LETTER • OPEN ACCESS

Evaluating nighttime lights and population distribution as proxies for mapping anthropogenic CO_2 emission in Vietnam, Cambodia and Laos

To cite this article: Andrea E Gaughan *et al* 2019 *Environ. Res. Commun.* **1** 091006

View the article online for updates and enhancements.

Recent citations

Environmental Research Communications

LETTER

Evaluating nighttime lights and population distribution as proxies for mapping anthropogenic CO$_2$ emission in Vietnam, Cambodia and Laos

Andrea E Gaughan[1,2] , Tomohiro Oda[4], Alessandro Sorichetta[2], Forrest R Stevens[1,2], Maksym Bondarenko[2], Rostyslav Bun[5,6], Laura Krauser[1], Greg Yetman[3] and Son V Nghiem[7]

1 University of Louisville, Department of Geography and Geosciences, Louisville, KY, United States of America
2 WorldPop, School of Geography and Environmental Science, University of Southampton, United Kingdom
3 CIESIN, Columbia University, New York, NY, United States of America
4 Universities Space Research Association, Columbia, MD, USA/NASA, Goddard Space Flight Center, Greenbelt, MD, United States of America
5 Lviv Polytechnic National University, Lviv, Ukraine
6 WSB University, Dabrowa Gornicza, Poland
7 Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, United States of America

E-mail: ae.gaughan@louisville.edu

Abstract

Tracking spatiotemporal changes in GHG emissions is key to successful implementation of the United Nations Framework Convention on Climate Change (UNFCCC). And while emission inventories often provide a robust tool to track emission trends at the country level, subnational emission estimates are often not reported or reports vary in robustness as the estimates are often dependent on the spatial modeling approach and ancillary data used to disaggregate the emission inventories. Assessing the errors and uncertainties of the subnational emission estimates is fundamentally challenging due to the lack of physical measurements at the subnational level. To begin addressing the current performance of modeled gridded CO$_2$ emissions, this study compares two common proxies used to disaggregate CO$_2$ emission estimates. We use a known gridded CO$_2$ model based on satellite-observed nighttime light (NTL) data (Open Source Data Inventory for Anthropogenic CO$_2$, ODIAC) and a gridded population dataset driven by a set of ancillary geospatial data. We examine the association at multiple spatial scales of these two datasets for three countries in Southeast Asia: Vietnam, Cambodia and Laos and characterize the spatiotemporal similarities and differences for 2000, 2005, and 2010. We specifically highlight areas of potential uncertainty in the ODIAC model, which relies on the single use of NTL data for disaggregation of the non-point emissions estimates. Results show, over time, how a NTL-based emissions disaggregation tends to concentrate CO$_2$ estimates in different ways than population-based estimates at the subnational level. We discuss important considerations in the disconnect between the two modeled datasets and argue that the spatial differences between data products can be useful to identify areas affected by the errors and uncertainties associated with the NTL-based downscaling in a region with uneven urbanization rates.

1. Introduction

Keeping track of spatiotemporal changes of the greenhouse gases (GHG) emissions is key to the successful implementation of the United Nations Framework Convention on Climate Change (UNFCCC) as monitoring emissions directly informs international climate change policy initiatives (Raupach *et al* 2007, Figueres *et al*

2018). While the Paris Climate Agreement recognized the importance of the climate mitigation actions at subnational levels (e.g. cities, private sectors), the emission inventories reported by countries, in the current inventory framework, do not fully allow us to monitor those efforts. This is because the current emission inventory framework is designed to quantify GHG emissions at the national scale with no current, globally consistent requirement for state or city-level emission inventory reporting.

Spatially-explicit emission estimates have been mainly developed for research application purposes, such as atmospheric modeling (e.g. (Andres *et al* 1996, Kurokawa *et al* 2013, Oda *et al* 2018, Janssens-Maenhout *et al* 2019). The spatial extent of emissions is often estimated and/or determined via spatial disaggregation of the emissions estimates made at larger, aggregated scales (e.g. countries and regions) and this is done using geospatial information. The first global CO_2 map was developed by Anders *et al* (1996) at the Carbon Dioxide Information Analysis Center, (CDIAC) at the Oak Ridge National Laboratory (ORNL). Andres *et al* (1996) used a global population density map from 1984 (Fung *et al* 1991) to disaggregate national level emission estimates to a $1° \times 1°$ grid cell. The approach comes with an inherent assumption of a good correlation between population and fossil fuel carbon dioxide (FFCO$_2$) emissions at a large aggregated spatial scale (e.g. state level). More recently, Anders *et al* (2016) updated their approach to include two additional gridded population products to cover more recent time periods (Andres *et al* 2016). Those two products are the Gridded Population of the World v3 (Center for International Earth Science Information Network (CIESIN) and (CIAT) 2005) and LandScan (Dobson *et al* 2000).

Additionally, in response to the strong need for a high-resolution CO_2 emission map for high-resolution atmospheric modeling and satellite CO_2 analyses, Oda and Maksyutov (2011) proposed an improved $1 km \times 1 km$ gridded emission model. Known as the Open Source Data Inventory for Anthropogenic CO_2 (ODIAC), their approach is based on a disaggregation of country level fuel based estimates, such as CDIAC estimates, using the combination of point source information and satellite-observed nighttime lights (NTL). NTL has been identified as a good indicator of human settlement and the intensity of human activities (e.g. (Elvidge *et al* 1999). Other studies have noted limitations for using NTL data as a 1:1 proxy for human population distribution, especially in low-lit, less developed regions of the world (Sutton *et al* 2001, Huang *et al* 2014, Pandey *et al* 2017). However, Oda and Maksyutov (2011) separately map point source emissions that are difficult to approximate by the spatial distribution of NTL and found that NTL is a useful remote sensing tool for disaggregating the remaining CO_2 emission estimates in timely, updatable manner at a global scale. Because of the high-resolution, updated emissions, ODIAC has been extensively used in atmospheric carbon budget studies across different scales from global (Takagi *et al* 2011, Maksyutov *et al* 2013, Feng *et al* 2017, Crowell *et al* 2019) to urban scales (Ganshin *et al* 2012, Oda *et al* 2012, Brioude *et al* 2013, Lauvaux *et al* 2016, Oda *et al* 2017, Wu *et al* 2018, Reuter *et al* 2019)

While emissions disaggregation using a wide variety of geospatial information is becoming more accessible, errors and uncertainties associated with the resulting emission fields remain fully unquantified. The evaluation of actual errors and uncertainties associated with gridded emissions is challenging fundamentally due to the lack of physical measurements (Andres *et al* 2016, Oda *et al* 2018, 2019). The research community has been studying the use of atmospheric measurements to objectively evaluate emission estimates from reported inventories (top-down analysis, e.g. (Vogel *et al* 2007, Lauvaux *et al* 2016, Nassar *et al* 2017)). However, the ability of constraining emissions highly depends on the atmospheric observation available and is limited to a small area (e.g. city) and particular sources (e.g. power plant). Thus, a common approach has been comparing inventories and using the differences as a proxy for errors and uncertainties in gridded emissions (e.g. (Gately and Hutyra 2017, Hutchins *et al* 2017, Oda *et al* 2018, 2019). Though such comparison does not provide any objective measure regarding the performance of the data used for emission downscaling, it allows for characterizing the differences that are unique to respective inventories (Oda *et al* 2019).

In this study, we evaluate the use of remotely-sensed, nighttime lights as an emission proxy by comparing non-point source ODIAC emissions to population data which are the direct proxy for the intensity of human activities (thus, CO_2 emissions). We are most interested in comparing gridded CO_2 emissions and spatially identifying the error and uncertainty of CO_2 emissions disaggregation techniques as a first step towards improving underlying data sets that inform not only the atmospheric modeling community but also CO_2 emission estimates at subnational levels. While NTL serves as an excellent proxy for developed countries (e.g. (Oda and Maksyutov 2011)), NTL is thought to perform poorly as a proxy for CO_2 emission disaggregation in low and middle income countries where the spatial pattern of emission distribution, as associated with population, is not directly proportional to NTL data. Raupach *et al* (2010) demonstrated that by showing correlations between NTL and population, with population acting as a proxy for CO_2 emissions (Raupach *et al* 2010). Thus, we use the NTL-disaggregated, non-point source CO_2 emissions from the ODIAC model to compare with a gridded population dataset driven by a set of ancillary geospatial data for 2000, 2005, and 2010. The objective is to examine how spatiotemporal changes in gridded CO_2 emission estimates based on residential population distributions compare with the changes in gridded CO_2 emission estimates informed only by NTL

data in order to identify areas of uncertainty in the ODIAC model. We examine the difference and similarities of these two datasets at multiple spatial resolutions for three emerging countries in Southeast Asia: Vietnam, Cambodia and Laos. This region represents an area where the NTL performance is poor and emission estimates will increase over time (Schneider *et al* 2015, Fulton *et al* 2017, Pandey *et al* 2017).The three countries also provide an ideal landscape of varying development patterns associated with different brightness intensities and population densities for highlighting areas of uncertainty in the ODIAC dataset.

Population data (POP) are traditionally based on censuses that are linked to areal administrative units of varying sizes, thus the subnational distributions are more constrained than CO_2 data that is typically based on a country level emission disaggregation (i.e. there are no constraints at subnational level). Some studies have explored the ways to combine the use of NTL and POP data to compensate for the weakness of relying only on NTL for emission disaggregation (Ghosh *et al* 2010, Rayner *et al* 2010). However, the literature lacks an evaluation of the combined use of these two highly-correlated datasets to characterize potential errors and uncertainties associated with the use of the proxy data, such as emission representation errors. We expect that the differences found can be largely explained by the impact of the additional covariates and modeling in the population data, which we consider as potential additions for better CO_2 emissions mapping.

2. Data and methods

2.1. CO_2 emission dataset
The ODIAC is a global high resolution (1 × 1 km) fossil fuel CO_2 emission data product (Oda and Maksyutov 2011, Oda *et al* 2018). The ODIAC is based on spatial disaggregation of CO_2 emission estimates made by the Carbon Dioxide Information Analysis Center (CDIAC) at the Oak Ridge National Lab (ORNL) (Boden *et al* 2016). CDIAC emissions are estimated by fuel type (solid, gas, and liquid fuels, bunker fuel, and gas flares) plus cement production, rather than the emission sector that is often used for the national inventory compilation (Marland and Rotty 1984). The ODIAC spatial disaggregation is done in two steps. First, emissions from point sources (mainly power plants) are estimated and mapped using the power plant emission estimates and geolocation taken from a global power plant database. The rest of the emissions (country total minus point source emissions), which we refer to a non-point source emissions, are distributed using the spatial distribution of satellite-observed nightlights (NTL) intensities (Oda and Maksyutov 2011, Oda *et al* 2018). Non-point source emissions are disaggregated to a 1 km × 1 km spatial resolution using Defense Meteorological Satellite Program (DMSP) calibrated radiance NTL data, with mitigated saturation effect, developed by NOAA's Earth Observation Group (Oda *et al* 2010). The calibrated radiance NTL data are a merged product of the regular DMSP NTL product and benefits from reduced gain observations (Ziskin *et al* 2012). Oda *et al* (2010) show an improved spatial emissions distribution from the original publication by Oda and Maksyutov (2011) due to the use of the calibrated radiance data.

Globally, this emission disaggregation is done for 65 individual countries and 5 aggregated geographical region groups (Oda and Maksyutov 2011). Vietnam, Cambodia and Laos are a part of the Asia Pacific geographical region group in the ODIAC country and region categorization. Since the country emissions for those three countries are once aggregated to the regional total before the emissions disaggregation, the country total emissions do not exactly match with the original CDIAC estimates. However, the geographic aggregation for Vietnam, Cambodia and Laos is necessary for an emission modeling framework that aims at quick, timely updates using the latest fuel statistical data from companies such as BP. Further details of the ODIAC approach and methodology are described elsewhere (Oda and Maksyutov 2011, Oda *et al* 2018). In this study, we excluded point source emissions and only use non-point (diffuse, area) source emissions (total minus - point source emissions) and specific to the study region of interest, focusing on the remotely-sensed NTL, nonpoint source ODIAC emissions. The version of ODIAC used in this study (ODIAC 2017) covers 2000 to 2016 but we use data for the years 2000, 2005 and 2010. The data product is available from http://db.cger.nies.go.jp/dataset/ODIAC/ (Oda and Maksyutov n.d.).

2.2. Population dataset
The gridded population data for this study is from the WorldPop Project (www.worldpop.org 2018). The underlying method is one of the more advanced techniques for gridding population distributions which relies on a hybrid approach of using a statistical weighting layer for estimating population density constrained by a dasymetric redistribution of the administrative unit population counts (Stevens *et al* 2015). Such an approach compares favorably or outperforms other techniques for producing gridded population maps (Sorichetta *et al* 2015, Stevens *et al* 2015, Reed *et al* 2018) while recognizing results are dependent on the spatial fidelity and accuracy of the ancillary data and census data used in the model.

The statistical component of the WorldPop model involves training a random forest (RF) model (Breiman 2001) to create a population density layer that will be used to disaggregate the administrative unit-based population counts to a regular grid of fixed spatial resolution. The RF-model is a non-parametric, non-linear statistical machine-learning approach that combines a set of decision trees into an 'ensemble' learner of multiple trees for a stronger output prediction. The ensemble of models is combined with random sampling of both training observations (bagging) for individual trees and covariate selection during tree growth. The individual trees and their prediction estimates can then be validated against withheld observations (out-of-bag errors, OOB), which can then be used across all trees in the RF to assess both variable importance and prediction performance (Breiman 2001, Liaw and Wiener 2002, Strobl *et al* 2009). The OOB error represents a robust and unbiased measurement of the prediction accuracy of the RF model (Breiman 2001) and a reliable proxy of the accuracy of the final gridded population datasets produced using the RF-based approach (Sorichetta *et al* 2015, Stevens *et al* 2015).

Predictions of population density from the RF model at the pixel level are inherently biased because the model is parameterized at the administrative unit level. Therefore, the pixel level estimates of population density are used as weights to anchor a dasymetric redistribution approach (Mennis 2003) of administrative census counts redistributed to a regular grid of fixed spatial resolution (100 m). Full details on the combined machine-learning and dasymetric modeling technique is found in (Gaughan *et al* 2015, Sorichetta *et al* 2015, Stevens *et al* 2015).

The WorldPop model is parameterized for each year of interest using an aggregated set of administrative units (Laos $n = 10035$, Cambodia $n = 1621$, Vietnam $n = 688$) and by maintaining the same consistent boundary definitions over time (Gaughan *et al* 2016). By relying on census data to inform the modeling process, the gridded population outputs effectively represent residential population counts which is in contrast to data such as Landscan which models ambient population counts (Dobson *et al* 2000). In addition, the 100 m gridded population datasets are eventually spatially aggregated (by summing population per grid cell) and coregistered to match the 1 km spatial resolution outputs from the ODIAC model. Population grids were done for 2000, 2005 and 2010. In terms of the RF model parameterization, subnational population counts for 2000, 2005 and 2010 were either extracted, interpolated, and extrapolated using two census dates (i.e., t_0 and t_1) for all three countries (i.e., 2005 and 2010 for Laos, 1998 and 2008 for Cambodia, 1999 and 2009 for Vietnam). This was done by calculating the growth rate of each administrative unit within each country and applying it to the corresponding population counts (Doxsey-Whitfield *et al* 2015). For each administrative unit, the corresponding exponential growth rate (r) was calculated using the following formula:

$$r = \frac{1}{t} \ln\left(\frac{P_1}{P_0}\right), \tag{1}$$

where r is the growth rate of a given administrative unit between t_0 and t_1, P_0 is its total population at time t_0, P_1 is its total population at time t_1, and t represents the number of years between t_0 and t_1.
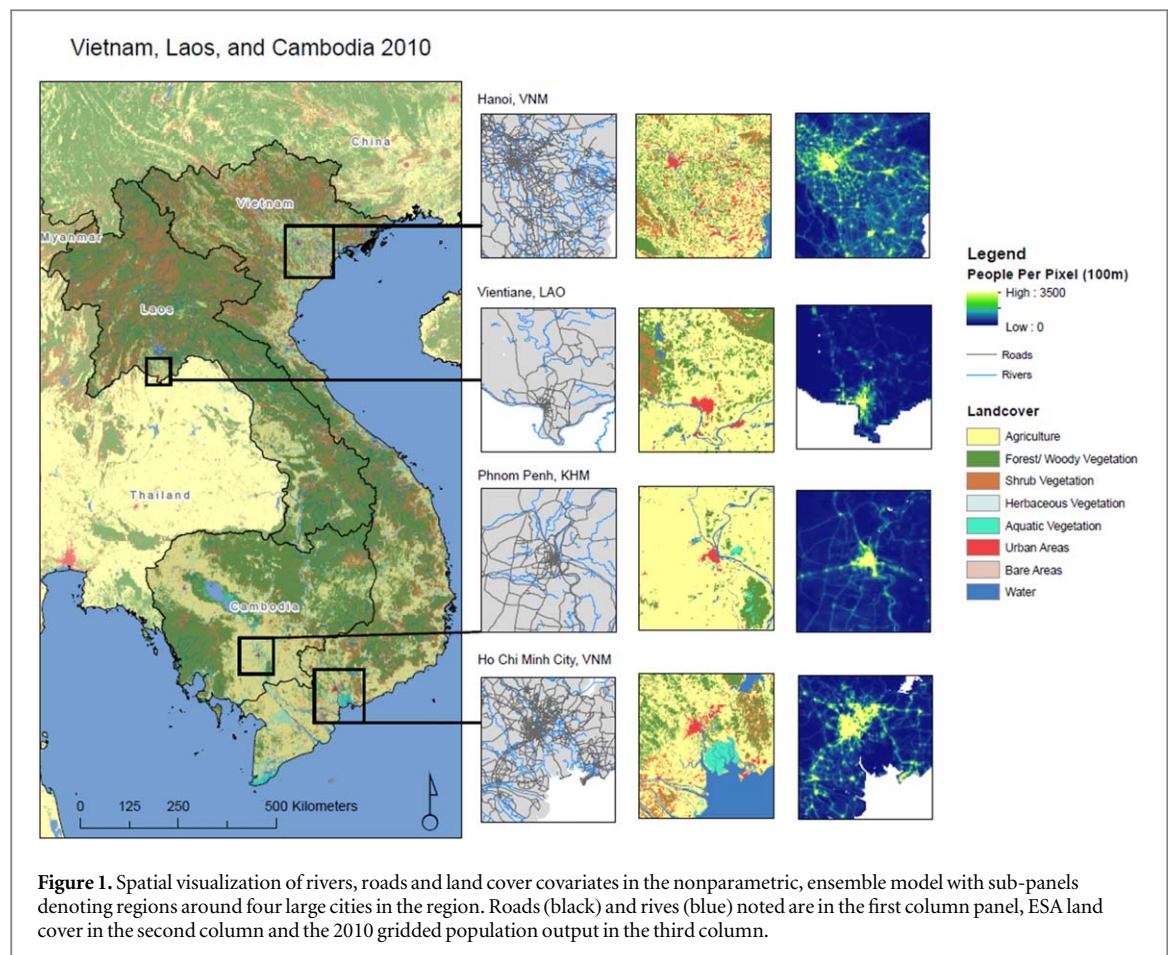
Geospatial covariates used as input to the RF have been standardized, harmonized, and co-registered, to the CIESIN Gridded Population of the World v4 archive of administrative-boundaries (http://sedac.ciesin.columbia.edu/downloads/docs/gpw-v4/gpw-v4-country-level-summary-rev10.xlsx) across all three countries of interest. We followed the approach outlined in Gaughan *et al* (2016) and used both temporally-invariant and temporally-explicit covariates (table 1) while excluding NTL data. Temporally-invariant data include topography and slope, average annual precipitation and temperature (representative of the current conditions), presence of main roads and their intersections, waterways, water bodies, and coastlines. Temporally-explicit data, for 2000, 2005 and 2010, include the European Space Agency (ESA) Climate Change Initiative (CCI) land cover layers and presence of protected areas (WDPA 2017). For all temporally-invariant categorical data we produced the corresponding 'distance-to' covariate dataset, while for the temporally-explicit categorical data we calculated the distance-to-edge covariates, where distances inside the edge are negative and distances outside the edge are positive. See (Lloyd *et al* n.d.) for the production methodology for how the various covariate datasets were assembled and harmonized. Figure 1 shows the entire study region with land cover, roads, and rivers displayed along with panels of main city area gridded population patterns for 2010.

2.3. Relationships between remotely-sensed nightlights and population

To determine the agreement between remotely-sensed NTL and gridded population, we examine the relationship between non-point source CO_2 emissions and population at multiple administrative levels over the three time points. We then compare the NTL-disaggregated CO_2 emission estimates (i.e. the ODIAC model) to a disaggregation of the CO_2 emission estimates by the gridded population (i.e. WorldPop model) for each year. By comparing disaggregations based on NTL versus a simple, per capita disaggregation we can directly interrogate how disassociation between nightlights and population estimates might affect the utility of either approach. Informing the underlying gridded population model is a set of geospatial covariates (table 1) that provide some

Table 1. Data description for geospatial covariates used in the population model.

Name	Source	Data type and nominal spatial resolution	Data product	Acquisition year
Viewfinder Panoramas	De Ferranti, J, http://viewfinderpanoramas.org/	Continuous raster, 3 arcseconds (~100 m at the Equator)	Topography, Slope	~2000
ESA CCI Land Cover Maps – V2.0.7	European Space Adency (ESA) & Université Catholique De Louvain (UCL)	Categorical raster, 10 arcseconds (~300 m at the Equator)	Land cover classes	2000–2015
Open Street Map	OpenStreetMap Foundation (OSMF) & Contributors	Categorical vector	Main roads, main road intersections and waterways	2016
ESA CCI WB v4.0	European Space Agency (ESA)	Categorical binary raster, 30 arcseconds (~150 m at the Equator)	Water bodies	2000–2012
WorldClim 2.0	Fick, S.E. and R.J. Hijmans	Continuous rasters, 30 arcseconds (~1 km at the Equator)	Mean annual temperature and precipitation	1970–2000
World Database of Protected Areas (WDPA)	UNEP-WCMC	Vector	Terrestrial and marine protected areas	2000–2014
Global Population of the World (GPWv4) Coastlines	CIESIN, Gridded Population of the World v4	Vector	Protected areas	

Figure 1. Spatial visualization of rivers, roads and land cover covariates in the nonparametric, ensemble model with sub-panels denoting regions around four large cities in the region. Roads (black) and rives (blue) noted are in the first column panel, ESA land cover in the second column and the 2010 gridded population output in the third column.

insight as to the disaggregation of population counts, over space and time, making the population-driven disaggregation approach useful to assess similarities and differences in how NTL, as a singular covariate, redistributes the CO_2 emission estimates.
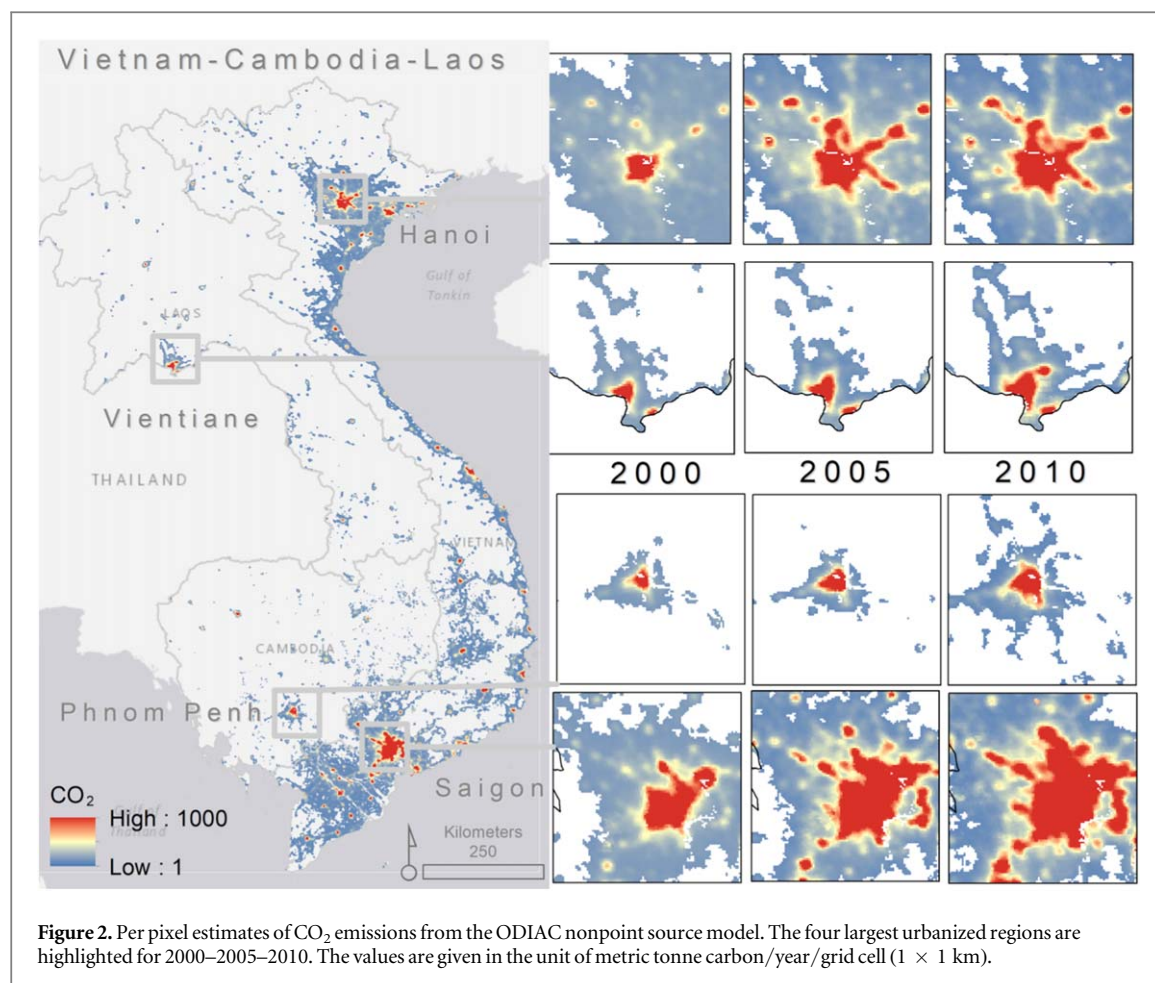
To disaggregate by population, we summed the total population count to the national level and created a population proportion layer for each year (e.g. 2010 people per pixel/total population by country). Next, we multiplied the proportion of population through by the total CO_2 by country for a given year. We did this for the three years of interest (2000, 2005, and 2010) resulting CO_2 emissions disaggregated by the population model described in section 2.2. We present different metrics to highlight differences in NTL-driven versus population-driven model outputs and identify spatial distributions related to the underlying data informing respective CO_2 models for each year. In addition, as done by previous studies (Hogue *et al* 2016), we examine the agreement of the two CO_2 emission maps at different spatial resolutions.

3. Results

3.1. The NTL model (ODIAC nonpoint source) and the POP model (WorldPop) for 2000, 2005, and 2010

The NTL model indicates expanding source areas for all three countries in a spatially-explicit manner from the 2000–2005–2010 period (figure 2). Vietnam, which is a more urbanized country, shows greater spatial distribution in the estimated CO_2 emissions as the number of lit grid cells in Vietnam are proportionally higher than in Laos and Cambodia.

Similar to the NTL maps, the annual gridded population data maps show a general increase in population counts over time (figure 3). The percentage of variance explained by the RF model and the prediction error associated to it are equal to 79% and 0.72, respectively, for all three years. The relative importance of a covariate in the RF model is captured by percent increase in mean squared error (%MSE). This means, in the ensemble model prediction phase, that the lower the %MSE, the more important a given covariate will be relative to other covariates when randomly selected and included in a node prediction. Results from the model in this study show that distance to roads and rivers are important to the model in terms of random effects for all three years (figure 1). Agriculture is also important in 2000 and 2005 but becomes slightly less so in 2010 (Supplemental (S1)). Full plot overviews for the model are provided in S1.
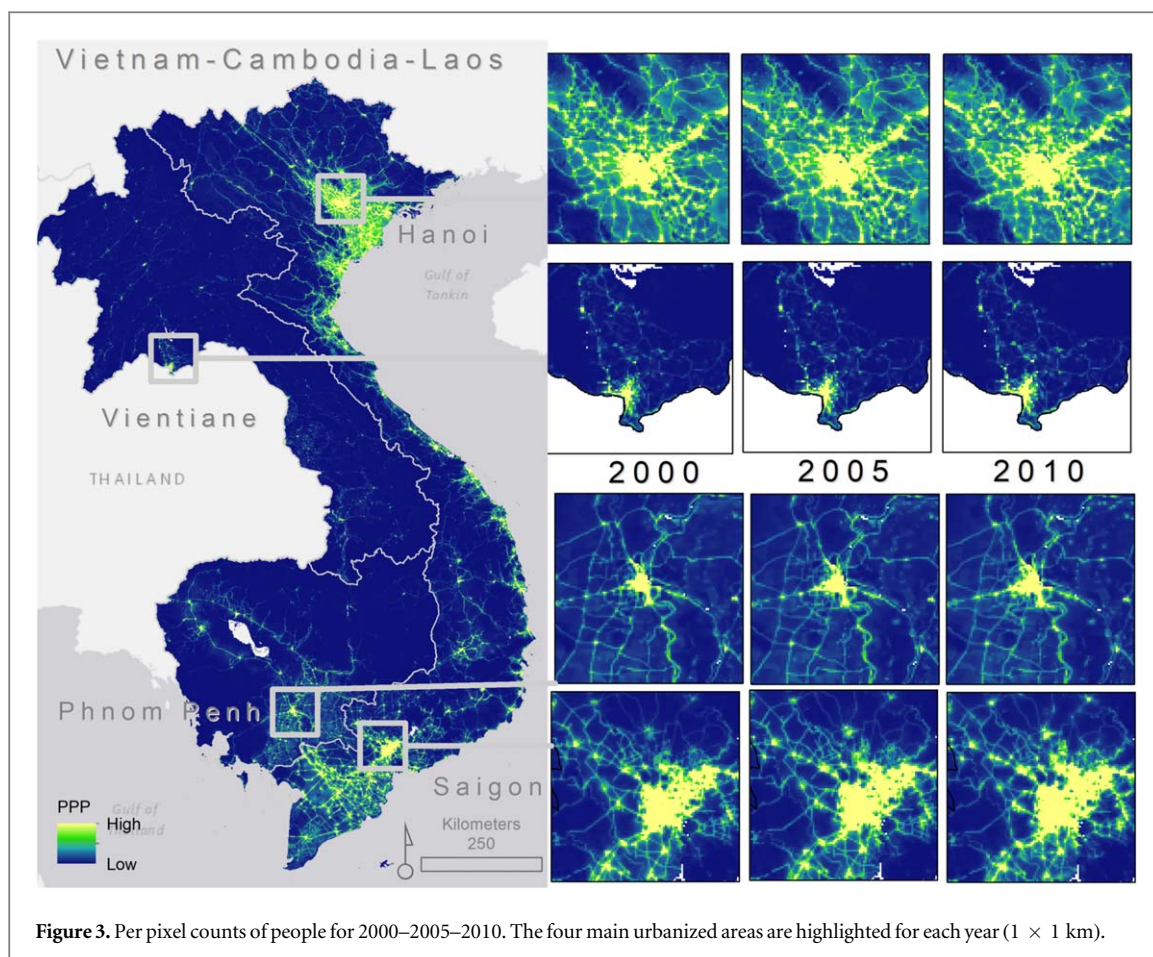
Figure 2. Per pixel estimates of CO_2 emissions from the ODIAC nonpoint source model. The four largest urbanized regions are highlighted for 2000–2005–2010. The values are given in the unit of metric tonne carbon/year/grid cell (1 × 1 km).

3.2. Characterizing spatiotemporal change between the NTL-driven model and POP-driven model for CO_2 emissions

Figure 4 highlights some important differences in per pixel estimates of CO_2 emissions from the NTL-driven (i.e. ODIAC) model versus the POP-driven model. The most noticeable difference between figures 4(a) and (b), is that the population-driven model allocates estimates of CO_2 emission to every grid cell while NTL does not. The gridded population modeling approach may estimate fractional population in a given grid cell (Stevens *et al* 2015) and will not predict zero in grid cells unless the administrative unit contains zero people. The type of approach performs favorably, and in many cases outperforms, other types of gridded population modeling approaches (Reed *et al* 2018), which may translate into better gridded residential CO_2 products. The NTL-driven model includes zero as a possible grid cell value (i.e. grey-shaded pixels in figure 4) and thus CO_2 emissions are more concentrated into certain areas for the NTL-driven model rather than spread across the entire study area.

When we difference the NTL-driven model estimates from the POP-driven model for disaggregating CO_2 emission estimates, the regional biases in the CO_2 emission from the NTL-driven model are more apparent. In figure 5, positive differences indicate grid cells where the NTL-driven model estimates higher emissions and negative differences indicate areas where population-based CO_2 emissions estimates are higher than NTL-based estimates. Thus, figure 5 illustrates that through time, nighttime lights-based emissions disaggregation tends to concentrate those estimates in different ways than population-based estimates. It is clear that population-driven model of emissions estimates, which excludes nighttime lights from ancillary data sources, tends to estimate higher emissions outside of the highly urbanized and developed areas of Vietnam, Cambodia, and Laos (shown by the orange-red tones in figure 5). These tend to be areas along major roads, intersections, and where population has been counted in census data but where nighttime lights may not be present. However, across 2000–2010 we see an expansion in the concentration of CO_2 source area estimates due to an increase in nighttime lighted areas most likely associated with development, as noted by the increase in 'purple' grid cells from 2000–2005–2010.

Figure 6 shows the level of disagreement between the two gridded CO_2 datasets by spatially aggregating at multiple spatial resolutions (i.e. by coarsening the two 1 km gridded datasets to 2, 3, 5, 10, 20, 50 and 100 km). As done in Oda *et al* (2019), we calculated the sum of the absolute differences at grid levels and defined the initial difference at 1 km as 100%. The results provide a means to assess how quickly the various datasets for each
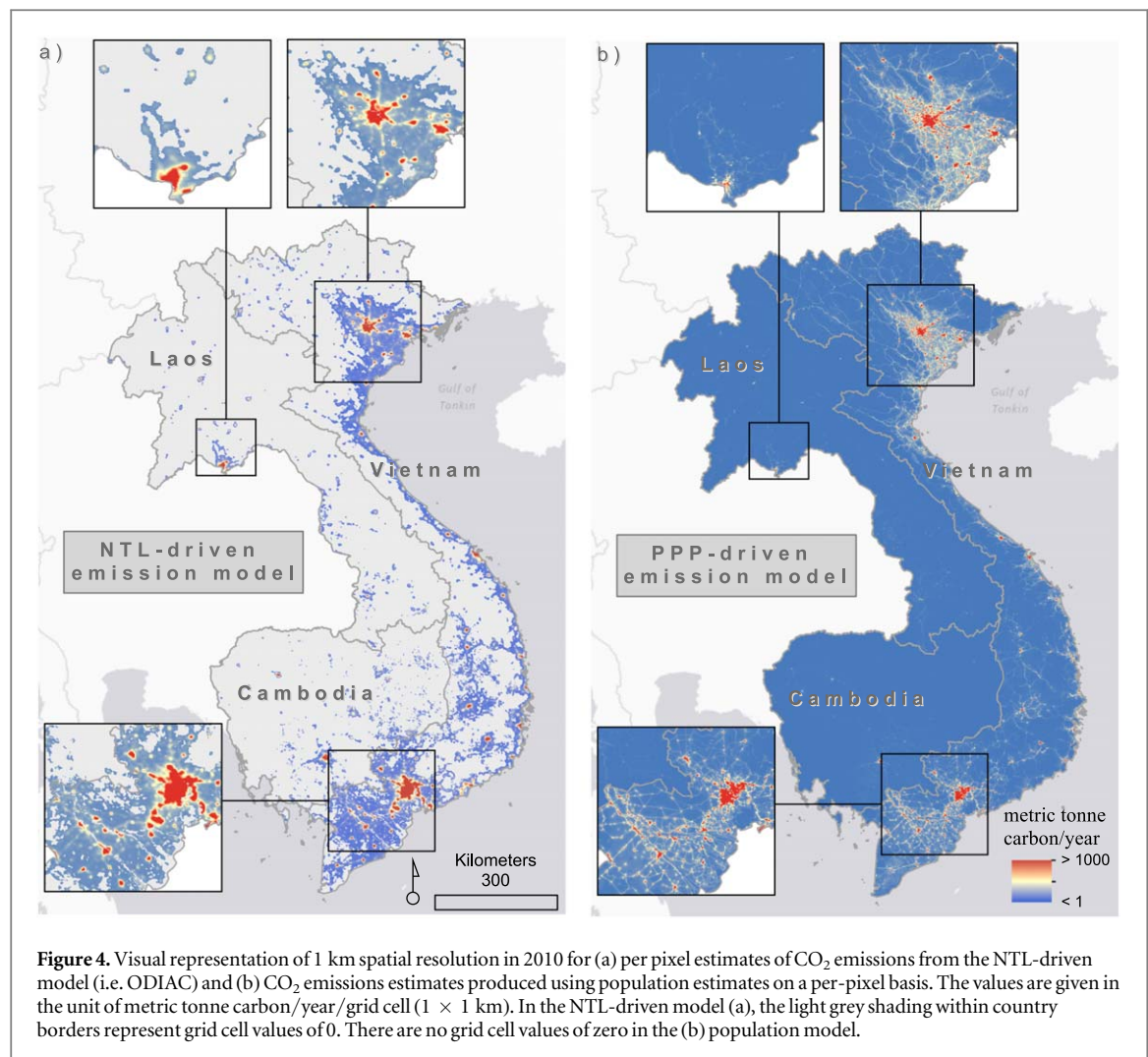
Figure 3. Per pixel counts of people for 2000–2005–2010. The four main urbanized areas are highlighted for each year (1 × 1 km).

country and year converge on similar patterns regardless of what data informs the CO_2 disaggregation. At the country level, while decreasing the spatial resolution of the two datasets especially at the first 25 km or so, there is a sharper gradient in the decrease of disagreement for Vietnam than for Laos and Cambodia. This suggests that, in general, there is a stronger correlation between the NTL and population spatial pattern in Vietnam than in the other two countries (with a better correlation in Laos than Cambodia), and that this correlation is stronger at lower spatial resolutions. In other words, the coarser the spatial resolution of the gridded CO_2 datasets, the more similar the outputs that are obtained by aggregating the NTL-based and the POP-based CO_2 grid.

Temporally, by decreasing the spatial resolution of the two gridded CO_2 datasets for Vietnam, the disagreement is lowest for 2000 than for 2005 and 2010. Quite the opposite is observed for Laos and Cambodia, with the disagreement between the two corresponding gridded CO_2 datasets generally lower for 2010 than for 2000 and 2005. In other words, while for Vietnam the outputs obtained by aggregating the NTL-based and POP-based CO_2 grids decrease 'faster' overall compared to Cambodia and Laos, the pattern across the different years suggests Vietnam has a different pattern of development and land use, which is reflected by the associations of NTL to population across time. We also note that the differences did not converge to zero. At a coarser resolution, the sum of differences is not largely impacted by the differences in small spatial patterns. We interpret the remaining differences at coarse spatial scale as a proxy bias due to the use of the NTL.

3.3. Relationships between the ODIAC-based CO_2 emissions and WorldPop population distribution for multiple subnational levels

To examine different spatial aggregations of the CO_2 emission estimates for 2000, 2005, and 2010, we sampled the NTL-based ODIAC and WorldPop population distribution outputs using different administrative unit level boundaries (i.e., $n = 11,163$ (level 3), $n = 678$ (level 2), and $n = 63$ (level 1) administrative units) and show the correlation coefficients for each level, by year, in table 2. The results highlight expected patterns for Vietnam, with improved correlation over time and improved correlation for coarser administrative unit levels (1st level is better than 2nd level, and better than 3rd level). This is consistent with results shown in figure 6 noting less disagreement when data sets are spatially coarsened. Laos and Cambodia show some deviation from this rule which we discuss further below. In addition, we plot the percent shares in the county total population to the NTL-based emissions from the ODIAC model which is located in the S1.

Figure 4. Visual representation of 1 km spatial resolution in 2010 for (a) per pixel estimates of CO_2 emissions from the NTL-driven model (i.e. ODIAC) and (b) CO_2 emissions estimates produced using population estimates on a per-pixel basis. The values are given in the unit of metric tonne carbon/year/grid cell (1 × 1 km). In the NTL-driven model (a), the light grey shading within country borders represent grid cell values of 0. There are no grid cell values of zero in the (b) population model.

We also examined the numbers of the administrative units that NTL failed to allocate CO_2 emissions for (administrative units with zero emissions) (S1). At the finest administrative level for each country, the numbers of the administrative units consistently decreased especially from 2005 to 2010. For example, 2,847 administrative units in Vietnam in 2000, which accounted for 54% of the total source area (approximated by the sum of admin unit areas) and housed 15% of the total population, were not identified by NTL (zero emissions allocated). In other words, using NTL as the only proxy data for disaggregation failed to identify those 15% of the population and the their CO_2 emissions were misallocated somewhere else. However, we believe these emissions might not be significant as this 15% people's activity do not seem to be CO_2 intensive (no lights, or not developed). In 2010, 31% of the source region remains zero emissions, but that only accounts for 6% of the population. The numbers gradually decreased to 42% in 2005 and then 31% in 2010. Cambodia showed a drastic change from 92% in 2000, 90% in 2005 and then 46% in 2010. Laos also showed the change in the similar way (62% in 2000, 63% in 2005 and 30% in 2010). This analysis is meaningful as this is not impacted by the differences in spatial patterns and population is constrained at these administrative levels. We present this as a loose estimate of error associated with NTL-based emission downscaling at the administrative unit level.

4. Discussion

Inherent in gridded emissions data like ODIAC are at least two big sources of uncertainties: (1) total emission errors and (2) spatial disaggregation errors. The analysis in this paper focuses on better understanding sources of spatial errors as it relates to subnational estimates and its direct relevance on climate mitigation policy. To examine the subnational level, we leverage another gridded data product, population data, to examine the uncertainties with the ODIAC model. The population estimates use source data that traditionally stems from censuses or surveys. That tabular information is then linked to irregular and varying sized administrative units for subnational spatial representation (Tatem *et al* 2012, 2014) and provides the source input for gridded
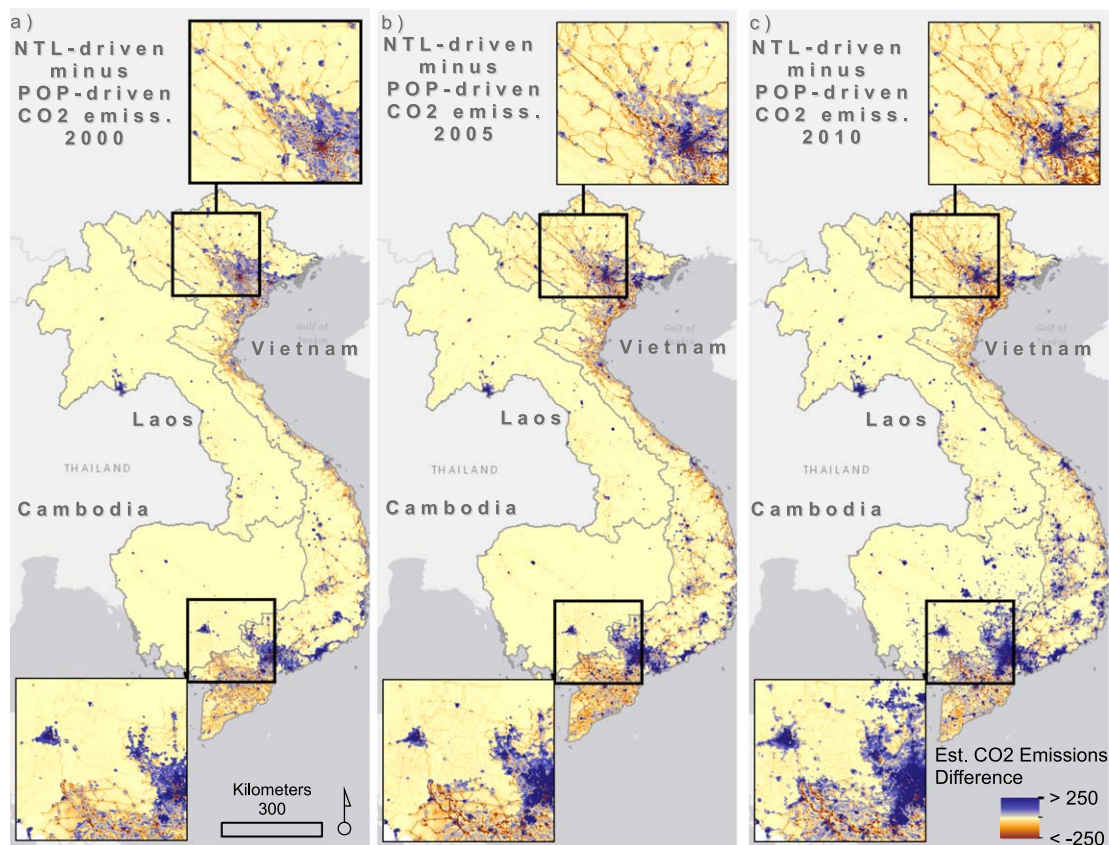
Figure 5. Per pixel differences in CO_2 emissions estimates produced using only nighttime light intensity, minus those produced using population estimates (per capita emissions). Units are expressed in tonne carbon/year/grid cell and results are separated by the years 2000 (a), 2005 (b), and 2010 (c).
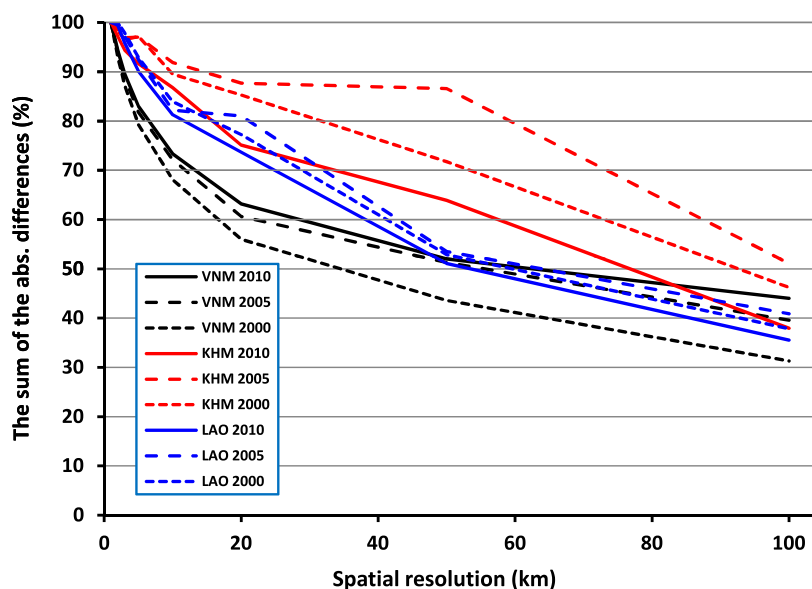

Figure 6. The level of disagreement between the two gridded CO_2 datasets by the spatial resolution (pixel size) in km versus the sum of the absolute differences at the spatial resolution divided by the sum of the absolute differences at 1 km, presented as a percentage.

population models. There will be inherent error in the gridded population data due to the uncertainty in the input data and the method and scale of model parameterization (Heuvelink 1998). However, there is still value in comparing the different data sources, in this case, the NTL-driven ODIAC model and the WorldPop population data, for estimating gridded CO_2 emissions and identifying areas of potential uncertainty in the ODIAC model.

Table 2. Correlation coefficients between the spatial aggregations of the CO_2 emission estimates based on the NTL-based ODIAC and World pop population distribution outputs for 2000, 2005, and 2010.

Administrative level	Number of units	2000	2005	2010
		Vietnam		
1st	63	0.857	0.877	0.881
2nd	678	0.665	0.709	0.729
3rd	11,163	0.585	0.653	0.708
		Laos		
1st	18	0.548	0.567	0.675
2nd	142	0.595	0.566	0.663
		Cambodia		
1st	25	0.502	0.544	0.409
2nd	178	0.548	0.653	0.431
3rd	1,576	0.419	0.471	0.245

At the various administrative unit levels, Vietnam shows the highest correlation between the distribution of lights and population (table 2, figure S5 is available online at stacks.iop.org/ERC/1/091006/mmedia), with the correlations improving across time. A higher correlation between the ODIAC model and population exists from 2000 to 2005 to 2010 due to increased development associated with more 'lit' and highly populated grid cells. (Hogue *et al* 2016, Oda *et al* 2019). In Laos and Cambodia we see low correlations across different administrative levels, with inconsistent trends temporally in their degree of association. We hypothesize this may relate to the concentration of both population and CO_2 emission estimates to the more urbanized areas for those countries and that there was strong population growth patterns outpacing changes in the emission processes during the time period analyzed (2000–2010) but further inquiry is needed to test those ideas.

For Vietnam, the fact that there is a sharper gradient of disagreement between the two gridded CO_2 datasets, especially in the first 25 km, for 2000 than for 2005 and 2010 (figure 6) could be due to the fact that the actual correlation between NTL and population distribution, at least at the 1 km grid cell level, is decreasing over time. This may be explained by the increasing use of lights for agricultural and aquacultural purposes (Chi Ling *et al* 2009). In other words, over time, lighted grid cells become increasingly associated to unpopulated grid cells corresponding, for example, to dragon fruit plantations and/or aquaculture installations (Kumari *et al* 2016). Conversely, in Laos and Cambodia, the disagreement between the two gridded CO_2 datasets decreases 'faster' for 2010 as the data is spatially coarsened than for the other two points in time suggest that the correlation between NTL and population distribution at the grid cell level is improving over time. One possible explanation might be due to faster rates of development-related electrification and urbanization-related population densification than Vietnam (Organisation for Economic Co-operation and Development 2018, United Nations 2019).

In figure 5, we see minimal differences in the use of the NTL data versus the population data as proxies for estimating subnational CO_2 emissions in less populated, more rural areas. However, the two proxies spatially and temporally differ in their disaggregation of CO_2 emissions for other areas due to the fact that the correlation is not always high between NTL and human residence. Interestingly, the NTL-based disaggregation tends to produce higher CO_2 emission estimates in more densely populated areas compared to the population-based disaggregation technique which is more liable to allocate a lower value in densely populated environments (Deville *et al* 2014, Dijkstra and Poelman 2014).

For the ODIAC model, the non-point source estimates from national CO_2 emission inventories are disaggregated solely based on NTL brightness, as captured by the DMSP-OLS sensor (Oda and Maksyutov 2011, Oda *et al* 2018). There are no subnational constraints imposed on the process of redistributing CO_2 emission estimates to 1 km × 1 km grid cells. For the corresponding population-driven estimate of CO_2 emissions, the underlying population model is parameterized based on subnational administrative units, along with a set of geospatial covariates, which influences the redistribution of CO_2 emission estimates in for each year.

The gridded population model has a more complex technique for weighting the disaggregation for non-point source estimates into 1 km × 1 km grid cells. By using an ensemble of trees, the RF approach provides flexibility for the type of data and type of relationship between predictor and response variables (section 2.3). NTL data is one possible option for inclusion in the set of covariates of the RF model, although it was excluded from the current model due to endogeneity concerns in comparing to the ODIAC data. The second part of the model, the dasymetric constraint, is based on subnational totals for population. Thus, the population model has two levels of additional information for the redistribution of non-point source estimates—(1) a pixel-level

population density weighting from the RF model and (2) a subnational constraint from administrative unit totals.

Ultimately, the reliability of the gridded population data is a function of the input population counts, which will vary on a country basis, the reliance of the use of ancillary variables with inherent error, and spatial grain of the subnational constraint applied in the model (Sinha *et al* 2019). There are multiple methods in the literature for modeling gridded population, including those datasets used in other gridded emissions studies (Leyk *et al* 2019). Each technique has its own pros and cons but consistency in the considerations for modeling across space and time is key for reliable comparisons of the gridded population products and any subsequent applications such as CO_2 emissions disaggregation models.

That said, the use of a gridded population product that has a set of underlying covariates informing the estimates for population distribution provides a useful assessment metric for examining error and uncertainty in the ODIAC model. Nighttime lights (NTL) will always be an imperfect proxy for capturing non-point source emissions associated for $FFCO_2$, and as past research has indicated, the knowledge of where people live in addition to NTL has potential to spatially refine these estimates (Oda *et al* 2019). It is particularly important, therefore, that further research leveraging finer spatial resolution emissions data, where available, be used to explore the uncertainty and sensitivity of these and similar approaches to CO_2 disaggregation.

5. Summary

Recognizing the different modeling methods that inform data proxies in the disaggregation of national CO_2 emissions estimates has important implications for the CO_2 modeling community. Findings from this study highlight differences in the NTL and population datasets to estimate subnational CO_2 emissions for a region where development and shifts in population distribution are uneven from 2000–2010 with varying growth trajectories that are country and region specific.

Geospatial data-driven techniques to disaggregate emissions is an important step forward for policy-relevant data products, even while recognizing challenges that exist in reconciling spatial and temporal considerations of data type, model application and methods used in analysis. However, more work is needed to identify the effective spatial resolution for such data and recognizing the underlying assumptions and data constraints of the geospatial proxies is paramount for judging the correct proxies to use in analysis. CO_2 gridded data are often used as an input for atmospheric CO_2 modeling. Depending on the problem setting, the spatial resolution of the modeling varies greatly from 1 km to much coarser spatial resolutions. Finally, considering that emission inventories are constructed for different compounds in a systematic way, the considerations of this study are applicable to other compounds such as air pollutants from fuel combustion.

Acknowledgments

ORCID iDs

Andrea E Gaughan ⬤ https://orcid.org/0000-0002-4898-1587

References

Andres R J, Boden T A and Higdon D M 2016 Gridded uncertainty in fossil fuel carbon dioxide emission maps, a CDIAC example *Atmos. Chem. Phys.* **16** 14979–95

Andres R J, Marland G, Fung I and Matthews E 1996 A 1° × 1°distribution of carbon dioxide emissions from fossil fuel consumption and cement manufacture, 1950-1990 *Global Biogeochem. Cycles* **10** 419–29

Boden T A, Andres R J and Marland G 2016 Global, regional, and national fossil-fuel CO_2 emissions *World Bank* **2017** 1751–2014

Breiman L 2001 Randomforest 2001 *Mach. Learn.*

Brioude J *et al* 2013 Top-down estimate of surface flux in the Los Angeles Basin using a mesoscale inverse modeling technique: assessing anthropogenic emissions of CO, NO_x and CO_2 and their impacts *Atmos. Chem. Phys.* **13** 3661–77

Chi Ling C, Yun Ham S, Chia Jen L and Yahn Chir L 2009 Effect of night illumination on growth and yield of soybean *J. Taiwan Agric. Res.* **58** 146–54 www.cabdirect.org/cabdirect/abstract/20093276425

Crowell S *et al* 2019 The 2015–2016 Carbon Cycle As Seen from OCO-2 and the Global *In Situ Network Atmos. Chem. Phys.* **19** 9797–9831 https://atmos-chem-phys-discuss.net/acp-2019-87/ Online

Deville P, Linard C, Martin S, Gilbert M, Stevens F R, Gaughan A E, Blondel V D and Tatem A J 2014 Dynamic population mapping using mobile phone data *Proc. Natl Acad. Sci. USA* **111**

Dijkstra L and Poelman H 2014 European Commission, Regional Working Paper, 2014, WP 01/2014 https://ec.europa.eu/regional_policy/sources/docgener/work/2014_01_new_urban.pdf

Dobson J E, Bright E A, Durfee R G and Worley B A 2000 LandScan: a global population database for estimating population at risk *Photogramm. Eng. Remote Sensing* **66** 849–57

Doxsey-Whitfield E, MacManus K, Squires J, Pistolesi L, Adamo S B, Baptista S R and Borkovska O 2015 Taking advantage of the improved availability of census data: a first look at the gridded population of the world, Version 4 *Pap. Appl. Geogr.* **1** 226–34

Elvidge *et al* 1999 Radiance calibration of DMSP-OLS low-light imaging data of human settlements *Remote Sens. Environ.* **68** Online https://pdfs.semanticscholar.org/4af4/227184b3add568466c629e5609c51ab9bd35.pdf

Feng L *et al* 2017 Consistent regional fluxes of CH_4 and CO_2 inferred from GOSAT proxy $XCH_4:XCO_2$ retrievals, 2010-2014 *Atmos. Chem. Phys.* **17** 4781–97

Figueres C, Quéré C L, Mahindra A, Bäte O, Whiteman G, Peters G and Guan D 2018 *Emissions are still rising: ramp up the cuts Nature* **564** 27–30 Online http://nature.com/articles/d41586-018-07585-6

Fulton L, Mejia A, Arioli M, Dematera K and Lah O 2017 Climate change mitigation pathways for Southeast Asia: CO_2 emissions reduction policies for the energy and transport sectors *Sustain.* **9** 1160

Fung I, John J, Lerner J, Matfhews E, Prather M, Steele L P and Fraser P J 1991 *Three-Dimensional Model Synthesis of the Global Methane Cycle* **96** Online https://pubs.giss.nasa.gov/docs/1991/1991_Fung_fu08000d.pdf

Ganshin A *et al* 2012 A global coupled Eulerian-Lagrangian model and 1 × 1 km CO_2 surface flux dataset for high-resolution atmospheric CO_2 transport simulations *Geosci. Model Dev.* **5** 231–43

Gately C K and Hutyra L R 2017 Large uncertainties in urban-scale carbon emissions *J. Geophys. Res. Atmos.* **122**

Gaughan A E *et al* 2016 Spatiotemporal patterns of population in mainland China, 1990 to 2010 *Sci. Data* **3** 160005

Gaughan A E, Stevens F R, Linard C, Patel N N and Tatem A J 2015 Exploring nationally and regionally defined models for large area population mapping *Int. J. Digit. Earth* **8**

Ghosh T, Elvidge C D, Sutton P C, Baugh K E, Ziskin D and Tuttle B T 2010 Creating a global grid of distributed fossil fuel CO_2 emissions from nighttime satellite imagery *Energies* **3** 1895–913

Heuvelink G B 1998 Uncertainty analysis in environmental modelling under a change of spatial scale *Nutrient Cycling in Agroecosystems* **50** 255

Hogue S, Marland E, Andres R J, Marland G and Woodard D 2016 Uncertainty in gridded CO_2 emissions estimates *Earth's Futur.* **4** 225–39

Huang Q, Yang X, Gao B, Yang Y and Zhao Y 2014 Application of DMSP/OLS nighttime light images: a meta-analysis and a systematic literature review *Remote Sens.* **6** 6844–66

Hutchins M G, Colby J D, Marland G and Marland E 2017 A comparison of five high-resolution spatially-explicit, fossil-fuel, carbon dioxide emission inventories for the United States *Mitig. Adapt. Strateg. Glob. Chang.* **22** 947

Janssens-Maenhout G *et al* 2019 EDGAR v4.3.2 Global Atlas of the three major Greenhouse Gas Emissions for the period 1970–2012 *Earth Syst. Sci. Data Discuss.* 1–55 https://earth-syst-sci-data-discuss.net/essd-2017-79/ Online

Kumari S A S M, Mannanyake M A D K, Warnasooriya W M S, Munasinghe M A L R and Kannangara K N 2016 Induced flowering with artificial light for year round production of dragon fruit in Sri Lanka *Ann. Sri Lanka Dep. Agric.* 18, 128–114

Kurokawa J, Ohara T, Morikawa T, Hanayama S, Janssens-Maenhout G, Fukui T, Kawashima K and Akimoto H 2013 Emissions of air pollutants and greenhouse gases over Asian regions during 2000-2008: Regional Emission inventory in ASia (REAS) version 2 *Atmos. Chem. Phys.* **13** 11019–11058

Lauvaux T *et al* 2016 High-resolution atmospheric inversion of urban CO_2 emissions during the dormant season of the Indianapolis flux experiment (INFLUX) *J. Geophys. Res.* **121** 5213–36

Leyk S *et al* 2019 Allocating people to pixels: a review of large-scale gridded population data products and their fitness for use *Earth Syst. Sci. Data Discuss.* (https://doi.org/10.5194/essd-2019-82)

Liaw A and Wiener M 2002 CRAN: R News Online, 2(3):18–22 https://cran.r-project.org/doc/Rnews/

Lloyd C T *et al* 2019 Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets *Big Earth Data, Rev.* **3** 108–39

Maksyutov S *et al* 2013 Regional CO_2 flux estimates for 2009–2010 based on GOSAT and ground-based CO_2 observations *Atmos. Chem. Phys.* **13** 9351–73

Marland G and Rotty R M 1984 Carbon dioxide emissions from fossil fuels: a procedure for estimation and results for 1950–1982 *Tellus B* **36** 232–61

Mennis J 2003 Generating surface models of population using dasymetric mapping *Prof. Geogr.* **55** 31–42

Nassar R, Hill T G, McLinden C A, Wunch D, Jones D B A and Crisp D 2017 Quantifying CO_2 emissions from individual power plants from space *Geophys. Res. Lett.* **44** 10045–10053

Oda T *et al* 2012 The use of a high-resolution emission dataset in a Global Eulerian-Lagrangian coupled model, 'Lagrangian Modeling of the Atmosphere' *AGU Geophysical Monograph Series* (https://doi.org/10.1029/2012GM001263)

Oda T *et al* 2019 Assessing errors and uncertainties in a global high-resolution gridded fossil fuel carbon dioxide emission inventory: implications for the use of spatially-explicit emissions in transport model simulations, top-down emission estimations, and climate mitig *Mitig. Adapt. Strateg. Glob. Chang.* **24**

Oda T, Lauvaux T, Lu D, Rao P, Miles N L, Richardson S J and Gurney K R 2017 On the impact of granularity of space-based urban CO_2 emissions in urban atmospheric inversions: a case study for Indianapolis IN *Elem. Sci. Anth.* **5**

Oda T and Maksyutov S 2011 A very high-resolution (1km×1km) global fossil fuel CO_2 emission inventory derived using a point source database and satellite observations of nighttime lights *Atmos. Chem. Phys.* **11** 543–56

Oda T and Maksyutov S Odiac - Fossil fuel CO_2 emissions dataset - Home *Cent. Glob. Environ. Res. Natl. Inst. Environ. Stud.* Online http://db.cger.nies.go.jp/dataset/ODIAC/

Oda T, Maksyutov S and Andres R J 2018 The open-source data inventory for anthropogenic CO_2, version 2016 (ODIAC2016): a global monthly fossil fuel CO_2 gridded emissions data product for tracer transport simulations and surface flux inversions *Earth Syst. Sci. Data* **10** 87–107

Oda T, Maksyutov S and Elvidge C D 2010 Disaggregation of national fossil fuel CO_2 emissions using a global power plant database and DMSP nightlight data *Proc. Asia-Pacific Adv. Netw.* **30**

Organisation for Economic Co-operation and Development OECD 2018 Economic Outlook for Southeast Asia, China and India 2019: Towards Smart Urban Transportation, OECD Publishing, Paris https://www.oecd.org/dev/economic-outlook-for-southeast-asia-china-and-india-23101113.html Online (https://doi.org/10.1787/saeo-2019-en)

Pandey B, Zhang Q and Seto K C 2017 Comparative evaluation of relative calibration methods for DMSP/OLS nighttime lights *Remote Sens. Environ.* **195** 67–78

Raupach M R, Marland G, Ciais P, Le Quéré C, Canadell J G, Klepper G and Field C B 2007 Global and regional drivers of accelerating CO_2 emissions *Proc. Natl Acad. Sci. USA* **104** 10288–93 Online http://ncbi.nlm.nih.gov/pubmed/17519334

Raupach M R, Rayner P J and Paget M 2010 Regional variations in spatial structure of nightlights, population density and fossil-fuel CO_2 emissions *Energy Policy* **38** 4756–64

Rayner P J, Raupach M R, Paget M, Peylin P and Koffi E 2010 A new global gridded data set of CO_2 emissions from fossil fuel combustion: methodology and evaluation *J. Geophys. Res. Atmos.* **115**

Reed F, Gaughan A, Stevens F, Yetman G, Sorichetta A and Tatem A 2018 Gridded population maps informed by different built settlement products *Data* **3** 33

Reuter M, Buchwitz M, Schneising O, Krautwurst S, O'Dell C W, Richter A, Bovensmann H and Burrows J P 2019 Towards monitoring localized CO_2 emissions from space: co-located regional CO_2 and NO_2 enhancements observed by the OCO-2 and S5P satellites *Atmos. Chem. Phys. Discuss.* **19** 9371–83 https://atmos-chem-phys-discuss.net/acp-2019-15/ Online

Schneider A *et al* 2015 A new urban landscape in East-Southeast Asia, 2000-2010 *Environ. Res. Lett.* **10** 034002

Sinha P, Gaughan A E, Stevens F R, Nieves J J, Sorichetta A and Tatem A J 2019 Assessing the spatial sensitivity of a random forest model: Application in gridded population modeling *Comput. Environ. Urban* **75** 132–45

Sorichetta A, Hornby G M, Stevens F R, Gaughan A E, Linard C and Tatem A J 2015 High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020 *Sci. Data* **2**

Stevens F R, Gaughan A E, Linard C and Tatem A J 2015 Disaggregating census data for population mapping using Random forests with remotely-sensed and ancillary data *PLoS One* **10**

Strobl C, Malley J and Tutz G 2009 An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests *Psychol Methods.* **14** 323–48

Sutton P, Roberts D, Elvidge C and Baugh K 2001 Census from Heaven: an estimate of the global human population using night-time satellite imagery *Int. J. Remote Sensing.* **22** 3061–76

Takagi H *et al* 2011 On the benefit of GOSAT observations to the estimation of regional CO_2 fluxes *Sola* **7**

Tatem A J, Adamo S, Bharti N, Burgert C R, Castro M and Dorelien A 2012 Mapping populations at risk: improving spatial demographic data for infectious disease modeling and metric derivation *Popul Health Metr.* **10**

Tatem A J *et al* 2014 Mapping the denominator: spatial demography in the measurement of progress *Int. Health* **6** 153–55

United Nations 2019 [World population prospects 2019] *United Nations. Dep. Econ. Soc. Aff. World Popul. Prospect. 2019.* https://population.un.org/wpp/ Last accessed March 2019

Vogel C, Moser S C, Kasperson R E and Dabelko G D 2007 Linking vulnerability, adaptation, and resilience science to practice: pathways, players, and partnerships *Glob. Environ. Chang.* **17** 349–64

WDPA 2017 Protected Planet Online https://protectedplanet.net/c/world-database-on-protected-areas

Wu D, Lin J C, Oda T, Ye X, Lauvaux T, Yang E G and Kort E A 2018 A lagrangian approach towards extracting Signals of Urban CO_2; emissions from Satellite Observations of Atmospheric Column CO_2 (XCO_2): X-St *Geosci. Model Dev.* **11** 4843–71

www.worldpop.org 2018 WorldPop, global high resolution population denominators project *Glob. High Resolut. Popul. Denominators Proj. Sch. Geogr. Environ. Sci. Univ. Southampt. Dep. Geogr. Geosci. Univ. Louisv. Département Géographie, Univ. Namur Cent.* Online https://worldpop.org/ Last accessed: Sep 2018

Ziskin D, Baugh K, Hsu F-C and Elvidge C D 2012 Methods used for the 2006 radiance lights *Proc. Asia-Pacific Adv. Netw.* 131–42