing, T. Zhao, D. Ji, and C. Xiong, "Combining XCO2 Measurements Derived from SCIAMACHY and GOSAT for Potentially Generating Global CO2 Maps with High Spatiotemporal Resolution," *PloS one*, vol. 9, no. 8, p. e105050, 2014.

[37] H. Watanabe, K. Hayashi, T. Saeki, S. Maksyutov, I. Nasuno, Y. Shimono, Y. Hirose, K. Takaichi, S. Kanekon, M. Ajiro *et al.*, "Global Mapping of Greenhouse Gases Retrieved from GOSAT Level-2 Products by Using a Kriging Method," *International Journal of Remote Sensing*, vol. 36, no. 6, pp. 1509–1528, 2015.

[38] Y. Liu, X. Wang, M. Guo, and H. Tani, "Mapping the FTS SWIR L2 Product of XCO2 and XCH4 Data from the GOSAT by the Kriging Method - A Case Study in East Asia," *International Journal of Remote Sensing*, vol. 33, no. 10, pp. 3004–3025, 2012.

[39] H. Nguyen, N. Cressie, and A. Braverman, "Multivariate Spatial Data Fusion for Very Large Remote Sensing Datasets," *Remote Sensing*, vol. 9, no. 2, p. 142, 2017.

[40] S. Bhattacharjee and S. K. Ghosh, "Time-Series Augmentation of Semantic Kriging for the Prediction of Meteorological Parameters," in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2015)*, July 2015, pp. 4562–4565.

[41] S. Bhattacharjee and S. K. Ghosh, "Exploring Spatial Dependency of Meteorological Attributes for Multivariate Analysis: A Granger Causality Test Approach," in *2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR)*. IEEE, 2015, pp. 1–6.

[42] G. Keppel-Aleks, P. Wennberg, C. O'Dell, and D. Wunch, "Towards Constraints on Fossil Fuel Emissions from Total Column Carbon Dioxide," *Atmospheric Chemistry and Physics*, vol. 13, no. 8, pp. 4349–4357, 2013.

[43] J. Heymann, M. Reuter, M. Buchwitz, O. Schneising, H. Bovensmann, J. Burrows, S. Massart, J. Kaiser, and D. Crisp, "CO2 Emission of Indonesian Fires in 2015 Estimated from Satellite-derived Atmospheric CO2 Concentrations," *Geophysical Research Letters*, vol. 44, no. 3, pp. 1537–1544, 2017.

[44] J. K. Hedelius, J. Liu, T. Oda, S. Maksyutov, C. M. Roehl, L. T. Iraci, J. R. Podolske, P. W. Hillyard, J. Liang, K. R. Gurney *et al.*, "Southern California Megacity CO2, CH4, and CO Flux Estimates using Ground-and Space-Based Remote Sensing and a Lagrangian Model," *Atmospheric Chemistry and Physics*, vol. 18, no. 22, pp. 16 271–16 291, 2018.

[45] R. Nassar, T. G. Hill, C. A. McLinden, D. Wunch, D. B. Jones, and D. Crisp, "Quantifying CO2 Emissions from Individual Power Plants from Space," *Geophysical Research Letters*, vol. 44, no. 19, pp. 10–045, 2017.

[46] J. Hakkarainen, I. Ialongo, and J. Tamminen, "Direct Space-based Observations of Anthropogenic CO2 Emission Areas from OCO-2," *Geophysical Research Letters*, vol. 43, no. 21, 2016.

[47] S. Bhattacharjee and S. K. Ghosh, "Performance Evaluation of Semantic Kriging: A Euclidean Vector Analysis Approach," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 6, pp. 1185–1189, 2015.

[48] N. Mendiratta, R. S. Kumar, and K. S. Rao, "Standards for Bio-geo Database," Natural Resources Data Management System (NRDMS) Division, New Delhi, India, Tech. Rep. 1, 2008.

[49] S. Bhattacharjee and S. K. Ghosh, "Measuring Semantic Similarity Between Land-cover Classes for Spatial Analysis: An Ontology Hierarchy Exploration Approach," *Innovations in Systems and Software Engineering*, vol. 12, no. 3, pp. 193–200, 2016.

[50] S. Bhattacharjee and S. K. Ghosh, "Measurement of Semantic Similarity: A Concept Hierarchy Based Approach," in *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics*. Springer India, 2015, pp. 407–416.

[51] S. Bhattacharjee, "Prediction of Meteorological Parameters: A Semantic Kriging Approach," in *23rd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL 2015) PhD Symposium*. ACM, November 2015, p. 1.

[52] S. Bhattacharjee and J. Chen, "Semantic kriging (SemK) example code." [Online]. Available: https://www.esm.ei.tum.de/index.php?id=105

[53] N. Cressie, "Statistics for Spatial Data," *Terra Nova*, vol. 4, no. 5, pp. 613–617, 1992.

[54] E. Pardo-Igúzquiza and P. M. Atkinson, "Modelling the Semivariograms and Cross-semivariograms Required in Downscaling Cokriging by Numerical Convolution-deconvolution," *Computers & Geosciences*, vol. 33, no. 10, pp. 1273–1284, 2007.

[55] M. Gunson and A. Eldering, "OCO-2 Level 2 Bias-Corrected XCO2 and Other Select Fields from the Full-physics Retrieval Aggregated as Daily Files, Retrospective Processing V9r, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC)," Accessed on: May 3, 2019. [Online]. Available: http://dx.doi.org/10.5067/W8QGIYNKS3JC

[56] "OCO-2 Data Center," Accessed on:March 26, 2020. [Online]. Available: https://oco.jpl.nasa.gov/oco-2-data-center/

[57] "OCO-2 Product Info," Accessed on: November 20, 2019. [Online]. Available: https://ocov2.jpl.nasa.gov/product-info/

[58] M. Kiel, C. W. O'Dell, B. Fisher, A. Eldering, R. Nassar, C. G. MacDonald, and P. O. Wennberg, "How Bias Correction Goes Wrong: Measurement of XCO2 Affected by Erroneous Surface Pressure Estimates," *Atmospheric Measurement Techniques*, vol. 12, no. 4, 2019.

[59] "Orbiting Carbon Observatory2 (OCO-2) Data Product User's Guide, Operational L1 and L2 Data Versions 8 and Lite File Version 9," Accessed on: June 5, 2019. [Online]. Available: https://docserver. gesdisc.eosdis.nasa.gov/public/project/OCO/OCO2_DUG.V9.pdf

[60] W. Li, P. Ciais, N. MacBean, S. Peng, P. Defourny, and S. Bontemps, "Major Forest Changes and Land Cover transitions based on Plant

Functional Types Derived from the ESA CCI Land Cover Product," *International Journal of Applied Earth Observation and Geoinformation*, vol. 47, pp. 30–39, 2016.

[61] "ODIAC Fossil Fuel Emission Dataset," Accessed on: June 10, 2019. [Online]. Available: http://db.cger.nies.go.jp/dataset/ODIAC/DL_odiac2018.html

[62] R. Nassar, L. Napier-Linton, K. R. Gurney, R. J. Andres, T. Oda, F. R. Vogel, and F. Deng, "Improving the Temporal and Spatial Distribution of CO2 Emissions from Global Fossil Fuel Emission Data Sets," *Journal of Geophysical Research: Atmospheres*, vol. 118, no. 2, pp. 917–933, 2013.

[63] "EDGAR - Emissions Database for Global Atmospheric Research," Accessed on: September 7, 2018. [Online]. Available: http://edgar.jrc.ec.europa.eu/

[64] D. Wunch, G. C. Toon, V. Sherlock, N. M. Deutscher, C. Liu, D. G. Feist, and P. O. Wennberg, "The Total Carbon Column Observing Network's GGG2014 Data Version," TCCON Data Archive, hosted by CaltechDATA, California Institute of Technology, Pasadena, CA, Tech. Rep., 2015. [Online]. Available: http://dx.doi.org/10.14291/tccon.ggg2014.documentation.R0/1221662

[65] "TCCON Data Description," Accessed on: September 7, 2018. [Online]. Available: https://tccon-wiki.caltech.edu/Network_Policy/Data_Use_Policy/Data_Description

[66] "TCCON Data Archive," Accessed on: September 7, 2018. [Online]. Available: https://tccondata.org/

[67] P. O. Wennberg, C. Roehl, J.-F. Blavier, D. Wunch, J. Landeros, and N. Allen, "TCCON Data from Jet Propulsion Laboratory (US), 2011, Release GGG2014R1," TCCON Data Archive, Hosted by CaltechDATA, Pasadena, California, 2016. [Online]. Available: https://tccondata.org

[68] P. O. Wennberg, D. Wunch, C. Roehl, J.-F. Blavier, G. C. Toon, and N. Allen, "TCCON Data from Caltech (US), Release GGG2014R1," TCCON Data Archive, Hosted by CaltechDATA, Pasadena, California, 2014. [Online]. Available: https://tccondata.org

[69] P. O. Wennberg, D. Wunch, C. Roehl, J.-F. Blavier, G. C. Toon, N. Allen, P. Dowell, K. Teske, C. Martin, and J. Martin., "TCCON Data from Lamont (US), Release GGG2014R1," TCCON Data Archive, Hosted by CaltechDATA, Pasadena, California, 2016. [Online]. Available: https://tccondata.org

[70] S. Kawakami, H. Ohyama, K. Arai, H. Okumura, C. Taura, T. Fukamachi, and M. Sakashita, "TCCON Data from Saga (JP), Release GGG2014R0," TCCON Data Archive, Hosted by CaltechDATA, Pasadena, California, 2014. [Online]. Available: https://tccondata.org

[71] F. Hase, T. Blumenstock, S. Dohe, J. Gross, and M. Kiel, "TCCON Data from Karlsruhe (DE), Release GGG2014R1," TCCON Data Archive, Hosted by CaltechDATA, Pasadena, California, 2014. [Online]. Available: https://tccondata.org

[72] I. Morino, T. Matsuzaki, and M. Horikawa, "TCCON data from Tsukuba (JP), 125HR, Release GGG2014.R2," TCCON Data Archive, Hosted by CaltechDATA, Pasadena, California, 2018. [Online]. Available: https://data.caltech.edu/records/958

[73] R. Sussmann and M. Rettinger, "TCCON Data from Garmisch (DE), Release GGG2014R0," TCCON Data Archive, Hosted by CaltechDATA, Pasadena, California, 2014. [Online]. Available: https://tccondata.org

[74] V. Sherlock, B. Connor, J. Robinson, H. Shiona, D. Smale, and D. F. Pollard, "TCCON data from Lauder (NZ), 125HR, Release GGG2014.R0," TCCON Data Archive, Hosted by CaltechDATA, Pasadena, California, 2014. [Online]. Available: https://data.caltech.edu/records/281

[75] T. Warneke, J. Messerschmidt, J. Notholt, C. Weinzierl, N. M. Deutscher, C. Petri, P. Grupe, C. Vuillemin, F. Truong, M. Schmidt, M. Ramonet, and E. Parmentier, "TCCON Data from Orléans (FR), Release GGG2014R0," TCCON Data Archive, Hosted by CaltechDATA, Pasadena, California, 2014. [Online]. Available: https://tccondata.org

[76] "TCCON Sites," Accessed on: November 25, 2019. [Online]. Available: https://tccon-wiki.caltech.edu/Sites/

[77] P. Mohan, X. Zhou, and S. Shekhar, "Quantifying Resolution Sensitivity of Spatial Autocorrelation: A Resolution Correlogram Approach," in *International Conference on Geographic Information Science*. Springer, 2012, pp. 132–145.

[78] S. Bhattacharjee, S. K. Ghosh, and J. Chen, *Semantic Kriging for Spatio-temporal Prediction*. Springer Nature Singapore, 2019.

[79] B. V. Srinivasan, R. Duraiswami, and R. Murtugudde, "Efficient Kriging for Real-time Spatio-Temporal Interpolation," in *Proceedings of the 20th Conference on Probability and Statistics in the Atmospheric Sciences*. American Meteorological Society Atlanta GA, 2010, pp. 228–235.

[80] B. Gräler, "Cokriging and Indicator Kriging," in *Course on Spatial Copulas*, 2011, pp. 2.1–2.20, Accessed on: February 26, 2020.

[Online]. Available: http://www.graeler.org/copulaIntro/02_Co-Kriging_Indicator-Kriging.pdf

[81] S. Bhattacharjee, J. Chen, L. Jindun, and X. Zhao, "Kriging-based Mapping of Space-borne CO2 Measurements by Combining Emission Inventory and Atmospheric Transport Modeling," in *Geophysical Research Abstracts*, vol. 22, 2020.

[82] B. Fasoli, J. C. Lin, D. R. Bowling, L. Mitchell, and D. Mendoza, "Simulating atmospheric tracer concentrations for spatially distributed receptors: updates to the Stochastic Time-Inverted Lagrangian Transport model's R interface (STILT-R version 2)," *Geoscientific Model Development*, vol. 11, no. 7, 2018.

[83] T. Nehrkorn, J. Eluszkiewicz, S. C. Wofsy, J. C. Lin, C. Gerbig, M. Longo, and S. Freitas, "Coupled weather research and forecasting–stochastic time-inverted lagrangian transport (WRF–STILT) model," *Meteorology and Atmospheric Physics*, vol. 107, no. 1-2, pp. 51–64, 2010.

[84] C. Gerbig, J. Lin, S. Wofsy, B. Daube, A. Andrews, B. Stephens, P. Bakwin, and C. Grainger, "Toward constraining regional-scale fluxes of CO2 with atmospheric observations over a continent: 2. Analysis of COBRA data using a receptor-oriented framework," *Journal of Geophysical Research: Atmospheres*, vol. 108, no. D24, 2003.

[85] M. Frey, M. K. Sha, F. Hase, M. Kiel, T. Blumenstock, R. Harig, G. Surawicz, N. M. Deutscher, K. Shiomi, J. E. Franklin *et al.*, "Building the COllaborative Carbon Column Observing Network (COCCON): Long-term Stability and Ensemble Performance of the EM27/SUN Fourier Transform Spectrometer," *Atmospheric Measurement Techniques*, vol. 12, no. 3, pp. 1513–1530, 2019.

[86] J. Chen, C. Viatte, J. K. Hedelius, T. Jones, J. E. Franklin, H. Parker, E. W. Gottlieb, P. O. Wennberg, M. K. Dubey, and S. C. Wofsy, "Differential Column Measurements Using Compact Solar-tracking Spectrometers," *Atmospheric Chemistry and Physics*, vol. 16, no. 13, pp. 8479–8498, 2016.

[87] L. Heinle and J. Chen, "Automated Enclosure and Protection System for Compact Solar-Tracking Spectrometers," *Atmospheric Measurement Techniques Discussions*, vol. 2017, pp. 1–21, 2017. [Online]. Available: http://dx.doi.org/10.5194/amt-2017-292

[88] F. Dietrich, J. Chen, B. Reger, J. Matzke, A. Forstmaier, X. Bi, A. Luther, M. Frey, F. Hase, and A. Butz, "First fully-automated differential column network for measuring GHG emissions tested in Munich," in *Geophysical Research Abstracts*, vol. 21, 2019, pp. EGU2019–13 327.

Shrutilipi Bhattacharjee (Member, IEEE) received the Ph.D. degree from IIT Kharagpur, Kharagpur, India, in 2016. She has worked as a Fulbright-Nehru Doctoral Research Fellow with the Department of Computer Science, University of Minnesota, Minneapolis, MN, USA, from July 2015 to January 2016. She is working as a Postdoctoral Fellow with the Environmental Sensing and Modeling (ESM) Group, Department of Electrical and Computer Engineering, Technical University of Munich (TUM), Munich, Germany. She has published nine peer-reviewed journal articles, twenty one conference papers, one book, and one book chapter. Her research interests include analyzing remotely sensed environmental parameters, semantic analysis, spatial statistics, and data mining.

Jia Chen (Member, IEEE) received the Dipl.-Ing. degree from Karlsruhe Institute of Technology, Karlsruhe, Germany, in 2006, and the Dr.-Ing. degree from Technical University of Munich (TUM), Munich, Germany, in 2011.

From 2011 to 2015, she was a Post-Doctoral Fellow with the Environmental Science and Engineering Department, Harvard University, Cambridge, MA, USA, where she holds an associate position at the Department of Earth and Planetary Sciences. She is also a Co-Investigator of the NASA Sponsored Project: Validation and Application of OCO-2 data in the Northeastern United States. She is head of the Environmental Sensing and Modeling (ESM) Group, Department of Electrical and Computer Engineering, Technical University of Munich (TUM). She published 2 books, 44 peer-reviewed journal articles and 106 conference papers. Her research interests include urban greenhouse gas monitoring, quantifying greenhouse gas emissions, establishing regional sensor networks, sensor development, and atmospheric transport modeling.